Practitioner's Guide To Statistical Sampling: Part 4

By **Brian Kriegler**
January 11, 2018, 10:43 AM EST

This is the final installment of four articles on statistical sampling in practice. In part 1 and part 2, I demonstrated that random sampling is a neutral and unbiased process. I also presented multiple methods for computing reliable confidence intervals. In part 3, I offered various solutions for dealing with missing sample selections so that statistical inferences remain valid. I discuss several common misperceptions about random sampling requirements in part 4 below.

Brian Kriegler

Three Myths About Random Sampling Requirements

We have established that random sampling is a neutral and unbiased scientific process. Every observation in the population has a known probability of selection. The "Reference Guide on Statistics" points out that "[p]robability sampling ensures that within the limits of chance … the sample will be *representative* of the sampling frame."[1] Random sampling therefore gives practitioners the "green light" to estimate population characteristics along with the applicable confidence intervals and margins of error.[2] Decades of research, textbooks and articles have shown that random sampling works.

Some practitioners nevertheless insist on requiring additional criteria to justify that statistical inference is permitted. It is tempting to prioritize having a "representative" sample.[3] But what does it mean for a sample to be "representative?" Well-known statistician David A. Freedman and attorney David H. Kaye, who authored the reference guide, acknowledge that "representative" is not a well-defined technical term.[4] Suggesting unnecessary conditions may be as innocent as a misunderstanding or intentional in an effort to confuse the audience. Three common perceived requirements include the following:

- Characteristics from a random sample must "mirror" those in the population
- The random sample must be large enough to produce a "sufficiently small" margin of error
- All observations must have the same probability of selection

If any of these criteria are not met, then the presumption is that the sample is biased and cannot be used to draw inferences about the population. In this article, we debunk these "myths."

We will refer to a "Medicare reimbursement example" over the course of this article.[5] The objective is to determine the extent of ineligible reimbursements submitted by a large medical provider. Medicare reimbursement ineligibility is determined by examining patient files. The population consists of 10,000 patients. It is prohibitively expensive to examine each patient’s file. Instead, total ineligible reimbursements can be estimated based on a random sample of patients. Suppose that 50 patients are randomly selected and analyzed; no patient files are missing. We will rely on this fact pattern to demonstrate that the above so-called criteria are not needed.

Myth 1: Characteristics from a random sample must “mirror” those in the population.

Assume that the patient database lists the geographic location for each patient in the population. We can observe data characteristics in both the population and sample.

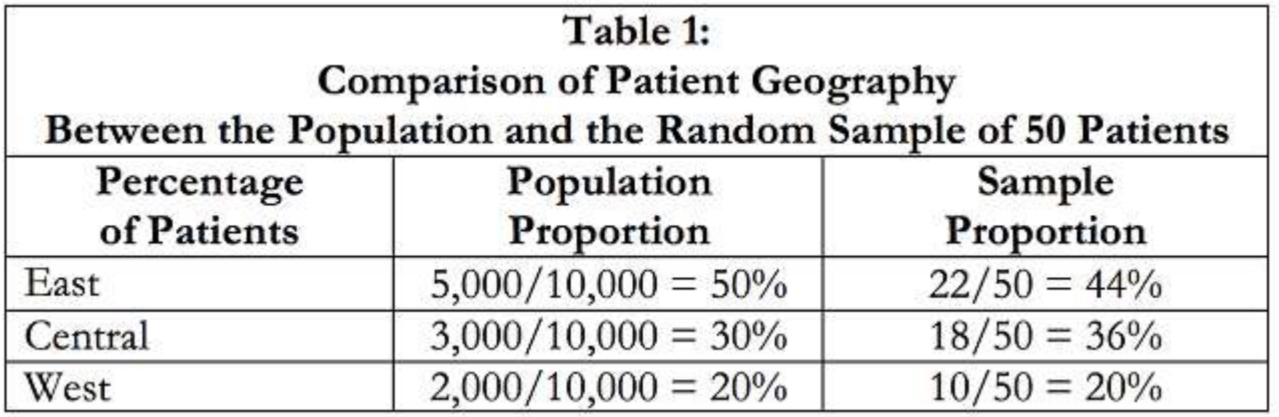

Table 1:
Comparison of Patient Geography
Between the Population and the Random Sample of 50 Patients

Percentage of Patients	Population Proportion	Sample Proportion
East	5,000/10,000 = 50%	22/50 = 44%
Central	3,000/10,000 = 30%	18/50 = 36%
West	2,000/10,000 = 20%	10/50 = 20%

The sample and population are not identical. Sample and population proportions differ by six percent in the Eastern and Central U.S. This raises two questions.

Do these differences demonstrate that the sample was not randomly selected?

The answer is no. One cannot conclude anything about how the data were selected based on a table of results. Determining whether the sample was random typically requires analyzing the computational program used to select the random sample. This program can be rerun to verify how the sample was selected if need be.

Could these differences simply be due to chance?

The answer is yes. No random sample is expected to look exactly like the population. One or more statistical tests are commonly used to gauge whether the sample could plausibly come from a defined population.[6] In this instance a “chi-square test” yields a probability of 61 percent. The interpretation is that six out of 10 random samples will produce samples that deviate from the population at least this much. The disparities between this sample and the defined population are not large enough to be concerned about the sampling process.[7]

In summary, showing the computational program and describing the results from the statistical hypothesis test(s) can provide strong evidence that the sample was selected at random. That confirmation is sufficient for purposes of calculating population estimates and margins of error.

Myth 2: The random sample must be large enough to produce a “sufficiently small” margin of error.

Another common misconception is that the sample must produce an estimate along with a prespecified degree of precision. The statistical practitioner oftentimes sets a goal for the margin of error. Suppose the target margin of error of +/- 10 percent.

Keeping with the Medicare reimbursement example, assume that claims for 25 of the 50 randomly selected patients were ineligible for reimbursement. This yields a population estimate of 50 percent and a margin of error of approximately 13.9 percent.[8]

Some practitioners would consider this to be an intolerably high margin of error. That may be true given the subject matter. However, it would be incorrect to conclude that the sample was not randomly selected or that the applicable confidence intervals were unreliable. The proper interpretation is that a larger sample is required to achieve a tolerable margin of error. In the absence of a larger sample, we are 95 percent confident (certain) that claims for 36.1 to 63.9 percent of patients were ineligible.

Myth 3: The probability of selection must be the same for all observations.

Random sampling is synonymous with probability sampling because every observation in the population has a known probability of selection. There is no prerequisite that each observation has the same probability of selection. We will demonstrate below how unequal probabilities of selection are incorporated into population estimate computations.

Let’s tweak the sampling design and objectives in the Medicare reimbursement example. The new objective is to gauge patient ineligibility both nationwide and within each geographic region using a random sample of 150 patients. 50 randomly selected patient files are analyzed from each region. The composition of the population remains the same: 50 percent in the East, 30 percent in the Central U.S., and 20 percent in the West. This is known as a “stratified sample” where patients are randomly selected from each of the three subpopulations, i.e., strata. Table 2 shows the results from this sampling design and review of patient files.

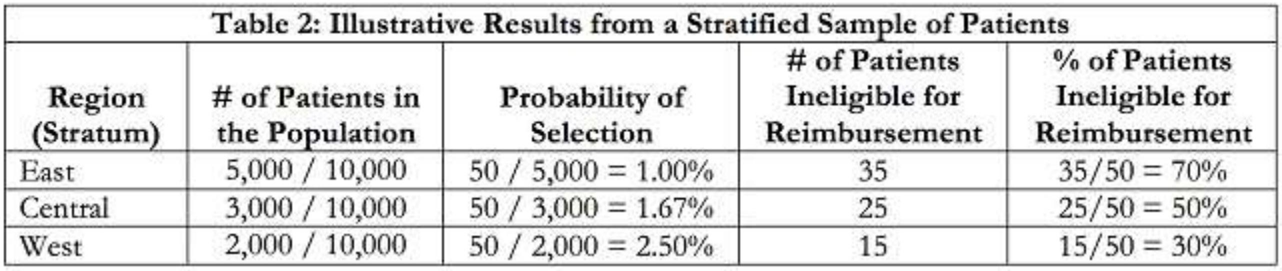

Table 2: Illustrative Results from a Stratified Sample of Patients

Region (Stratum)	# of Patients in the Population	Probability of Selection	# of Patients Ineligible for Reimbursement	% of Patients Ineligible for Reimbursement
East	5,000 / 10,000	50 / 5,000 = 1.00%	35	35/50 = 70%
Central	3,000 / 10,000	50 / 3,000 = 1.67%	25	25/50 = 50%
West	2,000 / 10,000	50 / 2,000 = 2.50%	15	15/50 = 30%

Table 2 reveals that the probability of selection is not the same across regions. For example, patients in the western region are 2.5 times more likely to be selected than patients in the east. Some practitioners might call this sample biased because this is not what we would expect a nationwide sample to look like. This does not prevent the sample from being used to derive valid population estimates and margins of error.

Estimated ineligibility is computed by dividing the number of ineligible sampled patients by the number of sampled patients. These estimates are reported in the far-right column of Table 2. We will refer to these estimates as “regional averages.”

The nationwide ineligibility estimate is computed using two pieces of information: the “regional averages” and the proportion of patients in each region.

[East] (35/50) x (5,000 / 10,000) +

[Central] (25/50) x (3,000 / 10,000) +

[West] (15/50) x (2,000 / 10,000) = 56%

This calculation of 56 percent is a “weighted average.” Each “regional average” is weighted — multiplied — by the proportion of patients in each region. This formula for a population estimate with unequal selection probabilities is found in numerous sampling textbooks.[9]

This calculation is statistically reliable for three reasons. First, a random sample is selected from each region. The estimate within each region is neutral and unbiased. Second, the population sizes within and across each region are fixed. The proportion of patients in each region dictates how much weight to assign to each region. Third, we know every patient’s geographic region using the patient database. In summary, unequal probabilities of selection in no way invalidate statistical inference calculations.

Concluding Remarks

“Representative” in a statistical context is intended to be an objective term of science. The reference guide equates random (probability) sampling with representativeness.[10] However, representativeness has morphed into a subjective term of art. The unscientific expectation is that a sample must be both random and “representative” of the defined population. This adds unnecessary confusion.

The three myths and misinterpretations are frequently used in support of claiming that a random sample is unreliable. Sometimes a random sample does look different than the population, the margin of error is larger than expected, or the probability of selection varies. It is wrong to conclude that the sample cannot be used to reach conclusions about the defined population under any of these circumstances. Since these myths can be deceptive it is imperative for the practitioner to have a solid understanding of these fundamental sampling principles.

Brian Kriegler, Ph.D., is a managing director at Econ One Research Inc. in Los Angeles.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of the firm, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] From Freedman, David A., and David H. Kaye (2011). Reference Guide on Statistics (within the Reference Manual on Scientific Evidence (ed. 3), p. 226 (italics emphasis added). Note that “probability sampling” synonymous with “random sampling” and “sampling frame” is synonymous with “defined population.”

[2] Reference Guide, p. 230 (“Randomness in the technical sense also justifies calculations of standard errors, confidence intervals, and p-values.”)

[3] It is generally appropriate to examine whether a sample is “representative” of a population when the sample is not randomly selected or is prone to bias. The estimated population average and confidence interval calculations are not automatic under these circumstances. The statistical practitioner is tasked with justifying why the sample nonetheless can be used to make statements about the population.

[4] Reference Guide, p. 295 (“[R]epresentative sample. Not a well defined term.”)

[5] This Medicare reimbursement example is similar but not identical to the example discussed in part 3 of this article series.

[6] These include “Kolmogorov-Smirnov tests,” “t-tests,” “chi-square tests,” and “z-tests.” Kolmogorov-Smirnov and t-tests are applied to numerical data. Chi-square tests and z-tests are applied to nonnumerical, categorical data.

[7] This difference is “statistically insignificant” because the probability is relatively high. Differences between the sample and population are “statistically significant” if the appropriate statistical test yields a relatively low probability. This threshold typically is 5 or 10 percent.

[8] The margin of error equals 1.96 x 0.5/√50 = 13.9 percent.

[9] See, e.g., Thompson, Steven K. (2002) Sampling. New York: Wiley. pp. 51-56; Cochran, William G. (1999) Sampling Techniques. New York: Wiley. pp. 258-259.

[10] Reference Guide, p. 226. (“Probability sampling ensures that within the limits of chance…the sample will be *representative* of the sampling frame.) (italics emphasis added)