THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 9, November 1982
Printed in U.S.A.

Planning and Conducting Field-Tracking Studies

By S. J. AMSTER, G. G. BRUSH, and B. SAPERSTEIN

(Manuscript received April 5, 1982)

This paper reviews the major aspects of planning and conducting field-tracking studies, including: (i) establishing well-defined, realistic objectives; (ii) designing data collection and analysis procedures to meet the objectives; and (iii) ensuring the successful implementation of these procedures. The paper gives general guidelines on matching study objectives and procedures, as well as detailed information on sample size selection for some common field-study situations. Several studies recently conducted by Bell Laboratories Quality Assurance Center are used to illustrate the principles of field-study planning and implementation.

I. INTRODUCTION

It is the function of Bell Laboratories Quality Assurance Center (QAC) to provide assurance that telecommunication products purchased by the Bell Operating Companies (BOCS) are of satisfactory quality and perform as required. This assurance is provided through the three primary activities of the Quality Assurance effort:

(*i*) Quality inspection and auditing at manufacturing, repair, and installation locations.

(*ii*) Qualitative feedback gathered through informal contacts with BOC personnel and a more formal engineering complaint procedure.

(*iii*) Quantitative field-tracking studies of selected products and systems.

This paper discusses the third activity from both a historical and tutorial point of view. The authors relate some lessons and principles learned through field-tracking studies in the past and offer suggestions for those planning to conduct a field-tracking study (FTS) in the future.

Formal field-tracking studies were undertaken during the 1960s. The studies that will be described in this paper began in 1973 with Product Performance Surveys (PPS)[1] on Western Electric station sets. PPSs are designed to track field performance of the sets, identify problems quickly, quantify the extent of those problems so that economic corrective action can be taken, and assure that the "fixes" are effective. Typically, PPSs on station sets are conducted concurrently in five or six BOC locations chosen to provide geographic and climatic diversity and good representation of a variety of set types. This permits approximately one million station sets to be tracked at any given time, and provides approximately 100,000 trouble events for recording and analysis each year.

PPS data on station sets have been instrumental in detecting and quantifying numerous field problems. Representative examples include a series of contact contamination problems in *Touch-Tone** dials, ringer failures in certain premium station sets, and lamp failures in key telephone sets.

The success of PPS has stimulated an increased effort into field studies of other products, such as PBX's, switching networks, channel bank equipment, switching machines—just about the entire range of telecommunications products purchased by the BOCs. Recently, this field-study effort has been extended to include selected general trade products manufactured by suppliers other than Western Electric. The remaining sections of this paper discuss principles learned by the authors in the process of conducting field-tracking studies and offer suggestions for those planning to conduct an FTS.

Section II discusses important considerations in planning an FTS; Section III discusses key steps in an FTS implementation program; Section IV is devoted to some illustrations from recent Quality Assurance Center studies.

II. PLANNING A FIELD-TRACKING STUDY

The principal steps involved in planning a successful FTS are:

(*i*) Defining study objectives
(*ii*) Planning data collection to meet those objectives
(*iii*) Planning for successful data analysis.

2.1 Defining study objectives

Perhaps the single most important requirement for a successful FTS is a clear statement of purpose that has been agreed to by the concerned parties. A study will frequently have an impact on many different organizations through its implementation, interpretation, and the use of its results. The designer, the manufacturer, and the user all have legitimate concerns in a given FTS. Obtaining their understanding and agreement is an important, but not necessarily a simple, task.

* Registered service mark of AT&T.

Early in the planning of a study, small changes can easily be made to accommodate the needs of potential users. But care must be taken not to try to answer all questions with a single study. Setting precise objectives that simplify implementation can avoid many pitfalls. For example, taking all the data that are easily accessible may initially seem reasonable, since we certainly don't want to miss anything that might be important. But, trying to ensure that "too many" pieces and types of data are good invariably leads to a degraded level of data quality. The topic of data collection is discussed in detail in Section 2.2.2.

Frequently, objectives change as data are collected. This implies the need to provide for such changes initially and to monitor the flow of data to determine when such changes are appropriate. For example, a study that has the objective of comparing the performance of products from three suppliers may quickly show that one supplier is an obvious noncontender. Rules for dropping such a candidate could result in a more efficient use of resources.

Objectives can be classified[2] as:

(*i*) Detecting problems
(*ii*) Quantifying known problems
(*iii*) Verifying quality audit information or reliability predictions
(*iv*) Establishing problem causes
(*v*) Measuring the impact of design or manufacturing change(s)
(*vi*) Evaluating the product.

A study can involve aspects of several of these, but procedures must be matched to purposes. For example, some studies are intended primarily to find and make a preliminary evaluation of problems. Once a problem has been identified, a more detailed study can be used to better quantify its economic impact.

Early thinking about a proposed study may be clarified by the following list of objectives, stated in a statistical framework:

(*i*) Point estimation (e.g., early failure rate)
(*ii*) Interval estimation (e.g., confidence or prediction intervals)
(*iii*) Comparisons (within study, with a standard or with results from a previous study)
(*iv*) Model testing (e.g., decreasing failure rate)
(*v*) Other information (previous list).

Failure to get agreement on specific objectives among all participants can easily lead to continuing disagreements regarding the implementation of the study and the interpretation of its results.

2.2 Planning data collection

Once the general objectives of a field study have been established, the work aimed at meeting those objectives begins with the planning of appropriate data collection procedures.

Most of this planning is aimed at answering the following questions: (*i*) *What* data will be collected? (*ii*) *How* will the data be collected? (*iii*) In what *study population* will the data be collected? and (*iv*) *How much* data (sample size) will be collected? Finding the appropriate answer to each of these questions for any given study is the key to its success. It is worthwhile examining each question separately and describing some of the answers that have been found appropriate in previous studies.

2.2.1 *What data will be collected?*

There are clearly many factors that will determine what data should be collected for any given field study. For purposes of this discussion, we assume that the study in question is directed at estimating the frequency of troubles occurring in a specified product population. This objective imposes the following minimum requirements on the data to be collected:

(*i*) The data must include the size of the study population.

(*ii*) The data must record or count every trouble "event" occurring in the study population during a specified time period, and must exclude or specifically identify events that are reported but occur outside the study population or specified time period.

Clearly, a field study satisfying only these minimum requirements will yield merely gross trouble rate information. However, there are a number of situations appropriate for such a minimal study.

First, for a larger, more detailed study, a preliminary estimate of the overall trouble rate is sometimes needed to determine the study population size. This topic will be further considered below, in the discussion on sample size (Section 2.2.4). Minimal data collection will usually suffice for such an estimate. Minimal data collection might also be appropriate after a detailed study to monitor the effectiveness of corrective actions that may have been taken in response to information obtained during the larger study.

A minimal program of data collection may also be justified in cases where the need for a larger, more detailed and more costly study must be demonstrated. Several tracking studies that we have conducted were operated in this way, with minimal trouble rate data collected until a need for more detailed information was indicated by observing higher than expected trouble rates.

For most field-tracking studies, however, minimal data collection falls short of what is needed in two important ways. First, since this approach provides no identification of the subpopulation in which any trouble occurs, it cannot yield specific trouble-rate estimates by subpopulations. Subpopulation, here, refers to a newly manufactured versus a repaired product, or to different manufacturing vintages of a given product that may reside within a single overall study population. Second, because this approach provides no information on the nature of each trouble event, it cannot yield estimates of the frequency with which the product under study fails for specific reasons.

Information on subpopulations and trouble types makes up virtually all of the detailed data that must be collected for any study; and determining the level of detail for each is a principal objective of study planning.

As noted, subpopulation data would ordinarily include information on whether a piece of equipment in which a trouble occurred was newly manufactured or repaired, the date of manufacture or repair (vintage), service life, and additional descriptive information on the product, such as the issue or series number for a product that has undergone changes in design or manufacture. (Specifying series or issue numbers for circuit packs is an example of detailed product specifications used in tracking studies that are currently under way). Included, too, under the general heading of subpopulation information would be data on how or by whom the trouble was reported, e.g., customers or employees.

In almost all FTS situations the more detailed the data asked for, the more complicated and costly the collection process will have to be. Therefore, it is important to limit to the extent possible the level of detail in subpopulation data requested. The guiding principle in choosing which characteristics should be included in data collection is straightforward: Include only characteristics for which it will be both useful and worthwhile to obtain separate subpopulation trouble-rate estimates when all the data have been collected. Since almost any level of detail can be viewed as potentially useful, the key is to choose only those characteristics that produce "partitions" that will be worthwhile, i.e., that will yield subpopulations of sufficient size to permit making accurate trouble-rate estimates and comparisons. In other words, do not waste time and money partitioning the trouble data into subpopulations so small that the individual data are insufficient to yield accurate and, therefore, useful results.

In many studies it is important to determine precisely when in the life of the equipment each trouble occurs. In those cases deciding when the lifetime of a product starts (so-called "zero time") is of crucial importance. This is particularly true when early life failure rates are to be estimated. For example, does lifetime begin when units arrive, are inspected, are installed, or first operated? Dead-on-arrivals may show up as defective initially or later in time, depending on the type of failure, its effect on the system, the extent of failure detection, and the procedure for collecting the data.

Electronic hardware frequently exhibits a decreasing failure rate during its early life. Here, failures tend to occur closer together during the early weeks of operation. Therefore, depending upon the "zero time" definition, much of the study's most useful data can be lost or misclassified. Particular care is required in defining zero time if units enter the study at different times, are turned on and off for testing, or are moved to different locations.

To relate a real-life incident, one of the authors was recently asked to analyze some data from a study where the objective was failure-rate estimation after six months of operation. But the records gave only the date of installation and failure. Plotting failures against time gave very strange results, solely because these units were turned on only intermittently and no record of actual operating time on each unit was available. In this case the ability to analyze important time-related failure characteristics was lost because of insufficient detail in the data collected.

Detailed data on the "nature" of troubles occurring during any study generally fall in one of two categories. The first category includes a description of the trouble symptoms, the particular portion or component of equipment in which the trouble was observed, and results of any detailed failure mode analyses performed on the failed components. The second category of detailed trouble information includes data on the particular circumstances or environmental conditions associated with any trouble. Whether equipment was observed to be initially defective or to fail in-service and usage conditions are examples of this second category. Below, we have listed some of the detailed items that may be included on the nature of subpopulations:

(*i*) Product vintage (date of manufacture or repair)
(*ii*) Source (new, repair, etc.)
(*iii*) Length in service
(*iv*) Issue, series number, or other product code identifiers.

Like the subpopulation information, the level of detail required on the nature of troubles can have a profound effect on the data collection process, including who will be involved in that process. We have listed the trouble types as follows:

(*i*) Component or equipment subcode
(*ii*) Trouble symptoms
(*iii*) Repair analysis results
(*iv*) Component failure mode analysis results
(*v*) Precise time of failure.

Obtaining data on failure-mode analyses, for example, may require the participation of technical organizations not directly involved in the field tracking itself. This, in turn, imposes additional requirements on the flow of hardware and paper (trouble tickets, analysis results, etc.) for a given study. At the end of this section we will illustrate some of these ideas with examples from recently conducted tracking studies. Now, we turn to a closer examination of the question, "How will the data be collected?"

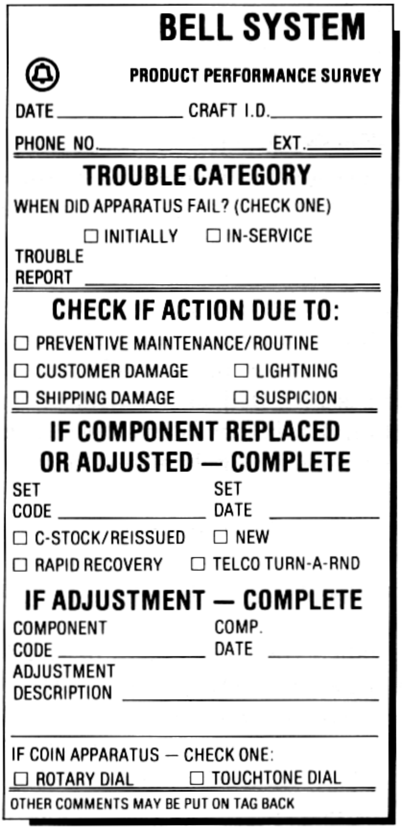
2.2.2 How will the data be collected?

There are as many answers to this question as there are products to be studied. Our aim in this paper, therefore, is to identify goals and procedures common to all or most field-tracking situations.

Probably the best way to start this discussion is the same way it is best to start planning a data collection process—by identifying existing procedures for recording, collecting, and storing information on the field performance of the product under study. It is a rare product on which no information is recorded in the field or at a repair center. Planning data collection should ideally be viewed as a process of either supplementing or tailoring existing data sources to suit the needs of a particular FTS.

At this point it would be helpful to distinguish between data collection carried out in the field (i.e., where the product under study is used), and that carried out in repair locations, and to discuss each separately.

In most tracking studies, the collection of field failure data involves the use of a trouble ticket that must be completed by people responsible for maintaining the equipment under study. As noted, completion of existing trouble tickets is frequently a part of the regular maintenance routine, and substitution of a more detailed study ticket, or "piggybacking" of the study ticket on an existing form, is preferable to burdening maintenance people with a new and separate piece of paper. Whether or not a separate or modified existing form is used, there are a number of basic rules that govern the design of trouble tickets. First, the tickets should be kept as short and as simple as possible. Those are the obvious rules. Less obvious, but equally important, are the following: Wherever possible, the trouble tickets should be formatted in "modular" fashion, with separate sections devoted to different types of information—e.g., time and place of the trouble in one section, equipment description in another, trouble description in still another. The most frequently used modules should appear first and most prominently; less frequently used modules should appear later. The trouble ticket used in the station set Product Performance Survey (PPS) (Fig. 1) illustrates these ideas. The top of the ticket gives information on when and where a trouble occurred. That information is required for each trouble report. Next comes information on the nature of the trouble, also needed for each event. Data on the type of set or component involved in the trouble come next; however, these data are not needed if the equipment in question is returned with the



BELL SYSTEM
PRODUCT PERFORMANCE SURVEY
DATE ____________ CRAFT I.D. ____________
PHONE NO. ____________ EXT. ______

TROUBLE CATEGORY
WHEN DID APPARATUS FAIL? (CHECK ONE)
☐ INITIALLY ☐ IN-SERVICE
TROUBLE REPORT ____________

CHECK IF ACTION DUE TO:
☐ PREVENTIVE MAINTENANCE/ROUTINE
☐ CUSTOMER DAMAGE ☐ LIGHTNING
☐ SHIPPING DAMAGE ☐ SUSPICION

IF COMPONENT REPLACED OR ADJUSTED — COMPLETE
SET CODE ____________ SET DATE ____________
☐ C-STOCK/REISSUED ☐ NEW
☐ RAPID RECOVERY ☐ TELCO TURN-A-RND

IF ADJUSTMENT — COMPLETE
COMPONENT CODE ____________ COMP. DATE ____________
ADJUSTMENT DESCRIPTION ____________

IF COIN APPARATUS — CHECK ONE:
☐ ROTARY DIAL ☐ TOUCHTONE DIAL
OTHER COMMENTS MAY BE PUT ON TAG BACK

Fig. 1—Station set Product Performance Survey trouble ticket.

trouble ticket. Finally, the last section of the ticket describes field adjustments, used only in those few cases where no hardware is returned along with the ticket.

As this last discussion of the station-set PPS implies, there is more to field data collection than the gathering of trouble tickets; there is frequently the gathering of failed hardware as well. The design of an effective, integrated hardware/trouble ticket data-flow system is as important as the design of the trouble ticket itself. The basic objectives of the data-flow system are:

(*i*) To ensure that each piece of returned hardware reaches the designated repair or diagnostic location and, in many cases, the designated individual responsible for hardware analyses in the study; and

(*ii*) To ensure that the information on the trouble tickets reaches the organization responsible for storing and analyzing the trouble data.

There are other important objectives, as well, primarily related to assuring compliance with study procedures and ensuring that hardware analysis results may be uniquely identified with reported trouble events. We will discuss the issue of compliance later. The ability to associate hardware analysis results with trouble symptom reporting is important in tracing down the causes of No Trouble Found (NTF) returns (e.g., diagnostics problems). The use of serialized, multipart tickets is the prime vehicle for making such associations and will be illustrated below.

We have already noted that the burden imposed by an FTS on field personnel can be minimized by using existing reporting forms, whenever possible. For some products, the burden can be even further reduced by exploiting automatic data collection procedures. We include in this category fully automatic data collection, such as that associated with accessing maintenance channel output of software-controlled equipment, and semiautomatic data collection, such as that associated with accessing computerized administrative data on customer trouble reports where the initial entry of the data into the data base depends on action by customers or field personnel. Access of existing data sources such as these has become an increasingly prominent mode of data collection in field-tracking studies. Access of repair location data bases serves an analogous function for hardware-repair analysis data.

2.2.3 In what study population will the data be collected?

In choosing the study population it is important to explicitly define the limits of the inferences to be made from the study. Are the results to be applied to all units, all units made in a given period or under given conditions, or used in a particular fashion, etc.? If the members of the study population received special care, were hand-made, produced at one plant, etc., then conclusions beyond these boundaries depend upon engineering judgment more than upon statistical inference. Confidence intervals reflect variability only in the population actually sampled and not from other sources. For example, increasing the sample taken in one operating area gives no information regarding inter-area differences. When sampling is performed by first selecting K operating areas and then sampling only within these, the formulas appropriate are those used in cluster sampling.[3] Here, the intra-area and inter-area variability are separated. Of course, looking at inter-area differences in detail can indicate important variables (maintenance procedures, environmental impact, etc.) that could be the focus of a follow-up study. Care must be taken before cause and effect relationships are assumed because of the multitudes of possible causes and interrelationships. As Cox relates:[4]

"If we wish to apply the conclusions to new conditions or units, some additional uncertainty is involved over and above the uncertainty measured by the standard error. The only exception . . . is when the units . . . are chosen from a well-defined population of units by a proper sampling procedure."

And later,

"... it is important to recognize explicitly what are the restrictions on the conclusions of any particular experiment."

In any tracking study there is a trade-off between more detailed conclusions regarding a smaller population and less detailed conclusions about a larger one. For example, a study may be aimed at determining whether a change in design has improved reliability in systems subject to certain load characteristics, or whether an overall reliability increase independent of load has occurred. A careful statement of objectives will greatly assist resolving such questions.

Once a population of interest has been defined and agreed upon, technical sampling questions can be addressed. There are certain population characteristics that require special attention. For example, if a small proportion of the units contribute a large proportion of the events under study, stratification and other specialized techniques may be required. Also, considerable gains in efficiency can sometimes be realized by the use of ratio or regression estimates. Here, known characteristics of products or systems under study are related to the characteristics of interest in the study.

2.2.4 How much data will be collected: sample size considerations

Selecting the appropriate number of units to be included in an FTS is very important. On the one hand, a sample size that is too large may add unnecessary expense to the study. On the other hand, a sample size that is too small may mean that any statistical test using study data may lack sufficient power to draw meaningful conclusions. Several authors[5,6,7] have addressed this problem. Reference 5 took the theory of Refs. 7 and 8 and transformed it into usable curves; these curves will be discussed in general in this section and in detail in the appendix.

The parameters of interest in a field study are summarized in Table I. In cases A and D a sample size will be chosen to control the precision of the estimates within certain bounds. In the remaining cases

Table I—Parameters of interest in a field study

	Proportion	Rate
Estimating one parameter	Case A	Case D
Testing hypothesis about one parameter	Case B	Case E
Comparing two parameters	Case C	Case F

the sample size will be chosen to control the probability of making incorrect conclusions. If we assume that failures associated with a proportion occur according to a binomial model and that failures associated with a rate occur according to a Poisson model, it is possible to develop excellent sample sizing guidelines for each of the cases A through F. (A discussion of model selection and use is included in the next section.) Each case is discussed in detail, with examples, in the appendix.

2.3 Planning for successful data analysis

In this section, we consider both the data analysis, itself, and the data storage and retrieval procedures that make the analysis possible.

2.3.1 Model building and data analysis

It requires no lengthy argument to establish that the payoff from any field study comes only with the successful analysis of the data from that study. And in a very real sense, all of the detailed planning on data collection is aimed at ensuring that at the conclusion of the study it will be possible to carry out all of the data analyses appropriate to the study objectives.

In broad terms, there are three things that generally get done with field-tracking data. These are:

(*i*) Estimating trouble or replacement rates, including the construction of confidence intervals, where appropriate and practical;

(*ii*) Searching the data for anomalies—equipment types or vintages that stand out, or trouble causes that stand out; and

(*iii*) Making comparisons of product performance among different types, or vintages, of equipment.

Each of these procedures requires careful planning and a close linkage between the setting of objectives, the design of the data collection process, and the data analysis itself.

During both planning and implementation of a study, the mechanism by which the study objectives, the actual data collection, and the data analysis are linked is the statistical "data model." It is through the data model that the nondeterministic (stochastic) nature of the data is described, and through the model that statistical inferences on the questions of interest to the study are made.

As noted above, most field-tracking studies concern themselves with counts of events (failures, replacements, etc.). It is for this reason that the simplest and most frequently used models in field studies are the binomial and Poisson models.

The binomial model relates the number of events of interest (failures, say), X, to the total number of "trials" (opportunities for failure), N, through the expression:

$$\text{Probability } [X = k] = \frac{N!}{k!(N-k)} p^k(1-p)^{N-k}, \qquad k = 0, 1, \cdots, N,$$

where p is the probability of a failure on an individual trial.

The Poisson model relates the number of events of interest, X, to the total amount of time during which those events can have occurred, t, through the expression:

$$\text{Probability } [X = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \qquad k = 0, 1, \cdots,$$

where λ is the rate at which the events occur in time.

Both models assume a uniform probability or intensity of occurrences—from trial-to-trial for the binomial, over time for the Poisson. For studies in which a model allowing for changing failure intensity seems appropriate (e.g., studies of equipment that may be subject to infant mortality), other models such as the Weibull and lognormal are commonly employed. Detailed information on the form and use of these models may be found in any one of several statistical/reliability texts[9] and we will not attempt to describe them here.

None of the models mentioned thus far is equipped to handle data collected under changing study conditions (e.g., changing environment, age, study locations, etc.), or so-called "nuisance factors."

To illustrate the problem of nuisance factors, suppose we wanted to compare the replacement of two types of equipment (called "old" and "new"), from a study in which the "old" equipment was observed, in one study location, while the "new" equipment was observed in that and other study locations. Here, the factor of interest is equipment type (old versus new); the nuisance factor is the difference that may exist between study locations, which could bias the comparison between the old and new equipment. It is at this point that the use of relatively sophisticated data-analytic techniques, employing tools such as the well-known linear (or log-linear) model, becomes necessary and worthwhile. These techniques allow for separating the effects (on replacement rates, for example) caused by equipment differences, study location differences, etc. and for getting at the factors of interest without ignoring potential biases introduced by the presence of nuisance factors. The use of linear models is well documented in both the statistical and engineering literatures.[10,11] However, when confronted with an apparent need to make use of such techniques, the study designer and data analyst should seek the assistance of a statistician who is thoroughly familiar with the application of these techniques.

The use of any of the models mentioned above involves making some assumptions about the data. For example, as noted, use of the binomial or Poisson models assumes a uniform failure probability or intensity. Use of a linear model generally involves some assumptions of independence between the way in which different factors affect the probability of equipment failure. If those assumptions are violated, the resulting data analysis can be invalid and, worse, misleading. For example, if the failure intensity changes with time (age) for a given type of equipment, use of the Poisson model in analyzing the data on that equipment could easily mask important information on both the short- and long-term reliability of the equipment. Invalid assumptions concerning the independence of various factors employed in a linear-model can mask or falsely create the impression of cause-and-effect relationships between various factors and the probability of failure. Rather than attempt to catalog all of the field-study conditions and assumptions associated with the use of any particular model, we will give some general guidelines on the choice and use of models in field-tracking studies.

Probably the simplest but most important rule to use in choosing a FTS model is "keep it simple." The more complicated a model is, the more parameters it will use that must be estimated during the data analysis, and the more assumptions it will require to make that analysis valid. As this last discussion implies, there are two additional rules that are closely related to the simplicity rule:

(*i*) Estimability—Since data analysis, at its core, involves making statistical inferences about parameters in the model from the available data, it is essential that the model and the collection process be matched to ensure that the right data are available in sufficient quantities to make inferences about all the parameters of interest. This is a point we have already touched on in the discussion on data collection.

(*ii*) Verifiability—The assumptions implicit in the use of any model must be verifiable or the results of the FTS will remain open to doubt. In some cases, engineering judgment can be used to justify certain model assumptions. In all cases, every effort must be made to verify assumptions from the data—either during a procedural trial (see Section 3.2 below), or as the first step in the data analysis stage of the study. A wide variety of statistical techniques are available for testing the uniformity and independence assumptions typically encountered in FTS model use; these techniques should be applied with the advice and assistance of a trained statistician.

In summary, successful data analysis is dependent on the choice of an appropriate FTS model that is matched to both the actual study conditions and to the data collection procedures employed in the study.

2.3.2 Data storage and retrieval

With the exception of very small-scale studies, involving perhaps fewer than 100 trouble events in all, computerized data storage is a great asset—if not a necessity—in permitting complete and timely analyses of field-tracking data. There are a number of data systems available [for example Data Management System (DMS)*, RAMIS®†, etc.] that lend themselves to constructing field-tracking data bases. Among the factors that must be considered are total eventual size, frequency of access required, and most important, flexibility of access—i.e., flexibility in retrieving and summarizing the data by one or more characteristics, such as equipment type or vintage, or type of trouble. It would be very difficult, for example, to compare the performances of different vintages of a given product if the data retrieval system did not permit easy, separate access to the trouble data for each vintage. On the other hand, it is important not to confuse a need for flexible data access with a need for an elaborate data retrieval system that turns out regular, detailed data summaries that display results in every conceivable way. The key is to retain flexibility without trying to preprogram every possible way of looking at the data.

III. FTS IMPLEMENTATION

In this section we briefly consider several topics in the actual implementation of an FTS:

(*i*) Developing procedures and training personnel
(*ii*) Assuring compliance with study procedures
(*iii*) Conducting a procedural trial.

3.1 Developing procedures and training personnel

Based on mutually agreed-upon objectives, specific procedures and forms for data collection need to be developed. Determining the extent of automatic data retrieval, checking the validity of the inputs, deciding exactly what data are necessary, etc., are detailed questions that require resolution.

Unless rules are provided to meet contingencies, people tend to either make up their own rules or just get discouraged about participation in the study. Although all possibilities cannot be provided for, care should be taken to anticipate the most common "unusual" events. As a default, a space for "additional comments" or "other" on data forms will alert the data analyst to the fact that the specified categories were ambiguous, not mutually exclusive, not exhaustive, etc.

The training of the field personnel who will actually perform the data collection is a very important step. Hands-on teaching with real situations will prepare them for being on their own. Giving them an

---

* Data Management System, developed by Bell Laboratories.
† RAMIS is a trademark of Mathematica, Inc.

indication of the reasons for the study and how important their participation is can improve their morale and impact on the quality of the data collected. A specific procedure to provide continuing contact and periodic feedback of results can also be a strong positive stimulus.

3.2 Compliance

It is difficult to overemphasize the importance of monitoring compliance with tracking-study procedures. The basic output of any FTS is a measure of the reliability of the equipment under study. In order for that measure to be useful and unbiased (by differences in the completeness of reporting for different products, trouble types, etc.), all or substantially all of the trouble events experienced by the equipment must be reported. It is the function of compliance procedures to ensure that this is the case.

Basically, compliance can be checked in one of two ways. If an independent (of the FTS) count of trouble events for the equipment under study is available, compliance can be checked by comparing that count to the number of troubles reported through the study procedure. This method is used in the station set PPS, where administrative counts of customer trouble reports serve as the independent count of station troubles in any PPS study location. If no such count is available, but the equipment under study is located in a geographically small, reasonably well controlled setting, such as a central office, serializing of the equipment under study and periodic mapping of the office inventory—when compared to the reported troubles—can serve as an effective compliance check. With either procedure, the key to maintaining good compliance is fast feedback to the people responsible for providing the field data and their management about the degree to which study procedures are being followed. It is for this reason, principally, that some identity of the field person reporting the trouble is included on most field-tracking study tickets.

As noted earlier, in addition to field data collection, many tracking studies involve the collection of data—usually from failure-mode analyses—at repair locations and/or diagnostic laboratories. Reporting forms for such analyses will usually have to be tailored to the particular equipment under study. But some of the general principles that govern field data collection apply to the hardware failure analysis data as well. The flow of hardware and paper must be designed to ensure that (*i*) each piece of hardware returned can be accounted for and checked off against reported field troubles, and (*ii*) individual hardware analyses can be associated with reported field trouble symptoms.

3.3 Procedural Trial

Once study procedures and forms have, at least tentatively, been developed, a trial is an excellent way to shake out unexpected problems. Here, an attempt is made to collect some actual data by people who will participate in the real study. Estimates of speed and accuracy of filling out forms, difficulties with interpreting procedures when faced with real situations, completeness of instructions, and potential usefulness of results are some of the possible outputs. If extensive revision of procedures, forms, etc., are required, a second trial may be necessary.

In addition to testing the data collection portion of the study, a trial of the data analysis methodology should also be made with simulated or actual data. It is useful to present possible conclusions, with their justification, to the users of the study results. Then, a comparison of their subjective impressions from the raw data with the quantitative results from the statistical analysis can be used to improve both. It is also at this point that model assumptions are to be verified or modified as needed.

IV. ILLUSTRATIONS

In this section, we briefly describe some recent field-tracking studies. Perhaps the longest running study is the Product Performance Survey on station sets, which we mentioned earlier in this paper. Figure 2 shows the flow of hardware and data in that study. The trouble ticket is shown in Fig. 1. Note the modularized design of the ticket described above. Analysis of returned equipment in this study is carried out by analysts in the Western Electric Quality Assurance organization who are dedicated to the study. These analysts encode the results of their analyses, as well as other information on the trouble tickets that accompany the returned hardware, for direct entry into a data base. Compliance is monitored by comparing the number of PPS trouble ticket returns to the total number of trouble reports tracked by administrative reporting systems in each study location.

A second example is illustrated in Figs. 3 and 4, which are the data-reporting form and a flow sheet, respectively, for the FTS of Northern

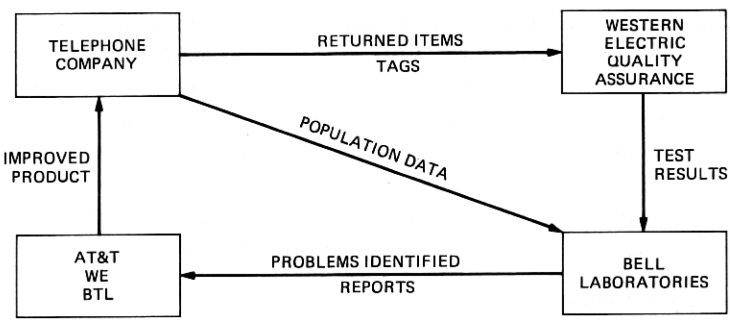


Fig. 2—Product Performance Survey data flow diagram.

DMS-10 Tracking Study Report AT&T / Bell-Northern Research

			M M D D Y Y	HR Min
Office ______		Start of Trouble/Activity		
Report No. ______		Maintenance Org. Called ?		
Generic Issue ______		Maintenance Org. Arrived ?		
Craftsman ______		Trouble Cleared ?		

REASON FOR REPORT

A □ Trouble B □ Class A Change C □ Service Disconnect Z □ Other – Specify

DESCRIPTION OF TROUBLE

A □ Total System Outage B □ Subsystem Outage C □ Automatic Sparing
D □ Excessive AUD Messages E □ Excessive BUG Messages F □ Overload
G □ Feature Problem (Specify) H □ CPF Message I □ Initially Defective
Further Description: J □ Line/Trunk Z □ Other – Specify

MEANS OF TROUBLE DETECTION

A □ Customer Complaints B □ System C □ Referred In D □ Alarm Given

RECOVERY A □ Automatic or B □ Manual

C □ SYSLOAD D □ SMART Recovery E □ INITIALIZATION

SPARING: F □ CPU G □ Memory H □ Network I □ PE Shelf J □ Line CCT K □ None

HARDWARE MAINTENANCE ACTION

A □ Unit Reseated B □ Unit Tested And Restored C □ Unit Replaced
Z □ Other (Specify)

UNIT IDENTIFICATION
Unit Location: Slot Shelf Bay

Serial Numbers

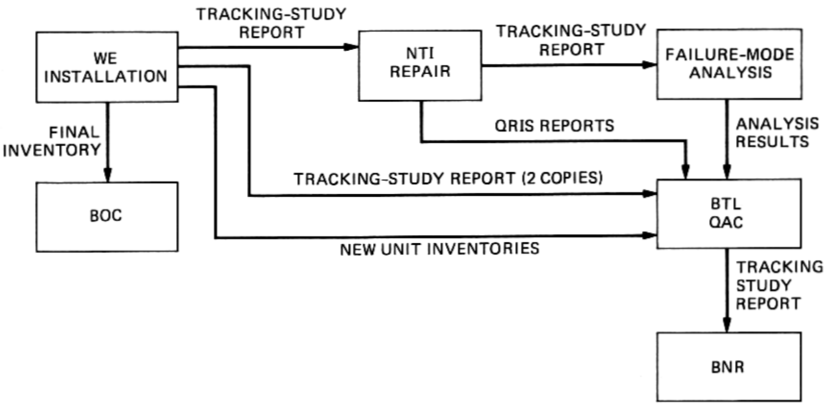
	Unit Code & Series		Study No.	NTI No.
Reported Unit				
Replacement Unit				

COMMENTS

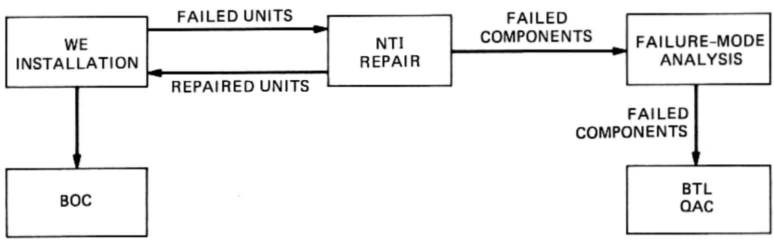
INSTRUCTIONS
1. Use a ball point pen and press firmly, you are making 4 copies.
2. Attach TTY printouts and other information (TTY printouts before and after trouble).
3. Report routing: copy 1-office file; copy 2-BTL/QAC; copy 3 -BNR via BTL/QAC copy 4 - with replaced hardware.

Fig. 3—DMS-10 Tracking Study Report.

Telecom's DMS-10 switching office. The flow sheet illustrates a point discussed in Section II, namely, that numerous organizations are often involved in an FTS. Cooperative planning among organizations involved played an important role in making this study run smoothly and produce meaningful results. The report form shows a completely different set of data fields and possible responses than did the PPS trouble ticket. Just as trouble tickets are compared with local administrative data in the station set study, report forms for this FTS are



(a)



(b)

Fig. 4—DMS-10 switching system installation tracking study. (a) Routing of information. (b) Routing of study units.

compared with maintenance and outage data automatically collected from the switching machine's maintenance output channel.

V. CONCLUSIONS

In this paper we have discussed several important aspects of planning and conducting an FTS. We have shown how careful planning beforehand in the areas of data collection, population definition, sample size, and stating of objectives is essential. We have also discussed means of ensuring that the study is producing the required ongoing data. If properly planned and conducted, FTSs can and do play a key role in assuring the quality and reliability of telecommunication products.

REFERENCES

1. V. S. Hiering and J. A. Hooke, "Product Performance Surveys: Field Tracking for Station Sets," Bell Lab Rec. (December 1978), special Quality Assurance issue.

2. W. A. Strong, unpublished work.
3. M. Hansen, W. Horowitz, and W. Madow, *Sample Survey Methods and Theory*, New York: John Wiley and Sons, Inc., 1953.
4. D. R. Cox, *Planning of Experiments*, New York: John Wiley and Sons, Inc., 1958.
5. G. G. Brush, unpublished work.
6. L. G. Lyon, unpublished work.
7. D. W. Swann, unpublished work.
8. A. Mace, *Sample Size Determination*, Melbourne, FL: Krieger Publishing Company, 1974.
9. J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, New York: John Wiley and Sons, Inc., 1982.
10. N. Draper and H. Smith, *Applied Regression Analysis*, New York: John Wiley and Sons, Inc., 1981.
11. J. Kalbfleisch, and R. Prentice, *The Statistical Analysis of Failure Time Data*, New York: John Wiley and Sons, Inc., 1980.
12. A. Duncan, *Quality Control and Industrial Statistics*, Homewood, IL: R. D. Irwin Inc., 1965.

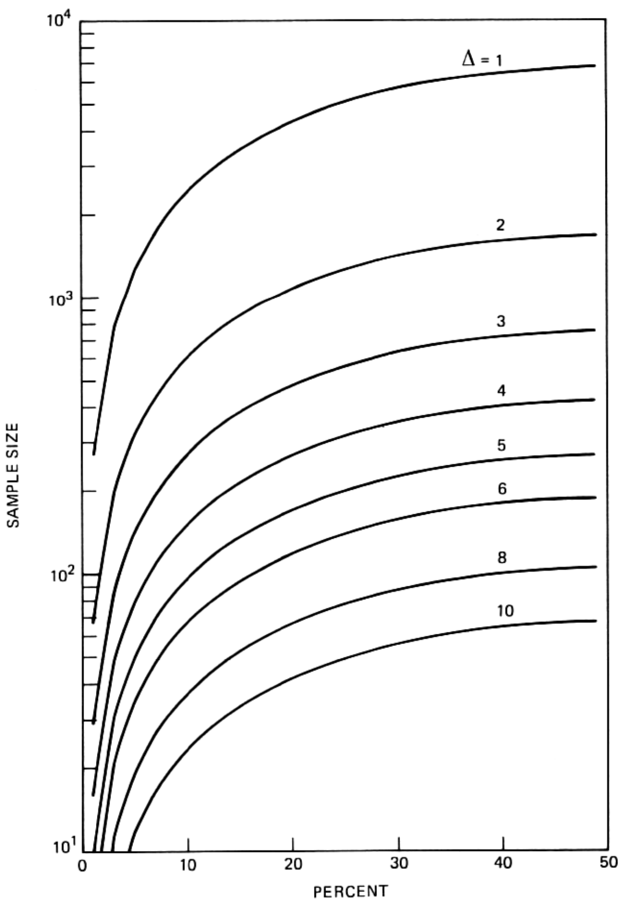


Fig. 5—Minimum sample sizes needed to generate 90-percent confidence intervals.

APPENDIX

Sample Size Selection

In this appendix we discuss in detail the six cases of sample size selection described in Section 2.2.4 of this article. These cases are:

(*i*) Estimating a parameter

(*ii*) Testing a hypothesis about one parameter

(*iii*) Comparing two parameters for both proportion and rates.

Each case is discussed in turn below. The six cases are shown in Table I, Section 2.2.4.

A.1 Case A

In Case A we wish to have a sample size to control the precision of the estimate of a percentage within certain bounds. The estimation

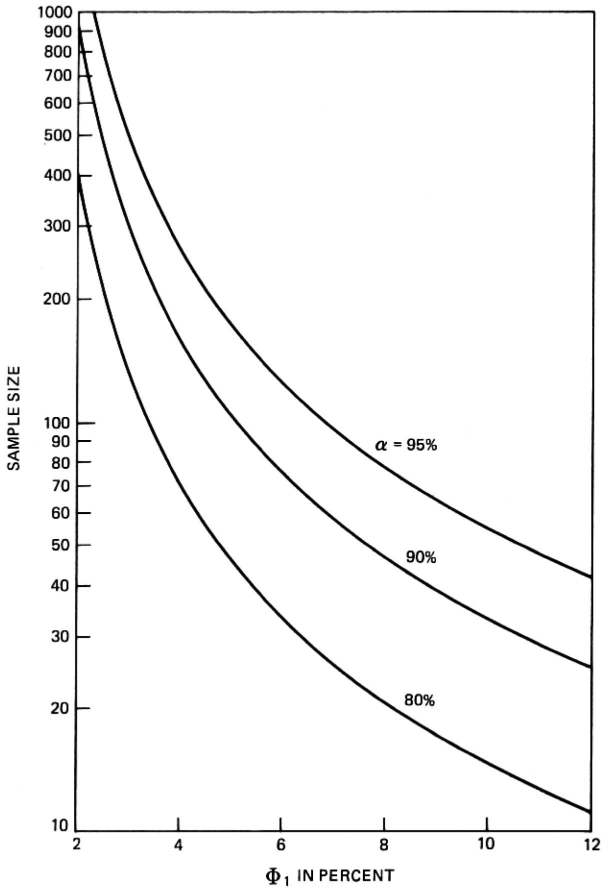


Fig. 6—Minimum sample sizes for $\Phi_0 = 1$ percent.

process is subject to imprecision; therefore, it is customary to express the estimate as an interval, say 2 to 6 percent, as opposed to a single point, say 4 percent. This interval is chosen so that if we were to repeat the process of data collection and interval construction, our intervals would cover the true, unknown percentage a very large proportion of the time. The shorter the interval, the more precise is our estimate. This interval will decrease in width as the sample size increases. We will then select the sample size before the FTS to obtain an anticipated width for our interval after the FTS. Figure 5 shows sample sizes necessary to generate 90-percent confidence intervals which are 2Δ wide. The sample size depends on the true percentage. The maximum sample size is required when the true percentage is 50 percent.

Example of Case A: Suppose we are only interested in estimating the percentage of units that are initially defective. We think that this percentage is less than 15 percent, and we want the estimated interval

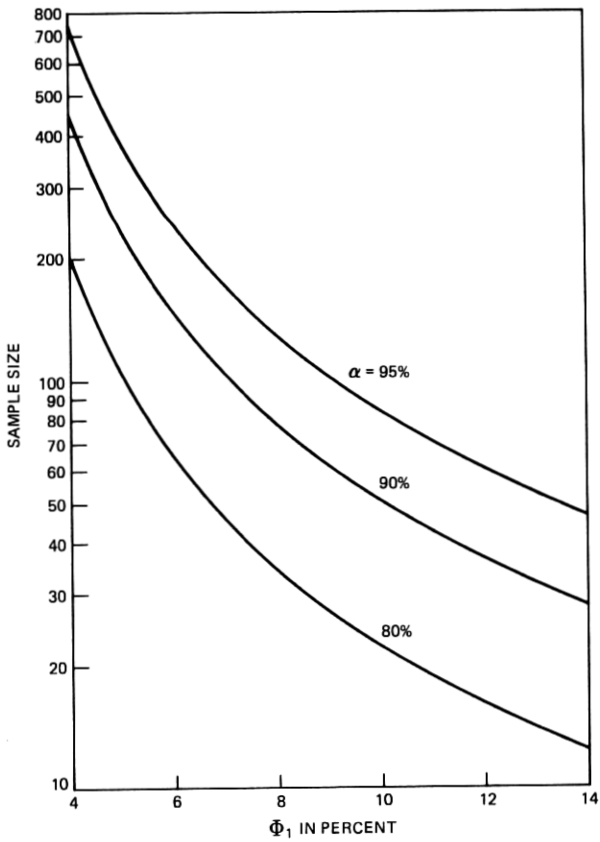


Fig. 7—Minimum sample sizes for $\Phi_0 = 2$ percent.

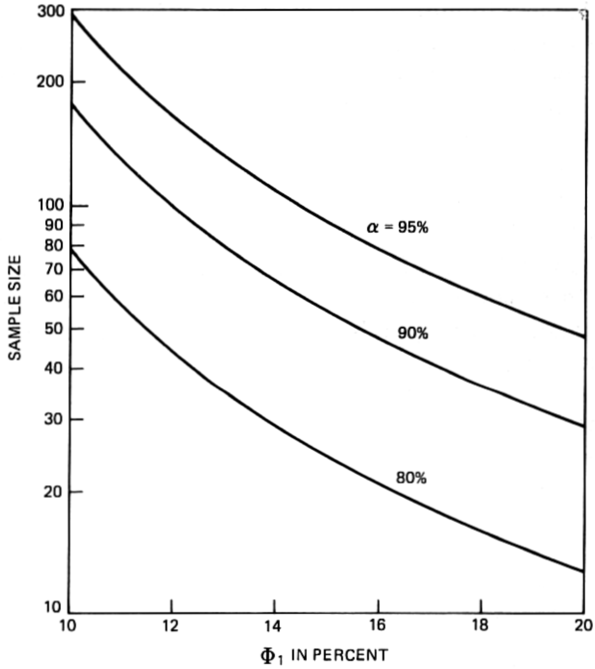


Fig. 8—Minimum sample sizes for $\Phi_0 = 5$ percent.

to be at most 6-percent wide. Therefore, $\Delta = 3$ and we see in Fig. 5 that a sample of size 400 is required. If we had no idea as to the true percentage we would use the maximum sample size for 50 percent, that is, 750. Note that the curves are symmetrical about 50 percent.

A.2 Case B

In Case B we wish to test the hypothesis that a proportion is less than or equal to Φ_0. We will look at a sample of n units, and make one of the two decisions:

(*i*) If we see that a number of units less than or equal to c, the "acceptance number", have the trait associated with the proportion, then we will accept the hypothesis that the proportion is less than or equal to Φ_0.

(*ii*) If we see that more than c of the units have the trait, then we will reject the hypothesis in favor of the alternative that the proportion is greater than Φ_0.

We wish to structure the test so that if the true value of the proportion is Φ_0, we will make decision *i* a large portion of the time, and if the true value of the proportion is Φ_1, we will make decision *ii* a large portion of the time. The reader more interested in acceptance

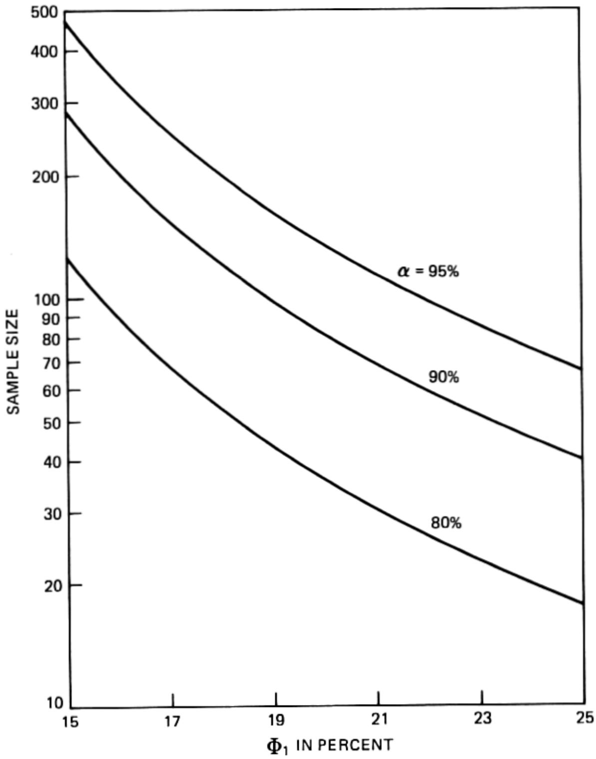


Fig. 9—Minimum sample sizes for $\Phi_0 = 10$ percent.

sampling plans, which is an example of such a situation, should refer to a specialized reference, e.g., Ref. 12.

Figures 6 through 9 show the required sample size for values of $\Phi_0 = 1, 2, 5$, and 10 percent for 80-, 90-, and 95-percent confidence levels. As an example of the use of the curves, let $\Phi_0 = 1$ and $\Phi_1 = 5$ percent. We see in Fig. 6 that for a 90-percent confidence level, a sample size of 100 is needed.

A.3 Case C

This case deals with comparing two percentages, call them percentage *A* and percentage *B*. These percentages might be similar characteristics on competing products, or competing designs. For example, we might be interested in percentages of circuit packs that are dead-on-arrival from two suppliers. After the FTS we may arrive at one of three conclusions:

(*i*) The two percentages are not significantly different

(*ii*) Percentage A is larger than percentage B

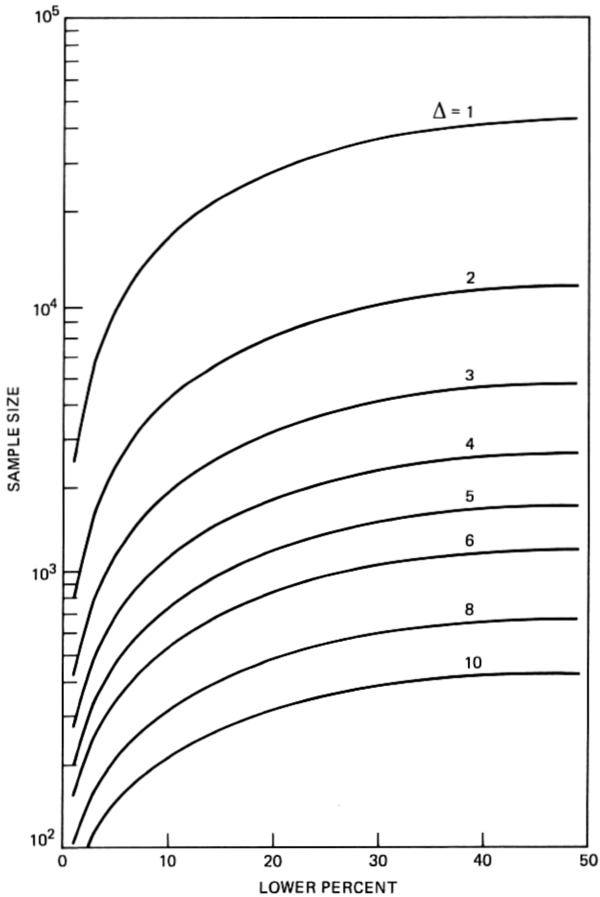


Fig. 10—Minimum sample sizes for comparing two proportions at the 90-percent confidence level.

(*iii*) Percentage B is larger than percentage A.
There are certain risks in arriving at incorrect conclusions. The risks decrease with increasing sample size. We wish to control, at a low level, the risk of not making conclusion (*i*) when percentages A and B are equal. And we wish to control, at a low level, the risk of not making conclusion (*ii*) when percentage A is Δ larger than percentage B [or, similarly, the risk of not making conclusion (*iii*) when percentage B is Δ larger than percentage A]. Figure 10 gives sample sizes necessary to accomplish this at the 90-percent confidence level.

Example of Case C: Suppose we wish to compare the percentages of plug-in units (from two suppliers) that fail during the warranty period. Further, we assume that the lower percentage will be less than 20 percent. We wish to have a high probability of concluding that the

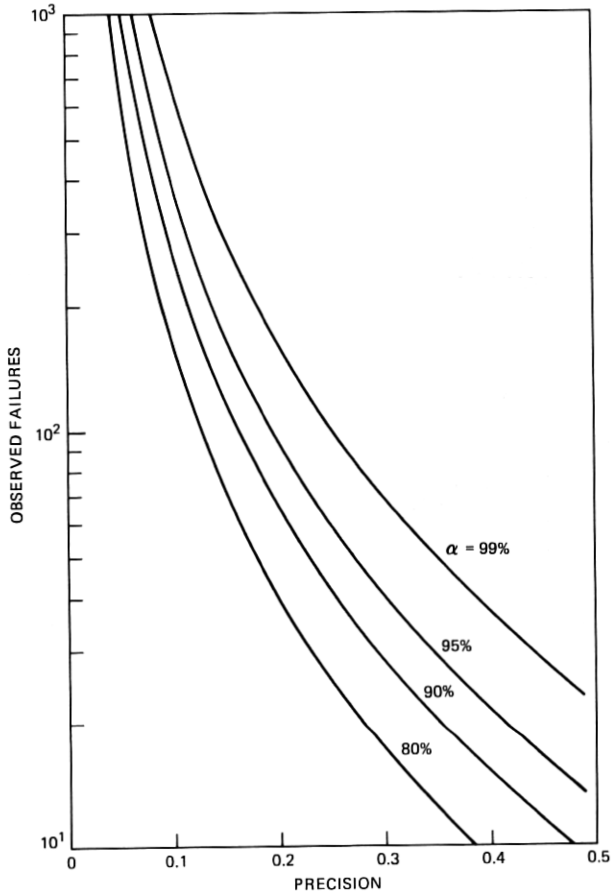


Fig. 11—Minimum observed failures for estimating a failure rate.

upper percentage is greater than the lower percentage when the upper percentage is 5 greater than the lower. For $\Delta = 5$ and a lower percent of 20, we need to look at 1300 units from each supplier. With no knowledge of the true percentages we would use the sample size for 50 percent, that is, 1700.

A.4 Case D

Cases D, E, and F deal with failure rates, as opposed to the percentages of Cases A, B, and C. (The results for Cases D, E, and F must be used subject to the cautions given at the end of this appendix.) Cases D, E, and F require the use of two curves. The first curve will tell us how many failures we need to see. The second curve will tell us how many units must be included in the FTS so that we are reasonably

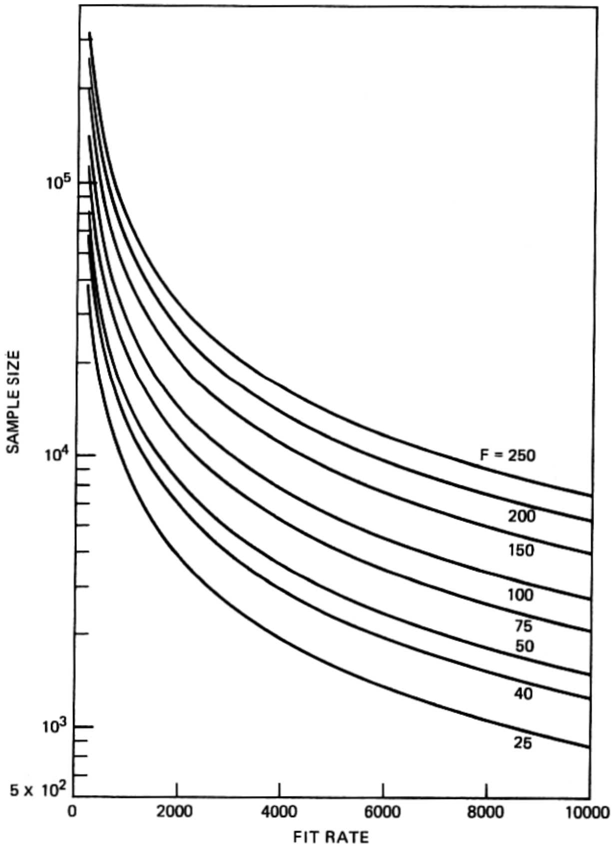


Fig. 12—Minimum sample sizes for failure-rate estimation (6-month interval).

certain that the failures occur in a prescribed time period. In Case A we measure the precision of our estimation by the width of the interval, expressed in *absolute* percentages. In Case D, we will measure the precision in terms of *relative* percentages. For example, if our interval is 1500 FITS* ± 5 percent = 1500 ± 75 FITS = (1425, 1575), then we will say that the precision is 5 percent. This interval corresponds to (1.25, 1.38) failures per 100 sets per year.

Example of Case D: Suppose that we wish to obtain a precision of 15 percent at the 90-percent confidence level in the estimate of the failure rate of a plug-in unit. In Fig. 11 at an abscissa of 0.15 (15 percent) we see that 120 failures must be observed. Suppose that the FTS is to last 12 months and that our reliability prediction gives us an

* FIT = Failures in 10^9 hours = 8.75×10^{-4} failures per 100 units per year.

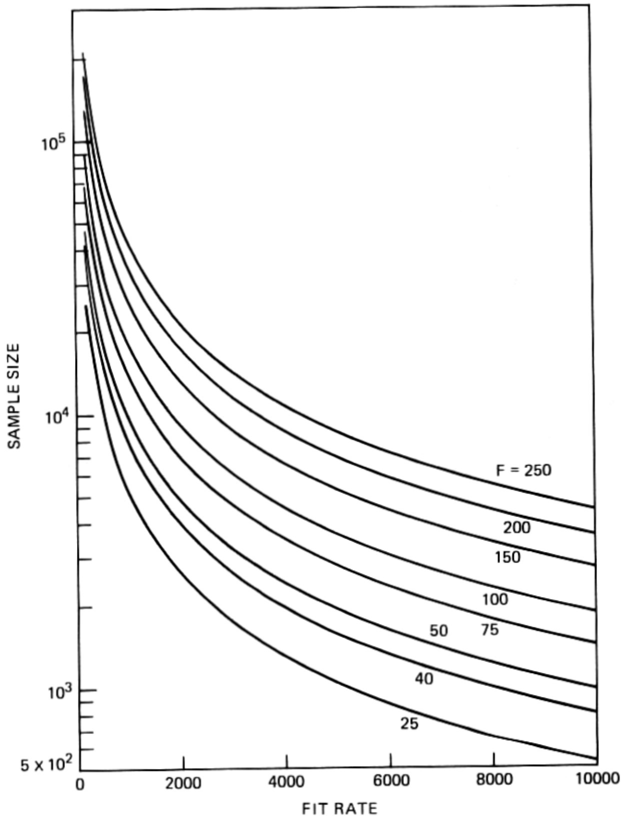


Fig. 13—Minimum sample sizes for failure-rate estimation (9-month interval).

estimated FIT rate of 2500. In Fig. 14 we see that about 7000 units need to be included in the study.

Figures 12 through 15 give required sample sizes for studies of lengths 6, 9, 12, and 18 months, and for FIT rates up to 10,000. If some other combination is needed, then the following formula should be used:

$$N = \frac{F + 1.645 \times \sqrt{F}}{(7.2 \times 10^{-7})\lambda T} + \frac{F}{2}, \tag{1}$$

where

F is the number of failures,
λ is the prior estimate of the failure rate in FITs (failures in 10^9 hours),
T is the number of months the study will last, and
N is the required sample size.

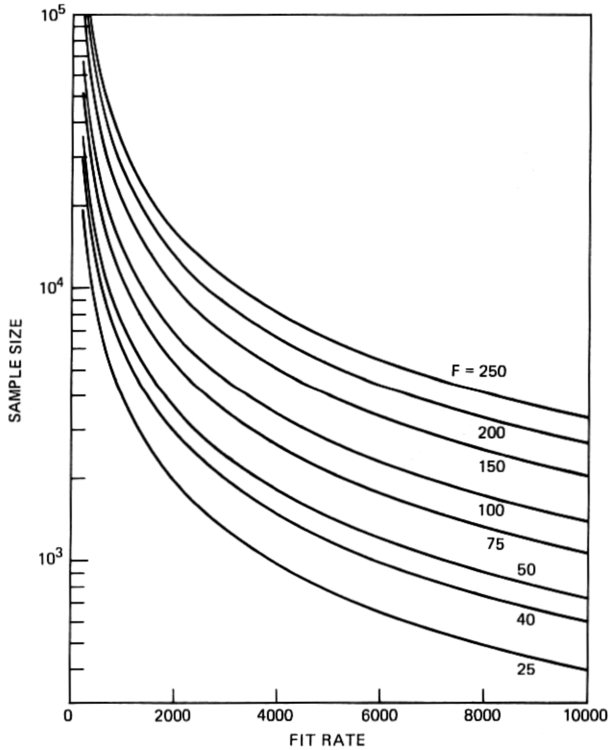


Fig. 14—Minimum sample sizes for failure-rate estimation (12-month interval).

This formula provides 95-percent confidence that the required number of failures will be observed.

A.5 Case E

In Case E we wish to test the hypothesis that a rate is less than or equal to a specified value, $V1$. Based upon the data observed, we will either

(*i*) Accept the hypothesis that the rate is less than or equal to $V1$, or

(*ii*) Reject the above hypothesis in favor of the alternative that the failure rate is greater than $V1$.

We wish to structure the test so that if the true value of the rate is $V1$, we make decision *i* with a high probability and if the true value of the rate is $(R)V1$, we make decision *ii* with a high probability.

Example of Case E: Suppose we wish to check to see how a newly designed part has changed the reliability of a piece of equipment. We are satisfied with $R = 2$ and the 90-percent confidence level. Figure 16 shows that 15 failures must be observed.

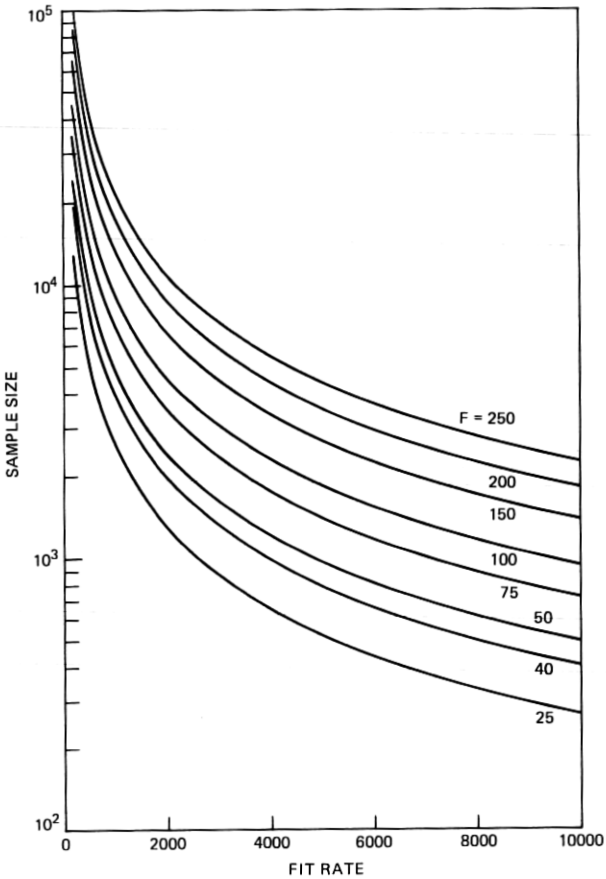


Fig. 15—Minimum sample sizes for failure-rate estimation (18-month interval).

A.6 Case F

Here we wish to compare failure rates of two competing products. At the end of the FTS we can arrive at one of three conclusions:

(*i*) Failure rate A and failure rate B are not significantly different
(*ii*) Failure rate A is larger than failure rate B
(*iii*) Failure rate B is larger than failure rate A

Again there are risks of arriving at incorrect decisions. As we increase the sample sizes, we can decrease these risks. We wish to control, at a low level, the risk of not making conclusion (*i*) when failure rates A and B are equal. And we wish to control, at a low level, the risk of not making conclusion (*ii*) when failure rate A is R times as large as failure rate B (or similarly the risk of not making conclusion (*iii*) when failure rate B is R times as large as failure rate A).

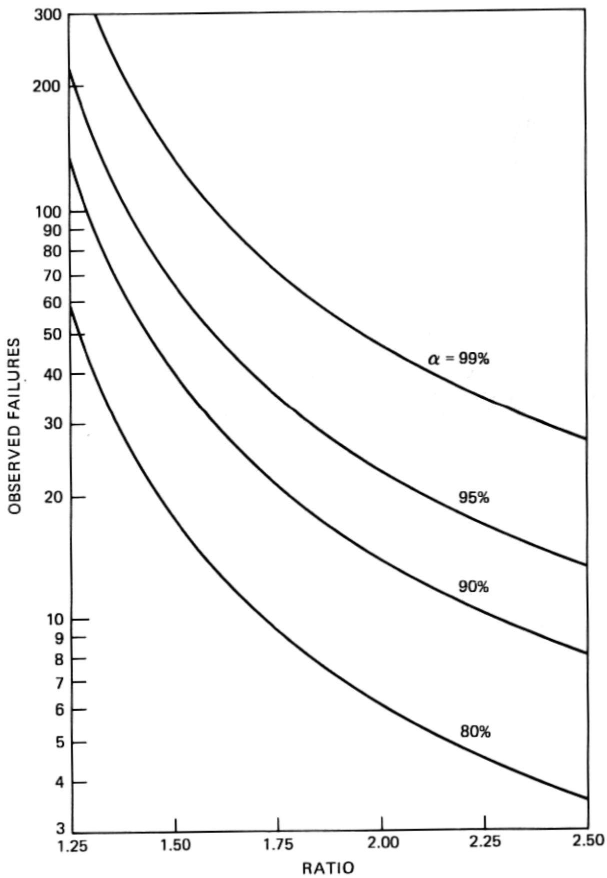


Fig. 16—Minimum observed failures for a hypothesis test (one failure rate).

Example of Case F: Suppose we wish to use the FTS to compare the failure rates of the channel units of two different suppliers. Suppose further that we wish to have a high chance (90-percent probability) of concluding that the larger failure rate is larger than the smaller failure rate when indeed the larger failure rate is twice the smaller. In Fig. 17 we see that we need to observe about 36 failures. If the study is to last 12 months and our reliability prediction yields an estimate of 6000 FITS, then Fig. 14 shows that a sample size of 900 is required to be 95 percent certain of observing the required number of failures. That is, we need 900 of each supplier's units in the study.

A.7 Cautions

In Cases D, E, and F, if the required number of failures is not observed in the nominal time period for the FTS, then the desired

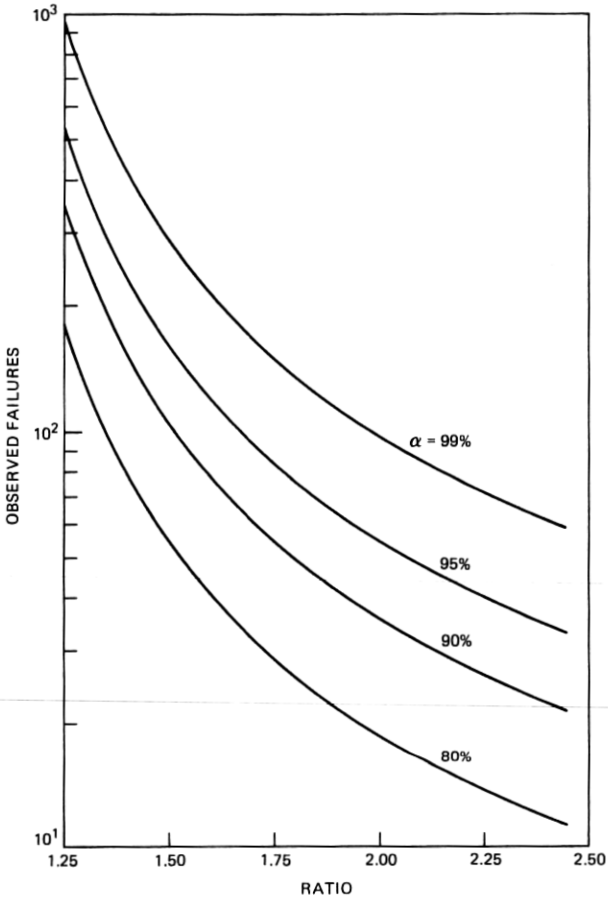


Fig. 17—Minimum observed failures comparing two failure rates.

precision will not be achieved. (This might occur if the reliability prediction is in error and yields a higher than actual FIT rate as a prediction. If the prediction is much higher than the actual, we will be incorrectly led to believe that the required number of failures will be observed in a shorter interval than is actually needed.) In this case it would be wise to extend the study period until the required number of failures is observed.

The theory developed for Cases D, E, and F requires that the failure rate be constant throughout the FTS. Even for very large sample sizes, the theory is sensitive to departures from this assumption. Therefore, if we know that the failure rate is high for one time period (e.g., early life) and low for a different time period (e.g., steady state), then we must do a separate analysis on each period, as shown in the following example.

Assume that the early failure period is 3 months. Our reliability predictions indicate that the early failure rate will be about 10,000 FITs and that the steady-state failure rate will be about 4,000 FITs. We wish to obtain a precision of 0.25 at the 90-percent confidence level in estimating each of the failure rates in an FTS that we wish to finish in 6 months or less. What sample size is needed? Figure 11 shows that we need to observe 41 failures, that is, we must observe 41 failures in the early-life period (months 1 to 3) and 41 failures in the steady-state period (months 4 to 6). Use of eq. (1) shows that we need at least 2400 units for the early-life period and 5980 for the steady-state period. Since we need to satisfy both requirements we will need a sample size of 5980.

The example above illustrates another important point. If you want to use the FTS to estimate several characteristics, then go through the sample size analysis for each characteristic. The FTS will satisfy all requirements if it has the maximum of the required sample sizes.

In Cases B, D, E, and F, curves for several confidence levels are placed on one page. However, for Cases A and C, each confidence level would take a separate page, so only the 90-percent confidence level was given. For other confidence levels, see Ref. 8.

LETTER TO THE EDITOR

Comments on "Voice Storage in the Network—Perspective and History," by E. Nussbaum*

In a recent article E. Nussbaum discussed the FCC's rejection of AT&T's petition for waiver to allow the offering of Custom Calling Services II in the U.S. under the Computer Inquiry II decision. Unfortunately, references were not given to these decisions for the benefit of those readers who may wish to learn more about this apparent frustration of technology and the policy issues involved. The FCC rejection can be found in 88 FCC 2d 1. The Computer Inquiry II decision is given in 47 CFR 64.702, adopted in 77 FCC 2d 384 (Final Decision) on reconsideration, 84 FCC 2d 50, appeal pending *sub nom* CCIA vs. FCC, Case No. 80-471 (D.C. Cir. 1980).

Michael J. Marcus
Acting Chief
Technical Analysis Division
Office of Science & Technology
Federal Communications Commission

---

* B.S.T.J., *61*, No. 5 (May–June 1982), pp. 811–13.