the AUDITOR

The Auditor e-Edition is published monthly by Paton Professional. It's contains exclusive auditor-related content that's not included in the print version of *The Auditor*.

December 2007 e-Edition

Feature Article

Statistically Rational Sampling Plans for Audits

by Steven Walfish

One of an auditor's biggest challenges is to select the best sample that represents the population being audited. Most auditors are trained to use rule-of-thumb methods, such as the 10-percent or square root of *n* plus one methods. Neither of these has statistical validity, but they're easy to implement, so auditors are inclined to use them.

Industries regulated by the Food and Drug Administration (FDA) have pushed to increase the statistical validity of audits. The FDA's Guide to Inspections of Quality Systems (QSIT Guide**)** presents different tables for selecting sample size using statistically valid methods. This article presents different strategies for selecting audit sample sizes using a risk-based strategy.

How do we collect our sample?

It's impossible to audit every document, record, or process. Usually, we need to make a decision based on our analysis of a sample document, record, or process. How we select the sample is important. The method must minimize bias, represent the population, and be of sufficient size to detect any abnormalities. There are several strategies that can help meet these objectives. The most common method is the simple random sample, in which each sampling unit has an equal probability of being selected. To use the simple random sample, you must specify every unit of the population and then take a random sample from it. This approach is similar to a raffle where the probability of any individual being selected is a function of the total population size. The drawback of this approach is that it's not always possible to know the scope and size of the population prior to the audit.

However, a second strategy, called “stratified random sampling,” allows an auditor to ensure that each major category is represented in the audit sample. This method requires that each category (or stratum) is specified, and that none of them overlap (i.e., items to be audited must fall in only one category). For example, you can select training records, batch records, complaints, and standard operating procedures as a stratum. The number of items sampled in each stratum doesn't have to be equal, which enables an auditor to concentrate on a few facets without overlooking others.

The third method is called “systematic sampling.” It's used when audits are time-based and an auditor wants to ensure that all time points are sampled adequately. Systematic sampling entails sampling every *nth* item in the audit. If we're auditing events during the last year, for example, we'd take a sample from every month in the last year. The difference between systematic sampling and stratified sampling is that each month would have the same number of records audited.

Which sampling strategy you use would depend on the purpose of the audit. Simple random sampling ensures that all samples are equally likely to be selected. Stratified random sampling ensures that each stratum is represented in the sample. Systematic sampling is a convenient sampling method for items where time is a factor of the audit.

Sampling error

Most auditors don't think about sampling error when conducting an audit. Sampling error can occur in two ways. The first is an audit during which no audit findings are cited, even though the quality system is truly ineffective. The second is an audit that has one audit finding in a quality system that *is* effective, but the auditor found it. If the sample size is either too small or too large, the audit result could either miss auditing findings or identify audit findings that don't represent the overall quality system.

Statisticians have a name for the two types of error: type I and type II. Type I errors are associated with failing an

audit when the quality system is functioning correctly. The errors cited from such an audit aren't indicative of the overall system. Type II errors are associated with passing an audit when the quality system is actually ineffective, but the audit failed to find the faults. Both of these errors have an effect on the audit process and product quality. Figure 1 shows these errors graphically.

	Actual Effectiveness of the Quality System		
Audit Conclusion		Effective	Ineffective
	Pass	Correct Decision	Type II Error
	Fail	Type I Error	Correct Decision

Figure 1: Sampling errors

Sample size calculations

Many audits either have too little or too many audit samples, which can lead to false conclusions. Typically, audits are performed using an attribute approach (i.e., number of records with an error), although during an audit the auditor might come across sampling plans for continuous data. For large sample sizes, the normal distribution can be used. Type I and type II errors must be specified in addition to the historical variability (i.e., standard deviation) and the difference deemed to be practically significant. Figure 2 gives the formula for calculating the sample size (n).

$$n = \frac{(Z_{\alpha} + Z_{\beta})^2 S^2}{\Delta^2}$$

Figure 2: Sample size formula

In the formula in figure 2, Z_a and Z_b are the type I and type II errors, respectively. S^2 is the historical variance. D^2 is the minimum difference to be detected from the null hypothesis. As the effect size decreases, the sample size increases. As variability increases, sample size increases. Sample size is proportional to risks taken.

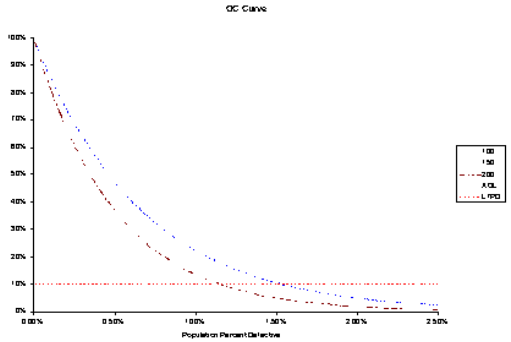
For example, if we have a type I error rate of 5 percent (Z_a=1.965), a type II error rate of 10 percent (Z_b=1.282), and a historical standard deviation of two (S^2=4), and we want to detect a difference from the null hypothesis of 0.4 (D^2=0.16), a sample size of 263 would be required.

Attribute sampling

Inspection by attributes involves classifying the sampled unit as either conforming or nonconforming. ANSI/ASQ Z1.4:2003 (which replaced MIL-STD-105E) is the most common standard used for inspection by attributes. It's indexed by the acceptable quality level (AQL), which is defined as "the maximum percent nonconforming (or the maximum number of nonconformities per hundred units) that, for purposes of sampling inspection, can be considered satisfactory as a process average." The AQL isn't lot or batch specific but rather a process average. The AQL is equivalent to the type I error in figure 1.

Under AQL sampling plans, if the process average is less than or equal to the AQL, then each lot has a high probability of passing inspection. The type II error is called the lot tolerance percent defective (LTPD). Although the standard selects the sample size as a function of the population size, the sampling statistics don't use the population size in calculating the AQL or LTPD. The increase in sample size as population size increases should be interpreted as decreasing the risk of failing a good quality system.

For any attribute sampling plan, an operating characteristic curve (OC curve) can be created that gives the probability of accepting a lot with a given defective percent in the population. The OC curve is determined by the sample size and the number of defectives allowed in the sample. For the purpose of audits, we can assume that the acceptable number of defects allowed is zero. Figure 3 gives the OC curves for different sample sizes with an accept number of zero.



Sample Size	AQL	LTPD
100	0.051%	2.279%
150	0.034%	1.524%
200	0.026%	1.147%

Figure 3: OC Curve (acceptance number = 0)

The AQL and LTPD decrease as the sample size increases. For a sample size of 100 units, for example, we expect to find no defects in the sample 95 percent of the time if the population percent defective is 0.051 percent or less. For the same sample size, we will find no defects in the sample 10 percent of the time if the population percent defective is 2.279 percent or less.

QSIT approach

The QSIT Guide outlines different sampling plans based on the upper confidence limit for the binomial distribution. This approach controls only the type I error, although the type II error can be calculated if necessary. Figure 4 gives the formula for the binomial distribution. For a given alpha level (i.e., type I error), maximum percent defective (p), and number of allowable defects in the sample, the formula can be solved for *n*.

$$\alpha = \binom{n}{x} \cdot p^{x} \cdot (1-p)^{(n-x)}$$

Figure 4: Binomial distribution

For example, at 95 percent confidence (the type I error is 5 percent) with a maximum percent defective of 30 percent and no allowable defects in the sample, we would require a sample size of 11. Figure 5 shows the tables from the QSIT document.

Confidence Limit .95≤	0 out of:	1 out of:	2 out of:
A .30 out*	11	17	22
B .20 out	18	29	37
C .20 out	17	25	34
D .15 out	24	45	65
E .10 out	35	52	72
F .03 out	72	115	157

Confidence Limit .99≤	0 out of:	1 out of:	2 out of:
A .30 out*	15	22	37
B .20 out	19	27	34
C .20 out	24	35	45
D .15 out	45	77	105
E .10 out	51	73	95
F .03 out	137	151	190

Figure 5: QSIT tables

Conclusions

Which approach is best for you? The answer is, “It depends.” It's clear that, based on the criticality of an audit, the sample size must reflect the associated risks. Rule-of-thumb methods like ten or the square root of *n* plus one does nothing to help mitigate risk, but rather gives convenient methods for selecting the sample size. Figure 6 compares the different sample size methods for attribute data.

Population Size	**10-Percent Method**			**Square root N +1**			**Z1.4**		
	Sample size	AQL	LTPD	Sample size	AQL	LTPD	Sample size	AQL	LTPD
1,000	100	0.052%	2.28%	33	0.160%	6.93%	80	0.064%	2.84%
1,0000	1,000	0.005%	0.23%	101	0.051%	2.25%	200	0.026%	1.15%
100,000	10,000	0.001%	0.02%	317	0.016%	0.72%	500	0.010%	0.46%

Figure 6: Comparison of sampling plans for acceptance number equal to zero

As shown in figure 6, just the sample size requirements in Z1.4 are a good compromise between the 10-percent and the square root of *n* plus 1 methods. The LTPD (type II) risk is the most critical risk to protect against. Accepting a quality system that's truly ineffective or inefficient is more costly than rejecting a good quality system. The method by that the sample is collected is as important as the sample size. Selecting the sample from the population correctly can minimize any bias from under-representing components of the system. Whenever possible, stratification should be employed to ensure that the population is well represented in the sample.

About the author

Steven Walfish is the founder and president of Statistical Outsourcing Services. He's spent the last 20 years of his career supporting the FDA-regulated industries. Walfish has a master's degree in statistics from Rutgers University and is the past chair of the ASQ Biomedical Division. He's a ASQ Certified Quality Engineer.

Click here to return to the e-Edition contents page.