Today: Chapter 11, confidence intervals for means

Announcements

• Useful summary tables:
  Sampling distributions: p. 353
  Confidence intervals: p. 439
  Hypothesis tests: p. 534
• Homework assigned today and Wed, due Friday.
• Final exam seat assignments will be sent soon.

Homework: (Due Fri March 15)
Chapter 11: #30bc, 48, 86*.
*Use R Commander for #86. Data file linked to website. Both 48 and 86 count double, 2 points each.

Overview from when we started Chapter 9:

<table>
<thead>
<tr>
<th>Parameter name and description</th>
<th>Population parameter</th>
<th>Sample statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>For Categorical Variables:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One population proportion (or probability)</td>
<td>p</td>
<td>( \hat{p} )</td>
</tr>
<tr>
<td>Difference in two population proportions</td>
<td>( p_1 - p_2 )</td>
<td>( \hat{p}_1 - \hat{p}_2 )</td>
</tr>
<tr>
<td>For Quantitative Variables:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One population mean</td>
<td>( \mu )</td>
<td>( \bar{x} )</td>
</tr>
<tr>
<td>Population mean of paired differences (dependent samples, paired)</td>
<td>( \mu_d )</td>
<td>( \bar{d} )</td>
</tr>
<tr>
<td>Difference in two population means (independent samples)</td>
<td>( \mu_1 - \mu_2 )</td>
<td>( \bar{x}_1 - \bar{x}_2 )</td>
</tr>
</tbody>
</table>

For each situation we will:
- Learn about the sampling distribution for the sample statistic
- Learn how to find a confidence interval for the true value of the parameter
- Test hypotheses about the true value of the parameter

Recall:

- A parameter is a population characteristic – value is usually unknown. We estimate the parameter using sample information. Chapter 11: C.I.s for means.
- A statistic, or estimate, is a characteristic of a sample. A statistic estimates a parameter.
- A confidence interval is an interval of values computed from sample data that is likely to include the true population value.
- The confidence level for an interval describes our confidence in the procedure we used. We are confident that most of the confidence intervals we compute using our procedure will include the true population value.

Recall from Chapter 10

A confidence interval or interval estimate for any of the five parameters can be expressed as

\[
\text{Sample estimate} \pm \text{multiplier} \times \text{standard error}
\]

where the multiplier is a number based on the confidence level desired and determined from the standard normal (z) distribution (for proportions) or Student’s t-distribution (for means).

\[
\text{Sample estimate} = \text{sample statistic.}
\]

Three Estimation Situations Involving Means

Situation 1. Mean of a quantitative variable.

Examples:
- What is the mean number of facebook friends UCI students have (for those on facebook)?
- What is the mean number of words a 2-year old knows?

Population parameter: \( \mu \) (spelled “mu” and pronounced “mew”) = population mean for the variable

Sample estimate: \( \bar{x} = \) sample mean for the variable, based on a sample of size \( n \).
Estimating the Population Mean of Paired Differences

**Situation 2.** Data measured in pairs, take differences, estimate the mean of the population of differences:
- What is the mean difference in blood pressure before and after learning meditation? ($d_i =$ difference for person $i$)
- What is the mean difference in hours/day spent studying and spent watching television for college students?

Population parameter: $\mu_d$ (called “mu” $d$)

Sample estimate: $\overline{d}$ = the sample mean for the differences, based on a sample of $n$ pairs, where the difference is computed for each pair.

Recall from Chapter 10

A **confidence interval** or **interval estimate** for any of the five parameters can be expressed as

Sample estimate $\pm$ multiplier $\times$ standard error

**Situation 1:** Standard Error of the Mean

$$s.d. (\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

In practice:
- We don’t know $\sigma$, so we estimate it using $s$.
- Replacing $\sigma$ with $s$ in the standard deviation expression gives us an estimate that is called the [standard error of $\overline{X}$](#).

$$s.e. (\overline{X}) = \frac{s}{\sqrt{n}}$$

Chapter 9 weight loss example:
$n = 25$ weight losses, $\sigma = 5$;  $s.d. (\overline{X}) = \frac{5}{\sqrt{25}} = 1$ pound

Suppose sample standard deviation is $s = 4.74$ pounds.
So the standard error of the mean is $4.74/5 = 0.948$ pounds.

Difference in two means

**Situation 3.** Estimating the difference between two population means for independent samples

Examples:
- How much difference is there in the means of what male students and female students expect to earn as a starting salary after graduation? (Question on 2011 class survey.)
- How much difference is there in the mean IQs for children whose moms smoked and didn’t during pregnancy?

Population parameter: $\mu_1 - \mu_2 =$ difference between the two population means.

Sample estimate: $\overline{X}_1 - \overline{X}_2 =$ difference between the two sample means. This requires independent samples.

Standard errors (in general)

**Rough Definition:** The **standard error** of a sample statistic measures, roughly, the **average difference** between the statistic and the population parameter. This “average difference” is over all possible random samples of a given size that can be taken from the population.

**Technical Definition:** The **standard error** of a sample statistic is the **estimated standard deviation** of the sampling distribution for the statistic.

Increasing the Size of the Sample

Suppose we take $n = 100$ weight losses instead of just 25.
The standard deviation of the mean would be

$$s.d. (\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = 0.5 \text{ pounds.}$$

- For samples of $n = 25$, sample means are likely to range between $8 \pm 3$ pounds => 5 to 11 pounds.
- For samples of $n = 100$, sample means are likely to range only between $8 \pm 1.5$ pounds => 6.5 to 9.5 pounds.

Larger samples tend to result in more accurate estimates of population values than smaller samples.
Standard Error of a Sample Mean

\[ s.e(\bar{x}) = \frac{s}{\sqrt{n}}, \quad s = \text{sample standard deviation} \]

Example: Mean number of Facebook friends

Class Survey, Winter 2011: Only those on Facebook.

\[ \text{About how many Facebook friends do you have?} \]

Minitab provides \( s.e \), R Commander doesn’t, but provides \( s \)

Statistics \( \rightarrow \) Summaries \( \rightarrow \) Numerical summaries, check Standard Deviation

Minitab for the example:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>StDev</th>
<th>SE</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>257</td>
<td>462.0</td>
<td>404.0</td>
<td>301.1</td>
<td>18.8</td>
<td></td>
</tr>
</tbody>
</table>

\[ s.e(\bar{x}) = \frac{301.1}{\sqrt{257}} = 18.8 \]

Situation 2: Standard Error of the mean of paired differences

\[ s.e(\bar{d}) = \frac{s_d}{\sqrt{n}} \]

where \( s_d = \text{sample standard deviation for the differences} \)

Example: How much taller (or shorter) are daughters than their mothers these days? \( s_d = 3.14 \) (for individuals)

\[ n = 93 \text{ pairs (daughter – mother)} \quad \bar{d} = 1.3 \text{ inches} \]

\[ s.e(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{3.14}{\sqrt{93}} = 0.33 \]

Recall from Chapter 10

A confidence interval or interval estimate for any of the five parameters can be expressed as

\[ \text{Sample estimate} \pm \text{multiplier} \times \text{standard error} \]

The multiplier is a number based on the confidence level desired and determined from the standard normal distribution (for proportions) or Student’s \( t \)-distribution (for means).

Student’s \( t \)-Distribution: Replacing \( \sigma \) with \( s \)

Dilemma: we generally don’t know \( \sigma \). Using \( s \) we have:

\[ t = \frac{\bar{x} - \mu}{s.e(\bar{x})} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n(\bar{x} - \mu)}}{s} \]

If the sample size \( n \) is small, this standardized statistic will not have a N(0,1) distribution but rather a \( t \)-distribution with \( n - 1 \) degrees of freedom (df).

NOTE: Use \( t^* \) for all 3 situations involving means, but different df formula for two independent samples.

Finding the \( t \)-multiplier

- Excel: See page 412.
- R Commander:
  - Distributions \( \rightarrow \) Continuous distributions \( \rightarrow \) t distribution \( \rightarrow \) t quantiles
  - Example: 95% CI for mean when \( n = 10 \)
    - Probabilities: \( \alpha/2 \) (for 95%, use .025)
    - Degrees of freedom (\( n = 10 \), so \( df = 9 \))
    - Lower tail
    - Gives negative of the t-multiplier
    - Ex: .025, 9, lower tail \( \rightarrow -2.262157 \), multiplier \( \approx -2.26 \)
  - Table A.2 (see page 411 for instructions)
  - Table A.2 is on page 670 or turn page inside back cover (easy to use compared to z!)}
95% C.I. for Mean Anticipated Starting Salary

Data from 2011 survey, what do you expect your starting salary to be after you graduate? (or after grad/prof school?)

\[ n = 244 \text{ (Some outliers at $200K, $250K, $500K)} \]

Sample mean = $63,075

Sample standard deviation = $46,607

Standard error of the mean = \[ \frac{46,607}{\sqrt{244}} = 2,984 \]

Multiplier = \( t^* \) with df of 100 = 1.98 (closest in Table A.2)

Sample estimate ± multiplier × standard error

\[ 63,075 ± 1.98 × 2,984 \]

\[ 63,075 ± 5908 \]

$57,167 to $68,983

C.I.s for some other survey Qs

<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>Facebook</td>
<td>257</td>
<td>462.0</td>
<td>301.1</td>
<td>18.8</td>
<td>(425.0, 499.0)</td>
</tr>
<tr>
<td>Income2010</td>
<td>260</td>
<td>3531</td>
<td>7197</td>
<td>446</td>
<td>( 2652,  4410)</td>
</tr>
<tr>
<td>StudentLoans</td>
<td>242</td>
<td>17529</td>
<td>33973</td>
<td>2184</td>
<td>(13227, 21831)</td>
</tr>
<tr>
<td>HoursStudy</td>
<td>264</td>
<td>5.362</td>
<td>4.203</td>
<td>0.259</td>
<td>(4.852, 5.871)</td>
</tr>
</tbody>
</table>

Note the extremely large standard deviations for all of these. Obviously they are not bell-shaped variables!

Paired Data Confidence Interval

Data: two variables for \( n \) individuals or pairs; use the difference \( d = x_1 - x_2 \).

Population parameter: \( \mu_d = \text{mean of differences for the population (same as } \mu_1 - \mu_2 \) .

Sample estimate: \( \bar{d} = \text{sample mean of the differences} \)

Standard deviation and standard error:

\[ s_d = \text{standard deviation of the sample of differences;} \]

\[ s.e(\bar{d}) = \frac{s_d}{\sqrt{n}} \]

Confidence interval for \( \mu_d \):

\[ \bar{d} ± t^* \times s.e(\bar{d}) \text{,} \]

where df = \( n - 1 \) for the multiplier \( t^* \).
Find 90% C.I. for difference: (daughter – mother) height difference

How much taller (or shorter) are daughters than their mothers these days?

\[ n = 93 \text{ pairs (daughter – mother), } \bar{d} = 1.3 \text{ inches} \]

\[ s_d = 3.14 \text{ inches, so } s.d.(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{3.14}{\sqrt{93}} = .33 \]

\[ \text{Multiplier } = t^* \text{ with 92 df for 90% C.I. = 1.66 (use df=90)} \]

\[ \frac{1.3}{0.33} = 3.93 \]

Sample estimate \( \pm \) multiplier \( \times \) standard error

\[ 1.3 \pm 1.66 \times 0.33 \]

\[ 1.3 \pm 0.55 \]

0.75 to 1.85 inches (does not cover 0)

Confidence interval interpretations

• We are 95% confident that the mean study hours per week for Stat 7, for all students over all time (who would complete a survey??) is between 4.85 and 5.87 hours.

• We are 90% confident that the mean height difference between female college students and their mothers is between 0.75 and 1.85 inches, with students being taller than their mothers.

Example: Small sample, so check for outliers

Data: Hours spent studying for those students who attend class 0 or 1 times a week; \( n = 10 \) students.

Create a 95% CI for study hours for students who don’t attend class. Small \( n \), so check for skewness and outliers.

Note: Boxplot shows no obvious skewness and no outliers.

Example, continued (study hours)

Results:

\[ \bar{x} = 3.7, s = 1.89, \text{ and } s.d.(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{1.89}{\sqrt{10}} = 0.60 \]

Multiplier \( t^* \) from Table A.2 with \( df = 9 \) is \( t^* = 2.26 \)

95% Confidence Interval:

\[ 3.7 \pm 2.26(0.6) \Rightarrow 3.7 \pm 1.36 \Rightarrow 2.34 \text{ to } 5.06 \text{ hours} \]

Interpretation: We are 95% confident that the mean of the study hours per week for Stat 7 for students who don’t attend class (and are represented by this sample) is covered by the interval from 2.34 to 5.06 hours per week.

(Compare to 95% C.I. for everyone, 4.85 to 5.87 hours.)

11.4 CI for Difference Between Two Means (Independent samples)

A CI for the Difference Between Two Means (Independent Samples, unpooled case):

\[ x_1 - x_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]

where \( t^* \) is the value in a \( t \)-distribution with area between \(-t^*\) and \( t^* \) equal to the desired confidence level.

Necessary Conditions

• Two samples must be independent.

Either …

• Populations of measurements both bell-shaped, and random samples of any size are measured.

or …

• Large (\( n \geq 30 \)) random samples are measured.
Degrees of Freedom

The t-distribution is only approximately correct and df formula is complicated (Welch’s approx):

\[
\frac{1}{n_1-1} \left( \frac{s_1^2}{n_1} \right) + \frac{1}{n_2-1} \left( \frac{s_2^2}{n_2} \right)
\]

Statistical software can use the above approximation, but if done by hand then use a conservative df = smaller of \((n_1 - 1)\) and \((n_2 - 1)\).

Example: Anticipated Starting Salary for Men/Women

Two-sample T for StartSalary

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>87</td>
<td>69356</td>
<td>44937</td>
<td>4818</td>
</tr>
<tr>
<td>Women</td>
<td>156</td>
<td>56772</td>
<td>31485</td>
<td>2521</td>
</tr>
</tbody>
</table>

Difference = mu (Men) - mu (Women)

Estimate for difference: 12585, df = 133

95% CI for difference: (1830, 23340)

\[
\tau_1 - \tau_2 \pm t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 12,585 \pm 87 \sqrt{\frac{44937^2}{87} + \frac{31485^2}{156}}
\]

Example 11.13 Diet vs Exercise

Study: \(n_1 = 42\) men on diet, \(n_2 = 47\) men exercise

Diet: Lost an average of 7.2 kg with std dev of 3.7 kg
Exercise: Lost an average of 4.0 kg with std dev of 3.9 kg

So, \(\bar{x}_1 - \bar{x}_2 = 7.2 - 4.0 = 3.2\) kg and \(s.e.(\bar{x}_1 - \bar{x}_2) = 0.81\) kg

Approximate 95% Confidence Interval:

\(3.2 \pm 2(0.81) \Rightarrow 3.2 \pm 1.62 \Rightarrow 1.58 \text{ to } 4.82\) kg

Note: We are 95% confident the interval 1.58 to 4.82 kg covers the additional mean weight loss for dieters compared to those who exercised. The interval does not cover 0, so a real difference is likely to hold for the population as well.

Using R Commander

Does tests and C.I.s in same step

- Data → New data set – give name, enter data
- One column for Drink or Don’t drink, one for Study hours
- Statistics → Means → Independent samples t-test
- Choose the alternative (≠, >, <) and confidence level

Welch Two Sample t-test

data: HoursStudy by Drink
t = 1.9923, df = 70.556, p-value = 0.05021
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:

\(-0.001651343 \text{ to } 3.450180754\)
sample estimates:

\[\text{mean in group Don’t drink} \quad \text{mean in group Drink} \]

\[6.500000 \quad 4.775735\]
Should check for outliers if small sample(s)
Graphs→Boxplot→Plot by Group
These are large samples, fortunately, because very skewed!
But still looks like those who drink have fewer study hours.

Confidence interval applet:
Illustrates the same concept as the hands-on team project last Friday.

http://www.rossmanchance.com/applets/Confsim/Confsim.html
http://www.rossmanchance.com/applets/NewConfsim/Confsim.html