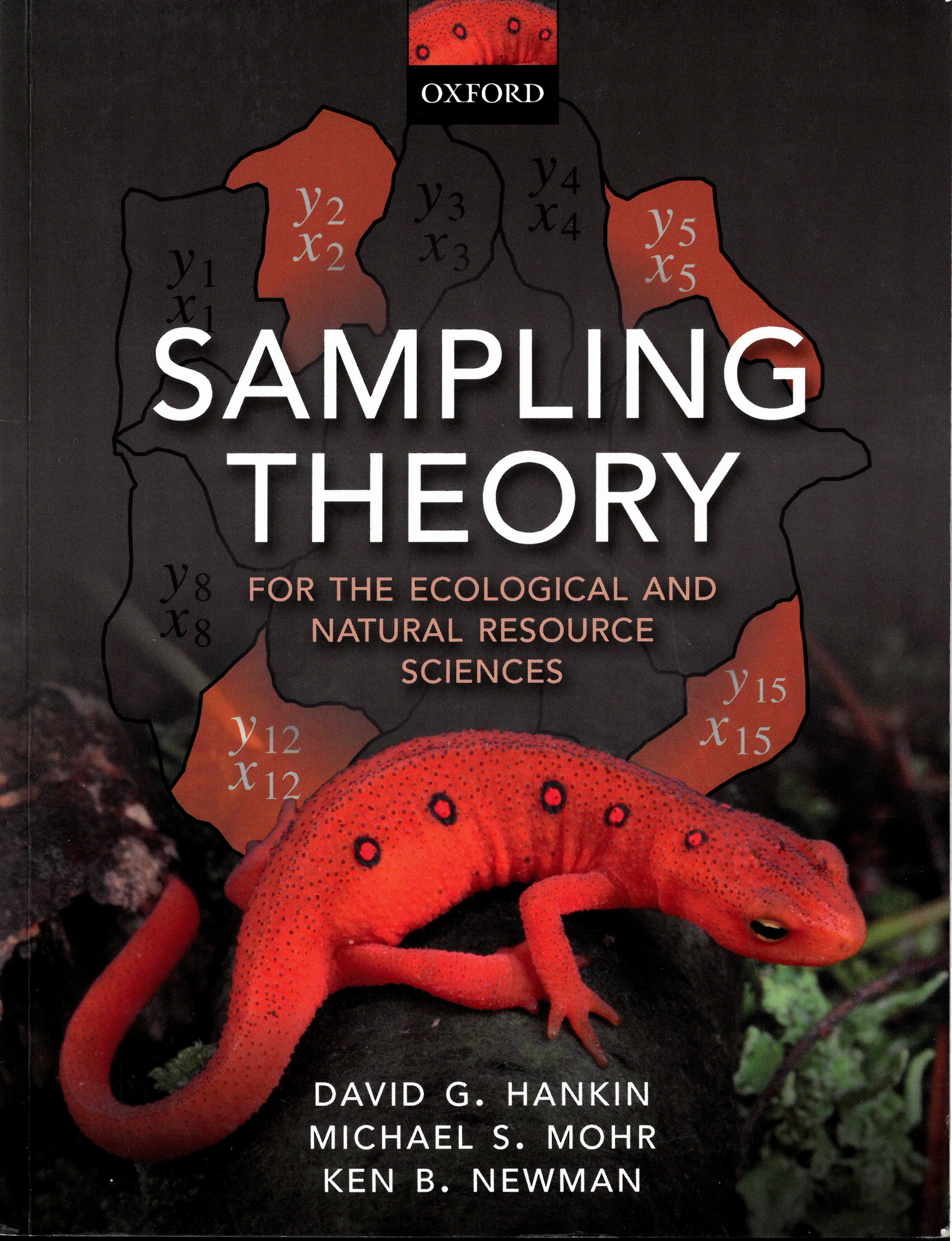
OXFORD

SAMPLING THEORY

FOR THE ECOLOGICAL AND NATURAL RESOURCE SCIENCES



DAVID G. HANKIN
MICHAEL S. MOHR
KEN B. NEWMAN

CHAPTER 2

Basic concepts

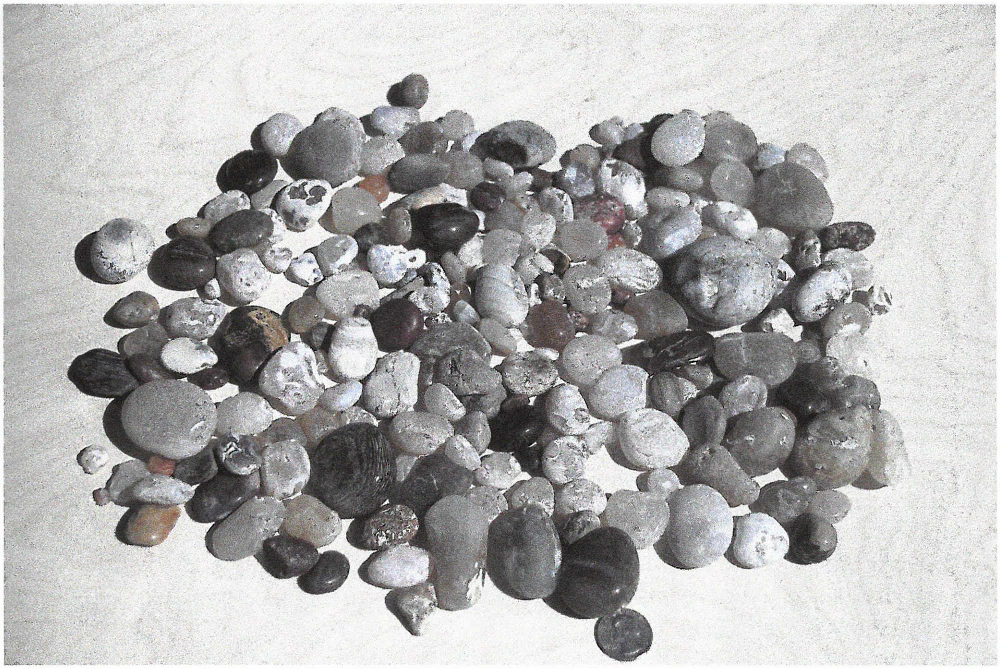


Fig. 2.1 An agate (chalcedony) population used for the judgment sampling exercise described in this chapter. Population also includes agatized petrified wood and jasper collected on the northern California shore. Photo credit: D. Hankin.

In this chapter, we briefly consider the most fundamental terms, concepts, and objectives of sampling theory in the context of a finite population.[1] Presentation is largely at a qualitative level in this chapter, with more sophisticated and better motivated quantitative treatment reserved for subsequent chapters. The intention is to quickly instill a rudimentary but sound conceptual grasp of the basic concepts of sampling theory, including the basic properties of estimators. In subsequent chapters, we will quantify these basic concepts and properties using formal arguments and examples. Important sampling theory terms appear in boldface type at first usage.

[1] The term *universe* is sometimes used in place of the term *population* in the sampling theory literature.

Sampling Theory: For the Ecological and Natural Resource Sciences. David G. Hankin, Michael S. Mohr, and Ken B. Newman, Oxford University Press (2019). © David G. Hankin, Michael S. Mohr, and Ken B. Newman. DOI: 10.1093/oso/9780198815792.001.0001

2.1 Terminology

Throughout this text, we will concern ourselves with estimation of a descriptive **population parameter** for a **variable** associated with a **population** of interest containing a finite number of **population units** based on examination of a **sample** (subset) of these units. Illustrative examples of population parameters (mean, proportion, total, respectively) of populations are

- the mean (average) age of male adults living in senior housing facilities throughout the U.S.;
- the proportion of 3rd grade students in North Carolina public schools that are obese;
- the total number of coastal redwood trees exceeding three feet diameter at breast height present on all National Forest lands in California.

We will begin our study of sampling theory with situations in which we can select samples directly from a listing of the units in a population. For example, for the first population, we might select a sample of male adults from a listing of all of the male adults residing in senior housing facilities in the US. For the second population, we might select a sample of students from a listing of all third grade students in North Carolina public schools. These are examples of the simplest kinds of samples that can be selected: samples selected directly from a complete listing of the population units.

A given population may have many different variables and related population parameters associated with it. For example, the population of male adults living in senior housing facilities will also have an average level of education, an average level of savings, an average IQ score, and so on. Typically, we wish to estimate a population parameter for a specific variable that holds primary interest and we refer to this variable as the **target variable.**

An example finite population consisting of $N = 16$ units is displayed in Figure 2.2(a). Each unit $i = 1, 2, \ldots, N$ (in this case, adjoining watersheds) may have multiple attributes or variables associated with it. We use y_i to denote the unit i value for the target variable, y, whereas we use x_i to denote the unit i value for an **auxiliary** variable, x, that may help us to estimate a population parameter of y. Figure 2.2(b) highlights a particular sample of size $n = 4$ units $(i = 3, 6, 10, 16)$.

It is not always feasible to select a sample directly from a complete listing of population units. For example, it might be very difficult or impossible to get a listing of the total

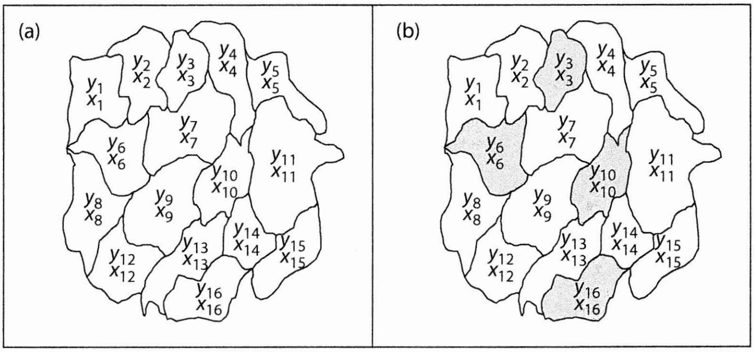


Fig. 2.2 Map (a) depicts an example population of $N = 16$ areal units (watersheds). Map (b) highlights a particular sample of $n = 4$ units. The $(x_i, y_i), i = 1, 2, \ldots, N$, denote the target and auxiliary variable values associated with each population unit.

number of male adults in all senior housing facilities throughout the U.S. Instead, it might be a relatively simple matter to get a listing of all of the senior housing facilities in the U.S. and to select a sample of n senior housing facilities from the N senior housing facilities. Each of the selected facilities would have a certain number of male adult occupants. We could determine the ages of all (or some) of those individuals within the selected facilities, and we could then use that information to estimate the average age of male adults in all senior home facilities. Because it may often not be practical to sample directly from a complete listing of the population units, sampling theorists consider formation of a **sampling frame** to be one of the most critical and challenging tasks. We use the term **simple frame** to describe a setting where the sampling frame consists of a listing of the **population units**, whereas we use the term **complex frame** to describe a setting where the sampling frame consists of a listing of **sampling units** which are well-defined groupings or clusters of the population units. Thus, if we selected a sample of senior home facilities from a listing of all such facilities, we would consider this to constitute selection of sampling units from a complex frame relative to the population units (male adults in all nursing homes).

In many natural resource contexts, the target of estimation may be expressed at a geographic level (as for Figure 2.2). For example, a private timber company may be interested in estimation of total abundance of an endangered species of owl or amphibian present on all N of its forest properties. In this case, sampling frame units consist not of individual owls or amphibians but of specific forest properties which are generally of highly variable size and shape. Here we might associate y_i, unit-specific owl or amphibian abundance, with property i. Similarly, in a stream survey context, where estimation of total fish abundance in a specified reach of stream might be the survey objective, there are many possible alternative sampling frames. One possibility would be to break the entire length of the reach into N 100 m sections, with a unit-specific y_i associated with each section. Another (and often better) sampling frame might consist of a listing of all natural habitat units, of varying size, in the stream reach (Hankin 1984).

2.2 Components of a sampling strategy

We use the term **sampling strategy**[2] to define the complete process whereby (1) a sample of population units is selected from a population (or sampling units from a complex frame), (2) relevant measurements are taken from sampled units, and (3) an estimate of a population parameter is made based on observed measurements. Assuming that the population and target variable of interest have both been carefully defined, development of a sampling strategy requires specification of four interrelated components.

Sampling frame. A listing of the units from which the (initial) sample will be selected. A *simple frame* consists of the basic population units, whereas a *complex frame* consists of sampling units which are well-defined groupings or clusters of the basic population units.

Sample selection methods(s). The method(s) used to select a subset or sample of the units listed in the sampling frame (e.g., selecting a sample of senior housing facilities, or selecting a sample of forest properties), and possibly specification of further methods for selecting subunits in the case of a complex frame (e.g., selecting

[2] This terminology follows the usage of, at least, Cassel et al. (1977), Brewer and Hanif (1983), Hedayat and Sinha (1991), Chaudhuri and Stenger (2005), and Brewer and Gregoire (2009).

samples of adult males from within selected housing facilities, or selecting samples of 1 m^2 plot units within selected forest properties).

Observation methods. The explicit protocol(s) that will be used to record or measure y (and sometimes x) of the selected units in a sample. This may be a straightforward measurement of y for the individual units (e.g., a direct response to a survey question), or it may involve use of some model-based approach for estimation of y for the individual units (e.g., use of mark-recapture estimation (Amstrup et al. 2005) to estimate forest property-specific abundance of an owl or amphibian species).

Estimation equations. The explicit formulas that will be used to estimate the population parameter(s) of interest and to calculate an associated error of estimation based on the observations made on the sampled units.

We use the term **sampling design** to refer to a particular choice of sampling frame and an associated sample selection method. These together lead to selection of a particular sample of frame units and/or basic population units. We will learn that the sampling design determines the set of all possible samples that can be selected and plays a key role in determination of errors of estimation.

The vast majority of the material considered in this text will assume that observation methods have been developed which will allow essentially error-free measurement of y (and also x) for all units that appear in a sample. In practice, of course, measurement errors are almost always non-zero and in some cases may be large. We therefore stress at the outset that it is critically important to establish rigorous and repeatable observation protocols that produce measurements of y (and x) that are, to the maximum extent practical, nearly error-free, and reproducible. When it is impossible to establish such protocols, it is critical that a survey has a sound basis for quantifying measurement errors and that these measurement errors are included in the overall error of estimation.

We will defer our consideration of complex sampling frames until later chapters, because it is easier to convey the basic principles of sampling theory for a simple frame. When measurement errors of y are close to zero and the sampling frame consists of the basic population units, then development of a sampling strategy depends only on specification of selection method(s) and specification of estimation equation(s). We therefore stress these two topics—specification of selection method(s) and estimation equation(s)—throughout this text.

2.3 Selection methods

The methods of selecting the units that appear in a sample may be quite simple or quite complex. Unit selection may be random and subject to chance (**probability sampling**), the approach that is the subject of this text, or unit selection may be purposive, based on professional judgment. We may thus refer to a **random sample** or to a **judgment** or **purposive sample**. Although **design-based** sampling theory, the subject that dominates the content of this text, is devoted entirely to randomized schemes for selection of units that appear in samples, some ecologists remain devoted to **representative sampling** in which professional expertise is used to select a particular location for research which is judged to adequately represent, say, a particular vegetation type (Mueller-Dombois and Ellenberg 1974 Chapter 5) or the *average* essential character of a small stream (Williams et al. 2004). A plausible argument may sometimes be made to justify such purposive or judgment sampling when sample sizes are extremely small. But for large sample sizes, it would be irrational to recommend this method of selection for two important reasons. First, judgment sampling lacks *objectivity*; selection of sample units involves subjective, personal choices. Second, it is impossible to describe or predict the statistical properties of judgment selection as a function of sample size or to attach a meaningful measure of uncertainty to estimates based on judgment samples.[3]

The random sample selection methods that we consider in this text all allow for calculation of the **inclusion probability**, π_i, that unit $i = 1, 2, \ldots, N$ in the population is included, by chance, in a random sample S of size n selected from a population of size N following some random selection method. If units are selected instead by purposive methods, then, for a given judge, the inclusion probability is 1 for those units that appear in the judgment sample and 0 for those that do not. Of course, a different judge would likely select a different judgment sample of the same fixed size n. Because inclusion probabilities for units vary among judges, selection of judgment samples lacks the kind of objectivity in sample selection that results when a randomized selection procedure is used instead, and unit selection is subject to chance rather than professional judgment. Randomized selection procedures allow an objective and statistically rigorous basis for inference from samples taken from finite populations.

In Section 2.6 we provide two illustrations of the dangers and limitations of judgment sampling. First, we provide a very brief vignette of a real-life setting where inferences based on judgment selection of representative sample locations appears to have generated seriously erroneous inferences. Second, we review the results from a classroom laboratory exercise that also suggest that judgment sampling can generate poor inferences. To fully appreciate these two illustrations, however, we first need to review the basic properties of estimators.

2.4 Properties of estimators

An **estimator** is a formula or calculation scheme that we use to calculate an **estimate** of a population parameter of interest based on measurements of y taken from units included in a random sample, S. For example, if we wish to estimate the mean of a target variable and we have used equal probability selection methods to select units from a simple sampling frame, then we may estimate the mean of the target variable, $\mu = \sum_{i=1}^{N} y_i / N$, using the following (no doubt familiar) calculation formula (estimator)

$$\hat{\mu} = \frac{\sum\limits_{i=1}^{n} y_i}{n}$$

where the index of the summation is over the $i = 1, 2, \ldots, n$ units that appear in the sample of size n selected from the finite population of size N, y_i denotes the value of the target variable associated with the i^{th} selected unit, and the caret or "hat" over μ indicates an estimator of μ. Note that the sample units in this estimator have been (implicitly) "relabeled" from the population to the sample. For example, if units 5, 7, and 13 had been selected in a sample of size $n = 3$ selected from the $N = 16$ units displayed in Section 2.1,

[3] In Chapter 7 we devote some attention to **model-based prediction** of finite population parameters. Although model-based prediction does not generally consider the probabilities associated with selection of sample units, purposive selection of *representative* locations would not generally be recommended for model-based inference, especially if the consequence were that the average values of auxiliary variables in the sample units were greatly different than the average values of these variables in the population (Valliant et al. 2000 Section 3.2).

then the target variable values in the population—y_5, y_7, and y_{13}—would be relabeled as the values—y_1, y_2, and y_3—in the sample.

To avoid the implicit "relabeling" of units implied by the familiar notation used for a sample mean, throughout this text we instead adopt the following equivalent notation for the same estimator

$$\hat{\mu} = \frac{\sum\limits_{i \in S} y_i}{n}$$

where $i \in S$ denotes the fixed unit labels of those n units that have been selected from N and are included in a random sample S. (A formal treatment of this topic is presented in Section A.6.1.) Using this notation, the y_i in a sample remain faithfully attached to their unit labels in the population.

One of the easiest ways to visualize the properties of estimators is to consider patterns of darts thrown at a dartboard. Imagine that the very center of the dartboard, the bullseye, represents the unknown value of a population parameter that we wish to estimate from a sample. There are typically a very large number of distinct samples that could possibly be selected. Each sample would contain a different set of units and associated y values, and would therefore produce a sample-specific estimate—the realized value of an estimator applied to a specific sample of units. (Note, however, that different samples may sometimes yield identical estimates.) Imagine that the different estimates resulting from selection of different samples could be visually represented by the scatter of darts thrown at the dartboard target. Dart locations that are close to the bullseye are *better* (i.e., have higher point values) than dart locations that are far from the bullseye, like an archery target.

Figure 2.3 displays four patterns of dart throws around a bullseye. Together, these patterns convey the most important properties of estimators. Imagine that Patterns (a) and (b) in Figure 2.3 represent the patterns of dart throws made by two highly skilled dart players who are competing at a local tavern, trying to hit the bullseye. They have both just arrived at the tavern; the dart player who produced Pattern (a) was using a set of darts that he had been using for years, whereas the dart player who produced Pattern (b) was using a new set of very expensive darts (never previously used). Pattern (a) is obviously the better pattern of dart throws. On average, the dart locations are centered around the bullseye and the locations are very similar to one another (i.e., the dart throws are highly reliable or precise). Overall, we would characterize this pattern of dart throws as having very high accuracy and very high point value. Pattern (b) is distinguished from Pattern (a) by the average location of the dart throws—on average they are quite far off target. But the reliability or precision of the throws is excellent and comparable to that of Pattern (a). We can recognize that Pattern (b) is less accurate than Pattern (a), because the average location of the darts is so far from the bullseye and the total point value would be less than for Pattern (a). Imagine that Patterns (c) and (d) reflect the locations of dart throws made by the same two individuals, but several hours after they first entered the tavern. (They will be relying on a designated driver to get them home!) On average, the dart throw locations for Pattern (c) remain centered about the bullseye, but the dart throws are now much less precisely located. Pattern (c) is obviously less accurate than Pattern (a) and possibly also less accurate than Pattern (b). For Pattern (d), the average location of dart throws remains far off target, and now the location of dart throws has become highly unreliable or imprecise. Clearly, this is the least accurate pattern of all and would have lowest point value.

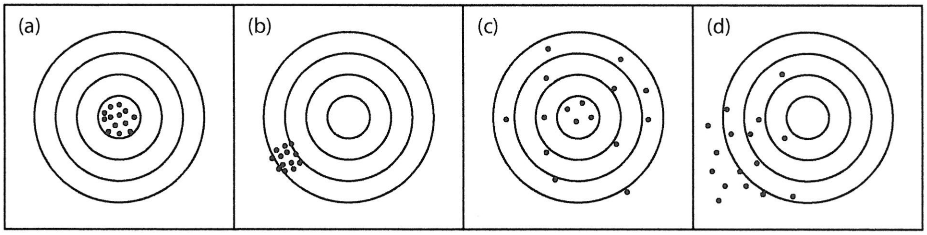


Fig. 2.3 Four patterns of dart throws aimed at the bullseye. Pattern (a) is approximately unbiased, has high precision and high accuracy. Pattern (b) has high precision, but because it is biased it is less accurate than Pattern (a). Pattern (c) is approximately unbiased but, because it has low precision, it has low accuracy. Pattern (d) is biased and has low precision and therefore has the lowest accuracy of the four dart patterns.

The clear conceptual differences among the patterns of dart throws in Figure 2.3, expressed in every day language, have important connections with the fundamental properties of estimators.

Expected value. The average value that an estimator takes on over all possible samples that can be selected according to some selection scheme. (Analogy: the average location of dart throws.)

Bias. The difference between the expected value of an estimator and the target population parameter. If the expected value is exactly equal to the target population parameter, then we have an **unbiased estimator.** (Analogy: the average dart location is exactly equal to location of the bullseye.) If instead the expected value differs from the population parameter (bullseye), then we have a **biased estimator.** Bias may be *positive* or *negative*, i.e., the expected value is greater than or less than the target population parameter.

Sampling variance. Sampling variance measures the variability among the estimates that arise from all possible samples that can be selected. This variability is measured as the average squared difference between estimates and the estimator's expected value. (Analogy: averaged squared distance from individual dart throw locations to the average location of darts.) Sampling variance is inversely related to the reliability or precision of an estimator. Sampling variance will be low when an estimator has high precision or reliability and vice versa.

Mean square error (MSE). Mean square error measures the overall accuracy of an estimator and is equal to the averaged squared difference between estimates and the target population parameter. (Analogy: average squared distance from individual dart locations to the bullseye.)

In Appendix A we show that mean square error equals sampling variance plus the square of the bias. Thus, for an unbiased estimator, mean square error is equivalent to sampling variance. Bias, if large, may make a substantial contribution to mean square error and thereby seriously reduce the accuracy of an estimator. Figure 2.4 uses a triangle analogy (Pythagorean theorem) to visually illustrate the relationships among bias, sampling variance, and mean square error. Note that a very reliable (low sampling variance) estimator with small bias may be a more accurate estimator than an unbiased but unreliable (high sampling variance) estimator.

If a serious attempt is made to "visually memorize" the patterns of dart throw locations displayed in Figure 2.3, the concepts of bias, sampling variance, and mean square error will

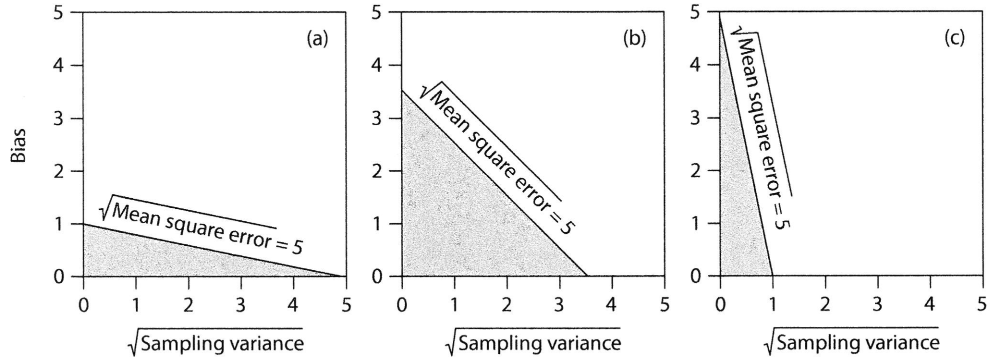


Fig. 2.4 Illustration of the Pythagorean relationship existing between an estimator's bias, sampling variance, and mean square error: bias^2 + sampling variance = mean square error. For each of the three examples, $\sqrt{\text{mean square error}} = \text{RMSE} = 5$ but the bias contributing to that RMSE differs: (a) bias = 1 accounts for a relatively small portion of the RMSE, (b) bias $\approx$ 3.536 and $\sqrt{\text{sampling variance}} \approx 3.536$ make equal contributions to the RMSE, and (c) bias $\approx$ 4.899 accounts for a relatively large portion of the RMSE.

become solidly imprinted in one's consciousness! The analogy of a pattern of darts thrown at a dartboard with the estimates generated from some sampling strategy is, however, deficient in two important respects. First, the "polar coordinate system" of the dartboard target makes it difficult to convey the notion of bias as being negative (on average, estimates are less than the target value) or positive (on average, estimates are greater than the target value). Second, this same coordinate system also complicates graphical representation of the distribution of possible sample estimates.

2.5 Sampling distribution of an estimator

Consider now a population consisting of $N = 240$ stones, with the variable of interest, y, being stone weight. The frequency distribution of this variable is displayed in Figure 2.5(a). Suppose that we wish to estimate the mean weight of a stone in this population, μ, by a "random selection" of 10 stones. Without getting into the details of the large number of ways that a sample of size $n = 10$ stones might be randomly selected, let us just imagine that we are using a sample selection method for which any particular set of 10 stones that might be selected is equally likely, and that we could in principle draw each of the possible distinct samples of 10 stones that could be selected according to this selection method (the Figure 2.5 caption provides additional details on sample selection). For each such sample, we could estimate μ by the mean weight of the sampled 10 stones, $\hat{\mu}$. We could then construct a frequency distribution of the estimates from all of these possible samples, as in Figure 2.5(b). If this frequency distribution were then scaled so that the total area was equal to one, we would have a graphical representation of the **sampling distribution** of the estimator $\hat{\mu}$. In this form, the histogram of the sampling distribution displays the probability that a sample estimate will fall within a specific bin range (Figure 2.5(b), probability axis). What do you think this distribution would look like if instead the sample size was, e.g., $n = 20$?

The sampling distribution of an estimator can be characterized by a minimum of two parameters: its mean (expected value) and its variance (sampling variance). If the mean

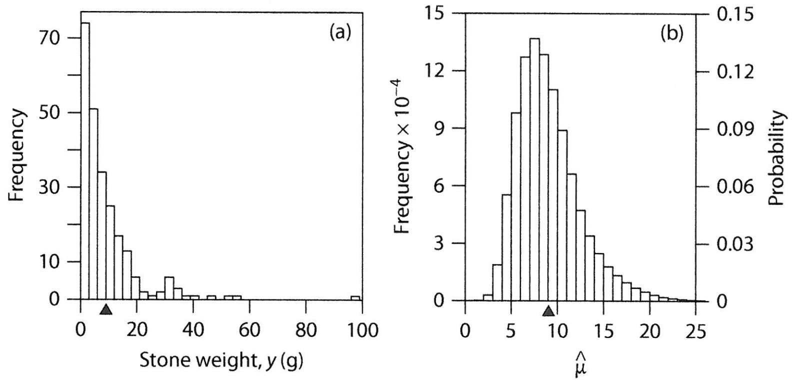


Fig. 2.5 Histograms of (a) population of $N = 240$ stone weights with mean weight $\mu = 9.03$ grams, and (b) sampling distribution of an estimator, $\hat{\mu}$, for the mean stone weight based on random sample selections of $n = 10$ stones, where the estimator was equal to the mean weight of the 10 sampled stones. Individual sample units were drawn at random from the population with equal probability (without replacement), and 10^6 such samples were drawn independently from each other using this same method. The mean of each distribution, 9.03 grams, is pointed to by the upper vertex of the black triangle located below each horizontal axis.

(expected value) of the sampling distribution is equal to the population parameter being estimated, then the estimator is unbiased (as in the example above). Otherwise, the estimator is biased, and the bias may be positive or negative. If the sampling distribution of the estimator has a very small range, then sampling variance will be small (the estimator is precise); if this range is very large, then sampling variance will be very large (the estimator is imprecise). An estimator's overall accuracy is measured by mean square error—sampling variance plus the squared bias.

Note that the sampling distribution of the estimator depicted in Figure 2.5(b) is not normally distributed. Normality is a highly desirable property of the sampling distribution of an estimator, but in Chapter 3 we will learn that the shape of the sampling distribution of an estimator depends on the shape of the distribution of the population variable *y* (from which the sample is drawn) and on sample size, *n*.

2.6 Judgment sampling versus random sampling

We begin our illustrations of the dangers of representative (purposive) selection of units with a brief (and greatly oversimplified) review of the history of surveys of spawning escapement (number of adult fish returning to freshwater streams to spawn) for Oregon coastal coho salmon (*Oncorhynchus kisutch*). Oregon coastal coho salmon historically supported extensive commercial and recreational fisheries. For the period 1958–1990, standardized survey protocols were executed in purposively selected *index reaches* ("representative" stream segments) (Jacobs and Cooney 1997). Mean densities (fish-per-mile) in these survey reaches were originally used primarily for qualitative assessment of trends in abundance, for which purpose they were no doubt useful (Chapter 13), but beginning with the passage of the federal Fishery Conservation and Management Act of 1976 and the subsequent introduction of intensive quantitative management of salmon fisheries, index reach densities were "scaled up" to produce quantitative estimates of spawning escapement for the Oregon coast, and these estimates were in turn used to develop harvest management policy. Beginning in 1990, a stratified random survey design was introduced for Oregon coastal coho, replacing the original index reaches by randomly selected reaches, while original index reaches were retained for comparison. Over the period 1990–1997, the mean density in these randomly selected reaches averaged just 27% of the mean density in the index reaches (Jacobs and Nickelson 1998), thereby generating greatly reduced estimates of coho salmon spawning escapement that were at odds with the overly optimistic index-based estimates from previous years that had been used for development of fishery management policy, and likely led to unintentional overfishing of Oregon coastal coho salmon. These much lower, but assumed unbiased, estimates of spawning escapement played an important role in the subsequent determination, in 1998, that Oregon coastal coho salmon were in a seriously depressed state, and they were subsequently listed as "threatened" under the federal Endangered Species Act. Presumably, the original index reaches had either been (a) deliberately selected to represent *excellent* or *good* rather than *average* spawning habitat, or (b) judgment selection of *representative/average* reaches had been quite seriously off the mark.

Our second illustration of the dangers and inadequacy of judgment sampling comes from a repeatable lab experiment that directly contrasts the performance of randomized selection of sample units with judgment selection. One of the simplest randomized selection procedures is **simple random sampling** or **SRS**. With SRS, one selects n sample units from N population units at random with equal probability, without replacement. For example, imagine a box with $N = 100$ balls that are physically identical except for an identifying number, i $(i = 1, 2, \ldots, 100)$, on the surface of the ball. To draw a simple random sample, we could reach into the box to select one ball at random with equal probability (i.e., all balls in the box being equally likely to be the one selected), record the ball number, and set the ball aside. We could then reach into the box again and select another ball at random with equal probability from the remaining 99 balls, record the ball number, set the ball aside, and so on until we have selected n balls. But, would this sampling approach actually result in a simple random sample? How likely is it that the balls positioned near the top, bottom, and sides of the box have the same probability of being selected as a ball near the middle of the box? This is one reason why casino and lottery games based on a random drawing of numbered ping-pong balls from a container continually mix up the balls (e.g., by rotating the container, injecting pulses of air, etc.). But what if the balls were instead objects—not identical in size and shape—how difficult would it then be to insure that the physical selection of objects was at random with equal probability? For these and other reasons, to insure that a sample is selected in a manner that meets the definition of the random sampling method (here SRS), the sample is usually selected as follows:

(1) the population units are listed by an identifying number, i $(i = 1, 2, \ldots N)$;
(2) a computer is used to draw a sample of n of these unit numbers according to the definition of the random sampling method using a random number generator; and
(3) these units are then selected for the sample.

Thus, with our box of balls example, we would draw an SRS sample of unit numbers first, and then select the associated balls from the box. In the R statistical/programming/graphics language and environment (R Core Team 2018), a simple random sample of unit numbers of size $n = 5$ from $N = 100$ can be selected using the following R expression: `sample(x=1:100, size=5, replace=FALSE)`. The result, e.g., {15, 48, 69, 67, 35}, is a listing of the units to include in the sample. If this R expression were executed again, a second SRS sample of unit numbers, selected independently of the first, might result in, e.g., {33, 12, 18, 51, 86}. (In Chapter 3 we present a more formal explanation of SRS and we show that there are about 75.3 million(!) distinct simple random samples of size 5 that can be selected from a population of size 100.)

Jessen (1978) describes an experiment with a population of stones that we have repeated in sampling theory lab sessions with success for many years. We bring to the lab room a large set of agates (a semi-precious gemstone composed of chalcedony) collected from northern California ocean beaches. The entire population of $N = 125$ agates is displayed on a table top (see Figure 2.1), so that students can view all of the population units. The population parameter to be estimated is the mean agate weight, μ. Individual students are asked to select, using their best *judgment*, samples of size $n = 1, 2, 3, 4, 5, 6, 8, 10, 20$ that they think, when the sample mean weight is taken, will result in their best estimate of μ for that value of n. There is considerable variation in the weights of the agates and the human eye has difficulty judging weight or volume, so this task is more difficult than one might suspect. During one of the more recent lab sessions, with results comparable to those on many previous occasions, $M = 22$ students recorded the total weight of agates for each sample that was selected and, as a subsequent lab exercise, students calculated the corresponding mean weight ($\hat{\mu}$) for each sample selected. The students were then informed of the population mean weight ($\mu = 14.843$ g), and for each n were asked to calculate the following performance measures for their collective judgment sampling effort. Letting $\hat{\mu}_j$ denote the sample mean weight obtained by student judge j, $j = 1, 2, \ldots, M$: (1) average value $= \sum_{j=1}^{M} \hat{\mu}_j / M$ (analogous to expected value), (2) bias $=$ average value $- \mu$, (3) proportional bias $=$ bias$/\mu$, (4) variance $= \sum_{j=1}^{M} (\hat{\mu}_j -$ average value$)^2 / M$ (analogous to sampling variance), and (5) variance $+$ bias^2 (analogous to mean square error). The analogies are a bit imperfect, of course, because the lab participants are only a small number of individuals out of billions of potential judges and it is risky to generalize from the performance of so few students.

As we will learn in Chapter 3, if SRS is used to select samples of size n from a population of size N, then the sample mean is an unbiased estimator of μ, with sampling variance equal to $\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$, where σ^2 is the finite population variance (a measure of the variation in agate weights, defined in Chapter 3), here equal to 307.383. With SRS, the bias and proportional bias of the sample mean is zero, and therefore its mean square error is the same as its sampling variance. But how might SRS perform compared to judgment sampling assuming the sample mean is used to estimate μ for each method of selection?

Figure 2.6 summarizes the collective performance of judgment sampling for the 22 students compared to that for SRS and its dependence on n. Panel (a) displays the sample mean weights obtained by the individual students at each value of n, along with a horizontal dashed line referencing the population mean value of $\mu = 14.843$ g. Note that the bias for judgment sampling [panel (b)] is at first very high (proportional bias is more than 90% for $n = 1$) and gradually declines with increasing n, but is negligible only for $n = 20$. In contrast, with SRS the sample mean is unbiased for all values of n. Although for small samples sizes ($n \leq 2$) the sampling variance for judgment sampling [panel (c)] is less than the sampling variance for SRS, for larger sample sizes the sampling variance for SRS is always less than that for judgment sampling. Finally, the mean square error for SRS is less than that for judgment sampling for all values of n [panel (d)]. Thus, the accuracy with SRS is greater than with judgment sampling when using the sample mean as an estimator of μ, at least for the values of n evaluated.

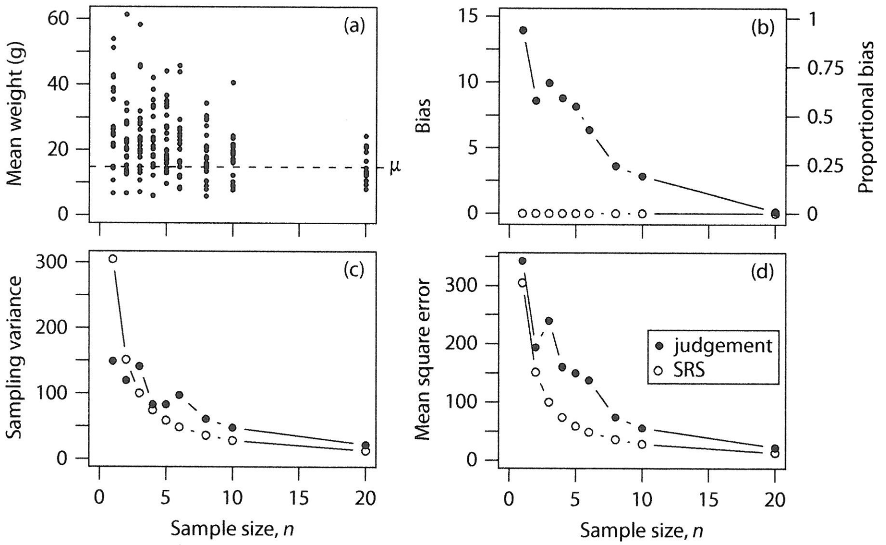


Fig. 2.6 Performance of 22 student judges in estimating the mean weight ($\mu = 14.843$ g) of agates in a small population ($N = 125$) with considerable variation in agate weights ($\sigma^2 = 307.383$), compared to performance for SRS. Both approaches used the sample mean as the estimator of μ. See text for definition of performance metrics.

Overall it appears that the only potential advantage of judgment sampling over SRS might be for very small sample sizes, when the variance among judges is less than the sampling variance for SRS. If judges could improve their skills in selection of agates, so as to dramatically reduce their bias, then judgment sampling might be recommended for the very smallest sample sizes. Otherwise, the performance of judgment sampling seems clearly inferior to SRS. It is also important to note that all units of the population were on full display in this sampling exercise. This is the most favorable kind of setting that one might possibly imagine for application of judgment sampling. When a field ecologist makes a judgment selection of a "representative" patch of vegetation or section of stream, etc., the choice is made based on his/her past experience which generally does not include a simultaneous view of the physical attributes of the entire population or even a view of these same physical attributes over the course of a professional career. Against this backdrop, it is all the more remarkable that the performance for SRS—in which units are selected purely at random, without regard to the y values—compares so favorably to that for judgment sampling. Wouldn't you think that an average student judge selecting a single agate could do a much better job purposively selecting a *single representative agate* than a randomized selection procedure that is just as likely to select the heaviest agate or the lightest agate as to select an "average" weight agate? Imagine how much more difficult it might be to purposively select a representative patch of a vegetation type or a representative section of a stream.