Practitioner's Guide To Statistical Sampling: Part 2

By **Brian Kriegler**
January 9, 2018, 12:21 PM EST


Brian Kriegler

This is the second of four articles on statistical sampling in practice. In part 1, I demonstrated that random sampling is a neutral and unbiased process. I also showed that the central limit theorem can be used to derive reliable confidence intervals when the sample size is "large enough." In part 2 below, I provide a method for constructing confidence intervals when the sample size is relatively small.

Looking ahead to part 3, I offer various solutions for dealing with missing sample selections so that statistical inferences remain valid. I conclude with part 4, in which I discuss several common misperceptions about random sampling requirements.

Resampling and Bootstrapping

A Method for Determining Confidence Intervals from Small Data Sets

We established in part 1 of this series of articles that random sampling is a neutral and unbiased process. A random sample's data characteristics will approximate those in the population. If the sample is "large enough," then confidence intervals for the population average can be calculated using well-established formulas based on the central limit theorem (CLT). CLT requirements are likely met if the data include at least 30 randomly selected observations.

An alternative method may be desirable if the practitioner cannot confirm that the data set is "large enough" to apply the CLT. Consider a class action lawsuit in which the court allows 25 randomly selected class members to testify. Testimony from these people is used to estimate characteristics about the class, e.g., damages per class member, along with the corresponding confidence interval. This sample may not be large enough to invoke the CLT. In this article we show one way to derive reliable confidence intervals under these circumstances.

For future reference this CLT-alternative method is called "bootstrapping." It builds on "resampling" the data (also referred to as "sampling with replacement"). We will study three questions in this article:

- How are confidence intervals derived using resampling and bootstrapping?
- How do we know that these confidence intervals are statistically reliable?

- How do confidence intervals based on resampling compare to those using the CLT?

Observations can be randomly selected multiple times, once or not at all when sampling with replacement. This is different from the "usual" random sampling without replacement in which each observation is selected once or not at all.

Table 1 lists three pragmatic differences between sampling with and without replacement.

Table 1: Differences between "Sampling with Replacement" and "Sampling without Replacement"

Sampling without Replacement	Sampling with Replacement
Each observation can only be selected once	There is no restriction on the number of times an observation can be selected
The "observed data" is a randomly selected subset from the defined population	The sample with replacement is selected from--and has the same number of observations as--the observed data
One subset of observations is selected from the defined population	*Many* samples with replacement can be selected from the observed data, *i.e.*, *as if it were the population*

Below is a simple demonstration of sampling with replacement. Consider the commuting time for five randomly selected people. These five people's commute times are 15, 20, 25, 30 and 35 minutes. The average of these times is 25 minutes. Assume that the commute times from the first three samples with replacement are as follows:

- First sample with replacement: 15, 15, 20, 35, 35 (average = 30 minutes)
- Second sample with replacement: 15, 15, 25, 30, 30 (average = 23 minutes)
- Third sample with replacement: 20, 30, 30, 35, 35 (average = 24 minutes)

We will explore how sampling with replacement a large number of times provides valuable results.

How is Resampling Used to Derive Confidence Intervals?

Six steps are used to derive confidence intervals from resampled data. Among statistical practitioners, this sequence is commonly referred to as "bootstrapping:"[1]

1. Determine which sampling procedure was used to select the observed data, e.g., simple random sampling.

2. Draw a random sample with replacement from the observed data using the sampling procedure identified in step 1.

3. Calculate the “resampled mean.”

4. Repeat steps 2 and 3 a few thousand times or more. With each repetition, the number of observations should be the same as in the observed data.

5. Identify the percentiles of sample means that align with the desired confidence interval. For example, assume that steps 2 and 3 are repeated 10,000 times with the assistance of a computer program. A common goal is to derive “one-sided” and “two-sided” 95 percent confidence intervals. A one-sided 95 percent confidence interval has a single lower bound at the fifth percentile of resampled means. With 10,000 observations, the fifth percentile is the 500th lowest value. A two-sided 95 percent confidence interval ranges from 2.5 percentile to the 97.5 percentile of resampled means. These are at the 250th lowest to the 250th highest value. These are “percentile confidence intervals.”[2]

6. Graph the histogram of resampled means. This histogram will be symmetric or asymmetric. A symmetric histogram usually indicates that bootstrapping and CLT-based formulas will produce approximately the same confidence intervals. An asymmetric histogram may suggest that bootstrapping produces more reliable confidence intervals.

There are pros and cons to bootstrapping. Bootstrapping only requires two pieces of information: a random sample of data and knowing how these data were selected. There are no requirements about the minimum sample size or population data characteristics. The drawback to bootstrapping is that there is no statistical formula to follow. The percentile confidence intervals typically are authenticated based on a review of the computer program that generated the resampled means.

Why Does Resampling from the Observed Data Produce Reliable Confidence Intervals?

Bootstrapping yields reliable confidence intervals because of three fundamental characteristics of random samples.

- The observed dataset is an unbiased approximation of the population because random sampling is inherently neutral and unbiased.

- Multiple random samples drawn from the same population will yield different results.

- One observed dataset can be resampled a large number of times to approximate the variation across multiple random samples.

Case Study: How Much Time Did Employees Work Off the Clock?

Consider a class action lawsuit consisting of 100,000 employees. The employees allege

that they were not paid for all time worked. Work performed “off the clock” must be estimated because the employer does not keep track of this time. It is prohibitively expensive to ask each employee how much time they worked off the clock. Instead a random sample of class members is deposed and asked how much time they worked off the clock. These employees’ testimony is used to estimate the average — and ultimately total amount — of unrecorded work time across all class members. In addition, there is a need to derive confidence intervals for the population average and total.[3]

Random sampling and bootstrapping are necessitated by the fact that the population average is not known. Setting aside that this fact is unknown to the practitioner, assume that the population consists of 90,000 class members who worked 10 minutes off the clock per workweek. The remaining 10,000 class members worked 60 minutes off the clock per workweek. This works out to an average of 15 minutes per workweek for each class member.

To learn about the population the practitioner selects a random sample of class members who report in deposition their off-the-clock time. This deposition testimony is analyzed after 20, 50 and 100 class members. Table 2 shows the results after each stage of sampling.

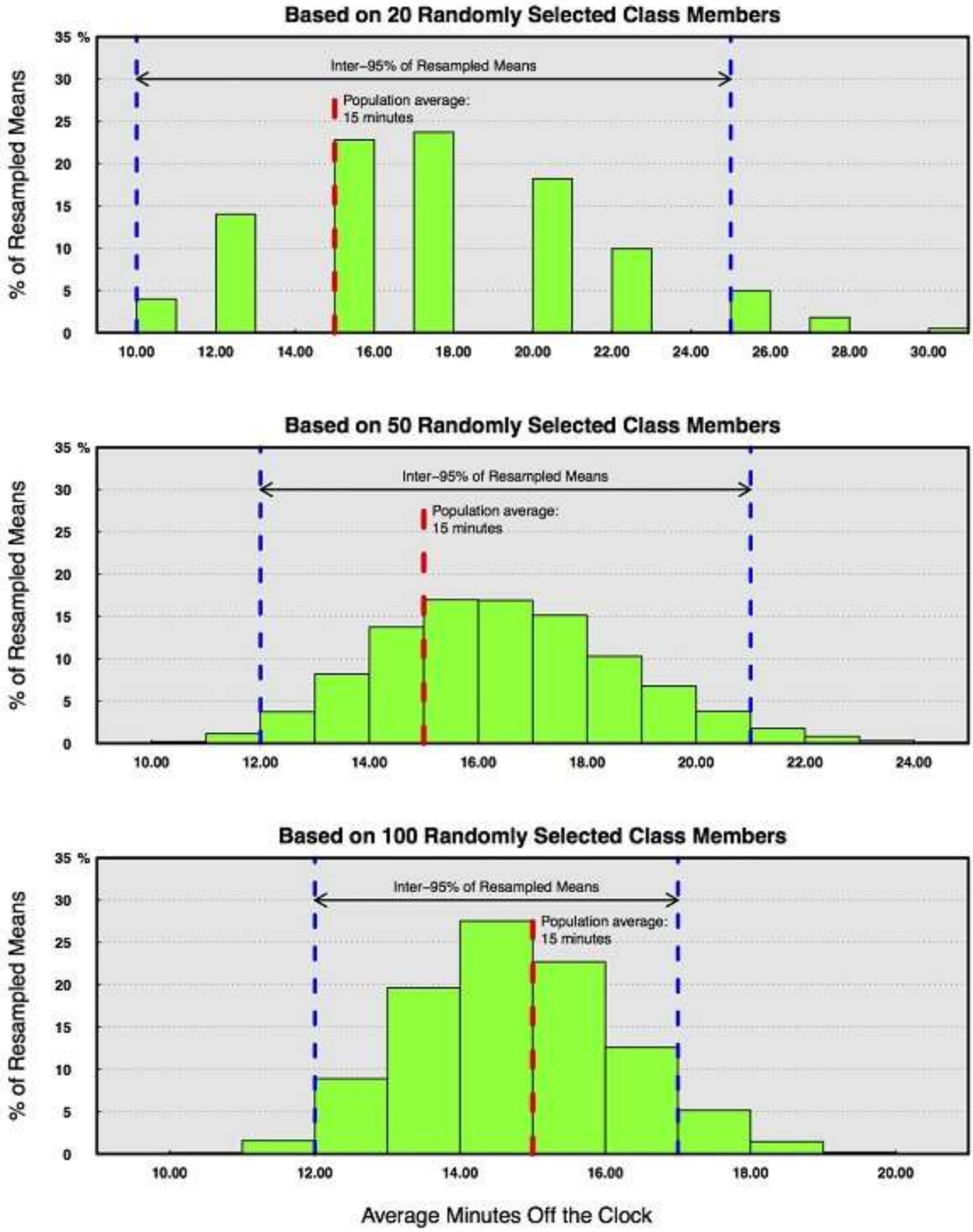
Table 2: Illustrative Results Based on a Random Sample of Class Members’ Deposition Testimony

Sample Size	# People with 10 Minutes Off the Clock	# People with 60 Minutes Off the Clock	Sample Average
20	17	3	(17 x 10 + 3 x 60) / 20 = 17.5
50	44	6	(44 x 10 + 6 x 60) / 50 = 16.0
100	91	9	(91 x 10 + 9 x 60) / 100 = 14.5
Population Average (in practice, not observed)	90,000	10,000	(90,000 x 10 + 10,000 x 60) / 100,000 = 15.0

Deriving Percentile Confidence Intervals

The practitioner resamples the observed data 10,000 times using the aforementioned bootstrapping steps. 10,000 resampled means are generated in the process. Figure 1 below shows three histograms of these resampled means.[4] The three graphs are based on samples of 20, 50 and 100 randomly selected class members. 95 percent of resampled means are between the two blue dotted lines. This is also the definition of a 95 percent confidence interval. The dotted red line shows the true population average of 15 minutes. The scales on each graph are different because the collection of resampled means gets narrower as the sample size increases.

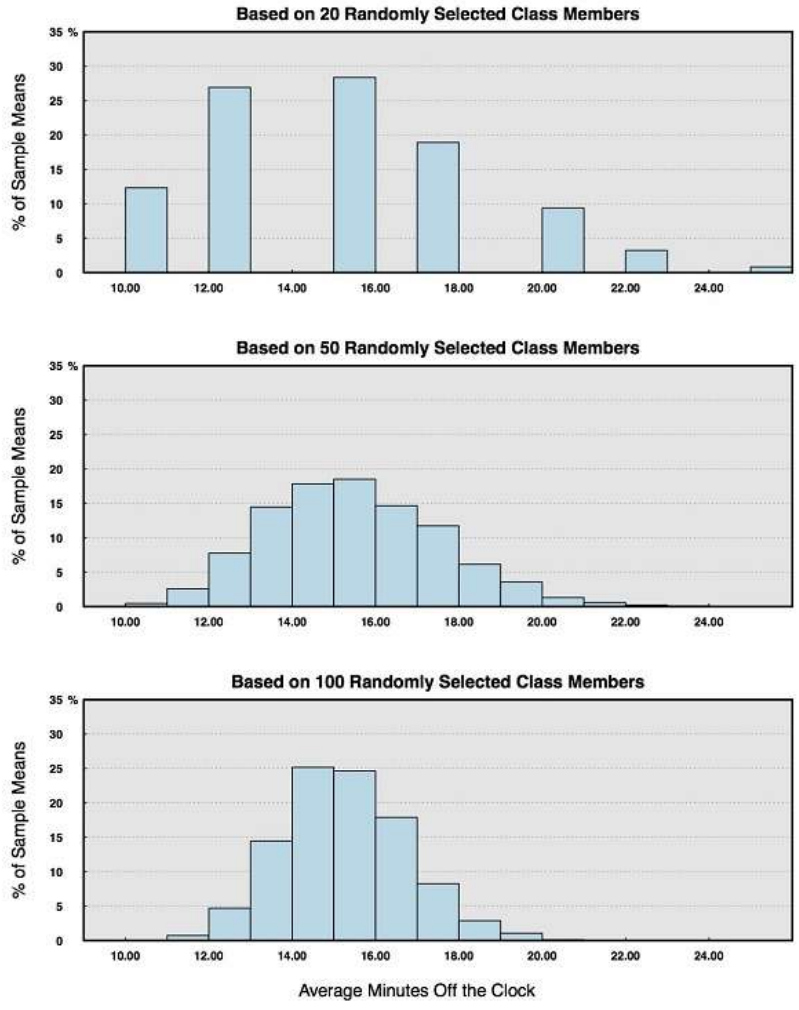
Figure 1: Histograms of 10,000 Resampled Averages And Percentile Confidence Intervals for the Classwide Average



Verifying that Percentile Confidence Intervals are Statistically Reliable

How do we know that the histograms of resampled means selected from the observed data are meaningful? We can find out by comparing Figure 1 to Figure 2 below. Figure 2 contains histograms of sample means selected from the population. Here we are assuming that we know the amount of time that every class member worked off the clock strictly for illustrative and comparative purposes. In practice random sampling is typically unnecessary if this information is known.

Figure 2: Histograms of 10,000 Sample Averages
Assuming that Data Were Available for All Class Members



The histograms in Figure 1 strongly resemble those in Figure 2. Both histograms based on 20 observations are asymmetric and disjointed. Both histograms based on 50 observations are slightly asymmetric. Both histograms based on 100 observations are approximately symmetric and bell-curved. The fact that these two figures look so similar confirms that bootstrapping can be used to approximate the true distribution of sample means. It follows

that bootstrapping also can be used to derive confidence intervals for the population mean.[5]

How Do Percentile Confidence Intervals Compare to CLT-based Confidence Intervals?

Using our practitioner's sample data, Table 3 compares percentile confidence intervals with CLT-based confidence intervals[6] for several commonly relied-upon confidence levels: one-sided 95 percent, two-sided 95 percent and two-sided 99 percent.

Table 3: Comparison of Percentile Confidence Intervals to CLT-based Confidence Intervals Using Case Study Data

Sample Size & Sample Mean	Confidence Interval Method	One-Sided 95% Interval	Two-Sided 95% Interval		Two-Sided 99% Interval	
		Lower Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
		(Number of minutes working off the clock)				
20 Depositions (Mean = 17.5)	Percentile	10.00	10.00	22.50	10.00	25.00
	CLT	9.05	7.80	22.20	5.16	24.85
	Difference	*0.95*	*2.20*	*0.30*	*4.84*	*0.15*
50 Depositions (Mean = 16.0)	Percentile	12.00	11.00	19.00	10.00	21.00
	CLT	11.41	10.69	19.31	9.26	20.74
	Difference	*0.59*	*0.31*	*0.31*	*0.74*	*0.26*
100 Depositions (Mean = 14.5)	Percentile	12.50	12.50	18.00	11.50	19.00
	CLT	12.50	12.01	17.99	11.04	18.96
	Difference	0.00	0.49	0.01	0.46	0.04

Note: The practitioner does not observe or know that the population mean is 15.0 minutes.

Differences between percentile and CLT-based confidence intervals are relatively large when based on 20 randomly selected class members. The 99 percent confidence interval lower bounds differ by nearly five minutes, and the 95 percent confidence interval lower bounds differ by more than two minutes.[7] These are relatively large differences considering the population average is 15 minutes. Such results suggest that a sample of 20 observations is not enough to apply the CLT. Bootstrapping is especially valuable under these circumstances.

Percentile and CLT-based confidence intervals tend to converge as more depositions are taken and analyzed. Random samples of 100 produce very similar results across confidence interval methodologies. This demonstrates that the practitioner can choose which confidence interval method to apply as the sample size grows. It also supports the notion that the CLT can be used to calculate confidence intervals if the random sample is "sufficiently large."

Concluding Remarks

Statistical practitioners have multiple tools for deriving confidence intervals. In part 1, we showed that the CLT has sample size requirements for calculating confidence intervals. Bootstrapping is an alternative method that can be applied to any random sample no matter how much data have been collected.

Resampling and bootstrapping are applicable if the observed data were selected using random sampling. It is tempting to conclude that a small number of observations — such as 20 to 50 depositions — cannot be used to draw inferences about thousands of people. A small sample may well look different than the defined population. This is why it is so important to calculate, interpret and rely on confidence intervals when drawing conclusions about the population. Bootstrapping provides an efficient and reliable tool for determining confidence intervals in a broad set of circumstances.

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

[1] See, e.g., Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York: Chapman & Hall. pp. 168-174. Note that the term “bootstrapping” implies getting into (or out of) a situation using existing resources. In a statistical context, the “existing resources” consist of one randomly selected dataset.

[2] Percentile confidence intervals are discussed in Efron and Tibshirani (1993), pp. 168-174. See also, e.g., Berk, R.A. (2004) Regression Analysis: A Constructive Critique. Thousand Oaks: Sage Publications. pp. 74-76.

[3] The estimated total amount of time that class members worked off the clock equals the number of class members multiplied by the estimated average.

[4] These computations were performed using the statistical software program “R.” The computational code used to generate the graphs and results included in this article is available upon request.

[5] Figure 2 does not include confidence intervals for the population mean. This is intentional. A confidence interval is only needed if the population mean is unknown. In this theoretical demonstration, data are available for every person in the population.

[6] The CLT-based formulas for confidence intervals applied in this article are approximately the following:

- One-sided 95 percent confidence interval with a lower bound: Sample Mean - 1.7 x Standard Deviation / √Sample Size.
- Two-sided 95 percent confidence interval: Sample Mean ± 2.0 x Sample Standard Deviation / √Sample Size.
- Two-sided 99 percent confidence interval: Sample Mean ± 2.6 x Sample Standard Deviation / √Sample Size.

[7] That is, 10.00 - 5.16 = 4.84 minutes, and 10.00 - 7.80 = 2.20 minutes.