2020

Clear-Sighted Statistics: Module 10: Sampling and Sampling Errors (slides)

Edward Volchok
CUNY Queensborough Community College

Sampling and Sample Errors

Module 10

CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK

0

"The basis of probability-based random sampling is that every member of the population must have a known, non-zero chance of being selected….*Sampling error* results from collecting data from some rather than all members of the population and is highly dependent on the size of the sample."* [Italics added]

-- Pew Research

*Pew Research Center, "Why Probability Sampling," https://www.people-press.org/methodology/sampling/why-probability-sampling/7/

1

Lecture Objectives

- Construct a sampling distribution of sample means
- Describe the implications of the Central Limit Theorem
- Use z-values to find probabilities of obtaining possible sample means, $\bar{X}$, from a normally distributed population

2

Sampling Error (One more time)

Statistic ≠ Parameter

- $\bar{X} \neq \mu$
- $p \neq \pi$
- $s \neq \sigma$

3

Sampling Distribution of $\bar{X}$

The Ubiquity of Sampling Errors

4

Dead Presidents Basketball Team

President	Av. Pts per Game
Washington	34
Adams	8
Jefferson	22
Madison	14
Monroe	12

$$\mu = \frac{\sum X}{N} = \frac{34+8+22+14+12}{5} = \frac{90}{5} = 18$$

5

How many samples of 2 are possible?

Use the Combinations formula

10 Samples

$$_{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{5!}{2!(5-2)!} = \frac{5!}{2(3)!} = \frac{120}{2(6)} = \frac{120}{12} = 10$$

6

Here are the 10 samples with their $\bar{X}$

Sample	Points	X
Washington/Adams	34, 8	21
Washington/Jefferson	34, 22	28
Washington/Madison	34, 14	24
Washington/Monroe	34, 12	23
Adams/Jefferson	8, 22	15
Adams/Madison	8, 14	11
Adams/Monroe	8, 12	10
Jefferson/Madison	22, 14	18
Jefferson/Monroe	22, 12	17
Madison/Monroe	14, 12	13

Samples properly drawn!

7

The 10 Samples and their means

90% of the samples have sampling error

$\bar{X} = \mu$

Sample	Points	$\bar{X}$
Washington/Adams	34, 8	21
Washington/Jefferson	34, 22	28
Washington/Madison	34, 14	24
Washington/Monroe	34, 12	23
Adams/Jefferson	8, 22	15
Adams/Madison	8, 14	11
Adams/Monroe	8, 12	10
Jefferson/Madison	22, 14	18
Jefferson/Monroe	22, 12	17
Madison/Monroe	14, 12	13

$\mu = 18$

Mean of the Sample Means, $\mu_{\bar{X}}$

$$\mu_{\bar{X}} = \mu \quad (\mu = 18)$$

$$\mu_{\bar{x}} = \frac{21+28+24+23+15+11+10+18+17+13}{10} = 18$$

Variability in population is greater than the sample

The sample means draw the data to the “center”

	Highest Value	Lowest Value	Range (H - L)
Population	34	8	26
Sample	28	10	18

Central Limit Theorem (CLT)

A central concept for the discipline of Statistics

History of the CLT

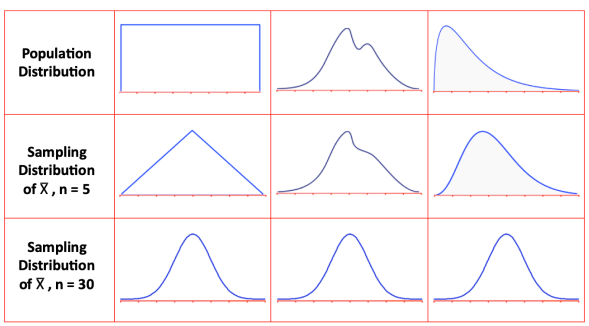
First proven in 1810 by Pierre-Simon Laplace

Revised in 1824 by Siméon-Denis Poisson

Key implications of the CLT

- Sampling distributions of the sample mean become more normally distributed as the sample size increases
- When the population is normally distributed, the sampling distributions of the sample mean will follow a normal distribution
- When the population is symmetrical, but not normally distributed, the sampling distribution of the sample means will emerge with a sample size as small as 10
- When the population is skewed, the normal shape of the sampling distribution will emerge with a sample size as small as 30

CLT Illustration



Based on the CLT

Samples of 30 or more are large enough to apply the CLT

We can assume the sample data follows normal distribution even when the population does not

Calculating z-values for samples

16

z-Values for the sample mean, $\bar{X}$

Population	Sample
$z = \frac{X - \mu}{\sigma}$	$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

17

Parts of the sample formula for z

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- $\bar{X} - \mu$ → Sampling Error
- $\sigma / \sqrt{n}$ → Standard Error of the Mean (SEM)*

*AKA Sigma sub X-Bar = $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

18

Structure of this formula is very important

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- Most NHST* formulas follow this structure
- Sampling error in the numerator
- Standard Error in the denominator

*Null Hypothesis Significance Tests

19

The more variable the data, the larger the SEM

$$\sigma / \sqrt{n} = \sigma_{\bar{x}}$$

$$5 / \sqrt{100} = 0.5$$

$$10 / \sqrt{100} = 1.0$$

$$15 / \sqrt{100} = 1.5$$

20

The larger the sample, the smaller the SEM

$$\sigma / \sqrt{n} = \sigma_{\bar{x}}$$

$$5 / \sqrt{100} = 0.500$$

$$5 / \sqrt{121} = 0.455$$

$$5 / \sqrt{144} = 0.417$$

21

22

CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK

23