Sampling in Assurance

Feb 2013

Introduction

Taking of samples is involved at every stage of the assurance process.

1) In assurance system design, when setting evaluation and surveillance frequency and in audit duration guidance.
2) At audit planning when deciding how the number of sites and which sites to visit, and which personnel to interview.
3) During an audit, when deciding how many samples of records, personnel for interview etc. are to be taken during the audit (most frequently driven by audit duration, not conscious decision).

Audit frequency and duration are interlinked. For example, two assurance schemes each run an audit program. One has 18 months between one and a half day long audits, the other 12 months between one day long audits. Which gives a better level of assurance? One has lower travel cost and time and the benefit of a more in-depth understanding of the factors contributing to audit risk due to longer time on-site, the other takes more frequent samples so sees the major risks more frequently.

Should sampling of sites to be visited and employees to be interviewed be randomised, or is visiting a number of sites in the same geographic area, or speaking to staff on the same shift to save time and travel cost acceptable?

Assuming that it takes an average five minutes to ask about an issue, get an answer / examine objective evidence and record it, an auditor may take 12 samples an hour. Allowing down time for opening and closing meetings, auditor review, and interruptions, in a full day audit perhaps 70 to 80 samples can be examined. Which 70 matters are the important ones to sample?

Most sampling is statistically based. Even if the methods of selecting a sample size and selecting the samples are not mathematically driven, statistics can be applied to the resulting sample to provide a hind-sighted view. Standards system owners may implement sampling programs that do not have a statistical base, but in doing so should bear this ability to apply the statistics in hindsight firmly in mind when designing those programs.

Judgemental Sampling

Judgmental, or non-statistical sampling, is appropriate to use in a limited range of situations:

1) When statistical results are not needed;
2) When there is a high degree of certainty that a conclusion can be drawn without further sampling;
3) The purpose is to take a survey in order to determine the necessity for and extent of substantive samples;

4) There is a desire to concentrate audit effort in a specific problem area revealed by a previous sample or other source of information (eg: risk assessment);

5) The entire population is very small and it would be quicker and easier to review all or most of the items in the population;

6) The area is very sensitive and there is no room for error (e.g. exact results are required and a 100% review is necessary).

Judgmental sampling is an appropriate method for activities such as:

- Investigating specific areas of concern – for example focussing on a limited number of standard clauses rather than all standard clauses.
- Sampling within an audit where the process has been validated, and the client is low risk so a decision on whether a full audit is needed or not can be made.
- Reviewing activities of particular interest or concern to determine whether more extensive testing is needed. For example a short unannounced audit of limited scope to see if the client conforms to requirements – and if not a full audit being held.
- Closing non conformities during a verification audit.

Judgmental sampling is often found within an audit, and can be used to sample a population. Low rates of sampling to verify self-declaration scheme operation is an example.

Judgmental sampling takes place as soon as random sampling is abandoned – for example when a sample of group members is selected to allow easy movement of auditors between sites in the same region.

Skewing sample selection to look for areas which are most likely to have non-conformity is common in judgmental sampling. This usually happens as a result of formal or informal frequency analysis. (Formal frequency analysis might be a Pareto analysis of non-conformities showing that 80% of the non-conformity happens on certain requirements. An informal frequency analysis might be collective view that “we know all clients have difficulty with this point”.)

Limits of Judgemental Sampling

If judgmental sampling allows a conclusion of conformity to be drawn without further sampling, no further work needs to be done. Positive results from a skewed sample could be interpreted to indicate that the population as a whole had a satisfactory level of conformity with requirements. However, this is a subjective judgment – no level of certainty can be applied to this statement.

If judgemental sampling identifies non-conformity, there is no way of knowing the frequency of non-conformity within the population sampled and hence the reliability of claims made about any member of that population. More work is required to do this.

As an example, 100 clients are sampled using judgmental sampling, with a sample of five (0.5 x square root n) skewed to those who are known to have prior problems. All five are found to have major non-conformity with requirements. This could be a result of the five worst clients being sampled, but statistically up to 100% of the population may be non-conforming. Further work must be done to increase certainty: a statistically relevant random sample should be taken to measure non-conformity levels in the population as a whole.

In the example above the results of further work may indicate that the population as a whole is performing acceptably, and skewing the sample identified the only five clients with problems. Equally it may indicate that the population was performing at an unsatisfactory level (e.g. the number of clients that are non-conforming is greater than is acceptable) and in that case further investigation is required.

The Assurance Code obliges standards system owners to develop an objective procedure for the actions to be taken on the discovery of non-conforming group members within a sample. That procedure can include both a quantitative and qualitative approach (e.g. is the group working to resolve the non-conformance?) and might include a table identifying the number of non-conforming group members that are allowed for different total sample sizes, for example:

Number of Group Members in a sample	Threshold Number of non-conforming members Allowed
2-5	1
6-10	2
11-15	3
16-20	4
21-25	5
26-30	6
31-40	7
41-50	9
51-60	11
61-70	13
71-80	15
80+	18

Source: adapted from ISO 2859 (via MSC CoC Methodology)

Representative Sampling

Sampling is defined as "*that part of statistical practice concerned with the selection of a subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference*."[1]

Sampling includes the size of the sample, and the frequency of sample taking. After a point a bigger sample size does not increase accuracy significantly (see Figure 5), but increasing audit frequency allows more snapshots in time to assess system stability.

The square root rule

Many certification bodies use a "square root rule" to determine sample size. This is a convenient and simple method to calculate the sample size. The rule specifies the sample to be a randomly selected sample equal to the square root of the lot size (in most cases the lot is the population under investigation), plus one. It has been adapted many times, and it is unclear where variations (e.g. 0.5 x square root) have come from.

[1] Wikipedia

The randomness of the selection of samples is critical. One problem with the rule, if non-conformities are not distributed evenly throughout the population it may result in acceptance where non-conformity still exists.

Setting sample size without referring to the population's risks has its own risks:

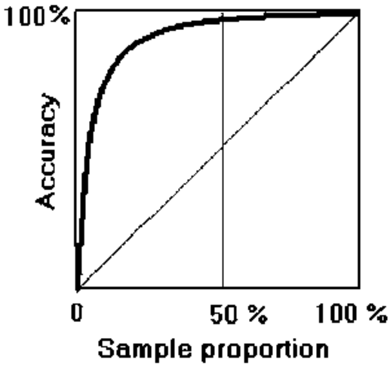
Determination of sample size as the square root of the lot size may create a sense of false security. Lot size alone is shown to provide an incomplete basis for determining sample size, whether by "square root" sampling or percentage sampling. Sampling plans based on quantitative statements of the risks involved are recommended[2].

For example, sampling a population with a high degree of diversity will require a larger sample size than sampling a relatively homogenous lot. Sampling a population with low control risk will allow a smaller sample than a population with higher control risk. Sampling plans should not exist in isolation from knowledge of the population, its possible strata and the risks associated with each stratum.

Sample size and accuracy

Figure 5 shows the relationship between sample size and the accuracy of information that the sample provides about the population sampled (or "lot").

Figure 1: Relationship between sample size and accuracy

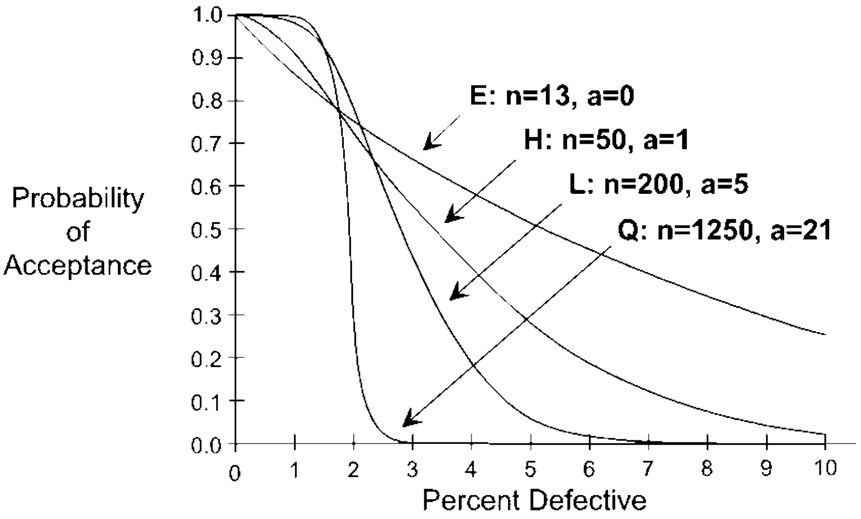


Accuracy can be achieved at relatively small sample sizes, provided that the samples are representative. Equally, too small a sample yields extremely low accuracy. Once more than about 10% of the lot is sampled, diminishing returns may set in.

A small sample can be used to test any size lot, but the results are not as reliable as with large samples. Figure 6 explains this. As the sample size (n in the figure) increases for an infinite population size, the probability of accepting a lot with a higher than specified number of defects (in this example 1%) decreases. For example, if the sample size is 13, there is a 50% probability that the lot will be accepted with more than 5% defects compared to the specified 1%. If the sample size is 50 there is a 50% probability that there are less than 3.5% defects in the population.

[2] Quoted from the abstract of Keith Borland (1950), "The Fallacy of the Square Root Sampling Rule," Journal of the American Pharmaceutical Association, 39, No. 7, p373-377

Figure 2: Relationship between sample size and probability of acceptance for AQL = 1



Sample sizes do not have to be related to the size of the population sampled. As sample size increases the chance that more than the acceptable number of defects will get through the audit / testing process reduces. Given this, standards system owners need to understand how accurate they wish their assurances to be before setting sample sizes.

Larger variations in populations require larger sample sizes

The greater the homogeneity of the population, the smaller the sample size can be. We take samples to get an estimate of some characteristic of the population – in most case conformity with standards. The accuracy of that estimate is proportional to the variability of the population divided by the sample size.

This means that if the population variability is high, we need either to:

- Take a larger sample, or
- Stratify the population – dividing the population into smaller populations and sampling each one. This does result in a larger total sample size.

Simple Random Sampling

When appropriate, the best sampling method to use is simple random sampling. Simple random sampling is appropriate to use in situations that require a single sample to be taken from a given population or a representative sample frame (the population to be sampled). The sample is then created by selecting randomly from the sample frame.

This method works fine when people in the sample frame are accessible. If it were to be applied to group assurance, then every group member (randomly chosen) would need to be audited; regardless of where they lived or if they are considered high or low-risk. The inability to audit any group member from the sample chosen would increase the margin of error of the audit.

Systemic Sampling

Systemic sampling is a method of randomly sampling each nth person (unit) in a list. It is useful for situations when it is difficult to identify units (people) in a sample frame.

For example, if there are 2,000 adult persons in a village and the sample size is 100, the researcher would divide the sample frame by the sample size (2000/100 = 20) to find the interval (20). The researcher would then organise the village sequentially (give everyone a number) and then (beginning with a random number between 1 & 20) draw every 20th person to survey.

This method only works when the sample frame is logically homogenous and the random draw is not missing or establishing some hidden order (e.g. drawing only women, and missing men in the survey).

Stratified Sampling

Stratified sampling is useful when there are smaller sub-groups that are to be researched: to achieve greater statistical significance in a smaller sample, and to reduce the standard error. To use stratified sampling, divide the population up into a set of smaller non-overlapping sub-groups (strata), then select a simple random sample in each sub-group.

Strata can be whatever groupings are useful for developing a homogenous group. Stratification helps reduce the standard error in cases when outcomes are expected to vary by strata. For example, if income is expected to vary by region, then stratified sampling with regional strata can help.

There are two subsets in stratified sampling:

- Proportionate stratified sampling takes the same proportion (sample fraction or sample size) from each stratum.
- Disproportionate stratified sampling takes a different proportion from different strata. This may be done to ensure minorities are adequately covered. If this is undertaken, it will be necessary to weight within-group estimates using the sampling fraction.

Q: When you stratify group members into relatively homogeneous categories (eg: large farms, small farms, cattle farms, cocoa farms) for sampling purposes, how do ensure that you sample at least the sq root throughout the entire group?

A: Conrad Guinea Rainforest Alliance Sustainable Agriculture Network: “It is generally possible to audit at least one farm of each category as part of the sample. It might get a little bit complicated when we are talking about a small group. However, in those cases, a farm will usually belong to at least two categories. For example, let’s say we have a group of 15 members. The sample size for the audit will be 4 farms… but at least 6 categories could be defined based on different criteria:

1) Small and medium-sized farms.
2) Those closer to the lake and those located in the hills.
3) Those who speak Spanish and those who speak a native language.
4) Those who use pesticides and those who don’t.
5) Farms that hire permanent workers and farms that don’t.

6) Farms that scored 80-85% in the previous audits and farms with a score of more than 85%.

It is evident that in this case it will be not possible to choose one farm from each category. However, a small farm (category 1) could also be located near the lake (category 2), in a Spanish-speaking region (category 3), with no use of pesticides (category 4) and with no external workers hired (category 5). This way, by choosing only one farm, we are covering 5 of the defined categories. In the practice, applying this kind of reasoning could improve the quality of the sample for small-sized groups."

Cluster Sampling

This method is useful when the sample frame is spread across a wide area such that simple random sampling would be difficult to implement and it is useful to reduce costs of survey implementation. To undertake cluster sampling:

- divide population into clusters (e.g. along geographic boundaries)
- randomly sample clusters

Survey all the units within sampled clusters. If this is not possible, then select a significant random sample and use the same selection rules in each cluster.

Some sampling notes:

- Representative (statistical) sampling allows conclusions about the entire population being sampled to be made
- Too small a sample has low certainty, and too large a sample results in diminishing returns
- Samples for diverse populations need to be larger – about 10% is a reasonable compromise.
- Sample sizes do not have to be related to the size of the population sampled. As sample size increases the chance that more than the acceptable number of defects will get through the audit / testing process reduces. Given this, standards systems owners need to understand how accurate they wish their assurances to be before setting sample sizes
- Standards systems should be precise about sampling and methods used – e.g. use of the word random implies its correct technical meaning; which is that each case had the same chance of being selected. Random does not mean, "just some cases I chose"