Assumptions for Statistical Inference

GERALD J. HAHN and WILLIAM Q. MEEKER*

In this article we overview and discuss some of the important practical assumptions underlying statistical inference. What we say, though not new, is stressed insufficiently in teaching statistical methods and applications. We build on the important conceptual difference between enumerative and analytic studies, emphasized by W. Edwards Deming. Our comments, however, go beyond the published views of Deming. We emphasize the assumptions needed for both types of studies and illustrate the concepts with examples.

KEY WORDS: Analytic study; Enumerative study; Sampling; Statistical education; Statistical interval.

1. INTRODUCTION

The important implicit assumptions in making statistical inferences and the resulting practical issues are, in our opinion, not discussed sufficiently in most textbooks on statistics and therefore are often not transmitted to students. Perhaps these issues are dismissed as “common sense” and therefore do not warrant elaboration. It has been our observation, however, that many analysts (both with and without statistical training) often fail to appreciate these assumptions; this can result in misleading or incorrect conclusions. The distinction between enumerative studies and analytic studies that has been emphasized by Deming (1950, 1953, 1975, 1986) is especially important. An understanding of this difference, and the potential consequences of making incorrect assumptions, is essential for correctly planning a statistical study, analyzing the resulting data, and taking action based on the results.

2. STATISTICAL INFERENCE

2.1 The Problem

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

1. 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.

2. A company needs to use data from a sample of five turbines to arrive at a guaranteed efficiency for a further turbine to be delivered to a customer.

3. A manufacturer uses tensile strength (and other) measurements obtained from a laboratory test on 10 samples of each of two types of material to select one of the two materials for future production.

An important part of the problem is to obtain data that are relevant to the decision to be made and that can be gathered under the ever-present constraints on time, money, and availability of sample units. The common textbook statements “Assume a random sample (or a simple random sample) from the population of interest” and “Assume a sample of independently and identically distributed observations from a normal distribution” are overly simplistic. Because these assumptions are so frequently glossed over, analysts tend to ignore them in practice. Unfortunately, these easy-to-state assumptions often provide only a crude approximation to reality, and if not met can result in seriously flawed conclusions.

2.2 Point Estimates and Statistical Intervals to Quantify Uncertainty

The sample data are often summarized by statements such as:

1. 293 out of the 1,000 sampled households were tuned in to the show.

2. The average efficiency for the sample of five turbines was 67.4%.

3. The samples using Material A had an average tensile strength 3.2 units higher than those using Material B.

The preceding “point estimates” provide a concise summary of the sample results, but 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 ±1% of the point estimate 67.4%? Or within ±1%? Or within ±10%?

Various types of statistical intervals may be calculated from the sample data. These intervals help quantify the uncertainty associated with our estimates. Moreover, if our knowledge, as reflected by the length of the un-

*Gerald J. Hahn is Manager, Management Science and Statistics at Corporate Research and Development, General Electric Company, Schenectady, NY 12301. William Q. Meeker is Professor, Department of Statistics, Iowa State University, Ames, IA 50011. The authors thank Tom Boardman, Bob Easterling, Emil Jere, Ramon Leon, Joseph Lu, Jock MacKay, Wayne Nelson, Bill Tucker, Mike Tveite, Steve Wardeman, Jack Wood, Bill Wunderlin, and a referee for their very helpful comments on an earlier version of this manuscript. In addition, they benefited from valuable insights gained from Dr. W. Edwards Deming and numerous participants at the Sixth Annual Deming Seminar for Statisticians, held at New York University in March 1992.
uncertainty interval, is too imprecise, then we may wish to obtain more data before making an important decision. It is, of course, imperative that the interval calculated addresses the problem at hand. In fact, the appropriate interval depends upon the specific application. Frequently used intervals are:

1. A confidence interval for an unknown characteristic of the sampled population or process. For example, based upon a sample of tensile strength measurements, we might wish to construct an interval to contain, with a specified degree of confidence, the mean or standard deviation of the sampled population or process, or to contain a function of such parameters, such as a percentage point (or percentile) of the sampled population or process, or the probability of exceeding a specified threshold value.

2. A statistical tolerance interval to contain a specified proportion of the units from the sampled population or process. For example, based upon a past 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 90% of units from the sampled process.

3. 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 past sample of tensile strength measurements, we might wish to construct an interval to contain, with a specified degree of confidence, the tensile strength of a randomly selected single future unit from the sampled process.

Most users of statistical methods are familiar with confidence intervals for the population mean and for the population standard deviation (but generally not for population percentage points or the probability of exceeding a specified threshold value). Some are also aware of tolerance intervals. However, despite their practical importance, most practitioners, and even many statisticians, know very little about prediction intervals except, perhaps, for their application to regression problems. Thus, 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 texts on statistics generally discuss confidence intervals, occasionally make reference to tolerance intervals, but usually consider prediction intervals only in the context of regression analysis. This is unfortunate because, in applications, tolerance intervals, prediction intervals, and confidence intervals on population percentiles 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 the better known confidence intervals. See Hahn (1970), Scheuer (1990), Vardeman (1992), and Hahn and Meeker (1991) for a description, comparison, and examples of these different statistical intervals.

In the subsequent discussion we emphasize statistical assumptions underlying the calculation of statistical intervals because we feel that such intervals can be especially important in practical applications. Our comments, however, apply to all forms of statistical inference, including point estimates, and not just statistical intervals alone.

3. ASSUMPTION OF SAMPLE DATA

Standard statistical methods for inference (e.g., statistical intervals) express the uncertainty due to sample variability in the (often limited) data. There are, of course, some situations where there is little or no statistical uncertainty. This is the case when the relevant information on every unit in a finite population has been recorded (without measurement, or other, error), or when the sample size is so large that the uncertainty in our estimates due to sampling variability is negligible. Examples of situations where one is generally dealing with the entire population are as follows:

1. There has been 100% inspection (i.e., all units are measured) of a performance property for a critical component used in a spacecraft.

2. A complete inventory of all the parts in a warehouse has been taken.

3. A customer has received a one-time order for five parts and has measured each of these parts. Even though the parts are a sample from a larger population or process, as far as the customer is concerned the five parts make up the entire population of interest.

Another, perhaps less obvious but common, example is where one has field data on the life performance (say, time to first failure for units that failed and running times for those that have not failed) of all units of a product, such as an aircraft engine or a locomotive, and interest centers only on units already in service. In this case, there may, however, still be much uncertainty about the time to failure distribution because most units may not yet have failed (i.e., the data are censored), despite the fact that one has (censored) data for every unit in the population.

For enumerative studies (see later discussion) in which the entire population is sampled, statistical inference methods such as statistical intervals, are unnecessary (and, in fact, inappropriate). Graphical methods and the use of summary statistics and, sometimes, probability statements to describe the population are, however, still useful.

4. CENTRAL ROLE OF PRACTICAL ASSUMPTIONS CONCERNING “REPRESENTATIVE DATA”

When making statistical inferences based on sample data, certain model and sampling assumptions are required. In the following sections, we discuss the major practical assumptions dealing with the “representative-
ness” of the sample data. Departures from these implicit assumptions are common in practice and can invalidate a formal statistical analysis. Ignoring such assumptions can provide a false sense of security, which, in many applications, is the weakest link in the inference chain. Thus, for example, production 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, such assumptions may be so far off the mark that it is inappropriate, and perhaps even misleading, to try to apply formal inference methods.

In the best situations, one can rely on physical understanding, or information from outside the study, to justify the practical assumptions. Such an evaluation is, however, principally the responsibility of the engineer—or “subject-matter expert.” Often, the assessment is far from clear-cut. In any case, one should keep in mind that statistical intervals reflect only the statistical uncertainties. Thus, although in theory this need not always be the case, in practice we have found that statistical intervals often provide a lower bound on the true uncertainty; the generally nonquantifiable deviations of the practical assumptions from reality add an unknown element of uncertainty to that given 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 usually be wider than the statistical interval alone. This observation does, however, lead to a general rationale for calculating a statistical interval for situations where the basic assumptions are questionable. In such cases, if it turns out that the statistical interval is wide, we then know that our estimates have much uncertainty, even if the assumptions were all correct—a point made previously by others, for example, Bert Gunter and Brian Joiner. On the other hand, a narrow statistical interval would imply a small degree of uncertainty only if the required assumptions hold.

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, e.g., Deming 1950). Textbooks in statistics have been slow in giving this distinction the attention it deserves—despite its central role in applications. Notable exceptions include the recent books by Gitlow, Gitlow, Oppenheim, and Oppenheim (1989, chap. 2) and by Moen, Nolan, and Provost (1991, chaps. 3 and 4) and, using different terminology, by Box, Hunter, and Hunter (1978, chaps. 1–3) and Snedecor and Cochran (1967, pp. 15 and 16).

To point out the differences between these two types of studies, we return to the examples of Section 2.1. As indicated, the statements in Section 2.2 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 average tensile strengths of the production units that would 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, exists at the time of sampling. Deming uses the term “enumerative study” to describe such situations. More specifically, Deming (1975, p. 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 (nonoverlapping) and exhaustive sampling units. The frame defines the population to be sampled in an enumerative study.

Further examples of enumerative studies are:

1. Public opinion polls to assess the current view, on some specified topic(s), of the entire adult U.S. population, or some defined segment thereof, such as all registered voters in a specified locality.
2. Sample audits to assess the correctness of last month’s bills so as to estimate the total error in such bills (for accounting purposes or for adjustment of individual bills should the audit suggest the need for reviewing all of the month’s bills). In this case, the population of interest consists of all of last month’s bills.
3. 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.

In an enumerative study, the correctness of statistical inferences requires a random sample from the frame. Such a sample is, at least in theory, often attainable; see Section 7.2.

In contrast, the second two examples of Section 2.1 (dealing, respectively, with the efficiency of a future turbine and the comparison of the future performance of two manufacturing processes) illustrate what Deming calls “analytic studies.” We no longer have an existing, finite, well-defined, unchanging population. Instead, we want to take actions (generally to improve), or make predictions about, the output of a, sometimes hypothetical future, process. We can, however, obtain data only from the existing (likely different) process.
Specifically, Deming (1975, p. 147) 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 centres in future product, not in the materials studied." He cited 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)." Similarly, we may wish to use data from an existing process to predict 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.

We take the liberty, for the sake of clarity in this exposition, to extend Deming's definition of an analytic study to encompass all situations that do not involve an enumerative study. That is, we define an analytic study to be a study in which one is not dealing with a finite, identifiable, unchanging collection of units, and, thus is concerned with a process, rather than a population.

We note that, at least in our experience, the great majority of applications encountered in practice, especially in industrial and in medical and other scientific applications, involve analytic, rather than enumerative studies. Moreover, it is inherently more complex to draw conclusions from analytic than from enumerative studies; 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.

Sometimes the differentiation between an analytic and an enumerative study is not clear-cut. In such cases, a "simple criterion to distinguish between enumerative and analytic studies" provided by Deming (1975) is useful. In particular, Deming stated that "A 100 per cent sample of the frame provides the complete answer to the question posed for an enumerative study, subject, of course, to the limitations of the method of investigation. In contrast, a 100 per cent sample . . . is still inconclusive in an analytic problem." For example, an "exit poll" to estimate the proportion of voters who have voted (or, at least, would 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). However, 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 as a result of information obtained in the study. Also adverse weather conditions on election day (not contemplated on the sunny day on which the survey was conducted) might deter some from 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 taken a 100% sample of eligible voters prior to the election, we still would not be able to predict the outcome of the election with certainty.

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 inventory, the study is enumerative; in fact, enumerative studies have sometimes literally been referred to as applications of "warehouse statistics." If, however, we wish to predict future performance of the product, perhaps after making changes, the study is analytic.

We shall return to the subject of differentiating between analytic and enumerative studies and give some further examples in Section 10.

6. STATISTICAL INFERENCE FOR ANALYTIC STUDIES

We differ from some published views of Deming's followers (e.g., p. 558 of Gitlow et al. 1989) who imply that statistical inference methods, such as statistical intervals, have no place whatsoever in analytic studies. Indeed, such methods have been used successfully in statistical studies in science and industry for decades and many of these studies have been analytic. Indeed, we feel that the decision of whether or not to use formal statistical inference methods, such as statistical intervals, need be made on a case-by-case basis and with great care. In addition, when such methods are used for analytic studies, they require assumptions different from those required for an enumerative study.

We will now consider, in further detail, the basic assumptions underlying inferences from enumerative studies and then comment on the assumptions made in analytic studies.

7. BASIC ASSUMPTIONS FOR INFERENCE FROM ENUMERATIVE STUDIES

7.1 Definition and Comparison of Target Population and Frame

In enumerative studies there is some "target population" about which it is desired to draw inferences. An important first step—although 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 any other defined time interval. 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 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," where exactly what constitutes such conditions also needs to be clearly stated.

As indicated in the Deming quote, the next step is that of establishing a frame from which the sample is to be taken, that is, obtaining or developing a specific listing, or other enumeration, of the population from which the samples will be selected. Examples of frames may be the serial numbers of all the automobile engines built over the specified time period, the listing of names in a telephone directory for a community, 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 telephone directory generally lists households, rather than individuals, and omits those who do not have a telephone, people with unlisted phones, and new arrivals in the community—and also may include businesses, which are not always clearly identified as such. If the telephone company wishes to estimate the proportion of telephones in working order at a given time, a complete listing (available to the phone company) will probably coincide with the target population. However, for most other studies, there may be an important difference between the frame (i.e., the telephone directory listing) and the target population.

The listing of all sample units provided by the frame can be thought of 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 the differences could have on the conclusions of the study. Moreover, it needs to be stated emphatically that these differences introduce uncertainties above and beyond those generally quantified by standard statistical intervals.

7.2 Assumption of a Random Sample

The data are assumed to be a random sample from the frame. Simple random sampling gives every possible sample of \(n\) units from the frame the same probability of being selected. There are also other random sampling methods beyond simple random sampling, such as stratified sampling, cluster sampling, and systematic sampling. Cochran (1977), Scheaffer, Mendenhall, and Ott (1979), Sukhatme, Sukhatme, Sukhatme, and Asok (1984), and Williams (1978) describe such other random sampling schemes.

The assumption of random sampling is critical. This is because the statistical intervals reflect only the variability due to the random sampling process and do not take into consideration biases that might be introduced by a nonrandom sample. It is especially important to recognize this limitation because in many studies one does not have a strictly random sample; see Section 9 for further discussion.

7.3 Other Statistical Assumptions

There are also a variety of other assumptions that are sometimes made in the analysis of specific enumerative studies, for example, the assumption of a normal distribution. These assumptions (although not always their practical importance) are discussed in standard textbooks.

8. ADDITIONAL ASPECTS OF ANALYTIC STUDIES

8.1 Characteristics of 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 usually 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 often does not even exist at the time of the study. As a result, the process that is sampled is likely to differ in various ways, including the distribution of the performance characteristic(s) of interest, from the one about which it is desired to draw inferences. As we have indicated, sampling prototype units produced in the lab or on a pilot line to draw conclusions about subsequent full-scale production is one common example of an analytic study.

We note also that in an analytic study, because one is no longer dealing with "an aggregate of identifiable units," there is no relevant frame from which one can take a random sample. Instead, one is dealing with a set of observations taken, in a sometimes "representative" manner, from an existing process. (In fact, as we emphasize in Section 8.5, the proper planning of an analytic study should be of major concern to statisticians.) However, the procedure for obtaining these observations may well involve some elements of random selection. For example, in building prototype units, one may be using raw materials randomly selected from last month's production. Similarly, in conducting a taste test on a new pastry product that is under evaluation, one may randomly select shoppers in a supermarket. However, despite these random elements, the basic units under evaluation—the prototype parts and the pieces of pastry in our examples—are generally not randomly selected from a frame.

The use of statistical inference methods implies a random sampling scheme. Statisticians sometimes get around the dilemma posed by taking observations from a process in analytic studies by suggesting that one is sampling at random from a "conceptual population." In any case, the relevance to the process of interest of the specific method for obtaining observations from such a "conceptual population" needs to be carefully assessed in evaluating the appropriateness of using statistical inference methods.
8.2 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 on 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 those for the process in the future. It also suggests that sequence of 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. Data from a process in strict statistical control correspond to the commonly (and sometimes misused) statistical assumption of “independently and identically distributed random variables.”

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. When a process can be assumed to be in statistical control, an analytic study might yield reasonable inferences. On the other hand, when the process is not in, or near, statistical control with respect to all characteristics of relevance, the applicability of statistical intervals, or other methods of statistical inference for characterizing the process, may be undermined by trends, shifts, cycles, and other variations. Finally, we observe that past data can be used to assess whether the process appears to have been in statistical control. However, the fact that a process has been in statistical control in the past provides no guarantee that it will continue to be so in the future.

8.3 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, production constraints, concerns for economy, and a variety of other considerations may lead one to conduct a laboratory-scale assessment, rather than perform direct on-line evaluations, 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.

8.4 Reiteration of the Basic Assumption for Analytic Studies

Obviously, any statistical inferences that we may make from an analytic study can apply only to the specific process that has been sampled and the specific manner in which the observations from that process were obtained. Thus, in the examples mentioned in Section 8.1, statistical inferences are limited to the “conceptual population” of the many other prototype parts, or pieces of pastry, that could, at least in theory, have been produced and tested at the same time and in the same manner as those in our study. Statistical inferences about the actually sampled process may, on occasion, be of some usefulness per se (at least to provide a lower bound on total uncertainty). However, in the overwhelming majority of cases, the investigator’s prime concern is not with the sampled process but with a different one. Any extrapolation of the inferences from the sampled process to some other process is generally beyond our area of competence. The validity of such an extrapolation needs to be assessed by the subject-matter expert. Under these circumstances, determining whether or not to use statistical inference methods requires a judgment call. If such methods are to be used, it is, moreover, the analyst’s responsibility to clearly point out their limitations.

8.5 Planning and Conducting an Analytic Study

In conducting an analytic study, because the process of real interest may not be available for sampling, one often has the opportunity and, indeed, the responsibility of defining the specific process that is to be sampled and how such sampling should proceed. As Deming (1975) and others (e.g., Gitlow et al. 1989; Moen et al. 1991) emphasized, in conducting analytic studies, one should generally aim to consider as broad an environment as possible. For example, one should include the wide spectrum of raw materials and operating conditions that might be encountered in the future. This is contrary to what is traditionally recommended for scientific investigations, namely to hold constant all variables except those that are key to the study itself. The reason for making the study sufficiently broad is to narrow the gap between the sampled process and the process of interest. This is especially pertinent if there are possible interactions between the factor(s) under evaluation in the study and the “background conditions.”

In some analytic studies, one might deliberately make evaluations under extreme conditions. In fact, Deming (1975) suggested 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” (in their occurrence in practice). He cited 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. Moreover, whenever possible, data on concomitant variables, such as operators and raw material, should be obtained (in time order), along with the response(s) of primary interest, for possible (graphical) analysis. One may think of this as setting up blocks or strata of interest and randomly sampling from the extreme strata.
When sampling over time, it is usually advisable to sample over relatively long periods, because observations taken over a short period of time are less likely to be representative of the process of interest (with regard to both average performance and long-run variability) than those obtained over a longer time period (unless the process is in strict statistical control). For example, in a study of the properties of a new alloy, specimens produced closely together in time may be more alike than those produced by the process in the long run due to variations in ambient conditions, raw material, operator, the condition of the machines, and the measuring equipment, and it is, of course, long-run performance in which we are generally most interested. This concept of “robust analytic studies” bears some resemblance to the concept of robust product development proposed by Genichi Taguchi, the eminent Japanese quality assurance practitioner; see Phadke (1989) or the August 1988 special issue of the International Journal of Quality and Reliability Engineering for further information on such methods.

9. 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, for example, 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 a roll, but often impractical to sample from any place else. For a production line, it is often more practical to sample material periodically, say every two hours during an eight-hour shift, than to select material at four randomly selected times. In this case, the results may also be used for process monitoring using control charts, for which periodic sampling is often actually desirable.

Selection of product from the top of the 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; one reason is that 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 inferences, strictly speaking, are not applicable for convenience sampling, even in an enumerative study. 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. This might, for example, be the case, even though samples were selected only from the top of the box, if the items were thoroughly mixed (deliberately or otherwise) before they were put into the box. Also, sampling from an end of a roll might yield information equivalent to that from random sampling if production is continuous, the process is in statistical control, and there is no “roll end effect.” Similar considerations hold in drawing conclusions about a production process from periodic samples. Our point is that treating a convenience sample as if it were a random sample in an enumerative study 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 making statistical inferences as if a random sample had been selected needs to be critically assessed—together with the other considerations discussed previously—based upon the specific circumstances.

Similar considerations apply in “judgment” or “pseudorandom” sampling. This occurs when personal judgment is used to choose “representative” sample units. For example, a foreman may, by eyeballing, take what appears to be a “random” selection of production units, without going through the necessary formalities for selecting a random sample. In many cases, this might yield results that are essentially equivalent to those of a random sample. Sometimes, however, this procedure will, either deliberately or nondeliberately, 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 actually lead, even unintentionally, to seriously biased samples and, therefore, invalid or misleading results. Thus, strictly speaking, the use of such judgment in place of random selection of sample units in enumerative studies invalidates the probabilistic basis for statistical inference and could render the results meaningless.

In analytic studies, on the other hand, one generally exercises some degree of judgment in sample selection, as illustrated by Deming’s thermometer example. Thus, in selecting production units in an analytic study, it might, under certain circumstances, be useful for the foreman to deliberately select “extreme” conforming and nonconforming units for scrutiny. Nevertheless, as indicated in Section 8.1, random sampling also plays a role in analytic studies. Thus, in the thermometer example, one would select test subjects at random within strata, and thermometers at random (possibly, also from strata) from among those available. See Deming (1975, 1976) for additional discussion.

10. A POSSIBLE APPROACH AND SOME FURTHER EXAMPLES

10.1 Steps in Evaluating Assumptions

In Figure 1 we suggest a possible approach for evaluating the assumptions underlying the calculation of
statistical intervals. A key step, of course, is the determination of whether the study is enumerative or analytic—a distinction that, as we have indicated, is not always clear-cut. However, the major issue is not whether the study is enumerative or analytic, but the applicability of statistical inference methods. The enumerative versus analytic study dichotomy just provides a useful paradigm for making this judgment.

We will now present, in some detail, further practical examples that illustrate the process shown in Figure 1. These reemphasize the importance of a full understanding of the goals of the study and of the available data (or, one hopes, the opportunities for obtaining appropriate data).

10.2 Light Bulb Life Characterization Example

A consumer or a consumer protection agency has purchased a package of six randomly selected light bulbs at a supermarket and is concerned with characterizing light bulb life. We consider three different cases.

Case 1. One bulb is selected at random out of the pack for future use in a critical application. The remaining five bulbs are life tested. From the data on the five tested bulbs, it is desired to predict the lifetime of the untested bulb. Moreover, all six bulbs are used in the same environment and “lifetime” is measured in the same way. We claim that in this investigation we are dealing with an enumerative study, consisting of a frame of six units, of which a random sample of five has been selected. Thus, methods described by Hahn and Meeker (1991) (both for a normal distribution and distribution-free) to obtain a prediction interval for a future observation apply. If, however, the test environment for the sixth bulb is different from that for the other five bulbs, one would be making a prediction to the new environment, resulting in an analytic study,
and the relevance of calculating a formal prediction interval would be questionable.

**Case 2.** Assume now that all six bulbs have been life tested, and that the goal is for a consumer protection agency to use the results to characterize average bulb life for “recent production,” under specified operating conditions (voltage, ambient temperature, and cycling frequency). Investigation suggests that it is reasonable to assume that the bulbs in the package were built some time during a three-month period, ending two weeks before the time at which the bulbs were purchased. Moreover, the agency agrees that this three-month period appropriately represents “recent production.”

In this case, we are again dealing with an enumerative study—the target population being a well-specified, finite group, that is, all such bulbs built during the three-month period. However, the frame for this study is, at best, ambiguous. One might, in fact, conclude that there was no frame, and, therefore, no way of assessing its relevance or of selecting a random sample. In this case, the calculation of a statistical interval seems meaningless (except, perhaps, as a coarse lower bound on the total uncertainty). If, on the other hand, the package had been selected randomly from all such packages in the store, one might claim that there is a frame—consisting of all such bulbs in the store—and the selected package is a cluster sample, with a single cluster of six observations. We would now be concerned both about how well this frame represents the population of interest (all bulbs made over the three-month period or all bulbs in stores) and the fact that there seems to be no way to estimate among-cluster variability. This would likely lead us to the conclusion that a statistical interval would again be meaningless, unless one could further assume that the production process is strictly in statistical control, as defined in Section 8.2.

Of course, most of the preceding “contortions” could have been avoided if a knowledgeable statistician had been involved in planning the study—and had been listened to. In that case, for example, all such bulbs currently in stores in the United States might have been selected as the frame. Then a cluster sample involving a random selection of stores and a random selection of bulbs (or packages of bulbs) within stores might have been recommended. This would have allowed a (correctly calculated) confidence interval to properly reflect the uncertainty in average bulb life for the population of interest.

A further complication arises if the results of the study were to be used to make statements about average bulb life in a “typical” consumer’s home. We would then be dealing with an analytic study, rather than an enumerative study, because the operating environment in the consumer’s home would, undoubtedly, differ from that in the study. Any statistical inferences would, of course, be limited to the test conditions used in the study, and might not be relevant for the real process of interest.

**Case 3.** Consider again the same situation as for Case 2, but now assume that the major interest is that of characterizing next year’s production. In this case, we are clearly concerned with the output of a future process and thus are dealing with an analytic study. Any statistical inferences would be limited to the process that was sampled, which is now different from the future process of interest.

### 10.3 Breast Cancer Self-Detection Study

We now cite (taking some liberties) a study conducted by the World Health Organization (WHO) to evaluate the effectiveness of self-examination for early detection of breast cancer. The study was conducted on a sample of female factory workers in Moscow and St. Petersburg. We assume that these women were selected for such practical reasons as the ready listing of potential participants and the willingness of factory management and workers to cooperate. We assume, for the purpose of discussion, that a major characteristic of interest is the time that self-examination saves in the detection of breast cancer.

Assume, initially, that the goal is the very narrow one of drawing conclusions (about breast cancer detection times) for female factory workers in St. Petersburg 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 St. Petersburg 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 assume further that the women selected by the random sample participate in the study and provide correct information. Then statistical inference methods, like statistical intervals, apply directly for this (very narrow) target population. (It is of course 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; as previously indicated, statistical inference is then appropriate.)

Actually, the World Health Organization 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 could well be 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 next 25 years. In this case, we are, in fact, dealing with an analytic study. In addition to the projection into the future, we need be concerned with such matters as the equivalence of learning skills and discipline, alternative ways of detecting breast cancer, and the possibility of different manifestations of breast cancer. In practice, the unquantifiable uncertainty involved in translating the results from the sampled population or process (female factory workers in Moscow and St. Petersburg today) to the (future) population or process of interest (e.g., all women in the next 25 years) may well be much greater than the quantifiable statistical uncertainty.
We hasten to add that our comments are in no way a criticism of the WHO study, the major purpose of which appears to be that of assessing, under a particular set of circumstances, and over a particular period of time, whether self-examination can be beneficial. We cite the study only as one example of an analytic study in which statistical intervals, or other methods of statistical inference, describe only part of the total uncertainty, and may, in fact, be of limited, if any, relevance.

10.4 Design of Experiments

Most designed experiments are analytic. One uses specially selected observations on a current process to improve, or make predictions about, a future process. A typical example is that of an experiment to assess the impact of various processing variables (e.g., temperature, pressure, catalyst addition, raw material batch) on the viscosity of a chemical. We would normally use the results to assess the (future) impact of the experimental variables on viscosity, and, perhaps, to determine appropriate “operating conditions” for the (future) operation of the process. Moreover, the scale and general environment in which the experiment is conducted is often quite different from those that are likely to be encountered in operations.

Such an experiment, then, certainly has all the hallmarks of an analytic study, and one might argue that the application of statistical inference to the experimental results is inappropriate. Our view, however, is that many experimenters will find it useful to quantify the statistical uncertainty with regard to the current process, using statistical intervals—even though we would insist that this limitation be clearly stressed in the reporting of the findings and that any statistical inferences be used to supplement, and not to replace, graphical summaries (which, of course, also reflect the current process).

10.5 Summation

It seems safe to say 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 that have one or more of the following properties:

1. One wishes to draw inferences concerning a process (and is dealing with an analytic, rather than an enumerative, study).
2. 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, one need be concerned with the implications of generalizing one’s conclusions beyond what is warranted from statistical theory alone—or, as we have repeatedly stated, the calculated statistical interval generally provides an optimistic lower bound on the total uncertainty, reflecting only the sampling variability.

Thus the prudent analyst will decide whether to calculate statistical intervals and stress the limitations of the resulting inferences, or to refrain from calculating such intervals under the belief that they may do more harm than good. In any case, these intervals may be secondary to the use of well-chosen statistical graphics to describe the data. Of course, similar limitations also apply to graphical evaluations and point estimates. These summaries apply to the frame or sampled process, rather than the population or (future) process of interest. However, such graphical assessments or point estimates are not accompanied by supposedly precise statements about quantified uncertainty, and, therefore, do not carry the same degree of, possibly misplaced, authority.

11. PLANNING THE STUDY

A logical, but hardly surprising, conclusion from the preceding discussion is that it is of prime importance to properly plan the study. Such planning is suggested by Figure 1, and by our discussion of enumerative and analytic studies. Planning the study helps assure that

1. The purpose of the study and the target population or process of interest are well defined initially.
2. For an enumerative study, the frame matches the target population as closely as practical, and the sampling is random (or as close to random as feasible).
3. 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, where appropriate).

Unfortunately, studies are not always conducted in this way. In some cases, in fact, the statistician is handed the results of a study and asked to analyze the data. This requires retrospectively defining both 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 highly frustrating, or even impossible, task and 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, it is imperative to plan the investigation statistically in the first place. One element of planning the study is determining an appropriate sample size. This technical consideration is, however, often secondary to the more fundamental issues described in this article. Further details on planning studies are provided in texts on survey sampling (dealing mainly, but not exclusively, with enumerative studies) and on experimental design (dealing mainly with analytic studies).
12. CONCLUSIONS

In this article, we have reviewed Deming’s differentiation between analytic and enumerative studies, and have noted that, in our experience and in that of others, analytic studies are far more prevalent than enumerative studies. In analytic studies, one is observing a current process, but, usually, is interested in taking actions on, or making predictions about, a future process. In enumerative studies, one is sampling (it is hoped in some random manner) from a frame, but often wishes to use the results to draw conclusions about a target population that generally differs from the frame.

Statistical inferences are limited to the current process and the frame, and any extrapolations need be justified by subject-matter expertise. In analytic studies especially, this often makes the use of statistical methods inappropriate, or even misleading (other than possibly as a lower bound on total uncertainty). However, there are also analytic studies, such as some involving designed experiments, in which we feel statistical inference methods can play a useful role—even though their applicability is limited to the current process. In addition there are enumerative studies for which, due to the “unrepresentativeness” of the frame or the inadequacy of the sampling, one needs to be very wary of using statistical inference.

In all studies when inferential methods are used, the statistician needs to make sure that the limitations are fully understood and considered. Moreover, irrespective of the type of study, the statistician’s most important role is the up-front contribution of planning the investigation—even though the planning process is different for analytic and enumerative studies. Then, given appropriate data, the results can often be well summarized graphically, and the discussion of the role of formal statistical inference may become irrelevant.

In conclusion, we again note that statisticians are naturally inclined to emphasize the formal aspects of inferential methodology, such as using efficient or unbiased estimators. The basic underlying assumptions, discussed in this article, are well known, but in the past have often been inadequately communicated. We urge a greater recognition of this highly important “soft side” of statistics in our training of students and in our support of clients.

[Received January 1991. Revised March 1992.]

REFERENCES


The American Statistician, February 1993, Vol. 47, No. 1