

A Sample/Population Size Activity:
Is it the sample size or the sample as a fraction of the population that matters?

By Margaret H. Smith*¹

Unless the sample encompasses a substantial portion of the population, the standard error of an estimator depends on the size of the sample, but not the size of the population. This is a crucial statistical insight that students find very counterintuitive. After trying several ways of convincing students of the validity of this principle, I have finally found a simple memorable activity that convinces students beyond a reasonable doubt. As a bonus, the data generated by this activity can be used to illustrate the central limit theorem, confidence intervals, and hypothesis testing.

The Problem

I have often led students in my introductory, college-level statistics classes through a discussion of how to conduct a poll of student opinion on some current controversial topic. The population might be the students at a large university, the students residing in a given state, or all students in the United States. At some point, I ask them how many students they would survey. Most base their answer on the size of the population; for example, “We should survey 10 percent of the students.” This reasoning also appears if I initially focus the discussion on a single university and then ask how many students should be surveyed if our population consisted of all the students in the state. Almost all students increase the sample size, because they believe that a larger population requires a larger sample. The only naysayers are persons familiar with actual public opinion polls who point out that only a few thousand voters are surveyed in presidential election polls.

I have tried many ways of explaining that it is the absolute size of the sample, not the size of

¹ I want to thank the editor and referees for their very helpful comments and suggestions.

the sample relative to the population, that matters for our confidence in an estimate. For example, I point out that we could surely get a pretty accurate estimate of the probability of heads from a few thousand coin flips, even though the population is theoretically infinite. Whether judged by their skeptical looks or their answers to test questions, many students remain unconvinced by these arguments. I now begin this discussion not with abstract arguments, but with a hands-on activity that lets students discover the principle for themselves.

The Activity

Early in the semester students do the homework exercise described in Appendix I using Smith's Statistical Package (SSP) <<http://www.economics.pomona.edu/StatSite/framepg.html>>, a free program that they can use with virtually no instruction; the exercise can also be done with most commercial statistics software packages, a spreadsheet program like Excel, or pocket calculators like the TI-83. The exercise can easily be modified to reflect a public opinion poll or some other design that appeals to students.

Students are told to fill in the 12 cells labeled A-L in the table in Appendix I and to answer the remaining questions with the understanding that they will be allowed to revise their answers after the classroom discussion. A very nice feature of this exercise is that individual students cannot see the overall answer in their results; it is only when students compile their collective data that the conclusions emerge. It is also nice that the parallel data collection by individual students drives home the elusive point that any random sample is but one of many samples that might have been obtained. And the conclusion is made more convincing by the fact that the students themselves gathered the data, rather than having the professor present results that might have been chosen selectively. It is much more powerful when an audience member reaches into his own hat and finds a rabbit.

When the students come to class with their results, I put the table in Appendix II on the board

and have all the students come to the board and put vertical marks in the appropriate spaces—thereby creating 12 bar graphs. As a class, we sit and stare at these graphs for a few minutes and jot down one or two observations. Then we have a discussion for 30 minutes and develop the idea that it is n , the absolute size of the sample, and not n/N , the size of the sample relative to the population, that determines the dispersion of the sample results.

I have found it useful to have students compare the graphs in three different ways. First, I have them increase the sample as a fraction of the population, n/N , from .10 to .25 to 1.00 (graphs A, B, C; E, F, G; or J, K, L). They all see that the sampling distribution becomes more concentrated around p as n/N rises. I want them to make these comparisons first because it is what they intuitively expect to see.

Second, I have them compare the sampling distributions for different values of n , holding n/N constant. First we compare $n = 40$, $N = 400$ (graph A); $n = 100$, $N = 1000$ (graph E); and $n = 1000$, $N = 10,000$ (graph J). Even though each sample is 10% of its population, the spread of these three sampling distributions varies quite a bit depending on the sample size, n . Then they compare the sampling distributions for the 25% samples (graphs B, F, K) and notice again that the accuracy of the sample does not seem to depend on n/N but rather on n . In each case, the sampling distribution becomes more concentrated around p when n increases.

Finally, I have them compare the sampling distributions for a given sample size n , varying the size of the population N . First we look at $n = 40$, with the population size varying from 400 to 1000 to 10000 (graphs A, D, and H). They see that the sampling distributions are similarly shaped, even though n/N is varying from 10% to 4% to 0.4%. This observation—that it is n , rather than n/N , that determines the shape of the sampling distribution—is confirmed by looking at sample size $n = 100$ for population sizes ranging from 400 to 1000 to 10,000 (graphs B, E, I).

In this way, I try to convince them that n is what drives our confidence in sample estimates,

because the distribution of \hat{p} narrows with n , not n/N . It helps to concede to them that if n/N is very large, of course the \hat{p} distribution will narrow considerably, which they all seem to understand intuitively (especially when they think about $n = N$!). This discussion gives student a richer appreciation of sampling distributions and the importance of n .

Reusing the Data

Another nice feature of this activity is that the data can be recycled to illustrate other points throughout the course, such as the central limit theorem, confidence intervals, and hypothesis testing. Not only do all the virtues described above reapply, but there is a very nice continuity to the discussion.

Central Limit Theorem

A rough bell shape can usually be seen in the student data. I tell the students that I have supplemented their data for $n = 1000$, $N = 10,000$ by drawing another 500 samples and collected the results in a data file that they can download from the course web site in order to do the homework assignment shown in Appendix III. (I use 500, instead of 1000, samples because the students can too easily become confused between the 1000 samples and the sample size of 1000). We can imagine that these additional data are what the results might look like if five hundred other students did the same activity. (After this activity has been done enough times, I will actually be able to combine the student data for many classes over the years.)

In addition to addressing the central limit theorem, this exercise gives students an opportunity to work with standardized variables.

Confidence Interval

The dispersion of the student results for each combination of n and N nicely illustrates two central principles of sampling. First, the sample results depend on the specific random sample chosen and therefore are subject to sampling error. Second, the fact that there is a distribution of

the sample results allows us to draw inferences from a single sample. In particular, the sample proportion is probably close to the population success probability. This observations sets up my promise that we will later see exactly how to quantify that probability—both theoretically and by using a lot more random samples than are currently on the board. Appendix IV shows a homework assignment that brings out these details.

Hypothesis Tests

Often, one or two students will report back a sample result that is plainly implausible. For example, that a sample of 1000 from a population of 10,000 with a success probability of 0.40 had a sample success frequency of 0.35. Presumably, the students made a mistake using the software or recording their results.

Such a mistake provides a wonderful opportunity to introduce hypothesis testing. As the students look at the results on the board, someone may remark how different one student's result is from everyone else's. If not, I make that observation. Either way, there is some laughter, attempts to identify the student with the odd result, and some nervous deer-in-the-headlights looks from some students.

I reassure everyone that no dire consequences are imminent, and then ask them to consider the plausibility of the odd result: "Maybe they just had an unlucky sample? Maybe, by the luck of the draw, they selected too many males? Or maybe there was a mistake in using the software or recording their results?" After a brief pause, I ask the punch line, "How could we draw a reasonable conclusion about whether this sample with only 35% successes was (a) an unlucky draw from a population with a 0.40 probability of a success; or (b) something else (like an inadvertent draw from a different population or a clerical error)?"

If no one recommends the hoped-for procedure, I suggest it: "If this was a random sample of size 1000 from an infinite population with a 0.40 success probability, what would be the

probability of so few successes?”

Using the binomial distribution, the exact probability of so few successes (fewer than 351) is 0.000645. The two-sided p-value is the probability that the number of successes would be so far from 40% (35% or fewer or 45% or more) is $2(0.000645) = 0.00129$. We can also use a normal approximation with or without a continuity correction to estimate this probability.

Does It Work?

To test the effectiveness of this activity, I asked the 30 students in my Fall 2003 introductory statistics class the two questions in Appendix V at the beginning of the semester and after doing the activity. The first statement is intended to represent the intuition students bring to a statistics class; the second statement is the intuition I would like them to have after taking the class. The distribution of before and after answers are shown in Tables 1 and 2. At the start of the semester, 53% of the students answered “This seems right” to Question 1 and 67% answered “I do not believe this” or “I am skeptical” to Question 2. After the activity, 73% answered “I do not believe this” to Question 1 and 87% answered “I believe this strongly” to Question 2.

Because the answers are categorical, the chi-square statistic can be used to assess whether the differences between the before and after responses are statistically persuasive. The exact p value for the chi-square statistic can be computed using the multivariate hypergeometric distribution (Agresti and Wackerly 1977). For Question 1, the exact p value is 4.3×10^{-7} ; for Question 2, the exact p value is 4.6×10^{-14} .

Reference

Agresti, A., & Wackerly, D. (1977). Some exact conditional tests of independence for $r \times c$ cross-classification tables. Psychometrika, 42, 111-125.

Appendix I Sample Size Assignment

How do samples of different sizes perform in terms of giving an accurate representation of the underlying population? Does it depend on how large the sample is relative to the size of the population or does it depend on the absolute size of the sample? Let's see. We will look at three populations that are 40% female and 60% male. First let's examine a population of size 400 ($N = 400$) and assume the students numbered 1-160 are female and the students numbered 161-400 are male. Second, assume a population size 1,000 ($N = 1,000$) and assume the students numbered 1-400 are female and the students numbered 401-1000 are male. Third, examine a population of size 10,000 ($N = 10,000$) and assume the students numbered 1-4000 are female and the students numbered 4001-10,000 are male. For each of these three populations, record the fraction of females you get in the samples of various sizes in the cells labeled A-L below. (Do not fill in any cells other than A-L.)

(HINT: Use SSP. First type in the variable names A-L at the top of the first 12 columns (in place of var1, var2, etc.) Then choose Uncertainty on the tool bar and choose the Random Numbers option. Choose "Draw a random sample without replacement..." Enter the population size in the first box and the sample size in the second box and choose to put your random sample in the spreadsheet. Finally, select the appropriate variable (A-L) for saving the data. The Numerical Order option on the screen will allow you to quickly determine how many females are in your sample. Repeat the process for all 12 samples.)

population N	400		1000		10,000	
Sample n	n/N	% female	n/N	% female	n/N	% female
40	40/400 = 0.10	A	40/1000 = 0.04	D	40/10000 = 0.004	H
100	100/400 = 0.25	B	100/1000 = 0.10	E	100/10000 = 0.01	I
250			250/1000 = 0.25	F		
400	400/400 = 1.00	C				
1000			1000/1000 = 1.00	G	1000/10000 = 0.10	J
2500					2500/10000 = 0.25	K
10000					10000/10000 = 1.00	L

Think about the following questions. Answer them the best you can by anticipating what you will see in class. And then modify your answers after you see the results for the whole class.

- a) Compare your results to others in the class. For which cells are the results similar and for which cells are they not? Explain.

- b) What do you notice as you compare the distributions of A to B to C or the distributions of E to F to G or the distributions of J to K to L? Explain what is going on as you move from low n/N samples to samples with $n/N = 1$?
- c) What do you notice as you compare these 10% samples: A, E, and J?
- d) What do you notice as you compare these 25% samples: B, F, and K?
- e) What do you notice as you compare these 3 distributions with sample size $n = 40$: A, D, and H?
- f) What do you notice as you compare these 3 distributions with sample size $n = 100$: B, E, and I?
- g) What happens to the concentration of the sampling distribution as n rises, holding n/N fixed (see questions c and d)?
- h) What happens to the concentration of the sampling distribution as n/N rises, holding n fixed (see questions e and f)?
- i) In general, for the more typical situation where the population is very large and the sample is less than a quarter of the population, does the accuracy of the point estimate depend on how big the sample is as a fraction of the population or on the absolute size of the sample?
- j) Based on what you learned from the exercise above, is this statement true or false?: “A 10% sample should be sufficient in all situations. For a population of size 400, a survey with 47 people should be more than enough to get a sample estimate that is representative of the population.”

Appendix II Classroom Table of Results

population N	400			1000				10,000				
sample n	40	100	400	40	100	250	1000	40	100	1000	2500	10000
n/N	.10	.25	1.00	.04	.10	.25	1.00	.004	.01	.10	.25	1.00
	A	B	C	D	E	F	G	H	I	J	K	L
.1251-.175												
.1751-.225												
.2251-.275												
.2751-.325												
.3251-.375												
.3751-.425												
.4251-.475												
.4751-.525												
.5251-.575												
.5751-.625												
.6251-.675												

Two Observations:

- 1.
- 2.

Takeaways:

- 1.
- 2.
- 3.
- 4.

Appendix III Central Limit Theorem Assignment

Discuss the sample estimates of percentage female for sample size, $n = 1000$ and population size $N = 10,000$ when $p = .40$, using data in HW6.xls. (These data are exactly like the data you collected in HW5.

- a) Draw a bar graph for \hat{p} with \hat{p} on the x-axis and %frequency on the y-axis.
- b) Explain how this bar graph compares to the theoretical binomial probability distribution of X/n when $p = .4$ and $n = 1000$.
- c) Explain why \hat{p} can be considered a random variable. What kind of random variable is this sample estimate? A mean or a proportion?
- d) Explain why this histogram is expected to have a normal distribution using the concept of the Central Limit Theorem.
- e) Calculate the mean and standard deviation of the random variable, \hat{p} .
- f) Standardize the random variable, \hat{p} , to have a mean of zero and a standard deviation of one. (Subtract the mean of the \hat{p} values from each \hat{p} , and divide each difference by the standard deviation of the distribution of \hat{p} .) Explain how to interpret the standardized values.
- g) Draw the bar graph for Z , the standardized \hat{p} , from part (f), with Z on the x-axis and %frequency on the y-axis. How does this graph compare to the graph in part (a) above? How is it similar and how is it different?
- h) What fraction of the values of \hat{p} is within one standard deviation of the mean of \hat{p} ? What fraction is within two standard deviations of the mean of \hat{p} ?
- i) Are your two histograms from (a) and (g) perfectly normally distributed? How can you use the information from (h) to tell?
- j) Why are they not perfectly normal?
- k) Summarize the main thing you learned from this exercise.

Appendix IV Confidence Interval Assignment

Suppose that the \hat{p} data in HW6.xls are several (500 to be precise) random samples of size 1000 from a population of 10,000 where $p = .4$ (40%) of all voters prefer candidate A.

- a. Plot the data in a bar graph where x/n is on the x-axis and % frequency is on the y-axis. What do you expect the shape to be? Draw a diagram of what you expect, labeling the axes appropriately. Explain why you expect this shape. Does the actual distribution of x/n look exactly like what you expected? How similar or different is it to what you expected?
- b. What is the mean of your distribution of x/n ? Label both the empirical mean (the average of the x/n values) and theoretical mean (p) on your graph (they might be quite close).
- c. What is the theoretical standard deviation of your histogram (for $p = 0.40$ and $n = 1000$)? Label ± 1 theoretical SD and ± 2 theoretical SD from the theoretical mean on your histogram.
- d. What fraction of your 500 \hat{p} or x/n values actually falls within 2 theoretical standard deviations of the true p ?
- e. Now calculate a 95% confidence interval (lower and upper bounds) for each of your 500 \hat{p} estimates of p .
- f. How many of your 500 confidence intervals actually capture the true p of 40%? Is this what you expected? Explain carefully.
- g. Draw a line under the histogram demarcating the position of the point estimate and width of its associated confidence interval for each of the confidence intervals that does not capture $p = 40\%$. What do you notice about the distance of those point estimates from the $p = 40\%$?
- h. Interpret the meaning of a 95% confidence interval for any one particular point estimate, based on what you've learned above.
- i. In practice, people do not know the true population mean and do not have several sample estimates of p ; rather they have only one sample estimate of p . Relate this exercise to a real world situation where *The Wall Street Journal* might take one political poll of 1000 people to anticipate the result of the next presidential election. What would be the margin of error for this poll (for a 95% confidence interval)? Explain how you would interpret the result of this political poll now that you understand about sampling distributions. How confident would you be in the results of this poll? Explain.

Appendix V Student Questionnaire

For each of these two statements, please underline the response that most closely represents your viewpoint:

1. “You need to obtain a sample that is at least 10% of the population in order to get a reliable estimate of the population parameter.”

1. I do not believe this
2. I am skeptical
3. Neutral/not sure
4. This seems right
5. I believe this strongly

2. “For large population sizes, the size of the population is irrelevant to the reliability of the sample estimate; what matters is the absolute size of the sample.”

1. I do not believe this
2. I am skeptical
3. Neutral/not sure
4. This seems right
5. I believe this strongly

Table 1 Distribution of Responses to Question 1

	Before	After
I do not believe this	3	22
I am skeptical	4	2
Neutral/not sure	7	3
This seems right	16	2
I believe this strongly	0	1
total	30	30

Table 2 Distribution of Responses to Question 2

	Before	After
I do not believe this	11	1
I am skeptical	9	0
Neutral/not sure	6	0
This seems right	4	3
I believe this strongly	0	26
total	30	30