Chapter 2
Sampling for Biostatistics

Angela Conley and Jason Pfefferkorn

Abstract Define the terms sample, population, and statistic. Introduce the concept of bias in sampling methods. Demonstrate how the sample size impacts the standard error. Illustrate several methods that exist on how to sample a population. Become familiar with the inherent advantages and disadvantages of the sampling methods.

Keywords Population • Sample • Statistic • Bias • Standard error • Random • Stratified • Survey • Systematic • Cluster • Tree

1 Introduction: Population and Samples

A primary goal in biostatistics is to derive conclusions about an entire group or population. These conclusions enhance our understanding of how the biological world responds and evolves based on behaviors and characteristics observed in the population. In practice, achieving this goal typically proves difficult, as one must mount a compelling argument from evidence collected from a subset that is typically a fraction of the population. A significant reason for collecting a subset of the population is that in many cases the population is too numerous and therefore, infeasible to collect data from every element in the entire set. In some instances, the population may be unknown. We call this subset a **sample** of the population. By carefully constructing the way in which this sample is collected, one may infer conclusions with little to no bias about the population based solely on the observations or measurements from the sample. Great statisticians are not only concerned with what the data reveals but also pay due diligence to how the data was obtained.

A. Conley, B.A., M.A. (✉)
Department of Mathematics, SEM Division, Cerritos College, 11110 Alondra Boulevard, Norwalk, CA, 90650, USA
e-mail: aconley@cerritos.edu

J. Pfefferkorn, B.A., M.A.
Space and Airborne Systems, Raytheon Company, 2000 East El Segundo Blvd, El Segundo, CA, 90245, USA
e-mail: Jason.pfefferkorn@gmail.com

M. Mićić (ed.), *Sample Preparation Techniques for Soil, Plant, and Animal Samples*, Springer Protocols Handbooks, DOI 10.1007/978-1-4939-3185-9_2

Ex. 1.1: A population to consider is the set of trunk diameters of a given species of tree throughout the entire world.

In this example, we could easily convince ourselves that we would need an extraordinary amount of resources (time, money, staff) to collect this colossal set of data—certainly an infeasible solution. However, if we were to select from this population a much smaller subset could we easily justify and extend our conclusions based on this sample to the rest of the population? Statisticians would cautiously say yes.

Note that we do not mean population as most people commonly understand the use of the term. Here **population** refers to a collection of features or measurements from an entire set of objects.

Observations or numerical measurements of the sample are used to compute a statistic. A **statistic** is a numerical measure derived from a sample. We use a statistic to estimate the parameter of the population in question. For instance, we may be interested in knowing the average life expectancy of a certain species of an insect. In this case the population parameter is the average, or mean, life expectancy. In order to estimate this parameter one collects a sample of these particular insects, measuring and recording each life expectancy of the insects in the sample and then deriving the average from this collection of measurements. This sample average is a fair approximation of the true average life expectancy.

However, samples must be selected in a way to ensure that the estimate of the population parameter is unbiased. Bias in estimates may lead to false or unjustifiable conclusions about the population in general. Various methods exist to collect samples and are discussed in the following section. Statisticians must weigh the benefits and drawbacks of each method.

2 Sampling Methods

In order to study sound sampling techniques that result in reasonable conclusions about a population, we turn our attention to some sampling techniques that have inherent flaws, which may ultimately misconstrue the results. Becoming aware of these methods will provide the reader with a background to identify unsustainable arguments in case studies that may make use of such techniques. It will also enforce the reader why sampling techniques are so crucial to developing conclusions about the population.

3 Biased Sampling Techniques

Ex. 1.2: A televised singing competition asks viewers to phone in to vote for their favorite contestant.

Here is a classic example of a **voluntary response survey**, where the decision to be included in the sample is made solely by the members in the sample. Generally,

this type of sample has the potential to be biased since those who take part usually do so because they feel strongly about the issue in order to influence the outcome of the survey. Typically, this sampling technique does not represent the entire population's preference.

Another sampling technique that may contain bias is estimating a characteristic of the population by posing loaded questions. In other words, questions regarding sensitive personal issues that may result in the subject providing false information should be avoided or carefully constructed to ensure anonymity. In addition, there may be the potential is a possibility for a respondent not providing an answer altogether.

Ex. 1.3: A question on an anonymous survey asks the following: Have you ever been convicted of a sex crime?

In this example, we have assured the subject their response is anonymous. However, the respondent may still choose to avoid answering the question altogether. A statistician must be aware of the non-response bias that could then alter the conclusions that the remaining data reveals. A controlled environment has a similar consequence where the subject may feel constrained in their response.

Ex. 1.4: An arborist samples the diameters of tree trunks of a certain species within the USA but for the sake of convenience only collects the data from the trees in close proximity to his place of work.

This example illustrates a form of sampling that has a high potential of creating biased results. Here the arborist collects data simply and efficiently because it is particularly convenient. It's not difficult to convince yourself how this approach to sampling may unfairly portray the true population parameter, in this case, the average diameter of tree trunks. For instance, there may exist regional differences that result in the average diameter of tree trunks being larger than other regions. Potential reasons for these differences among regions could include rainfall totals, soil content, and tree age to name just a few.

4 Random Sampling

Sampling techniques that include some aspect of randomness provide the statistician a greater confidence in our estimate of the true population parameter. However, the best sampling technique is a random sample.

A random sample is one in which every member of the population has the same chance of being included in the sample.

Ex. 1.5: Consider a population that consists entirely of a certain species of trees that reside in seven parks in the nearby region. We wish to estimate the average height of these trees. For the sake of argument, we shall consider collecting the data from the population to be a significant challenge due to a lack of tools, staff and money. In order to circumvent this problem, we estimate the average height from a sample. Now, the question becomes: which trees do we select and from which parks?

The best way to obtain a random sample would be to identify and locate each tree with a number. Use a uniform random number generator to select from the numbered trees a small subset of the population. This random number generator ensures that each tree has the same chance of being included in the sample. A number of software suites exist to perform the calculations described in this chapter. One such tool is Microsoft's Excel spreadsheet program. Familiarize yourself with the following functions as they will undoubtedly prove useful (Table 2.1).

In Fig. 2.1 we have identified the seven parks with dashed lines that contain the particular species of trees we are interested in. Each tree is identified with a number from 1 to 84. In this case our population is 84. Using a uniform random number generator, we have decided to sample the heights of 21 trees. The selection we have made is completely random ensuring that each tree could have been selected with equal probability. It is fairly interesting to note, that in this instance, the

Table 2.1 List of useful Excel functions

Function	Description
RANDBETWEEN(lower bound, upper bound)	Generates a uniform random number between, and inclusive of, the specified lower and upper bounds.
AVERAGE(number1, number2, …, number*N*)	Returns the average from the specified set of numbers

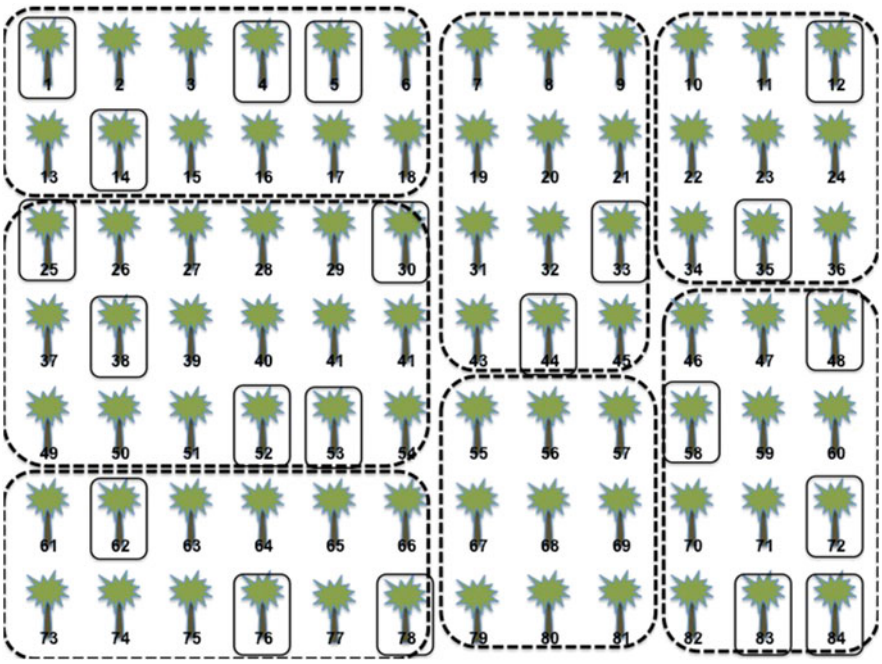

Fig. 2.1 21 of 84 trees selected from 7 parks

sample selected at random does not include a tree from every park. Since every tree has an equal probability of being selected at random, an outcome like this is just as likely to occur as a sample where at least one tree from every park was included in the sample.

5 Sample Size and the Standard Error

It now behooves us to introduce the concept of a sample size and what is an appropriate size to use. Until now we've mentioned that a sample is a subset. However, by definition this subset could include just one observation or the entire population. (A subset could be the empty set, but that wouldn't be a very interesting sample!) So, the question is what is an appropriate sample size? The easy answer to this question is it depends on the population under study. If the population is small, then the population parameter may be easily obtained with no sample necessary. If the population were so large that its size is unknown, then a sample would be required. A general rule of thumb that most within the statistics community agree with is that a minimum sample size of 30 may be sufficient. This is not to say that one should merely stop at 30 elements in the sample. As you might expect, the estimate on the population parameter becomes more accurate as the sample size increases for a random sample. So one should size the sample according to available resources.

In cases where an estimate of the proportion of the population having a particular characteristic is desired, there exists a formula to compute the standard error of the sample. If p represents the proportion of the population that possesses some characteristic we are interested in measuring, then the standard error for a sample of size n is given by Eq. (2.1).

$$\mathrm{SE} = \sqrt{\frac{p \times (1-p)}{n}} \tag{2.1}$$

However, Eq. (2.1) only holds true under certain conditions. These conditions are that the sample must be random and that if we are sampling without replacement then the population generally must be ten times the sample size.

When we say sample with replacement, what we mean is that an element from the population is randomly selected, the data is collected from that element, and then subsequently returned to the population. By returning this element to the population, we are ensuring that every trial is independent and the probability of being selected remains the same. However, you will note that by returning the element you may then be able to select that same element again. In this case, the sample may contain redundant information. This is likely to be true when the population is rather small compared to the sample size.

In contrast to sampling *with* replacement, there is sampling *without* replacement. In this case, once an element is selected from the population it is removed from the

population to prevent it from being sampled again. This process doesn't necessarily ensure that the probability of an element being selected at random remains the same as the sample is collected. However, if the population is large enough, the change in probabilities are negligible that for all intents and purposes it is the same as sampling with replacement. Consider the following examples that help to illustrate these points.

Ex. 1.6: **Large population with replacement**—Our population is the trunk diameters of a certain species of trees in the USA. Here the population is so vast and innumerable that we require a sample to estimate the average diameter of the tree trunks. Ideally we would like to draw trees at random across the USA to minimize the amount of bias. Suppose we had the resources to collect data from a random sample of 500 trees. Sequentially, we select one of these trees from the sample of 500, record its trunk diameter, and then subsequently return it to the population for the possibility of being sampled again. However, if we have some inclination that the population numbers in the millions, then the probability of the same tree being selected again remains so infinitesimally small that we may avoid the possibility of recording that same tree's trunk diameter a second or third time. Probability theory tells us the possibility exists but the chances of that happening are so rare that it is hardly a concern to us.

Ex. 1.7: **Small population with replacement**—Our population is the trunk diameters of a certain species of trees in a nearby park. Suppose that this population is known to contain 60 trees and we do not have the resources to collect the population parameter. Let us further suppose that we only had the time to collect a random sample of 15 trees. In contrast with **Ex. 1.6**, we are illustrating that since the population is rather small, there is a much greater chance of acquiring redundant data. This redundancy could bias the results. Therefore, we suggest that when working with a relatively small population, consider sampling without replacement to prevent a bias in the results.

Notice that in Eq. (2.1) as the sample size increases, the precision of the estimate of the population parameter improves.

Ex. 1.8: If the proportion of trees infected in a given park is known to be 11 % and the number of trees in our population is 1200. Then using Eq. (2.1), the standard error in our estimate from a random sample of 50 trees is

$$\text{SE} = \sqrt{\frac{0.11 \times (1 - 0.11)}{50}} = 0.0442 \text{ or } 4.4\%$$

However, if we were to sample 200 trees notice that the standard error decreases and therefore, the precision of our estimate of the population parameter improves.

$$\text{SE} = \sqrt{\frac{0.11 \times (1 - 0.11)}{200}} = 0.0221 \text{ or } 2.2\%$$

We leave it as an exercise to the reader to confirm the standard error of a sample size of 400 as shown in Table 2.2.

Table 2.2 Standard error as it relates to sample size

Sample size, n	Proportion of population w/disease, p	Standard error
50	0.11 (11 %)	0.0442 (4.42 %)
200	0.11 (11 %)	0.0221 (2.21 %)
400	0.11 (11 %)	0.0156 (1.56 %)

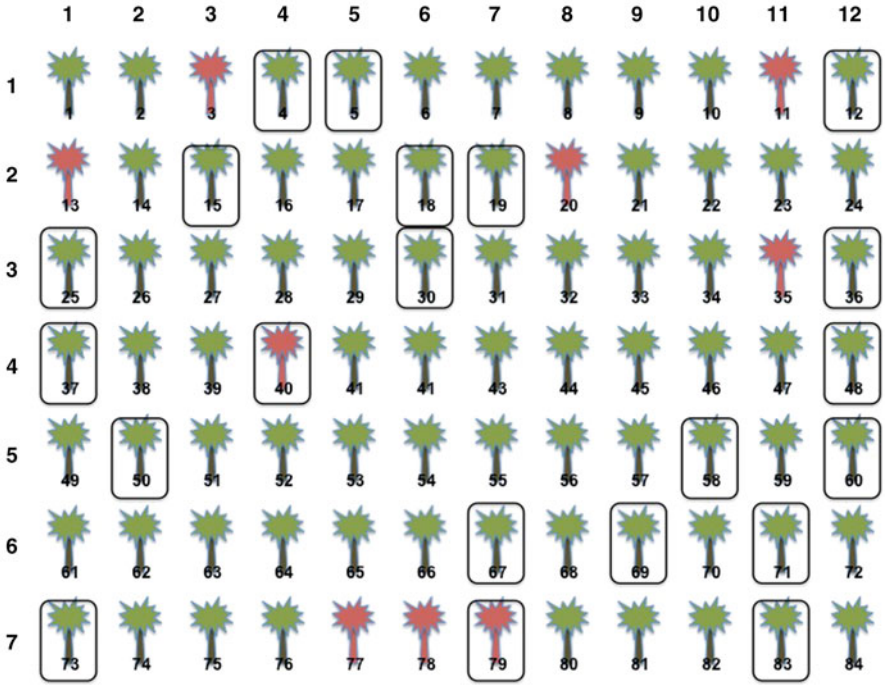

Fig. 2.2 Trees infected (*red*) and trees not infected (*green*)

6 Alternative Sampling Techniques

Consider the population of trees in Fig. 2.2. Green trees represent healthy trees and red trees represent trees infected with a disease.

In this case, we can easily compute the actual population proportion, which is 9/84 or 10.7 %. To illustrate some alternative sampling techniques, let's suppose that we do not know the actual population proportion.

Stratified sampling begins by breaking the population up into distinct groups, called strata. Data is then collected from randomly selected elements of each stratum. A variety of factors can determine the strata. The elements of each stratum are usually grouped by a common attribute.

Ex. 1.9: To demonstrate, we chose to simply stratify the population by rows. Within each stratum, we randomly select three elements to obtain a sample proportion for trees infected, illustrated in Fig. 2.3.

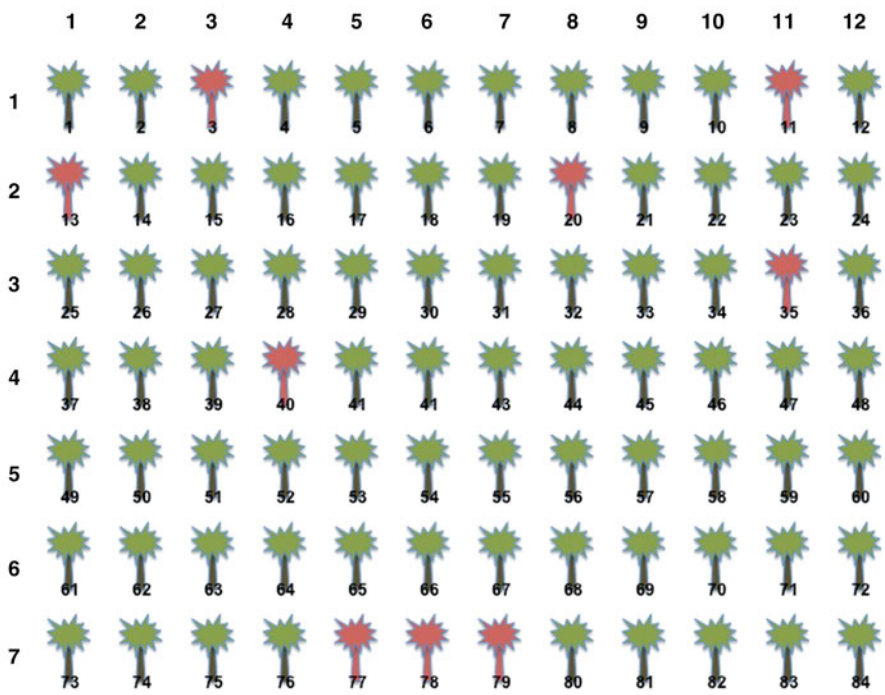

Fig. 2.3 Stratified sampling of tree population

Here we obtained a sample proportion of 2/21 or 9.5 % using stratified sampling. Again notice that *every* stratum (row) was represented in calculating the sample proportion. We leave it as an exercise for the reader to stratify the population by columns and compute the sample proportion by randomly selecting two trees in each column.

Cluster sampling begins in the same manner as stratified sampling—the population is divided into distinct clusters. However, from that point some clusters are randomly selected and all of the elements in those clusters are included in the sample.

Ex. 1.10: Suppose we cluster the population by rows.

In this example, we have randomly selected rows 2, 3, 5 (see Fig. 2.4). In this sample of 36 trees, three had the disease. Therefore the sample proportion is 3/36 or 8.3 %. Note that only the rows that were selected are captured in the sample proportion.

Systematic sampling begins with a randomly selected starting point, and then includes each *k*th element in the sample.

Ex. 1.11: Suppose we randomly select our starting point as the first tree and then systematically select every fifth element from this set as demonstrated in Fig. 2.5. Ultimately we end up with a sample of 16 elements in this particular example.

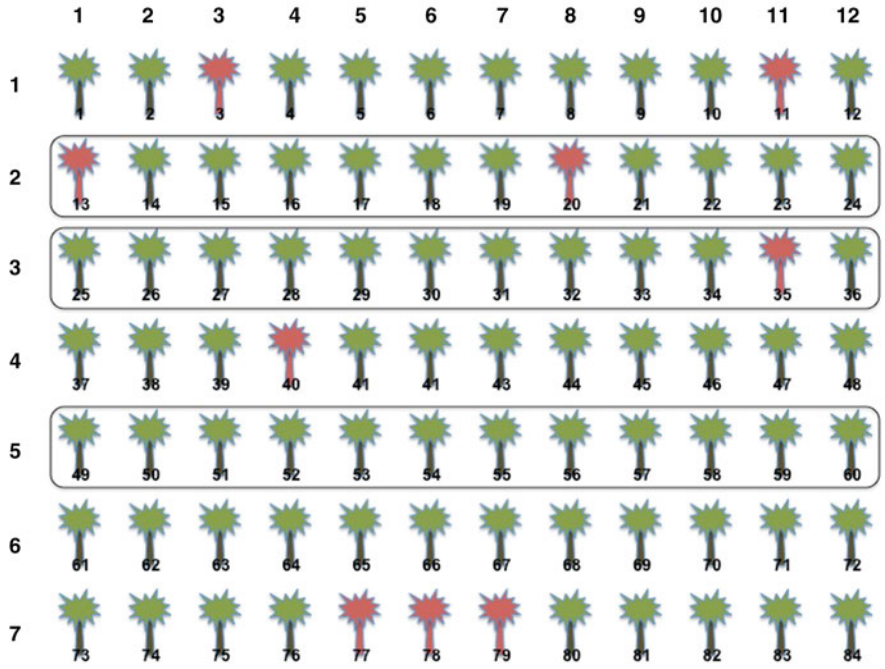

Fig. 2.4 Cluster sampling of tree population

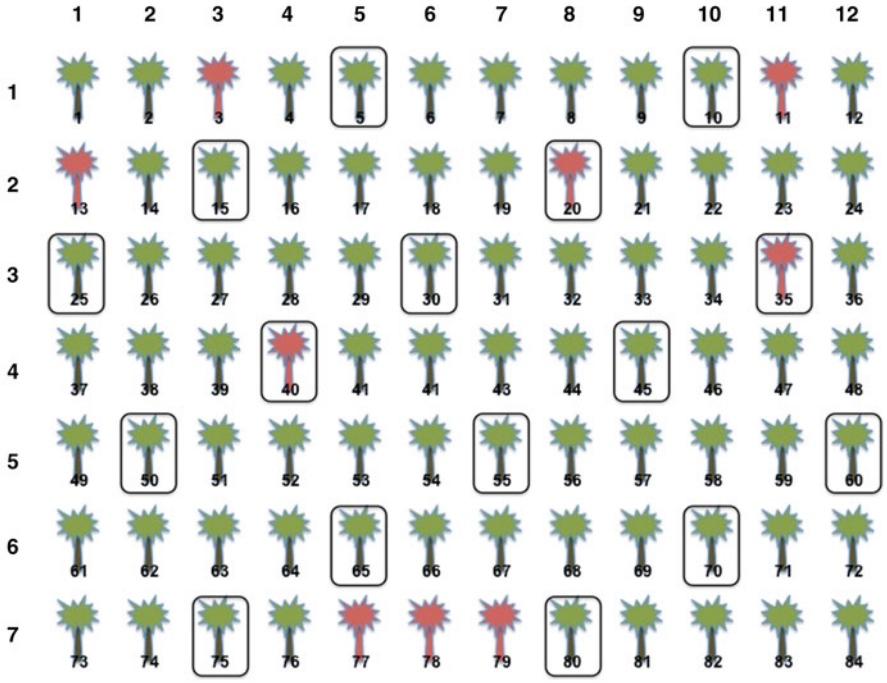

Fig. 2.5 Systematic sampling of tree population

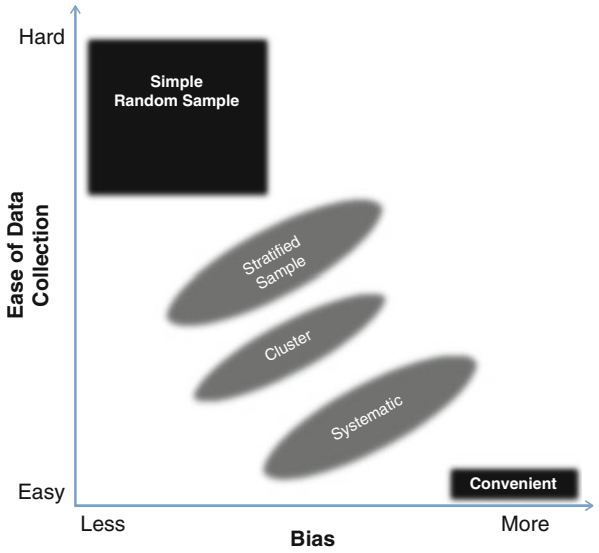

Fig. 2.6 Contrasting sampling techniques

Of the 16 trees systematically selected, the sample proportion is 2/16 or 12.5 %.

The previous three examples not only demonstrate various sampling techniques but also reinforce the fact that statistics vary while the parameter remains the same.

Each sampling technique has its advantages. Each technique has an element of randomness while offering an organized way to collect the data. However, if the sample selected is not significant in size the estimate in the population parameter may not be best. In Fig. 2.6, we have summarized the variety of sampling methods we discussed, contrasting their potential for bias with their ease of collection.

Suggestions for Further Reading

1. Gould R, Ryan C (2011) Introductory statistics: exploring the world through data. Pearson Education, New York, NY
2. Groves R (2004) Survey errors and survey costs. Wiley-Interscience, New York, NY
3. Sowder J, Sowder L, Nickerson S (2012) Reconceptualizing mathematics. W.H. Freeman and Company, New York, NY

http://www.springer.com/978-1-4939-3184-2

Sample Preparation Techniques for Soil, Plant, and Animal Samples

Micic, M. (Ed.)
2016, XXIII, 406 p. 96 illus., 65 illus. in color., Hardcover
ISBN: 978-1-4939-3184-2
A product of Humana Press