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.

Chapter 11

Estimating Means with Confidence

Overview from when we started Chapter 9:

Parameter name and description	**Population parameter**	**Sample statistic**
For Categorical Variables:		
One population proportion (or probability)	p	$\hat{p}$
Difference in two population proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
For Quantitative Variables:		
One population mean	μ	$\overline{x}$
Population mean of paired differences (dependent samples, paired)	μ_d	$\overline{d}$
Difference in two population means (independent samples)	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$

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

Sample estimate ± multiplier × 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).

Sample estimate = 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*:** μ (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**:** $\boldsymbol{\mu_d}$ (called "*mu*" *d*)

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

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**:** $\bar{x}_1 - \bar{x}_2$ = difference between the two sample means. This requires *independent* samples.

Recall from Chapter 10

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

Sample estimate ± multiplier × standard error

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.

Situation 1: Standard Error of the Mean

$$\text{s.d.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

In practice:

•We don't know σ, so we estimate it using s.

•Replacing σ with s in the standard deviation expression gives us an estimate that is called the **standard error of** $\bar{x}$.

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Chapter 9 weight loss example:

n = 25 weight losses, σ = 5; $s.d.(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}}$ = 1 pound

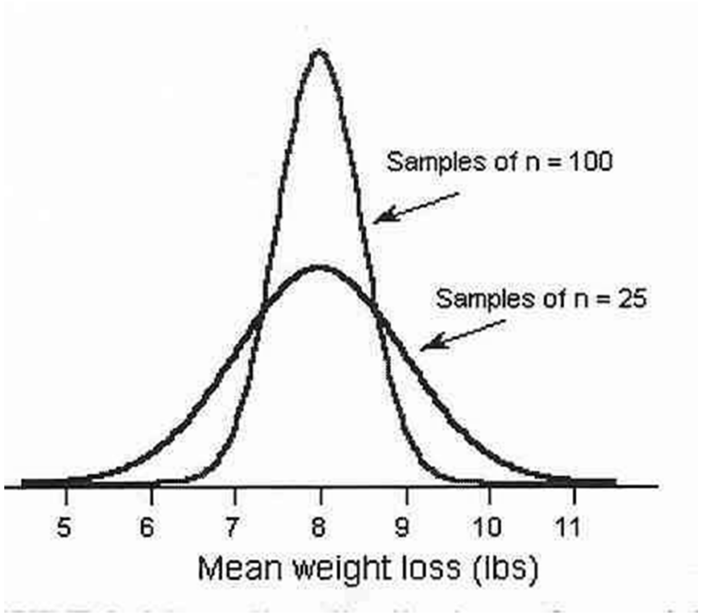
Suppose *sample* standard deviation is s = 4.74 pounds.
So the standard *error* of the mean is 4.74/5 = 0.948 pounds.

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

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

- For samples of ***n* = 25**, sample means are likely to range between 8 ± 3 pounds => 5 to 11 pounds.
- For samples of ***n* = 100**, sample means are likely to range only between 8 ± 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.

About how many Facebook friends do you have?

Minitab provides **s.e.**, R Commander doesn't, but provides ***s***

Statistics → Summaries → Numerical summaries, check Standard Deviation

Minitab for the example:

Variable	N	Mean	Median	**StDev**	**SE Mean**
Facebook	257	462.0	404.0	301.1	18.8

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

Situation 2: Standard Error of the mean of paired differences

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

where s_d = sample standard deviation for the *differences*

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

n = 93 pairs (daughter – mother) $\bar{d}$ = 1.3 inches

$$s.e.(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{3.14}{\sqrt{93}} = .33$$

Situation 3: Standard Error of the Difference Between Two Sample Means (*unpooled)*

$$s.e.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 11.3 *Lose More Weight by Diet or Exercise?*

Study: n_1 = 42 men on diet, n_2 = 47 men on exercise routine

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 $$s.e.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{(3.7)^2}{42} + \frac{(3.9)^2}{47}} = 0.81$$

Recall from Chapter 10

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

Sample estimate ± multiplier × standard error

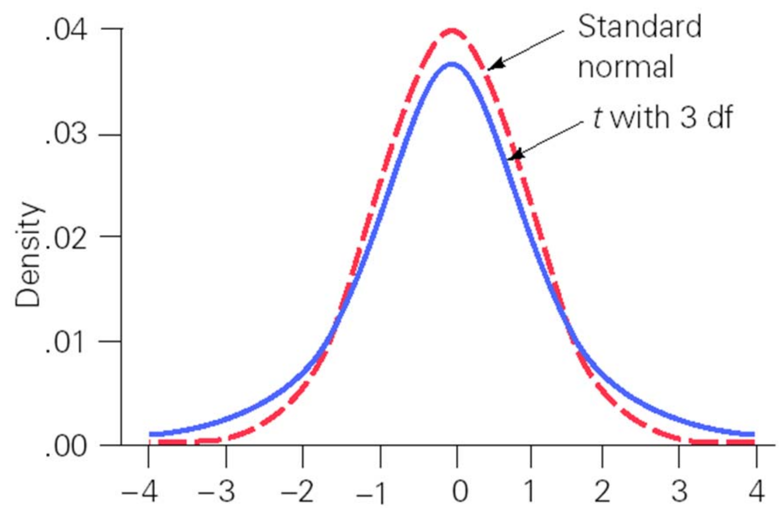
The **multiplier** is a number based on the confidence level desired and determined from the standard normal distribution (for proportions) or Student's *t*-distribution (for means).

Student's *t*-Distribution: Replacing σ with *s*

Dilemma: we generally don't know σ. 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.

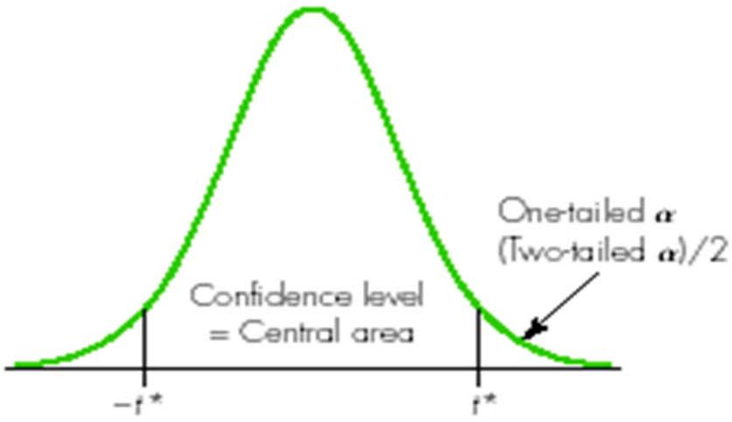
Finding the *t*-multiplier

- Excel: See page 412.
- R Commander:

  Distributions → Continuous distributions → t distribution → 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 → −2.262157, multiplier ≈ 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!)

Table A.2 t^* Multipliers for Confidence Intervals and Rejection Region Critical Values



Example: df = 9
95% confidence
t* = 2.26

	Confidence Level						
df	*.80*	*.90*	*.95*	*.98*	*.99*	*.998*	*.999*
1	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	1.89	2.92	4.30	6.96	9.92	22.33	31.60
3	1.64	2.35	3.18	4.54	5.84	10.21	12.92
4	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	1.48	2.02	2.57	3.36	4.03	5.89	6.87
6	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	1.41	1.89	2.36	3.00	3.50	4.79	5.41
8	1.40	1.86	2.31	2.90	3.36	4.50	5.04
9	1.38	1.83	2.26	2.82	3.25	4.30	4.78
10	1.37	1.81	2.23	2.76	3.17	4.14	4.59

etc…

Confidence Interval for One Mean *or* Paired Data

A Confidence Interval for a Population Mean

$$\bar{x} \pm t^* \times s.e.(\bar{x}) \quad \Rightarrow \quad \bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

where the **multiplier *t**** is the value in a *t*-distribution with degrees of freedom = df = n - 1 such that the area between -t^* and t^* equals the desired confidence level. (Found from Excel, R Commander or Table A.2.)

Conditions:

- Population of measurements **is bell-shaped** (no major skewness or outliers) and r.s. of any size > 2; OR
- Population of measurements is ***not*** bell-shaped, **but** a ***large*** random **sample** is measured, $n \geq 30$.

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 (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)

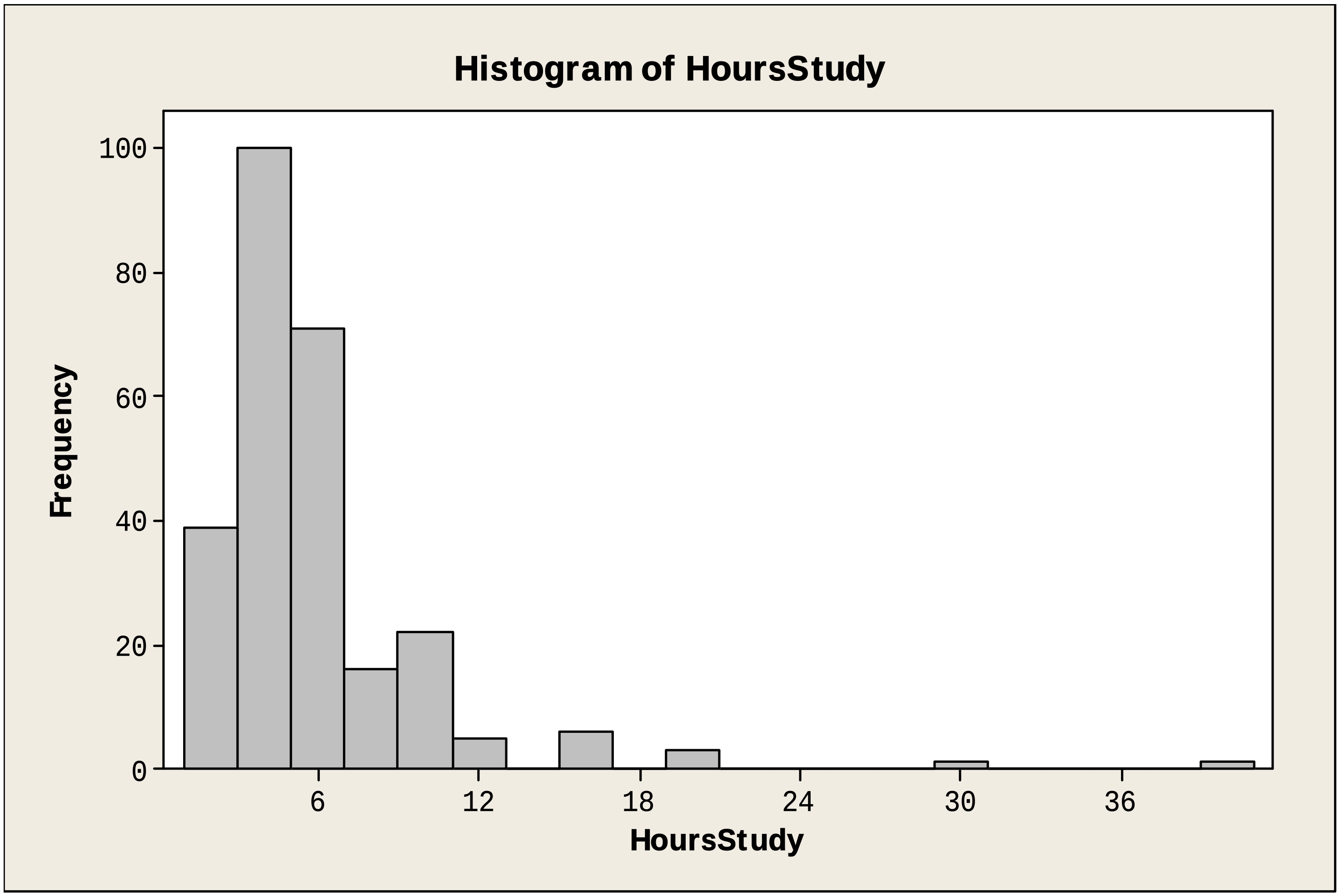
Sample estimate ± multiplier × standard error
63,075 ± 1.98 × 2984
63,075 ± 5908
\$57,167 to \$68,983

C.I.s for some other survey Qs

Variable	N	Mean	StDev	SE Mean	95% CI
Facebook	257	462.0	301.1	18.8	(425.0, 499.0)
Income2010	260	3531	7197	446	(2652, 4410)
StudentLoans	242	17529	33973	2184	(13227, 21831)
HoursStudy	264	5.362	4.203	0.259	(4.852, 5.871)

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

Example: Histogram for study hours



Paired Data Confidence Interval

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

Population parameter: μ_d = mean of differences for the population (same as $\mu_1 - \mu_2$).

Sample estimate: $\bar{d}$ = sample mean of the differences

Standard deviation and standard error:

s_d = standard deviation of the sample of differences;

$$s.e.\left(\bar{d}\right) = \frac{s_d}{\sqrt{n}}$$

Confidence interval for μ_d: $\bar{d} \pm t^* \times s.e.\left(\bar{d}\right)$,

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 pairs (daughter – mother), $\bar{d}$ = 1.3 inches

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

Multiplier = t^* with 92 df for 90% C.I. = 1.66 (use df=90)

Sample estimate ± multiplier × standard error

1.3 ± 1.66 × 0.33

1.3 ± 0.55

0.75 to 1.85 inches (does *not* cover 0)

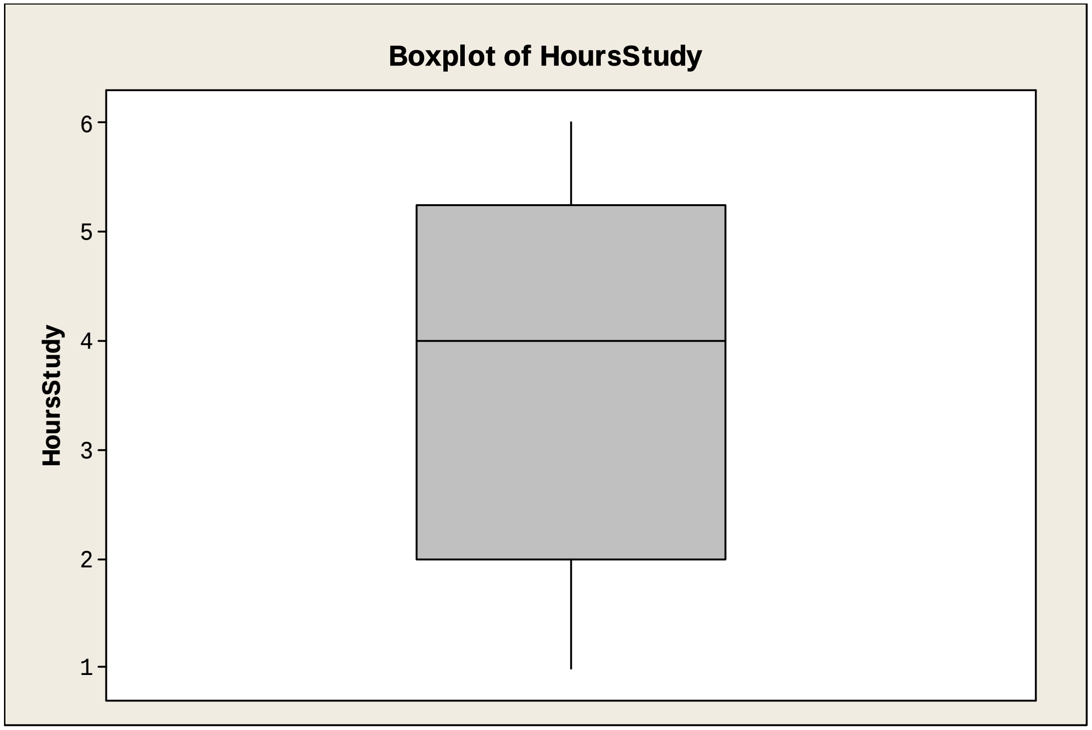
Confidence interval interpretations

- 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:

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

Multiplier t^* from Table A.2 with df = 9 is $t^* = 2.26$

95% Confidence Interval:

$3.7 \pm 2.26(0.6)$ => 3.7 ± 1.36 => 2.34 to 5.06 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, <u>unpooled</u> case):

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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):

$$\text{df} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

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***

(Minitab output)

```
Two-sample T for StartSalary
Group      N    Mean   StDev   SE Mean
Men       87   69356   44937      4818
Women    156   56772   31485      2521
Difference = mu (Men) - mu (Women)
Estimate for difference:  12585, df = 133
95% CI for difference:  (1830, 23340)
```

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \Rightarrow \quad 12{,}585 \pm 1.98\sqrt{\frac{(44937)^2}{87} + \frac{(31485)^2}{156}}$$

Interpretation: We are 95% certain that the mean anticipated starting salary for men is between \$1830 and \$23,340 *higher* then the mean anticipated starting salary for women, for the population of students represented by this sample.

Approximate 95% CI

For <u>sufficiently large samples</u>, the interval

Sample estimate ± 2 × Standard error

is an **approximate 95% confidence interval** for a population parameter.

Note: Except for very small degrees of freedom, the multiplier t^* for 95% confidence interval is close to 2. So, 2 is often used to approximate, rather than finding degrees of freedom. For instance, for 95% C.I.:

$t^*(30) = 2.04$, $t^*(60) = 2.00$, $t^*(90) = 1.99$, $t^*(\infty) = z^* = 1.96$

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 ± 2(.81) => 3.2 ± 1.62 => 1.58 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

- Read in or enter data set
- Statistics → Means →
  - Single sample t-test
  - Independent samples t-test (requires data in one column, and group code in another)
  - Paired t-test (requires the original data in two separate columns)

Example: Compare study hours for those who drink and those who don't

- 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

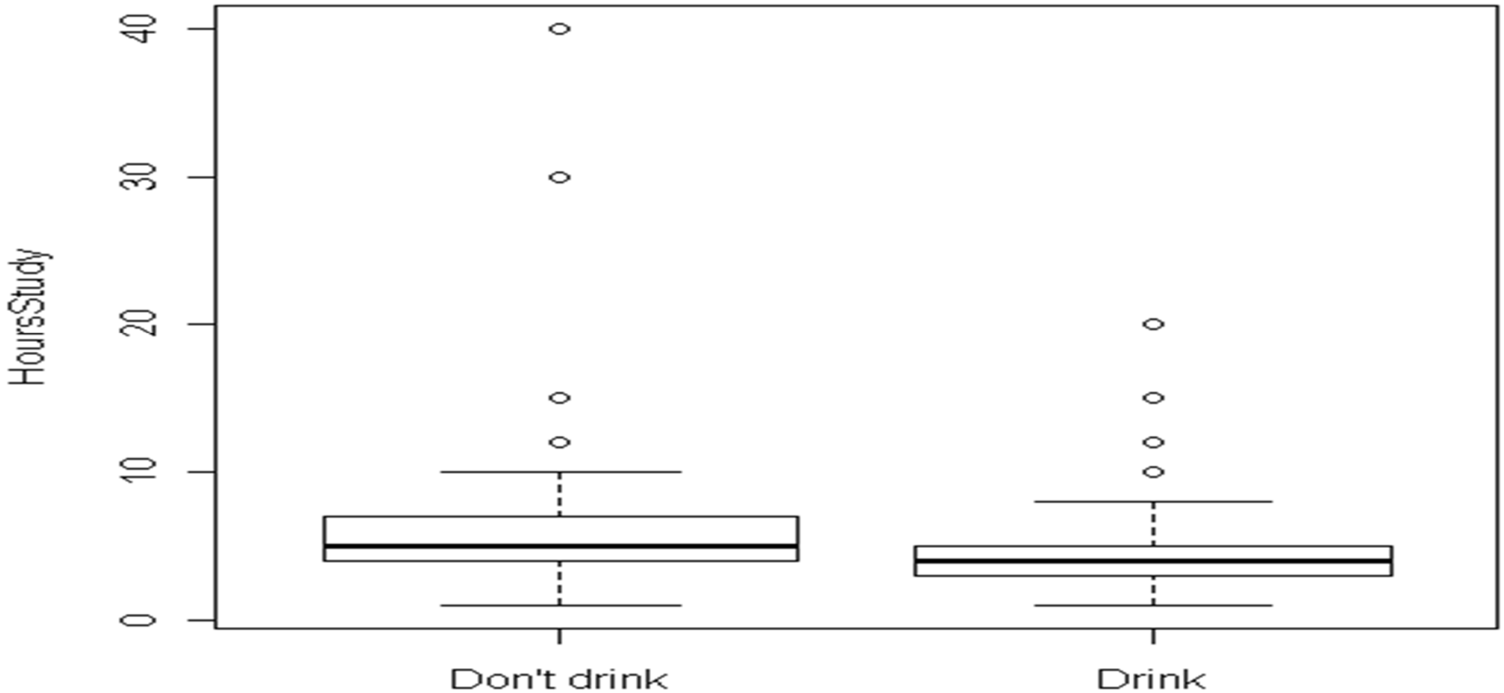
```
  Welch Two Sample t-test
data:  HoursStudy by Drink
t = 1.9923, df = 70.556, p-value = 0.05021
alternative hypothesis: true difference in means is not
  equal to 0
95 percent confidence interval:
 -0.001651343  3.450180754
sample estimates:
mean in group Don't drink       mean in group Drink
                 6.500000                  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