Chapter 1

Introduction, Basic Concepts, and Assumptions

OBJECTIVES AND OVERVIEW

This chapter provides the foundation for our discussion throughout the book. Its emphasis is on basic concepts and assumptions. The topics discussed in this chapter are:

- The concept of statistical inference (Section 1.1).
- An overview of different types of statistical intervals: confidence intervals, tolerance intervals, and prediction intervals (Section 1.2).
- The assumption of sample data (Section 1.3) and the central role of practical assumptions about the data being "representative" (Section 1.4).
- The need to differentiate between enumerative and analytic studies (Section 1.5).
- Basic assumptions for inferences from enumerative studies, including a brief description of different random sampling schemes (Section 1.6).
- Considerations in conducting analytic studies (Section 1.7).
- Convenience and judgment samples (Section 1.8).
- Sampling people (Section 1.9).
- The assumption of sampling from an infinite population (Section 1.10).
- More on practical assumptions (Sections 1.11 and 1.12).
- Planning the study (Section 1.13).
- The role of statistical distributions (Section 1.14).

Statistical Intervals: A Guide for Practitioners and Researchers, Second Edition.
William Q. Meeker, Gerald J. Hahn and Luis A. Escobar.
 Published 2017 by John Wiley & Sons, Inc.
Companion Website: www.wiley.com/go/meeker/intervals

- The interpretation of a statistical interval (Section 1.15).
- The relevance of statistical intervals in the era of big data (Section 1.16).
- Comment concerning the subsequent discussion in this book (Section 1.17).

1.1 STATISTICAL INFERENCE

Decisions frequently have to be made from limited sample data. For example:

- A television network uses the results of a sample of 1,000 households to determine advertising rates or to decide whether or not to continue a show.
- A company uses data from a sample of five turbines to arrive at a guaranteed efficiency for a further turbine to be delivered to a customer.
- A manufacturer uses tensile strength and other measurements obtained from a laboratory test on ten samples of each of two types of material to select one of the two materials for use in future production.

The sample data are often summarized by statements such as:

- 293 out of the 1,000 sampled households were tuned to the show.
- The mean efficiency for the sample of five turbines was 67.4%.
- The samples using material A had a mean tensile strength 3.2 units larger than those using material B.

The preceding "point estimates" provide a concise summary of the sample results, but they give no information about their precision. Thus, there may be big differences between such point estimates, calculated from the sample data and what one would obtain if unlimited data were available. For example, 67.4% would seem a reasonable estimate (or prediction) of the efficiency of the next turbine. But how "good" is this estimate? By noting the variation in the observed efficiencies of the five turbines, we know that it is unlikely that the turbine to be delivered to the customer will have an efficiency of *exactly* 67.4%. We may, however, expect its efficiency to be "close to" 67.4%. But how close? Can we be reasonably confident that it will be within $\pm 0.1\%$ of the point estimate 67.4%? Or within $\pm 1\%$? Or within $\pm 10\%$? We need to quantify the uncertainty associated with our estimate or prediction. An understanding of this uncertainty is an important input for decision making, for example, in providing a warranty on product performance. Moreover, if our knowledge, as reflected by the width of the uncertainty interval, is too imprecise, we may wish to obtain more data before making an important decision.

The example suggests quantifying uncertainty by constructing statistical intervals around the point estimate. This book shows how to obtain such intervals. We describe frequently needed, but not necessarily well-known, statistical intervals calculated from sample data, differentiate among the various types of intervals, and show their applications. Methods for obtaining each of the intervals are presented and their use is illustrated. We also show how to choose the sample size so as to attain a desired degree of precision, as measured by the width of the resulting statistical interval. Thus, this book provides a comprehensive guide and reference to the use of statistical intervals to quantify the uncertainty in the information about a sampled population or process, based upon a possibly small, but randomly selected sample. The concept of a random sample is discussed further in Section 1.6.3.

1.2 DIFFERENT TYPES OF STATISTICAL INTERVALS: AN OVERVIEW

Various types of statistical intervals may be calculated from sample data. The appropriate interval depends upon the specific application. Frequently used intervals are:

- A *confidence interval* to contain an unknown characteristic of the sampled population or process. The quantity of interest might be a population property or "parameter," such as the mean or standard deviation of the population or process. Alternatively, interest might center on some other property of the sampled population, such as a quantile or a probability. Thus, depending upon the question of interest, one might compute a confidence interval that one can claim, with a specified high degree of confidence, contains (1) the mean tensile strength, (2) the standard deviation of the distribution of tensile strengths, (3) the 0.10 quantile of the tensile strength distribution, or (4) the proportion of specimens that exceed a stated threshold tensile strength value.
- A *statistical tolerance interval* to contain a specified proportion of the units from the sampled population or process. For example, based upon a random sample of tensile strength measurements, we might wish to compute an interval to contain, with a specified degree of confidence, the tensile strengths of at least a proportion 0.90 of the units from the sampled population or process. Hereafter we will generally simply refer to such an interval as a "tolerance interval."
- A *prediction interval* to contain one or more future observations, or some function of such future observations, from a previously sampled population or process. For example, based upon a random sample of tensile strength measurements, we might wish to construct an interval to contain, with a specified degree of confidence, (1) the tensile strength of a randomly selected single future unit from the sampled process (this was of interest in the turbine efficiency example), (2) the tensile strengths for *all* of five future units, or (3) the mean tensile strength of five future units.

Most users of statistical methods are familiar with (the common) confidence intervals for the population mean and for the population standard deviation, but often not for population quantiles or the probability of exceeding a specified threshold value. Some, especially in industry, are also aware of tolerance intervals. Despite their practical importance, however, most practitioners, and even many professional statisticians, know very little about prediction intervals except, perhaps, for their application to regression problems. A frequent mistake is to calculate a confidence interval to contain the population mean when the problem requires a tolerance interval or a prediction interval. At other times, a tolerance interval is used when a prediction interval is needed. Such confusion is understandable, because statistics textbooks typically focus on the common confidence intervals, occasionally make reference to tolerance intervals, and consider prediction intervals only in the context of regression analysis. This is unfortunate because in applications, tolerance intervals, prediction intervals, and confidence intervals on distribution quantiles and on exceedance probabilities are needed almost as frequently as the better-known confidence intervals. Moreover, the calculations for such intervals are generally no more difficult than those for confidence intervals.

1.3 THE ASSUMPTION OF SAMPLE DATA

In this book we are concerned only with situations in which uncertainty is present because the available data are from a random sample (often small) from a population or process. There are, of course, some situations for which there is little or no such statistical uncertainty. This is

the case when the relevant information on *every unit* in a finite population has been recorded without measurement error, or when the sample size is so large that the uncertainty in our estimates due to sampling variability is negligible (as we shall see, how large is "large" depends on the specific application). Examples of situations in which one is generally dealing with the entire population are:

- The given data are census information that have been obtained from all residents in a particular city (at least to the extent that the residents could be located and are willing to participate in the study).
- There has been 100% inspection (i.e., all units are measured) of a performance property for a critical component used in a spacecraft.
- A complete inventory of all the parts in a warehouse has been taken.
- A customer has received a one-time order of five parts and has measured each of these parts. Even though the parts are a random sample from a larger population or process, as far as the customer is concerned the five parts make up the entire population of interest.

Even in such situations, intervals to express uncertainty are sometimes needed. Suppose, for example, that based upon extensive data, we *know* that the weight of a product is approximately normally distributed with a mean of 16.10 ounces and a standard deviation of 0.06 ounces. We wish an interval to contain the weight of a single unit randomly selected by a customer, or by a regulatory agency. The calculation of the resulting *probability interval* is described in books on elementary probability and statistics. Such intervals generally assume complete knowledge about the population (e.g., the mean and standard deviation of a normal distribution). In this book, we are concerned with the more complicated problem where, for example, the population mean and standard deviation are *not* known but are estimated, subject to sampling variability. In particular, tolerance intervals and prediction intervals converge to probability intervals as the sample size increases. On the other hand, because there is no statistical uncertainty remaining, the width of confidence intervals converges to zero with increasing sample size.

Statistical uncertainty also exists, even though the entire population has been evaluated, when the readings are subject to measurement error. For example, one might determine that in measuring a particular property, 971 out of 983 parts in a production lot are found to be within specification limits. Due to measurement error, however, the actual number of parts within specifications may not exactly equal 971. Moreover, if something is known about the statistical distribution of measurement error, one can then also quantify the uncertainty associated with the estimated number of parts within specifications (e.g., Hahn, 1982).

Finally, we note that even when there is no quantifiable statistical uncertainty, there may still be other uncertainties of the type suggested in the discussion to follow.

1.4 THE CENTRAL ROLE OF PRACTICAL ASSUMPTIONS CONCERNING REPRESENTATIVE DATA

We have briefly described different statistical intervals that a practitioner might use to express the uncertainty in various estimates or predictions generated from sample data. This book presents the methodology for calculating such intervals. Before proceeding, we need to make clear the major practical assumptions dealing with the "representativeness" of the sample data. We do this in the following sections. Departures from these implicit assumptions are common in practice and can invalidate any statistical analyses. Ignoring such assumptions can lead to a false sense of security, which, in many applications, is the weakest link in the inference process. Thus,

for example, product engineers need to question the assumption that the performance observed on prototype units produced in the lab also applies for production units, to be built much later, in the factory. Similarly, a reliability engineer should question the assumption that the results of a laboratory life test will adequately predict field failure rates. In fact, in some studies, the assumptions required for the statistical interval to apply may be so far off the mark that it would be inappropriate, and perhaps even misleading, to use the formal methods presented here.

In the best of situations, one can rely on physical understanding, or information from outside the study, to justify the practical assumptions. Such evaluations, however, are principally the responsibility of the subject-matter expert. Often, the assessment of such assumptions is far from clear-cut. In any case, one should keep in mind that the intervals described in this book reflect only the statistical uncertainty due to limited data. In practice, the actual uncertainty will be larger because the generally unquantifiable deviations of the practical assumptions from reality provide an added *unknown* element of uncertainty beyond that quantified by the statistical interval. If there were formal methods to reflect this further uncertainty (occasionally there are, but often there are not), the resulting interval, expressing the *total* uncertainty, would be wider than the statistical interval alone. This observation suggests a rationale for calculating a statistical interval for situations where the basic assumptions are questionable. If it turns out that the calculated statistical interval is wide, we then know that our estimates have much uncertainty—even *if* the assumptions were all correct. A narrow statistical interval would, on the other hand, imply a small degree of uncertainty *only if* the required assumptions hold.

Because of their importance, we feel it appropriate to review, in some detail, the assumptions and limitations underlying the use and interpretation of statistical intervals before proceeding with the technical details of how to calculate such intervals.

1.5 ENUMERATIVE VERSUS ANALYTIC STUDIES

Deming (1953, 1975, 1986) emphasizes the important differences between "enumerative" and "analytic" studies (a concept that he briefly introduced earlier in Deming, 1950). Despite its central role in making inferences from the sample data, many traditional textbooks in statistics have been slow in giving this distinction the attention that it deserves.

To point out the differences between these two types of studies, and some related considerations, we return to the examples of Section 1.1. The statements there summarize the sample data. In general, however, investigators are concerned with making inferences or predictions *beyond* the sample data. Thus, in these examples, the real interest was, not in the sample data per se, but in:

1. The proportion of households in the *entire country* that were tuned to the show.
2. The efficiency of the, *as yet not manufactured*, turbine to be sent to the customer.
3. A comparison of the mean tensile strengths of the *production units to be built* in the factory *some time in the future* using material A and material B.

In the first example, our interest centers on a finite identifiable unchanging collection of units, or population, from which the sample was drawn. This population, consisting of all the households in the country with access to the show, exists at the time of sampling. Deming uses the term "enumerative study" to describe such situations. More specifically, Deming (1975, page 147), defines an enumerative study as one in which "action will be taken on the material in the frame studied," where he uses the conventional definition of a frame as "an aggregate of identifiable units of some kind, any or all of which may be selected and investigated. The frame may be lists of people, areas, establishments, materials, or other identifiable units that

would yield useful results if the whole content were investigated." Thus, the frame provides a finite list, or other identification, of distinct (non-overlapping) and exhaustive sampling units. The frame defines the population to be sampled in an enumerative study.

Some further examples of enumerative studies are:

- Public opinion polls to assess the *current view* on some specified topic(s) of the entire US adult population, or some defined segment thereof, such as all registered voters in a specified locality.
- Sample audits to assess the correctness of last month's bills and to estimate the total error in such bills. In this case, the population of interest consists of all of last month's bills.
- Product acceptance sampling to decide on the disposition of a particular production lot. In this case, the population of interest consists of all units in the production lot being sampled.

In an enumerative study, the correctness of statistical inferences requires a random sample from the frame. Such a sample, is, at least in theory, generally attainable; see Section 1.6.

In contrast, the second two examples of Section 1.1 (dealing, respectively, with the efficiency of a future turbine and the comparison of two materials) illustrate what Deming (1975, page 147) calls "analytic studies." We no longer have an existing, finite, well-defined, unchanging population. Instead, we want to take action to improve or make predictions about the output of a, sometimes hypothetical, future *process* based upon data from an existing (likely different) process.

Specifically, Deming (1975) defines an analytic study as one "in which action will be taken on the process or cause-system . . . the aim being to improve practice in the future . . . . Interest centers in future product, not in the materials studied." He cites as examples "tests of varieties of wheat, comparison of machines, comparisons of ways to advertise a product or service, comparison of drugs, action on an industrial process (change in speed, change in temperature, change in ingredients)." We may wish to use data from an existing process to predict the characteristics of future output from the same or a similar process. Thus, in a prototype study of a new product, interest centers on the process that will manufacture the product in the future.

These examples are representative of many encountered in practice, especially in engineering, medical, and other scientific investigations. In fact, the great majority of applications that we have encountered in practice involve analytical, rather than enumerative studies. It is, moreover, inherently more complicated to draw inferences from analytic studies than from enumerative studies because analytic studies require the critical (and often unverifiable) added assumption that the process about which one wishes to make inferences is statistically identical to that from which the sample was selected.

1.5.1 Differentiating between Enumerative and Analytic Studies

What one wishes to do with the results of the study is often a major differentiator between an enumerative and an analytic study. Thus, if one's interest is limited to describing an existing population, one is dealing with an enumerative study. On the other hand, if one is concerned with a process that is to be improved, or is otherwise subject to change, perhaps as a result of the study, then one is clearly dealing with an analytic study.

Deming (1975) presents a "simple criterion to distinguish between enumerative and analytic studies. A 100% sample of the frame answers the question posed for an enumerative study, subject of course to the limitations of the method of investigation. In contrast, a 100% sample . . . is still inconclusive in an analytic problem." This is because for an analytic study our real interest is in a process that will be operating in the future. Deming's rule can be useful when the differentiation between an analytic and an enumerative study does not seem clear-cut. For example, an "exit poll" to estimate the proportion of voters who have voted (or, at least, assert

that they have voted) for a particular candidate, based upon a random sample of individuals leaving the polling booth, is an example of an enumerative study. In this case, a 100% sample provides perfect information (assuming 100% correct responses). In contrast, estimating, before the election, the proportion of voters who will *actually* go to the polls and vote for the candidate involves an analytic study, because it deals with a future process. Thus, between the time of the survey and election day, some voters may change their minds, perhaps as a result of some important external event—or even as a consequence of action taken by one or more of the candidates based upon information obtained in the study. Also, extraneous factors, such as adverse weather conditions on election day (not contemplated on the sunny day on which the poll was conducted), might stop some going to the polls—and the "stay-at-homes" may well differ in their voting preferences from those who do vote. Thus, even if we had sampled every eligible voter prior to the election, we still would not be able to predict the outcome with certainty, because we do not know who will actually vote and who will change their mind in the intervening period. (Special considerations in sampling people are discussed in Section 1.9.)

Taking another example, it is sometimes necessary to sample from inventory to make inferences about a product population or process. If interest focuses merely on characterizing the current inventory, the study is enumerative. If, however, we wish to predict the future performance of the product, perhaps after making design changes, the study is analytic. Finally, drawing conclusions about the performance of a turbine to be manufactured in the future, based upon data on turbines built in the past, involves, as we have indicated, an analytic study. If, however, the measured turbines *and* the turbine to be shipped were all independently and randomly selected from inventory (unlikely to be the case in practice), one would be dealing with an enumerative study.

1.5.2 Statistical Inference for Analytic Studies

We do not agree with the views of some (e.g., Gitlow et al., 1989, page 558) who imply that statistical inference methods, such as statistical intervals, have no place whatsoever in analytic studies. Indeed, such methods have been used successfully for decades in science and industry in studies that have been predominantly analytic. Many statistical methods were, in fact, developed with such studies in mind. Instead, we feel that the decision of whether or not to use statistical intervals in analytic studies needs to be made on a case by case basis. Use of statistical intervals in such studies requires a keen understanding and assessment of the additional assumptions that are being made.

1.5.3 Inferential versus Predictive Analyses

In addition to differentiating between analytic and enumerative studies, it is also useful to differentiate between inferential and predictive analyses. Broadly speaking, the goal of an inferential analysis is to gain an understanding of the mechanism that underlies or resulted in the observed data. A typical example is that of a *manufacturer* wishing to determine how different processing (and possibly environmental) variables impact the performance of a product with the goal of building an improved product in the future.

In contrast, the goal of a predictive analysis is typically to predict future performance—without necessarily understanding the underlying mechanism. In the preceding example, such analyses might be of principal interest to the *purchaser* of the product who wishes to predict future performance.

Both inferential and predictive analyses can involve either enumerative or analytic studies. In our example, the study is analytic when the underlying conditions under which the data were obtained differ from those under which one wishes to draw conclusions—irrespective of whether one is interested in gaining an understanding of the impact of different processing

variables or predicting future performance. Thus, the available data might be from in-house testing on early production units, but the inferences or predictions to be made deal with field exposure of high volume production, making the study analytic in both cases. We will now consider, in further detail, the basic assumptions underlying inferences from enumerative and analytic studies.

1.6 BASIC ASSUMPTIONS FOR INFERENCES FROM ENUMERATIVE STUDIES

1.6.1 Definition of the Target Population and Frame

In enumerative studies there is some "target population" about which it is desired to draw inferences. An important first step—though one that is sometimes omitted by analysts—is that of explicitly and precisely defining this target population. For example, the target population may be all the automobile engines of a specified model manufactured on a particular day, or in a specified model year, or over some other defined time period. In addition, one need also make clear the specific characteristic(s) to be evaluated. This may be a measurement or other reading on an engine, or the time to failure of a part on a life test, where "failure" is precisely defined. Also, in many applications, and especially those involving manufactured products, one must clearly state the operating environment in which the defined characteristic is to be evaluated. For a life test, this might be "normal operating conditions," and exactly what constitutes such conditions needs to be clearly stated.

The next step is that of establishing a frame from which the sample is to be taken. Establishing a frame requires obtaining or developing a specific listing, or other enumeration of the population from which the sample will be selected. Examples of frames are the serial numbers of all the automobile engines built over the specified time period, the complete listing of email addresses of the members of an organization, the schedule of incoming commercial flights into an airport on a given day, or a tabulation of all invoices billed during a calendar year. Often, the frame is *not* identical to the target population. For example, a listing of land-line telephone numbers generally corresponds to households, rather than individuals, and omits those who do not have a telephone or who have only a cell phone, people with unlisted phone numbers, new arrivals in the community, etc.—and also may include businesses, which are not always clearly identified as such. If the telephone company wishes to estimate the proportion of listed land-line phones in working order at a given time, a complete listing of such telephones (available to the phone company) will probably coincide with the target population about which inferences are desired. For most other studies, however, there may be an important difference between the frame (i.e., the telephone directory listing) and the target population.

The listing provided by the frame will henceforth be referred to as the "sampled population." Clearly, the inferences from a study, such as those quantified by statistical intervals, will be on the frame and—when the two differ—not on the target population. Thus, our third step—after defining the target population and the frame—is that of evaluating the differences between the two and the possible effect that such differences could have on the conclusions of the study. Moreover, it warrants repeating that these differences introduce uncertainties above and beyond those quantified by the statistical intervals provided in this book. If well understood, these differences can, at least sometimes and to some degree, be dealt with.

1.6.2 The Assumption of a Random Sample

The data are assumed to be a random sample from the frame. Because we deal only with random samples in this book, we will sometimes use the term "sample" to denote a random sample. In

enumerative studies, we will be concerned principally with the most common type of random sampling, namely *simple random sampling.* We briefly describe other types of random samples in Section 1.6.3.

Simple random sampling gives every possible sample of n units from the frame the same probability of being selected. A simple random sample of size n can, at least in theory, be obtained from a population of size N by numbering each unit in the population from 1 to N, placing N balls bearing the N numbers into a bin, thoroughly mixing the balls, and then randomly drawing n balls from the bin. The units to be sampled are those with numbers corresponding to the n selected balls. In practice, tables of random numbers generated by computer algorithms (e.g., Rizzo, 2007; Ripley, 2009; Gentle, 2009, 2013) and by statistical computing software (e.g., R Core Team, JMP or Minitab), provide easier ways of obtaining a random sample.

The assumption of random sampling is of critical importance in constructing statistical intervals. This is because such intervals reflect only the randomness due to the sampling process and do not take into consideration biases that might be introduced by not sampling randomly. It is especially important to recognize this limitation because in many studies, and especially ones involving sampling people, one does not have a strict random sample; see Section 1.9.

1.6.3 More Complicated Random Sampling Schemes

There are also other random sampling methods beyond simple random sampling, such as stratified random sampling, cluster random sampling, and systematic random sampling. These are used frequently in such applications as sampling of human populations, auditing, and inventory estimation. For such samples, rather than every possible sample of n units from the sampled population having the *same* probability of being selected, each possible sample has a *known* probability of being selected. Statistical intervals can also be constructed for such more complicated sampling schemes; these intervals are generally more complicated than the ones for simple random samples. The interested reader is referred to books referenced in the Bibliographic Notes section at the end of this chapter.

The more complicated *random* sampling schemes, described briefly below, need to be differentiated from various *nonrandom* sampling schemes, which we describe in Section 1.8 under the general heading of "convenience and judgment sampling."

Stratified random sampling

In some sampling applications, the population is naturally divided into non-overlapping groups or *strata*. For example, a population might be divided according to gender, job rank, age group, geographic region, or manufacturer. It may be important or useful to take account of these strata in the sampling plan because:

- Some important questions may focus on individual strata (e.g., information is needed by geographic region, as well as for the entire country).
- Cost, methods of sampling, access to sample units, or resources may differ among strata (e.g., salary data are generally more readily available for individuals working in the public sector than for those working in the private sector).
- When the response variable has less variability within a stratum than across strata (i.e., across the entire population), stratified random sampling provides more precise estimates than one would obtain from a simple random sample of the same size. The increase in precision results in narrower (but more complicated to construct) statistical intervals concerning the entire population.

In stratified sampling, one takes a simple random sample from each stratum in the population. The methods presented in this book can be used with data from a single stratum to compute statistical intervals for that stratum. However, when stratified sample data are combined across strata (e.g., to obtain a confidence interval for the population mean or total) special methods are needed, as described in textbooks on survey sampling. These books also discuss how to allocate units across strata and how to choose the sample size within each stratum.

Cluster random sampling

In some studies it is less expensive to obtain samples by using "clusters" of "elements" (generic terminology) that are conveniently located together in some manner, instead of taking a simple random sample from the entire population. For example, when items are packed in boxes, it is often easier to take a random sample of boxes and either evaluate all items in each selected box—or take further random samples within each of the selected boxes—rather than to randomly sample individual items irrespective of the box in which they are contained. Also, it may be more natural and convenient to interview some or all adult members of randomly selected families rather than randomly selected individuals from a population of individuals. Finally, it is often easier to find a frame for, and sample groups of, individuals clustered in a random sample of locations rather than taking a simple random sample of individuals spread over, say, an entire city. In other cases, only a listing of clusters, but not of the individuals or items they contain, may be available; clusters are then the natural sampling units. In each of these cases, one needs to define the clusters, obtain a frame that lists all clusters, and then take a random sample of clusters from that frame. Sometimes, as previously indicated, responses are then obtained for all elements (i.e., individuals or items) within each selected cluster, although often subsampling within clusters is conducted.

The value of the information for each additional sample unit within a cluster can be appreciably less than that of an individual unit for a simple random sample, especially if the items in a cluster tend to be similar and many units are chosen from each cluster. On the other hand, if the elements in the clusters are a well-mixed representation of the population, the loss in precision due to the use of cluster sampling may be slight; often, the lower per-element cost of cluster sampling will more than compensate for the loss in statistical efficiency. Thus, given a specified total cost to conduct a study, the net result of cluster sampling can be an improvement in overall statistical efficiency, as evidenced by narrower statistical intervals compared to those for a simple random sample—even though the simple random sample requires a smaller total sample size. Also, when the investigator has a say in choosing the cluster size, the loss of efficiency might be mitigated by taking a larger sample of smaller clusters. (When clusters contain only one element, one is back to simple random sampling.) Books on survey sampling, such as those mentioned in the Bibliographic Notes section at the end of this chapter, provide details of cluster sampling and sample size selection.

Systematic random sampling

It is often much easier to select a sample in a systematic manner than to take a simple random sample. For example, because there is no readily available frame, it might be difficult to obtain a simple random sample of all the customers who come into a store on a particular day. It would, however, be relatively simple to sample every 10th, or other preselected number, person entering the store. Similarly, it might be much easier and more natural to have a clerk examine every 10th item in a file cabinet, instead of choosing a simple random sample of all items. In both examples, the ratio of cost incurred to information gained to conduct the study might be appreciably smaller with systematic sampling than with simple random sampling. In both

examples, it would usually be cost effective to use systematic, rather than simple, random sampling. In some situations a systematic sample may, in fact, be the only feasible alternative.

Systematic samples are random samples as long as a random starting point is used. Special methods and formulas, however, are needed to compute statistical intervals. Also, the systematic pattern that is to be used in sampling must be chosen carefully. Serious losses of efficiency or biases may result if there are periodicities in the sampled population and if these are in phase with the systematic sampling scheme. For example, if a motor vehicle bureau measures traffic volume each Wednesday (where Wednesday is a randomly selected weekday), the survey results would likely provide a biased estimate of average weekday traffic volume. On the other hand, if such sampling took place every sixth weekday, the resulting estimate would likely be a lot more reasonable. Books on survey sampling, such as those mentioned in the Bibliographic Notes section at the end of this chapter, provide details of systematic sampling.

1.7 CONSIDERATIONS IN THE CONDUCT OF ANALYTIC STUDIES

1.7.1 Analytic Studies

In an enumerative study, one generally wishes to draw inferences by sampling from a well-defined existing population, the members of which can be enumerated, at least conceptually—even though, as we have seen, difficulties can arise in finding a frame that adequately represents the target population and in obtaining a random sample from that frame. In contrast, in an analytic study one wishes to draw conclusions about a process that may not even exist—or may not be accessible—at the time of the study. As a result, the process that is sampled is likely to differ, in various ways, from the one about which it is desired to draw inferences. As we have indicated, sampling prototype units, made in the lab or on a pilot production line, to draw conclusions about subsequent full-scale production is one common example of an analytic study.

1.7.2 The Concept of Statistical Control

A less evident example of an analytic study arises if, in dealing with a mature production process, one wishes to draw inferences about future production, based upon sample data from current or recent production. Then, if the process is in so-called "statistical control," *and remains so*, the current data may be used to draw inferences about the future performance of the process. The concept of statistical control means, in its simplest form, that the process is stable or unchanging. It implies that the statistical distributions of the characteristics of interest for the current process are identical to these for the process in the future. It also implies that the sequence of data from production is not relevant. Thus, units selected consecutively from production are no more likely to be alike than units selected, say, a day, a week, a month, or even a year, apart. All of this, in turn, means that the only sources of variability are "common cause" within the system, and that variation due to "assignable" or "special" causes, such as differences between raw material lots, operators, and ambient conditions, have been removed.

The concept of statistical control is an ideal state that, in practice, may exist only approximately, although it may often provide a useful working approximation. If a process is in statistical control, then samples from the process are (or can be modeled as) independent and identically distributed. When a process is in statistical control, the statistical intervals provided in this book should yield reasonable inferences about the process. On the other hand, when the process is not in, or near, statistical control, the applicability of the statistical intervals given here for characterizing the process may be undermined by trends, shifts, cycles, and other variations unless they are accounted for in a more comprehensive model.

1.7.3 Other Analytic Studies

Although analytic studies frequently require projecting from the present to a future time period, this is not the only way an analytic study arises. For example, practical constraints, concerns for economy, and a variety of other considerations may lead one to conduct a laboratory scale assessment, rather than perform direct evaluations on a production line, even though production is up and running. In such cases, it is sometimes possible to perform verification studies to compare the results of the sampled process with the process of interest.

1.7.4 How to Proceed

The following operational steps are appropriate for potentially constructing statistical intervals for many analytic studies:

- Have the engineer, scientist or subject-matter expert define the process of interest.
- Determine the possible sources of data that will be useful for making the desired inferences about the process of interest (i.e., define the process to be sampled or evaluated).
- Clearly state the assumptions that are required for the results of the study on the sampled process to be applicable for the process of interest.
- Collect well-targeted data and, to the extent possible, check the assumed model and any other assumptions.
- Jointly decide, in light of the assumptions and the data, and an understanding of the underlying cause mechanism, whether there is value in calculating a statistical interval, or whether this might lead to a false sense of security, and should, therefore, be avoided.
- If it is decided to obtain a statistical interval, ensure that the underlying assumptions are fully recognized and make clear that this interval represents only the uncertainty associated with the random sampling and does not include uncertainties due to differences between the sampled process and the process of interest. Therefore, the actual uncertainty will be greater than that expressed by the width of the interval and in some applications, could be substantially greater.

1.7.5 Planning and Conducting an Analytic Study

In conducting an analytic study, one typically cannot sample directly from the process of interest. This process may, as we have seen, not yet exist. Instead, one needs to define the specific process that is to be sampled and how the sampling should proceed. In so doing, as broad an environment as possible should be considered. For example, in characterizing a production process, one should include the wide spectrum of raw materials and operating conditions that might be encountered in the future. This will require fewer assumptions when using the resulting data to draw inferences about the actual process of interest. It is usually advisable to sample over relatively long time periods because observations taken over a short period are less likely to be representative of the process of interest with regard to both average performance and long-run variability unless the process is in strict statistical control. For example, in studying the properties of a new alloy, specimens produced closely together in time may be more alike than those produced over longer time intervals due to variations in ambient conditions, raw material, operators, machine condition, measuring equipment, etc.

In some analytic studies, one might deliberately make evaluations under extreme conditions. In fact, Deming (1975) asserts that in the early stages of an investigation, “it is nearly always

the best advice to start with strata near the extremes of the spectrum of possible disparity in response, as judged by the expert in the subject matter, even if these strata are rare." He cites an example that involves the comparison of the speed of reaching equilibrium for different types of thermometers. He advocates, in this example, an initial study on two groups of people: those with normal temperature and those with high fever. In addition, it is important that information on relevant concomitant variables is recorded, whenever feasible, for inclusion in, possibly graphical, subsequent analyses. For example, in dealing with a production process, data identifying operator, raw material lot, ambient conditions and other factors that might potentially impact the performance of the process should generally be retained for potential future analyses.

1.8 CONVENIENCE AND JUDGMENT SAMPLES

In practice, it is sometimes difficult, or impossible, even in an enumerative study, to obtain a random sample. Often, it is much more convenient to sample without strict randomization. Consider again a product packaged in boxes whose performance is to be characterized. If the product is ball bearings, it might be easy to thoroughly mix the contents of a box and sample randomly. On the other hand, suppose the product is made up of fragile ceramic plates, stacked in large boxes. In this case, it is much easier to sample from the top of the box than to obtain a random sample from among all of the units in the box. Similarly, if the product is produced in rolls of material, it is often simple to cut a sample from either the beginning or the end of the roll, but often impractical to sample from anyplace else. (This is not a systematic random sample, because there is not a random starting point.) Also, when sampling from a production process, it is often more practical to sample periodically, say every 2 hours during an 8-hour shift, than to select samples at four different randomly selected times during each shift. In this, and other applications, a further justification for periodic sampling is the need to consistently monitor the process for changes by the use of control charts, etc.

Selection of product from the top of a box, from either end of a roll, or at prespecified periodic time intervals for a production process, without a random starting point, are examples of what is sometimes referred to as "convenience sampling." Such samples are generally *not* strictly random because some units (e.g., those not at either end of the roll) have no chance of being selected. Because one is not sampling randomly, statistical intervals, strictly speaking, are not applicable for convenience sampling. In practice, however, one uses experience and understanding of the subject matter to decide on the applicability of applying statistical inferences to the results of convenience sampling. Frequently, one might conclude that the convenience sample will provide data that, for all practical purposes, are as "random" as those obtained by a simple random sample. Sampling from an end of a roll might, for example, yield information equivalent to that from simple random sampling *if* production is continuous, the process is in statistical control, and there is no roll end effect. Similar assumptions apply in drawing conclusions about a process based upon selecting samples from production periodically. Thus, treating a convenience sample as if it were a random sample *may sometimes* be reasonable from a practical point of view. However, the fact that this assumption is being made needs to be recognized, and the validity of using statistical intervals as if a random sample had been selected needs to be critically assessed based upon the specific circumstances.

Similar considerations apply in "judgment" or "pseudo-random" sampling. This occurs when personal judgment is used to choose "representative" sample units; a foreman, for example, might by eyeball take what appears to be a "random" selection of production units, without going through the formalities that we have described for selecting a random sample. In many cases, this might yield results that are essentially equivalent to those obtained from a random sample. Sometimes, however, this procedure will result in a higher probability of selecting, for

example, conforming or nonconforming units. In fact, studies have shown that what might be called "judgment" can lead to seriously biased samples and, therefore, invalid or misleading results. Thus, the use of judgment, in place of random selection of sample units, invalidates the probabilistic basis for statistical inference and could render statistical intervals meaningless.

Judgment is, of course, important in planning studies, but it needs to be applied carefully in the light of available knowledge and practical considerations. Moreover, where possible, judgment should *not* be used as a substitute for random sampling or other randomization needed to make probabilistic inferential statements, such as constructing statistical intervals. Thus, returning to Deming's example of comparing the speed of reaching equilibrium for different type thermometers, it might well be advantageous to make comparisons for strata of people with normal temperature and with high fever. Within these two strata, however, patients and thermometers should be selected at random, to the degree possible. This will provide the opportunity for valid statistical inferences within strata, even though these inferences may be in a severely limited domain.

1.9 SAMPLING PEOPLE

In many important applications, such as public opinion polls, marketing studies and TV program viewing ratings, the subject of the study is not a product, but people from whom we wish to solicit verbal responses. Such studies typically present added issues, including special considerations to ensure that the frame resembles the target population as closely as possible and the added problem of nonrespondents (e.g., individuals who choose not to participate in a study). If these issues are not handled appropriately, they could make formally constructed statistical intervals meaningless or even misleading. We elaborate below.

Telephone surveys are used extensively in people-response studies. However, such studies can result in the frame seriously failing to represent the target population. This was made evident in the 1936 US Presidential election in which telephone surveys predicted the election of Al Landon over Franklin Roosevelt. An important reason for this erroneous prediction was the fact that in 1936 the characteristics—and, most importantly, the voting preferences—of the frame of the then telephone owners differed appreciably from the target population of the voting electorate. Today, surveys of land-line telephone users can similarly miss the mark by excluding the increasing number of households that rely only on cell phones.

Sampling people at shopping malls or other on-site locations might be used instead of, or in addition to, telephone surveys, especially in marketing studies. If, however, one is interested in a population beyond shopping mall visitors, such studies may again lead to a frame that fails to adequately represent the population of interest by tending to exclude the elderly, poor people, wealthy people, and other demographic groups that visit shopping malls infrequently.

Once defined, it is often possible to secure a random sample from the frame. For example, in telephone surveys, random digit dialing is frequently used. However, unlike sampling a manufactured product, the subjects selected for a people study can choose whether or not to respond—and whether or not to provide a truthful response. Moreover, willingness to participate in a TV viewing survey may be correlated with a respondent's viewing preferences. Also, the fact that the person who answers the phone—or even who happens to be at home at the time of the call—is unlikely to be a random family member, can create additional bias.

Alternatives to telephone surveys tend to have similar or other difficulties. For example, response rates on mail surveys tend to be especially low. On-site studies might yield a higher response rate, but, as already suggested, present other challenges. As already suggested, the problem with nonrespondents is that those who choose to respond might differ in characteristics, views or behavior from those who do not. In particular, people who hold strong views on a

subject may be more likely to respond. Special inducements, such as financial compensation, can reduce the number of nonrespondents, but such inducements may also bias the results.

The preceding difficulties are, of course, well known to experts who conduct people-response studies, and various procedures have been developed to mitigate the resulting problems or to compensate for them. Thus, one approach to address the nonresponse problem is to take a follow-up sample of nonrespondents, making a special effort to solicit a response. The results are then compared with those of the initial sample which, if needed, are adjusted accordingly.

Another more general approach that aims to address both nonresponse and inadequacy of the frame is to compare the demographics—especially with regard to variables that are likely to be related to the survey response—of the respondents in the selected sample with those of the population of interest, and then correct for disproportionalities (in a manner similar to the analysis of results from a stratified sample). As a consequence, the results from a particular respondent might be weighted more heavily in the analysis than those of other respondents if the respondent's demographics appear to be underrepresented in the sample.

1.10 INFINITE POPULATION ASSUMPTION

The methods for calculating statistical intervals discussed in this book, except for those in Section 6.3, are based on the assumption that the sampled population or process is infinite, or, at least, very large relative to the sample size. However, in most enumerative studies, the assumption of an infinite population is not met. With a finite population, the sampling itself changes the population available for further sampling by depletion and therefore samples are no longer truly independent. For example, if the population consisted of 1,000 units, selection of the first sampled unit reduces the population available for the second sample to 999 units.

Books on survey sampling show how to use a "finite population correction" to adjust, approximately, for the finite population size. Using such a correction generally results in a narrower statistical interval than one would obtain without such a correction. Thus, ignoring the fact that one is sampling from a finite population usually results in conservative intervals (i.e., intervals that are wider than required for the specified confidence level). In an analytic study, the population or process is conceptually infinite; thus, no finite population correction is needed.

In practice, if the sample size is a small proportion of the population (10% or less is a commonly used figure), ignoring the correction factor will give results that are approximately correct. Thus, in many enumerative studies, it is not unreasonable to assume an infinite population in calculating a statistical interval.

The preceding discussion leads to a closely related, and frequently misunderstood, point concerning sampling from a finite population—namely, that the precision of the results depends principally on the *absolute*, and not the *relative*, sample size. For example, say you want to make a statement about the mean strengths of units produced from two lots from a stable production process. Lot 1 consists of 10,000 units and lot 2 consists of 100 units. Then, a simple random sample of 100 units from lot 1 (a 1% sample) provides more precise information about the mean of lot 1 than a random sample of size 10 from lot 2 (a 10% sample) provides about the mean of lot 2. We will describe methods for sample size selection in Chapters 8, 9, and 10.

1.11 PRACTICAL ASSUMPTIONS: OVERVIEW

In Figure 1.1 we summarize the major points of our discussion in Sections 1.3–1.10 and suggest a possible approach for evaluating the assumptions underlying the calculation of the statistical intervals described in this book. Although it is, of course, not possible to consider all

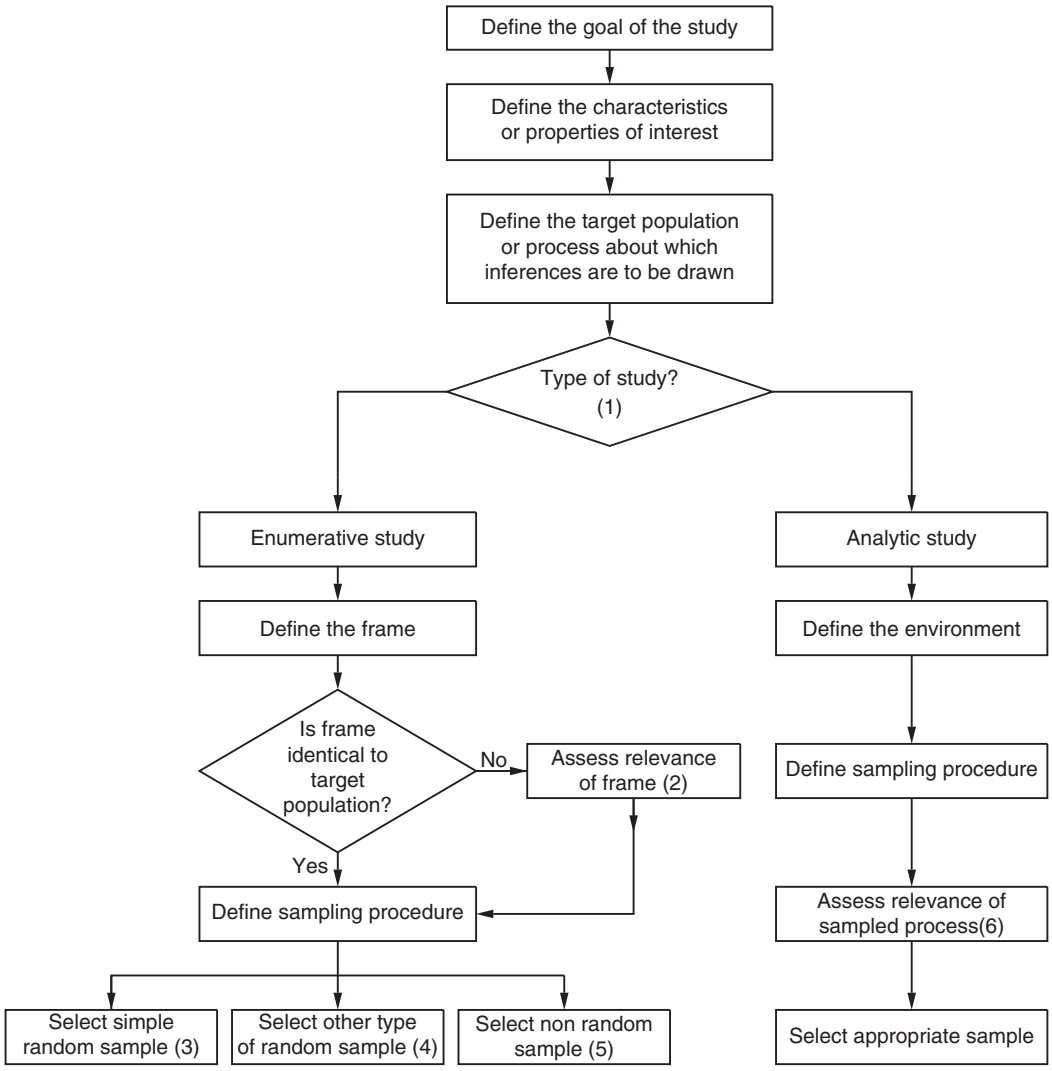

Figure 1.1 Possible approach to evaluating assumptions underlying the calculation of a statistical interval. See the text for explanations of the numbered items.

possible circumstances in such a diagram, we believe that Figure 1.1 provides a useful guide for practitioners on how to proceed for many situations.

The following comments pertain to the numbers shown in parentheses in Figure 1.1:

(1) Is the purpose of the study to draw conclusions about an existing finite population (enumerative study) or is it to act on and/or predict the performance of a (frequently future) process (analytic study)?

(2) Statistical intervals apply to the frame from which the sample is taken. When the frame does not correspond to the target population, inferences about the target population could be biased. A statistical interval quantifies only the sampling uncertainty due to a limited sample size. Actual uncertainty will frequently be larger.

(3) Most statistical intervals in this book assume a simple random sample from the frame.

(4) More complicated statistical intervals than those for simple random samples apply; see Section 1.6.3 and the books on survey sampling referred to in the Bibliographic Notes section at the end of this chapter.

(5) Statistical intervals do not apply. If calculated, they describe uncertainty to the extent that the nonrandom sample provides an approximation to a random sample. Resulting intervals in this case, as in other cases, do not account for uncertainty due to sampling bias.

(6) Statistical intervals apply to the sampled process, and not necessarily to the process of interest. Thus, any statistical interval does not account for uncertainty due to differences between the sampled process and the process of interest.

1.12 PRACTICAL ASSUMPTIONS: FURTHER EXAMPLE

We now cite, taking some liberties, a study (see Semiglazov et al., 1993) conducted for the World Health Organization (WHO) to evaluate the effectiveness of self-examination by women as a means of early detection of breast cancer. The study was conducted on a sample of factory workers in Leningrad (now St. Petersburg) and Moscow. This group of women was presumably selected for such practical reasons as the ready listing of potential participants and the willingness of factory management and workers to cooperate. A major characteristic of interest in this study is the time that self-examination saves in the detection of breast cancer.

Suppose, initially, that the goal was the very limited one of drawing conclusions about breast cancer detection times for female factory workers in Leningrad and Moscow at the time of the study. The frame for this (enumerative) study is the (presumably complete, current, and correct) listing of female factory workers in Leningrad and Moscow. In this case, the frame coincides with the target population, and it may be possible to obtain a simple random sample from this frame. We suppose further that the women selected by the random sample participate in the study and provide correct information and that the sample size is small (i.e., less than 10%), relative to the size of the population. Then the statistical intervals, provided in this book, apply directly for this (very limited) target population. (It is possible in an enumerative study to define the target population so narrowly that it becomes equivalent to the "sample." In that case, one has complete information about the population and, as previously indicated, the confidence intervals presented in this book degenerate to zero width, and the tolerance and prediction intervals become probability intervals.)

Extending our horizons slightly, if we defined the target population to be all women in Moscow and Leningrad at the time of the study, the frame (of female factory workers) is more restrictive than the target population. The statistical uncertainty, as expressed by the appropriate statistical interval, applies only to the sampled population (i.e., the female factory workers), and its relevance to the target population (i.e., all women in Moscow and Leningrad) needs to be assessed.

In actuality, the WHO is likely to be interested in a much wider group of women and a much broader period of time. In fact, the basic purpose of the study likely was that of drawing inferences about the effects of encouraging self-examination for women throughout the world, not only during the period of study, but, say, for the subsequent 25 years. In this case we are, in fact, dealing with an analytic study. In addition to the projection into the future, we need to be concerned with such matters as differences in self-examination learning skills and discipline, alternative ways of detecting breast cancer, the possibility of different manifestations of breast cancer, and many others. The unquantifiable uncertainty involved in translating the results from the sampled population or process (i.e., female factory workers in Moscow and Leningrad at the time of the study) to the (future) population or process of major interest (e.g., all women

throughout the world in the subsequent 25 years) may well be much greater than the quantifiable statistical uncertainty.

Our comments are in no way a criticism of the WHO study, the major purpose of which appears to be that of assessing whether, under a particular set of circumstances and over a particular period of time, self-examination can be beneficial. We cite the study only as one example of an analytic study in which statistical intervals, such as those discussed in this book, describe only part of the total uncertainty, and may, in fact, be of very limited relevance.

Fortunately, not all studies are as global in nature and inference as this one. It seems safe to say, however, that in applications, the simple textbook case of an enumerative study in which the frame is in good agreement with the target population, and in which one has a random sample from this frame, is the exception, rather than the rule. Instead it is more common to encounter situations in which:

- One wishes to draw inferences concerning a process (and, thus, is dealing with an analytic, rather than an enumerative, study).
- One is dealing with an enumerative study, but the frame differs from the target population in important respects, and/or sampling from the frame is not (strictly) random.

As indicated, in each of these cases, we need to be concerned with the implications in generalizing our conclusions beyond what is warranted from statistical theory alone—or, as we have repeatedly stated, the calculated statistical interval generally provides an optimistic quantification of the total uncertainty, reflecting only the sampling variability. Thus, in studies like the WHO breast cancer detection evaluation, the prudent analyst needs to decide whether to calculate statistical intervals at all—and, if so, stress their limitations—or to refrain from calculating such intervals in the belief that they may do more harm than good. In any case, such intervals need to be supplemented by, and often are secondary to, the use of statistical graphics to describe the data—as illustrated in the subsequent chapters.

1.13 PLANNING THE STUDY

A logical conclusion from the preceding discussion is that it is of prime importance to properly plan the study to help assure that:

- The target population or process of interest is well defined initially.
- For an enumerative study, the frame matches the target population as closely as practical and the sampling from this frame is random or as close to random as feasible.
- For an analytic study, the investigation is made as broad as possible so as to reduce the almost inevitable gap between the sampled process and the process of interest and randomization is introduced to the degree feasible.

Unfortunately, studies are not always conducted in this way. Often, analysts are handed the results and asked to analyze the data. This requires *retrospectively* defining the target population or process of interest and the frame or process that was actually sampled, and determining how well the critical assumptions for making statistical inferences apply. This is often a frustrating, or even impossible, task because the necessary information is not always available. In fact, one may sometimes conclude that in light of the deficiencies of the investigation or the lack of knowledge about exactly how the study was conducted, it might be misleading to employ any method of statistical inference.

The moral is clear. If one wishes to perform statistical analyses of the data from a study, including calculation of the intervals described here, it is essential to plan the investigation statistically in the first place. One element of planning the study is determining the required sample size; see Chapters 8, 9, and 10. This technical consideration is, however, often secondary to the more fundamental issues described in this chapter. Further details on planning studies are provided in texts on survey sampling (dealing mainly, but not exclusively, with enumerative studies) and books on experimental design (dealing mainly with analytic studies). See the references in the Bibliographic Notes section at the end of this chapter.

1.14 THE ROLE OF STATISTICAL DISTRIBUTIONS

Many of the statistical intervals described in this book assume a distributional model, such as a normal distribution, for the measured variable, possibly after some transformation of the data. Frequently, the assumed model only approximately represents the population or process, although this approximation is often adequate for the problem at hand. With a sufficiently large sample (say, 40 or more observations), it is usually possible to detect important departures from the assumed model and, if the departure is large, to decide whether there is a need to reject or refine the model. For more pronounced departures, fewer observations are needed for such detection. When there is not enough data to detect important departures from the assumed model, the model's correctness must be justified from an understanding of the physical situation and/or past experience. Such understandings, of course, should enter the assessment, irrespective of the sample size.

Some intervals—notably confidence intervals to include the population mean—are relatively insensitive to the assumed distribution; other intervals strongly depend on this assumption. We will indicate the importance of distributional assumptions in our discussion of specific intervals. Hahn (1971) discusses "How Abnormal is Normality?", and numerous books on statistics, including Hahn and Shapiro (1967, Chapter 8), describe graphical methods and statistical tests to evaluate the assumption of normality. We provide a brief introduction to, and example of, this subject in Section 4.11.

Statistical intervals that do not require any distributional assumptions are described in Chapters 5 and 13 and in several of the case studies in Chapters 11 and 18. Such "nonparametric" intervals have the obvious advantage that they do not require the assumption of a specific distribution. Nonparametric intervals are, therefore, especially appropriate for those situations for which the results are sensitive to assumptions about the underlying distribution. A disadvantage of nonparametric intervals is that they tend to be wider (i.e., less precise) than the corresponding interval under an assumed distributional model. Moreover, frequently there is not sufficient data to compute a nonparametric interval at the desired confidence level. Such intervals also still require the other important assumptions discussed in the preceding sections.

1.15 THE INTERPRETATION OF STATISTICAL INTERVALS

Because statistical intervals are based upon limited sample data that are subject to random sampling variation, they will sometimes not contain the quantity of interest that they were calculated to contain, even when all the necessary assumptions hold. Instead, they can be claimed to be correct only a specified percentage (e.g., 90%, 95%, or 99%) of the time they are calculated, that is, they are correct with a specified "degree of confidence." The percentage of such statistical intervals that contain what one claims they contain is known as the confidence level associated with the interval. The selection of a confidence level is discussed in Section 2.6.

The confidence level, at least from a traditional point of view, is a property *of the procedure* for constructing a particular statistical interval, and not a property of the computed interval itself. Thus, the confidence level is the probability that, in any given study, the random sample will result in an interval that contains what it is claimed to contain. Using this (classical) interpretation, a particular confidence interval to, for example, contain the mean of a population *cannot* correctly be described as being an interval that contains the actual population mean with a specified probability. This is because the mean is an unknown fixed characteristic of the population which, in a given situation, is either contained within the interval or is not (i.e., the probability is either one or zero). All we can say is that in calculating many different confidence intervals to contain population means from different (independent) random samples, the calculated confidence interval will actually contain the actual population mean with a specified probability—known as the *coverage probability*—and, due to the vagaries of chance, will fail to do so the other times. We provide further, more specific, elaboration in Sections 2.2.5, 2.3.6, and 2.4.3.

For many of the best-known statistical interval procedures, because of the simplicity of the model assumptions, (e.g., the procedures given in Chapters 3 and 4 for the normal distribution) the coverage probability of the interval procedure is *exactly* equal to the nominal confidence level that is input to the procedure. Outside this relatively narrow set of circumstances, however, statistical interval procedures have coverage probabilities that are only approximately equal to the nominal confidence level. Indeed, it has been a vigorous area of statistical research to find new and better statistical interval procedures that provide better coverage probability approximations. The results of some of this research are used for the interval procedures presented in Chapters 5–7 and 12–18.

Finally, we note that the philosophy of inferences in constructing Bayesian intervals that we present in Chapters 15, 16, 17, and some of the examples in Chapter 18, differs from the preceding non-Bayesian (sometimes referred to as "frequentist") approach to constructing and evaluating statistical intervals. In particular, Bayesian inference methods require the specification of a joint prior distribution to describe our *prior knowledge* about the values of the model parameters (but the use of the probability distribution in this context does *not* imply that the unknown parameter is random). Using conditional probability operations (known as Bayes' theorem), the prior distribution is combined with the data to generate a joint posterior distribution, representing the updated state of knowledge about the parameters. Based on the joint posterior distribution, it is possible, for example, to generate intervals to contain a specific parameter or a particular function of the parameters with a specified *probability*. As a result, such intervals based on Bayesian methods are often referred to as "credible intervals" and not "confidence intervals" (and we will use such terminology in subsequent chapters). One should, however, keep in mind that for given data the probability in the credible interval statement comes directly from the prior distribution. The actual parameter value (fixed, but unknown) is, again, either contained in the interval or not.

Some make a distinction between "subjective Bayesian" and "objective Bayesian" analyses. At the risk of oversimplifying the distinction, in a subjective Bayesian approach, one uses some combination of previous experience, expert opinion, and other subjective information to choose the joint posterior distribution. The objective Bayesian approach, on the other hand, attempts to specify a prior distribution that uses little or no prior information to set the prior distribution (variously referred to as "default," "reference," "noninformative," "vague," or "diffuse" prior distributions) that might, for example, have the objective of producing a Bayesian procedure with good frequentist properties (e.g., that the coverage probability be close to the nominal confidence level). Indeed, the vast majority of Bayesian analyses in practical applications use the objective approach. In such cases one could argue that the use of the term "confidence interval" is still warranted. In many (if not most) practical problems where an informative prior

distribution is to be used, the analysis will involve a combination of the subjective and the objective approaches because useful prior information may be available for only one of several parameters.

1.16 STATISTICAL INTERVALS AND BIG DATA

Much has been said about the technological changes that have brought us into the "big data" era. Big data is a consequence of the availability and accessibility of enormously large and complicated data sets. "Big" is a relative term, and how big is "big" is often characterized by the volume, variety, and velocity (know as the "three Vs") of a data set. The arrival of big data, as well as our ability to analyze such data and potentially gain useful information therefrom, were made possible by advances in sensor, communications, data storage, and computational technology.

How do statistical intervals pertain to big data? An immediate answer might be that the confidence intervals discussed in this book have little relevance for big data. This is because confidence intervals are used to quantify the uncertainty in estimating some characteristic(s) of interest due to the (random) sampling of a population or process. This uncertainty is most pronounced when one has limited sample data. In dealing with big data there is little, or often essentially no, sampling uncertainty. In addition, as previously indicated (and as will be discussed in more detail in Chapter 2), tolerance and prediction intervals converge to becoming probability intervals describing a distribution as the sample size increases.

Thus, our major concern in dealing with big data needs to be not with statistical variability, but with the "representativeness" of the data with regard to the population of processes of interest, as discussed throughout this chapter.

So why, in this age of big data, should we be concerned with statistical intervals—much less update a book on the subject? Our answer is simple. Even though the era of big data is, indeed, upon us and raises numerous new challenges, there are still many, and perhaps even most, situations in which circumstances limit us to small data. In assessing a rare disease, one might, for example, have only a relatively few recorded cases. In conducting a sample survey, budget limitations typically restrict us to a small sample. In new product reliability assessments, sample availability and cost concerns often impose severe restrictions on the amount of testing that can be conducted. And these are just a few examples. In fact, in most situations in which we generate new data, such as designed experiments or sample surveys, as opposed to analyzing existing data, one is restricted to relatively small samples and, therefore, it is appropriate to use statistical intervals to quantify the associated statistical uncertainty. In summary, even though this surely is the age of big data, let us not forget the continued need for drawing meaningful inferences from small sets of data.

We also note that the ability to do computer-intensive analyses, which have helped bring about the age of big data, has also led to the development of improved methods for constructing statistical intervals for more complicated inference problems dealing mainly with small to medium-size data, as described in Chapters 12–18, thus making our analyses appreciably more powerful.

1.17 COMMENT CONCERNING SUBSEQUENT DISCUSSION

The assumptions that we have emphasized in this chapter apply throughout this book and warrant restatement each time we present an interval or an example. We have decided, however, to relieve the reader from such repetition. Thus, frequently, we limit ourselves to saying that

the resulting interval applies "to the sampled population or process" or more generically, "to the sampled distribution" and often we omit altogether making any such restrictive statement. The reader needs, however, to keep in mind in all applications the underlying assumptions and admonitions stated in this chapter.

BIBLIOGRAPHIC NOTES

General treatment of statistical intervals

Hahn (1970b) describes confidence intervals, tolerance intervals, and prediction intervals for a normal distribution. Both Scheuer (1990) and Vardeman (1992) discuss in detail confidence intervals, prediction intervals, and tolerance intervals that are distribution-free and also ones that depend on the assumption of a normal distribution.

Enumerative and analytic studies

Using different terminology than Deming, the books by Snedecor and Cochran (1967, pages 15–16) and Box et al. (2005, Chapters 1–3) discuss the differences between enumerative and analytical studies. Also, this distinction is explicitly discussed in detail in the book by Gitlow et al. (1989). Some relevant discussion of enumerative and analytical studies appears in Hahn and Meeker (1993), Chatfield (1995, 2002), Hillmer (1996), Wild and Pfannkuch (1999), and MacKay and Oldford (2000). Emphasis on making the distinction between enumerative and analytic studies seems to have waned in recent years—which, in the authors' opinion, is unfortunate.

Books about survey sampling

Statistical intervals can also be constructed for random sampling schemes that are more complicated than simple random sampling. Such intervals generally require special methods for estimating variances and are described in books on survey sampling such as Cochran (1977), Levy and Lemeshow (2008), Groves et al. (2009), Lohr (2010), Lumley (2010), and Little (2014).

Books about experimental design

There are numerous books on the subject of experimental design. These include, for example, Box et al. (2005), Montgomery (2009), Wu and Hamada (2009), Goos and Jones (2011), and Morris (2011).