W. EDWARDS DEMING

On a Problem in Standards of Auditing From the Viewpoint of Statistical Practice

THE ROSS INSTITUTE OF NEW YORK UNIVERSITY

JOURNAL OF ACCOUNTING, AUDITING & FINANCE

Volume 2, Number 3 PP. 197–208 Spring 1979

On a Problem in Standards of Auditing From the Viewpoint of Statistical Practice

W. Edwards Deming*

Every auditor relying on tests to form an opinion on a set of financial statements faces the possibility that he used a poor method of sampling or that he examined an insufficient number of transactions. Statistical theory has therefore been drawn upon to assist practicing auditors in the ex ante decisions they must make regarding the items to be tested, and certain statistical concepts have been incorporated into generally accepted auditing standards.

Whenever technical concepts are adopted across professional boundaries, a risk exists that the members of the profession from whom the concepts are taken will disagree with the application. In this article, an eminent statistician expresses such a disagreement with a detailed auditing standard of the AICPA. Editor

Introduction

The author writes from the standpoint of the theoretical statistician, that is, a statistician that guides his practice with the aid of statistical theory.[1] The author is not an economist, nor a doctor of medicine, nor an expert in transportation, not an accountant, but works with experts in these fields and in many others.

The immediate incentive for preparation of this article lies in the need for comment on the following recommendations made by an auditing standard of the American Institute of Certified Public Accountants, and to explain why these recommendations conflict with statistical practice.

> ".34 The auditor's judgment concerning the reliance to be assigned to internal accounting control and other relevant factors should determine the reliability level to be used for substantive tests.

---

* W. Edwards Deming, Ph.D., is a consultant in statistical studies in Washington.

[1] W.A. Shewhart, *Statistical Method From the Viewpoint of Quality Control* (Graduate School, Department of Agriculture, 1939).

Such reliability should be set so that the combination of it and the subjective reliance on internal accounting control and other relevant factors will provide a combined reliability level conceptually equal to that which would be used in the circumstances described in paragraph .32. Thus, the reliability level for substantive tests for particular classes of transactions or balances is not an independent or isolated decision; it is a direct consequence of the auditor's evaluation of internal accounting control, and cannot be construed properly out of this context.

".35 The concept expressed in paragraph .34 can be applied by use of the following formula:

$$S = 1 - \frac{(1 - R)}{(1 - C)}$$

Where

S = Reliability level for substantive tests.
R = Combined reliability level desired (e.g. 95 percent as illustrated in paragraph .32).
C = Reliance assigned to internal accounting control and other relevant factors.

This concept is illustrated in the following table, for which the combined reliability level desired is assumed, for illustrative purposes, to be 95 percent:

Auditor's Judgment Concerning Reliance to Be Assigned to Internal Accounting Control and Other Relevant Factors	*Resulting Reliability Level for Substantive Tests*
90%	50%
70%	83%
50%	90%
30%	93%" [2]

Recommendations Conflict With Statistical Practice

There is a frame of accounts to be examined. This is a stratum of what appear to be homogeneous accounts, which the auditor has set

[2] AICPA Professional Standards, "Precision and Reliability for Statistical Sampling in Auditing," App. B, AU Sec. 320B, p. 301.

aside for examination for possible serious faults, such as backup papers missing or found to be fraudulent if the examination be pushed hard enough. The frame is so large that adequate examination of every account therein is, in the judgment of the auditor, impracticable under the circumstances, and unnecessary. Examination of a sample less than 100 percent is his decision.

We are not concerned here with petty mistakes in the nature of nuisances, such as errors of posting or crediting or debiting the wrong account. These mistakes will eventually be resolved, with little effect on the total assets of record in the frame. If they are too numerous in the judgment of the auditor, he may express his concern on the matter, and may suggest better supervision through use of statistical controls.

Our concern here is the possibility of serious faults in the frame, the kind that could be characterized as fatal. We require some terminology for communication. An account in the frame that the auditor declares to be in serious fault will be a red account. An account that is cleared of all serious faults will be white.

We must remember that an account will be red or white depending on the methods and diligence that the auditor elects to apply. He will declare any given account to be red or white, and we accept his decision. It might be, for example, that further examination of an account on which nothing unusual has turned up so far might turn out to be fraudulent were the auditor to push further his examination. Thus, he might find nothing wrong if he were to accept the papers present, yet he might learn, if he were to push further, that the bank that issued a certificate of deposit is nonexistent. How far to push the examination is his responsibility.

We make the supposition here that the auditor's decision on an account is repeatable, that his performance is independent of the size of the sample (even though, in practice, performance will suffer on huge samples).

The auditor will examine a sample of size n. He will find no serious error, or he may find one or more. Another sample may well contain some other number of serious errors. There are N accounts in the frame, numbered 1, 2, 3, and onward to N. The serial numbers in the frame are unalterable. In my own practice, I may require a tape or microfilm to identify every serial number with a particular account. The completeness of the frame is essential. The sample of n will be selected by reading out n unduplicated random numbers between 1 and N (simple random sampling without replacement). Any random number selects

uniquely one account or case. The question is, how big should n be? Some notation and a diagram are necessary.

N	Total number of accounts in the frame
R	Number of accounts in the frame that the auditor would declare to be red, were he to examine them all. (This R is not related to the R in the quoted paragraphs.)
$P = R/N$	Proportion of red accounts in the frame
$Q = 1 - P$	
n	Size of sample, fixed
r	Number red in the sample, a random variable
$p = r/n$	Proportion red in the sample, a random variable

Cause-System, Frame, Sample

The diagram shows a frame of N accounts. These accounts were produced by a myriad of operations that we call the cause-system. From a given frame, we may draw any number of samples of size $n \leq N$ provided each sample be restored to the frame before we draw the next sample.

For any given frame containing R red accounts, the probability (Pr) of getting r red accounts in the sample of n will be

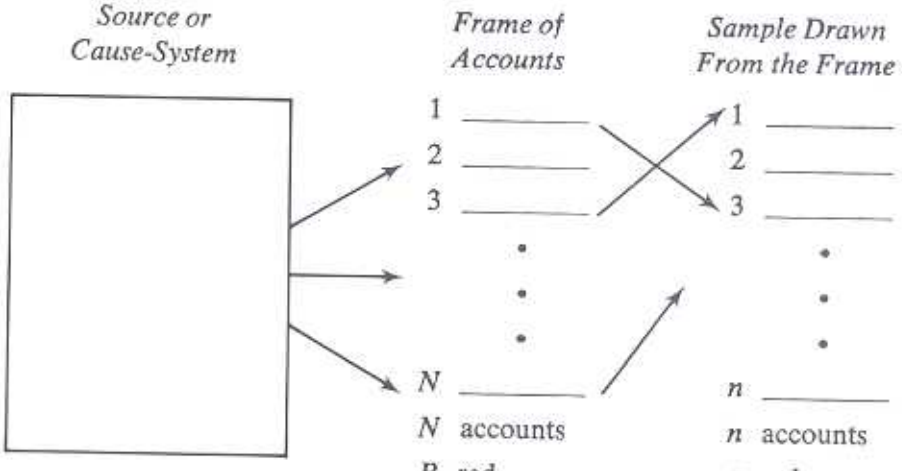
$$(1) \qquad Pr\ (r|R) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

$$(2) \qquad \underset{P}{\text{Cond}}\ Ep = P \qquad P \text{ constant}$$

$$(3) \qquad \underset{P}{\text{Cond}}\ \text{Var}\ p = \frac{N-n}{N-1}\,\frac{PQ}{n} \qquad P \text{ constant}$$

These conditional expectations exist for repeated samples from the same frame because the sampling procedure is stable, in statistical con-

trol, guaranteed by the use of random numbers and the presumed acceptable performance of the auditor.



If the cause-system be stable, let π be average proportion red in repeated production of frames.

N accounts	n accounts
R red	r red
$N - R$ white	$n - r$ white
$P = R/N$ proportion red in the frame	$p = r/n$ proportion red in the sample
$Q = 1 - P$	$q = 1 - p$

Suggestions for Statistical Aid to the Auditor's Problems

Some Difficulties

One might try to calculate from the sample in hand the probability that R could be as large as any specified value. This calculation requires knowledge of $\phi(R)$, which describes the cause-system that produced the frame. $\phi(R)$ is the relative frequency with which the cause-system, if stable, puts R red accounts into repeated frames. Thus, if the cause-system were stable, giving the distribution $\phi(R)$ in repeated frames, it would then be possible with the help of the sample to refine the probability that the frame contains R red accounts. The need for knowledge of $\phi(R)$ in order to calculate $Pr(R|r)$ was recognized by Thomas Bayes.[3]

[3] Thomas Bayes, "An Essay Toward Solving a Problem in the Doctrine of Chances," *Transactions of the Royal Society* (23 Dec. 1763). Reproduced in *Two Papers by Bayes* (W. Edwards, Deming, Ed., with a commentary by E.C. Molina) (Graduate School, Department of Agriculture, 1940; Washington: Hafner, 1965). An example in which the prior probabilities are known appears in Deming's *Some Theory of Sampling* (New York: John Wiley & Sons, 1950), p. 326.

Let there be r red accounts in the sample of n. The conditional probability $Pr(R|r)$ of R red accounts in the frame would then be

$$(4) \quad Pr(R|r) = \phi(R)\ Pr\ (r|R) / \sum_{R=0}^{N} \phi(R)\ Pr(r|R)$$

where $Pr(r|R)$ is the probability of drawing r red accounts into the sample from a frame that contains R red accounts.

$$(5) \quad \begin{cases} Pr(r|R) \to 1 & \text{for the particular value of } R \\ Pr(r|R) \to 0 & \text{for any other } R \end{cases}$$

It follows from Eq. 4 that

$$(6) \quad \begin{cases} Pr\ (R|r) \to Pr\ (r|R) \to 1 \\ \quad \text{as } n \to N \end{cases}$$

regardless of the form of $\phi(R)$. Thus, for the calculation of $Pr(R|r)$, one need not know the form of $\phi(R)$, provided n is big enough.

The fundamental difficulty with any attempt to calculate $Pr(R|r)$, even for large samples, is the statistical requirement of a stable cause-system for the production of red accounts in frame after frame. What we need in this problem is a method of inference that does not depend on the stability of $\phi(R)$. The likelihood ratio offers a solution, to be treated later.

We pause for three trivial observations. If the sample shows $r > 0$ red accounts, then regardless of the existence of $\phi(R)$, the frame contains at least r red accounts. Another obvious statement is that if the frame contains no red account, the sample will not show any either. (Later, we make use of this theorem in the likelihood ratio.) Further, if the frame is totally red, $R = N$ and $r = n$ in every sample.

Stable Cause-System in Manufacturing

It is important not to confuse a stable system in manufacturing with the problem that the auditor faces in the problem at hand. In the manufacture of product, the cause-system is a production process, stable or unstable. If stable, it has been made so with the help of statistical methods. Statistical control is not a natural state in manufacturing. It is an achievement. In the state of statistical control, the main characteristics of the product in lot after lot satisfy well enough chosen criteria of randomness. Examples of criteria of randomness might be (1) no runs of five or more points up or down, and no runs of five or more points above or below the median; or (2) no point out of control in the $\bar{x}$- or R-charts.

(See any book in the statistical control of quality.) In a stable cause-system, P has a predictable probability distribution $\phi(P)$. P has an expected value $EP = \pi$, and a variance.

In practice, there may be more than one type of defect. We could then work with a predictable probability distribution $\phi_i(P_i)$ for type i of defect, and we could let $\phi(P)$ be the probability distribution of defective items from all causes. The mathematics would be the same for two types of irregularity combined as it would be for either one alone.

In manufacturing, the quality-control records and statistical tests of measurements on samples of a product that was made under statistical control of the most important quality characteristics provide a far safer basis for purchase of lot after lot of the end product than tests of samples of the end product could provide. Moreover, there accrues to manufacturer and to purchaser the advantages of better communication and decreased cost. The average quality π of lots exists (by achievement, not by accident), and moreover, is known. The cause-system, under statistical control, has a known distribution $\phi_i(P_i)$ for any type of defect. This distribution is not a matter of judgment; it is a matter of record in the form of statistical evidence. It is of course necessary that the quality-control records be bona fide, that the inspectors know their job, and that instruments be compared with standards at frequent intervals. The quality-control records are not merely run-charts and statistical tests that signal special causes of variation. They include also records of action taken to find and remove special causes of variation so indicated, and evidence of success in removal. Small control samples of the end product are, nevertheless, taken as precautionary measures to ensure, for example, that the product being purchased is what was ordered; to test for such events as avoidable damage in transit; and to detect drift from standard of the manufacturer's measurements sufficient to render questionable acceptance of his quality-control records as evidence of quality. The statistical evidence of control must be indisputable on record.

In the state of statistical control, EP and Var P exist.

$$(7) \qquad EP = \pi$$

$$(8) \qquad Ep = EP = \pi$$

$$(9) \qquad \text{Var } p = \text{Var } P + E\,\frac{N-n}{N-1}\,\frac{PQ}{n}$$

In the special case where $\phi(R)$ is binomial,

$$(10) \qquad \phi(R) = \binom{N}{R}(1-\pi)^{N-R}\,\pi^{R} \qquad \text{[Binomial distribution]}$$

$$(11) \qquad \text{Var } P = \frac{\pi\ (1 - \pi)}{N}$$

whence Eq. 9 gives

$$(12) \qquad \begin{aligned} \text{Var } p &= \frac{\pi\ (1 - \pi)}{N} + \text{E}\ \frac{N - n}{N - 1}\ \frac{PQ}{n} \\ &= \frac{\pi\ (1 - \pi)}{N} + \frac{N - n}{N - 1}\ \frac{1}{n}\ E\ PQ \\ &= \frac{\pi\ (1 - \pi)}{N} + \frac{N - n}{N - 1}\ \frac{1}{n}\ \frac{N - 1}{N}\ \pi\ (1 - \pi) \\ &= \frac{\pi\ (1 - \pi)}{N} + \left(\frac{1}{n} - \frac{1}{N}\right) \pi\ (1 - \pi) \\ &= \frac{\pi\ (1 - \pi)}{n} \end{aligned}$$

exactly as if the sample were drawn with replacement directly from the cause-system. The lot (the frame) plays no role in the use of the sample to estimate the average level of the production process.[4]

Some Indefensible Methods in Auditing

Statistical control of operations, even if it is excellent, only provides a stable cause-system for petty, nuisance mistakes, reducing them to a low level. It cannot be relied on to detect systematic errors, amongst which could be a serious defect in internal control. Unfortunately, a numerical value for the reliance to place on the internal control in a particular situation cannot be stated with operational meaning. The recommendations cited[5] are accordingly not admissible statistical practice.

A solution must be found that does not depend on the existence of a stable cause-system. The solution does not lie in plans of acceptance sampling, for acceptance or rejection lot by lot for a succession of lots of a manufactured product. Plans of acceptance sampling afford only the guarantee of (1) a specified AOQL; (2) on the average, acceptance of only a certain proportion (β) of lots that do not meet the specified acceptable level of quality; and (3) detection of complete breakdown of the system of production. "AOQL" means the expected overall poorest possible quality (in the statistical sense) of the pool of a long

---

[4] Deming, note 3 *supra*, p. 258.
[5] AICPA Professional Standards, note 2 *supra*, p. 276.

succession of lots. The specified AOQL may be drastically exceeded in a short run of lots, and for one lot, one frame, the AOQL has no meaning.

Difficulties that attend the use of confidence intervals are accentuated in the problem at hand.[6]

The Likelihood Ratio[7]

Take a frame of 10,000 accounts. None of the 10,000 stands out far above the average in dollar value. If it did, it would attract instant attention. It would have been in another stratum for intense treatment, perhaps 100 percent, as statistical theory would indicate.

There is only one way out if the auditor must guarantee that the number of serious errors, if any, cannot be greater than half a dozen in a frame of 10,000. Only a detailed examination of every account, 100 percent, would find them. If this is not a practicable procedure, everyone concerned must accept some risk.

The statistical problem is to find what kind of valid statement can be made about this particular frame on the basis of a sample of size n that shows no serious fault (i.e., $r = 0$), and to evaluate the risk of this statement. This evaluation must retain all the information in the sample, and it must be acceptable to all people. It is the auditor's responsibility to decide just what risk to take.

The likelihood ratio may be helpful on the problem. It retains all the information that there is in the sample. It asks no questions about the cause-system. Moreover, it is easy to understand. We are interested here only in the case where $r = 0$: The auditor examines a sample of accounts and finds no serious error in the sample. We define the "likelihood ratio" as the ratio of the probability that the sample would show $r = 0$ were there R red accounts in the frame, to the probability that the sample would show $r = 0$ were $R = 0$. (This latter probability is of course unity.) This ratio of probabilities is as follows:

$$(13) \quad L = \left\{ \frac{\binom{R}{0}\binom{N-R}{n}}{\binom{N}{n}} \right\} \div \left\{ \frac{\binom{0}{0}\binom{N}{n}}{\binom{N}{n}} \right\}$$

[6] W.A. Shewhart, note 1 *supra*, at 62.

[7] Likelihood was invented by Sir Ronald Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*, Vol. 222 (1922), pp. 309-368. A readable treatise on likelihood is the book by A.F.W. Edwards, *Likelihood* (Cambridge University Press, 1972).

$$= \frac{\binom{N - R}{n}}{\binom{N}{n}} = \frac{(N - n)!}{N!} \frac{(N - R)!}{(N - R - n)!}$$

$$= \frac{N - R}{N} \frac{N - 1 - R}{N - 1} \frac{N - 2 - R}{N - 2} \cdots \frac{N - n + 1 - R}{N - n + 1}$$

$$= (1 - P)\left(1 - \frac{P}{1 - 1/N}\right) \times$$

$$\left(1 - \frac{P}{1 - 2/N}\right) \cdots \left(1 - \frac{P}{1 - (n - 1)/N}\right)$$

[n factors]

The table shows the likelihood ratio for a few values of P for several sizes n of the sample that show $r = 0$. The table was constructed with $N = 10{,}000$. The table would for practical purposes be unaffected were N taken as 20,000 or 50,000, or as low as 5,000. The number N is not important, unless it is only a few times as big as the sample. The correct likelihood would then be smaller than the likelihood shown in the table.

The Likelihood Ratio L as Defined in Eq. 13 Constructed With $N = 10{,}000$

The proportion of samples that will show $r = 0$ when the frame contains the proportion P of red accounts.

P is the Proportion of Red Accounts in the Frame

n	$P = 0$	$P = .001$	$P = .005$	$P = .01$	$P = .02$
50	1	.9511	.7778	.6043	.3633
100	1	.9043	.6043	.3642	.1313
200	1	.8170	.3633	.1313	.0169
300	1	.7373	.2172	.0468	.0021
400	1	.6647	.1292	.0165	.0003

Use of the table is simple, and its interpretation likewise. Take the entry .1313 under $P = .01$ (e.g., 100 red accounts in a frame of 10,000) for $n = 200$. The entry .1313 means that $P = .01$ is .1313 as likely as $P = 0$. Let there be a long series of samples of size $n = 200$ drawn from a frame in which $P = .01$ (one percent in serious fault), and another long

series of samples of size $n = 200$ drawn from a frame in which $P = 0$ (from which the samples can only show $r = 0$). The proportion of samples from the former frame that show $r = 0$ will be .1313 as numerous as those from the latter frame, or about 1:7 or 1:8. That is, roughly 6 times out of 7, on the average, or 7 times out of 8, a sample of 200 will contain at least one example of the serious type of error for the auditor to pursue further by other means. The likelihood ratio is not a confidence interval.

The table shows that, for the same sample, the likelihood of $P = .02$ is .0169, which is much smaller than .1313. These likelihoods might indicate to the auditor that as many as 2 percent improper serious errors in the frame is highly unlikely, although the possibility of one percent must be admitted. The interpretation of the table is his responsibility.

If we increase the size of sample to $n = 300$, the likelihood of $P = .01$ drops to .0468, or about 1 to 21. The auditor might wish to say that $P = .01$ is unlikely. If he increases the sample to $n = 400$, the likelihood of $P = .01$ drops to .0165, or about 1 to 60, wherefore he might wish to say that $P = .01$ is highly unlikely.

A percentage of possible red accounts in the frame translates directly into dollars. A proportion P of serious faults in the frame indicates that about the same proportion of dollars could be in trouble.

The presumption set up earlier that none of the accounts stands out far above the average must mean that the smaller the proportion of serious errors, the stronger will be the presumption that their effect on the financial statements is immaterial.

It is obvious that the likelihood of a very low proportion of red accounts in the frame is nearly as big as the likelihood for $P = 0$, unless the sample be made very large. Thus, for $P = .001$ and $n = 200$, the likelihood is .8170. A sample of $n = 400$ reduces the likelihood only to .6647, and a sample of $n = 800$ only reduces it to .4418 (not shown in the table).

There is thus no hope to detect by a sample a fault that occurs in some very small proportion, except by making the sample, for practical purposes, almost equal to the whole frame. A more efficient statistical procedure would be to segregate the frame into small clusters of accounts from a common source or of a common type, and to study first those that in the auditor's judgment could be in serious error. The remaining clusters could be pooled to form a new frame.

The table is made up for $r = 0$. The case of $r = 1$, in which a single account shows a serious fault, leads to a whole new approach in which

a table of likelihoods would be of no interest in statistical practice. If a single serious fault ($r = 1$) turned up in a sample, the auditor, if he followed statistical practice, would sequester all the accounts that could possibly have come from the same source and would study every one till he arrived at an explanation and conclusion. He might then turn to sampling the remainder of the frame. What is a suitably small likelihood ratio? The answer is up to the auditor. The statistician's job is to help him to understand the calculations.

Dr. Robert J. Brousseau and Dr. R. Clifton Bailey have kindly furnished a useful approximation for L as defined in Eq. 13. In logarithmic form, the approximation appears as

$$(14) \qquad ln\ L = n\ \ ln\ Q - \frac{n(n-1)}{2N}\ \frac{1-Q}{Q} - \frac{n\,(n-1)(2n-1)}{12 \qquad N^2}\ \frac{1-Q^2}{Q^2} - \ldots$$

where $Q = 1 - P$. The series converges. The t-th term is

$$-\ \frac{1}{t}\ \ \frac{1-Q^t}{Q^t}\ \ \frac{1}{N^t}\ \sum_{i}^{n-1}\ i^t$$

The three terms written produce with exact agreement the values of L in the table.