Announcements:

- Discussion today is review for midterm, no credit. You may attend more than one discussion section.
- Bring 2 sheets of notes and calculator to midterm. We will provide Scantron form.

Homework: (Due Wed)

Chapter 10: #5, 22, 42

Chapter 10

Estimating Proportions with Confidence

Reminder from when we started Chapter 9

Five situations we will cover for the rest of this quarter:

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	$\bar{x}$
Population mean of paired differences (dependent samples, paired)	μ_d	$\bar{d}$
Difference in two population means (independent samples)	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

For each situation will we:

√ 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

Confidence interval example from Fri lecture

Gallup poll of $n = 1018$ adults found 39% believe in evolution. So $\hat{p} = .39$

A **95% confidence interval** or **interval estimate** for the proportion (or percent) of *all* adults who believe in evolution is **.36** to **.42** (or **36%** to **42%**).

Confidence interval: an interval of estimates that is *likely* to capture the population value.

Goal today: Learn to calculate and interpret confidence intervals for p and for $p_1 - p_2$ *and* learn general format.

Remember population versus sample:

- **Population proportion:** the fraction of the ***population*** that has a certain trait/characteristic or the probability of success in a binomial experiment – denoted by p. The value of the ***parameter*** p is not known.
- **Sample proportion:** the fraction of the ***sample*** that has a certain trait/characteristic – denoted by $\hat{p}$. The ***statistic*** $\hat{p}$ is an estimate of p.

The **Fundamental Rule for Using Data for Inference:** Available data can be used to make inferences about a much larger group *if the data can be considered to be representative with regard to the question(s) of interest.*

Some Definitions:

- **Point estimate:** A *single number* used to estimate a population parameter. For our five situations:

point estimate = sample statistic = sample estimate

$= \hat{p}$ for one proportion

$= \hat{p}_1 - \hat{p}_2$ for difference in two proportions

- **Interval estimate:** An *interval* of values used to estimate a **population parameter**. Also called a **confidence interval**. For our five situations, always:

Sample estimate ± multiplier × standard error

Details for proportions:

Sample estimate ± multiplier × standard error

Parameter	Sample estimate	Standard error
p	$\hat{p}$	$s.e.(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	See p. 424 for formula

Multiplier and Confidence Level

- The **multiplier** is determined by the desired confidence level.
- The **confidence level** is the probability that the procedure used to determine the interval *will* provide an interval that includes the population parameter. Most common is .95.
- If we consider *all possible* randomly selected samples of same size from a population, the *confidence level* is the fraction or percent of those samples for which the confidence interval includes the population parameter.

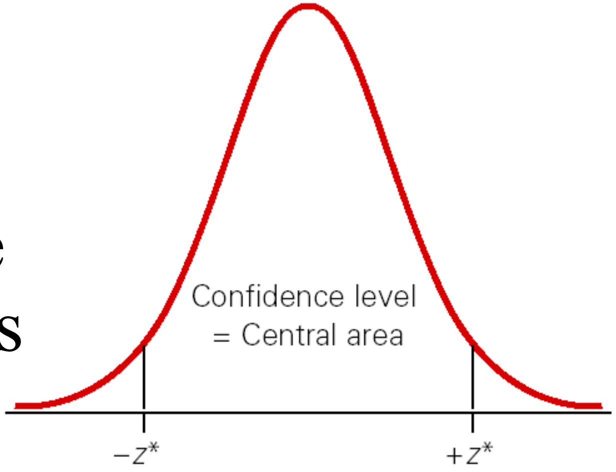
  See picture on board.
- Often express the confidence level as a percent. Common levels are 90%, 95%, 98%, and 99%.

More about the Multiplier

Confidence Level	Multiplier	Confidence Interval
90	1.645 or 1.65	$\hat{p} \pm 1.65$ standard errors
95	1.96, often rounded to 2	$\hat{p} \pm 2$ standard errors
98	2.33	$\hat{p} \pm 2.33$ standard errors
99	2.58	$\hat{p} \pm 2.58$ standard errors

Note: Increase confidence level => larger multiplier.

Multiplier, denoted as z^*, is the standardized score such that the area between $-z^*$ and z^* under the standard normal curve corresponds to the desired confidence level.



Formula for C.I. for proportion

Sample estimate ± multiplier × standard error

For one proportion: A confidence interval for a population proportion p, based on a sample of size n from that population, with sample proportion $\hat{p}$ is:

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

Example of different confidence levels

Poll on belief in evolution:

$n = 1018$

Sample proportion = .39

Standard error = $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{.39(1-.39)}{1018}} = .0153$

90% confidence interval

$.39 \pm 1.65(.0153)$ or $.39 \pm .025$ or .365 to .415

95% confidence interval:

$.39 \pm 2(.0153)$ or $.39 \pm .03$ or .36 to .42

99% confidence interval

$.39 \pm 2.58(.0153)$ or $.39 \pm .04$ or .35 to .43

Interpretation of the confidence interval and confidence level:

- We are ***90% confident*** that the proportion of *all* adults in the US who believe in evolution is between **.365 and .415**.
- We are ***95% confident*** that the proportion of *all* adults in the US who believe in evolution is between **.36 and .42.**
- We are ***99% confident*** that the proportion of *all* adults in the US who believe in evolution is between **.35 and .43**.

Interpreting the confidence level of 99%:

The interval .35 to .43 ***may or may not*** capture the true proportion of adult Americans who believe in evolution

But, ***in the long run*** this ***procedure*** will produce intervals that capture the unknown population values about 99% of the time. So, we are 99% confident that it worked this time.

Notes about interval width

- Higher confidence <=> wider interval
- Larger *n* (sample size) <=> more narrow interval, because *n* is in the *denominator* of the standard error.
- So, if you want a more narrow interval you can either *reduce* your confidence, or *increase* your sample size.

Reconciling with Chapter 3 formula for 95% confidence interval

Sample estimate $\pm$ Margin of error

where (conservative) margin of error was $\dfrac{1}{\sqrt{n}}$

Now, "margin of error" is $2\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

These are the *same* when $\hat{p} = .5$. The new margin of error is *smaller* for *any* other value of $\hat{p}$ So we say the old version is *conservative.* It will give a *wider* interval.

Comparing three versions (Details on board)

For the evolution example, $n = 1018, \ \hat{p} = .39$

- *Conservative* margin of error = $.0313 \approx .03$
- *Approximate* margin of error using z* = 2

$$2 \times .0153 = .0306 \approx .03$$

- *Exact* margin of error using z* = 1.96

$$1.96 \times .0153 = .029988 \approx .03$$

All very close to .03, and it really doesn't make much difference which one we use!

New example: compare methods

Marist Poll in Oct 2009 asked "How often do you text while driving?" $n = 1026$

Nine percent answered "Often" or "sometimes" so and $\hat{p} = .09$

$$s.e.(\hat{p}) = \sqrt{\frac{.09(.91)}{1026}} = .009$$

- *Conservative* margin of error = .0312
- *Approximate* margin of error = 2 × .009 = .018.

This time, they are quite different!

The conservative one is too conservative, it's double the approximate one!

Comparing margin of error

- Conservative margin of error will be okay for sample proportions near .5.
- For sample proportions far from .5, closer to 0 or 1, don't use the conservative margin of error. Resulting interval is wider than needed.
- Note that using a multiplier of 2 is called the *approximate* margin of error; the *exact* one uses multiplier of 1.96. It will rarely matter if we use 2 instead of 1.96.

Factors that Determine Margin of Error

1. ***The sample size, n.***
   When sample size *increases*, margin of error *decreases*.
2. ***The sample proportion,*** $\hat{p}$**.**
   If the proportion is close to either 1 or 0 most individuals have the same trait or opinion, so there is little natural variability and the margin of error is smaller than if the proportion is near 0.5.
3. ***The "multiplier"* 2 or 1.96.**
   Connected to the "95%" aspect of the margin of error. Usually the term "margin of error" is used only when the confidence level is 95%.

General Description of the Approximate 95% CI for a Proportion

Approximate 95% CI for the population proportion:

$$\hat{p} \pm 2 \text{ standard errors}$$

The standard error is $s.e.(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

Interpretation: For about 95% of all randomly selected samples from the population, the confidence interval computed in this manner captures the population proportion.

Necessary Conditions: $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 10, and the sample is randomly selected.

Finding the formula for a 95% CI for a Proportion – use Empirical Rule:

For 95% of all samples, $\hat{p}$ is within 2 st.dev. of p

Sampling distribution of $\hat{p}$ tells us *for 95% of all samples*:

-2 standard deviations $< \hat{p} - p <$ 2 standard deviations

Don't know true standard deviation, so use standard error.
For approximately 95% of all samples,

-2 standard errors $< \hat{p} - p <$ 2 standard errors

which implies for approximately 95% of all samples,

$\hat{p}$ – 2 standard errors $< p <$ $\hat{p}$ + 2 standard errors

Same holds for *any* confidence level; replace 2 with z*

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

where:

- $\hat{p}$ is the sample proportion
- z^* denotes the multiplier.
- . $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the standard error of $\hat{p}$.

Example 10.3 *Intelligent Life Elsewhere?*

Poll: Random sample of 935 Americans

Do you think there is intelligent life on other planets?

Results: 60% of the sample said "yes", $\hat{p} = .60$

$$s.e.(\hat{p}) = \sqrt{\frac{.6(1-.6)}{935}} = .016$$

90% Confidence Interval: .60 ± 1.65(.016), or .60 ± .026

.574 to .626 or 57.4% to 62.6%

98% Confidence Interval: .60 ± 2.33(.016), or .60 ± .037

.563 to .637 or 56.3% to 63.7%

Note: entire interval is above 50% => high confidence that a majority believe there is intelligent life.

Confidence intervals and "plausible" values

- Remember that a confidence interval is an *interval estimate* for a population parameter.
- Therefore, any value that is covered by the confidence interval is a *plausible value* for the parameter.
- Values *not* covered by the interval are still possible, but not very likely (depending on the confidence level).

Example of plausible values

- 98% Confidence interval for proportion who believe intelligent life exists elsewhere is:

  .563 to .637 or 56.3% to 63.7%

- Therefore, 56% is a *plausible value* for the population percent, but 50% is not very likely to be the population percent.
- Entire interval is above 50% => high confidence that a *majority* believe there is intelligent life.

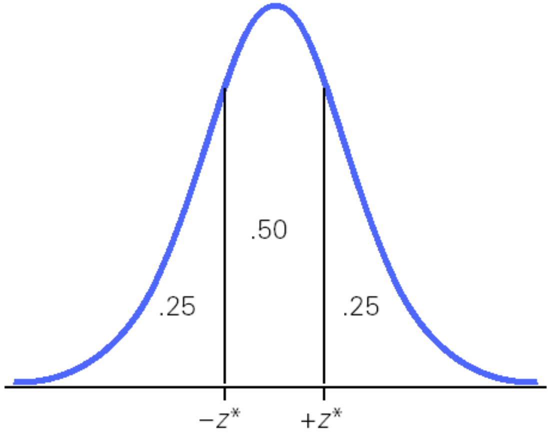
New multiplier: let's do a confidence level of 50%

Poll: Random sample of 935 Americans

"Do you think there is intelligent life on other planets?"

Results: 60% of the sample said "yes", $\hat{p} = .60$

We want a **50% confidence interval**.
If the area between -z^* and z^* is .50,
then the area to the left of z^* is .75.
From Table A.1 we have $z^* \approx .67$.
(See next page for Table A.1)



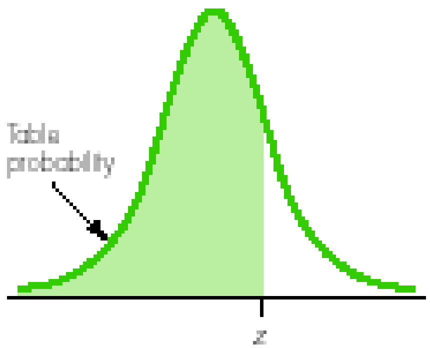
50% Confidence Interval: $.60 \pm .67(.016)$, or $.60 \pm .011$

.589 to .611 or 58.9% to 61.1%

Note: Lower confidence *level* results in a narrower *interval*.

Here is the relevant part of Table A.1. We want the z* value with area 0.75 below it. The closest value is the one with .7486 below it, which corresponds to a z value of 0.67. So z* = 0.67 is the multiplier for a 50% confidence interval. (Note: We could average the two z values with .7486 and .7517 below them and use 0.675, but .67 is close enough.)

Standard Normal Probabilities (for $z > 0$)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

Remember conditions for using the formula:

1. Sample is **randomly selected** from the population. ***Note***: Available data can be used to make inferences about a much larger group *if the data can be considered to be representative with regard to the question(s) of interest.*
2. Normal curve approximation to the distribution of possible sample proportions assumes a **"large" sample size**. Both $n\hat{p}$ and $n(1-\hat{p})$ should be at least 10 (although some say these need only to be at least 5).

In Summary: Confidence Interval for a Population Proportion *p*

General CI for *p*: $$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Approximate 95% CI for *p*: $$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Conservative 95% CI for *p*: $$\hat{p} \pm \frac{1}{\sqrt{n}}$$

Section 10.4: Comparing two population proportions

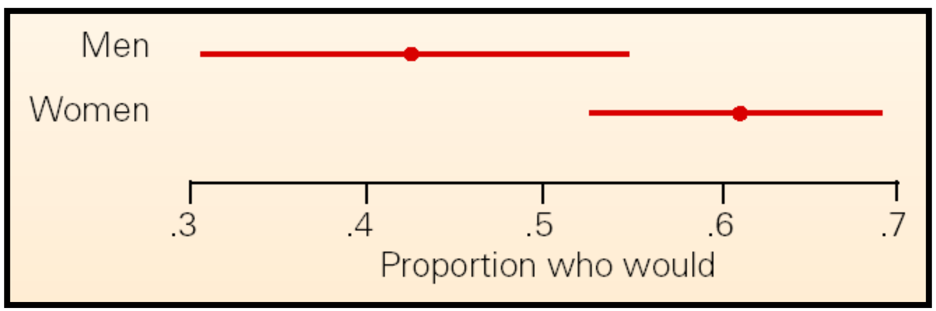
- Independent samples of size n_1 and n_2
- Use the two *sample* proportions as data.
- Could compute separate confidence intervals for the two population proportions and see if they overlap.
- Better to find a confidence interval for the *difference* in the two population proportions,

Case Study 10.3 *Comparing proportions*

Would you date someone with a great personality even though you did ***not*** find them attractive?

Women: .611 (61.1%) of 131 answered "yes."
95% confidence interval is .527 to .694.

Men: .426 (42.6%) of 61 answered "yes."
95% confidence interval is .302 to .55.



Conclusions:

- ***Higher proportion*** of ***women*** would say yes. CIs slightly overlap
- Women CI ***narrower*** than men CI due to larger sample size

Compare the two proportions by finding a CI for the difference

C.I. for the difference in two population proportions:

Sample estimate ± multiplier × standard error

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Case Study 10.3 *Comparing proportions*

Would you date someone with a great personality even though you did ***not*** find them attractive?

Women: .611 of 131 answered "yes."
95% confidence interval is .527 to .694.

Men: .426 of 61 answered "yes."
95% confidence interval is .302 to .55.

Confidence interval for the difference in *population proportions* of women and men who would say yes.

$$(.611 - .426) \pm z^* \sqrt{\frac{.611(1-.611)}{131} + \frac{.426(1-.426)}{61}}$$

95% confidence interval

- A 95% confidence interval for the difference is .035 to .334 or 3.5% to 33.4%.
- We are 95% confident that the *population* proportions of men and women who would date someone they didn't find attractive *differ* by between .035 and .334, with a lower proportion for men than for women.
- We can conclude that the two *population* proportions *differ* because 0 is not in the interval.

Section 10.5: Using confidence intervals to guide decisions

- A value *not* in a confidence interval can be rejected as a likely value for the population parameter.
- When a confidence interval for $p_1 - p_2$ does not cover 0 it is reasonable to conclude that the two population values differ.
- When confidence intervals for p_1 and p_2 do not overlap it is reasonable to conclude they differ, but if they do overlap, no conclusion can be made. In that case, find a confidence interval for the difference.

From the Midterm 2 review sheet for Chapter 10 - you should know these now

1. Understand how to interpret the *confidence level*
2. Understand how to interpret a *confidence interval*
3. Understand how the sampling distribution for $\hat{p}$ leads to the confidence interval formula (pg. 417-418)
4. Know how to compute a confidence interval for one proportion, including conditions needed.
5. Know how to compute a confidence interval for the difference in two proportions, including conditions needed.
6. Understand how to find the multiplier for desired confidence level.
7. Understand how margin of error from Chapter 3 relates to the 95% confidence interval formula in Chapter 10
8. Know the general format for a confidence interval for the 5 situations defined in Chapter 9 (see summary on page 483).