2020

Clear-Sighted Statistics: Module 10: Sampling and Sampling Errors

Edward Volchok
CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 10: Sampling and Sampling Errors

> Probability sampling, where a small randomly selected sample of the population can be used to estimate the distribution of an attitude or opinion in the entire population with statistical confidence, had traditionally provided the foundation for survey research and political polling. The basis of probability-based random sampling is that every member of the population must have a known, non-zero chance of being selected. Probability sampling provides the means by which the margin of sampling error can be calculated and the ***level of confidence*** in survey estimates reported. ***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.[1] [italics added]
>
> -- Pew Research

I. Introduction

In Module 3, we reviewed basic probability and non-probability sampling techniques. We explained that sampling is nearly always the only realistic way to learn about a population because compared to taking a census—counting every element in the population—sampling is faster, less expensive, and generally as reliable as a census when properly conducted. We distinguished random sampling errors from systematic errors. We found that systematic errors are due to flaws in the research design, human error, or fraudulent behavior on the part of researchers or respondents. Sampling error—when the parameter of interest does not equal the statistic—is not due to human error. Random sample errors often occur when samples are conducted. In this sense, the risk of random sampling error is ubiquitous.

After completing this module, you will be able to:

- Construct a sampling distribution of sample means.
- Understand why random sampling error is not due to human error.

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

In Module 11 we will turn to constructing confidence intervals for the purpose of estimating unknown population parameters using sample statistics.

II. Random Sampling Error (One More Time)

Sampling error occurs when the statistic does not equal the parameter; that is when $\bar{X} \neq \mu$. Every sample has a risk of sampling error because not every variable is included in a sample. These random errors are not due to carelessness or human errors. In Module 11 on Confidence Intervals, we will set the acceptable limits on sampling error when we estimate unknown population parameters.

III. The Sampling Distribution of the Sample Means

Due to sampling error, the sample mean, $\bar{X}$, varies from sample to sample. The best way to demonstrate this is to construct a sampling distribution of the sample means, which is a probability distribution of all sample means. Let's construct one from a very small population.

Imagine a version of heaven, hell, or alternate universe, where the first five presidents of the United States formed an intramural basketball team. Here are the average points scored per game during the last season for each president on the team:

Table 1: Average Points per Game

President	Points
Washington	34
Adams	8
Jefferson	22
Madison	14

Monroe	12

The population mean for the score per player per game is 18 points.

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

Equation 1: Population Mean

We now create all possible samples of two players. The combinations will help us determine how many samples are possible. This formula is appropriate because the order of selection is unimportant. There are 10 possible samples, as shown in Equation 2.

$$\text{nCr} = \frac{n!}{r!\,(n-r)!} = \frac{5!}{2!\,(5-3)!} = \frac{120}{2(6)} = \frac{120}{12} = 10$$

Equation 2: 10 Possible Samples Using the Combinations Formula

Here is the sampling distribution of the sample means:

Table 2: Sampling Distribution of the Sample Means

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

Of the ten samples, only the sample mean for Jefferson/Madison equals the population mean. This is, therefore, the only sample *without* sampling error. The nine other samples have sampling error.

Let's see what happens when we take the mean of the sample means, $\mu_{\bar{x}}$, which we pronounce as "mu sub X-Bar."

$$\mu_{\bar{x}} = \frac{\text{Sum of the Sample Means}}{\text{Number of Samples}} = \frac{21 + 28 + 24 + 23 + 15 + 11 + 10 + 18 + 17 + 13}{10} = 18$$

Equation 3: Mean of the Sampling Distribution of Sample Means

The population mean, μ, and the mean of the sampling distribution, $\mu_{\bar{x}}$, are equal. When the population mean and the mean of the sampling distribution are not equal, they should be very close.

The variability in the population, as measured by the range, is greater than that in the sampling distribution of sample means.

Table 3: Range for the Population and Sample Distribution of the Sample Means

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

The sampling distribution of the sample means will always have less variability than the population distribution. This is because the sampling distribution of the sample means is created using the sample means that draw the data towards the "center."

IV. The Central Limit Theorem (CLT)

The Central Limit Theorem is a central concept for the discipline of statistics. Without it modern statistical methods would not exist. The CLT was first proven by Pierre-Simon Laplace in 1810. In 1824, the French mathematician, Siméon-Denis Poisson, refined the theorem. The mathematics underlying the CLT are difficult. But we need only concern ourselves with the implications of the CLT.

Here is why CLT has such important implications:

- Sampling distributions of the sample means 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 3.**

Conclusion: Samples of 30 or more are large enough to apply the CLT. We can then assume the sampling distribution of sample means is normally distributed even when the population is not. Figure 1 shows what happens to the sampling distribution when the sample size is increased.

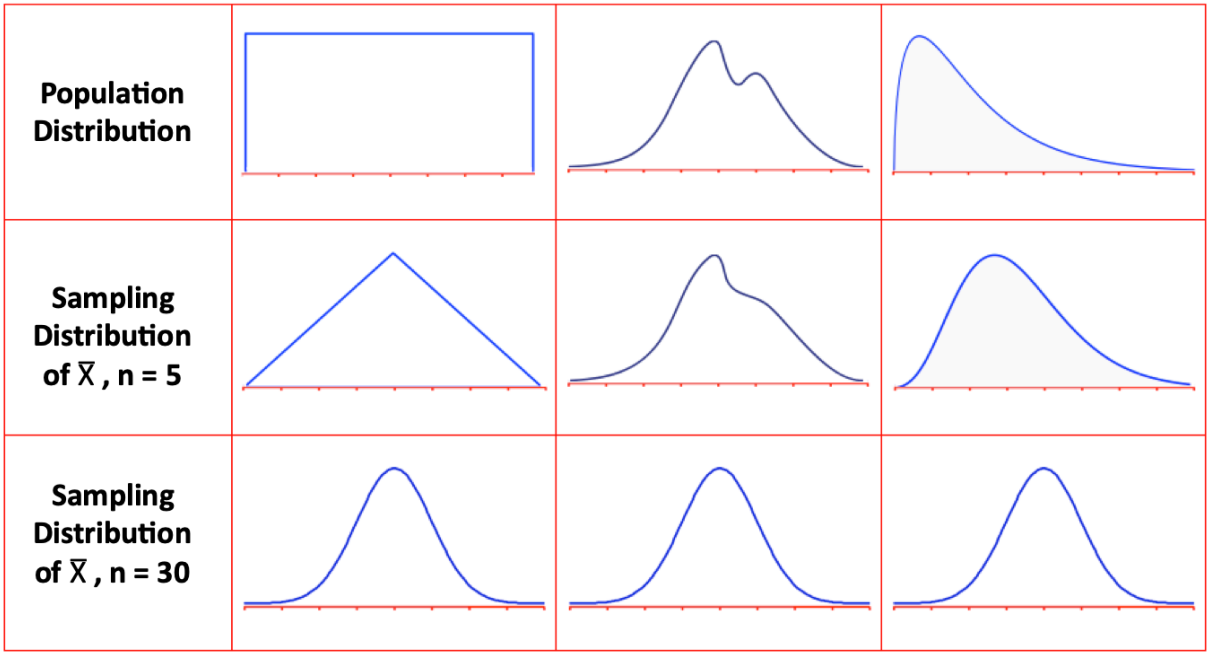

Figure 1: As sample size increases, the sampling distribution of sample means becomes more normal

The CLT has enormous implications: Under most circumstances, we assume the data will be normally distributed. Of course, the assumption of normality should always be tested. In Module 17, Chi-Square Tests, we will cover a test to determine whether the data are normally distributed. When the data are not normally distributed, nonparametric methods we are used to analyze the data. Chi-Square is the only nonparametric method typically covered in introductory statistics class.

V. Standardizing Sampling Error with z-values

In Module 9, we introduced z-values, which require interval or ratio data, and used the formula for a population. Now we modify this formula for comparing a sample mean to a population mean, for measuring sampling error. Table 4 shows the formulas for a population and a sample:

Table 4: Formulas for z-Values for a Population and a Sample

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

Let's examine the new equation for z-values. The numerator, $\bar{X} - \mu$, measures sampling error. The smaller the result, the lower the sampling error. The denominator measures the standard error of the mean, $\sigma_{\bar{X}}$ or SEM:

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

Please note: $\sigma_{\bar{X}}$ is pronounced: "Sigma sub X – Bar"

Equation 4: Formula for the Standard Error of the Mean

With the standard error of the mean, $\sigma_{\bar{X}}$, will be larger when the data are more variable; which is to say, when the standard deviation, σ, is the larger.

Table 5: The larger the standard deviation, σ, the larger the standard error

$\boldsymbol{\sigma} / \sqrt{n} = \sigma_{\bar{X}}$	$\mathbf{5} / \sqrt{100} = 0.5$	$\mathbf{10} / \sqrt{100} = 1.0$	$\mathbf{15} / \sqrt{100} = 1.5$

Similarly, the larger the sample size, n, the smaller the standard error of the mean, $\sigma_{\bar{X}}$.

Table 6: The larger the sample size, n, the smaller the standard error of the mean

$\sigma / \sqrt{\mathbf{n}} = \sigma_{\bar{X}}$	$5 / \sqrt{\mathbf{100}} = 0.5$	$5 / \sqrt{\mathbf{121}} = 0.45$	$5 / \sqrt{\mathbf{144}} = 0.42$

The z-value formula for sample is very important. It is the essence of many formulas used in Null Hypothesis Significance Testing (NHST)—showing the strength of the statistical evidence in an analysis— which we will discuss in later modules. It is a fraction with sampling error in the numerator and the standard error of the mean, or the proportion, in the denominator.

VI. Calculating z-value for Samples

We usually use samples to make decisions about how different a parameter, μ, is from the statistic, $\bar{X}$. When we discuss NHST, we will calculate z-values for samples to determine whether the probability that the difference between $\bar{X}$ and μ is due to sampling error.

Let's calculate a couple of z-values for samples. According to the National Center for Education Statistics, the mean SAT score, μ, in 2017 for high school students in New York State was 1052 with a σ of 188. We take a sample of 36 students at Queens College. This sample shows the sample mean, $\bar{X}$, is 1140. Compare the New York State SAT scores to the survey results conducted among students at Queens College. The z-value for that sample is 2.81.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1140 - 1052}{188/\sqrt{36}} = \frac{1140 - 1052}{188/6} = \frac{88}{31.33} = 2.81$$

Equation 5: z-value for SAT Score of Students at Queens College Compared to the New York State Average

What is the probability of having a z-value as high as 2.81? To answer this question, we look up this z-value on the Area Under the Curve Table. A z-value of 2.81 is 49.75 percent above the mean. Only 0.25 percent of the data are higher than this score. It seems reasonable to conclude that the difference between the Queens College average of 1140 and the New York State average of 1052 is more than what we would expect from sampling error. This difference is likely a statistically significant difference; which means that the difference between $\bar{X}$ and μ is bigger than what we would expect with sampling error.

Area between the Mean and z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986

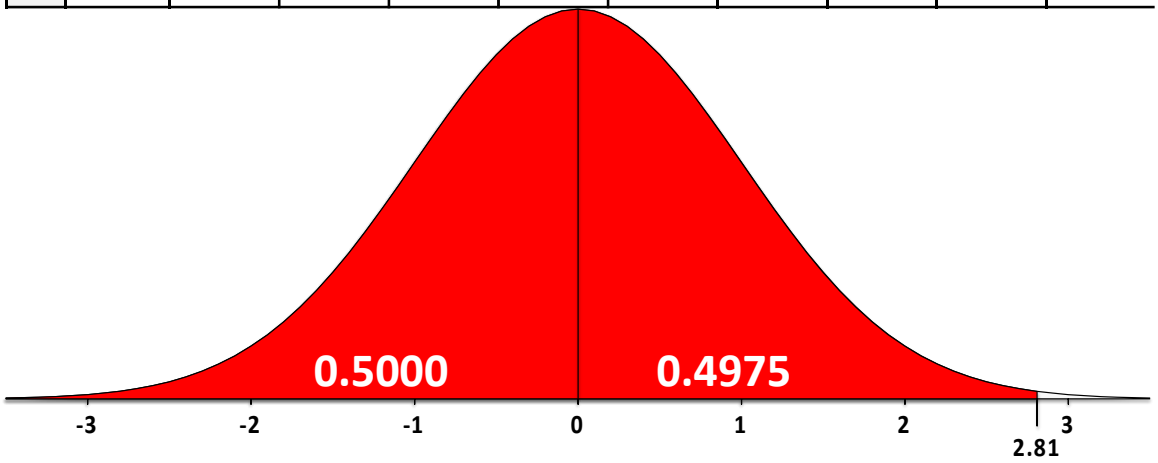

Figure 2: z-value for 2.81 = 0.4975 and a shaded normal curve

Let's calculate the z-value for another college. A survey of 36 students at York College reveals a sample mean of 1040. The z-value these students at is -0.38, which is very close to the population mean for students residing in New York State. Using the Area Under the Curve Table, we conclude that the students at York College are only 14.8 percent below the mean for New York State.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1040 - 1052}{188/\sqrt{36}} = \frac{-12}{188/6} = \frac{-12}{31.33} = -0.38$$

Equation 6: z-value for SAT Scores for Students at York College Compared to the New York State Average

Area between the Mean and z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879

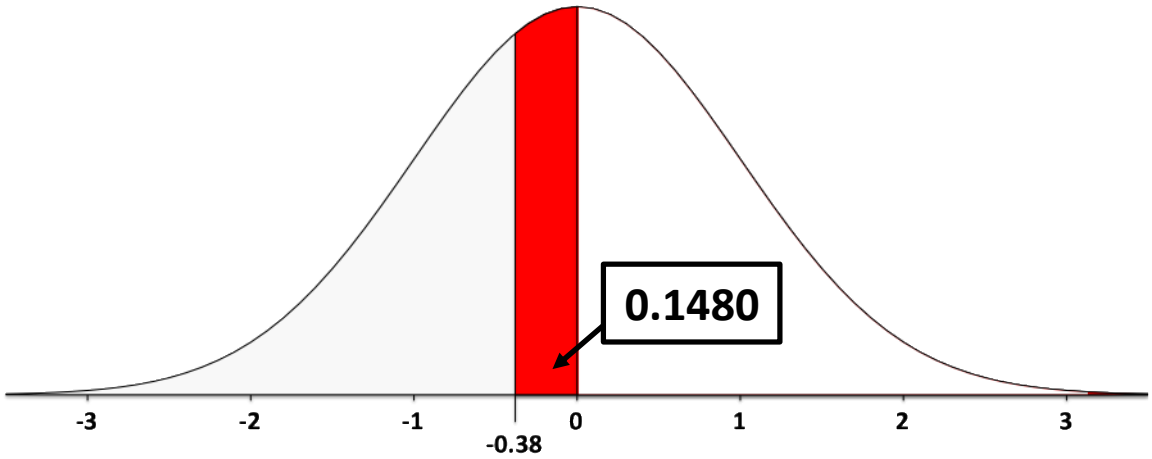

Figure 3: z-value for 0.38 and a shaded normal curve

The difference between SAT scores of 1040 and 1052 appears to be merely sampling error.

IX. Summary

With the z-value formula for samples, we can now begin to conduct inferential statistics or methods of inferring findings about populations using samples. From here on, our focus will be inferential statistics. In our next module, Module 11, we will estimate the population mean, μ, and population proportion (π) using confidence intervals, which are a range of values where, over the long-term, we would expect to find the population parameter. We will also introduce another continuous probability distribution called *student-t*. In Module 12, we will review some fundamental ways of determining sample size, which is an important issue for null hypothesis significance testing. Modules 13 through 17 will cover NHST.

Module 18, which will cover Linear Correlation and Regression, has a number of null hypothesis tests.

X. Exercises

Data for these exercises can be found in 10_Exercises.xlsx.

Exercise 1: Sample Distribution of the Sample Means

After a disappointing season, President Adams, who is 5'7" tall, is voted off the Presidents' basketball team. His replacement is a lanky newcomer, President Abraham Lincoln, who is 6'4" tall. The average points per game for the Presidents' team after President Lincoln's first thirty game are shown in Table 7.

Table 7: Average Points Scored per Game

President	Points
Monroe	10
Madison	14
Jefferson	20
Washington	30
Lincoln	36

A) Calculate the population mean, μ;

B) How many samples of two players from the five players are possible?

C) Calculate the samples means, show your results as a sampling distribution of sample means;

D) How many of the samples have sample error? Which ones, if any, do not?

E) Calculate the mean of the sampling distribution of sample means;

F) Using the range compare variability in the population to variability in the sampling distribution.

Exercise 2: Area Under the Curve

According to the National Center for Education Statistics, the mean SAT score (μ) in 2017 for high school students in New York State was 1052 with a σ of 188. You conducted a random sample of 36 students at five colleges around New York City. Here are your results:

Table 8: SAT Score for Five New York City Area Colleges

College	**n**	**SAT, $\bar{X}$**
Brooklyn College	36	1160
School of the Visual Arts	36	1110
Saint Francis College	36	950
Touro College	36	1015
Vaughn College of Aeronautics	36	1010

A. Calculate the z-values for each school;

B. Report the probability of finding these z-values for each school;

* * *

* * *

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