Notes 3: Statistical Inference
Supplemental Presentation Notes

Recall from Notes 1:
1 and 2. Inference, and Population vs. Sample Illustrated

Population – Example: Graduate Class in Statistics with 10 students
Sample – Subset of population (less than entire population selected)

	Population = Entire Class	Age		Sample = Subset of Population (e.g., n = 4)	Age
Student	Name				
1	Bryan	47		Bryan	47
2	Marijke	15			
3	Gunther	13		Gunther	13
4	Marlynn	48			
5	Bob	73			
6	James	73			
7	Diana	67		Diana	67
8	Linda	70			
9	Gary	40		Gary	40
10	Eric	45			
	Mean (Parameter) μ =	49.10		Mean (Statistic) $\bar{X}$ or M =	41.75

Parameter – Example: Mean Age of Population is μ = 49.10; use Greek letters to symbolize parameters
Statistics – Example: Mean Age of Sample is M = 41.75; use Roman (or Latin) letters to symbolize statistics

3. Randomness and Sampling

Types of Samples

Probability – can identify all in population and can calculate probability of selection; uses randomize selection
- Simple Random Sampling – randomized selection with equal and independent chance of selection
  - Example: randomly selecting students from a class using random numbers of picking names from box
- Stratified Sampling – random selection of units from defined strata
  - Example: dividing class by sex then randomly selecting number of males and then females
- Cluster Sampling – random selection of clustering units
  - Example: random selection of 4 classes from among 15 classes at a school

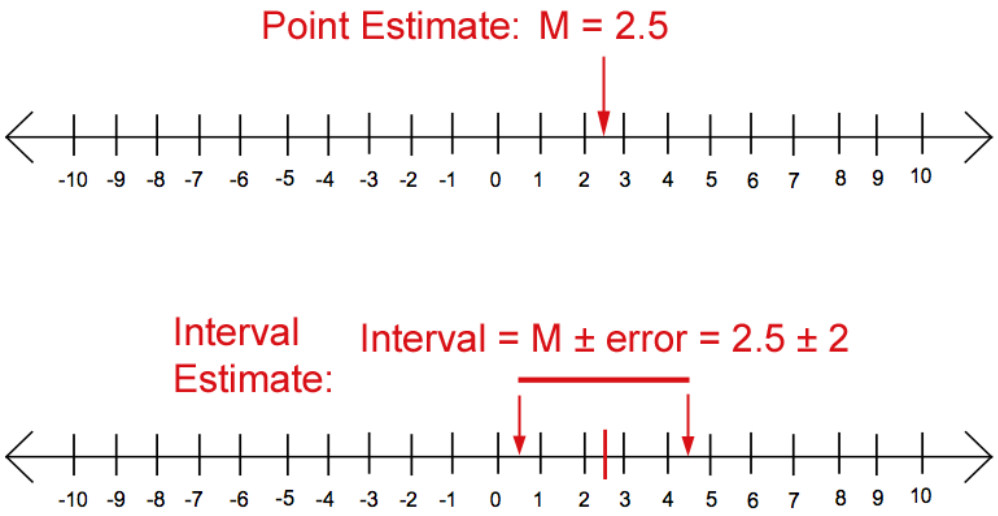
Non-probability – often cannot identify chance of selection or population, may not use randomized selection
- Convenience Sampling – selection based upon availability; randomized selection typically not employed

Sampling error – the difference between a parameter and the statistic used to estimate that parameter where the difference is the result of random chance, random fluctuation, in the sample selected.

Example: M - μ = 41.75 – 49.10 = -7.35

Bias – difference between a parameter and statistic (i.e., M - μ), but the difference is due to systematic problems with sample selection such that the result obtained often either overestimates or underestimates the parameter.

4. Point and Interval Estimates



5. Population Distribution, Sample Distribution, and Sampling Distribution

Population – raw scores in population (census)
Sample – raw scores in sample taken from population
Sampling – distribution of a statistic taken from multiple samples

	Population = Entire Class	Age		Sample = Subset of Population (e.g., n = 4)	Age
Student	Name				
1	Bryan	47		Bryan	47
2	Marijke	15			
3	Gunther	13		Gunther	13
4	Marlynn	48			
5	Bob	73			
6	James	73			
7	Diana	67		Diana	67
8	Linda	70			
9	Gary	40		Gary	40
10	Eric	45			
	Mean (Parameter) μ =	49.10		Mean (Statistic) $\bar{X}$ or M =	41.75

		1st Age Selected	2nd Age Selected	3rd Age Selected	4th Age Selected		Mean for Sample
Sample	1	47	13	67	40		41.75
Sample	2	15	48	73	45		45.25
Sample	3	13	73	67	70		55.75
Sample	4	73	67	40	45		56.25
Sample	5	47	15	67	70		49.75
Sample	6	15	13	67	40		33.75
Sample	7	48	73	67	70		64.50

The *sampling distribution of the mean* is listed above in light gray.

6. Sampling Distribution of the Sample Mean (M or $\overline{X}$)

Variance Error of Sample Mean: variance of sample means = $\sigma_{\bar{X}}^2 = \sigma^2/n$

Standard Error of Sample Mean: standard deviation of sample means = $\sigma_{\bar{X}} = \sqrt{\sigma^2/n} = \frac{\sigma}{\sqrt{n}}$

Calculate the standard error of the sample mean for the following:

(a) M = 25, population SD, σ, is 25, and n = 25.

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ = 25 / 5 = 5.00

(b) Test scores for sample of students: 78, 81, 86, 93, 64, 68, 71, 75, 83 and σ = 9.00

N = 9
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ = 9 / 3 = 3.00

(c) Participants' age: 25, 23, 26, 29, 33, 21, 28, and population variance, σ^2, is 36.

N = 7
σ = SQRT(36) = 6
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ = 6 / SQRT(7) = 3.00 / 2.646 = 1.133

8. Confidence Intervals

Notes 3 Summary

(a) Sampling and Inference

We sample so we can make inferences from sample to population, or more specifically, from statistics to parameters, i.e.,

$$\bar{X} \rightarrow \mu$$

(b) Random Error, Sampling Error

We know that each random sample will produce a statistic that does not precisely equal the parameter it is designed to estimate. The difference between a statistic and its corresponding parameter, if this difference is due to random chance, is known as sampling error, i.e.,

$$Sampling\ Error = \bar{X} - \mu$$

(c) Point vs. Interval Estimate

Statistics are point estimates that don't provide information of imprecision; statistics appear to be precise when in fact we know that statistics are estimated with error. It is more useful to provide both a point estimate, the statistic, and an interval estimate for that statistic to help folks understand the precision of the estimate.

(d) Sampling Distribution of Sample Mean $\bar{X}$

Due to the central limit theorem, we know that with sufficiently large sample sizes the Sampling Distribution of the Sample Mean ($\bar{X}$) will take an approximately normal distribution in shape.

We also know that we can estimate the standard deviation of sample means, called a standard error, using this formula for known population standard deviation:

$$\sigma_{\bar{X}} = \sqrt{\sigma^2/n} = \frac{\sigma}{\sqrt{n}}$$

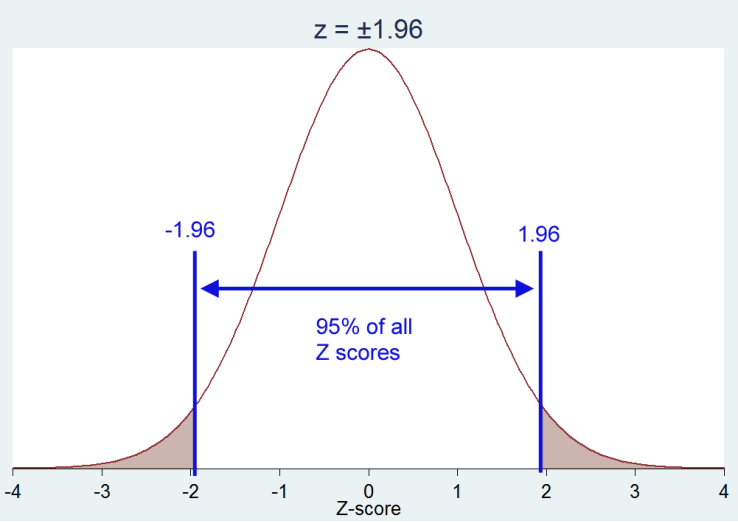
Therefore, since

- statistics are designed to estimate parameters,
- statistics have error when estimating parameters,
- interval estimates provide an indication of the amount of error, or precision of the estimates,
- and we know that the sampling distribution of the sample mean is approximately normal in shape,
- it would be helpful to estimate an interval that takes into account the above to help with estimating parameters.

95% Confidence Interval

Some logic for confidence intervals:

- One may use the Z table to calculate probabilities for select Z scores about the normal distribution,
- about 95% of all Z values lie between ± 1.96,



- an interval constructed with ± 1.96 Z scores about the sample mean should include the population mean, μ, about 95% of the time
- thus, this formula should enable one to construct a 95% confidence interval that will include the population mean 95% of the time:

$$95\%CI\ (or\ .95CI) = \overline{X} \pm 1.96\ \sigma_{\overline{X}}$$

where the two limits for the interval are defined as

Upper Limit: $\overline{X} + 1.96\ \sigma_{\overline{X}}$

Lower Limit: $\overline{X} - 1.96\ \sigma_{\overline{X}}$

Example 1: 95%CI for Mean Age

	Population = Entire Class	Age		Sample = Subset of Population (e.g., n = 4)	Age
Student	Name				
1	Bryan	47		Bryan	47
2	Marijke	15			
3	Gunther	13		Gunther	13
4	Marlynn	48			
5	Bob	73			
6	James	73			
7	Diana	67		Diana	67
8	Linda	70			
9	Gary	40		Gary	40
10	Eric	45			
	Mean (Parameter) μ =	49.10		Mean (Statistic) $\overline{X}$ or M =	41.75
	SD (Parameter) σ =	21.14			

(a) Find standard error of mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{21.14}{\sqrt{4}} = \frac{21.14}{2} = 10.57$$

(b) Calculate upper and lower limited based upon the sample mean:

Upper Limit: $\bar{X} + 1.96\,\sigma_{\bar{X}}$ = 41.75 +1.96 (10.57) = 41.75 +20.72 = 62.47
Lower Limit: $\bar{X} - 1.96\,\sigma_{\bar{X}}$ = 41.75 −1.96 (10.57) = 41.75 −20.72 = 21.03

The 95% CI limits are 21.03 to 62.47.

We can be 95% confident that the unknown population mean age for students in this class lies within the interval of 21.03 to 62.47.

Example 2: 95%CI for Mean SAT

A sample of 50 undergraduate students at GSU reported a mean verbal SAT score of 537. The College Board, producers of the SAT, reports that each section of the SAT has a population SD of 100. Construct a 95%CI for GSU's mean verbal SAT score.

(a) Find standard error of mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{50}} = \frac{100}{7.071} = 14.142$$

(b) Calculate upper and lower limited based upon the sample mean:

Upper Limit: $\bar{X} + 1.96\,\sigma_{\bar{X}}$ =537 +1.96 (14.142) =537 + 27.72 = 564.72
Lower Limit: $\bar{X} - 1.96\,\sigma_{\bar{X}}$ = 537 − 1.96 (14.142) =537 - 27.72 = 509.28

The 95% CI limits are 509.28 to 564.72

We can be 95% confident that the unknown population verbal SAT mean from which these students were sampled lies within the interval of 509.28 to 564.72.

Example 3: 95%CI for Mean IQ

One study reports that the mean IQ for undergraduate students who graduate is 116. The sample for this study was 63. The population SD for IQ is 15. Construct a 95% CI for this mean IQ.

(a) Find standard error of mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{63}} = \frac{15}{7.937} = 1.8898$$

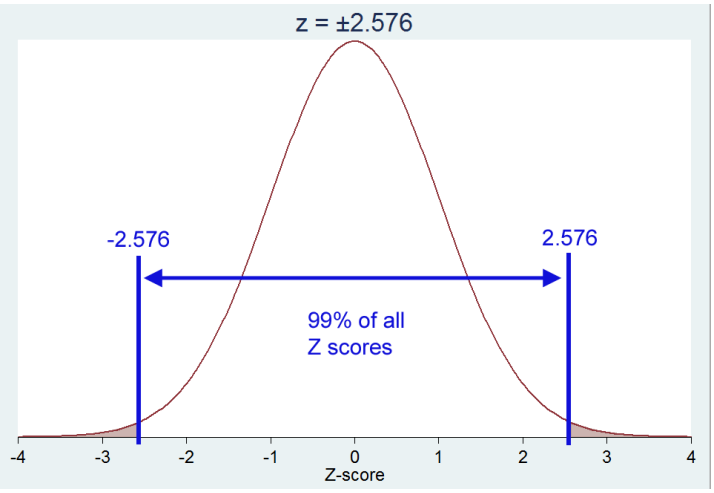
(b) Calculate upper and lower limited based upon the sample mean:

$$\text{Upper Limit: } \overline{X} + 1.96\,\sigma_{\bar{X}} = 116 + 1.96\,(1.8898) = 116 + 3.704 = 119.70$$
$$\text{Lower Limit: } \overline{X} - 1.96\,\sigma_{\bar{X}} = 116 - 1.96\,(1.8898) = 116 - 3.704 = 112.30$$

The 95% CI limits are 112.30 to 119.70

We can be 95% confident that the unknown population mean IQ from which these students were sampled lies within the interval of 112.30 to 119.70.

99% Confidence Interval



Same logic and formula as for the 95% CI, only the Z score used changes:

$$99\%CI\ (or\ .99CI) = \overline{X} \pm 2.576\,\sigma_{\bar{X}}$$

Example 1: 99%CI for Mean Age

Using the class data for age, find the 99% CI about the sample mean of 41.75. Recall that n = 4 and σ = 21.14.

(a) Find standard error of mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{21.14}{\sqrt{4}} = \frac{21.14}{2} = 10.57$$

(b) Calculate upper and lower limited based upon the sample mean:

$$\text{Upper Limit: } \overline{X} + 2.576\,\sigma_{\bar{X}} = 41.75 + 2.576\,(10.57) = 41.75 + 27.23 = 68.98$$
$$\text{Lower Limit: } \overline{X} - 2.576\,\sigma_{\bar{X}} = 41.75 - 2.576\,(10.57) = 41.75 - 27.23 = 14.52$$

The 99% CI limits are 14.52 to 68.98.

We can be 99% confident that the unknown population mean age for students in this class lies within the interval of 14.52 to 68.98.

Example 2: 99%CI for Mean SAT

A sample of 50 undergraduate students at GSU reported a mean verbal SAT score of 537. The College Board, producers of the SAT, reports that each section of the SAT has a population SD of 100. Construct a 95%CI for GSU's mean verbal SAT score.

(a) Find standard error of mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{50}} = \frac{100}{7.071} = 14.142$$

(b) Calculate upper and lower limited based upon the sample mean:

Upper Limit: $\overline{X} + 2.576\ \sigma_{\bar{X}} = 537 + 2.576\ (14.142) = 537 + 36.43 = 573.43$

Lower Limit: $\overline{X} - 2.576\ \sigma_{\bar{X}} = 537 - 2.576\ (14.142) = 537 - 36.43 = 500.57$

The 99% CI limits are 500.57 to 573.43

We can be 99% confident that the unknown population verbal SAT mean from which these students were sampled lies within the interval of 500.57 to 573.43.

Final Note

A serious limitation with the above Confidence Interval formula is reliance upon the population standard deviation, σ, being known. Normally one does not know σ so an alternative formula for the confidence interval will be needed. The formula for unknown σ is introduced with the one-sample t-test.

Excel Example

See link on Course Index for Excel file to test 95% accuracy rate.

9. Margin of Error

Is just the standard error multiplied by critical Z value.

Example – often media will report some is likely to get 49% of the vote in an election, with a margin of error of +/- 2%

This means the range of the estimated vote lies between 47% and 51%, 49-2 = 47% and 49+2 = 51%. So we don't know exactly what percent of the population will vote for this candidate, our best estimate is that percent who will vote for the candidate is between 47% and 51%.