Assumptions for Statistical Inference

Gerald J. Hahn
Corporate Research and Development
General Electric Company
Schenectady, NY 12301

William Q. Meeker
Dept. of Statistics
Iowa State University
Ames, IA 50011

March 14, 1992

Abstract

In this paper 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 and, on occasion, depart from those of some of his followers. We emphasize that the assumptions needed for valid inferences from an analytic study are fundamentally different from those needed in an enumerative study and illustrate the basic ideas 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" which 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) is especially important. An understanding of this difference is essential for correctly planning a statistical study, analyzing the resulting data, and for 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:

- A television network uses the results of a sample of 1000 households to decide whether or not to continue a show.
- 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.
- A manufacturer uses tensile strength 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 is both relevant to the decision to be made and that can be gathered under the ever-present constraints on time, money, availability of sample units, etc. The common textbook statements "Assume a random sample (or a simple random sample) from the population of interest" or "Assume a sample of independently and identically distributed observations from a normal distribution" is 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:

- 293 out of the 1000 sampled households were tuned in to the show.
- The average efficiency for the sample of five turbines was 67.4%.
- 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 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? From 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 $\pm 10\%$?

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 uncertainty interval, is too imprecise, then we may wish to obtain more data before making an important decision. The appropriate interval depends upon the specific application. Frequently used intervals are:

- A *confidence interval* for an unknown characteristic of the 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 mean or standard deviation of the population or process, or to contain a function of such parameters, such as a percentage point (or percentile) of the population or process, or the probability of exceeding a specified threshold value.
- 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 the units from the sampled process.

- 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, often 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 most texts on statistics generally discuss confidence intervals, occasionally make reference to tolerance intervals, but generally 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 tolerance intervals and prediction intervals are generally no more difficult than those for confidence intervals. See Hahn (1970), Scheuer (1990), Vardeman (1990), or Hahn and Meeker (1991) for a description, comparison, and examples of these different statistical intervals.

In our subsequent discussion we will be emphasizing statistical assumptions underlying the calculation of statistical intervals because these are of particular interest to us. Our comments, however, apply to all forms of statistical inference, and not just statistical intervals alone.

3 The Assumption of Sample Data

Standard statistical methods for inference (e.g., statistical intervals) express uncertainty that is present because of sample variations 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:

- 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).
- 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 for 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 at hand 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 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 still appropriate in such cases.

4 The Central Role of the Practical Assumptions Concerning "Representative Data"

When making 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 "representativeness" of the sample

data. Departures from these implicit assumptions are common in practice and can invalidate the entire 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 would be inappropriate, and, perhaps even misleading, to try to apply formal inference methods.

In the best of situations, one can rely on physical understanding, or information from outside the study, to justify the practical assumptions. Such assessment 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 often statistical intervals provide a *lower bound* on the true uncertainty; the, generally non-quantifiable deviations of the practical assumptions from reality provide an added *unknown* element of uncertainty in addition 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* uncertainly, 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 long, we then know that our estimates have much uncertainty—even *if* the assumptions were all correct. On the other hand, a narrow statistical interval would imply a small degree of uncertainty *only if* the required assumptions hold, a point made perviously by others (e.g., Bert Gunter).

5 Enumerative Versus Analytic Studies

Deming (1953, 1975) emphasizes the important differences between "enumerative" and "analytic" studies (a concept that he briefly introduced earlier, e.g., Deming 1950). Despite its central role in making inferences from the sampled data, textbooks in statistics have been slow in giving this dis-

tinction the attention that it deserves. Some exceptions include the recent book by Gitlow et al. 1989 (Chapter 2) and, using different terminology, Box, Hunter and Hunter 1978 (Chapters 1 to 3) and Snedecor and Cochran 1967 (pages 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:

- The proportion of households in the *entire country* that were tuned to the show,
- The efficiency of the, *as yet not manufactured*, turbine to be sent to the customer,
- A comparison of the average tensile strengths of the *production units to be built* in the factory during the *forthcoming year* using Material A and Material B.

In the first example, our interest centers on a finite identifiable 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) 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.

Some further examples of enumerative studies are:

- Public opinion polls; in this case, the population of interest might be the entire adult U.S. population, or some defined segment thereof, such as all registered voters,
- Sample audits to assess the correctness of last month's bills; in this case, the population of interest consists of all of last month's bills,

- Product acceptance sampling; in this case, the population of interest consists of all units in the production lot being sampled.

In an enumerative study, one wishes to draw conclusions about an existing well-defined *population* that is being sampled directly. In such a study, the correctness of statistical inferences requires a random sample from the target population. Such a sample is, at least in theory, generally attainable in such a study; 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 two manufacturing processes next year) illustrate what Deming calls "analytic studies." We no longer have an existing, well-defined, finite population. Instead, we want to draw inferences, or make predictions, about a future *process*. We can, however, obtain data only from the existing (most likely somewhat 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 centres 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)." 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 part, the process (or conceptual population) of interest consists of parts of that type that may be manufactured in the future.

These examples are representative of many—indeed, the great majority, in our experience of—applications encountered in practice, especially in engineering, medical, and other scientific investigations. Moreover, it is inherently more complex to draw inferences from analytic studies than from enumerative studies; analytic studies require the important (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.

What one wishes to do with the results of the study is 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 still to be improved, or otherwise to be acted upon, perhaps as a result

of the study, then we are clearly dealing with an analytic study. Deming (1975) presents the following "simple criterion to distinguish between enumerative and analytic studies. 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." This is because for an analytic study our real interest is in a process that is not available for sampling. Thus, a 100% sample of the process that is available could miss the mark.

Frequently, the differentiation between an analytic and an enumerative study does not seem clear cut. In such cases, Deming's rule may be applied. For example, a public opinion "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 is 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. Also adverse weather conditions on election day (not contemplated on the sunny day on which the survey was conducted) might deter many from going to the polls and the "stay-at-homes" are, therefore, likely to 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, because we do not know who will actually vote and who will change their minds in the intervening period. (Special consideration in sampling people will be discussed later.)

Taking another example, it is sometimes necessary to sample from inventory to make inferences about a product population or process. If interest focuses on the inventory, the study is enumerative; in fact, enumerative studies have sometimes literally been referred to as applications of "warehouse statistics." If, however, interest focuses on the future output from the process, the study is analytic. Finally, drawing conclusions about the performance of a turbine to be manufactured in the future, based upon data on five turbines built in the past, involves, as we have indicated, an analytic study. On the other hand, if the five measured turbines *and* the further turbine to be shipped were independently and randomly selected from in-

ventory (unlikely to be the case in practice), one would be dealing with an enumerative study.

6 Statistical Inference For Analytic Studies

We differ from some published views of Deming's followers (e.g. page 558 of Gitlow, Gitlow, Oppenheim, and Oppenheim 1989) who imply that statistical inference methods, such as statistical intervals, have no place whatsoever in analytic studies. (We hasten to note, however that Deming (1986) only says that "application to analytic problems ... [is] ... unfortunately, however, in many textbooks deceptive and misleading"—a statement with which we completely agree.) Indeed, such methods have been used successfully in statistical studies in science and industry for decades and most of these studies have been analytic. Instead, 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 (see subsequent discussion). 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 Enumerative Studies

7.1 Definition of Target and Sampled Population

In every enumerative study 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 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 a "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, i.e., developing a specific listing, or other enumeration of the population from which the sample can subsequently be selected. Examples of such frames may be the serial numbers of all the automobile engines built over the specified time period, the complete listing 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, new arrivals in the community, etc.—and also may include businesses, which are not always clearly identified as such. If one were trying to conduct an evaluation of the proportion of listed telephones in working order at a given time, a complete listing of telephones (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 sampling frame (i.e., the telephone directory listing) and the target population of real interest.

The listing provided by the frame will henceforth be referred to as the "sampled population." Clearly, the inferences from a study such as those expressed by statistical intervals, will be on this sampled population, and—when the two differ—not on the target population. Thus, our third step—after defining the target population and the sampled population—is that of evaluating the differences between the two, and the possible consequences of the difference on the results 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 The Assumption of a Random Sample

The data are assumed to be a random sample from the sampled population. Simple random sampling gives every possible sample of n units from the population the same probability of being selected. A simple random sample 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, such as those given in existing tabulations of such numbers (e.g., Rand 1955) and generated by computer algorithms [e.g., IMSL 1987 and Kennedy and Gentle 1980) or statistical computing software (e.g., MINITAB), provide easier ways of obtaining a random sample.

There are also other random sampling methods for enumerative studies beyond simple random sampling, such as stratified sampling, cluster sampling, and systematic sampling. Cochran (1977), Scheaffer, Mendenhall, and Ott (1979), Sukhatme et al. (1989), 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 randomness introduced by 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 subsequent 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, e.g., the assumption of a normal distribution. These assumptions (although not always their practical importance) are discussed in standard textbooks; we will, therefore, not comment further on them here.

8 Additional Aspects of Analytic Studies

8.1 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 usually be enumerated, at least conceptually—even though, as we have seen, difficulties can arise in obtaining a random sample from the target population. In contrast in an analytic study one wishes to draw conclusions about a process which often does not even exist at the time of the study. As a result, the process that is sampled may differ, in various ways, 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 and very obvious example of an analytic study.

8.2 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 valid 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 distribution of the characteristics of interest for the current process are identical to those for the process in the future. It also suggests that units selected consecutively from production are no more likely to be alike than the 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, ambient conditions, etc, have been removed. The concept of statistical control is an ideal state that, in practice, may exist only infrequently, though it may often provide a useful working approximation. When a process is in statistical control, an analytic study might yield reasonable inferences about the process of real interest. 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.

8.3 Random Samples in Analytic Studies

In applying statistical inference methods in an analytic study, one assumes that observations are independently and identically distributed. This assumption is, sometimes, referred to as having a "random sample," and can be achieved most readily for a process in "statistical control." For analytic studies of processes that can change over time, the timing of the selection of sample units becomes critical.

8.4 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 or pilot line 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 pilot process with the actual production process.

8.5 How to Proceed

The following steps provide appropriate guidelines for many analytic studies:

- Have the engineer or subject matter expert describe the process of interest.
- Determine existing or new sources of data for making the desired inferences about the process of interest, i.e., define the process to be sampled or evaluated.
- Have the engineer or subject matter expert clearly state the assumptions that are required for the results on the sampled process to be translatable to the process of interest.
- Collect well-targeted data (see Section 12) and, to the extent possible, check the model and 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 conducting statistical inferences, such as calculating a statistical interval, or whether this might lead to a false sense of security, and should, therefore, be avoided (sometimes, one may be able to use the data to check the validity of the assumptions, by, for example, comparison with some known population characteristic),
- If it is decided to obtain a statistical interval, ensure that the underlying assumptions are fully recognized and make clear that the length of this interval generally represents only a lower bound on the total uncertainty. That is, it deals only with the uncertainty associated with the random sampling—and does not include uncertainties due to

the differences between the sampled process and the process of real interest.

8.6 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. As Deming (1975) and others (e.g., Gitlow, Gitlow, Oppenheim and Oppenheim 1989, Moen, Nolan and Provost 1991) emphasize, 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, operating conditions, etc. that might be encountered in the future. This is contrary to what one might do in some scientific investigations for which it may be desirable to hold constant all variables except those that are key to the study itself. The reason for making the study sufficiently broad is that one has to make fewer assumptions in subsequently using the results of the sampled process to draw inferences about the process of interest. Moreover, in 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 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, the measuring equipment, etc. and it is, of course, long run performance in which we are generally most interested.

In some analytic studies, one might deliberately make evaluations under extreme conditions. In fact, Deming (1975) believes 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 cites an example that involves the comparison of "speed of reaching equilibrium" for different types of thermometers. He advocates, in this example, to perform 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, raw material, etc. should be obtained (in time order), along with

the response(s) of primary interest, for possible (graphical) analysis.

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 8-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, etc. for which periodic sampling is 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; for example 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. 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 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 "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 *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 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. For example, a foreman may, by eyeball, 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 non-deliberately, 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 invalidates the probabilistic basis for statistical inference methods, and could render the results 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, such judgment should *not* be used as a substitute for the random sampling or other randomization needed to make probabilistic inferential statements. 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. However, within these two strata, we should select patients at random. Also, thermometers of each type should be randomly selected (possibly within strata) to the greatest degree possible. This provides the opportunity for valid statistical inferences, even though these inferences may be in a severely limited domain. Deming (1975, 1976) provides additional discussion on the use of judgment in analytic studies.

10 Sampling People

Additional considerations frequently arise in sampling human populations, such as in a public opinion poll or a television rating survey. In this case, in contrast to sampling a manufactured product, the subject selected for the study generally chooses whether or not to respond, and whether or not to provide a truthful response. As a result, response rates for mail surveys, without special inducements, have been found to be extremely low, and the results correspondingly unreliable. (Although special inducements might increase the sample size, they might also lead to a less "representative" sample.) "On site" surveys, such as at shopping malls, are likely to result in higher response rates, but such haphazard sampling results in a nonrandom selection from the population of interest by tending to exclude the very elderly, poor people, wealthy people, and others who rarely visit shopping malls.

Telephone surveys might obtain a "more random" selection of households, but, require respondents to be home and might result in a biased selection of certain family members. In TV rating surveys, one is interested in determining the viewing habits of, say, the entire viewing audience. However, a particular survey generally provides information for only those individuals or households who can be induced, perhaps, in part, by financial rewards, to participate in the survey.

Non-response, and related problems, clearly defeat the goal of strict random sampling, and, thus, again compromise the use of statistical intervals that are calculated under the assumption of such a sample. Thus, if no adjustments are made, non-respondents will introduce biases into the evaluations if, as is often the case, the respondents and the non-respondents tend to have different views. For example, willingness to participate in a TV rating survey is likely to be correlated with programming preferences. These difficulties are well known to experts who conduct such surveys, and various procedures have been developed to mitigate them or to compensate for them. These include:

- A follow-up sample from among the initial non-respondents, and a comparison of the results with those from the initial respondents,

- Comparison between respondents and non-respondents with regard to "demographics," or other variables, that are likely to be related to the response variable.

11 Practical Assumptions: Summary and Further Example

In Figure 1 we summarize the major points of this paper, and suggest a possible approach for evaluating the assumptions underlying the calculation of statistical intervals.

To illustrate this approach, we 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 St.Petersburg and Moscow. We assume that this group was 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 methods are then inappropriate, but descriptive statistics and statistical graphics could still be used to describe the population.)

Extending our horizons only slightly, if we defined the target population to be all women in Moscow and St. Petersburg at the time of the study, the frame (of female factory workers) is more restrictive than the target population. As we have indicated, the statistical uncertainty, quantified, for example, by a confidence interval, applies only to the sampled population (i.e., the female factory workers), and, in this case, its relevance to the target population needs to be assessed.

In actuality, 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, not only is the frame highly restrictive, but 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, the possibility of different manifestations of breast cancer, etc. In practice, the unquantifiable uncertainty involved in translating the results from the sampled population or process (female factory workers in Moscow and St.Petersurg today) to the (conceptual) population of major interest (e.g., all women in the next 25 years) may well be much greater than the quantifiable statistical uncertainty.

If it were decided that the population of interest were all the women in the world *today*, one could debate whether this is an analytic study (requiring inferences about a population different from the one being sampled) or an enumerative study (for which the sampled population is a subset of the target population). Such discussion would, however, be somewhat academic—because, irrespective, of one's preferences in terminology, the bottom line is still the need to assess the relevance of the results of the sample to the much larger population or process of actual interest.

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, self-examination can be beneficial. It is, moreover, highly important to have a careful statistical plan for such a study (see subsequent discussion). We only use this study as an example of a case 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.

Fortunately, not all studies are as global in nature and inference as this one. However, it seems safe to say that, in applications, the simple case of an enumerative study in which one is randomly sampling from the target population, is the exception, rather than the rule. Instead it is more common to encounter situations that have one or more of the following properties:

- One wishes to draw inferences concerning a future process (and, thus, is dealing with an analytic, rather than an enumerative, study),

- The sampled population differs from the target population,

- The sampling 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 needs to 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 statistical graphics to describe the data.)

12 Planning the Study

A logical conclusion from the preceding discussion is that it is of prime importance to properly plan the study, if at all possible. Such planning is suggested by Figure 1, and by our discussion in Section 5 dealing with analytic studies. Planning the study helps assure that

- The target population or process of interest is defined initially,
- The frame or sampling scheme is established to match the target population or process as closely as practical,
- A sampling procedure which is as close to random as feasible is used.

Unfortunately, studies are not always conducted in this way. In some cases, in fact, the analyst 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 population or process that was actually sampled, determining how well these match, and deciding whether, and, to what degree, it is reasonable to assume that the sample was randomly selected from the target population or process. This is often a highly frustrating, or even impossible, job, 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 the sampling method, 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 most important 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, usually secondary to the more fundamental issues described in this paper. 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 the analytic studies of processes, etc.)

13 Conclusions

Statisticians are naturally inclined to emphasize the formal aspects of inference methodology, such as using efficient or unbiased estimators. The basic underlying assumptions, as discussed in this article, are, of course, well-known, but are often not properly communicated. This results in wrong or misleading conclusions. We urge a greater recognition of these basic assumptions by statisticians—especially in their dealings with students and clients.

14 Acknowledgments

We would like to thank Tom Boardman, Bob Easterling, Emil Jebe, Wayne Nelson, Bill Tucker, Mike Tveite, Steve Vardeman, Jack Wood, Bill Wunderlin, and a referee for their very helpful comments on an earlier version of this manuscript.

References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters*, New York, NY: John Wiley and Sons, Inc.

Cochran, W. G. (1977), *Sampling Techniques*, (Third Edition) New York: John Wiley and Sons, Inc.

Deming, W. E. (1950), *Some Theory of Sampling*, New York: John Wiley and Sons.

Deming, W. E. (1953), On the distinction between enumerative and analytic surveys, *Journal of the American Statistical Association* **48**, 244-255.

Deming, W. E. (1975), On probability as a basis for action, *The American Statistician* **29**, 146-152.

Deming, W. E. (1976), On the use of judgment samples, *Reports of Statistical Applications, Japanese Union of Scientists and Engineers* **23**, 25-31.

Deming, W. E. (1986), *Out of Crisis*, Cambridge, MA:MIT Center for Advanced Engineering Study.

Gitlow, H., Gitlow, S., Oppenheim, A., and Oppenheim, R. (1989), *Tools and Methods for the Improvement of Quality*, Homewood, IL: Irwin.

Hahn, G. J. (1970), Statistical intervals for a normal population. Part I. Tables, examples and applications. *Journal of Quality Technology* **2**, 115-125.

Hahn, G.J. and Meeker, W.Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York, NY: Wiley

IMSL (1987), *IMSL STAT/LIBRARY User's Manual* (Version 1.0), Houston, TX: IMSL, Inc.

Kennedy, W. J., and Gentle, J. E. (1980), *Statistical Computing*, New York: Marcel Dekker Inc.

Moen, R., Nolan, T. W. and Provost, L.P. (1991), *Improving Quality through Experimentation*, New York: McGraw-Hill.

Rand Corporation (1955), *A Million Random Digits with 100,000 Normal Deviates*, New York: The Free Press.

Scheaffer, R. L., Mendenhall, W., and Ott, L. (1979), *Elementary Survey Sampling*, (Second Edition), North Scituate, MA: Duxbury Press.

Scheuer, E.M., (1990), *Let's teach more about prediction*, Proceedings of the ASA Section on Statistical Education, Washington, DC: American Statistical Association.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984), *Sample Surveys with Applications*, (Third Edition), Ames, IA: Iowa State University Press.

Vardeman, S. (1990), *What about the other intervals?* Preprint 90-29, Department of Statistics, Iowa State University Ames, Iowa.

Williams, B. (1978), *A Sampler on Sampling*, New York: Wiley-Interscience, Inc.