2020

Clear-Sighted Statistics: Module 12: Estimating Sample Sizes

Edward Volchok
CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 12: Estimating Sample Sizes

> "Once upon a time, there was a little girl named Goldilocks. She went for a walk in the forest. Pretty soon, she came upon a house. She knocked and, when no one answered, she walked right in. At the table in the kitchen, there were three bowls of porridge. Goldilocks was hungry. She tasted the porridge from the first bowl. '**This porridge is too hot!**' she exclaimed. So, she tasted the porridge from the second bowl. '**This porridge is too cold**,' she said. So, she tasted the last bowl of porridge. '**Ahhh, this porridge is just right**,' she said happily and she ate it all up."[1]
>
> -- "Goldilocks and the Three Bears"

I. Introduction

Researchers are like Goldilocks when it comes to the size of their samples. They want them sized just right; not too large, not too small. When samples are too large, money and time are wasted. But when they are too small, researchers are concerned that they might miss uncovering an important effect. We call this a Type II error. Or, the effect "discovered" may only be due to random sample error, which we call a Type I error. Type I and Type II errors will be discussed in the modules on Null Hypothesis Significance Testing. We will see that this is an issue of *statistical power*, the ability of the selected hypothesis test to uncover important findings, given the variability of the data *and* the **size of the sample**. In essence, properly determining the size of the sample helps researchers conduct tests that have sufficient power to find an important effect when one actually exists while avoiding results that show an effect when one does not exist.

In this module we will use z-values to determine the sample size for studies about the population mean, μ, and the population proportion, π. After completing this module, you will:

- Understand the basic method to estimate the necessary sample size for the mean using z-values.
- Understand the basic method to estimate the necessary sample size for the proportion using z-values.
- Be cognizant of the problem of shrinkage.
- Be aware that there are more sophisticated methods for estimating sample size.

This module is accompanied by two Excel files that you should download and use:

- **12_Estimating_n.xlsx**: Excel workbook for estimating sample sizes for the mean and proportion.
- **12_Exercises.xlsx**: Excel workbook to be used for the end-of-module exercises.

II. Factors Affecting the Size of the Sample

Three factors determine sample size:

1. The Chosen Confidence Level

A 95 percent confidence level is used most frequently. Occasionally we see sample sizes based on 90, 98, 99, or 99.9 percent confidence levels. The higher the confidence level, the larger the sample. Here are the z-values for these confidence levels found using the Area Under the Curve table printed on paper and Microsoft Excel:

Table 1: Common Confidence Level Using in Determining Sample Size

Confidence Level	z-value (Table)	z-value (Excel)
90%	1.65	1.645
95%	1.96	1.960
98%	2.33	2.326
99%	2.58	2.576

Please note: Excel's calculations for the critical value are more accurate than those found on a paper critical values table.

2. The Maximum Allowable Error (E)

The *maximum allowable error* (E) is also known as *error tolerance*. It can be that same value as the *margin of error* for the confidence interval. The larger the maximum allowable error, the smaller the sample size. The allowable error is the amount added to and subtracted from the sample mean, $\bar{X}$, or sample proportion, p, to locate the end-points of the confidence interval. The smaller the allowable error, the narrower the confidence interval and the larger the sample size. Sample size increases when the allowable error is decreased.

Without knowing the sample size, it is not possible to calculate the margin of error. Researchers, therefore, have to estimate the amount of error they can tolerate.

3. Variability of the Data (When Estimating Sample Sizes for the Mean)

The variability of the data, as measured by population standard deviation, σ, also affects the size of the sample. The more variable the data, the larger the required sample size. The problem we face is that the population standard deviation is usually not known. Typically researchers use one of three methods to estimate the population standard deviation:

1. Estimate the population standard deviation based on available studies on the topic under investigation.

2. Conduct a pilot study, and use the sample standard deviation, s, as the estimate for the population standard deviation.

3. Based the estimate of the population standard deviation on the empirical rule, which states that nearly all observations will lie plus or minus three standard deviations from the mean. Our estimate of the population standard deviation would be the difference between the highest and lowest values divided by six: (H – L)/6.

III. Estimating the Sample Size for Means

Here is the formula for determining the sample size for means:

$$n = \left(\frac{z\sigma}{E}\right)^2$$

Equation 1: Formula for Calculation the Sample Size for the Mean

Where: z = z-value that corresponds to the selected confidence level
σ = Sample standard deviation
E = Allowable error
n = Sample size

Here is an example of estimating sample size for the mean. The Metropolitan Transit Authority wants to determine how much the average passenger spends on public transportation during a 90-day period. Available data suggest that the standard deviation is $20. The allowable error is $4.50. How large a sample is required?

Table 2: Sample Size Calculations for the Mean at a 95% and 99% Confidence Levels

Steps	**95% CL**	**99% CL**
Step 1:	$n = \left(\frac{1.96 * 20}{4.50}\right)^2$	$n = \left(\frac{2.58 * 20}{4.50}\right)^2$
Step 2:	$n = \left(\frac{39.2}{4.50}\right)^2$	$n = \left(\frac{51.6}{4.50}\right)^2$
Step 3:	$n = (10.356)^2$	$n = (11.47)^2$
Step 4:	$n = 75.88 = 76$	$n = 131.484 = 132$

The required sample size is 76 passengers when using a 95 percent confidence level and 132 when using a 99 percent confidence level. **Please note:** Whenever the answer is not a whole number, we round up to the next highest whole number. **Never round down:** Doing so will result in a sample size that is too small.

While Microsoft Excel does not have built-in sample size functions, it is still a very useful tool for calculating sample sizes. Figure 1 shows the sample size calculations done

using Excel. This problem is in the workbook titled 12_SampleSize.xlsx under the Mean worksheet.

Enter the values for the standard deviation (s) and allowable error (E)
Sample Size Estimation for Means

σ = 20.000
E = 4.500

z-values from found on table

CL	z	σ	E	zσ/E	$n=(zs/E)^2$	Rounded
90%	1.65	20.000	4.500	7.333	53.7778	54
95%	1.96	20.000	4.500	8.711	75.8835	76
98%	2.33	20.000	4.500	10.356	107.2375	108
99%	2.58	20.000	4.500	11.467	131.4844	132

z-values from calculated using Excel

CL	z	σ	E	zσ/E	$n=(zs/E)^2$	Rounded
90%	1.645	20.000	4.500	7.310	53.4428	54
95%	1.960	20.000	4.500	8.711	75.8807	76
98%	2.326	20.000	4.500	10.339	106.9016	107
99%	2.576	20.000	4.500	11.448	131.0597	132

Figure 1: Sample Size Calculations for the Mean in Microsoft Excel

IV. Estimating the Sample Size for Proportions

Here is the formula for determining the sample size for proportions:

$$n = p(1-p)\left(\frac{z}{E}\right)^2$$

Equation 2: Formula for Calculation the Sample Size for the Proportion

Where: z = z-value that corresponds to the selected confidence level
p = Estimate of the proportion based on available data or a pilot study
E = the allowable error
n = Sample size, n

A major pet food company wants to survey dog owners in large metropolitan areas about a new raw meat line of dog food. Available data suggest that 25 percent of households have at least one dog. Researchers at the company want the sample to have an

allowable error of 2.5 percent. They also intend to use a 95 percent confidence level. How large a sample is needed?

Here are the inputs for our formula:

p = 0.25
z = 1.96 for a 95 percent CL, 2.58 for a 99 percent CL
E = 0.025

Table 3: Sample Size Calculations for the Proportion at a 95% and, 99% Confidence Levels

Steps	**95 CL**	**99% CL**
Step 1:	$\mathrm{n} = 0.25(1-0.25)\left(\frac{1.96}{0.025}\right)^2$	$\mathrm{n} = 0.25(1-0.25)\left(\frac{2.58}{0.025}\right)^2$
Step 2:	$\mathrm{n} = 0.25(0.75)\left(\frac{1.96}{0.025}\right)^2$	$\mathrm{n} = 0.25(1-0.25)\left(\frac{2.58}{0.025}\right)^2$
Step 3:	$\mathrm{n} = 0.1875(78.4)^2$	$\mathrm{n} = 0.1875(103.2)^2$
Step 4:	$\mathrm{n} = 0.1875(6{,}146.56)$	$\mathrm{n} = 0.1875(10{,}650.24)$
Step 5:	$\mathrm{n} = 1{,}152.48 = 1{,}153$	$\mathrm{n} = 1.996.92 = 1{,}997$

The answer: The required sample size is 1,153 when using a 95 percent confidence level and 1.997 when using a 99 percent confidence level. **Remember:** Whenever the result of the calculation is not a whole number, **round up any fractional number to the next highest whole number. Never round down:** Doing so will result in a sample size that is too small.

Figure 2 shows the sample size calculations for proportions at various confidence levels.

Sample Size Estimation for Proportions

Enter the values for the proportion (p) and allowable error (E)

p = 0.2500
E = 0.0250

$$n = p(1-p)\left(\frac{z}{E}\right)^2$$

Where:
n = sample size
p = proportion
E = Allowable Error
p = proportion

z-values from table — **Sample Size**

CL	z	p	(1 - p)	p(1 - p)	E	(z/E)	$(z/E)^2$	Sample Size	n Rounded
90%	1.65	0.2500	0.75	0.1875	0.0250	66.00	4,356.00	816.75	817
95%	1.96	0.2500	0.75	0.1875	0.0250	78.40	6,146.56	1,152.48	1,153
98%	2.33	0.2500	0.75	0.1875	0.0250	93.20	8,686.24	1,628.67	1,629
99%	2.58	0.2500	0.75	0.1875	0.0250	103.20	10,650.24	1,996.92	1,997

Figure 2: Calculation for Sample Size for the Proportion

Remember: Round up any fraction number. Never round down.

When we have no estimate of the population proportion, we use 50 percent, 0.50. This will result in the largest possible sample size at the selected level of confidence.

Table 4: Size Calculation for the Proportion Where There is No Estimate of the Proportion

Steps	**95% CL**
Step 1:	$n = 0.5(1-0.5)\left(\frac{1.96}{0.025}\right)^2$
Step 2:	$n = 0.5(0.5)\left(\frac{1.96}{0.025}\right)^2$
Step 3:	$n = 0.25(78.4)^2$
Step 4:	$n = 0.25(6{,}146.55)$
Step 5:	$n = 1{,}536.64$

The required sample size is 1,5367.

V. Shrinkage

The formulas shown above provide the minimum sample sizes for the assumption used in the calculations. Remember that whenever we conduct research with human subjects, we

get non-response errors because people drop-out of the study or they fail to answer important questions for a variety of reasons. Non-response error, in effect, reduces the size of the sample. This is called shrinkage. Good researchers must ensure that the sample size accounts for non-response errors. Techniques used to do this are beyond the level of an introductory statistics course.

VI. Limitations of these Methods and G*Power

The two methods shown above for determining sample size are limited to normal distributions for the mean and proportion when using z-values is appropriate. As we will discuss in future modules on Null Hypothesis Significance Testing (NHST), using z-values is not always appropriate. In addition, important considerations for NHST like statistical power and the probability of a Type II error are not considered with these calculations. Statistical power is the ability of a test to detect an effect when one exists. A Type II error is a false negative or the failure to reject the null hypothesis when there is a an effect. These terms will be explained in the next module.

There is a wonderful free software package called G*Power that calculates sample size requirements for a wide variety of analyses: z-tests, t-tests, F-tests, Chi-Square tests, linear regression, logistic regression, as well as a variety of nonparametric tests. G*Power reports the sample size needed to achieve a certain level of statistical power. We will discuss what statistical power is in our next module. Statistical power is the ability of a statistical analysis to detect an important effect when one actually exists.

We will use G*Power to select sample size in the context of NHST.

VII. Summary

We have completed our discussion of calculating the sample size for analyses of the mean and proportion. We are now ready to begin our discussion of hypothesis testing.

VIII. Exercises

Complete the following problems. You can use Excel to calculate your solutions by using the file titled Module12_Exercises.xlxs.

Exercise 1: The Association of Used Car Dealers wants to study the maintenance costs for the first year after cars have come off the manufacturer's warranty. How large a sample is needed when s = $575 and E = $100? Calculate the sample size using a 90%, 95%, 98%, and 99% confidence level.

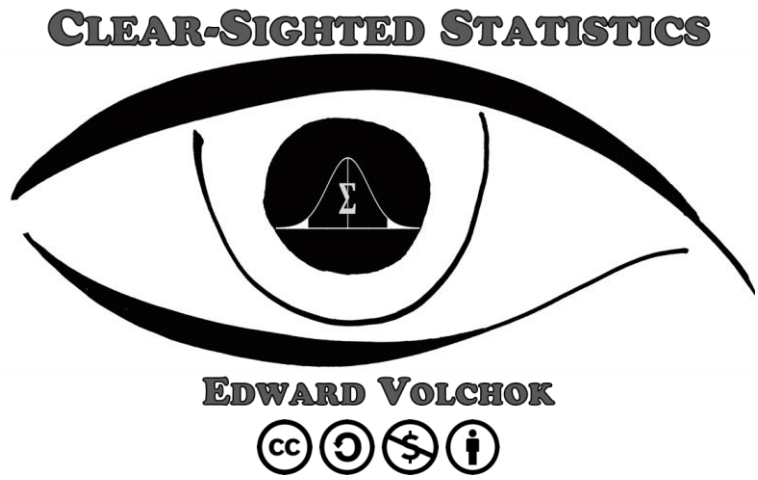
Exercise 2: The national law firm of Dewey, Cheatem, and Howe intends to conduct a study of salaries for newly graduated lawyers who have just passed a state bar exam. How large a sample is needed if we assume the standard deviation is $20,555 and the allowable error is $3,500? Calculate sample sizes using 90%, 95%, 98%, and 99% confidence levels.

Exercise 3: The American Association of Realtors and the U.S. Association for the Paranormal are interested in conducting a survey among people who believe that ghosts are real. The results of a HuffPost/YouGov poll suggest that 43% of the population think ghost are real. How large a sample is required if you set the proportion at 43% and the allowable error at 3%? Calculate a 90, 95, 98, and 99% levels of confidence. Then conduct the calculations assuming that you have no information of the proportion of the population that believes in ghosts.

Exercise 4: The American Association of Realtors and the US Association for the Paranormal are interested in conducting a survey among people who believe ghosts

are real and not dangerous. The results of a HuffPost/YouGov poll suggest that 30% of the population thinks ghosts are real and not dangerous. These people would be willing to buy a home that is haunted. How large a sample is required? Calculate the sample size using a 90, 95, 98, and 99% levels of confidence with an allowable error of 0.03.

* * *



* * *

[1] "Goldilocks and the Three Bears," https://www.dltk-teach.com/rhymes/goldilocks_story.htm.