1961

Statistical sampling and independent auditing

Oscar S. Gellein

Follow this and additional works at: https://egrove.olemiss.edu/dl_hs
Part of the Accounting Commons, and the Taxation Commons

Recommended Citation
Haskins & Sells Selected Papers, 1961, p. 195-208
Statistical Sampling and Independent Auditing

by Oscar S. Gellein
Partner, Executive Office

Presented before a joint meeting of the Colorado and Wyoming Societies of Certified Public Accountants, Boulder, Colorado—May 1961

Here in the shade of the towering Flat Irons and at the edge of the immeasurable Rockies it should be easy to get into the spirit of thinking about quantity or size. Today I propose to consider the quantity of our work; I prefer to think of it as the quality of quantity. How do we decide on the right amount of audit work? Other professional men probably are not as concerned with quantity as we are.

The lawyer often, I am sure, asks himself: Have I sought the best remedy for my client? Have I found the right precedents, the right arguments, the best way to present them? Occasionally, he asks himself: Have I found enough precedents? But the number is not as critical as the kind.

The physician is keenly concerned with the diagnosis and often asks himself: Have I made the right tests at the right time? He doesn’t often ask himself: Have I made the right number of tests? Have I taken enough blood to make a satisfactory test? The amount of blood that he takes generally just doesn’t make any difference as long as he has the minimum amount required to make up the slides and the like. An important exception, of course, relates to medical research, such as occurred in establishing the effectiveness of the Salk serum.

The auditor, however, has to answer two questions in almost every engagement, both of which are critical. Have I applied the right procedures? Have I applied them to enough items to obtain the desired satisfaction? Quality and number are critical. The number of items to be tested is critical from both directions. Too few means that the basis of our opinion may be inadequate. Too many means that the economics of our practice is upset. It means further, and more importantly, that the auditor would be charging for work that does not add significantly to the satisfaction that he needs for his opinion.

Things have been stirring over the years in shaping the principles that underlie the financial statements and in selecting the procedures to be applied in gaining audit satisfaction.
Study of accounting principles continues apace. All of us know about the Institute's Accounting Principles Board and the project groups formed in connection with it to study such matters as postulates, broad principles, income tax allocation, pension costs, funds statements, leaseholds, nonprofit institutions, and the like. This is good. It is necessary.

Attention to auditing standards and procedures has had considerable emphasis too, particularly during the period 1939 to 1950. This attention has been directed mainly to the quality of the work—to standards and to procedures. The report or our opinion, too, has had attention—probably the greater attention. The quantity of work, on the other hand, has had little attention. Here we have said that the circumstances are determining. We have said further that the circumstances can be appraised only in subjective terms; that the experience of the auditor manifests itself in a proper judgment on the extent of tests. We have said further that the auditor's appraisal of client employees, of how they are organized and supervised, and of the personal qualities of the management, bears importantly on the determination of the right amount of work to do. I agree with all of this, but at the same time insist that even the most experienced auditor often needs some guides either to bolster his judgment or to help him form one concerning the extent of audit tests.

Auditors like others have whims and temperaments. These are not made uniform by experience, even the same experience. The view of any ten competent auditors as to the quantity of work to be done in connection with a particular procedure on a given engagement will spread out over a considerable span. None of us, I am sure, will deny that there is merit in attempting to eliminate the differences in the extent of audit tests relating to arbitrary factors such as whim and temperament. All of us, I am equally confident, wish to retain the differences relating to variations in conditions, such as those concerning internal control and the competence of client employees.

VARIATIONS IN EXTENT OF TESTS

Concerning these differences relating to arbitrary factors: Are they significant? I think so. You will recall that a few years ago the Institute issued a case study on the Extent of Audit Samples. Eight accountants of different firms were given facts and figures about a hypothetical business concern and were asked to suggest the extent
of testing appropriate in the circumstances. Some of the results were as follows:

Test of payrolls relating to hours, rates, deductions and the like: The suggestions ranged from 30 per cent of one payroll to 100 per cent of two payrolls.

Accounts receivable confirmation: The range was from 30% of accounts and 60% of value to 100%.

Examination of sales invoices: The range was from 20% of one month to 100% of two months.

Inventory test counts: The range was from 5% of value to 60% of value.

Inventory pricing: The range was from 10% of value to 50% of value.

Cash transactions (Check inspection and tracing to cash register; footings, etc.): The range was from 100% of one month to 100% of 3 months.

These variations mirror significant differences in audit time. I would be the first to agree that if these eight accountants had been developing their audit programs in the field with first-hand knowledge of the conditions, as well as the opportunity to make inquiries, the dispersion would have been lessened materially. At the same time, I would be the last to agree that the same accountants would have arrived at the same test sizes if they had been working with an actual case. Audit judgment like any other judgment needs some guides. An unguided judgment is not a judgment, it is a hunch. Of course, some people have better hunches than others, probably because they have formulated some hunch guides and know how to use them, and this means they are not hunches but are judgments.

All of my comments so far have had a single purpose, namely, to point up the merit of searching for better guides for answering the question: When has the right amount of work been done? The rest of my comments concern one such search. It is still a search, too. I am not prepared to say that the solution has been found. You can view my comments, if you like, as some of the observations of one who has experimented with statistical sampling, but has not made up his mind whether it offers a practicable solution to the broad problem of "how much work?" I do conclude, however, that it is a valuable tool for the auditor and that it offers enough promise of broad-scale application to warrant a real test.
My observations, accordingly, are tentative, but only because they have not been validated by all of the experimental exposure that I think desirable for final conclusions.

GENERAL CONCEPTS OF STATISTICAL SAMPLING

I do not intend to dwell on the statistical considerations, particularly the sophisticated ones. It has been my experience that an auditor does not need extreme depth of statistical knowledge to apply statistical sampling to his tests, but that he does need a general understanding of the principal concepts underlying it.

So let me comment on some of its features.

Actually, statistical sampling is a type of sampling that permits an objective statement (one expressed in mathematical terms) to be made about the results of the sample. Its basis rests in the laws of chance or probability. These laws are concerned with the probable occurrence of a given event as compared with the total possible occurrences of all events in a given situation.

For example, the laws of probability simply state what our intuition tells us about the probability of drawing, say, a spade from a deck of cards. We know that the probability is $\frac{13}{52}$, or $\frac{1}{4}$ since a spade can occur thirteen ways out of fifty-two possible occurrences.

Carry this trite example one more step. Skipping the mathematics (but they are the same in concept as in the foregoing illustration), the probability of drawing no spades in two draws is about 56%, of drawing one spade in two draws about 38%, and of drawing two spades about 6% or a total of 100%. This means, of course, that 56 times out of 100, no spades will appear in two draws and one or two will appear 44 times, but only 6 of the 44 times will both cards be a spade. Now from this situation let us be a bit more specific about this objective statement that one can make as a result of a statistical sample.

Suppose we are handed fifty-two cards, not a deck mind you, and are told that there may be some spades included among them. Then we are asked to draw two cards, one after the other, and make any statements we can about the number of spades included among the 52 cards. We do so. We draw first a heart, then a diamond. What can we say? Well, we are 100% sure that there are not more than 50 spades. How sure are we that there are not more than 49 spades? Quite sure, but not 100% so. What about 48? 47? 46? We are less
and less sure as the number decreases. What are the chances that there is not more than 1 spade—actually the probability is $\frac{1}{2}$, or a little less than 4%. But for each number from 1 to 50 there is a calculable degree of sureness. As a matter of fact, by backing up through the mathematics previously mentioned we see that we are 44% sure that there are not more than 13 spades. For if there were thirteen spades we would have drawn at least one spade 44% of the time, and if there were more than thirteen the percentage would be higher. These two figures—the 44% and the 13—illustrate the two basic measurements of statistical sampling, namely, reliability (or confidence level) and precision. The 44% is the measure of reliability and the 13 is the measure of a precision limit. As indicated previously, these measures go in pairs, there being an indefinite number of pairs in any given situation.

Let me hasten to add that I see few cases in auditing where we would be satisfied with 44% reliability. But I think it also obvious that the reliability can be increased to any desired degree by increasing the sample size, or for that matter by being satisfied with a higher precision limit. But this is enough of the mathematics. This part, I am sure, sounds like a lot of hocus-pocus, but it isn’t. It isn’t mysterious. It isn’t hypothetical, it is mathematical. Further, the matters of risk and satisfaction are inherent in testing whether based on statistical sampling or some other kind. Nothing new is introduced by statistical sampling except the means for measuring satisfaction.

RANDOM SELECTION

One thing needs to be said about the condition that must be present before precision and reliability can be measured meaningfully, and that is that the items must be drawn randomly. This means, loosely, that there must be no circumstances favoring the selection of one item over another. A simple way of doing it in many circumstances is to make a random selection of a starting point, and then to select every nth item following it; another to use random-number-digit tables.

I think it is at this point that some auditors begin to get bothered. They reason that judgment is relinquished when items are selected randomly—they feel they have lost control of selection. They argue that they are not inclined to permit any statistical device to select items for their examination. But this generalization is far too broad.
If there is an audit purpose in looking at a particular item, the auditor should look at it. There is no question about that. Random selection just shouldn't be used when there is reason to look at particular items and, as far as I know, nobody is saying that it should be so used. On the other hand, in almost every audit, there are a number of tests where representativeness is desired. We as auditors have said so in our statements on generally accepted auditing standards. Representativeness is achieved by a random selection. Let nobody be mistaken, statistical sampling applies only when examination of particular items is not required.

This doesn't mean that statistical sampling is not applicable when the auditor wishes to look at substantially all of the larger items. This can be achieved statistically, and as a matter of fact may be statistically desirable since it often results in greater reliability or tighter precision with a smaller over-all sample.

Let us look at some experiences with statistical sampling. The first case relates to a client application to accounting information. This is thrown in simply to get a feel for what we are talking about.

**PHYSICAL INVENTORY COUNT**

This was a case where a multi-unit manufacturing company sought to solve a problem relating to the taking of a physical inventory at one of its locations. The inventory, a work-in-process inventory, was not large in relation to the company's total inventory, but it was made up of about 40,000 lots located in a number of factory departments. Since this particular inventory was ordinarily taken during a vacation shutdown, the company was seeking some way to shorten inventory-taking time without sacrificing significant accuracy.

After studying the characteristics of the inventory, with statistical assistance, it was decided that statistical sampling could be applied. Preliminary study showed that satisfactory reliability could be obtained by counting all of the high-valued lots and 10 per cent of the low-valued lots, or about 4,500 or the 40,000 lots. The 10 per cent sample was selected in such a way that the precision and reliability were within the limits previously agreed upon as to the desired level of accuracy. Incidentally, there was a dry run on a deck of 40,000 IBM cards representing the inventory. This dry run showed that, as a matter of fact, the result of the sample was well within the limits of accuracy desired.
It is interesting to note the way in which the lots were selected randomly for the 10 per cent sample. Ahead of the counting, about 2,000 pairs of random numbers ranging from 1 to 20 were selected and placed in 2,000 sealed envelopes. Inventory crews were provided with pads of stickers or tags, numbered from 1 to 20, with sealed envelopes containing random numbers, and with specific instructions on the order to be observed in identifying lots. With this in hand, they proceeded to place on the first 20 lots, numbered slips identifying them. Then a sealed envelope was opened. If it contained, say, numbers 4 and 17, they counted the items in lots 4 and 17, but not the other 18. They then proceeded to the next 20 lots and repeated the process; this time the lots counted may have been numbered, say, 1 and 13. After completing the count, the low-valued lots counted were priced at standard costs; the total dollar figure thus obtained was multiplied by 10; and the physical inventory at standard cost was the figure so determined increased by the standard cost of the high-valued lots that were counted 100%. On this basis, the 4,400 lots that were counted out of the total of 40,000 lots represented about 60% of the dollar inventory. Incidentally, as a part of the review of the accuracy of the count a statistical sample was taken from the lots included in the first sample to estimate errors in counting. This in turn was used to determine whether recounts were necessary.

The goal sought was achieved; the time required for counting was shortened considerably. The cost of inventory-taking was not decreased materially, however, since the time required for inventory planning was increased. I found it interesting that the company believed it had a more accurate inventory on a sample basis than it had when the inventory was counted 100%. Be assured that sampling does not furnish more accuracy. The only way this could be true would be for the human error eliminated to be greater than the sampling error introduced.

APPLICATION TO AN AUDIT

In commenting on applications of statistical sampling to auditing I shall refer to one engagement where it was done on an experimental basis and comment on some of the problems that arose as well as some of my impressions. I shall mention some of the audit tests where it was applied. In doing so I do not intend to imply that there may not be other audit tests where it also can be applied.
Our approach has been to develop a method that provides us with a given degree of reliability that either the rate of errors or the dollar amount of them, does not exceed a specified amount, which statistically, of course, is the upper limit of precision.

RELIABILITY

My comments are not intended to explore the question of the degree of reliability sought. This is a matter of audit satisfaction and must be resolved at a policy level for a firm, or for that matter, for the profession as a whole. Without exploring this matter let me rather dogmatically state a few of my personal views. I think generally a firm will wish to fix the level of reliability to be sought in all engagements and to allow for varying conditions of materiality through the measure of precision. For example, I would think that as between audits A and B a firm would wish to have a constant degree of reliability, 95% or whatever, for similar conditions of internal control but probably different limits of precision to allow for differences in size and the like. In audit A, for example, $5,000 might be material while in audit B materiality might not be a consideration below $20,000.

This is not to say, however, that the statistical reliability should be constant among engagements. Instead, it means that, over all, considering the reliance that can be placed on internal control, the satisfaction gained from analytic review, and the statistical reliability to be obtained, a constant level of audit satisfaction should be sought. It is difficult for me to see any rational basis for different degrees of audit satisfaction among engagements, at least as to those where the standard short-form report is rendered.

PRECISION LIMIT

Precision is something else again. This is the measure of the limit of the error. Its meaning bears directly on materiality and audit purpose. It must be geared to the particular engagement and to the particular audit step. Since it is the expression of the maximum error that might be present, it is gauged by materiality in relation to financial position or results of operations, or perhaps to the amount of an error that might subject the auditor to criticism should the client find it rather than the auditor.
AUDIT OPERATIONS

Among the steps where we applied statistical sampling on one engagement were the following:

- Comparison of details of mail lists of cash receipts with the cash book
- Examination of paid checks and comparison of them with the cash book
- Testing items of payroll with underlying records
- Checking sales invoices against the sales records and against underlying records, such as price lists, contracts, and the like
- Confirming accounts receivable by direct correspondence with customers
- Checking the trial balance of vouchers payable to related subsidiary records
- Examining vouchers and checking them to the voucher register
- Testing inventory pricing and extensions

You will note that these procedures vary considerably as to nature and purpose. Further, they vary considerably as to the total number of items from which the test was to be selected.

As mentioned before, I do not want to get entangled in the statistical considerations. The application in the field required no complicated statistical formulas to be used. Instead, we were making use of tables that had been prepared for application generally, and only simple arithmetic calculations were required. But much preparatory work had been done.

Let me repeat. We were seeking satisfaction that the aggregate amount of any errors that might be present did not exceed a specified figure. We were not attempting to make estimates of the amount of the error, or for that matter of any thing else except the maximum error. For example, we did not set out to estimate from confirmation responses the dollar amount of accounts receivable. Instead, we sought to send out enough confirmations to satisfy us that any dollar error in accounts receivable did not exceed some predetermined figure. I submit that this is what we ordinarily do now when we make tests without statistical sampling, where there is some reliance on internal control. This is unlike the case that I mentioned concerning the physical-inventory count by statistical sampling. In that case the
statistical sample was used to furnish an estimate of the inventory. Now let us look at some of the problems and some of the conclusions. The first one may surprise you.

AUDITING PROBLEMS

Most of the problems encountered were auditing problems, not statistical ones. And they were ones that it does us good to look at, with or without statistical sampling. I suspect that one of the principal benefits of efforts to apply statistical sampling to auditing will be a sharpening of auditing purposes and ways of achieving them. I think, as a matter of fact, that it may require us to develop an articulated auditing philosophy. This will be a plus—a real plus.

What were these auditing problems? Well first, assuming we have resolved the problem of the degree of confidence that we want, we must set for ourselves a limit of precision—and this must be done for each test. Bear in mind, by the way, that there is a precision limit in our present practice of testing, we just don't know how much it is—but it is there. This must be approached by asking ourselves at least two questions, both of which are auditing questions: If there were an error in the accounts that a particular procedure might be designed to uncover, but did not, how long could it be and still allow a fair presentation of financial position or results of operations? Having resolved this one, we would still have to ask ourselves another question: Is there some smaller maximum amount that I must have in mind to avoid being criticized by my client if I don't find the error and he does? This, I think you will agree is a real problem, a practical one. For example, would you feel a greater need to find an error of a given size in cash than in, say, accounts receivable, or property, or cost of goods sold. I think you would agree there is an order of tolerable errors. So we must resolve the question of the maximum undetected error that we can live with, and state it in dollars. This is just materiality all over again.

Then we must identify the population from which the sample is to be drawn. This sounds statistical in nature, actually it is an auditing problem because it is keyed to the purpose of the audit test. For example, take the step of comparing subsequent disbursements with accounts payable at the examination date. Is the population the accounts-payable trial balance? Is it the subsequent dis-
bursements? Is it something else? The answer must be found in the
definition of the audit purpose. Why were we applying the procedure?
If we are doing it to satisfy ourselves that the recorded accounts
payable were subsequently paid in the regular way, then our popula­
tion will be the accounts-payable list at the balance-sheet date. If on
the other hand, as is more likely, we agree that the purpose is to test
for unrecorded liabilities, the population represents subsequent dis­
bursements to some cut-off date. Resolution of questions concerning
the precision limit and the population require penetrating inquiry of
audit purposes. Just once around the track with such inquiries on a
given engagement makes it apparent why so many juniors ask their
seniors why a particular audit procedure is applied and why so many
seniors respond with, “Don’t ask so many silly questions, just do
the work.”

Now for some impressions based on our experimental work.
I am content to call them impressions. Some of them, however, are
close to being conclusions; particularly the first one.

**IMPRESSIONS AND OBSERVATIONS**

*Will statistical sampling impinge on the auditor’s judgment?* An
unequivocal no. It will supplement our judgment, not weaken it, not
dilute it, not supplant it. None of the judgmental factors shift from
the auditor to the statistician. Definition of the audit purpose, fixing
the desired level of audit satisfaction, selecting the procedure to be
applied, doing the audit work itself, deciding whether to extend the
test or do something else—all of these matters remain with the
auditor. Statistical sampling can be one of the tools to assist him
in gaining his audit satisfaction and in knowing how much he has.

*Will it be necessary for all auditors to be trained statisticians?* No. It
will be necessary for those applying statistical sampling to have a
general knowledge of its concepts, and during the development stage
the profession will require the services of expert statisticians—as a
matter of fact, will require considerable assistance from statisticians.
In my opinion, this is where the profession is now. Pooling of the
results of the research should enable accountants to apply statistical
sampling with little more than a general knowledge of fundamentals.
It will certainly not be necessary for a statistician to accompany the
auditor in carrying out the field work. For example, the senior that
carried out some of the applications I referred to earlier was able to do so after only a few hours of instruction. But providing him with the tools took much doing.

Are there difficult mechanical problems to resolve in getting a random sample? I suppose the usual understanding is that a random selection can be made only by using random number digit tables that furnish one with hit-or-miss numbers which in turn are hard to use because of the way vouchers, checks, and other documents are filed. Our experience has been that with a little imagination a method of random selection can be adapted to the method of filing or listing. Frequently, for example, the client either has or can prepare a tab run that makes the random selection very easy. In other circumstances pages and lines on pages can be selected randomly. Further, in many cases the document numbers can be used even though they are not filed in numerical order. For example, a test consisting of all documents whose numbers end in 7 and in 3 preceded by an odd number would give a 15 per cent sample. The mechanics of random selection will not, I think, present insurmountable problems. In this regard, I find it comforting that the statisticians tell us that we are on the conservative side as to precision and reliability when we use a random start and each nth item thereafter rather than an unrestricted random selection.

Are there significant assumptions in the statistical method that should concern the auditor? There may be. For instance, in applications where an estimate is made by statistical sampling of some feature of a population, such as the average amount of an item, the measures of precision and reliability and the sample size are based on certain assumptions about the population itself. This has caused some auditor wag to say: “Well, this sounds like the fellow that got a prescription to stop his rather frequent sneezes and the instruction was: Take a spoonful five minutes before each sneeze.” It isn’t that bad. In the first place a preliminary sample, which becomes a part of the over-all sample, may be used to get satisfactory information for determination of sample size. But more importantly, it is possible to apply sampling to auditing where satisfaction is gained about the maximum possible error in such a way that no assumptions are made about the characteristics of the population.

Will statistical sampling result in smaller audit tests or will it show that the extent of our present tests is inadequate? I have little more than a feel for this. More experimentation will provide the answer. Tenta-
tively, I am inclined to think there will not be a significant change in the amount of work we are doing on an over-all basis. Some tests will be increased, some decreased. I am convinced statistical sampling will not result in significant reductions in audit time. Its value will relate, as to audit time considerations, to furnishing some guides on whether we are spending the right amount of time on particular audit operations of a given engagement and, accordingly, the right amount of time on the engagement as a whole.

I think it very significant that one thing you learn quite early is that the absolute size of the sample is more significant than its size relative to the size of the population. I fear that as auditors we have been too concerned with percentage samples, such as inspecting 10% of the items or some such figure. Generally, the result has been that we are doing too much on some large populations.

For instance, in a fairly large population a sample of 77 will give 90% assurance that no more than 3% of the items in the population are erroneous and a sample of 100 will give 95% assurance. The extra reliability that one gets with additional items is relatively small, and falls off quite fast.

Will statistical sampling cast auditing in a new form? It will not revolutionize the audit process. It isn't going to provide a statistical device whereby an auditor goes to line so and so and column so and so of table so and so and performs some mumbo jumbo to obtain answers as to whether the financial statements are fairly presented. It will be a valuable tool to have in your kit, however. Before much can be said about the likely frequency of its use, there are some difficult problems to be resolved in fitting statistical sampling into the internal-control circle of basing the extent of tests on the evaluation of internal control and of evaluating internal control on the basis of some tests.

Is statistical sampling applicable only to relatively large populations? Not necessarily. Obviously, there is some point at which the size of the population is so small that it just doesn't make sense to run a test at all, but instead to look at all of the items. This is true, with or without statistical sampling. In connection with the procedures that I mentioned earlier, the populations ranged from 200 to 48,000, most of them being in the range of 400 to 2,000. Large masses of items are not required for applicability of statistical sampling. On the other hand, its greatest usefulness may concern procedures where the mass is large.
SUMMARY

In summary then I think that there are places in auditing where statistical sampling will be useful, probably many places. I am not in a position to know how extensive the applications will become in auditing or how fast they will come about. I am convinced, however, that there are more pluses than minuses in the results likely to be gained by studying possible uses of it. I am inclined to think the pluses are quite a good deal larger than the minuses.