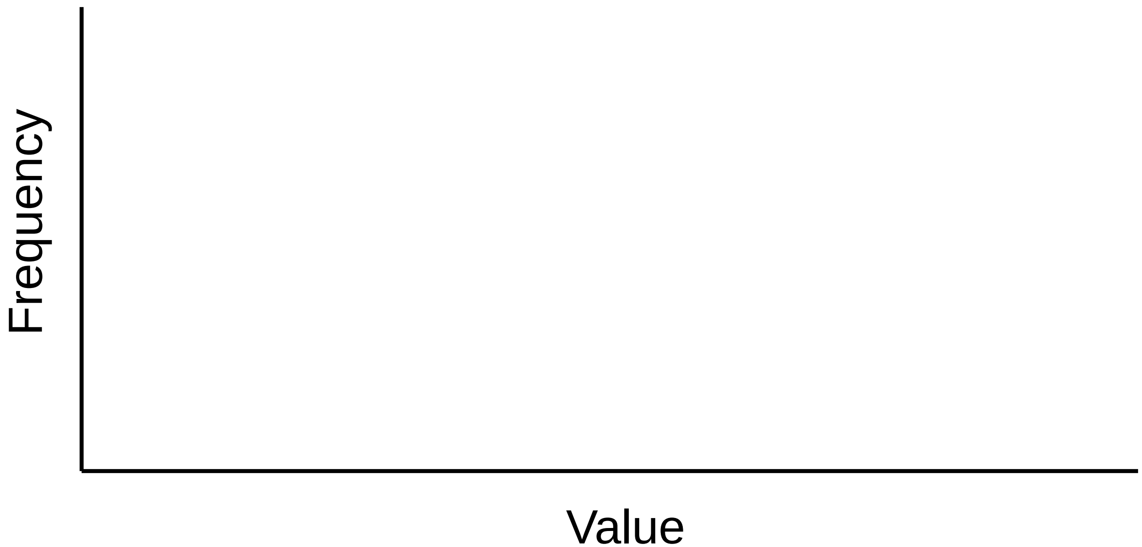
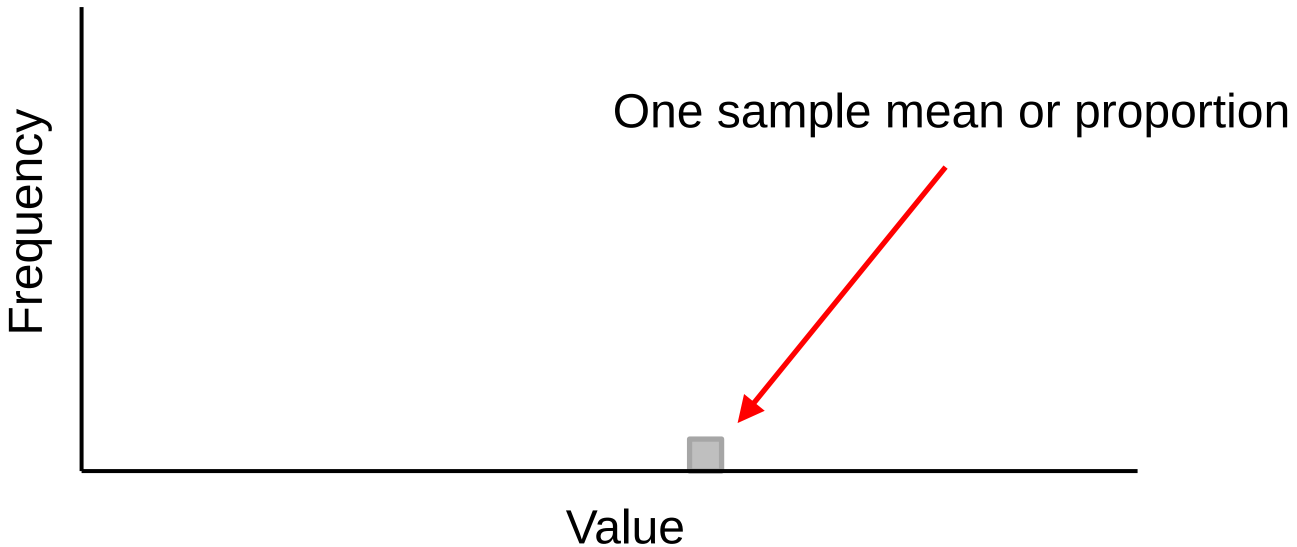
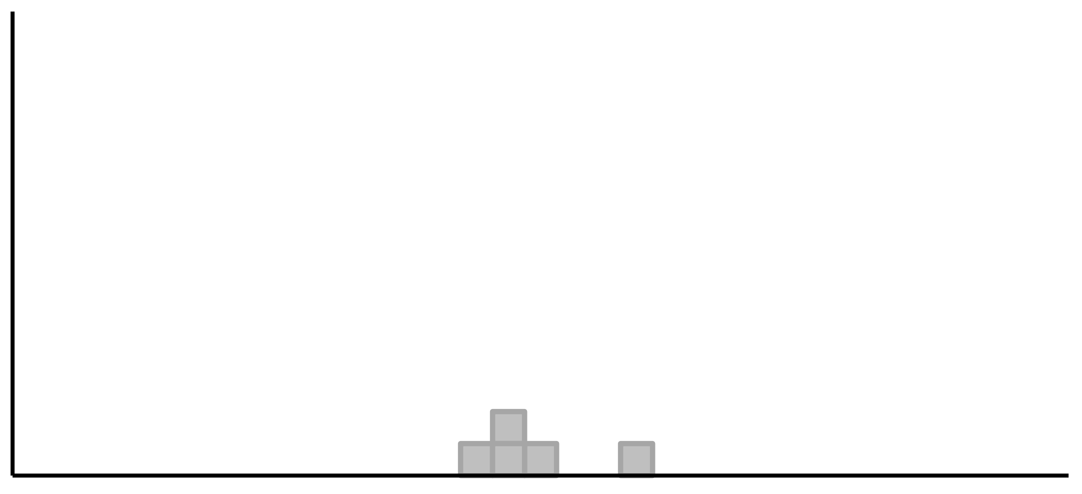
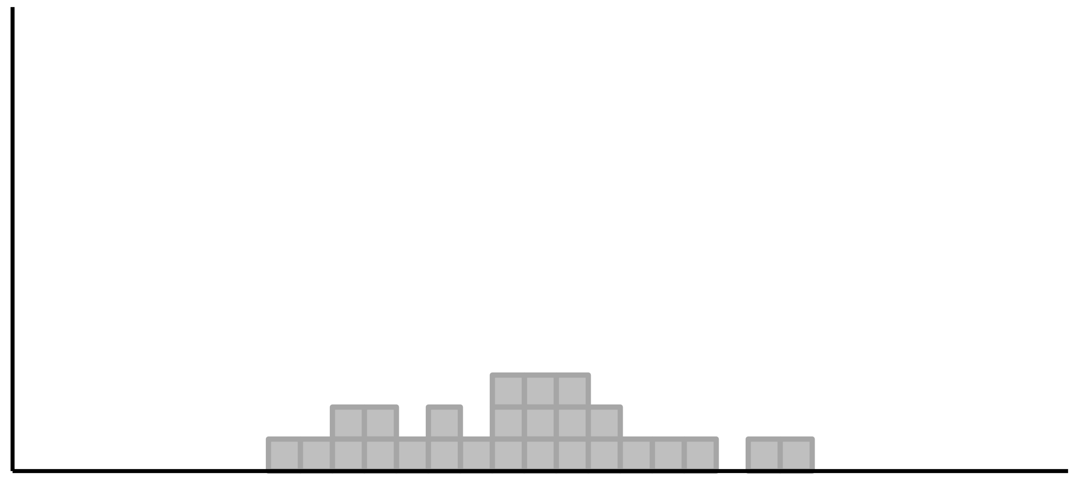
SOC 3811/5811: BASIC SOCIAL STATISTICS

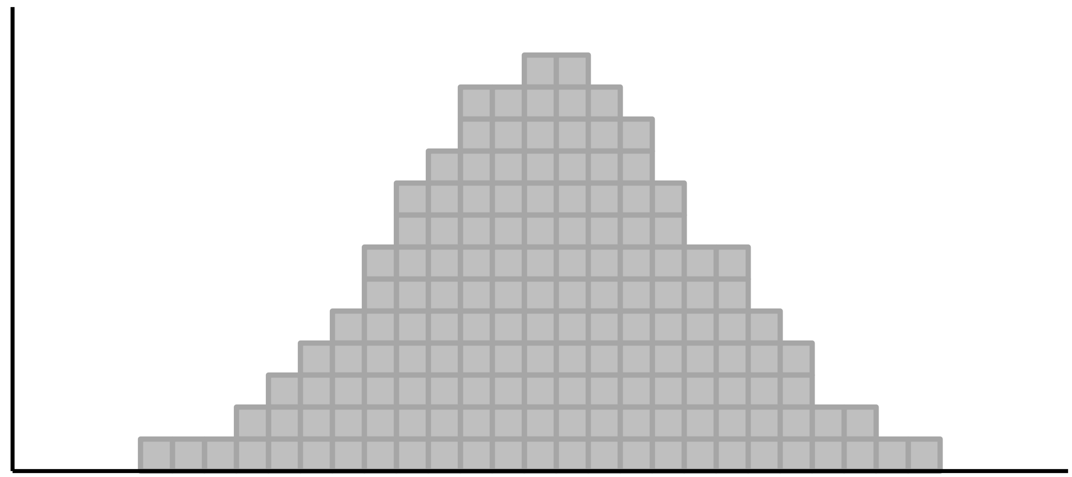
Confidence Intervals

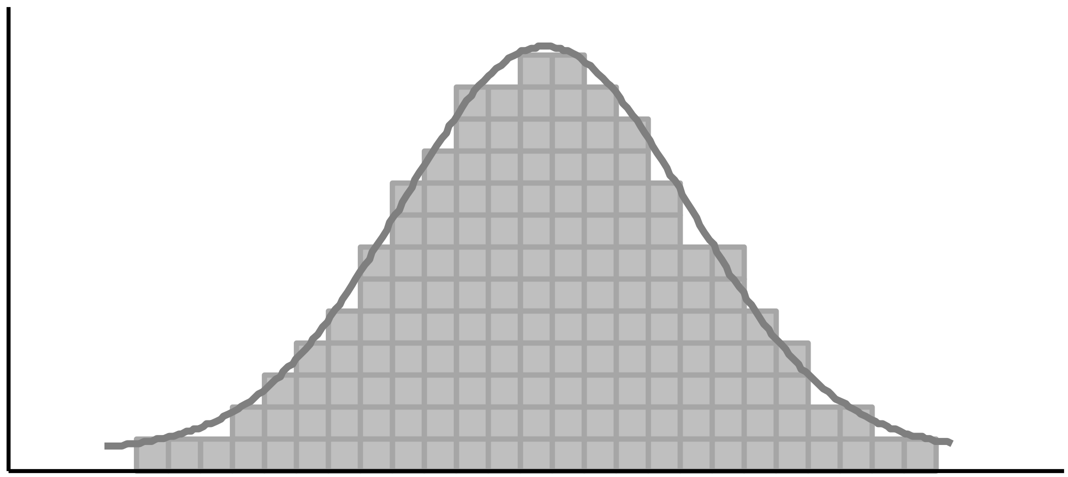

Frequency
Value

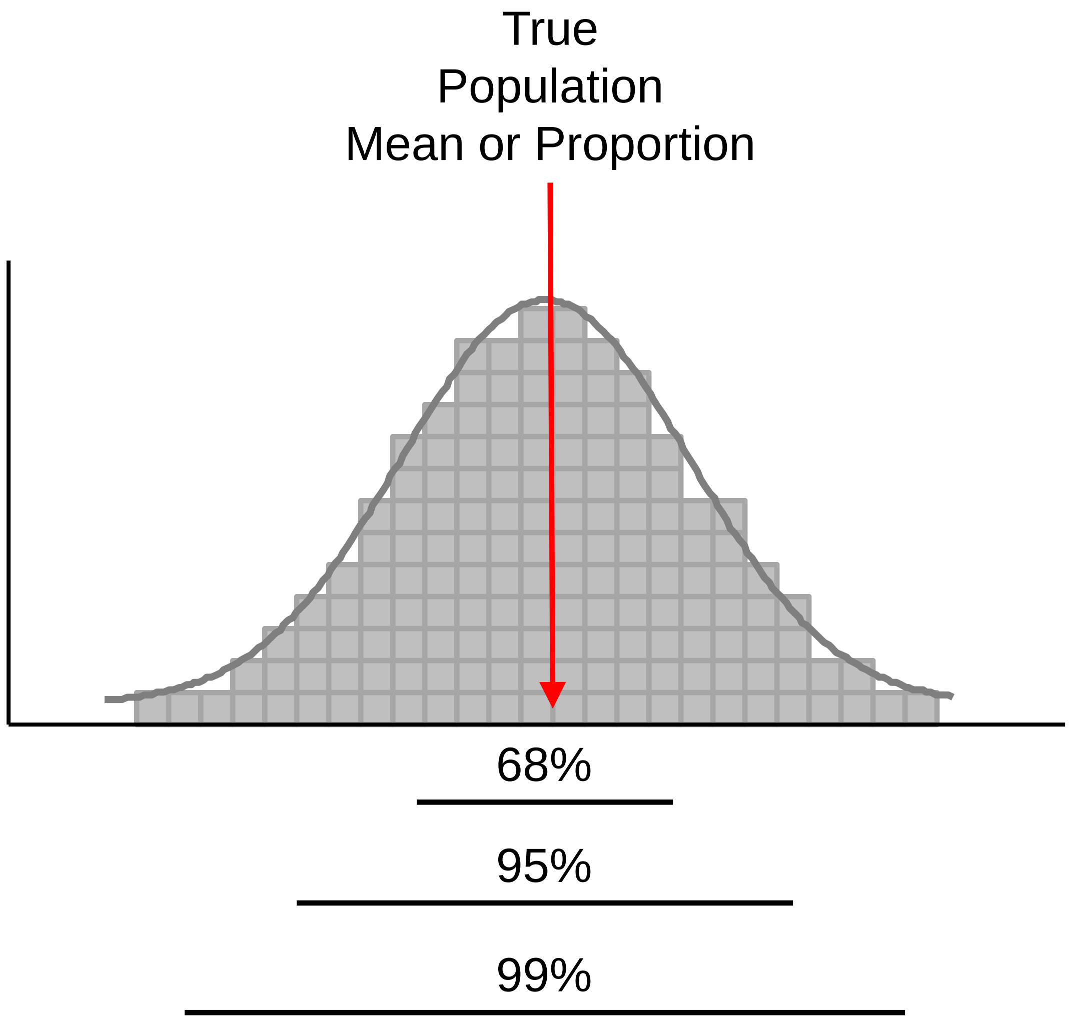

One sample mean or proportion
Frequency
Value

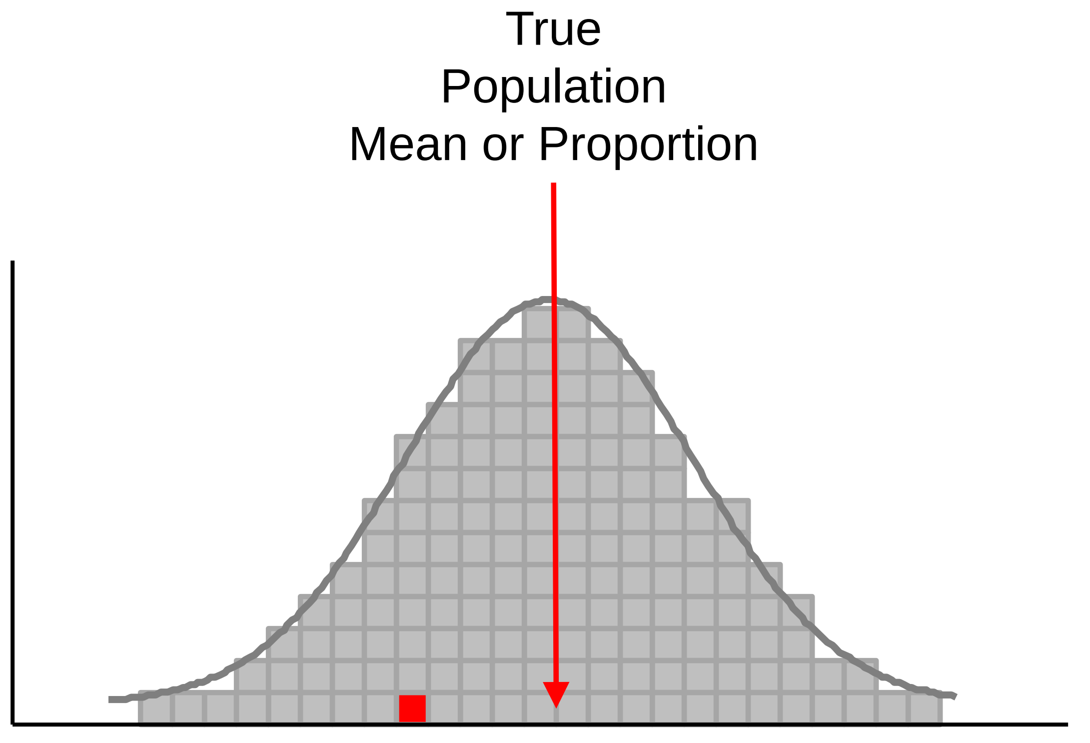


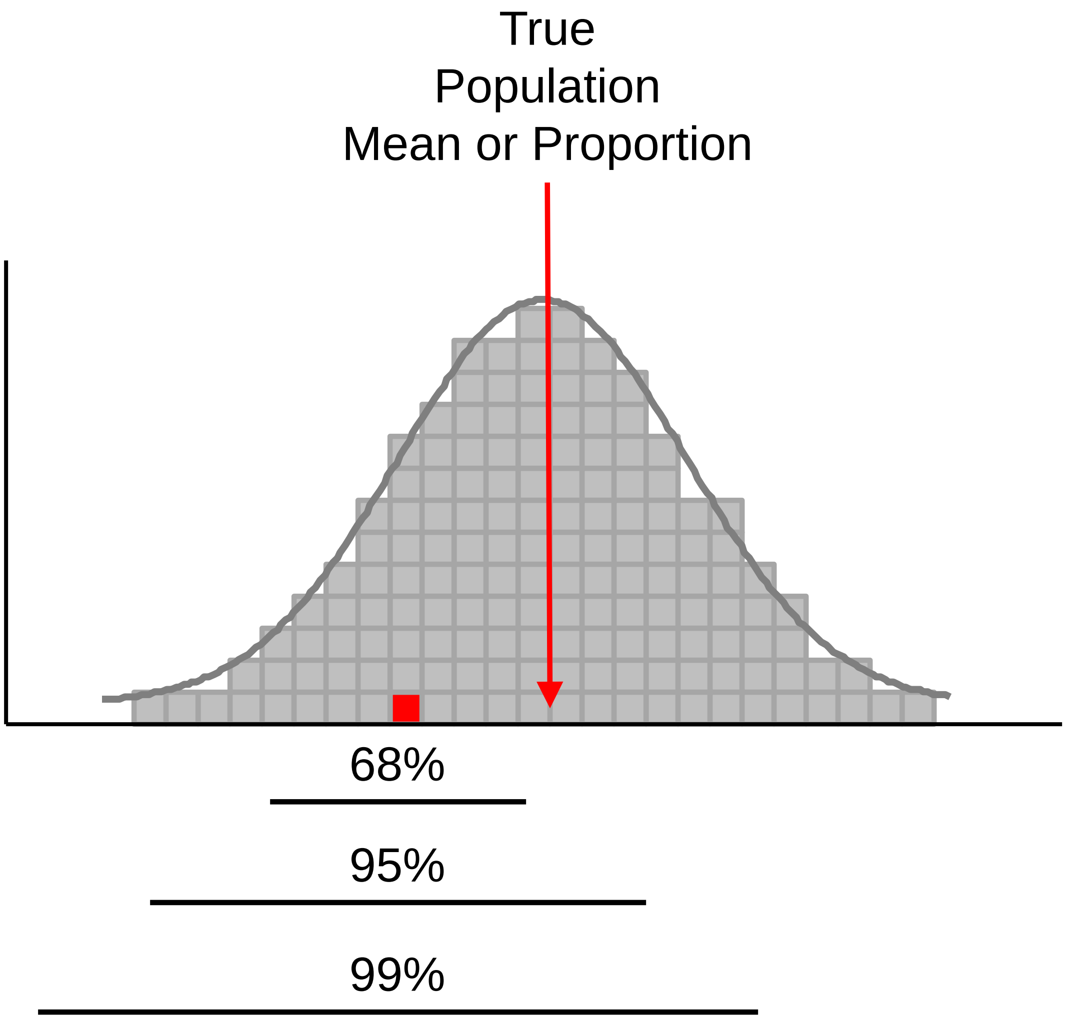







True
Population
Mean or Proportion
68%
95%
99%


True
Population
Mean or Proportion


True
Population
Mean or Proportion
68%
95%
99%

Inferences for Population Parameters

TODAY

When we compute **confidence intervals** we use sample data to specify a range of values within which we are confident that the population parameter falls

Example: "With 95% certainly we conclude that the population proportion of people who own a car is between 0.401 and 0.420"

Example: "With 99% certainty we conclude that the mean income in the population is somewhere between $31,200 and $32,000"

Inferences for Population Parameters

NEXT WEEK

When we conduct **hypothesis testing** (or **significance testing**) we use sample data to test particular claims about the value of a population parameter

Example: "Do our sample data support the assertion that more than 41% of people in the population own cars?"

Example: "Do our sample data support the assertion that the mean income in the population is greater than $31,900?"

Confidence Intervals

Confidence Interval

A range of values that is "likely" to contain the population parameter (for example, a mean or proportion)

Just how "likely" it is that the confidence intervals contains the population proportion is called the *confidence level*

If our confidence level is C%, then we are saying that if we drew many random samples and computed many confidence intervals, then the true population parameter would be contained within the resulting confidence intervals C% of the time

Confidence Intervals

We will consider four different sorts of confidence intervals, all of which follow the same logic

Confidence Intervals for <u>Proportions</u>

Use $\hat{p}$ to infer p, the population proportion

Confidence Intervals for <u>Means</u>

Use $\overline{Y}$ to infer μ_Y, the population mean of Y

Confidence Intervals for <u>Differences in Proportions</u>

Use $\hat{p}_1 - \hat{p}_2$ to infer the difference between two population proportions, p_1 and p_2

Confidence Intervals for <u>Differences in Means</u>

Use $\overline{Y}_1 - \overline{Y}_2$ to infer the difference between two population means, μ_{Y1} and μ_{Y2}

Confidence Intervals for Proportions

EARLIER: We said that the sample percentage was within plus or minus one "margin of error" of the population proportion 95% of the time

We defined the "conservative margin of error" as:

$$\frac{1}{\sqrt{N}} \times 100\%$$

The "conservative margin of error" is just a quick, informal way of computing a 95% confidence interval

NOW: Define the margin of error more formally

Confidence Intervals for Proportions

For a 95% confidence interval for **proportions**:

$$\text{Margin of Error} = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where

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

is the standard error of the sampling distribution of $\hat{p}$

For a 95% confidence interval, the margin of error equals plus or minus 1.96 standard errors

Confidence Intervals for Proportions

Based on the logic of sampling distributions:

95% of the sample proportions should fall within plus or minus 1.96 standard errors of p

This is exactly the same as saying that there is a 95% chance that any particular sample proportion falls within plus or minus 1.96 standard errors of p

(Likewise, for a 68% confidence interval we would be 68% confident that the population proportion falls within plus or minus one standard error of any particular sample proportion)

Confidence Intervals for Proportions

Informally, the formula for the confidence interval for a population proportion is $\hat{p}$ plus or minus the margin of error; we control the size of the margin of error by adjusting the desired confidence level

More formally:

$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The $Z_{\alpha/2}$ is called the "multiplier"

For a 95% confidence interval, $Z_{\alpha/2}$ equals 1.96

For a 68% confidence interval, $Z_{\alpha/2}$ equals 1

Etc.

Confidence Intervals for Proportions

What proportion of American adults (age 25+) has a physical disability?

I drew one random sample with n=1,000; in this sample, the proportion with a disability was 0.121

What is the 95% confidence interval for the population proportion of people with disabilities?

$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.121 \pm 1.96 \times \sqrt{\frac{0.121(0.879)}{1{,}000}}, \text{or } 0.121 \pm 0.020$$

Confidence Intervals for Proportions

Thus with 95% certainty we conclude that…

…the population proportion of people with a disability equals 0.121 plus or minus 0.020

…the population proportion likely falls between 0.101 and 0.141

What would a 68% confidence interval look like?

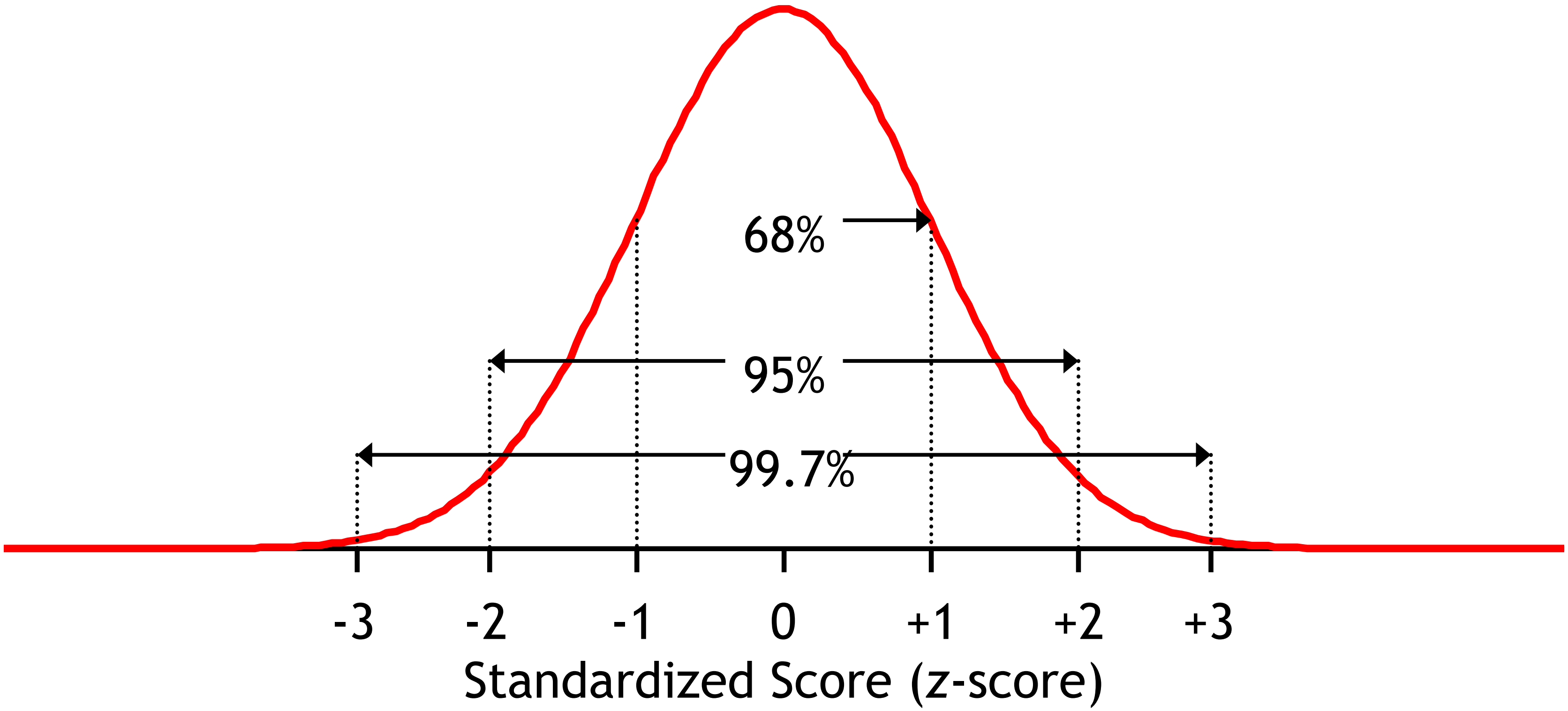
$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.121 \pm 1 \times \sqrt{\frac{0.121(0.879)}{1{,}000}}, \text{or } 0.121 \pm 0.010$$

Note that the margin of error got smaller

Confidence Intervals for Proportions

For any sample of size n, the width of the confidence interval increases (or becomes less precise) as the confidence level increases



Confidence Intervals for Proportions

We would like to have a very precise confidence interval

- It wouldn't be very useful, for example, to say that we are confident that the population proportion of people with disabilities is between 0.01 and 0.23

We can have a more precise confidence interval if we accept a lower confidence level

- But it also wouldn't very useful to say, for example, that we are 10% confident that the population proportion of people with disabilities falls between 0.120 and 0.122

How can we have a high confidence level **and** a narrow (that is to say, precise) confidence interval?

Confidence Intervals for Proportions

The formula for the confidence interval for a population proportion includes n

$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The larger the value of n, the smaller the margin of error

- We set the confidence level
- If n is large, we can have a high confidence level <u>and</u> a narrow confidence interval
- Of course, a larger n usually costs more money…

Confidence Intervals for Proportions

Returning to our example, what would 95% confidence intervals for the population proportion of people with disabilities look like with different sample sizes?

n	*p*-hat	margin of error
10	0.121	0.206
100	0.121	0.065
1,000	0.121	0.020
10,000	0.121	0.007

Confidence Intervals for Proportions

In fact many researchers decide how many people to include in their sample based on how precise of a confidence interval they desire

- Imagine that we want to be able to construct a 95% confidence interval that has a margin of error of plus or minus 0.01
- What sample size should we choose?
- We would first need to make a guess about the value of *p*-hat (let's say 0.12, using the disability example)

$$0.01 = 1.96 \times \sqrt{\frac{0.12(0.88)}{n}}, \text{so } n = 4{,}057$$

Worksheet

How many Americans cannot name the governor of the state in which they live?

In 1987, the General Social Survey asked 1,819 people for the name of the governor of their state

447 people gave incorrect answers

Construct and interpret a 99% confidence interval for p

Confidence Intervals More Generally

So far we have examined confidence intervals for proportions, but the fundamental logic is the same for other types of confidence intervals

All confidence intervals can be written generally as:

$$\text{Sample Estimate} \pm \text{Multiplier} \times \text{Standard Error}$$

Or:

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}})$$

Confidence Intervals More Generally

$$\text{Sample Estimate} \pm \text{Multiplier} \times \text{Standard Error}$$

For **proportions**: $se_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

For **means**: $se_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$

For **differences in proportions**: $se_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

For **differences in means**: $se_{Y-X} = \sqrt{\frac{s_Y^2}{n_Y} + \frac{s_X^2}{n_X}}$

Confidence Intervals for Means

Given this general formula for any confidence interval:

$$\text{Sample Estimate} \pm \text{Multiplier} \times \text{Standard Error}$$

we can then specify a confidence interval for a population mean as:

$$\bar{Y} \pm t_{a/2} \frac{s_Y}{\sqrt{n}}$$

Recall that to use the sample standard deviation (s_Y) in place of the unknown population standard deviation (σ_Y) we use the t-distribution with n-1 degrees of freedom instead of the Z distribution

Confidence Intervals for Means

The sample must be representative of the population from which it was drawn

One of these two things must be true:

The random sample is small (<30) and symmetrically distributed with no outliers, and the population of measurements is bell-shaped

--- or ---

The size of the random sample is large ($\geq$30), regardless of the shape of the distribution of the measurements in the population

Confidence Intervals for Means

1,000 adults are randomly selected from the population of the United States. Their mean personal income was \$33,529, with a standard deviation of \$40,609

In general, a confidence interval for μ equals

$$\$33{,}529 \pm t_{\alpha/2} \frac{\$40{,}609}{\sqrt{1000}}$$

According to a **t Table**,

$t_{\alpha/2}$ = 1.645 yields a confidence level of 0.90

$t_{\alpha/2}$ = 1.960 yields a confidence level of 0.95

$t_{\alpha/2}$ = 2.576 yields a confidence level of 0.99

Confidence Intervals for Means

A 90% confidence interval for the mean equals:

$$\$33{,}529 \pm 1.645 \times \frac{\$40{,}609}{\sqrt{1000}}, \text{or } \$33{,}529 \pm \$2{,}119$$

A 95% confidence interval for the mean equals:

$$\$33{,}529 \pm 1.960 \times \frac{\$40{,}609}{\sqrt{1000}}, \text{or } \$33{,}529 \pm \$2{,}517$$

A 99% confidence interval for the mean equals:

$$\$33{,}529 \pm 2.576 \times \frac{\$40{,}609}{\sqrt{1000}}, \text{or } \$33{,}529 \pm \$3{,}313$$

Worksheet

We sampled 1,600 people and observed their IQ scores. We got a sample mean of 103 and a standard deviation of 15.

Construct a 90% confidence interval for the population mean μ

Confidence Intervals for Differences in Proportions

Given this general formula for any confidence interval:

$$\text{Sample Estimate} \pm \text{Multiplier} \times \text{Standard Error}$$

we can then specify a confidence interval for the difference between two population proportions as:

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Note that here we use the $z_{\alpha/2}$ multiplier and the standard normal distribution instead of the $t_{\alpha/2}$ multiplier and the *t*-distribution

Confidence Intervals for Differences in Proportions

When computing a confidence interval for differences in proportions, it must be the case that the two samples are independent such that measures in one sample are not related to measures in the other sample

- If our samples were from populations of (1) men and (2) women, then measures of blood pressure might be independent
- If our samples were from populations of (1) wives and (2) their husbands, then measure of blood pressure might be dependent

Also, $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ must all be at least 5 (and preferably 10)

Confidence Intervals for Differences in Proportions

Do high school dropouts and high school graduates experience different rates of disability?

- I randomly selected 1,022 people from the 2000 U.S. Census
- Of the 1,022 people, n_1=141 had completed less than high school and n_2=881 had at least completed high school
- Among the 141 high school non-completers, 20.2% had a disability
- Among the 881 high school completers, 9.1% had a disability

In the population, how do rates of disability differ between people who have completed high school and those who have not?

Confidence Intervals for Differences in Proportions

Given this formula for a confidence interval for differences in proportions:

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

we can construct a 95% confidence interval using a $Z_{\alpha/2}$ value of 1.96. With n_1=141, n_2=881, $\hat{p}_1$=0.202, and $\hat{p}_2$= 0.091, we have

$$0.202 - 0.091 \pm 1.96 \times \sqrt{\frac{0.202(1-0.202)}{141}+\frac{0.091(1-0.091)}{881}}$$

$$0.111 \pm 0.069 \text{ (between } 0.042 \text{ and } 0.180)$$

Worksheet

Every year the General Social Surveys asks people whether they agree or disagree that "a working mother can establish just as warm and secure a relationship with her children as a mother who does not work."

In 1977, 735 of 1,503 respondents agreed. In 2012, 939 of 1,301 respondents agreed.

Construct a 99% confidence interval for the difference in population proportions between 1977 and 2012

Confidence Intervals for Differences in Means

Given this general formula for any confidence interval:

$$\text{Sample Estimate} \pm \text{Multiplier} \times \text{Standard Error}$$

we can then specify a confidence interval for the difference between two population means as:

$$\bar{Y} - \bar{X} \pm t_{\alpha/2} \sqrt{\frac{s_Y^2}{n_Y} + \frac{s_X^2}{n_X}}$$

What value do we use for the df of the t-distribution?

- The mathematically correct number of degrees of freedom for the t-distribution in this case is complex
- Good Approximation: Use the smaller of n_Y-1 or n_X-1

Confidence Intervals for Differences in Means

When computing a confidence interval for differences in means, it must again be the case that the two samples are <u>independent</u> such that measures in one sample are not related to measures in the other sample

One of these two things must be true for each sample:

The random sample is small (<30) and symmetrically distributed with no outliers, and the population of measurements is bell-shaped

--- or ---

The size of the random sample is large ($\geq$30), regardless of the shape of the distribution of the measurements in the population

Confidence Intervals for Differences in Means

Do high school dropouts and high school graduates have different mean earnings?

- I randomly selected 1,022 people from the 2000 U.S. Census
- Among the n_Y=141 high school non-completers, income averaged \$16,259 with a standard deviation of \$14,672.
- Among the n_X=881 high school completers, income averaged \$36,284 with a standard deviation of \$42,697

In the population, how do earnings differ between people who have completed high school and those who have not?

Confidence Intervals for Differences in Means

Given this formula for a confidence interval for differences in means:

$$\bar{Y} - \bar{X} \pm t_{\alpha/2}\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_X^2}{n_X}}$$

we can construct a 95% confidence interval using a $t_{\alpha/2}$ value of 1.98. With n_Y=141, n_X=881, $\bar{Y}$ =\$16,259, s_Y=\$14,672, $\bar{X}$ =\$36,284, and s_X=\$42,697, we have

$$\$16{,}259 - \$36{,}284 \pm 1.98 \times \sqrt{\frac{\$14{,}672^2}{141} + \frac{\$42{,}697^2}{881}}$$

$$-\$20{,}025 \pm \$3{,}755 \text{ (between } -\$16{,}270 \text{ and } -\$23{,}780\text{)}$$

Worksheet

Every year the General Social Surveys administers a 10-item vocabulary knowledge test

In 1974, the 1,447 respondents had a mean score of 6.0 (our of 10) with a standard deviation of 2.2

In 2012, the 1,280 respondents had a mean score of 5.9 with a standard deviation of 2.0

Construct a 99% confidence interval for the difference in population means between 1974 and 2012

Want More?

Parts 1 through 5 of David Lane's book

http://onlinestatbook.com/2/estimation/confidence_ov.html

This section of Jerry Dallal's book

http://www.jerrydallal.com/LHSP/ci.htm

Stat Trek's discussion

http://stattrek.com/estimation/confidence-interval.aspx