Statistical Science
1989, Vol. 4, No. 1, 2–33

Statistical Models and Analysis in Auditing

Panel on Nonstandard Mixtures of Distributions

Abstract. This report is a study of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing. It was prepared by the Committee on Applied and Theoretical Statistics of the Board on Mathematical Sciences, National Research Council, through its Panel on Nonstandard Mixtures of Distributions. A comprehensive survey of the various methodologies that have been provided in the literature is presented, together with numerical examples. A detailed annotated bibliography of statistical practice in auditing is included.

Key words and phrases: Accounting, auditing, Bayesian models, confidence bounds, dollar unit sampling, mixtures of distributions, nonstandard mixtures, sampling theory.

1. INTRODUCTION

One of the first problems of national importance that was considered by the Committee on Applied and Theoretical Statistics (CATS) was posed to it by staff members of the Internal Revenue Service (IRS). They were concerned with the lack of appropriate statistical methodologies for certain nonstandard situations that arise in auditing where the distributions appropriate for modeling the data are markedly different from those for which most statistical analyses were designed.

The quality of the procedures used in a statistical analysis depends heavily on the probability model or distributions assumed. Because of this, considerable effort over the years has been expended in the development of large classes of standard distributions, along with relevant statistical methodologies, designed to serve as models for a wide range of phenomena. However, there still remain many important problems where the data do not follow any of these more "standard" models. The problem raised by the IRS provides a strikingly simple example of data from a nonstandard distribution for which statistical methodologies have only recently begun to be developed, and for which much additional research is needed. The example is of such national importance, both for government agencies and for business and industry, that it is the primary focus of this report. The potential monetary losses associated with poor statistical practice in this auditing context are exceedingly high.

It is the purpose of this report to give a survey of the available statistical methods, to provide an annotated bibliography of the literature on which the survey is based, to summarize important open questions, to present recommendations designed to improve the level and direction of research on these matters and to encourage greater interaction between statisticians and accountants. This report is primarily directed toward researchers, both in statistics and accounting, and students who wish to become familiar with the important problems and literature associated with statistical auditing. It is hoped that this report will stimulate the needed collaborative research in statistical auditing involving both statisticians and accountants. It is also hoped that practitioners will benefit from the collection of methodologies presented here, and possibly will be able to incorporate some of these ideas into their own work.

Although this report is centered upon a particular nonstandard distribution that arises in auditing, the original proposal for this study recognized that this

The members of the Panel on Nonstandard Mixtures of Distributions were Donald Guthrie, University of California, Los Angeles, Chairman; Z. W. Birnbaum, University of Washington; Wilfrid J. Dixon, University of California, Los Angeles; Stephen E. Fienberg, Carnegie Mellon University; Jane F. Gentleman, University of Waterloo, Canada; James M. Landwehr, AT&T Bell Laboratories; Nancy R. Mann, University of California, Los Angeles; Paul W. Mielke, Colorado State University; John Neter, University of Georgia; Donald M. Roberts, University of Illinois; John Van Ryzin, Columbia University (deceased); H. Tamura, University of Washington; Arthur J. Wilburn, A. J. Wilburn Associates; and James S. Williams, Colorado State University. The corresponding author for this article is H. Tamura, Graduate School of Business Administration, University of Washington, Seattle, Washington 98195.

same type of nonstandard model arises in many quite different applications covering almost all other disciplines. Three general areas of application (accounting, medicine and engineering) were initially chosen for consideration by the Panel. Later in this Introduction we list several examples in order to illustrate the widespread occurrence of similar nonstandard models throughout most areas of knowledge. These examples will, however, primarily reflect the original areas of emphasis of the Panel. Before describing these examples, however, we briefly discuss the general concept of a mixture of distributions because it appears in the name of the Panel.

Nonstandard Mixtures

The phrase "mixture of distributions" usually refers to a situation in which the jth of k (taken here to be finite) underlying distributions is chosen with probability $p_j, j = 1, \cdots, k$. The selection probabilities are usually unknown and the number of underlying distributions k may be fixed or random. The special case of two underlying distributions is an important classical problem that encompasses this report's particular problem in which, with probability p, a specified *constant* is observed whereas, with probability $1 - p$, one observes a random measurement whose distribution has a density function. That is, it is a mixture of a degenerate distribution and an absolutely continuous one.

There are many examples of probability models that are best described as mixtures of two or more other models in the above sense. For example, a probability model for the heights of 16 year olds would probably best be described as the mixture of two unimodal distributions, one representing the model for the heights of girls and one for the boys. Karl Pearson in 1894 was possibly the first to study formally the case of a mixture of two distributions; in this case they were two normal distributions thereby providing one possible mixture model for the above example of heights. Following this, there were few if any notable studies until the paper of Robbins and Pitman (1949) in which general mixtures of χ^2 distributions were derived as probability models for quadratic forms of normal random variables. Since then, there have been many other papers dealing with particular mixture models. The published research primarily deals with mixtures of distributions of similar types, such as mixtures of normal distributions, mixtures of χ^2 distributions, mixtures of exponential distributions, mixtures of binomial distributions and so on. However, the literature contains very few papers that provide and deal with special "nonstandard" mixtures that mix discrete (degenerate, even) and continuous distributions as emphasized in this report.

In general, the word mixture refers to a convex combination of distributions or random variables. To illustrate, suppose X and Y are random variables with distribution functions F and G, respectively. Let $0 \leq p \leq 1$. Then $H = pF + (1 - p)G$ is a distribution function that may be called a mixture of F and G. The interpretation of H is that it represents a model in which the distribution F is used with probability p while G is used with probability $1 - p$. In terms of random variables, one may say that H models an observation Z that is obtained as follows: With probability p observe X having distribution F, and with probability $1 - p$ observe Y having distribution G. Such mixtures may then be viewed as models for data that may be interpreted as the outcomes of a two-stage experiment. In the first stage, a population is randomly chosen and then in the second stage an observation is made from the chosen population.

It is not necessary to limit oneself to mixtures of just two or even a finite number of distributions. In general, one may have an arbitrarily indexed family of distributions, for which an index is randomly chosen from a given mixing distribution. It should also be emphasized that there is considerable ambiguity associated with mixtures; every distribution may be expressed as a mixture in infinitely many ways. Nevertheless, when mixture models are formulated reasonably, they can provide useful tools for statistical analysis. There is by now a large literature pertaining to statistical analyses of mixtures of distributions; for a source of references, see Titterington, Smith and Makov (1985). Problems and applications of mixtures also appear in the literature associated with the term heterogeneity; see Keyfitz (1984).

Applications Involving Nonstandard Mixtures

The interpretation of the nonstandard mixtures emphasized in this report is quite simple. If F, the degenerate distribution, is chosen in the first stage, the observed value of the outcome is zero; otherwise the observed value is drawn from the other distribution. In what follows we illustrate several situations in which this type of nonstandard mixture may arise, and indicate thereby its wide range of applications. There are of course fundamental differences among many of these applications. For example, in some of these applications, the mixtures are distinguishable in the sense that one can tell from which population an observation has come, whereas in others the mixtures are indistinguishable. In many applications it is necessary to form restrictive parametric models for the nondegenerate distribution G; in at least one example, G is itself seen to arise from a mixture. In some cases G admits only positive values of X; in other cases G presents positive, negative or even zero

values. Of course, if G also permits zero values with positive probability, then the mixture is clearly indistinguishable.

The descriptions of the following applications are brief and somewhat simplified. They should suffice, however, to indicate the broad diversity of important situations in which these nonstandard mixtures arise. We begin with the auditing application that is the focus of this report.

1. In auditing, some population elements contain no errors, whereas other population elements contain errors of varying amounts. The distribution of errors can, therefore, be viewed as a mixture of two distinguishable distributions, one with a discrete probability mass at zero and the other a continuous distribution of non-zero positive and/or negative error amounts. The main statistical objective in this auditing problem is to provide a statistical bound for the total error amount in the population. The difficulty inherent in this problem is the typical presence of only a few or no errors in a given sample. This application will be the main focus of this report; it is studied at length in Section 2.

Independent public accountants often use samples to estimate the amount of monetary error in an account balance or class of transactions. Their interest usually centers on obtaining a statistical upper bound for the true monetary error, a bound that is most likely going to be greater than the error. A major concern is that the estimated upper bound of monetary error may in fact be less than the true amount more often than desired. Governmental auditors are also interested in monetary error—the difference between the costs reported and what should have been reported, for example. Because the government may not wish to overestimate the adjustment that the auditee owes the government, interest often centers on the lower confidence limit of monetary error at a specified confidence level allowed by the policy.

The mixture problem affects both groups of auditors as well as internal auditors who may be concerned with both upper and lower limits. In all cases there is a serious tendency for the use of standard statistical techniques, that are based upon the approximate normality of the estimator of total monetary error, to provide erroneous results. Specifically, as will be reviewed in the next section, both confidence limits tend to be too small. Upper limits being too small means that the frequency of upper limits exceeding the true monetary error is less than the nominal confidence level. Lower limits being too small means that the frequency of lower limits being smaller than the true monetary error is greater than the nominal confidence level. To the auditors these deficiencies have important practical consequences.

Most of the research to date has been directed toward the independent public accountants' concern with the upper limit. For example, the research outlined in Section 2 that is concerned with sampling of dollar units represents a major thrust in this direction. By contrast, very little research has been done on the problem of the lower confidence bound. This represents an area of considerable importance where research is needed.

2. In a community a particular service, such as a specific medical care, may not be utilized by all families in the community. There may be a substantial portion of nontakers of such a service. Those families who subscribe to it do so in varying amounts. Thus the distribution of the consumption of the service may be represented by a mixture of zeros and positive values.

3. In the mass production of technological components of hardware, intended to function over a period of time, some components may fail on installation and therefore have zero life lengths. A component that does not fail on installation will have a life length that is a positive random variable whose distribution may take different forms. Thus, the overall distribution of lifetimes which includes the duds is a nonstandard mixture.

4. In measuring precipitation amounts for specified time periods, one must deal with the problem that a proportion of these amounts will be zero (i.e., measured as zero). The remaining proportion is characterized by some positive random variable. The distribution of this positive random variable usually looks reasonably smooth, but in fact is itself a complex mixture arising from many different types of events.

5. In the study of human smoking behavior, two variables of interest are smoking status—Ever Smoked and Never Smoked—and score on a "Pharmacological Scale" of people who have smoked. This also is a bivariate problem with a discrete variate—0 (Never Smoked), 1 (Ever Smoked)—and a continuous variate "Pharmacological Score." A nontrivial conditional distribution of the second variate can be defined only in association with the 1 outcome of the first variate. This problem can be further complicated by nonresponse on either of the first or second variates.

6. In the study of tumor characteristics, two variates may be recorded. The first is the absence (0) or presence (1) of a tumor and the second is

tumor size measured on a continuous scale. In this problem, it is sometimes of interest to consider a marginal tumor measurement that is 0 with nonzero probability an example of a mixture of unrelated distributions. The problem can be further complicated by recognizing that the absence of a tumor is an operational definition and that in fact patients with nondetectable tumors will be included in this category.

7. In studies of genetic birth defects, children can be characterized by two variates, a discrete or categorical variable to indicate if one is not affected, affected and born dead, or affected and born alive, and a continuous variable measuring the survival time of affected children born alive. The conditional distribution of survival time given this first variable is undefined for children who are not affected, a mass point at 0 for children who are affected and born dead, and nontrivial for children who are born alive. In some cases it may be necessary to consider the conditional survival time distribution for affected children as a mixture of a mass point (at 0) and a nontrivial continuous distribution.

8. Consider measurements of physical performance scores of patients with a debilitating disease such as multiple sclerosis. There will be frequent zero measurements from those giving no performance and many observations with graded positive performance.

9. In a study of tooth decay, the number of surfaces in a mouth which are filled, missing or decayed are scored to produce a decay index. Healthy teeth are scored 0 for no evidence of decay. The distribution is a mixture of a mass point at 0 and a nontrivial continuous distribution of decay score. The problem could be further complicated if the decay score is expressed as a percentage of damage to measured teeth. The distribution should then be a mixture of a discrete random variable (0—healthy teeth, 1—all teeth missing) with nonzero probability of both outcomes and a continuous random variable (amount of decay in the (0, 1) interval).

10. In studies of methods for removing certain behaviors (e.g., predatory behavior or salt consumption), the amount of the behavior which is exhibited at a certain point in time may be measured. In this context, complete absence of the target behavior may represent a different result than would a reduction from a baseline level of the behavior. Thus, one would model the distribution of activity levels as a mixture of a discrete value of zero and a continuous random level.

11. Time until remission is of interest in studies of drug effectiveness for treatment of certain diseases. Some patients respond and some do not. The distribution is a mixture of a mass point at 0 and a nontrivial continuous distribution of positive remission times.

12. In a quite different context, important problems exist in time-series analysis in which there are mixed spectra containing both discrete and continuous components.

In some of the above examples, the value zero is a natural extension of the possible measurements, and in other examples it is not. For example, in measuring behavioral activity (Example 9), a zero measurement can occur because the subject has totally ceased the behavior, or because the subject has reduced the behavior to such a low average level that the time of observation is insufficient to observe the behavior. This indecision might also occur in the example concerning tumor measurement or in rainfall measurement. In other examples, however, it is possible to determine the source of the observation. The very fact that the service lifetime of a component in Example 3 is zero identifies that component as a dud, and in Example 7 there is a clear distinction between stillborn and liveborn children. These two kinds of examples represent applications of indistinguishable and distinguishable mixtures, respectively.

2. STATISTICAL MODELS AND ANALYSES IN AUDITING

2.1 The Beginnings

The field of accounting encompasses a number of subdisciplines. Among these, two important ones are *financial accounting* and *auditing*. Financial accounting is concerned with the collection of data about the economic activities of a given firm and the summarizing and reporting of them in the form of financial statements. Auditing, on the other hand, refers to the independent verification of the fairness of these financial statements. The auditor collects data that is useful for verification from several sources and by different means. It is very evident that the acquisition of reliable audit information at low cost is essential to economical and efficient auditing.

There are two main types of audit tests for which the acquisition of information can profitably make use of statistical sampling. First, an auditor may require evidence to verify that the accounting treatments of numerous individual transactions comply with prescribed procedures for internal control. Second, audit evidence may be required to verify that reported monetary balances of large numbers of individual items are not materially misstated. The first audit test, collecting data to determine the rate of procedural errors of a population of transactions is called a

compliance test. The second, collecting data for evaluating the aggregate monetary error in the stated balance, is called a *substantive test of details.* The auditor considers an error to be *material* if its magnitude "is such that it is probable that the judgement of a reasonable person relying upon the report would have been changed or influenced by the inclusion or correction of the item" (Financial Accounting Standards Board, 1980).

Current auditing standards set by the American Institute of Certified Public Accounts (AICPA) *do not* mandate the use of statistical sampling when conducting audit tests (AICPA, 1981, 1983). However, the merits of random sampling as the means to obtain, at relatively low cost, reliable approximations to the characteristics of a large group of entries, were known to accountants as early as 1933 (Carman, 1933). The early applications were apparently limited to compliance tests (Neter, 1986). The statistical problems that arise, when analyzing the type of nonstandard mixture of distributions that is the focus of this report, did not surface in auditing until the late 1950s. At about that time, Kenneth Stringer began to investigate the practicality of incorporating statistical sampling into the audit practices of his firm, Deloitte, Haskins & Sells. It was not until 1963 that some results of his studies were communicated to the statistical profession. The occasion was a meeting of the American Statistical Association (Stringer, 1963, 1979).

Before summarizing Stringer's main conclusions, we describe the context as follows. An item in an audit sample produces two pieces of information, namely, the *book* (recorded) amount and the *audited* (correct) amount. The difference between the two is called the *error amount.* The percentage of items in error may be small in an accounting population. In an audit sample, it is not uncommon to observe only a few items with errors. An audit sample may not yield any nonzero error amounts. For analyses of such data, in which most observations are zero, the classical interval estimation of the total error amount based on the asymptotic normality of the sampling distribution is not reliable. Also, when the sample contains no items in error, the estimated standard deviation of the estimator of the total error amount becomes zero. Alternatively, one could use the sample mean of the audited amount to estimate the total mean audited amount for the population. The estimate of the mean is then multipled by the known number of items in the population to estimate the population total. In the audit profession, this method is referred to as *mean-per-unit estimation* (AICPA, 1983). Because observations are audited amounts, the standard deviation of this estimator can be estimated even when all items in the sample are error-free. However, because of the large variance of the audited amount that may arise in simple random sampling, the mean-per-unit estimation is imprecise. More fundamentally, however, when the sample does not contain any item in error, the difference between the estimate of the total audited amount and the book balance must be interpreted as the sampling error. The auditor thus evaluates that the book amount does not contain any *material error.* This is an important point for the auditor. To quote from Stringer (1963) concerning statistical estimates ("evaluations") of total error:

> Assuming a population with no error in it, each of the possible distinct samples of a given size that could be selected from it would result in a different estimate and precision limit under this approach; however, *from the view point of the auditor, all samples which include no errors should result in identical evaluations.*

Stringer then reported in the same presentation that he, in collaboration with Frederick F. Stephan of Princeton University, had developed a new statistical procedure for his firm's use in auditing that did not depend on the normal approximation of the sampling distribution and that could still provide a reasonable inference for the population error amount when all items in the sample are error-free. This sampling plan is apparently the original implementation of the now widely practiced dollar (or monetary) unit sampling and is one of the first workable solutions proposed for the nonstandard mixtures problem in accounting. However, as it is studied later in this report, the method assumes that errors are overstatements with the maximum size of an error of an item equal to its book amount. Another solution, using a similar procedure, was devised by van Heerden (1961). His work, however, was slow to become known within the American accounting profession.

In the public sector, statistical sampling has also become an integral part of audit tools in the Internal Revenue Service since the issuance of the 1972 memo by their Chief Council (IRS, 1972, 1975). In a *tax examination* the audit agent uses statistical sampling of individual items to estimate the adjustment, if necessary, for an aggregate expense reported in the tax return. Statistical auditing may also be utilized by other governmental agencies. For example, the Office of the Inspector General of the Department of Health and Human Services investigates compliance of the cost report of a state to the Medicaid policy by using statistical sampling of items. In these cases, a large proportion of items in an audit sample requires no adjustment, i.e., most sample items are allowable deductions. Because an individual item adjustment is seldom negative, the audit data for estimation of the total adjustment is a mixture of a large percentage of zeros and a small percentage of positive numbers.

Thus, the mixture model and related statistical problems that are important to accounting firms in auditing also arise in other auditing contexts such as those associated with IRS tax examinations. Significant differences also exist in these applications, however, and these will be stressed later.

For concise accounts of the problems of statistical auditing one is referred to Knight (1979), Smith (1979) and Neter (1986); the last reference also includes recent developments. Leslie, Teitlebaum and Anderson (1980) also provide an annotated bibliography that portrays the historical development of the subject through 1979. In the sections which follow, however, we provide a comprehensive survey of the research efforts that have contributed to the identification and better understanding of problems in statistical auditing. We include brief descriptions of many of the solutions that have been proposed for these problems along with their limitations. It will be noticed that the solutions thus far proposed are mainly directed toward the special need for good upper bounds on errors when errors are overstatements. This is an important and common audit problem for accounting firms but in the case of tax examinations, although the mixture distribution is similar, the interest is in the study of lower bounds. Thus in statistical auditing, whether in the private or public sector, the investigator's interest is usually concerned with *one-sided* problems, i.e., of an upper or a lower bound, rather than *two-sided* problems as currently stressed in many texts.

The next section provides the definitions and notations that are used. Then in Sections 2.3 through 2.7, we present various methodologies that have been provided in the literature. Numerical examples are given in the last section to illustrate some of the alternative procedures.

2.2 Definitions and Notation

An *account*, such as accounts receivable or inventory, is a population of individual accounts. To distinguish the use of the word "account" in the former sense from the latter, we define the constituent individual accounts, when used as audit units, as line items. Let Y_i and X_i, the latter not usually known for all values of i, denote the book (recorded) amount and the audited (correct) amount, respectively, for the ith line item of an account of N line items. The book and audited balances of the account are, respectively,

$$Y = \sum_{i=1}^{N} Y_i, \tag{2.2.1}$$

called the *population book amount*, and

$$X = \sum_{i=1}^{N} X_i, \tag{2.2.2}$$

called the *population audited amount.* The *error amount* of the ith item is defined to be

$$D_i = Y_i - X_i. \tag{2.2.3}$$

When $D_i > 0$, we call it an overstatement and, when $D_i < 0$, an understatement. When $Y_i \neq 0$, the fractional error,

$$T_i = D_i/Y_i, \tag{2.2.4}$$

is called the *tainting* or simply the *taint* of the ith item. It is the error amount per dollar unit of the ith item. We may then write

$$D_i = T_i Y_i. \tag{2.2.5}$$

The error of the book balance of the account is thus

$$D = Y - X = \sum_{i=1}^{N} D_i = \sum_{i=1}^{N} T_i Y_i. \tag{2.2.6}$$

As emphasized in Section 2.1, a large proportion of items in an audit population will likely be error-free, so that $D_i = 0$ for many values of i. Similar populations are common in many disciplines as discussed in Section 1.

Aitchison (1955) was the first to consider an inference problem for such a population. Following his approach, the error d of an item randomly chosen from an accounting population may be modeled as

$$d = \begin{cases} z & \text{with probability } p, \\ 0 & \text{with probability } (1-p), \end{cases} \tag{2.2.7}$$

where p is the proportion of items with errors in the population and $z \neq 0$ is a random variable representing the error amount. z may depend on the book amount. The nonstandard mixture problem that is the focus of this report is the problem of obtaining confidence bounds for the population total error D when sampling from the model (2.2.7).

A useful sampling design for statistical auditing is to select items without replacement with probability proportional to book values. This sampling design can be modeled in terms of use of individual dollars of the total book amount as sampling units and is commonly referred to as *Dollar Unit Sampling* (DUS) or *Monetary Unit Sampling* (MUS). (Anderson and Teitlebaum, 1973; Roberts, 1978; Leslie, Teitlebaum and Anderson, 1980). The book amounts of the N items are successively cumulated to a total of Y dollars. One may then choose systematically n dollar units at fixed intervals of I $(= Y/n)$ dollars. The items with book amounts exceeding I dollars, and hence items that are certain to be sampled, are separately examined. Items with a zero book amount should also be examined separately as they will not be selected. If a selected dollar unit falls in the ith item, the tainting T_i $(= D_i/Y_i)$ of the item is recorded. Namely, every

dollar unit observation is the tainting of the item that the unit falls in. The model (2.2.7) may then be applied for DUS by considering d as an independent observation of tainting of a dollar unit. p is, then, the probability that a dollar unit is in error. Thus, (2.2.7) can be used for sampling individual items or individual dollars. In the former, d stands for the error amount of an item, and in the latter for the tainting of a dollar unit.

In the next section we present some results from several empirical studies to illustrate values of p and the distribution of z, for both line item sampling and DUS designs.

2.3 Error Distributions of Audit Populations—Empirical Evidence

Why do errors occur? Hylas and Ashton (1982) conducted a survey of the audit practices of a large accounting firm in order to investigate the kinds of accounts that are likely to show errors, to obtain alternative audit leads for detection of these errors and to attempt to identify their apparent causes. Not surprisingly, their study shows that unintentional human error is the most likely cause of recording errors. The remainder of this section reports the results of several empirical studies about actual values of the error rate p and actual distributions of the non-zero error z in the model (2.2.7). The sample audit populations are from a few large accounting firms and each contains a relatively large number of errors. Therefore, the conclusions may not represent typical audit situations.

A. *Line Item Errors.* Data sets supplied by a large accounting firm were studied by Ramage, Krieger and Spero (1979) and again by Johnson, Leitch and Neter (1981). The orientations of the two studies differ in some important respects. The latter provides more comprehensive information about the error amount distributions of the given data sets. It should be noted that the data sets are not chosen randomly. Instead, they have been selected because each data set contains a large number of errors, enough to yield a reasonable smooth picture of the error distribution.

According to the study by Johnson, Leitch and Neter (1981), the median error rate of 55 accounts receivables data is 0.024 (the quartiles are: $Q_1 = 0.004$ and $Q_3 = 0.089$). On the other hand, the median error rate of 26 inventory audits is 0.154 ($Q_1 = 0.073$ and $Q_3 = 0.399$). Thus the error amount distribution of a typical accounts receivable in their study has a mass 0.98 at zero. A random sample of 100 items from such a distribution will then contain, on the average, only two non-zero observations. On the other hand, the error amount distribution of a typical inventory in their study has a mass 0.85 at the original and sampling of 100 items from such a distribution will contain, on the average, 15 non-zero observations. The items with larger book amounts are more likely to be in error than those with smaller book amounts. The average error amount, however, does not appear to be related to the book amount. On the other hand, the standard deviation of the error amount tends to increase with book amount.

Ham, Losell and Smieliauskas (1985) conducted a similar study using data sets provided by another accounting firm. Besides accounts receivable and inventory, this study also included accounts payable, purchases and sales. Four error rates are defined and reported for each category of accounts. It should be noted that their study defines errors broadly, because they include errors that do not accompany changes in recorded amounts.

The distribution of non-zero error amounts again differs substantially between receivables and inventory. The error amounts for receivables are likely to be overstated and their distribution positively skewed. On the other hand, errors for inventory include both overstatements and understatements with about equal frequency. However, for both account categories, the distributions contain outliers. Graphs in Figure 1 are

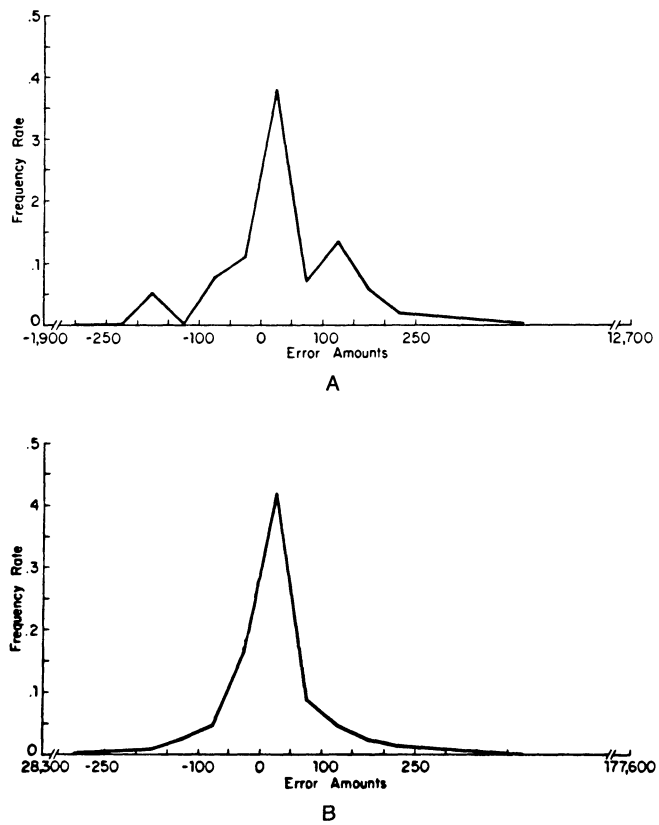

FIG. 1. *Examples of distribution of error amounts. A, accounts receivable audit* 69 (106 *observations*); *B, inventory audit* 23 (1139 *observations*). *From Johnson, Leitch and Neter* (1981).

taken from Johnson, Leitch and Neter (1981) and illustrate forms of the error amount distributions of typical receivables and inventory audit data. The figures show the nonnormality of error distributions.

Similar conclusions are also reached by Ham, Losell and Smieliauskas (1985). Their study also reports the distribution of error amounts for accounts payables and purchases. The error amounts tend to be understatements for these categories. Again, the shape of distributions are not normal.

B. *Dollar Unit Taintings.* When items are chosen with probability proportional to book amounts, the relevant error amount distribution is the distribution of taintings weighted by the book amount. Equivalently, it is the distribution of dollar unit taintings. Table 1 tabulates an example. Neter, Johnson and Leitch (1985) report the dollar unit tainting distributions of the same audit data that they analyzed previously. The median error rate of receivables is 0.040 for dollar units and is higher than that of line items (0.024). Similarly, the median dollar unit error rate for inventory is 0.186 (0.154 for line items). The reason is, they conclude, that the line item error rate tends to be higher for items with larger book amount for both categories. Because the average line item error amount is not related to the book amount, the dollar unit tainting tends to be smaller for items with larger book amounts. Consequently, the distribution of dollar unit tainting tends to be concentrated around the origin. Some accounts receivable have, however, a J-shaped dollar unit taint distribution with negative skewness.

TABLE 1

Illustration of dollar unit tainting distribution using a hypothetical accounting population of five items: the difference between the error amount distribution (A) and the tainting distribution (B) is illustrated

A. Composition of audit population

Line item (i)	Book values (Y_i)	Error (D_i)	Taint (T_i)
1	300	30	0.10
2	800	40	0.05
3	600	60	0.10
4	200	50	0.25
5	100	100	1.00
Total	2000	280	

B. Distribution of tainting

Tainting	Proportion line item	Proportion dollar unit
0.05	0.20	0.40
0.10	0.40	0.45
0.25	0.20	0.10
1.00	0.20	0.05
Total	1.00	1.00

One significant characteristic of the dollar unit tainting distribution that is common for many accounts receivable is the existence of a mass at 1, indicating that a significant proportion of these items has a 100% overstatement error. Such an error could arise when, for example, an account has been paid in full but the transaction has not been recorded. A standard parametric distribution such as normal, exponential, gamma, beta and so on, alone may not be satisfactory for modeling such distribution. Figure 2 gives the graphs of the dollar unit tainting distributions for the same audit data used in Figure 1. Note that the distribution of taintings can be skewed when that of error amounts is not. Note also the existence of an appreciable mass at 1 in the accounts receivable example. The situation here may be viewed as a nonstandard mixture in which the discrete part has masses at two points.

2.4 The Performance of Estimators Commonly Used for Human Populations When Applied to Accounting Populations

In this section we introduce, within the auditing context, the estimators commonly used in the survey sampling of human populations. We then review their relative performances when used in the sampling of accounting populations.

Suppose that a sample of n items is taken. We denote the book amount, audited amount, error amount and tainting of the kth item in the sample by analogous lower case letters, namely, y_k, x_k, $d_k = y_k - x_k$ and $t_k = d_k/y_k$ (if $y_k \neq 0$), respectively. Denote their sample means by $\bar{y}$, $\bar{x}$, $\bar{d}$ and $\bar{t}$, respectively. Many estimators of the population audited amount have been proposed. First of all, the mean-per-unit estimator is

$$\hat{X}_m = N\bar{x}. \tag{2.4.1}$$

We may also consider *auxiliary information estimators* to improve precision. For example, one may use the difference estimator

$$\hat{X}_d = Y - N\bar{d}. \tag{2.4.2}$$

Alternatively, one may use the ratio estimator

$$\hat{X}_r = Y(\bar{x}/\bar{y}). \tag{2.4.3}$$

Another possibility is a weighted average of $\hat{X}_m$ and either $\hat{X}_d$ or $\hat{X}_r$, namely,

$$\hat{X}_{w,1} = w\hat{X}_m + (1 - w)\hat{X}_d, \tag{2.4.4a}$$

or

$$\hat{X}_{w,2} = w\hat{X}_m + (1 - w)\hat{X}_r. \tag{2.4.4b}$$

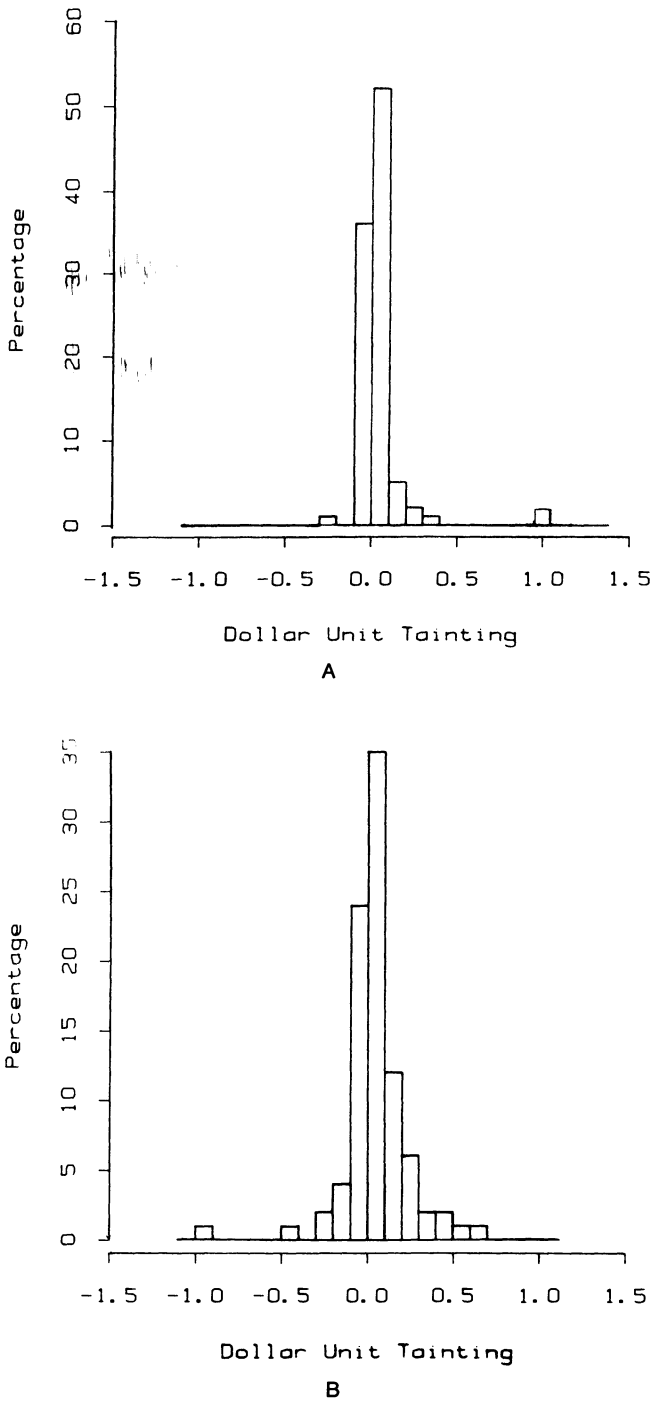

FIG. 2. *Examples of distribution of dollar unit tainting. A, accounts receivable audit 69; B, inventory audit 23. The histograms are drawn using the data in Table 2 of Neter, Johnson and Leitch* (1985).

One may also construct weighted combinations of $\hat{X}_m$, $\hat{X}_d$ and $\hat{X}_r$. Because the book amount of each item in the population is available, we may sample items with probability-proportional-to-size (PPS). As introduced in Section 2.2, this sampling method is commonly referred to as dollar unit sampling. We may then consider using an unbiased mean-per-dollar unit estimator

$$\hat{X}_{pps} = \left(\frac{Y}{n}\right) \sum_{k=1}^{n} \left(\frac{x_k}{y_k}\right). \tag{2.4.5}$$

The auditor's main concern is with the population total error amount D and the above estimators lead, respectively, to

$$\hat{D}_m = Y - \hat{X}_m = Y - N\bar{x}, \tag{2.4.6a}$$

$$\hat{D}_d = Y - \hat{X}_d = N\bar{d}, \tag{2.4.6b}$$

$$\hat{D}_r = Y - \hat{X}_r = Y(\bar{d}/\bar{y}), \tag{2.4.6c}$$

$$\hat{D}_{w,1} = Y - \hat{X}_{w,1} = w\hat{D}_m + (1 - w)\hat{D}_d, \tag{2.4.6d}$$

$$\hat{D}_{w,2} = Y - \bar{X}_{w,2} = w\hat{D}_m + (1 - w)\hat{D}_r \tag{2.4.6e}$$

and

$$\hat{D}_{pps} = Y - \hat{X}_{pps} = \left(\frac{Y}{n}\right) \sum_{k=1}^{n} \left(\frac{d_k}{y_k}\right) = Y\bar{t}. \tag{2.4.6f}$$

Note that the last estimator, $\hat{D}_{pps}$, may be interpreted as the mean error-per-dollar unit estimator using a random sample of n dollar units. These estimators may be used with stratification on book amount in more sophisticated sampling designs. Plots of the bivariate data, either (X_i, Y_i) or (D_i, Y_i) may have some value in selecting appropriate auxiliary information estimators.

In the remainder of the section we describe the performance of these estimators when they are used in sampling from the distributions characterized by the mixture defined by (2.2.7). For each estimator, its precision, measured by the square root of the mean squared error, as well as the confidence levels, based on normal approximations, of the associated two-sided confidence interval and upper and lower confidence bounds are stated.

In auditing practice, the performance of an upper or a lower confidence bound is often more meaningful than that of a two-sided confidence interval. For example, when estimating the audited amount of an asset, the auditor would like to know, with a known confidence level, the lower bound of the true asset amount because of the potential legal liability that may follow as a consequence of overstating the measure. He, therefore, will be concerned if the true level of confidence of the lower bound is actually lower than the supposed level because this implies that he is assuming greater risk than intended. On the other hand, when a government agency, such as the Internal Revenue Service, applies statistical auditing procedures when examining a firm's expense account, it is more concerned with estimating the upper (lower) bound of the audited amount (proposed adjustment), because it wants to avoid the failure to recognize allowable expenses because this could lead to

overassessment of taxes. In this case, what matters most is whether the actual confidence level of the upper (lower) bound of the audited amount (adjustment) is close to the supposed level. If the actual confidence level is substantially higher than the stated level, the agency is assuming a much lower risk than allowed by the policy.

With financial support by the American Institute of Certified Public Accounts and Touche Ross & Co. (an accounting firm in the United States) and with computing support by the University of Minnesota, Neter and Loebbecke (1975, 1977) conducted an extensive study to examine the performance of alternative estimators in sampling audit populations. An important feature of this study is that the populations used in the experiment were constructed from real audit data. The study essentially confirms the observations that Stringer reported in 1963. Namely, the estimators commonly used when sampling from human populations perform poorly and are inadequate for statistical auditing when the populations are contaminated by rare errors. First, as discussed in Section 2.1, they provide the auditor with no means to make inferences about the total error amount when all items in the sample are error-free. Second, they are either imprecise, as in the case of the mean-per-unit estimator, because the audited amount has a large standard deviation, or, as in the case of auxiliary information estimators, the confidence intervals and bounds may not provide planned levels of confidence for the sample sizes commonly used in auditing practice.

Table 2 gives a summary, based on the Neter-Loebbecke study, of the performance of the mean-per-unit estimator $\hat{X}_m$, the difference estimator $\hat{X}_d$, their combination $\hat{X}_{w,1}$ with $w = .1$ and the unbiased estimator $\hat{X}_{pps}$. The first three of these are under simple random sampling of items whereas the last one uses PPS. The ratio estimator $\hat{X}_r$ performed almost identically with the difference estimator and so it is not included. Note that the confidence level of the lower bound for the audited amount is the confidence level of the upper bound of the error amount because the latter is obtained by subtracting the former from the known book value. Their conclusions can be summarized as follows:

- The mean-per-unit estimator is imprecise compared to the difference estimator. Also when the population audit amount is highly skewed, the confidence interval does not provide the nominal level. The confidence level of the lower bound, however, is larger than the stated.
- The difference estimator is precise but produces a confidence interval that does not provide the nominal level of confidence when the error rate is low or the errors are overstatements. For overstatement errors this failure to meet the stated confidence level for the two-sided confidence interval is caused by that of the lower (upper) bound for the audited amount (the error amount). The upper (lower) bound for the audited amount (the error amount), however, is overly conservative.
- The combination of the two estimators alleviates the problem but does not provide a satisfactory solution to all cases. Also, finding a proper weight seems to present a problem.
- The performance of the unbiased estimator using PPS (DUS) is generally poor even for a high error rate. When errors are overstatements, the performance is almost identical with that of the difference estimator. Namely, the lower (upper) bound for the audited value (the error amount) does not provide the stated confidence level, whereas the upper (lower) bound is conservative.

As expected, stratification improves the precision of the mean-per-unit estimator dramatically, but not of the performance of the confidence interval of the auxiliary information estimators described above following (2.4.1).

The audit data sets used by Neter and Loebbecke were then made public, and several studies followed that extended their results. Burdick and Reneau (1978) applied PPS without replacement and studied the performance of other estimators. Baker and Copeland (1979) investigated the performance of the stratified regression estimator. Beck (1980) focused on the effect of heteroscedasticity on the reliability of the regression estimator. Frost and Tamura (1982) applied the jackknife to reduce bias in the standard error to improve the reliability of the ratio interval estimation. The main conclusion from these studies is that commonly used auxiliary information estimators provide precise point estimates, but their confidence intervals based on the asymptotic normality of the estimators do not provide confidence levels as planned, when used in sampling audit populations contaminated by rare errors.

Kaplan (1973a, b) was the first to recognize that the poor performance of auxiliary information interval estimation may be caused by the mixed nature of the audit population. He observed that the sampling distribution of the pivotal statistic (point estimator − true mean)/(standard deviation of estimator) may not follow the t-distribution for an auxiliary information estimator when sampling from nonstandard mixtures such as these arising in audit populations. Frost and Tamura (1986, 1987) extended Kaplan's observation and showed that the mixture may cause the population

TABLE 2

Precision and reliability of confidence interval of commonly used estimators for the audited amount (sample size = 100)

Population	Population error rates as percentages					Population	Population error rates as percentages				
	0.5	1	5	10	30		0.5	1	5	10	30
	A. The mean-per-unit estimator, $\hat{X}_m$						B. The difference estimator, $\hat{X}_d$				
1 (22)	24.1	24.0	—	24.0	23.9	1 (+/−)	0.1	0.2	0.4	0.6	0.9
	81.8	81.8	—	81.7	81.7		30.5	37.3	96.8	94.0	96.3
	100.0	100.0	—	100.0	100.0		38.0	61.5	97.2	99.7	99.5
	81.8	81.8	—	81.7	81.7		92.5	75.8	99.6	94.3	96.8
2 (3.5)	18.2	18.2	18.3	18.2	18.3	2 (+/−)	0.2	0.4	1.0	1.3	3.0
	93.7	93.7	93.7	93.7	93.0		31.2	41.8	82.3	97.2	95.5
	99.5	99.5	99.5	99.5	99.3		41.3	41.8	99.7	99.7	98.8
	94.2	94.2	94.2	94.2	93.7		89.9	100.0	82.6	97.5	96.7
3 (7.9)	35.7	35.7	35.7	35.8	36.1	3 (+)	0.1	0.1	0.1	0.2	0.4
	82.5	82.5	82.5	82.5	82.5		23.3	36.8	73.7	80.3	90.8
	99.7	99.7	99.7	99.7	99.7		23.3	36.8	73.7	80.3	91.2
	82.8	82.8	82.8	82.8	82.8		100.0	100.0	100.0	100.0	99.6
4 (3.3)	20.2	20.2	20.4	20.7	21.4	4 (+)	1.1	1.1	1.6	3.5	9.2
	92.7	92.7	92.3	92.8	93.2		21.2	30.0	58.2	62.0	74.8
	99.5	99.5	99.3	99.5	99.8		21.2	30.0	58.2	62.0	74.8
	93.2	93.2	93.0	93.3	93.4		100.0	100.0	100.0	100.0	100.0
	C. The combination, $\hat{X}_{w.1}$, of $\hat{X}_m$ and $\hat{X}_d$ with $w = .1$						D. Unbiased estimator with PPS, $\hat{X}_{pps}$ (mean-per-unit estimator with dollar unit sampling)				
1	2.4	2.4	—	2.5	2.5	1	0.2	0.2	—	0.6	1.0
	82.3	82.0	—	83.3	84.3		17.5	49.7	—	90.3	92.2
	100.0	100.0	—	99.8	99.8		20.2	60.7	—	99.7	99.8
	82.3	82.0	—	83.5	84.5		97.3	89.0	—	90.6	92.4
2	1.8	1.9	2.1	2.2	3.6	2	—	—	1.0	1.5	—
	93.8	94.2	94.3	94.5	93.8		—	—	80.2	94.5	—
	99.5	99.5	99.5	99.2	99.2		—	—	99.2	100.0	—
	94.3	94.7	94.8	95.3	94.6		—	—	81.0	94.5	—
3	3.6	3.6	3.6	3.6	3.6	3	0.1	—	0.3	0.5	0.9
	82.7	82.5	83.0	82.8	82.3		5.2	—	31.5	44.8	77.0
	99.7	99.7	99.7	99.7	99.5		5.2	—	31.5	44.8	77.0
	83.0	82.7	83.3	83.1	82.8		100.0	—	100.0	100.0	100.0
4	2.2	2.3	2.5	3.8	8.5	4	0.5	—	1.0	1.8	3.9
	93.7	94.0	94.7	94.8	78.7		30.7	—	69.7	86.5	94.5
	99.3	99.3	99.0	95.3	78.7		30.7	—	69.7	87.0	95.5
	94.4	94.7	95.7	99.5	100.0		100.0	—	100.0	99.5	99.0

Note: Population numbers correspond to the Neter-Loebbecke study populations. The error rates are adjusted within each population by randomly eliminating errors in the 30% error population. The number in the parenthesis is the skewness of the audited amount for (A) and the sign of the error amount for (B). For (A) through (D), the first entry is the root mean squared error of the estimator in terms of the percentage of the total audited amount. The second entry is the estimated true level of confidence of a two-sided 95.4% confidence interval. The third entry is the estimated true confidence level of a one-sided lower 97.7% bound. The fourth entry is the estimated true confidence level of a one-sided upper 97.7% bound computed from the second and the third entries. These estimates are based on 600 independent samplings of size 100 from the corresponding population. — indicates that the entry is not available from the study.

Sources: Neter and Loebbecke (1975): Tables 2.3, 2.8, 2.11, 2.14, 3.1, 3.2, 4.2, 4.5, 5.3, 5.5, 10.1 and 10.3. Also tables in the Appendix.

error distribution to be highly skewed, especially when the error rate is low and errors are overstatements, and that this population skewness causes in turn the sampling distribution of the pivotal statistic to be skewed in the opposite direction. Therefore, the auxiliary information interval estimation based on the asymptotic normality of the sampling distribution may perform poorly for the sample sizes used in statistical auditing.

Table 3, taken from Frost and Tamura (1987), gives estimates of the probability that the population error is outside of the upper and lower bounds, respectively,

TABLE 3
The reliability of the upper and the lower bounds of 95% two-sided confidence intervals for the difference estimator of the population error

Sample size	Population error rates as percentages				
	1	2	4	8	16
50	0.650	0.483	0.324	0.213	0.140
	0.000	0.000	0.000	0.001	0.002
100	0.485	0.329	0.216	0.145	0.097
	0.000	0.000	0.001	0.002	0.004
200	0.327	0.215	0.142	0.101	0.072
	0.000	0.001	0.002	0.004	0.007
400	0.219	0.144	0.101	0.070	0.055
	0.001	0.002	0.004	0.007	0.010
800	0.144	0.099	0.070	0.054	0.045
	0.002	0.004	0.007	0.010	0.013

Note: Exponential distribution is used to model overstatement errors. The first entry is the probability that the population error exceeds the upper bound of a two-sided 95% interval using normal approximation. The second entry is the probability that the population error is lower than the lower bound of the two-sided 95% interval. The standard errors of these estimates are within .001.

of the 95% two-sided confidence interval for a difference estimator that is appropriate under assumptions of normality. In this example, the errors are assumed to be overstatements and are sampled from an exponential distribution. If the normal approximation were good, each entry would be close to 0.025. It is clear from Table 3 that the unwarranted use of normal approximations results in two-sided confidence intervals that have unreliable upper bounds and overly conservative lower bounds. The population of accounts receivable often is contaminated by overstatement errors. As indicated before, the auditor is concerned with assuming a greater risk than intended in setting an upper bound of the population error amount in an asset account. Namely, he wants to control the risk of overstating the true asset amount which may lead to legal liability for the auditor. Much effort has therefore been expended on developing alternative methods that might provide more satisfactory upper bounds for the error amount of an accounting population with overstatement errors. In the next section, we will review some of these methods. The parallel problem of tightening the lower bound for the population error amount is equally important because of its implications for improving the efficiency of estimating the proposed adjustment of reported expenses to government agencies. However, this problem has been relatively overlooked by academic researchers, and merits much greater attention.

2.5 Confidence Bounds Using Attribute Sampling Theory

Beginning with this section, we survey alternative solutions that have been proposed for the problem of estimating total error amounts in accounting populations. There are two main approaches: (1) approaches that utilize attribute sampling theory and (2) approaches that utilize Bayesian inference. Combinations of approaches 1 and 2 have also been proposed. Other approaches include the modeling of the sampling distribution by distributions other than the normal.

A. *Line Item Attribute Sampling.* The simplest application of attribute sampling theory is to audit situations in which it is assumed that all errors are overstatements with the maximum size of the error amount equal to the book amount, namely, $0 \le D_i \le Y_i$ (or equivalently, $0 \le T_i \le 1$) for $i = 1, \cdots, N$. Let Y_L be the known maximum book amount, i.e., $Y_i \le Y_L$ for $i = 1, \cdots, N$. Because N is large relative to the sample sizes used in practice, we may propose the following model for random sampling of n items from the accounting population with or without replacement. Let p be the proportion of items with errors in the population. Then, for any item in the sample, the observed error,

$$(2.5.1) \qquad d = \begin{cases} z & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

Because $z \le Y_L$,

$$(2.5.1a) \qquad E(z) = \theta \le Y_L,$$

and

$$(2.5.1b) \qquad E(d) = \mu_D = p\theta \le pY_L.$$

Hence, the total error amount D as defined in (2.2.6) becomes

$$(2.5.1c) \qquad D = Np\theta \le NpY_L.$$

Suppose that a sample of n items contains m items with errors and $n - m$ error-free items. Let $\hat{p}_u(m; 1 - \alpha)$ be a $(1 - \alpha)$-upper confidence bound for p. That is,

$$(2.5.1d) \qquad \text{prob}\{p \le \hat{p}_u(m; 1 - \alpha)\} = 1 - \alpha.$$

We may use a binomial distribution or, when p is small, a Poisson distribution as widely practiced. Then a $(1 - \alpha)$-upper confidence bound for D is

$$(2.5.2) \qquad \hat{D}_u(m; 1 - \alpha) = N\hat{p}_u(m; 1 - \alpha)Y_L.$$

Because observed values of z are not used, the upper bound (2.5.2) can be overly conservative, and stratification on book amount may be used to tighten the bound (Fienberg, Neter and Leitch, 1977).

B. *Dollar Unit Attribute Sampling.* Let us suppose that the sample is taken by means of Dollar Unit Sampling (Section 2.2). As before, all errors are assumed to be overstatements with the maximum size of the error amount equal to the book amount. Thus $0 \le T_i \le 1$. The model (2.5.1) may then be applied for the analysis of the DUS data by considering d as an independent observation of dollar unit tainting with $0 < z \le 1$. θ is the mean dollar unit tainting. In this case, p is then the proportion of dollar units in error and is equal to Y_d/Y, where Y_d is the total book amount of items in error. Thus, in this case,

$$D = Yp\theta \le Yp.$$

Given m nonzero tainting observations, a $1 - \alpha$ upper bound for the total error amount D is

$$\hat{D}_{u,\mathrm{DUS}}(m, 1 - \alpha) = Y\hat{p}_u(m; 1 - \alpha). \tag{2.5.3}$$

Because $Y \le NY_{\mathrm{L}}$,

$$\hat{D}_{u,\mathrm{DUS}} \le \hat{D}_u.$$

The confidence level of the bound (2.5.3) is at least $1 - \alpha$ but the bound is still conservative because it assumes that all taintings are equal to 1.

C. *The Stringer Bounds.* When $m \ge 1$, we have dollar unit taintings z_j, $j = 1, \cdots, m$, in the sample that may be less than 1 and this information can be used to tighten the bound. There are several ways of accomplishing this task and these procedures are commonly referred to as *combined attributes and variables* (CAV) *estimation* (Goodfellow, Loebbecke and Neter, 1974). The best known and most widely used CAV estimation is credited to Stringer and is called the *Stringer bound.* Let $0 < z_m \le \cdots \le z_1 \le 1$ be ordered observations from a random sample of size m from z. Then a $(1 - \alpha)$-Stringer bound is defined by

$$\hat{D}_{u,\mathrm{st}} = Y\Big\{\hat{p}_u(0; 1 - \alpha) + \sum_{j=1}^{m} [(\hat{p}_u(j; 1 - \alpha) - \hat{p}_u(j - 1; 1 - \alpha)]z_j\Big\}. \tag{2.5.4}$$

Here, $\hat{p}_u(0; 1 - \alpha)$, being the $(1 - \alpha)$ upper bound for the error rate p when the sample contains no error, is equal to $1 - \alpha^{1/n}$ using the binomial distribution. The Poisson approximation $3/n$ is widely used among practitioners. When $z_j = 1$ for all j, (2.5.4) reduces to (2.5.3).

It is as yet unknown whether the Stringer bound always provides the confidence level at least as large as the nominal for overstatement errors. Indeed, commonly used CAV estimations are heuristic and it is difficult to determine theoretically their sampling distributions. Many simulation studies, however, have been performed using various error distributions ranging from the populations used by Neter and Loebbecke (1975) to standard parametric models. These studies have provided strong empirical evidence that the confidence level of the Stringer bound is at least the nominal level. In fact these studies indicate that it may be overly conservative (see, for example, Leitch, Neter, Plante and Sinha, 1982; Roberts, Shedd and MacGuidwin, 1982; Reneau, 1978). However, the formulation of the Stringer bound has never been satisfactorily explained. Not even an intuitive explanation can be found in auditing literature.

For audit populations contaminated by low error amounts, the bound may grossly overestimate the total error amount, causing the auditor to conclude that the total book amount contains a material error when it does not. The ensuing activities, e.g., taking additional samples, or requesting the client to adjust the account balance, etc., may be costly to the client. The development of a more efficient CAV estimation procedures has thus become an important research goal.

D. *Multinomial Bounds.* One approach toward this goal, involving an interesting application of attribute sampling theory for estimation of the total error amount, has been developed by Fienberg, Neter and Leitch (Fienberg, Neter and Leitch, 1977; Neter, Leitch and Fienberg, 1978). Because their model uses the multinomial distribution, the resulting upper bound is commonly called a *multinomial bound.* The observation of dollar unit taint is categorized, for example, in 101 classes ranging from 00 to 100 cents. Let p_i be the probability that an observation on d falls in the ith category (i cents), where $0 \le p_i < 1$ and $\sum p_i = 1$. Then, instead of (2.5.1), we have

$$d = \frac{i}{100} \quad \text{with probability } p_i, \qquad i = 0, \cdots, 100. \tag{2.5.5}$$

Then

$$E(d) = \mu_D = \sum_{i=0}^{100} \frac{i}{100} p_i, \tag{2.5.5a}$$

so that

$$D = Y\mu_D = Y \sum_{i=0}^{100} \frac{i}{100} p_i. \tag{2.5.5b}$$

Let w_i be the number of observations in a sample of n dollar units that fall in the ith category; $\sum w_i = n$. Clearly $\mathbf{w} = (w_0, w_1, \cdots, w_{100})$ follows a multinomial distribution with the parameters $(n, \mathbf{p})$, $\mathbf{p} = (p_0, p_1, \cdots, p_{100})$, if the sampling is done with replacement. (If the sampling is done without replacement, then this may still be used as an approximate model.) The sample mean is

$$\bar{z} = \hat{\mu}_D = \sum_{i=0}^{100} \frac{i}{100} \frac{w_i}{n} \tag{2.5.6}$$

and a point estimate of D is given by

$$\hat{D} = Y\hat{\mu}_D.$$

The following procedure is then proposed for an upper bound for μ_D and hence for D. Let S be a set of outcomes $\mathbf{v} = (v_0, \cdots, v_{100})$ that are "as extreme as or less extreme than" the observed results $\mathbf{w}$. The concept of extremeness must be specified. Clearly, S is not unique and computational simplicity must be taken into account for its definition. The particular S proposed by Fienberg, Neter and Leitch, called the step down S, is the set of outcomes such that (1) the number of errors does not exceed the observed number of errors and (2) each error amount does not exceed any observed error amount. Given this step down S, a $(1 - \alpha)$-joint confidence set for $\mathbf{p}$ is determined by those values of p_i that satisfy

$$(2.5.7) \quad \sum_S \frac{n!}{v_0! \cdots v_{100}!} \prod_{i=0}^{100} p_i^{v_i} \geq \alpha, \quad \sum v_i = n.$$

A $(1 - \alpha)$-upper bound for θ, and hence also for D by multiplying the former by Y, is obtained by maximizing (2.5.5a) over those $\mathbf{p}$ defined by (2.5.7). However, the true level of confidence of the multinomial bound using the step down S is not known.

When the sample does not contain any errors, the multinomial bound is the same as the bound given in (2.5.3). When the sample contains errors, the multinomial bound is considerably tighter than the widely used Stringer bound. However, the computation of the bound quickly becomes unmanageable as the number of errors increases. Neter, Leitch and Fienberg (1978) reported that the array size of the step down S set is 429×9 for 6 errors, 1430×9 for 7 and 4862×10 for 8 errors. When using a computer of the size of a large IBM 370, they only did the computations for up to 7 errors. Software for personal computers that will compute the bound for up to 10 errors is now available from Plante (1987).

Leitch, Neter, Plante and Sinha (1981, 1982) propose to cluster the observations for improving the computational efficiency of the multinomial bound for many errors. The observed errors d_i are grouped into g groups of similar sizes. Then all errors in the same cluster are given the maximum error amount of the cluster. In forming an optimum set of clusters, the algorithm that minimizes

$$(2.5.8) \qquad C = \sum_k (\max_j (d_{k_j}) - d_{k_j}),$$

where d_{k_j} is the jth tainting in the kth cluster, is recommended. The loss of efficiency due to this clustering cannot easily be assessed because the bound has not been computed without the grouping of observations when the number of errors are many. Leitch, Neter, Plante and Sinha (1981) reports, however, that with 20 to 25 errors in the sample, the multinomial bound with five to six clusters still compares favorably with the Stringer bound.

Plante, Neter and Leitch (1985) provides a study that compares the multinomial upper bound with the two CAV upper bounds, i.e., the Stringer and cell bounds (Leslie, Teitlebaum and Anderson, 1980). The multinomial bound is the tightest of the three and the observed confidence level is not significantly different from the nominal level 0.95 used for the study.

If the auditor knows the maximum size of the understatement error, it is possible to apply the multinomial approach to set the upper bound. Also, although conservative, a lower bound can be set (see Neter, Leitch and Fienberg, 1978).

2.6 Other Developments for the Analysis of Dollar Unit Sample Data

In this section, we give a brief account of some other approaches to the statistical analysis of dollar unit sample data. Firstly, there are proposals for approximating the sampling distribution of the mean tainting using models other than normal distributions. Secondly, in order to set a tighter bound, the use of parametric models for the distribution of tainting has been suggested. We will discuss these attempts below.

A. *Nonnormal Sampling Distributions.* Along the line of improving the reliability of large-sample, classical interval estimation, Garstka and Ohlson (1979) suggested a modification of the constant by which the estimated standard deviation is multiplied to make it dependent upon the number of errors found in the sample. Tamura (1985) comments, however, that this modification does not take into account the skewness of the sampling distribution and may not always produce a satisfactory result. Dworin and Grimlund (1984) propose approximating the sampling distribution by a three-parameter gamma distribution. The method of moments is used to estimate the parameter values of the approximating gamma distribution. A number of heuristics are invoked including the introduction of a 'hypothetical tainting observation' for computing the sample moments. The method of computing this added data point varies slightly depending on whether the audit population is accounts receivables or inventory. The method can handle both over- and understatement errors, however. Through extensive simulation tests, they show that the upper bound computed by their method provides the confidence level close to the stated. Moreover, they show that the moment upper bound is about as tight as the multinomial upper bound.

B. *Parametric Models.* A natural way to improve the efficiency of a bound is to describe the error

distribution using a parametric model, following Aitchison (1955). This point was illustrated by Garstka (1977a, b). By treating an observation of dollar unit tainting in terms of smaller units, say 10-cent units, he uses a geometric distribution as a model. A parametric bound can be sensitive to the choice of the model. Lillestol (1981), using a logarithmic series distribution instead of a geometric distribution, demonstrates the point. Recently, Tamura and Frost (1986) have proposed a power function density to model taintings. Efron's parametric bootstrap is used to approximate the sampling distribution of the estimator for setting the bound. The study shows that the bootstrap bound is reliable and much tighter than the nonparametric Stringer bound when the data are generated from the correct specification. Research to compare performance of different models of tainting, including their robustness to parametric assumptions, may prove to be fruitful for achieving economical and efficient auditing.

2.7 Bayesian Models for the Analysis of Audit Data

From the discussions presented so far, we may summarize two basic problems associated with statistical auditing. Firstly, it is difficult to determine the small sample sampling distribution of the estimator when the population is characterized by the mixture. Secondly, because of the low error rate the sample does not provide sufficient information about the characteristics of the nonzero error distribution of the audit population to set a good bound.

Earlier in Section 2.3, this report reviewed the results of empirical studies of the error distribution of various accounting populations. These results have provided auditors with considerable evidence about the audit environment. Using such results, an auditor may make a more intelligent prediction about the error distribution of certain audit populations. By incorporating this prior information into the analysis of the sample data, the auditor should usually be able to obtain a more efficient bound for a total population error. Bayesian inference provides a useful framework to incorporate the auditor's informed judgment with the sample information and we will review in this section developments among this line of audit data analysis. Empirical Bayes methods do not seem to have been used on these problems, a direction that may be worth investigation.

A. *Normal Error Models.* Felix and Grimlund (1977) propose a parametric Bayesian model. They assume audit item sampling but their model has also been applied for dollar unit sampling by Menzefricke and Smieliauskas (1984). In their formulation the error amount z in the sampling model (2.5.1) is assumed to be normally distributed dependent on the mean μ_Z and the precision h, the inverse of the variance σ^2. μ_Z is given a normal prior that depends on h. h is given a gamma prior. The joint distribution of μ_Z and h is often referred to as a normal gamma distribution. The prior distribution for the error rate p is given a beta distribution and is independent of (μ_Z, h). These prior specifications are conjugate with the likelihood function of the data as determined by the sampling model, i.e., the posterior distribution of (μ_Z, h) is again a normal gamma and that of p is beta. The two posterior distributions are again independent.

In order to develop the posterior distribution for the population mean $\mu_D = p\mu_Z$, first, h is integrated out from the posterior distribution of (μ_Z, h), resulting in a Student distribution for the marginal distribution of μ_Z. Because $\mu_D = p\mu_Z$, substituting μ_Z with μ_D/p in the marginal distribution, and integrating out p, the posterior distribution for μ_D is obtained. The result of this integration cannot be written explicitly and has to be numerically obtained. However, the expected value and the variance can be derived. (Felix and Grimlund (1977) derive the posterior distribution by using a different approach than described here. However, Menzefricke and Smieliauskas (1984) show that in their approach, a source of variability has been overlooked, which leads to a smaller variance than should be the case.)

B. *Infinite Population Models.* The probability models that underly the methods discussed so far may be referred to as *finite population models.* By this nomenclature one stresses the fact that the lists of book values $Y_1, \cdots, Y_N$ and audited amounts $X_1, \cdots, X_N$ that are associated with the financial statements at the time of audit, are finite in number and considered fixed. There is no randomness associated with these values. Randomness, and hence the need for a probability model, enters only by the act of sampling n book values from the population's N values. The method of sampling determines which probability model must be considered.

One may consider other probability models in which the Y_i's and X_i's are themselves viewed as having been randomly generated. For example, at the start of a fiscal year the line items in a company's books are yet to be determined, and from that point in time it might be appropriate to view them as random variables subject to some probability distribution. To contrast them with the sampling models previously discussed, these globally random models can be referred to as *infinite population models.*

In 1979, Cox and Snell proposed such an infinite population model. The importance of their proposal is that it provides a theoretical basis for DUS methodology, something that had not previously been available. In their method, the account entries, Y_i, and their correct values X_i, are viewed as being outcomes

from N *independent* repetitions of an experiment whose probabilistic description is therefore completely specified by the common joint distribution function of the pairs (Y_i, X_i) for $i = 1, \cdots, N$. Equivalently, one could specify the joint distribution of (Y_i, D_i) because $Y_i - D_i = X_i$.

As stressed before, a large portion of the errors D_i will be zero in auditing situations. Consequently, when modeling the distribution functions of the D_i, one should use distributions that place a large probability at zero. One way to think of the generation of such errors is in two stages as follows: first, determine whether there is in fact an error, and then, if there is an error, determine its value. More specifically, introduce δ_i to be 0 or 1 according as to whether or not $D_i = 0$. To describe the model for (Y_i, δ_i, D_i), one may begin by specifying the marginal distribution of Y_i, F_Y say; then the conditional distribution of $\delta_i = 1$ given Y_i, $p(Y_i)$ say; and then the conditional distribution of D_i given Y_i and $\delta_i = 1$, $F_{D|Y}$ say. The conditional distribution of D_i given Y_i and $\delta_i = 0$ is degenerate at zero by definition because D_i is zero when δ_i is zero.

If both F_Y and $F_{D|Y}$ are assumed to have densities, $f_Y(y)$ and $f_{D|Y}(d, y)$ say, then the triple (Y_i, δ_i, D_i) has the density function $f_{Y,\delta,D}$ defined by

$$f_{Y,\delta,D}(y, 0, 0) = f_Y(y)\{1 - p(y)\}, \tag{2.7.1a}$$

and

$$f_{Y,\delta,D}(y, 1, d) = f_Y(y)p(y)f_{D|Y}(d, y). \tag{2.7.1b}$$

Although this model allows for the probability of an error to depend on the magnitude of the observed book value, the most common practice is to assume that this error probability $p(y)$ is a constant.

In order to complete the description of the probability model, it remains to specify how the sampling selection is determined. For this, Cox and Snell (1979) introduce a "sampling" variable, S_i say, which equals 1 if the ith item is to be sampled, and 0 otherwise. In general, one would then specify the conditional distribution of S_i, given the other observations Y_i, δ_i and D_i. An important special case is that of probability-proportional-to-size sampling and for this, one would set

$$\text{prob}(S_i = 1 \mid Y_i, \delta_i, D_i) = cY_i. \tag{2.7.2}$$

Under the PPS sampling design, there is a particularly simple relationship between the conditional means of the taintings, $T_i = D_i/Y_i$, and the ratio of conditional means of the errors D_i to the book amounts Y_i. It can be shown, by straightforward manipulation of conditional expectations, that as a consequence of (2.7.2)

$$E\left(\frac{D_i}{Y_i} \,\middle|\, S_i = 1, \delta_i = 1\right) = \frac{E(D_i \mid \delta_i = 1)}{E(Y_i \mid \delta_i = 1)}. \tag{2.7.3}$$

Equation (2.7.3) is of fundamental importance in this model. From it, one can derive an expression, (2.7.7b) below, for the population mean error in terms of the population error rate and the conditional mean of the tainting. This is the relationship that is used in the analysis of DUS data. To do this, begin by multiplying both numerator and denominator of the ratio in (2.7.3) by $\text{prob}(\delta_i = 1)$. The ratio becomes $E(D_i\delta_i)/E(Y_i\delta_i)$, where we have made use of the zero-one nature of δ_i. If Z denotes a random variable whose distribution is the same as the conditional distribution of the tainting $T_i = D_i/Y_i$, given that it is nonzero ($\delta_i = 1$) and is sampled, ($S_i = 1$), then the lefthand side of (2.7.3) is the mean of Z, say μ_Z. Thus, (2.7.3) may be written as

$$E(D_i\delta_i) = E(Y_i\delta_i)\mu_Z. \tag{2.7.4}$$

Moreover, if $\mu_D = E(D_i)$ *denotes the mean error,*

$$\mu_D = E(D_i\delta_i) + E\{D_i(1 - \delta_i)\} = E(D_i\delta_i), \tag{2.7.5}$$

because $D_i = 0$ when $\delta_i = 0$. Now introduce p_S to be the probability of an item being in error given that the item is sampled; that is, $p_S = \text{prob}(\delta_i = 1 \mid S_i = 1)$. By (2.7.1b) and (2.7.3), direct computation yields

$$\begin{aligned} p_S &= \frac{E\{cY_ip(Y_i)\}}{c\mu_Y} \\ &= \frac{E\{E(Y_i\delta_i \mid Y_i)\}}{\mu_Y} = \frac{E(Y_i\delta_i)}{\mu_Y}. \end{aligned} \tag{2.7.6}$$

Upon substitution of this in (2.7.4) one obtains the important relationship

$$\mu_D = \mu_Y p_S \mu_Z. \tag{2.7.7a}$$

In many situations it is assumed that $p(y)$, the conditional probability that an error is present in a book amount of magnitude y, is a constant, say p. In this case, Y_i and δ_i are independent so that $p_S = p$ and $E(Y_i \mid \delta_i = 1) = \mu_Y$. Then, (2.7.7a) becomes

$$\mu_D = \mu_Y p \mu_Z. \tag{2.7.7b}$$

It should be noted, however, that the empirical evidence reported in Section 2.3 indicates that the assumption of constant $p(y)$ may not be justified in all cases.

Relations involving higher moments of the taintings may be derived for this model by similar analyses. In particular, for $k = 1, 2, \cdots$, the analogue of (2.7.3) is

$$E(Z^k) = \frac{E(D_i^k/Y_i^{k-1})}{p_S\mu_Y}, \tag{2.7.8}$$

from which information about the variance and skewness of Z can be obtained.

The proportionality constant c in (2.7.2) satisfies $\text{prob}(S_i = 1) = c\mu_Y$, where $\mu_Y = E(Y_i)$ for all i. The

number of line items sampled n in this infinite population model is the *random* quantity $S_1 + \cdots + S_N$. Thus n is a binomial $(N, c\mu_Y)$ random variable with expected value N prob$(S_i = 1) = Nc\mu_Y$. Thus c plays the role of determining the sample size. The important difference is that whereas n is fixed in the finite population model, the sample size in this infinite model is random. If prob$(S_i = 1)$ is small while N prob$(S_i = 1)$ is moderate, a Poisson $(Nc\mu_Y)$ approximation might be used for the exact binomial distribution, as was done by Cox and Snell.

Suppose that a PPS sample of n items contains m items with errors and $n - m$ items error-free. Let $\mathbf{z} = (z_1, \cdots, z_m)$ be taintings of the m items with error. For estimation purposes, let us set $\hat{p} = m/n$ and $\bar{z} = \sum z_i/m$. In view of (2.7.5), a natural point estimate of the population total error is then

$$\hat{D} = N\mu_Y\hat{p}\bar{z}. \tag{2.7.9}$$

The known total book amount Y is used to estimate $N\mu_Y$ in practice. This can also be viewed as fitting the infinite model to the existing finite audit population. Using Y to stand for the known total book amount as defined in Section 2.2, we get

$$\hat{D}_{CS} = Y\hat{p}\bar{z}. \tag{2.7.10}$$

C. *The Cox-Snell Parametric Models.* Cox and Snell proceed to formulate a parametric Bayesian model in a fashion similar to the Grimlund and Felix model. The prior distribution of p is specified by a gamma distribution with parameters a/p_0 and a, where p_0 is the prior mean of p. Z is assumed to have an exponential density and its parameter $1/\mu_Z$ has also a gamma distribution $\{(b-1)\mu_0, b\}$, where μ_0 is the prior mean of μ_Z. These two prior distributions are assumed independent. Then it can be shown that the posterior distribution of μ_D is a scalar transformation of an F distribution. Specifically (see Cox and Snell, 1982; Moors, 1983), if $=_L$ indicates equality of probability laws, and if F_{v_1,v_2} denotes a random variable having an F distribution with the numbers of degrees of freedom v_1 and v_2,

$$\mu_D =_L \frac{m\bar{z} + (b-1)\mu_0}{n + a/p_0}\,\frac{m+a}{m+b}\,F_{2(m+a),2(m+b)}. \tag{2.7.11}$$

Godfrey and Neter (1984) investigate the sensitivity of the Cox-Snell bound to its parametric assumptions. For example, because $0 \le p \le 1$, the effects of truncating the gamma prior for p at 1 as well as replacing it with the beta prior are investigated. Similarly, the effects of truncating the exponential distribution at 1 for Z are considered. Because the distribution of tainting often has a discrete mass at 1, its effect is also studied. For these moderate parametric modifications, the bound appears relatively stable compared to the effect of the prior parameter settings on the bound. The practitioners' interest may, however, lie in the sensitivity of the performance of the bound to the prior parameter settings under repetitive sampling. Their study, using 21 hypothetical audit populations, shows that the reliability of the Cox-Snell bound is, as expected, sensitive to changes in prior parameter values but that it is possible to set these values conservatively so that the bound has a confidence level close to the nominal and still tighter than the Stringer bound for this set of study populations. Neter and Godfrey (1985), extending their earlier study (Godfrey and Neter, 1984), show that the sensitivity of the Cox-Snell bound to parameter settings does not disappear when the sample size is increased to 500, a size seldom exceeded in current audit practice. (The sample size of 100 is used in their previous study.) The study goes on to use another set of 9 study populations to identify some conservative prior parameter settings for which the bound is reliable and tighter than the Stringer bound.

Menzefricke and Smieliauskas (1984) investigated the gain in tightness resulting from parametric modeling of tainting. The performance of Bayesian parametric bounds is compared with that of the Stringer bound and other CAV bounds. The Bayesian bounds include the Cox-Snell bound and two versions of the normal error model introduced earlier. Only one parameter setting is used for each model. Their study uses audit populations contaminated by both positive and negative taintings. Because the Cox-Snell model assumes taintings to be positive, *ad hoc* adjustments are tried. Using simulation, they show that the Bayesian bound, using the normal distribution to model errors, outperforms both CAV bounds.

D. *Nonparametric Bayesian Models.* The empirical evidence reported in Section 2.3 shows that standard distributions may not work for modeling the distribution of dollar unit tainting. A Bayesian nonparametric approach may then provide a necessary flexibility for modeling of available audit information. In Section 2.5 an example of a nonparametric error distribution was introduced, where the dollar unit tainting of an item in the sample is treated as a random observation from a discrete distribution (i, p_i) for $i = 0, \cdots, 100$ in cents and $p_i \ge 0$, $\sum p_i = 1$. Tsui, Matsumura and Tsui (1985) propose the Dirichlet distribution to incorporate the auditor's prior prediction of the unknown p_i. In their model the auditor is assumed to provide the best prediction $\rho = (\rho_0, \cdots, \rho_{100})$ of $\mathbf{p} = (p_0, \cdots, p_{100})$ and a weight K for the prediction. It is then suggested that the prior distribution of $\mathbf{p}$ is a Dirichlet $(K\rho)$. The distribution of p_i is thus a beta $\{K\rho_i, K(1-\rho_i)\}$ with

$$\mathrm{E}(p_i) = \rho_i \tag{2.7.12a}$$

and

$$\text{(2.7.12b)} \qquad \text{var}(p_i) = \frac{\rho_i(1-\rho_i)}{K+1}.$$

Let $\mathbf{w} = (w_0, \cdots, w_{100})$ with $\sum w_i = n$ be the sample data of n items. $\mathbf{w}$ is distributed as a multinomial distribution $(n, \mathbf{p})$ when sampling is with replacement (if sampling is without replacement, approximately). Because the Dirichlet prior distribution is conjugate with the multinomial sampling model, the posterior distribution of $\mathbf{p}$ is again a Dirichlet distribution with the parameter $(K\rho + \mathbf{w})$. We may define

$$\text{(2.7.13a)} \qquad K' = K + n,$$

$$\text{(2.7.13b)} \qquad \hat{p}_i = \frac{w_i}{n},$$

and

$$\text{(2.7.13c)} \qquad \rho_i' = \frac{(K\rho_i + n\hat{p}_i)}{K'}, \quad i = 0, \cdots, 100.$$

Then the posterior distribution of $\mathbf{p}$ is Dirichlet $(K'\rho')$, where $\rho' = (\rho_0', \cdots, \rho_{100}')$. By the definition of μ_D,

$$\text{(2.7.14)} \qquad \mu_D = \sum_{i=0}^{100} \frac{i}{100}\rho_i,$$

the posterior distribution of μ_D is derived as a linear combination of ρ_i'. It can be shown that

$$\text{(2.7.15a)} \qquad E(\mu_D) = \sum_{i=0}^{100} \frac{i}{100}\rho_i',$$

and

$$\text{(2.7.15b)} \qquad \text{var}(\mu_D) = \frac{1}{K'+1}\left\{\sum_0^{100}\left(\frac{i}{100}\right)^2\rho_i' - \left(\sum_{i=0}^{100}\frac{i}{100}\rho_i'\right)^2\right\}.$$

The exact distribution of μ_D is complicated and therefore is approximated by a beta distribution having the same mean and the variance. Using simulation, Tsui, Matsumura and Tsui (1985) suggest that $K = 5$, $\rho_0 = 0.8$, $\rho_{100} = 0.101$ and remaining 99 ρ_i's being 0.001 be used as the prior setting for their upper bound to perform well under repeated sampling for a wide variety of tainting distributions.

McCray (1984) suggests another nonparametric Bayesian approach using the multinomial distribution as the data-generating model. In his model, μ_D has been discretized, involving a number of categories, say μ_{D_j}, $j = 1, \cdots, N_\mu$. The auditor is to provide his assessment of the prior distribution by assigning probabilities q_j to the values, μ_{D_j}. Then the posterior distribution of μ_D is determined to be

$$\text{(2.7.16)} \qquad \text{prob}(\mu_D = \mu_{D_j} \mid \mathbf{w}) = \frac{q_j L(\mathbf{w} \mid \mu_{D_j})}{\sum q_h L(\mathbf{w} \mid \mu_{D_h})}$$

where

$$\text{(2.7.17)} \qquad L(\mathbf{w} \mid \mu_{D_j}) = \max \prod_{i=0}^{100} p_i^{w_i},$$

in which the maximum is taken over all probabilities $\{p_i\}$ satisfying

$$\text{(2.7.18)} \qquad \sum_{j=1}^{N_\mu} \mu_{D_j} p_j = \mu_D.$$

It should be noted that the two nonparametric models introduced above can incorporate negative taintings; that is, the auditor defines any finite lower and upper limits for tainting and divides the sample space into a finite categories.

Simulation studies have been performed to compare performances of these Bayesian bounds with the procedures described in earlier sections. Dworin and Grimlund (1986) compares the performance of their moment bound with that of McCray's procedure. Several Bayesian and non-Bayesian procedures are also compared in Smieliauskas (1986). Grimlund and Felix (1987) provide results of an extensive simulation study that compares the long run performances of the following bounds: Bayesian bounds with normal error distribution as discussed in A above, the Cox and Snell as discussed in C, the bound of Tsui, Matsumura and Tsui as discussed in D and the moment bound discussed in Section 2.6.

Recently, Tamura (1988) has proposed a nonparametric Bayesian model using Ferguson's Dirichlet process to incorporate the auditor's prior prediction of the conditional distribution of the error. It is hypothesized that the auditor cannot predict the exact form of the error distribution, but is able to describe the expected form. Let $F_0(z)$ be the expected distribution function of z representing the auditor's best prior prediction. The auditor may use any standard parametric model for F_0. Alternatively, F_0 may be based directly on past data. The auditor assigns a finite weight α_0 to indicate his uncertainty about the prediction. Then the auditor's prior prediction is defined by the Dirichlet process with the parameter

$$\text{(2.7.19)} \qquad \alpha(z) = \alpha_0 F_0(z).$$

This means that $\text{prob}(z \le z')$ is distributed according to the beta distribution beta $\{\alpha(z'), \alpha_0 - \alpha(z')\}$. The posterior prediction given m observations on z, say $\mathbf{z} = (z_1, \ldots, z_m)$, is then defined by the Dirichlet

process with the parameter

$$\alpha(z \mid \mathbf{z}) = \{\alpha_0 + m\} \cdot \{w_m F_0 + (1 - w_m)F_m\}(z), \tag{2.7.20}$$

where

$$w_m = \frac{\alpha_0}{\{\alpha_0 + m\}} \tag{2.7.21}$$

and $F_m(z)$ is the empirical distribution function of z. The distribution function of the mean θ of z is given by

$$G_\theta(v) = \text{prob}(\theta \le v) = \text{prob}(T^{(v)} \le 0), \tag{2.7.22}$$

where the characteristic function of $T^{(v)}$ is

$$\phi(v)(u) = \exp\left[-\int_{-\infty}^{\infty} \log\{1 - iu(t - v)\}\, d\alpha(t)\right]. \tag{2.7.23}$$

The distribution of θ is obtained by numerical inversion of (2.7.23). The distribution function of the mean tainting μ is, then, given by

$$H_\mu(d) = \text{prob}(\mu \le d) = \text{prob}(p\theta \le d) = E(\theta \le d/p \mid p). \tag{2.7.24}$$

This integration can be done numerically. In this work a beta distribution is proposed to model p.

2.8 Numerical Examples

In Section 2.5 through 2.7 various methods for setting a confidence bound for the accounting population error were described. They differ from the classical methods of Section 2.4 in the sense that these methods do not assume that the sampling distributions of their estimators are normal. Among these new developments, we illustrate in this section the computation of the following upper bounds for the total population error: the Stringer bound, the multinomial bound, parametric bounds using the power function and the moment bound. In addition, computation of two Bayesian models developed by Cox and Snell and Tsui, Matsumura and Tsui will also be illustrated. Software for computing all but one of these bounds can be developed easily. The exception is the multinomial bound, which requires extensive programming unless the number of errors in the sample is either 0 or 1. These methods are designed primarily for setting an upper bound of an accounting population error contaminated by overstatements in individual items. The maximum size of the error amount of an item is assumed not to exceed its book amount. These methods also assume DUS. Under this sampling design the total population error amount is equal to the known book amount Y times the mean tainting per dollar unit $\mu_D = p\mu_Z$. We will, therefore, demonstrate the computation of a 95% upper bound for μ_D using each method. The data used for these illustrations are hypothetical. Our main objectives are to provide some comparisons of bounds using the same audit data and also to provide numerical checks for anyone who wishes to develop software for some of the bounds illustrated in this section.

A. *No Errors in the Sample.* When there are no errors in a sample of n dollar units, the Stringer, multinomial and power function bounds are identical and are given by the 95% upper bound for the population error rate p. The bound is therefore directly computed by

$$\hat{p}_\mu(0; 0.95) = 1 - (0.05)^{1/n}, \tag{2.8.1}$$

using the binomial distribution. For $n = 100$, $\hat{p}_\mu(0; 0.95) = 0.0295$. In practice, the Poisson approximation of $3/n$ is often used. The computation of the moment bound is more involved but gives a very similar result.

For Bayesian bounds, the value of a 95% confidence bound depends on the choice of the prior about the error distribution. Using extensive simulation, Neter and Godfrey (1985) discovered that for certain priors the Cox and Snell bound demonstrates a desirable relative frequency behavior under repeated sampling. One such setting is to use the following values for the mean and the standard deviation for the gamma prior of p and μ_Z, respectively: $p_0 = 0.10$, $\sigma_p = 0.10$, $\mu_0 = 0.40$, and $\sigma_\mu = 0.20$. These can be related to the parameters a and b in (2.7.11) as follows:

$$a = (p_0/\sigma_p)^2, \tag{2.8.2a}$$

$$b = (\mu_0/\sigma_\mu)^2 + 2. \tag{2.8.2b}$$

Thus for no errors in the sample, i.e., $m = 0$, using the above prior values, we compute

$$a = (0.10/0.10)^2 = 1,$$

$$b = (0.40/0.20)^2 + 2 = 6.$$

The degrees of freedom for the F distribution are $2(m + a)$ and $2(m + b)$, so for $m = 0$ they are 2 and 12, respectively. Because the 95th percentile of $F_{2,12}$ is 3.89, and the coefficient, when $n = 100$, is

$$\frac{m\bar{z} + (b-1)\mu_0}{n + a/p_0}\,\frac{m + a}{m + b} = \frac{(6-1)0.40}{100 + 1/0.10}\,\frac{1}{6} = 0.00303,$$

the 95% Cox and Snell upper bound is $0.00303 \times 3.89 = 0.01177$.

For another Bayesian bound proposed by Tsui, Matsumura and Tsui we use the prior given in Section 2.7, namely, the Dirichlet prior with parameters $K = 5$, $\rho_0 = 0.8$, $\rho_{100} = 0.101$ and $\rho_i = 0.001$ for

$i = 1, \cdots, 99$. Given no error in a sample of 100 dollar unit observations, the posterior values for these parameters are $K' = K + n = 105$, and $\rho_0' = (K\rho_0 + w_0)/K' = (5(0.8) + 100)/105 = 0.99048$. Similarly, $\rho_{100}' = 5(0.101)/105 = 0.00481$, and $\rho_i' = 5(0.001)/105 = 0.00004762$ for $i = 1, \cdots, 99$. The expected value for the posterior μ_D is then

$$E(\mu_D) = \left(\frac{1}{100} + \frac{2}{100} + \cdots + \frac{99}{100}\right) 0.00004762 + \frac{100}{100} 0.00481$$

$$= 0.007167.$$

To obtain $\text{var}(\mu_D)$, we compute $E(\mu_D^2) = 0.0063731$ so that

$$\text{var}(\mu_D) = \frac{E(\mu_D^2) - \{E(\mu_D)\}^2}{K' + 1} = 0.00005964.$$

The posterior distribution is, then, approximated by the beta distribution having the expected values and the variance computed above. The two parameters α and β of the approximating beta distribution $B(\alpha, \beta)$ are

$$\alpha = E(\mu_D)\left[\frac{E(\mu_D)\{1 - E(\mu_D)\}}{\text{var}(D)} - 1\right] = 0.848$$

and

$$\beta = \{1 - E(\mu_D)\}\left[\frac{E(\mu_D)\{1 - E(\mu_D)\}}{\text{var}(D)} - 1\right] = 117.46.$$

The upper bound is then given by the 95th percentile of the beta distribution with parameters 0.848 and 117.46, which is 0.00227.

B. *One Error in the Sample.* When the DUS audit data contain one error, each method produces a different result. First of all, for computation of the Stringer bound, we determine a 95% upper bound for p, $\hat{p}_u(m, 0.95)$ for $m = 0$ and 1. Software is available for computing these values (e.g., BELBIN (BINES in the latest version) in International Mathematical and Statistical Libraries). We compute $\hat{p}_u(0, 0.95) = 0.0295$ and $\hat{p}_u(1, 0.95) = 0.0466$. Suppose that the observed tainting is $t = 0.25$. Then a 95% Stringer bound is

$$\hat{p}_u(0, 0.95) + t(\hat{p}_u(1, 0.95) - \hat{p}_u(0, 0.95))$$
$$= 0.0295 + 0.25(0.0466 - 0.0295) = 0.0338.$$

Second, the multinomial bound has an explicit solution for one error. It is convenient to express the observed tainting in cents so set $t' = 100t$. Denote also a 95% lower bound for p as $\hat{p}_l(m, 0.95)$ when a sample of n observations contain m errors. Then a 95% multinomial bound for $m = 1$ is given by $(t'\hat{p}_{t'} + 100\hat{p}_{100})/100$, where $\hat{p}_{t'}$ and $\hat{p}_{100}$ are determined as follows. Let

$$\hat{p}_0 = \max\left[\left\{\frac{0.05}{1 + t'n/[(100 - t')(n - 1)]}\right\}^{1/n}, \hat{p}_l(n - 1, 0.95)\right]. \tag{2.8.3}$$

Then

$$\hat{p}_{t'} = \frac{1}{n}\left\{\frac{0.05}{\hat{p}_0^{n-1}} - \hat{p}_0\right\} \tag{2.8.4}$$

and

$$\hat{p}_{100} = 1 - \hat{p}_0 - \hat{p}_{t'}. \tag{2.8.5}$$

To illustrate the above computation, using $t' = 25$ and $n = 100$, we compute that $\hat{p}_l(99, 0.95) = 0.9534$ and

$$\left\{\frac{0.05}{1 + 25(100)/[(100 - 25)(100 - 1)]}\right\}^{1/100} = 0.96767,$$

so that $\hat{p}_0 = 0.96767$. Then by (2.8.4),

$$\hat{p}_{t'} = \frac{1}{100}\left(\frac{0.05}{0.96767^{99}} - 0.96767\right) = 0.00326.$$

Hence, $\hat{p}_{100} = 1 - 0.96767 - 0.00326 = 0.0291$. A 95% multinomial upper bound, when $m = 1$, is then $0.25(0.00326) + 0.0291 = 0.02988$.

Third, we discuss computation of the parametric bound using the power function for modeling the distribution of tainting. The density of z is

$$f(z) = \lambda z^{\lambda - 1} \quad \text{for } 0 < z \leq 1. \tag{2.8.6}$$

The mean tainting $\mu_Z = \lambda/(\lambda + 1)$, and hence

$$\mu_D = \frac{p\lambda}{\lambda + 1}. \tag{2.8.7}$$

Given a sample of $n = 100$ dollar units and the same single error of $t = 0.25$, we compute the maximum likelihood estimates of parameters p and λ,

$$\hat{p} = \frac{m}{n} = 0.01 \tag{2.8.8a}$$

and

$$\hat{\lambda} = -\frac{m}{\sum_{i=1}^{m} \log t_i} = 0.7214, \tag{2.8.8b}$$

respectively. Using these estimates, we construct the following parametric bootstrap estimate of the distribution of the error d of the population:

$$d = \begin{cases} 0, & \text{with probability } 0.99, \\ z, & \text{with probability } 0.01, \end{cases} \tag{2.8.9}$$

where z has the density

$$\hat{f}(z) = 0.7214z^{-0.2786}, \quad 0 \le z \le 1. \tag{2.8.10}$$

A random sample of size n from the distribution (2.8.9) and (2.8.10) is called the bootstrap sample. Denote $\hat{\mu}_D^*$ as the value of $\hat{\mu}_D = \hat{p}\hat{\lambda}/(\hat{\lambda}+1)$ computed from a single bootstrap sample. The distribution of $\hat{\mu}_D^*$ under sampling from (2.8.9) and (2.8.10) is the bootstrap distribution of $\hat{\mu}_D^*$. The 95 percentile of the bootstrap distribution is used to set a bound for μ_D. To approximate the bootstrap sampling distribution, we may use simulation. Let B be the number of independent bootstrap samples. Then an estimate of a 95% upper bound is U_B such that

$$\frac{(\#\ of\ \hat{\mu}_D^* < U_B)}{B} \ge 0.95. \tag{2.8.11}$$

B should be sufficiently large. For our example, using $B = 1000$, we get $U_B = 0.01481$.

The computation of the two Bayesian bounds can follow the steps given above for the case when $m = 0$. Thus, for the Cox-Snell bound, we compute, using the same prior settings, $F_{2(1+1),2(1+6)}(.95) = 3.112$. Hence the desired 0.95 upper bound of μ_D is

$$\frac{0.25 + 5(0.40)}{100 + 1/0.10}\frac{1+1}{1+6}3.112 = 0.0182.$$

The bound of Tsui, Matsumura and Tsui can also be computed in the same way as before. For this sample, $K' = 105$ as before. But $\rho_0' = (5(0.8) + 99)/105 = 0.98095$, $\rho_{25}' = (5(0.001) + 1)/105 = 0.009571$, $\rho_{100}' = 5(0.101)/105 = 0.00481$ and $\rho_i' = 5(0.001)/105 = 0.00005$ for the rest. The mean and the variance of this posterior distribution are 0.0095 and the distribution has $\alpha = 1.382$ and $\beta = 143.37$. A 95% upper bound is 0.0255.

C. *More than One Error in the Sample.* When the number of errors m exceeds 1, the multinomial bound computation is more involved and requires substantial programming. At the time of this report, a standard package does not appear to be available. (Note: A copyrighted PC software, that uses clustering discussed in Section 2.5, is available from R. Plante of Purdue University.) For the bounds for which the computation has been illustrated, one can follow the steps shown for the $m = 1$ case in a straightforward manner. We will, however, describe the computation of the moment bound for $m = 2$. Suppose that observed taintings are 0.25 and 0.40. In this method, the sampling distribution of the estimator $\hat{\mu}_D$ is approximated by a three parameter gamma distribution, $\Gamma(x; A, B, G)$, where $A > 0$, $B > 0$ and $x > G$. The method of moments is used to estimate these parameters. Let m_i, $i = 1, 2$ and 3, be the sample central, second and third moments. Then the moment estimators are

$$\tilde{A} = 4m_2^3/m_3^2, \tag{2.8.12}$$

$$\tilde{B} = \tfrac{1}{2}\, m_3/m_2 \tag{2.8.13}$$

and

$$\tilde{G} = m_1 - 2m_2^2/m_3. \tag{2.8.14}$$

For computation of m_i of the sample mean tainting, a number of heuristic arguments are introduced. First of all, we compute the average tainting $\bar{t} = 0.325$ of the two observations. Suppose that the population audited is a population of accounts receivables. Then, we *compute*, without any statistical explanation being given, the third data point, t^*,

$$t^* = 0.81[1 - 0.667\tanh(10\bar{t})] \cdot [1 + 0.667\tanh(n/10)] = 0.3071.$$

The term in the second pair of brackets will not be used when the population is inventory. t^* is so constructed that when there is no error in a sample, the upper bound is very close to the Stringer bound. Using, thus, three data points (two observed and one constructed) the first three noncentral moments are computed for z, i.e., the tainting of items in error. They are

$$v_{z,1} = (0.25 + 0.40 + 0.3071)/3 = 0.31903,$$

$$v_{z,2} = (0.25^2 + 0.40^2 + 0.3071^2)/3 = 0.1056$$

and

$$v_{z,3} = (0.25^3 + 0.40^3 + 0.3071^3)/3 = 0.03619.$$

The noncentral moments of d are simply p times the noncentral moments of z. Using well known properties of moments, the population central, second and third moments can then be derived from noncentral moments. These population central moments are used to determine the three noncentral moments of the sample mean. Throughout these steps the error rate p is treated as a nuisance parameter but at this stage is integrated out using the normalized likelihood function of p. Then, the noncentral moments of the sample mean are shown to be

$$v_{d,1} = \frac{m+1}{n+2}v_{z,1}, \tag{2.8.15}$$

$$v_{d,2} = \frac{\mathscr{A}}{n}, \tag{2.8.16}$$

where

$$\mathscr{A} = \frac{m+1}{n+2}v_{z,2} + (n-1)\frac{m+1}{n+2}\frac{m+2}{n+3}v_{z,1}^2,$$

and

$$v_{d,3} = \frac{\mathscr{B}}{n^2}, \tag{2.8.17}$$

where

$$\mathscr{B} = \frac{m+1}{n+2} v_{z,3} + 3(n-1)\frac{m+1}{n+2}\frac{m+2}{n+3} v_{z,1}v_{z,2} + (n-1)(n-2)\frac{m+1}{n+2}\frac{m+2}{n+3}\frac{m+3}{n+4} v_{z,1}^3.$$

Using (2.8.15) through (2.8.17), we compute $v_{d,1} = 0.93831 \times 10^{-2}$, $v_{d,2} = 0.14615 \times 10^{-3}$ and $v_{d,3} = 0.29792 \times 10^{-5}$. Then

$$m_1 = v_{d,1} = 0.93831 \times 10^{-2}, \tag{2.8.18}$$

$$m_2 = v_{d,2} - v_{d,1}^2 = 0.58104 \times 10^{-4}, \tag{2.8.19}$$

$$m_3 = v_{d,3} - 3v_{d,1}v_{d,2} + 2v_{d,1}^3 = 0.51748 \times 10^{-6}. \tag{2.8.20}$$

Using these values, we compute $\tilde{A} = 2.93$, $\tilde{B} = 0.00445$ and $\tilde{G} = -0.00366$. These parameter estimates are used to determine the 95th percentile of the gamma distribution to set a 95% moment bound. The bound is 0.0238. For comparison, for the same audit data, the Stringer bound = 0.0401, the parametric bound = 0.0238, and using the prior settings previously selected, the Cox and Snell bound = 0.0248 and the Tsui, Matsumura and Tsui bound = 0.0304. Table 4 tabulates the results. Note that when there is no error in the sample ($m = 0$), the two Bayesian bounds, under the headings C&S and Tsui, are considerably smaller than the four other bounds. The reason is that the four other bounds assume that all taints are 100% when there is no error in the sample. When the sample does contain some errors, the bounds are closer, as shown for $m = 1$ and 2.

3. SUMMARY AND RECOMMENDATIONS

The purpose of this report has been to review and evaluate the state of statistical practice in accounting and auditing. In it, we have emphasized, (1) the importance of the problem as one of national interest, (2) the nonstandard nature of the statistical problem relative to the main body of existing statistical methodology, (3) the lack of adequately reliable procedures and (4) the generally scattered and ad hoc nature of the existing methodology. It is clear that much additional research is needed. For this reason, the report has been directed primarily toward researchers and graduate students, both in statistics and accounting. However, the following summary and recommendations should be of interest to a much wider audience within our respective disciplines.

- Auditing is an essential activity in a society with advanced capital markets. In such a society, investors and government officials base many important decisions on accounting information. Those decisions affect the welfare of all citizens. Auditing is a costly activity and statistical procedures can play an important role in reducing those costs.
- Basic statistical problems in auditing arise when one wishes to estimate the total population error in an account. Relative to the main body of statistical methodology, these problems are nonstandard due to a unique feature of the data; audit data usually contains mostly zeros! Existing statistical methods do not offer satisfactory solutions for inferences based on such information.
- This report's survey of the existing literature and practices points up several important observations. First of all, statistical methods have only recently begun to be developed for analyzing this nonstandard type of data; in the annotated bibliography, all but five of the references are dated after 1972. The first significant contribution was that of Aitchison (1955) and the key idea of Dollar Unit Sampling (DUS) was reported by Stringer in 1963.
- One of the main factors that serves to retard progress in the development of new methodology for auditing problems is the high degree

TABLE 4

Comparison of six 95% upper confidence bounds for μ_D: the Stringer bound, the multinomial bound, the moment bound, the parametric bound, the Cox and Snell bound and the Tsui, Matsumura and Tsui bound (sample size is $n = 100$)

No. of errors	Stringer	Multinomial	Moment	Parametric	C & S	Tsui
$m = 0$	0.0295	0.0295	0.0295	0.0295	0.0118	0.0023
$m = 1$ $t = 0.25$	0.0338	0.0299	0.0156	0.0152	0.0182	0.0255
$m = 2$ $t_1 = 0.40$ $t_2 = 0.25$	0.0401	0.0315[a]	0.0239	0.0238	0.0248	0.0304

[a] This value was computed by the software made available by Plante (1987).

of confidentiality placed upon accounting information. The resultant lack of good data prevents the characteristics of accounting populations from being adequately known in all but a limited number of cases. In order to improve the quality and applicability of statistical auditing procedures, it is essential that much more data be made available by both public and private sectors. In particular, one's confidence in the outcomes of statistical analyses depends heavily upon the suitability of the models that have been postulated for the situations in question. However, the selection of appropriate models relies critically upon the availability of adequate data from such situations. Until there is adequate data available to give guidance and justification for model selections, there will be less than the desired confidence in the analyses based upon them.

- A survey of existing approaches to the statistical problems of auditing reveals that one of the most important ideas is that of DUS. This sampling design selects items from an account with probability proportional to their book amounts. Items with large book amounts, therefore, are more likely to be selected than items with smaller amounts. Because the items with larger book values are considered relatively more important than those with smaller book values, DUS is an appealing sampling design when the auditor places primary emphasis on overstatements. The DUS design does have some limitations, however. For example, items with a zero book amount will not be selected under this sampling.
- The dollar unit sampling design also permits the auditor to incorporate into the analysis prior knowledge that the errors are overstatements and that the maximum size of an error of an item is equal to its book amount. This assumption, when applicable, sets an upper limit of 1 and a lower limit of 0 for a DUS error. This error, referred to by accountants as tainting, is the ratio of the error amount to the recorded book amount. Under this assumption, an auditor can set a conservative upper bound for the population error with a confidence level at least as large as the stated one. The upper bound for the population error amount is equal to the $1 - \alpha$ upper bound for the error rate, multiplied by the known total book amount of the population. Cox and Snell (1979) provides a theoretical framework for this method. In this report, it is concluded that this bound based on attribute sampling theory is the only procedure available that has a theoretically known sampling distribution. This means that the long run performance of all other currently available procedures must be investigated by means of simulation. Consequently, it is not easy to obtain information about the performance of these procedures in a wider audit situation.
- A significant weakness of the upper bound defined in this way is that it is far too conservative in that the author's confidence coefficients are much larger than intended. A heuristic method, credited to Stringer, has been widely used and it produces a tighter bound. It is now about 25 years since the Stringer bound was proposed. In spite of the fact that extensive simulation demonstrates that it is far too conservative, no theoretical justification has as yet been obtained! This remains an important and interesting open question.
- Several alternative methods, mostly heuristic and sometimes totally *ad hoc*, have been proposed in recent years and these are reviewed in this report. Based on limited investigations, the upper bounds set by some of these procedures are shown to be considerably tighter than the Stringer bound. Much more research along these lines is needed. It is also recommended that extensive testing be carried out using real data in order to evaluate adequately these and other procedures.
- Sequential methods would seem to be appropriate for some of these problems, and yet there is a noticeable lack of such methods in the relevant literature. This is in spite of the fact that general sequential methodology is available in statistical monographs directed toward accounting, for example, Cyert and Davidson (1962). In particular, simple two-stage sampling schemes could be considered as a possible way to improve the performance of some of the statistical procedures.
- Empirical studies indicate that negative errors caused by understatements are also quite common in auditing populations. Very little research, however, has been done on the problem of determining bounds for these cases, and this needs to be corrected. Note that DUS may not be an effective sampling design when understatements are present because items with larger audited amounts may have smaller chances of selection than desired.
- Except for the procedures that utilize Bayesian methods, existing procedures are not effective for setting a good lower bound for accounting population errors. This failure is extremely serious; one particularly important situation involves the estimation of the adjustment

of a firm's expense accounts by the Internal Revenue Service. The current IRS procedure is to apply standard sampling methods such as those used in surveys of human populations. Investigation of the performance of these estimators for certain audit populations indicates that such IRS practice is too conservative in the sense that the IRS is assuming much lower risk than allowed in the policy. That is, the actual level of confidence is substantially higher than the nominal level. Such practice tends to underestimate the potential tax revenue due the government. Similar problems also arise in other governmental agencies, e.g., the Office of Inspector General of the Department of Health and Human Services, in their investigation of compliance with government guidelines of reported expenses by local governments. It is important that intensive research be carried out for the purpose of developing more reliable procedures for determining lower confidence bounds. The financial benefits to the government from such research should be significant.

- The development of valid statistical methods for setting confidence bounds for accounting populations is of national interest and importance, in major part because of the considerable economic benefits that would accrue to both the public and private sectors.
- In developing methodologies, primary emphasis should be placed upon the derivation and performance of one-sided confidence intervals and not the two-sided confidence intervals commonly discussed in standard statistical texts. Texts should be revised to reflect this.
- In this age of widely available high speed computing equipment, it is reasonable to expect significantly greater use of computer-intensive statistical methodologies. There is also a need for greater use of computers in the simulation of performance characteristics of existing methodologies, particularly as increased data sets become available to suggest more realistic simulation models.
- The survey of the existing literature that is given in the annotated bibliography below reveals that the statistics profession as a whole has not been heavily involved with the important statistical problems that arise in auditing. This may be due in part to the fact that there has not been adequate nor regular interaction between researchers from the accounting and the statistics professions.
- It is recommended that a series of workshops, conferences and courses be set up at which theoretical and applied statisticians can meet and exchange expertise and problems with accountants and auditors from each of the sectors of government, business and academia. It would be expected that proceedings of some of these activities would be published with the purpose of improving the communication between the two disciplines. There would also be important benefits from holding a conference that would bring researchers together from the many diverse areas of applications that exist throughout all of the sciences in which problems of nonstandard mixtures arise. Several such areas are briefly described in Section 1. Primarily, the exchange of problems and the transfer of relevant methodologies and references could expedite progress in all areas.
- The initial emphasis of coordinated research activities between the auditing and statistics professions should focus upon seeking ways to encourage statisticians to become directly involved in the auditing environment. In this way, statisticians would become more familiar with the statistical problems in auditing and especially with the characteristics of the data bases in this setting. It is also recommended that accounting firms make audit data available for a wider research community than its own profession.
- Large private accounting firms have both economic incentives and resources to carry on research for the purpose of developing better statistical procedures for their audit problems. However, concerted efforts must be made to improve the statistical methodologies used in the public sector.

ACKNOWLEDGMENTS

This report is the result of the efforts of many. The survey of existing methodologies, Section 2, and the post-1981 updating of the annotated bibliography were prepared by H. Tamura, who undertook the responsibility for the final preparation of the report since his appointment to the Panel in 1985. We express our appreciation to W. F. Felix, University of Arizona; J. Godfrey, University of Georgia; L. Heath, University of Washington; K. W. Stringer, New York University; and R. Bartyczak, T. Lieb, R. Weakland and D. Wilt of the Internal Revenue Service. Their considerable assistance is gratefully acknowledged.

The completion of this project would not have been possible without the determined oversight of Ralph Bradley and Ronald Pyke during their terms as chairman of the Committee on Applied and Theoretical Statistics. In particular, Ronald Pyke worked closely with H. Tamura, providing leadership and support

throughout the development of the report. We also gratefully acknowledge the support of the staff of the Board on Mathematical Sciences under the direction of Frank Gilfeather and Lawrence H. Cox.

Support for this project was provided by the Treasury Department and the Defense Logistics Agency, and by core funds for the Board on Mathematical Sciences, National Research Council, provided by the National Science Foundation, the Air Force Office of Scientific Research, the Army Research Office, the Department of Energy and the Office of Naval Research.

This report is reprinted from *Statistical Models and Analysis in Auditing*, 1988, with permission of the National Academy Press, Washington, D. C.

ANNOTATED BIBLIOGRAPHY

AITCHISON, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *J. Amer. Statist. Assoc.* **50** 901–908.

This paper is the first to address the problem of the estimation of parameters from data containing many zeros. The best unbiased estimators of the population mean and the variance are derived. However, the paper does not consider the sampling distribution of these estimators and thus results are not of immediate use to auditors.

AMERICAN INSTITUTE OF CERTIFIED PUBLIC ACCOUNTANTS (AICPA) (1981). *Statement on Auditing Standards (SAS No. 39)*. AICPA, New York.

AMERICAN INSTITUTE OF CERTIFIED PUBLIC ACCOUNTANTS (1983). *Audit Sampling*. AICPA, New York.

Auditing practices are regulated by a number of national organizations. Among them, the most influential organization is the American Institute of Certified Public Accountants, a national organization of practicing certified public accountants. The current auditing standards are stated in *Statement on Auditing Standards (SAS No. 39)*. *Audit Sampling* is a detailed interpretation of SAS No. 39 and describes procedures and practicing guides that auditors should adhere to when the auditing is performed based on examination of less than 100% of the items of an accounting balance.

ANDERSON, R. J. and LESLIE, D. A. (1975). Discussion of considerations in choosing statistical sampling procedures in auditing. *J. Accounting Res.* **13** (suppl.) 53–64.

The discussion focuses partly on the paper by Loebbecke and Neter and partly on the then-forthcoming AICPA Monograph No. 2, *Behavior of Major Statistical Estimators in Sampling Accounting Populations*. With respect to the former, Anderson and Leslie question the distinction between whether an audit should have "attributes" or "variables" objectives, arguing that all audit objectives should be expressed in monetary terms. With respect to the latter study, Anderson and Leslie argue that the AICPA study should have included dollar-unit sampling with the Stringer bound, rather than the combined attributes-variables bound because the Stringer bound is less conservative and more widely used. The discussants believe that dollar-unit sampling with the Stringer bound is appropriate in almost all circumstances so that the auditor need not consider alternative sampling approaches and fall back procedures if the anticipated environmental conditions are not met, as proposed by Loebbecke and Neter.

ANDERSON, R. J. and TEITLEBAUM, A. D. (1973). Dollar-unit sampling. *Canad. Chartered Accountant* (after 1973, this publication became *CA Magazine*) April 30–39.

This expository article introduces dollar-unit sampling in a way understandable to a broader group of researchers and practitioners.

BAKER, R. L. and COPELAND, R. M. (1979). Evaluation of the stratified regression estimator for auditing accounting populations. *J. Accounting Res.* **17** 606–617.

This study supplements the Neter and Loebbecke AICPA Monograph No. 2 by studying the behavior of the stratified regression estimation for the same four accounting populations used in the AICPA study. The precision of the regression estimator in general tends to be almost the same as the precision of the stratified difference, ratio and regression estimators. As for these other estimators, the reliability of the nominal confidence coefficient for the stratified regression estimator is poor at low error rates, with the regression estimator performing even more poorly than the other estimators.

BARKMAN, A. (1977). Within-item variation: A stochastic approach to audit uncertainty. *Accounting Rev.* **52** 450–464.

The author proposes that the audit amount to be established by the auditor for a line item be treated as a random variable, such as when the line item is the amount of bad debt for an account receivable. The author assumes that the distribution reflecting the uncertainty of the line item audit amount is given by the beta distribution and that the distribution of the total amount is normal. A simulation study was carried out to study the behavior of sample estimates under different population conditions for mean-per-unit and difference estimators.

BECK, P. J. (1980). A critical analysis of the regression estimator in audit sampling. *J. Accounting Res.* **18** 16–37.

This study supplements the Neter and Loebbecke study in AICPA Research Monograph No. 2 by examining the behavior of the stratified and unstratified regression estimators for the accounting populations considered in the AICPA monograph. In addition, one of these populations was further manipulated in order to vary the extent of heteroscedasticity in the population. The results obtained were similar to those previously reported for the difference and ratio estimators in the Neter and Loebbecke study. The author concludes that heteroscedasticity appears to be a significant factor in the behavior of the regression estimator in two of the four accounting populations and that stratification cannot always be relied upon to provide a confidence level close to the nominal one. The authors also made a limited study of the power of statistical tests based on the regression estimator.

BURDICK, R. K. and RENEAU, J. H. (1978). The impact of different error distributions on the performance of selected sampling estimators in accounting populations. *Proc. Bus. Econ. Statist. Sec.* 779–781. Amer. Statist. Assoc., Washington.

This paper reports on a simulation study based on one of the accounting populations employed in the Neter and Loebbecke (1975) study. Errors were injected into the population at different error rates, with equal probability for each line item, with probability proportional to book amount and with probability inversely proportional to book amount. A number of estimators were studied as to their precision and the closeness of the actual confidence level to the nominal one based on large-sample theory. The authors conclude that an estimator developed by Hartley is to be preferred over the other estimators studied.

CARMAN, L. A. (1933). The efficacy of tests. *Amer. Accountant* December 360–366.

This paper proposes application of a simple probability model for computing the sampling risk in auditing and is the first publication of such an attempt in accounting.

COX, D. R. and SNELL, E. J. (1979). On sampling and the estimation of rare errors. *Biometrika* **66** 124–132.

After describing a theoretical model of monetary unit sampling, the paper presents a Bayesian analysis of the problem. Specifically, the authors consider the case where the number of errors has a Poisson distribution and the proportional error density is exponential. Using a simple conjugate prior, they derive the posterior distribution and discuss some possibilities for the parameters of the prior distribution.

COX, D. R. and SNELL, E. J. (1982). Correction to "On sampling and the estimation of rare errors." *Biometrika* **69** 491.

Corrections to their 1979 paper are announced in this note.

CYERT, R. M. and DAVIDSON, H. J. (1962). *Statistical Sampling for Accounting Information.* Prentice-Hall, Englewood Cliffs, N.J.

This basic text in statistical sampling for auditing introduces among other standard topics the application of sequential sampling for compliance test.

DEAKIN, E. B. (1977). Discussant's response to "Computing upper error limits in dollar-unit sampling," by S. J. Garstka. In *Frontiers of Auditing Research* (B. E. Cushing and J. L. Krogstad, eds.) 195–201. Bureau of Business Research, Univ. Texas at Austin, Austin, Tex.

This critique of Garstka (1977b) mentions that the conservatism of upper error bounds should be reduced by research on the handling of unobserved errors, that the models proposed by Garstka lack empirical support, that the proposed use of alternative models provides no rationale for selecting an appropriate model in a given situation and that the research does not provide sufficient evidence to support the assumption that generalized or compound Poisson models would be any more useful in auditing than the present dollar-unit sampling models.

DUKE, G. L., NETER, J. and LEITCH, R. A. (1982). Power characteristics of test statistics in the auditing environment: An empirical study. *J. Accounting Res.* **20** 42–67.

The power characteristics of eight test statistics, used with both the positive and negative testing approaches, are studied for four different accounting populations. Two-error characteristic linkage models were developed for systematically creating different total error amounts in an accounting population. It was found that no one test statistic and either of the two testing approaches is uniformly superior, and that audit decisions based on a sample of 100 observations tend to involve high sampling risks for the auditor.

DWORIN, L. and GRIMLUND, R. A. (1984). Dollar unit sampling for accounts receivable and inventory. *Accounting Rev.* **59** 218–241.

The development of the moment bound is discussed in detail in this article. The authors state that their methods are based on the assumption that the dollar unit tainting follows a mixture of two χ^2 distributions. A helpful chart is provided in their Table 1 for computing the bound. Its performance is compared with that of the multinomial bound.

DWORIN, L. and GRIMLUND, R. A. (1986). Dollar unit sampling: A Comparison of the quasi-Bayesian and moments bounds. *Accounting Rev.* **61** 36–57.

This article reports the results of comparing the performance of the moments bound with that of McCray's quasi-Bayesian bound using the uniform prior. A slight modification is proposed in the original moment bound to make the bound more efficient without any noticeable loss of the reliability of the bound. For the populations considered, the true level of confidence tends to be higher than the nominal level of 95% used in the study for both bounds. Still, both bounds are considerably tighter than the Stringer bound. However, between the two, the performance is relatively comparable.

FELIX, W. L., JR. and GRIMLUND, R. A. (1977). Sampling model for audit tests of composite accounts. *J. Accounting Res.* **15** 23–42.

This article discusses an alternative statistical sampling model which avoids some of the assumptions of conventional methods. It is a Bayesian approach where the book value and audit value are analytically combined with the auditor's prior judgments. A single combined "beta-normal" probability distribution for the total error in an account balance is derived. Several properties of this distribution are presented, followed by a discussion of how the auditor may use it to make probabilistic statements about the total error amount in a population. The computational procedure for using the beta-normal distribution is cumbersome, thus several alternative computational procedures are suggested, followed by a brief discussion of how the analysis may be used to preselect a sample size.

FELIX, W. L., JR., LESLIE, D. A. and NETER, J. (1982). University of Georgia Center for Audit Research monetary unit sampling conference, March 24, 1981. *Auditing: J. Practice Theory* **1** 92–103.

This paper reports the results of the conference on dollar unit sampling held at the University of Georgia in March 1981. A short summary of existing methods for computing bounds is given. Also, the advantages and disadvantages of DUS compared to line item sampling are discussed. Several DUS methods are presented and compared. Research issues are also summarized.

FESTGE, M. O. (1979). Discussion of an empirical study of error characteristics in audit populations. *J. Accounting Res.* **17** (suppl.) 103–107.

This is a discussion of the study by Ramage, Krieger and Spero (1979) by a practicing auditor. One of his comments is that the study may be biased because the audit data base used in the study is supplied by one of the major accounting firms and thus may reflect their audit objectives which may also vary from one case to another.

FIENBERG, S. E., NETER, J. and LEITCH, R. A. (1977). Estimating the total overstatement error in accounting populations. *J. Amer. Statist. Assoc.* **72** 295–302.

This paper presents a statistical sampling approach based on the multinomial distribution for obtaining a bound for either the total population overstatement or understatement or both. The bound is derived by using an optimization routine that finds the maximum monetary error subject to constraints representing the joint confidence region for the multinomial parameters. The key element is the definition of the S set which denotes the set of all outcomes as extreme or less extreme than the observed outcomes. The multinomial bound is numerically compared to the Stringer bound and the results indicate that the multinomial bound is less conservative than the Stringer bound.

FINANCIAL ACCOUNTING STANDARDS BOARD (FASB) (1980). *Statement of Financial Accounting Concepts No. 2 (SFAC2: Qualitative Characteristics of Accounting Information)*. FASB, Stamford, Conn.

The Securities Exchange Act of 1934 gave the SEC the authority to promulgate financial reporting standards (generally accepted accounting principles or GAAP) for those companies subject to the jurisdiction of the SEC. The SEC, in turn, has delegated this authority to the FASB. The definition of a *material error* is from paragraph 132 of SFAC2 issued in May, 1980.

FROST, P. A. and TAMURA, H. (1982). Jackknifed ratio estimation in statistical auditing. *J. Accounting Res.* **20** 103–120.

One recent development in statistics is the use of computer-intensive methods for data analysis. In this article, the performance of the ratio estimation is studied when the standard error is computed using the conventional method and using the jackknife. The accounting data used in the Neter and Loebbecke study are employed as the populations for simulation. Their conclusion is that when error rates are not too small, so that the problem of the mixture as discussed in the main text is not severe, the jackknife clearly gives a better performance and should be used.

FROST, P. A. and TAMURA, H. (1986). Accuracy of auxiliary information interval estimation in statistical auditing. *J. Accounting Res.* **24** 57–75.

This work extends Kaplan's 1973 investigation of the performance of the auxiliary information interval estimators and traces the cause of the poor performance of these estimators to the skewness of the accounting population induced by the mass of probability at the origin. Analysis is done based on the difference estimator but indicates that the result can be applied to the ratio estimator.

FROST, P. A. and TAMURA, H. (1987). Accuracy of auxiliary information interval estimation in statistical auditing. Working Paper Series 2–87, Dept. Management Science, School and Graduate Schools of Business Administration, Univ. Washington.

This working paper contains additional results to those that are reported in Frost and Tamura (1986).

GARSTKA, S. J. (1977a). Models for computing upper error limits in dollar-unit sampling. *J. Accounting Res.* **15** 179–192.

This paper investigates alternative methods of computing upper error limits for the monetary error in an accounting population. The compound Poisson process is used to model the error rate and the distribution of error sizes in the population. Simulations are used to demonstrate that tighter upper error limits can be achieved using Bayesian procedures compared to the Stringer bound.

GARSTKA, S. J. (1977b). Computing upper error limits in dollar-unit sampling. In *Frontiers of Auditing Research* (B. E. Cushing and J. L. Krogstad, eds.) 163–182. Bureau of Business Research, Univ. Texas at Austin, Austin, Tex.

This paper poses a number of models to be used in conjunction with dollar-unit sampling. In particular, six compound Poisson models and three generalized Poisson models are considered and upper error bounds developed for each. A simulation study is then used to examine the properties of the various bounds. Use of prior information in selecting the appropriate Poisson model can lead to tighter upper error limits.

GARSTKA, S. J. (1979). Discussion of an empirical study of error characteristics in audit populations. *J. Accounting Res.* **17** (suppl.) 108–113.

Based on the argument that the characteristics of an accounting population should assist the auditor in selecting a proper estimator, the author comments that the measures reported in the study by Ramage, Krieger and Spero (1979) may not be useful for the auditors. For example, he points out that the fractional errors are computed in terms of the audited amount as the base. However, the audited amount is not available for most of the items.

GARSTKA, S. J. and OHLSON, P. A. (1979). Ratio estimation in accounting populations with probabilities of sample selection proportional to size of book values. *J. Accounting Res.* **17** 23–59.

The authors propose a modification of the standard PPS estimator of the population total monetary error. The modification involves deriving a factor to use as a multiple of the standard error in constructing an upper confidence limit for the total monetary error. The basis for the factor is largely heuristic. Limited simulation is used to test the performance of the procedure.

GODFREY, J. T. and NETER, J. (1984). Bayesian bounds for monetary unit sampling in accounting and auditing. *J. Accounting Res.* **22** 497–525.

In 1979, Cox and Snell published a Bayesian model for analysis of dollar unit sample data. This work investigates the sensitivity of the Cox and Snell bound if the auditor's knowledge is incorporated by using different prior distributions, *e.g.*, by using a beta distribution instead of the gamma distribution as proposed by Cox and Snell for the error rate. The authors observe that the effects are moderate. The authors, however, report that the Cox and Snell bound is sensitive to the choice of the prior parameter values. Using simulation, they conclude, it is possible to find the prior parameter values for which the bound demonstrates a desirable relative frequency property.

GOODFELLOW, J. L., LOEBBECKE, J. K. and NETER, J. (1974). Some perspectives on CAV sampling plans. Part I. *CA Magazine* October 23–30; Part II. *CA Magazine* November 46–53.

Combined attributes-variables (CAV) sampling plans were developed to overcome inadequacies in both attributes and variables sampling plans. CAV plans seek to combine the two approaches to obtain effective and efficient estimates of dollar errors in audit populations with low error rates. Part I of the article discusses the basic concepts of CAV sampling. It explains how an attributes sampling plan of unstratified audit units can lead to upper dollar precision limits for the population total overstatement and how stratification of the units improves the efficiency of the estimate. Finally, it considers units selected with probabilities proportional to the book amounts and the essentially equivalent procedures of unstratified random selection of dollar units. Part II explains how the combined attributes variables approach provides tighter precision limits. The strengths of CAV sampling plans are that they provide dollar precision estimates even when the sample contains no error, incorporate the efficiency advantages of stratification without requiring stratified selection and rely on simple conceptual foundations. The weaknesses of the CAV approach include the unbalanced treatment of overstatement and understatement errors, inapplicability to sampling nondollar audit units, ineffective design for disclosing errors, inadequacy of one-sided precision limits for determining the amount of adjustment required, assumption of a zero error rate for planning sample size and emphasis on conservation of precision limits.

GRIMLUND, R. A. and FELIX, W. L. (1987). Simulation evidence and analysis of alternative methods of evaluating dollar-unit samples. *Accounting Rev.* **62** 455–479.

The long run performances of three Bayesian bounds and the moment bound by Dworin and Grimlund are compared. Three Bayesian models are: the normal error model as developed by Grimlund and Felix, the Cox and Snell bound and the multinomial bound with the Dirichlet prior by Tsui, Matsumura and Tsui. The populations used for simulation utilize the model described in Dworin and Grimlund (1984). The non-zero taintings are specified by a mixture of χ^2 distributions and include negative values. The performance of a Bayesian bound depends on the prior parameter values. However, for the prior settings used in this study the Bayesian normal error model and the multinomial bound indicate more consistent performance than the Cox and Snell bound. The multinomial bound with the Dirichlet prior is reported to be too conservative.

HAM, J., LOSELL, D. and SMIELIAUSKAS, W. (1985). An empirical study of error characteristics in accounting populations. *Accounting Rev.* **60** 387–406.

The empirical study of audit populations is scarce and this work is one of four such studies published. Although the previous three studies used the data base supplied by one major accounting firm, this work is based on the data from another major accounting firm. Three factors are considered as possibly affecting the error distribution: (1) account category, (2) company size and (3) industry.

HYLAS, R. E. and ASHTON, R. H. (1982). Audit detection of financial statement errors. *Accounting Rev.* **57** 751–765.

Based on 152 audit cases of one major public accounting firm the causes of errors are traced. In these 152 audits, 281 errors requiring financial statement adjustments were found. The error causes are classified into seven categories and their frequencies are reported.

INTERNAL REVENUE SERVICE (1972). Audit assessments based on statistical samples (Memorandum to Assistant Commissioner from Chief Counsel). IRS, Washington.

INTERNAL REVENUE SERVICE (1975). Audit assessments based on statistical samples. Supplemental Memorandum (March 6 Memorandum to Chief Counsel from Director, Refund Litigation Division and Acting Director, Tax Court Litigation Division). IRS, Washington.

The legal ramifications of statistical sampling for tax audit are studied in these documents and the opinion of the Chief Counsel is stated. It is concluded that "although the propriety of the use of such techniques is not free from doubt, there is sufficient merit in the proposal to warrant judicial testing."

JOHNSON, J. R., LEITCH, R. A. and NETER, J. (1981). Characteristics of errors in accounts receivable and inventory audits. *Accounting Rev.* **56** 270–293.

Auditors and accountants require empirical information about the characteristics of audit populations and error distributions to plan the audit strategy. There is a need for information about the relative frequency, magnitude, distribution and possible causes of errors. In this article, the error characteristics and the relationship between errors and book values in 55 accounts receivable and 26 inventory audits are examined. The distributions of the error amounts and error taintings were studied, as well as the relation between error amounts and book amounts. A summary of the findings is (i) there is great variability in error rates, with those of inventory audits tending to be much higher; (ii) evidence suggests that the error rates may be higher for larger accounts and for accounts with larger line items; (iii) most errors in receivable audits are overstatements, whereas in inventory audits, overstatements and understatements are more balanced in number; (iv) the distribution of error amounts are far from normal, with peak near the mean and farther tails in the upper direction, with receivable errors tending to be larger and less variable than inventory errors; (v) the distributions of error taintings are characterized by pronounced discontinuities at 100%, especially for receivables audits where 100% overstatement errors are frequent; (vi) the mean taintings for receivables are surprisingly large, whereas those for inventories are smaller, but inventories show large negative taintings which occur frequently; (vii) the distributions of taintings are variable and depart substantially from a normal distribution, with some negatively skewed tainting distributions for inventories; and (viii) a study of 20 audits failed to disclose any strong linear relation between error amount and book value, but errors for larger book amounts tend to be more variable. Because the study was based on data from only one CPA firm, the authors emphasize the need for replication studies of the issues raised in the article.

KAPLAN, R. S. (1973a). Stochastic model for auditing. *J. Accounting Res.* **11** 38–46.

A stochastic model is proposed for variable estimation in auditing. The model is based on the use of ratio and regression estimators. These estimators are valuable in audit applications because they utilize the recorded or book value of sample items in the estimation procedure. A disadvantage of ratio or regression estimates is that they are biased, but the bias becomes small as the sample size increases. The model can be used in conjunction with classical techniques to estimate sample size and obtain standard errors of estimates. The sample size would be a function of the auditor's estimates of the error rate and the first two moments of the error distribution, as well as the distribution of book values which is known at the time of audit. The model focuses on the need to estimate two different parameters in an audited population—the error rate and the distribution of errors. Kaplan contends that techniques (such as mean-per-unit estimation using sample values only), which fail to recognize this underlying structure, will probably be of little value to auditors.

KAPLAN, R. S. (1973b). Statistical sampling in auditing with auxiliary information estimators. *J. Accounting Res.* **11** 238–258.

Much of the literature applying statistical sampling to auditing is usually based on techniques developed for sample surveys, such as the simple mean-per-unit estimator. But the auditor typically has more information about the population than is available to those conducting sample surveys. The article indicates how the auditor should use statistical estimators which explicitly use all the available auxiliary information. A class of auxiliary information estimators (difference, ratio, unbiased ratio-type, mean ratio, regression and audit models) and their variance estimates are investigated. Kaplan concludes that to use relatively small sample sizes, while working with stringent materiality factors, auditors must use auxiliary information estimators. Classical techniques are designed for homogeneous populations, whereas audit populations consist of two parts: one of all correct items and the other of items in error. Therefore, techniques which do not explicitly recognize this seem inadequate for auditing applications, and thus, there is a challenge to develop statistically valid techniques which utilize this information.

KAPLAN, R. S. (1975). Sample size computations for dollar-unit sampling. *J. Accounting Res.* **13** (suppl.) 126–133.

This paper discusses a procedure which computes dollar-unit sample size as a function of materiality and the risks of making

alpha and beta errors. In order to control for alpha risk and also allow for some errors in the sample, a low rather than zero error rate is specified. This low error rate is chosen such that one would not expect to reject a population with an error rate this low more than alpha per cent of the time. Given a materiality percentage, a specified low error rate and the alpha and beta risks levels, the procedure derives the sample size required and the critical number of total errors before rejecting the population. The sample sizes generated by this procedure are much larger than those which are based on an assumption of a zero error rate.

KEYFITZ, N. (1984). Heterogeneity and selection in population analysis. Statistics Canada Research Paper No. 10, September 1984.

The concept and effect of heterogeneity in populations are discussed with examples. Heterogeneity and mixtures are closely related. Here the emphasis is upon its effect on error structure and bias in data as well as upon the analysis of data that arises from statistically following groups over time.

KNIGHT, P. (1979). Statistical sampling in auditing: An auditor's view point. *Statistician* **28** 253–266.

Statistical sampling for auditing is reviewed in the auditor's context. Various terms commonly used among practicing auditors are explained.

LEITCH, R. A., NETER, J., PLANTE, R. and SINHA, P. (1981). Implementation of upper multinomial bound using clustering. *J. Amer. Statist. Assoc.* **76** 530–533.

The multinomial bound proposed by Fienberg, Neter and Leitch (1977) is difficult to compute when the number of errors in the sample increases. The authors suggest grouping error observations to reduce the number of errors to be used for computation of the bound. This, of course, leads to losing some efficiency. However, the loss is shown to be not too large for the number of errors between five and eight. Beyond eight errors, the comparison with unclustered bound is not available because of the difficulty in computing the latter.

LEITCH, R. A., NETER, J., PLANTE, R. and SINHA, P. (1982) Modified multinomial bounds for larger numbers of errors in audits. *Accounting Rev.* **57** 384–400.

A modification of the multinomial bound is presented that enables the auditor to obtain bounds for substantially larger numbers of errors in audit samples than was possible with the basic methodology for the multinomial bound. The modification consists of clustering taintings found in the sample and obtaining a conservative bound by assuming all taintings in a cluster are as large as the largest tainting in the cluster. It is found that the modified multinomial bound is usually considerably tighter than the Stringer bound. A simulation study indicated that the confidence level for the modified multinomial bound exceeds or is close to the nominal level for all populations studied.

LESLIE, D. A. (1977). Discussant's response to "Computing upper error limits in dollar-unit sampling," by S. J. Garstka. In *Frontiers of Auditing Research* (B. E. Cushing and J. L. Krogstad, eds.) 183–191. Bureau of Business Research, Univ. Texas at Austin, Austin, Tex.

Some criticisms of the Garstka paper include the fact that many of the populations used in the simulation study did not contain material total error amounts and that the Poisson models are unrealistic in not taking into account bunchings of 100% taintings found in actual accounting populations.

LESLIE, D. A., TEITLEBAUM, A. D. and ANDERSON, R. J. (1980). *Dollar-Unit Sampling—A Practical Guide for Auditors.* Pitman, London.

This is the first book on dollar-unit sampling, a procedure in which the sampling unit is defined as an individual dollar. Aside from the appendices, which account for more than one-third of the publication, it is divided into four parts: auditing foundations for sampling, dollar-unit sampling, planning and evaluation, and applications and practical guidance. Much of the book is devoted to extolling the superiority of dollar-unit sampling over audit unit sampling procedures. The primary advantages of dollar-unit sampling are that it requires no assumption regarding the distribution of errors, and it provides an upper monetary error limit when there are no non-zero differences between the reported value and the audited value.

LILLESTOL, J. (1981). A note on computing upper error limits in dollar unit sampling. *J. Accounting Res.* **19** 263–267.

This paper comments on the work of Garstka (1977a) and demonstrates that if the logarithmic series distribution is used to model the tainting, instead of the geometric series, as proposed by Garstka, the upper bound could change noticeably. Yet, it is difficult to determine which model to use when the auditor expects only several error observations in the sample.

LOEBBECKE, J. K. and NETER, J. (1975). Considerations in choosing statistical sampling procedures in auditing. *J. Accounting Res.* **13** (suppl.) 38–52.

It is suggested that the sampling procedure to be used in a particular auditing application be determined after consideration of audit objectives and environmental factors expected to be encountered in the application. Some of the characteristics of the audit procedure to be considered in choosing an appropriate one include the ability to enlarge the sample, the nature of the sampling frame and the bias of the audit procedure. The authors suggest that the auditor's plan include a provision for a fall back procedure in case the anticipated environmental factors differ from the actual ones.

MCCRAY, J. H. (1984). A quasi-Bayesian audit risk model for dollar unit sampling. *Accounting Rev.* **59** 35–51.

Multinomial modeling of the audit data by Fienberg, Neter and Leitch (1977) appears to provide various extensions. In this work the author treats the mean tainting as a discrete variable and develops a heuristic Bayesian approach to the problem. The work is reviewed in Section 2.7.

MCRAE, T. W. (1974). *Statistical Sampling for Audit and Control.* Wiley, New York.

This text covers the major topics in statistical sampling for auditing, including the basics of statistical sampling, methods of sample selection, estimating population means and proportions, acceptance and discovery sampling and monetary-unit sampling. In addition, the text considers more specialized topics such as cluster, multistage and replicated sampling and the Bayesian approach to making inferences. The text contains a bibliography and a number of tables, including tables for discovery sampling, estimation of population proportion and acceptance sampling. The text is written at a nontechnical level and does not contain significant elements of theory.

MEIKLE, G. R. (1972). *Statistical Sampling in an Audit Context.* Canadian Institute of Chartered Accountants, Toronto.

This monograph discusses an early version of monetary unit sampling where a stratified design is employed.

MENZEFRICKE, U. (1983). On sampling plan selection with dollar-unit sampling. *J. Accounting Res.* **21** 96–105.

An approach for determining the sample size in dollar unit sampling is developed. All errors are assumed to be 100% overstatements.

MENZEFRICKE, U. (1984). Using decision theory for planning audit sample size with dollar unit sampling. *J. Accounting Res.* **22** 570–587.

Using Bayesian models for the error distribution, an approach for sample size determination in dollar unit sampling is developed.

MENZEFRICKE, U. and SMIELIAUSKAS, W. (1984). A simulation study of the performance of parametric dollar unit sampling statistical procedures. *J. Accounting Res.* **22** 588–603.

The performances of certain Bayesian parametric models are investigated in the presence of both over- and understatements. Their performances are compared with those of the Stringer bound and the load and spread bound, which is another nonparametric bound used by practitioners. One Bayesian bound is an application of the Felix and Grimlund's normal error model; the second bound also uses the same structure but develops the bound using a different approach than Felix and Grimlund. (The two bounds show different results.) Another Bayesian model is Cox and Snell's exponential error model. Only one parameter value configuration was used for each Bayesian model. The study concludes that in the presence of understatement error, the normal model, using their own derivation of the bound, appears to show the best performance.

MOORS, J. J. A. (1983). Bayes' estimation in sampling for auditing. *Statistican* **32** 281–288.

This paper reports an error in the Cox and Snell model and presents an alternative derivation of the parametric bound.

NETER, J. (1986). Boundaries of statistics—Sharp or fuzzy? *J. Amer. Statist. Assoc.* **81** 1–8.

In this 1985 American Statistical Association Presidential Address the author reviews the problems of statistical auditing. He comments that the existing solutions often contain heuristic elements, and calls for more active participation of professional statisticians to solve the problem.

NETER, J. and GODFREY, J. (1985). Robust Bayesian bounds for monetary unit sampling in auditing. *Appl. Statist.* **34** 157–168.

The main problem in using a Bayesian bound is to identify proper prior parameter values. Using extensive simulation studies, the authors find certain alternative prior configurations for the Cox and Snell bound that produce desirable relative frequency performance from the bound.

NETER, J., JOHNSON, J. R. and LEITCH, R. A. (1985). Characteristics of dollar unit taints and error rates in accounts receivables and inventory. *Accounting Rev.* **60** 488–499.

The distribution of dollar unit taints is studied using the same data used in the authors' previous study (Johnson, Leitch and Neter, 1981).

NETER, J., LEITCH, R. A. and FIENBERG, S. E. (1978). Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors. *Accounting Rev.* **53** 77–93.

The results cited in Fienberg, Neter and Leitch (1977) are presented here in language more suited to auditors. Additionally, the paper contains a good survey of previous research in the area of upper bounds on monetary unit sampling.

NETER, J. and LOEBBECKE, J. (1975). *Behavior of Major Statistical Estimators in Sampling Accounting Populations—An Empirical Study.* American Institute of Certified Public Accountants, New York.

This monograph is an empirical study of the behavior of statistical estimators commonly used in auditing, based on simulated audit populations with varying error rate patterns constructed by extrapolating error characteristics found in audits of four actual populations. These four populations represent a variety of skewness, error rates and mixture of overstatement and understatements for two types of accounts. The authors report on both the shapes of the book value distributions and error characteristics including rate, magnitude and their relation to book value. The major conclusion of the study is that many widely used statistical procedures may not be reliable when applied to audit populations, especially when the error rate is low. Some other conclusions from the study are that the standard errors for many estimators tend to increase with increases in the error rates, contradicting the assumption that the standard error of the estimator is constant; no one statistical procedure is optimal under all audit circumstances; and further research is needed involving larger samples, different numbers of strata, other estimators, 100% examined stratum truncation, effective hypothesis testing procedures and new statistical procedures especially useful for auditing.

NETER, J. and LOEBBECKE, J. K. (1977). On the behavior of statistical estimators when sampling accounting populations. *J. Amer. Statist. Assoc.* **72** 501–507.

This article is a condensed version (with emphasis on the statistical aspects) of the Neter and Loebbecke (1975) AICPA empirical study on the precision and reliability of several statistical estimators in sampling four accounting populations with various error rates. Problems in using difference and ratio estimators, as well as other estimators, for constructing large-sample normal confidence intervals when the population error rate is low are explored empirically. The findings indicate the need for great care in using large-sample normal confidence intervals for sample sizes of 100 or 200, which are frequently used in auditing practice. The authors conclude that the conditions governing the appropriateness of large-sample normal theory results for ratio and difference estimators need more research, including investigations of the sources of unreliability of the standard large-sample procedures.

PLANTE, R. (1987). Personal communication.

Plante developed a personal computer software to compute the multinomial bound for up to 25 errors. For the number of errors exceeding 10, the program uses clustering of errors. The program is available from Plante, Krannert School of Management, Purdue University, West Lafayette, Ind. 47909.

PLANTE, R., NETER, J. and LEITCH, R. A. (1985). Comparative performance of multinomial, cell and Stringer bounds. *Auditing: J. Practice Theory* **5** 40–56.

In this simulation study, superiority of the multinomial bound is demonstrated as compared to popularly used nonparametric bounds. The effect of alternative dollar unit sampling is also investigated when the population line items are randomly ordered. For comparison, stratified difference estimator using line item sampling is also included.

RAMAGE, J. G., KRIEGER, A. M. and SPERO, L. L. (1979). An empirical study of error characteristics in audit populations. *J. Accounting Res.* **17** (suppl.) 72–102.

The authors contend that audit population error characteristics, aside from distributional shape, can be described by three

rates and a ratio. The rates are overall error rate, F, the fraction of errors which are overstatements, FOV, and contamination—the fraction of errors with relative magnitudes greater than one, CON. The ratio is the error magnitude relative to audit value, denoted by RM. The results of an empirical study of these characteristics indicate that estimates of the three rates vary widely among populations; there is little evidence that either FOV or CON varies systematically with the error rate; for a specific population, none of three rates appears to vary systematically as book value increases; error-absolute magnitude increases roughly in proportion to both book and audit values, but the error magnitude relative to audit value is nearly constant as the audit value increases in magnitude; and inventory populations have widely ranging error rates, typically higher than accounts receivable. However, the usefulness of CON and RM as systematic measures of error magnitude would seem to be limited, because audit sample planning and selection are related to book value, not audit value.

RENEAU, J. H. (1978). CAV bounds in dollar unit sampling: Some simulation results. *Accounting Rev.* **53** 669–680.

This article presents the results of a simulation designed to examine the behavior of five procedures for computing an upper bound on monetary error. The simulation involves three population error direction-conditions, seven error rate conditions and five sample size conditions. The study population was generated to resemble a population previously studied by Neter and Loebbecke. Summary results of the simulation are included in the paper.

ROBBINS, H. and PITMAN, E. J. G. (1949). Application of the method of mixtures to quadratic forms in normal variates. *Ann. Math. Statist.* **20** 552–560.

The distributions of linear combinations of independent χ^2 random variables with possibly different degrees-of-freedom are obtained as mixtures of χ^2 distributions. This use of mixtures is further applied to the ratio of independent quadratic forms of normal random variables.

ROBERTS, D. M. (1978). *Statistical Auditing.* American Institute of Certified Public Accountants, New York.

This reference book describes the standard statistical techniques used by auditors. Attention is given to the special problems faced by the auditor, namely, that monetary errors may be confined to a relatively small proportion of the population. Some rules of thumb are suggested to guide the auditor in selecting an appropriate technique. A simple stochastic model is presented to describe the error-generating process. Based upon this model, the author suggests some modifications of the standard techniques to adapt to the case of rare monetary errors.

ROBERTS, D. M. (1986). Stratified sampling using a stochastic model. *J. Accounting Res.* **24** 111–126.

The author uses a model similar to Kaplan (1973) in order to develop a procedure to test materiality of the total overstatement. Stratification is used to improve the normality of the test statistic.

ROBERTS, D. M., SHEDD, M. D. and MACGUIDWIN, M. J. (1982). The behavior of selected upper bounds of monetary error using PPS sampling. *Symp. Auditing Res. IV.* The Center for International Education and Research in Accounting, Univ. Illinois, Urbana-Champaign, Ill. (Reviewed in *Auditing: J. Practice Theory* **2** 112, 1983.)

This paper describes the results of a simulation study of six PPS estimators used to compute the upper limit of monetary error. One bound represents a variation of the Stringer bound, two of the bounds represent variations of the Garstka-Ohlson bound and two represent variations of the used estimators based upon normal theory. The paper presents detailed analysis of the performance of the bounds as a function of the number of errors observed, thus permitting the reader to observe the behavior of using a combination of bounds. A limitation of the study was the fact that no randomization of the population order was made between simulation trials, thus raising the possibility that population order might be a factor in affecting the observed result.

SMIELIAUSKAS, W. (1986). A note on a comparison of Bayesian with non-Bayesian dollar-unit sampling bounds for overstatement errors of accounting populations. *Accounting Rev.* **61** 118–128.

This work compares the long run performance of various Bayesian and non-Bayesian bounds.

SMITH, T. M. F. (1976). *Statistical Sampling for Accountants.* Accountancy Age Books, London.

This text covers the major topics in statistical sampling for accountants and auditors. Additionally, the last chapter in the book (Chapter 14) describes the problem of sampling for rare events. Monetary units sampling is described as a technique for coping with the auditor's problem of determining the monetary error when those errors are rare. Criticism of monetary unit sampling is also presented, particularly as related to the effect of the selected dollar being a part of an account balance or transaction that represents the audit unit.

SMITH, T. M. F. (1979). Statistical sampling in auditing: A statistician's viewpoint. *Statistician* **28** 267–280.

The statistical problems that arise in auditing are summarized and the validity of dollar unit sampling is discussed. The Cox and Snell infinite population model is presented as the only available theoretical justification for DUS.

STRINGER, K. W. (1963). Practical aspects of statistical sampling in auditing. *Proc. Bus. Econ. Statist. Sec.* 405–411. Amer. Statist. Assoc., Washington.

This paper describes some difficulties of using statistical procedures based on normal theory in many auditing situations where monetary errors are rare. Although few details are given there is a brief description of the methodology now known as monetary unit sampling.

STRINGER, K. W. (1979). Statistical sampling in auditing. The state of the art. *Ann. Accounting Rev.* **1** 113–127.

It is fair to say that the author is most instrumental in introducing statistical sampling in auditing. This article reviews its historical development. He also predicts that use of statistical sampling, particularly of dollar unit sampling, will expand in auditing practice and calls for research to develop more efficient bounds.

TAMURA, H. (1985). Analysis of the Garstka-Ohlson bounds. *Auditing: J. Practice Theory* **4** 133–142.

This article comments on a property of the Garstka-Ohlson bound. It is demonstrated that the bound may not work because it does not take into account the skewness of the sampling distribution of the estimator.

TAMURA, H. (1988). Estimation of rare errors using expert judgement. *Biometrika* **75** 1–9.

A nonparametric Bayesian model is proposed using Ferguson's Dirichlet process to specify the prediction of the conditional distribution of the error. The distribution of the conditional mean of the error is obtained by numerically inverting the

characteristic function. The error rate is modeled by a beta distribution. The distribution of the mean error is derived by taking the expectation of the mean of the conditional error over the error rate. Numerical examples are given and comparisons with parametric models are discussed. The model is discussed in Section 2.7.

TAMURA, H. and FROST, P. A. (1986). Tightening CAV (DUS) bounds by using a parametric model. *J. Accounting Res.* **24** 364–371.

A potentially profitable application of computer-intensive data analysis is in approximating the small-sample sampling distribution. In this article the authors apply the parametric bootstrap to determine the sampling distribution of the estimator of the mean tainting. A power function is proposed for modeling the taintings. Their model is described in Section 2.6.

TEITLEBAUM, A. D., LESLIE, D. A. and ANDERSON, R. J. (1975). An analysis of recent commentary on dollar-unit sampling in auditing. McGill Univ. working paper. March.

This paper is a response to the two-part article in the October and November 1974 issues of *CA Magazine* by Goodfellow, Loebbecke and Neter. Issues under contention to which responses are made in this paper include the planning of sample size with dollar-unit sampling, the handling of over- and understatement errors, the method of sample selection, the evaluation of an upper bound and a comparison of monetary-unit sampling with the Stringer bound and line-item sampling with variables estimation.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

An extensive list of references on mixtures is provided in this well organized exposition of the subject.

TSUI, K. W., MATSUMURA, E., M. and TSUI, K. L. (1985). Multinomial-Dirichlet bounds for dollar unit sampling in auditing. *Accounting Rev.* **60** 76–96.

The multinomial bound developed by Fienberg, Neter and Leitch (1977) is difficult to compute. It is also subject to the definition of the S set. In this article, the authors develop a Bayesian approach to the problem. The model is described in Section 2.7.

VAN HEERDEN, A. (1961). Steekproeven als Middel van Accountantscontrolex (Statistical sampling as a means of auditing). *Maandblad voor Accountancy en Bedrijfshuishoudkunde* **11** 453.

This is the earliest known publication proposing the use of monetary unit sampling in auditing. The suggested evaluation technique involved regarding the monetary units within any audit as being either correct or in error. For example, if an audit unit with a recorded amount of 100 dollars had an audited amount of 80 dollars, the 80 dollars of the 100 dollars were regarded as correct and the last 20 dollars were regarded as incorrect.