Chapter 8 – Estimation

- Estimation
- Review of Central Limit Therum
- Confidence Levels
- Confidence Intervals
- Confidence Interval Precision
  - Standard Error of the Mean
  - Sample Size
  - Standard Deviation

Estimation Defined:

Estimation – A process whereby we select a random sample from a population and use a sample statistic (such as the mean of the sample) to estimate a population parameter (such as the mean of the population).

(1) Why do we use a sample to estimate a population parameter instead of simply using the actual population parameter?

(2) Why not treat the sample statistic as representing the true population parameter?

Point and Interval Estimates

- **Point Estimate** – A single sample statistic used to estimate a population value (for example, a sample mean used as an estimate for the population mean).
- **Interval Estimate** – identifies a range of values that are used as an estimate of a single population value.
- For example, while a point estimate for estimating a population statistic would be a single number such as "3", the interval estimate for estimating a population value would be a range—we would say the actual population value is somewhere between, say, "1 and 5".

Why do we prefer an Interval Estimation?

- We know that if we use the point estimate there will be some degree of error between our sample estimate and the true population parameter. However, we don't know how much error there is.
- If we use an interval estimate (also referred to as a confidence interval), we can calculate the likelihood (or confidence) that the true population parameter falls within the interval we have identified.

Thus, the interval estimate identifies a range of values within which the population parameter is likely to fall. This range is referred to as the confidence interval.

We can determine the likelihood that the population parameter falls within the confidence interval. This is referred to as the confidence level (for example, we might be 95% confident or 99% confident that the actual population parameter falls within our confidence interval).

In other words:

We can evaluate the accuracy of the confidence interval by assessing the likelihood that the interval will contain the actual population parameter.

Typically, the steps taken are to first determine what level of confidence you want to have and then determine from your sample what the confidence interval is for that level of confidence.

For example, we might want to be 99% confident that our confidence interval contains the population parameter. Or, perhaps we decide that we want our confidence level to be 95%.

Once we decide the confidence level we want, we then determine the confidence interval for that level of confidence.

Determining a Confidence Interval (CI)

$$CI = \overline{Y} \pm Z\ (s_{\overline{Y}})$$

If we took a random sample of 150 college presidents and then wanted to know <u>the average age</u> for all college presidents throughout the U.S. (the population), we could use our sample to estimate a range or interval within which the average age for all presidents is likely to fall. To accomplish this:

(1) decide the level of confidence that we want
(2) calculate the standard error for the sample,
(3) multiply it by our confidence level chosen (such as 1.96 if we want to be 95% confident) and then
(4) add and subtract this from the sample mean to obtain our confidence interval.

Determining a Confidence Interval (CI) with a particular level of confidence (e.g., 95% confident or 99% confident)

$$CI = \overline{Y} \pm Z\ (s_{\overline{Y}})$$

where:

$\overline{Y}$ = sample mean
Z = Z score (for example 2 SE's)
$s_{\overline{Y}}$ = estimated standard error

Estimating the standard error

Since the SD of the population is not available to calculate the SE, we use the sample SD in its place and we call it the "estimated standard error".

Estimated Standard Error Based On Large, Random Sample $= s_{\overline{Y}} = \dfrac{s_Y}{\sqrt{N}}$ ← (standard deviation of sample)

Confidence Levels and Corresponding *Z* Values

Confidence Level	*Z* Value
90%	1.65
95%	1.96
99%	2.58

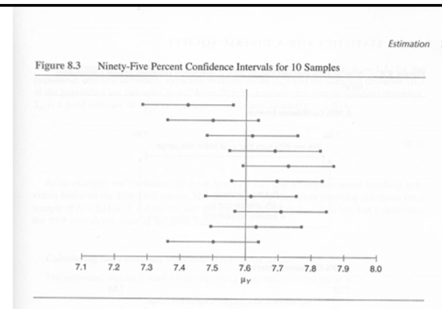
What is the advantage of using a 99% confidence level instead of using a 95% confidence level?

Advantage: we have more confidence that we know the interval (or range) within which the true population parameter is located.

What does it mean to be 99% confident?

In other words, calculating and using a 99% confidence level means that, if we take 100 different random samples, only one of the 100 samples would not provide a confidence interval that includes the true population parameter.

We are willing to take the risk that our one sample is NOT that 1 in a 100 samples but rather is one of the 99 samples whose confidence interval does include the population parameter.



The vertical line represents the true population mean. Each horizontal line represents a single sample and for that sample shows the confidence intervals for a 95% confidence level (i.e., 2 SE's + and – the sample mean)

What is the <u>disadvantage</u> of using a 99% confidence level instead of using a 95% confidence level?

Disadvantage: by using the 99% confidence level, the confidence interval will be wider so we lose "precision" of the estimate.

Reducing risk Taking When Estimating

(that is, how can we improve our chances that the population parameter is within our CI)

$$CI = \bar{Y} \pm Z\left(\frac{s_Y}{\sqrt{N}}\right)$$

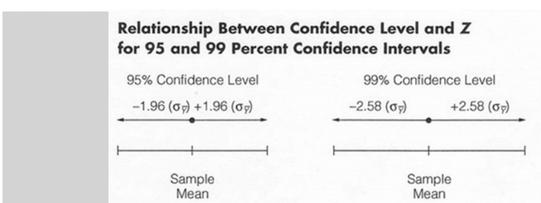
- **Confidence Level** – The confidence level affects the risk of being wrong. For example, increasing our confidence level from 95% to 99% means we are less likely to draw the wrong conclusion (that is, less likely to have a CI that does not include the population parameter).

  – we take a 1% risk (rather than a 5%) that the specified interval does not contain the true population mean.

Confidence Interval and Risk Taking

$$CI = \bar{Y} \pm Z\left(\frac{s_Y}{\sqrt{N}}\right)$$

Unfortunately, when we reduce our risk of being wrong, we will also create a wider range of values . . . ***So the interval becomes <u>less precise</u>.***

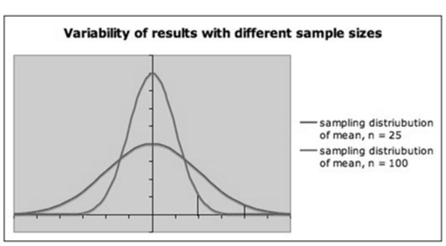


Increasing Precision When Estimating

$$CI = \bar{Y} \pm Z\left(\frac{s_Y}{\sqrt{N}}\right)$$

Sample Size – Researchers can increase the precision of the estimate by increasing the sample size. Larger samples result in smaller standard errors, and therefore, result in the distribution of cases being more closely clustered around the mean.

A more closely clustered distribution results in our confidence intervals being narrower and more precise (i.e., a CI with less distance/a smaller range).



Example: Sample Size and Confidence Intervals

95 Percent Confidence Interval and Width for Mean Income for Three Different Sample Sizes

Sample Size	Confidence Interval	Interval Width	s_Y	$s_{\bar{Y}}$
N = 472	\$27,259-\$31,421	\$4,162	\$23,067	1061.53
N = 945	\$27,869–\$30,811	\$2,942	\$23,067	750.39
N = 1,890	\$28,300–\$30,380	\$2,080	\$23,067	530.64

Standard Deviation and Risk Taking

CI = $\overline{Y} \pm Z\ (\frac{s_Y}{\sqrt{N}})$

A smaller standard deviation results in a more precise confidence interval (i.e., the CI has a smaller range or width).

Standard Deviation and Risk Taking

CI = $\overline{Y} \pm Z\ (\frac{s_Y}{\sqrt{N}})$

Unlike sample size and confidence level, the researcher plays no role in determining the standard deviation of a sample.

Why?

Example: Determining 95% CIs for Hispanic Migration and Earnings

$CI = \overline{Y} \pm Z\ (s_{\overline{Y}})$

From 1980 Census samples:

- **Cubans** had an average income of $16,368 ($S_y$ = $3,069), N=3,895
- **Mexicans** had an average of $13,342 ($S_y$ = $9,414), N=5,726
- **Puerto Ricans** had an average of $12,587 ($S_y$ = $8,647), N=5,908

Example: Determining 95% CIs for <u>Average Cuban Earnings</u> where a sample of Cubans had an average income of $16,368; S_y = $3,069, N=3,895

$(s_{\overline{Y}})$ = S_y / Sqrt of N $\qquad CI = \overline{Y} \pm Z\ (s_{\overline{Y}})$

$s_{\overline{Y}}$ = Standard error = $3{,}069/\sqrt{3895}$ = 49.17

1.96 * 49.17 = $96.37

$16,368 – $96.37 to $16,368 + $96.37
(mean-SE to mean+SE)

= $16,272 to $16,464

Determine 95% confidence interval for Mexicans where a random sample of 5,726 Mexicans had an average income of $13,342 and standard deviation (S_y) of $9,414.

$s_{\overline{Y}}$ = S_y / Sqrt of N $\qquad CI = \overline{Y} \pm Z\ (s_{\overline{Y}})$

In-class assignment: calculate the confidence interval with a 95% confidence?

Determining 95% confidence for a sample of 5,726 Mexicans with an average income of $13,342 and standard deviation (S_y) of $9,414.

$CI = \overline{Y} \pm Z\ (s_{\overline{Y}})$

$s_{\overline{Y}}$ = Standard error = $9414/\sqrt{5726}$ = 124.41
(stand.dev./$\sqrt{N}$) (9414/75.7)

1.96 * 124.41 = 243.8

$13,342 – 243.8 to $13,342 + 243.8
(mean-SE to mean+SE)

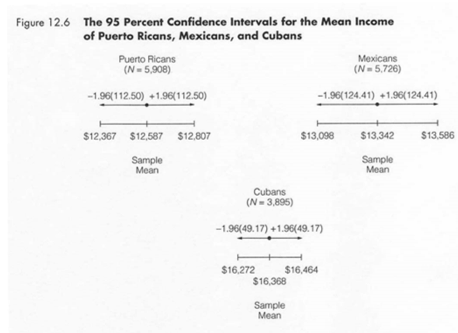
= $13,098 to $13,586

Do the Cuban and Mexican immigrants have the same average incomes?

We are 95% confident that the Cuban average income is between:
\$16,272 to \$16,464

We are 95% confident that the Cuban average income is between:
\$13,098 to \$13,586

Example: Hispanic Migration and Earnings



In Sum:

Now you can <u>estimate</u> a population's parameter (such as the mean) based on a sample.

Simply calculate the standard error, determine the level of confidence that you want, and then calculate the confidence interval using the given formula. This will tell you the probability (eg., 95% probability) that the true population mean falls within your confidence interval.

凩訝

(see you later)