Chapter 14—Sampling

A. Purpose of Sampling

The systematic investigation of files is an integral part of market regulation. While it is rarely feasible to review all files of an examinee, the examination must nevertheless produce credible judgments about all files. For example, a judgment might assume the form of "claims processing errors for all claims in the state of x during period y exceeding z percent," even though all claims in a given state cannot reasonably be reviewed. Fortunately, it is not necessary to review all claim files in a given state to make such a judgment: applied statistics, a branch of probability theory in higher mathematics, provides an answer through sampling. It is important that both market conduct examiners and market analysts understand and properly apply sampling techniques. This chapter focuses primarily on sampling as it relates to proportions or percentages, although the same concepts generally apply to other statistics. For ease of reference, the term "regulator" will be used to refer to all insurance department staff that may use sampling.

Done properly, sampling permits valid generalizations or inferences about a wider population because the statistical properties governing the production of samples are known, via abstract mathematical probability theories, as well as countless empirical experiments and observations. The principles of sampling are not conceptually difficult; indeed, they are very nearly intuitive.

For example, the probability that the toss of a fair coin will result in "heads" is known to be 5. After 100 coin tosses, the proportion of tosses resulting in "heads" will be very close to 50 percent. If the proportion were 30 percent, one would likely reject the idea (or "hypothesis") that the coin is indeed fair, thus making a statistical inference[19] about the underlying process based on a sample. Because the coin has been deemed unfair, a valid generalization based on the sample (of 100 tosses) is that *future* coin tosses will also fail to produce a balanced ratio of head and tails. Of course, the inference could be wrong: in a tiny fraction of cases, even a fair coin will produce only 30 percent heads. But because the probability of this occurrence is remote (it would only happen in 0.002 percent of cases if the coin were really fair), one feels confident in making a judgment. Time and resources spent investigating the coin can be further reduced by reducing the sample size, or number of coin tosses, from 100 to 50. After 50 tosses the probability of 30 percent heads is 87 percent, and one can still be highly confident that the coin is unfair given the result.

Sampling is governed by the same principles of probability as those of a simple coin toss:

1. The probabilities of the underlying process must be known. In this context, the probability of selecting any given file from the entire population must be known. Therefore, sampling must be random. The relevant probabilities associated with non-random sampling techniques are generally unknown, and generalizations about a population from which the sample is taken cannot be made with a known probability (or confidence) of being correct.

[19] Statistical inferences are made by *rejecting* a proposition or hypothesis and thereby accepting a contrary, mutually exclusive alternative, with some known probability of being correct. Rarely is the process the other way around, whereby a statistical test affirmatively establishes a proposition. This is because failing to reject a hypothesis at a probability of x does *not* indicate that the hypothesis is correct with a probability of 1-x. Even an unfair coin will produce 50 percent heads some times, so that an outcome of 50 percent heads does not affirmatively establish that the coin is fair, even though the hypothesis that "the coin is fair" is not rejected.

© 2006-2007 National Association of Insurance Commissioners

2. Sampling methods should minimize the possibility of departures from randomness, or the introduction of statistical "bias." Significant bias will invalidate statistical inferences. For example, if a skilled magician could manipulate coin tosses in such a way that one outcome is more likely than another, the inference that the coin is unfair would be incorrect. Rather, the sampling process was "biased," or non-random.

3. Inferences from samples are never made with certainty, but only with some known and calculable probability of being correct. This probability is called *confidence*. After 50 tosses of a coin resulting in 30 percent heads, the coin can be declared unfair with a confidence of 99.13 percent (or 100 – 0.87).

4. The level of confidence is largely dependent on the size of the sample. Inferences about a coin can be made with greater confidence after 100 tosses compared to 50. A confidence level of 95 percent or greater is generally accepted by most professions as sufficient to support conclusions. In some instances, a 90 percent confidence level may be acceptable; however, for regulatory purposes, a 95 percent confidence level is the initial acceptance sample size recommended.

5. Inferences are made only about the population from which a sample is taken. An inference that one coin is not "fair" does not indicate that all coins are not fair.

These principles do not mean that errors found in a non-statistical sample are not errors, but it does mean that great care must be exercised to not suggest that the errors are representative of any broader population or process. Random sampling is universally recognized in all regulatory venues as a valid science. Findings based upon non-statistical sampling methodologies may be subject to legitimate challenges when the jurisdiction attempts to affect a resolution without being able to show that the errors are representative. For example, it is unlikely that any court would accept generalized findings based on improper sampling procedures.

This chapter is designed for a non-technical readership. Technical details that may not be of concern to the general reader can be found at the end of this chapter. In addition, the discussion here is confined to the fundamentals of sampling, and regulators may confront situations where the sampling strategies presented here require modification.

Those wishing to pursue the subject further can avail themselves of a variety of textbooks. Two recommended texts are the *Handbook of Sampling for Auditing and Accounting* by Herbert Arkin (McGraw-Hill, publisher) and *Statistical Methods for Rates and Proportions* by Joseph L. Fleiss (John Wiley and Sons, Publisher).

B. Sampling Generally

A sample should be a microcosm of the population or field from which it is drawn. It should be representative of all the relevant insurer processes under analysis, such as claims processing, cancellation notifications or complaint handling. The regulator should adhere to the methodology prescribed in this chapter to ensure that the sample is representative and that generalizations or conclusions about insurer processes are credible. Sampling should follow five steps:

1. Clearly and precisely define the population from which the sample will be taken. Any conclusions based on sample evidence can only be generalized back to the target population.

© 2006-2007 National Association of Insurance Commissioners

Population definition should include the following parameters: time period under review, functional definition of the process under review, location and origin of the process.

2. Determine a sampling strategy, such as the level of confidence necessary to support conclusions, and the appropriate sample size necessary to achieve the confidence level.

3. Examine the files in such a way that conclusions about each file can be quantified in binary form, such as "pass / fail" or "deviates from statute / complies with statute."

4. Calculate the percentage of deviations or failures present in the sample.

5. Based on the sample results, determine a numeric interval that contains the true or population deviation rate with a known level of probability or confidence. The "confidence interval" will form the basis of any conclusions about a process.

This chapter explains a two-stage sampling method that deviates slightly from this general format. An initial sample is taken which, due to its smaller size, is used only to determine whether further investigation is merited. If so, the regulator proceeds to a second larger sample capable of supporting conclusions about overall error rates with reasonable precision. This method is designed solely for efficiency, or as a labor-saving device, since in many instances the regulator can reasonably conclude that further investigation is unnecessary after reviewing only a relatively small sample.

In a number of states, it has become common practice to start with a standard size sample, such as 50 or 100 items, based on the overall field size of the matter under review. If the examinee challenges the error ratio that results from the standard size sample, the regulator must then consider pulling a larger sample. In some cases, the use of a standard size sample is a sufficient screening sample to detect anomalies. In all cases, a sample size should be selected which supports conclusions with a 95 percent confidence level.

In all cases, regulators should balance the costs and benefits of the sampling method used.

The most common calculations necessary to make inferences are included in various tables in this chapter. In addition, computer programs are available which can randomly select files, compute statistical formulas, develop probabilities, make complex computations and even make a sample selection. One such program, ACL, is described in further detail later in this chapter.

C. Sampling Methods

The validity of random sampling depends to a large degree on knowledge of the population. No one method works well in all cases and different methods should be tailored for the individual circumstances presented.

1. **Random Sampling.** The most widely known method of sampling is "random sampling." All items in the target population or field (before selection) have an equal chance of appearing in a random sample. No items or units have been "preselected" out of the field. Random selection may be attained through use of a random numbers table or a random numbers generator in computer software.

© 2006-2007 National Association of Insurance Commissioners

2. **Systematic Sampling.** Another method of sampling is to employ a systematic interval throughout a listing of all files. To sample 50 files drawn from 5,000 files, select every hundredth file after a random start number—say the third file. There are other methods for systematic sampling, such as changing the interval after each file selection, so that, on average, every one-hundredth file is selected.

3. **Stratified Sampling.** A variety of other sampling methods can be employed to adapt the principles of random sampling to more complex situations. For example, a regulator may have reason to focus on various subgroups, or *strata*, in an overall population. If the *stratum* is not large, its members may not appear in sufficient numbers in a sample of the overall population to support credible inferences about the subpopulation. *Stratified sampling* is designed for such instances. A stratified sample is obtained by performing a separate and independent random sample on each subpopulation of interest. The results are then combined into a single sample. For example, if a regulator is concerned about the impact of a specific processing center on overall claims settlement practices in a state, a random sample may be drawn from the center of interest, and a separate sample drawn from the remainder of claims in a state. The items in the resulting sample must be *weighted* to reflect the proportion of each subpopulation in the general or overall population before inferences can be made about the statewide claims processes.

   If only a single claims processing center is of interest, rather than the overall population, a random sample may be drawn solely from the specific center. However, since the remainder of the population was not sampled, any conclusions based in the sample should be confined to the subpopulation from which the sample was taken. Broader generalizations will not be valid.

D. Standards

The sampling method used must be subject to the following standards:

1. **Preselection and Statistical Bias.** Preselection can introduce statistical biases into the sampling procedure, which, if significant, will invalidate results. Generally, the term deals with the avoidance of files within a universe of files from which a sample is drawn. Note that the term does **not** pertain to the process of selecting a target subpopulation of interest, a strategy that is perfectly valid. Rather, the term refers to biases introduced into the sampling process *after* the target population has been defined. Once defined, the sample should be randomly selected from all of the files in the target population.

   Thus, homogeneity of the files in a sample should not be confused with preselection. Homogeneity is a means of defining the universe of files from which a sample will be drawn. The tests to be applied in a particular examination may in part define the universe of files from which the sample will be drawn. The distinction between preselection and targeting a specific *stratum* is made through a description of the universe of files. For example, if the test in an examination is focused on redlining for a particular geographic area, files outside of the particular geographic area would not be made part of the universe from which a sample is drawn. That does not represent preselection as used here, since no inferences based on the sample will be made about geographic areas that were excluded from the initial universe of files.

   A famous example of preselection resulting in significant statistical bias in a sample is the 1936 *Literary Digest* poll of voting intentions. The *Literary Digest* predicted a large victory for challenger Alfred Landon over incumbent Franklin Roosevelt, a result unambiguously refuted by

© 2006-2007 National Association of Insurance Commissioners

Roosevelt's victory with more than 60 percent of the popular vote. The *Literary Digest* had employed the same sampling techniques that had successfully predicted the outcome of prior elections: namely, pulling a sample from list of telephone numbers and registered vehicle owners. Unfortunately, the sampling universe (telephone and vehicle owners) was significantly unrepresentative of the target population (presumably consisting of all voters), since both telephone and vehicle ownership were highly correlated with income in the 1930s. Prior to the election of 1936, voting preference was not strongly correlated with income, so that, while the bias was present in prior samples, it did not significantly impact the validity of the survey. However, in 1936, the electorate became far more polarized along socioeconomic lines, rendering the statistical bias of the sampling so significant as to produce wildly inaccurate results. Contemporary pollsters take great pains to identify not only individuals of voting age or even registered voters, but *likely* voters, since the preferences of voters differ in significant ways from non-voters.

Preselection thus occurs due to the non-random selection of files within a given universe of files, whether or not the *purpose* is to attain a biased result. No preselection can be permitted. Generally, sample selection by the examinee should be avoided due to the difficulty in demonstrating that preselection has not occurred. Preselection is not the same as prior selection, where a sample is selected in advance of the arrival of the examination team. Should an Examiner-in-Charge (EIC) choose to select a sample in advance, precautions must be taken to ensure that the sample files are not disturbed prior to the examination review.

In a market regulation context, preselection is demonstrated by the regulator who avoids all files in the bottom shelf because they are inconvenient. The files on the bottom shelf may all belong to one claims person or underwriter, and that individual would thereby be deleted from the sample. Another example is the case where all complaints for a particular policy form are kept in the branch office and are consequently deleted because the regulator does not want to travel to that site. These examples are preselected based on location, but the same application is present for time, procedure or any of several other variables. The central point here is that after a target population has been defined, no selection biases should contaminate the sampling process such that some items in the target population have a different probability of being selected than other items. Such biases can render the sample unrepresentative and unsuitable for making inferences about the target population.

Preselection can also occur due to the use of company computers and programmers to develop a "pull list." If company programmers reduce a field of 500,000 policies to a list of 500 files from which the regulators make their selection of 50 files, there may be preselection. Examples of this might be where no files appear in ZIP code XXXXX, or in time frame May 11 to May 23, or for claims closed without payment. Regulators can guard against this outcome by reconciling data obtained during the examination with other available data sources, or via simple reasonability reviews of the data. For example, some insurance departments collect ZIP code data, which can be used to assess whether the pull list contains the entire population of interest. All states have access to statewide financial data, which may also be used to verify the accuracy of pull lists.

The EIC should note that it is his/her responsibility to ensure that no preselection has occurred. If a regulator places total reliance on the company, there would be no need for regulators to be there at all—and a self-report of the results of any sample drawn would be adequate. In all cases, the EIC should work closely with the company coordinator, system analysts and/or programmers to ensure that no preselection of files occurs.

© 2006-2007 National Association of Insurance Commissioners

2. **Confidence Level**. As discussed earlier, a confidence level is a measure of the probability that a conclusion about the true and unknown value in the overall population is correct, based on what is observed in a representative and unbiased sample. In many instances, the level of confidence is associated with a numeric interval within which, with a probability equal to the confidence level, the true value is likely to lie.

   Confidence is directly related to sample size, but it is also related to the true proportion of errors within a population of files. Larger proportions are associated with a higher level of sampling variability and, therefore, require larger sample sizes to support the same level of confidence as smaller proportions. For example, other things being equal, the confidence interval will be widest for proportions of 50 percent (or conversely, the given interval will be associated with less confidence). Smaller samples are required when the true proportion moves away from 50 percent in either direction, or toward 0 percent and 100 percent. For example, for large populations, a sample of size 1,067 is necessary to produce a 95 percent confidence interval of ± 3 percentage points when the population proportion is 50 percent. A sample of only 203 supports an estimate of the same interval at the same confidence when the proportion is reduced to from 50 percent to 5 percent (or increased to 95 percent). A regulator may have sufficient experience to know what proportions to reasonably expect for a specific process, and determine the minimum sample size necessary to support credible estimates.

   For the first-stage acceptance sample, a minimum confidence level of 95 percent should be selected. For the second-stage sample, the regulator should use discretion in selecting an appropriate confidence level, although it should never be less than 90 percent.

   While regulators may instinctively have negative feelings about certain company procedures, those instinctive feelings will not be valid in an administrative proceeding or in court unless findings can be shown valid with a high confidence level. A determination of the confidence level and margin of error should be made during the planning stage, prior to taking a sample. These two factors largely determine the appropriate sample size, and regulators should weigh the costs and benefits associated with increasing the sample size vs. acceptance of less precise estimates or a larger margin of error.

3. **Tolerance Level.** The tolerance level represents a critical threshold used during the initial acceptance sample to determine whether a process requires additional investigation. If the results of an initial sample cannot confidently rule out the possibility that the true processing error rate is above the tolerance level, a second sample of sufficient size to estimate the actual rate of processing errors should be taken. The tolerance level is thus used to provide parameters for a mathematical construction. This expression of tolerance has little to do with the real tolerance that a jurisdiction may have for error. For instance, the tolerance for deliberate violations of certain statutes may effectively be zero. The mathematical error tolerance standard used in this manner for all procedures except those tied to claim resolution should not be more than 10 percent. For all claim procedures, the error tolerance standard should not be more than 7 percent. The use of the 10 percent and 7 percent tolerance levels does not signify that the regulator is tolerant of that level of error. Nevertheless, the severity with which inadvertent and unintentional errors are viewed should decrease (significantly) as error rates fall below established tolerance levels. Smaller mathematical tolerances may be used, although sample sizes should be subsequently adjusted according the methods described in this chapter.

© 2006-2007 National Association of Insurance Commissioners

4. **Extrapolation**. Generalization or extrapolation of results beyond the field of files from which the sample is selected is not acceptable. If files are sampled from the Chicago branch underwriting office, results cannot logically be extrapolated to the branch office in Philadelphia. A sample can only be representative of the population from which it was drawn—and no other. Any alternative assumptions are very frail, insupportable and probably invalid.

E. Data Verification

In recent years, data verification processes have evolved into highly sophisticated, rigorous and organized systems for ensuring the integrity and accuracy of data. No amount of rigor in sample design can surmount data that is inaccurate: a valid sample drawn from inaccurate data will still produce invalid conclusions. A variety of data problems can introduce serious statistical biases and distortions into the sampling process. The examination process should incorporate a systematic investigation into the accuracy of data collected as part of the examination.

The most frequently used data verification procedures are related to *completeness, validity, internal consistency, duplicated or missing records and reasonability.* If a data problem cannot be remedied, procedures should be adopted to minimize the risk of statistical bias, and such procedures, along with their justification, should be explained in the examination report.

1. **Completeness.** Data from which a sample will be taken should include the entire universe of files or target population. To ensure completeness, such data should be reconciled with control totals, if available. Most states have access to a variety of data that can serve this purpose. All states can obtain statewide data from the financial annual statement, including aggregate annual premiums written and earned, losses paid and incurred and additional expense items. Where the population to be sampled matches that captured on the financial annual statement, the amounts should reconcile. For example, if the target population is all paid claims, the amounts in the examination data file should roughly reconcile with the paid loss amounts reported on the annual statement. Similarly, regulator complaint data provided by an insurer can be reconciled to each insurance department's recorded complaints.

   Reconciliation is a time-consuming, and thus expensive, process for insurance regulators and companies. Reporting systems and data call parameters change over time and it can be difficult to precisely reconstruct some records. Market regulation records are, by their nature, different from financial records. Reconciliation of market regulation records to the annual financial statement is a difficult and expensive process. The regulator should consider whether reconciliation is necessary in all samples. For example, if the regulator has a high confidence level in the company data, and the initial numbers provided are roughly consistent with annual statement data, it may not be necessary to precisely reconcile all data. On the other hand, if the regulator has evidence that the data provided from the company is incomplete or inaccurate, the need for reconciliation is increased.

   Many states collect data beyond the data available to all states. If an insurance department collects policy or exposure counts, these data can reliably verify the completeness of any analogous data provided during an examination. Each state should systematically determine which control totals may be available, and implement a quality control strategy that utilizes such data.

© 2006-2007 National Association of Insurance Commissioners

2. **Validity.** Data fields should be systematically checked to determine that all values are valid and that all codes used correspond to the reporting specifications. Validity is generally determined in a *prima facie* sense: values are wrong "on their face" in that the true value cannot logically be as reported. For example, data that include codes that are not specified on the reporting protocols are simply "wrong," and must be recoded. A payment reported under an automobile no-fault policy for an accident that occurred in an at-fault state will generally be incorrect.

   As with all data problems, data records containing invalid values should not be discarded, because doing so would pose a risk of significantly biasing the subsequent sample. Rather, every effort should be made to determine the true values, and then recode the data as necessary.

3. **Internal consistency.** Examiners should identify ways to ensure that each data record is internally consistent, such that values reported in different data fields are not logically contradictory to others. Similar to *validity,* inconsistency is determined on a *prima facie* basis: a data record is internally inconsistent when two or more values cannot logically be *simultaneously* correct. For example, if a data record for a private automobile insurance policy reported policy limits of $50,000 per occurrence, but the paid loss amount is reported as $70,000, the necessary conclusion is that one or both of these values are incorrect. Such inconsistencies, *when relevant to defining the universe to be sampled or to a process under investigation,* should be recoded to correct values prior to taking a sample.

4. **Duplication of data elements.** Duplicate items in a population from which a sample will be taken must be removed prior to sampling. The presence of a significant amount of duplicate data fields could introduce significant statistical bias into the sampling procedure. Random sampling is predicated on the fact that each item in a population has an identical probability of being selected. If an item appears three times in a dataset, the probability that it will be sampled is three times larger than that for a single item.

   Duplication is a particular challenge in performing analysis of accident and health carriers. The process to remove all duplications can be extremely challenging and time-consuming. In this type of examination, the regulator must balance the time and cost of attempting to remove all duplications with the information sought by the query. If the regulator has a high degree of confidence in the overall data provided, it may make sense to factor in the existence of duplicates.

   Duplication is defined with respect to the universe being sampled. For example, some insurers capture data by *claimant* rather than by *occurrence.* Three claims arising from a single automobile accident may appear in triplicate in a dataset. This does not constitute duplication if the intent is to sample the universe consisting of *all claimants.* However, if the target population consists of *all occurrences* from which claims arise, the duplicate records must be removed prior to sampling. Failure to do so could bias the sample in a number of obvious and not so obvious ways. For example, payouts for claims consisting of multiple claimants are very likely to be significantly higher than overall average payouts. There may very well be geographic correlates associated with the types of accidents likely to produce multiple claimants. Since it is extremely unlikely than all possible biases associated with duplication can be identified and corrected, the most prudent strategy is to remove the duplicates from the data prior to sampling.

5. **Missing Data Elements.** Missing data elements can potentially bias a sample in the same manner as duplicate items, if the data elements are relevant to the definition of the population

© 2006-2007 National Association of Insurance Commissioners

from which the sample will be taken. For example, if the target population is *paid claims*, but the dataset contains a portion of claims for which payment status is not recorded and so are excluded from the sample, the sample will potentially be biased. Bias will occur if the relevant characteristics of the subset of items for which the information is missing differs on average from the overall population of paid claims. Since both the likelihood and degree of such potential differences are generally unknown, potential bias cannot be ruled out in a non-arbitrary way.

Ideally, no relevant data elements should be missing, although some small amount is often tolerated in many data quality control systems. If the percentage of missing elements is believed to be tolerable, an explicit explanation should be provided in the examination report, including a specification of the percentage of data that it was necessary to discard.

6. **Reasonability.** Reasonability checks identify anomalous data values that deviate significantly from averages, or "what one would expect to see." Reasonability checks can be performed by examining the upper and lower extreme values for each data element, and comparing these values to the average value for the entire dataset. Values that appear unreasonable should be investigated to determine that they are correct. For example, an average annual premium for an automobile policy for an examinee may be $800, with the highest extreme reported as $5,000 and the lowest extreme reported as $30. Such values are not *prima facie* invalid, but they are anomalous to such an extent as to merit further investigation.

   Data elements that are missing or inaccurate, but which are not relevant to defining a population, drawing a sample, or to the process under investigation, can safely be ignored. For example, if it appears that a substantial proportion of paid loss amounts are reported incorrectly in the data, but the sampling universe consists of all closed claims regardless of payment status, the data inaccuracies will not bias the resulting sample. Sampling proceeds without respect to reported loss amounts, and all files from the population "all closed claims" still have an equal chance of selection. An exception to this rule may be those instances in which the data reporting is so inaccurate as to suggest that errors are systematic and that an examinee's core data handling capacities are significantly flawed.

F. Problem Data and Departures from Random Sampling

In some cases, complete and accurate data forming the universe of files to be sampled cannot be obtained. In these instances, the regulator has one of two choices:

1. Redefine the target population to accommodate the portions of the data that are complete, accurate and available. If the new target population is narrower than the original population, conclusions based on the sample can be made only about the narrower population. If the new target population is broader than the initial population, conclusions can still be made about the initial population if their members are in the sample in sufficient numbers.

   Examples:

   Narrower population: The initial desired population is all claims in a state. However, data from one claim processing center is found to be corrupt and cannot be repaired. The new population to be sampled then becomes all claims in a state, except those processed at the center producing the corrupt data. The subsequent sample indicates that 13 percent of claim files contain errors (+/– x percent). The only valid generalization from this sample is that 13 percent of the claims from the

© 2006-2007 National Association of Insurance Commissioners

centers sampled, not 13 percent of all claims in the state, contain errors. Nothing can be meaningfully said, based on the sample, of the processing center that was excluded from the initial population.

Broader population: The initial target population was all *paid claims*, but the data elements relating paid claims failed numerous data integrity checks and the problems could not be remedied. Because the data fails to reliably distinguish claims closed with payment from claims closed without payment, the population may be redefined as all *closed claims.* Paid claims, of course, are an element of the new population. If paid claims appear in the subsequent sample in sufficient numbers, generalizations about error rates associated with paid claims can still be valid. However, the confidence interval and margin or error for the subpopulation (paid claims) must be calculated separately based on their numbers in the sample, error rate and population size. This procedure is not uncommon, and the reader has no doubt seen polls in the popular press that provide estimates for subpopulations in an overall sample, such as those defined by ethnicity or gender.

2. If data is corrupt and the population cannot be meaningfully redefined in a way to effectuate the purposes of the examination, no valid sampling can occur. This chapter does not recommend any form of non-random sampling from a given population. As discussed throughout this chapter, departures from randomness can introduce significant statistical biases into the sample, rendering the sample unrepresentative of the general population. In addition, since the probabilities of non-random sampling outcomes are unknown, no calculable level of confidence can be attached to conclusions.

   Even in this instance, the regulator is not totally without recourse. Every effort should be made to investigate essential insurer processes—even in those instances when valid sampling cannot be performed. However, a strong caveat is that generalizations or extrapolations from findings will be invalid. Evidence of errors is strictly limited to the actual errors identified, and no claims about overall error rates can be made. If 10 violations are identified in 20 files that are non-randomly selected, the examinee can only be meaningfully cited for 10 discrete violations, not for 50 percent of the entire population of files.

   There may be many situations suitable for non-random investigative techniques. Random sampling is unnecessary for processes in which each discrete violation is the target of the investigation, rather than an overall violation rate in a defined population. For example, it may not be possible to obtain data for a population consisting of all advertising materials used in a state over a specified time period, and thus no way to randomly sample from this population. However, if seven violations are identified among the advertising materials that are available, the examinee is noncompliant in seven known instances, even though no knowledge is gained about the overall rate of noncompliance.

   The examination process is heavily reliant on random sampling, since market conduct audits are generally tailored to identify systemic process failures rather than discrete or incidental violations. Nevertheless, there may still be many instances in which other investigative techniques are appropriate. The caveats repeated throughout this chapter are intended to alert regulators to the lack of validity of generalized conclusions derived from non-random samples. Nothing in this chapter, however, precludes non-random investigative techniques, so long as generalizations are avoided.

© 2006-2007 National Association of Insurance Commissioners

G. Sample Sizes

As with the example of the coin, larger sample sizes lend themselves to greater confidence in conclusions. One would feel less confident basing conclusions about the fairness of a coin after only five flips as compared to 50. Because probabilities are known, a precise level of confidence can be calculated for any given sample size, if the sample is produced by a random process. Generally, statisticians accept a 95 percent confidence level as sufficient to support scientific findings. Very rarely are confidence levels below 90 percent considered "statistically significant."

The term confidence, in the statistical sense, is always related to a specified level of precision (or margin of error) of an estimate calculated from a sample. Confidence and precision are inversely related: other things being equal, less confidence is associated with more precise estimates. For example, in many popular presentations of sample results, an estimate is presented with a confidence of x percent and a margin of error (or confidence interval) of $\pm y$ percentage points. That is, the real (and unknown) population proportion is known to lie within the margin of error with a probability of x percent. Conversely, the probability that the true value lies outside of the margin of error is (100 – x percent), since the two outcomes are mutually exclusive and jointly exhaustive (i.e., the proportion must either lie inside or outside of the interval). A given sample will support a conclusion with a narrower margin of error, but with *less* confidence. For example, for a sample size of 500 for which a proportion is calculated at 50 percent, one is more confident that the population proportion is between 45 and 55 (or ± 5 percentage points) than between 49 and 51 (± 1 percentage point). Both precision and confidence are governed by sample size.

The sample size, confidence level and margin of error are always calculated in the context of a specific target population, and are not applicable to any specific subpopulation within the target population. For example, if an EIC attempts to sample all fire claims of 20XX for a company on a countrywide basis, and even if a rather large sample of 500 files is selected, very few files for any one jurisdiction will likely be present. Let's assume only seven fields were reviewed for Jurisdiction A. Although one can make generalized statements about the overall claims practices of the company countrywide, very little can be said of its practices in Jurisdiction A on the basis of only seven files. To make accurate statements on the procedures in Jurisdiction A, a much larger sample of Jurisdiction A claims must be reviewed.

Large all-purpose samples, intended to give blanket coverage over a wide range of variables, will usually fail in testing specifics. When gross categories are used (countrywide) little can be deduced about specifics. The same is also true of time sampling for a three-year period, then discussing a single year, and category "sampling" for all fire coverages, then attempting a discussion of homeowners policies. Thus, the regulator should carefully delineate the target population prior to the adoption of a particular sampling strategy. If necessary, a particular subpopulation can be oversampled or specifically targeted to produce sample sizes necessary to support conclusions, as per the discussion of stratified sampling.

H. Initial Sample

A minimum confidence level of 95 percent is used to make inferences from the small first-stage sample. Due to the relatively small sample size, the estimate made from the first-stage sample has a wide confidence interval (or margin of error). Thus, the small sample is insufficient to produce an accurate "point estimate," or a precise estimate of the true population proportion. Instead, the first-stage sample is designed to rule out the possibility that a given error rate is above a specified threshold. If this

© 2006-2007 National Association of Insurance Commissioners

possibility cannot confidently be ruled out, the regulator proceeds to the larger second-stage sample capable of supporting more precise estimates with a high degree of confidence.

The sample sizes indicated in the Acceptance Samples Table (AST) will produce a one-tailed lower 95 percent confidence limit of no more than 4.5 percentage points for claims, and 5 percent for non-claims. In some instances, the samples were adjusted somewhat to reduce the likelihood of "false positives," or instances in which a process that is in compliance will still trigger a second sample. In addition, sample sizes for non-claims processes are larger than the corresponding samples for claims.[20]

The "p-values" in columns E and J are equal to (100 percent confidence level), and represent the probability that the number of errors found in the sample would have occurred if the true or population error rate were at least equal to the tolerance level (i.e., 7 percent of claims, 10 percent of non-claims). For example, in a sample of 76 drawn from a population of 200, the probability of finding two *or fewer* errors is 4.8 percent *if the sample were taken from a population with an error rate of 7 percent.* A second sample is triggered when the p-value exceeds 5 percent. This is the point at which the confidence level (100 percent p-value) is less than 95 percent, and that the process error rate is below the critical threshold. Column F indicates when an additional sample is necessary for tests utilizing a 7 percent tolerance level, while Column K uses 10 percent. The p-values are cumulative probabilities derived from the hypergeometric distribution.

The AST represents the generally recommended sample size for most applications. However, the regulator has some discretion in the selection of the initial sample size. There may very well be instances in which greater precision is desired, particularly if examining a critical issue or process likely to represent a high probability of consumer harm. If sample sizes significantly different from those listed in the AST are selected, the regulator should be prepared to provide explicit justification with respect to the substantive issues being investigated.

Slightly larger samples can reduce the likelihood that an initial sample fails to detect a practice that is noncompliant (i.e., "false negatives"). Ideally the likelihood of false positives, where a compliant process fails the first round of sampling, should also not be high. Increased precision associated with larger sample sizes can reduce the likelihood of both types of inference errors (sometimes referred to as "false alarms" or "failed alarms"). If the regulator is less concerned about the risk of false positives, significantly smaller samples can be used.

[20] The 95 percent confidence limits become wider as the true population proportion increases, and are at their widest when the population proportion is at 50 percent. The value of the interval is symmetrical for proportions greater or less than 50 percent (i.e. the margin of error or confidence interval will be the same for proportions of 30 and 70 percent, 20 and 80 percent, etc. This result may seem counterintuitive, but it is attributable to the fact that the sampling variability of a proportion is greatest when the population proportion is 50 percent, and at its minimum when the true proportion is 0 or 100 percent (in which case, there would be 0 variability in the sample estimate across different samples, all of which would precisely replicate the population). Thus, a slightly larger sample is required for a tolerable error of 10 percent compared to 7 percent.

© 2006-2007 National Association of Insurance Commissioners

Acceptance Samples Table										
	Claims					Other				
A	B	C	D	E	F	G	H	I	J	K
Pop. (N)	Sample (n)	Error	% Error	P-Value, Pop=.07	Add for Claims	Sample (n)	Error	% Error	P-Value, Pop=.10	Add for Other
200	76	0	0.0%	0.1%	No	79	0	0.0%	0.0%	No
200	76	1	1.3%	1.0%	No	79	1	1.3%	0.0%	No
200	76	2	2.6%	4.8%	No	79	2	2.5%	0.3%	No
200	76	3	3.9%	14.9%	Yes	79	3	3.8%	1.4%	No
200	76	4	5.3%	32.6%	Yes	79	4	5.1%	4.7%	No
200	76	5	6.6%	54.9%	Yes	79	5	6.3%	12.2%	Yes
400	82	0	0.0%	0.1%	No	84	0	0.0%	0.0%	No
400	82	1	1.2%	1.1%	No	84	1	1.2%	0.1%	No
400	82	2	2.4%	4.8%	No	84	2	2.4%	0.4%	No
400	82	3	3.7%	13.6%	Yes	84	3	3.6%	1.6%	No
400	82	4	4.9%	28.3%	Yes	84	4	4.8%	4.8%	No
400	82	5	6.1%	47.0%	Yes	84	5	6.0%	11.4%	Yes
500	83	0	0.0%	0.1%	No	86	0	0.0%	0.0%	No
500	83	1	1.2%	1.2%	No	86	1	1.2%	0.1%	No
500	83	2	2.4%	4.9%	No	86	2	2.3%	0.4%	No
500	83	3	3.6%	13.5%	Yes	86	3	3.5%	1.5%	No
500	83	4	4.8%	27.9%	Yes	86	4	4.7%	4.5%	No
500	83	5	6.0%	46.0%	Yes	86	5	5.8%	10.6%	Yes
1,000	105	0	0.0%	0.0%	No	113	0	0.0%	0.0%	No
1,000	105	1	1.0%	0.3%	No	113	1	0.9%	0.0%	No
1,000	105	2	1.9%	1.5%	No	113	2	1.8%	0.0%	No
1,000	105	3	2.9%	5.0%	Yes	113	3	2.7%	0.2%	No
1,000	105	4	3.8%	12.0%	Yes	113	4	3.5%	0.7%	No
1,000	105	5	4.8%	23.4%	Yes	113	5	4.4%	2.0%	No
1,000	105	6	5.7%	38.2%	Yes	113	6	5.3%	4.8%	No
1,000	105	6	5.7%	53.0%	Yes	113	7	6.2%	9.8%	Yes
2,000	107	0	0.0%	0.0%	No	114	0	0.0%	0.0%	No
2,000	107	1	0.9%	0.3%	No	114	1	0.9%	0.0%	No
2,000	107	2	1.9%	1.5%	No	114	2	1.8%	0.0%	No
2,000	107	3	2.8%	4.9%	No	114	3	2.6%	0.2%	No
2,000	107	4	3.7%	11.7%	Yes	114	4	3.5%	0.7%	No
2,000	107	5	4.7%	22.5%	Yes	114	5	4.4%	2.1%	No
2,000	107	6	5.6%	36.7%	Yes	114	6	5.3%	5.0%	No
2,000	107	7	6.5%	52.2%	Yes	114	7	6.1%	10.0%	Yes

© 2006-2007 National Association of Insurance Commissioners

Acceptance Samples Table										
	Claims					Other				
A	B	C	D	E	F	G	H	I	J	K
Pop. (N)	Sample (n)	Error	% Error	P-Value, Pop=.07	Add for Claims	Sample (n)	Error	% Error	P-Value, Pop=.10	Add for Other
3,500	108	0	0.0%	0.0%	No	115	0	0.0%	0.0%	No
3,500	108	1	0.9%	0.3%	No	115	1	0.9%	0.0%	No
3,500	108	2	1.9%	1.5%	No	115	2	1.7%	0.0%	No
3,500	108	3	2.8%	4.8%	Yes	115	3	2.6%	0.2%	No
3,500	108	4	3.7%	11.5%	Yes	115	4	3.5%	0.7%	No
3,500	108	5	4.6%	22.1%	Yes	115	5	4.3%	2.1%	No
3,500	108	6	5.6%	35.9%	Yes	115	6	5.2%	4.9%	No
3,500	108	7	6.5%	51.2%	Yes	115	7	6.1%	9.8%	Yes
5,000	108	0	0.0%	0.0%	No	115	0	0.0%	0.0%	No
5,000	108	1	0.9%	0.3%	No	115	1	0.9%	0.0%	No
5,000	108	2	1.9%	1.6%	No	115	2	1.7%	0.0%	No
5,000	108	3	2.8%	4.9%	Yes	115	3	2.6%	0.2%	No
5,000	108	4	3.7%	11.6%	Yes	115	4	3.5%	0.8%	No
5,000	108	5	4.6%	22.2%	Yes	115	5	4.3%	2.1%	No
5,000	108	6	5.6%	36.0%	Yes	115	6	5.2%	4.9%	No
5,000	108	7	6.5%	51.2%	Yes	115	7	6.1%	9.9%	Yes
10,000	109	0	0.0%	0.0%	No	116	0	0.0%	0.0%	No
10,000	109	1	0.9%	0.3%	No	116	1	0.9%	0.0%	No
10,000	109	2	1.8%	1.5%	No	116	2	1.7%	0.0%	No
10,000	109	3	2.8%	4.8%	No	116	3	2.6%	0.2%	No
10,000	109	4	3.7%	11.3%	Yes	116	4	3.4%	0.7%	No
10,000	109	5	4.6%	21.6%	Yes	116	5	4.3%	2.0%	No
10,000	109	6	5.5%	35.2%	Yes	116	6	5.2%	4.7%	No
10,000	109	7	6.4%	50.2%	Yes	116	7	6.0%	9.5%	Yes
20,000	109	0	0.0%	0.0%	No	116	0	0.0%	0.0%	No
20,000	109	1	0.9%	0.3%	No	116	1	0.9%	0.0%	No
20,000	109	2	1.8%	1.5%	No	116	2	1.7%	0.0%	No
20,000	109	3	2.8%	4.8%	No	116	3	2.6%	0.2%	No
20,000	109	4	3.7%	11.3%	Yes	116	4	3.4%	0.7%	No
20,000	109	5	4.6%	21.7%	Yes	116	5	4.3%	2.1%	No
20,000	109	6	5.5%	35.2%	Yes	116	6	5.2%	4.8%	No
20,000	109	7	6.4%	50.2%	Yes	116	7	6.0%	9.6%	Yes
50,000	109	0	0.0%	0.0%	No	116	0	0.0%	0.0%	No
50,000	109	1	0.9%	0.3%	No	116	1	0.9%	0.0%	No
50,000	109	2	1.8%	1.6%	No	116	2	1.7%	0.0%	No
50,000	109	3	2.8%	4.8%	No	116	3	2.6%	0.2%	No

© 2006-2007 National Association of Insurance Commissioners

Acceptance Samples Table										
	Claims					Other				
A	B	C	D	E	F	G	H	I	J	K
Pop. (N)	Sample (n)	Error	% Error	P-Value, Pop=.07	Add for Claims	Sample (n)	Error	% Error	P-Value, Pop=.10	Add for Other
50,000	109	4	3.7%	11.4%	Yes	116	4	3.4%	0.8%	No
50,000	109	5	4.6%	21.7%	Yes	116	5	4.3%	2.1%	No
50,000	109	6	5.5%	35.2%	Yes	116	6	5.2%	4.8%	No
50,000	109	6	5.5%	48.8%	Yes	116	7	6.0%	9.6%	Yes

I. Additional Sample

If the initial acceptance sample indicates that an additional sample is necessary to more precisely estimate the level of error in the field of files from which the sample was drawn, several options are available. There are a variety of ways to select such an additional sample. The sampling method selected should be described in the examination report. In conformity with generally accepted practice, the report should also include the confidence limits associated with any estimate.

J. Sampling Topics and Tables

1. **Sample Sizes**

   Numerous software packages can easily calculate necessary sample sizes. Alternatively, sample sizes can be estimated with the formula presented in Section M, although the formula is only an approximation to the more complex algorithm used to produce the table, and which is implemented in most auditing software.[21]

   Sample size for testing proportions, such as error rates in a population of files, is governed entirely by four parameters:

   a. **Population size:** The larger the population, the larger the necessary sample. When the population is sufficiently large, further population increases have minimal impact on sample size.
   b. **Desired margin of error or precision:** Sample size is inversely related to the margin of error. The smaller the desired margin, the larger the necessary sample.
   c. **Confidence level:** More confidence requires larger samples.
   d. **The (unknown) error rate or proportion in the population to be estimated:** Necessary sample sizes are largest when the actual error rate in a population to be sampled is 50 percent and declines as the error rate approaches 0 percent and 100 percent (see below).

   Of these four parameters, values for only two are established by the regulator: the margin of error and the confidence level. These two parameters can have a significant impact on necessary

[21] The table was produced in SAS via an iterative algorithm that employed the cumulative hypergeometric probability function (SAS function *"probhypr"*). Most auditing software generates sample sizes using the same or closely similar probability distributions. Because the hypergeometric calculations are complex and labor intensive, the normal approximation to the hypergeometric is often employed when the sample sizes must be calculated manually. This is the formula presented below following the table. Since the formula is only an approximation, sample sizes produced by it will differ somewhat from those displayed in the table, as well as sample sizes generally returned by computer software.

© 2006-2007 National Association of Insurance Commissioners

sample sizes. Regulators should carefully weigh the costs and benefits when making sampling decisions, such as whether gains in precision or higher confidence or merited by the cost of producing and investigating a larger sample of files.

In selecting a sample size, the regulator must estimate the true population proportion, or the actual percentage of files in the population that that contain errors. Differences in the sampling variability associated with different proportions can be substantial. For example, when sampling from a population of 5,000 files, a sample size of 200 is necessary to obtain a margin of error of 3 percent when the population proportion is 5 percent. However, the necessary sample size to achieve the same margin of error increases to 917 when the true proportion is 50 percent. If the initial guess about the population proportion is far off, the resulting estimates produced from the sample may have a significantly wider margin of error than anticipated. One conservative approach is to always select the sample size associated with a proportion of 50 percent. However, a significant amount of labor can be saved by using any information available that indicates that the true population proportion is greater than or less than 50 percent. For example, the estimate produced from the initial acceptance sample may be used in calculating the subsequent sample size. Final confidence limits must be calculated *after* the sample is obtained, using the sample proportion as a substitute for the (unknown) actual proportion.

K. Considerations for Selecting Sample Sizes

The rationale for the two-stage acceptance sampling technique discussed in this handbook is that the possibility that a process exceeds a specified error rate can be ruled out without having to draw a large sample in every case. In some instances, a small sample can effectively identify insurer processes that are likely to be compliant, even though the sample cannot produce very precise point estimates of the actual population proportion since the confidence interval or margin of error will be large. For example, the recommended stage-one sample sizes in the AST are designed to accommodate a confidence level of 95 percent, with the corresponding (one-sided) confidence limit (or margin of error) of 4.5 percent for claims and 5 percent of non-claims.[22]

However, there is an additional decision risk associated with the first-stage sample. While one can be reasonably confident that a process is compliant if the sample proportion is less than the lower bound of the confidence limit, the converse is not true. The fact that a sample proportion exceeds the lower confidence limit does *not* indicate that a company process is *noncompliant*. Rather, all that is determined in this situation is that the possibility that the process is noncompliant cannot be ruled out with much certainty (but it is not thereby "ruled in," as it were). The stage-one sample is, therefore, generally unsuitable for making a determination that a process is noncompliant.

There are, therefore, two types of risks associated with inferences based on the initial sample. First, the process may in fact be compliant, but the process fails the initial test, leading a regulator to draw the larger second sample. Alternatively, the process may be noncompliant, even though the sample indicates that it is compliant. This second probability is minimized by use of the 95 percent confidence limits, but the risk is not reduced to zero. Statisticians call these types of incorrect conclusions *Type I* and *Type II* errors:

[22]A one-sided or "one-tailed" 95 percent confidence interval is essentially a one-sided interpretation of a two-sided 90 percent confidence interval. This is a valid interpretation since inferences are made only about whether the sample error rate exceeds the lower bound. The likelihood that the true value exceeds the upper bound is not relevant to the decision at hand.

© 2006-2007 National Association of Insurance Commissioners

1. **Type I Error:** "False alarm" or "false positive." A "null hypothesis" is rejected when it should be accepted. For regulatory purposes, this error occurs when a regulator proceeds to the second larger sample when, in fact, the process is compliant.

2. **Type II Error:** "Failed alarm" or "false negative." A "null hypothesis" is inappropriately accepted. For example, the insurer process passes the initial test and is not further investigated, even though the process is not compliant.[23]

A Type I error results in wasted time and resources, in the sense that a large sample is gathered to investigate a company process that was, in fact, compliant. A Type II error leads to a failure of regulatory oversight, in that problem areas of company operations will remain uninvestigated. Unfortunately, there is a strict trade-off associated with the two categories of inference errors: *for a given sample size, minimizing the risk of a Type I error maximizes the risk of a Type II error, and vice versa.* However, inference risks are calculable, and can be managed by altering decision rules for a given sample size. Alternatively, the risk of both types of errors can generally be reduced by increasing sample sizes. If the initial sample size is substantially increased, the whole rationale of two-stage sampling is defeated.

The trade-off between Type I and Type II risks might be clarified by a more mundane example. If the sensitivity of a smoke alarm is calibrated too high, there is a high probability of "false alarms." The alarm may sound in the presence of normal environmental smoke, such as that produced from cooking. Clearly, it is unlikely that the alarm will fail in the event of a house fire, but it is also very likely that a high number of false alarms will reduce the alarm's efficacy. In response, a frustrated homeowner might decide to remove the battery, thus reducing the risk of a Type I error to zero. In the event of a real fire, the probability of a Type II error is thereby increased to one, since the now powerless alarm will necessarily fail to detect a hazardous fire.

A rational sampling approach should carefully balance the costs and benefits associated with each type of risk, such as regulatory resources diverted from noncompliant areas or additional expense associated with unnecessary sampling versus the potential consumer harm resulting from regulatory oversight failures. Indeed, regulators may rationally adopt differing sampling strategies to alter Type I and Type II trade-offs depending on the context. For example, given a company process for which failure would entail a high risk of consumer harm, a regulator may tolerate an elevated Type I risk in order to reduce a Type II risk.

Researchers are generally concerned with Type I risks to an extent that Type II risks are commonly ignored in a wide variety of research fields. When Type II risks are made an explicit part of research design, a level of 20 percent is generally considered acceptable, although often a much higher level is tolerated (compared to a 5 percent maximum for Type I risks).

Regulators, however, are much more concerned with Type II risks, or the risk of failing to detect a noncompliant process (a "failed alarm"). Therefore, the risk trade-off associated with permitting a larger number of errors in the sample is generally unacceptable. As such, it is generally preferable to negotiate

[23] Strictly speaking, the "null hypothesis" in this handbook is that "the true proportion is greater than 7 (or 10) percent" rather than "the true proportion is less than or equal to 7 percent." Thus, the terms "Type I" and "Type II" above should really be reversed. For expository reasons, a terminology consistent with the verbal meaning of the terms "false alarm" or "false positive" is adopted to avoid conceptual confusion. In reality, a Type I error or "false positive" in this context is the erroneous rejection of the hypothesis that a process is noncompliant (or H0: p>.07), though this might be better thought of conceptually as a Type II error ("false negative" or "failed alarm").

© 2006-2007 National Association of Insurance Commissioners

inference risks by adjusting sample sizes rather than altering decision rules for a given sample size. In many instances, efficiency gains can be obtained with modest sample size increases, which can reduce the risk of *both* types of inferences.

Smaller Samples

One method used by some auditors (see, for example, the *Financial Condition Examiners Handbook*) utilizes much smaller samples than the sample sizes recommended by this handbook. A process is deemed reliable if zero errors are found in the sample. Given the decision rule, a sample size is selected that reduces the probability of zero sample errors to less than 5 percent when the true error rate equals the tolerable level or critical threshold. Thus, the Type I risk of the method equals that of the method prescribed in this chapter.

For example, if zero errors were found in a sample size of 38 drawn from a population of 200, one could be at least 95 percent confident that the true error rate is not greater than or equal to 7 percent (p-value is 4.7 percent). Similarly, with a sample size of 27 taken from the same population, zero errors would occur only 4.7 percent of the time if the true error rate was 10 percent.

However, the Type I risk is significantly greater than risks associated with larger samples. For claims, the Type I risk for each population size exceeds 71 percent for processes that have an actual error rate of only 3 percent, which is well below the 7 percent critical threshold. Similarly, the Type I risk for non-claims processes is more than 83 percent for processes with a 6 percent error rate. Large Type I risks are inherent in small samples due to a large margin of error.

This method is not recommended for general use, since it does not finely discriminate between compliant and noncompliant processes, except when the true error rate is well below 3 percent.[24] As such, significant efficiency gains—which constitute the rationale of two-stage sampling—are unlikely to be realized, since a second-stage sample will be triggered in most instances. Regulators should use this method only in those instances in which they have reason to believe that the true error rate is low.

L. ACL and Sampling

One common auditing software package widely used by regulators is audit command software (ACL). This section discusses the sampling routines available in ACL.

ACL employs the same one-tailed confidence methodology for acceptance sampling that is described in this chapter, and it is, therefore, well-suited for examination and analysis purposes. Sample sizes in ACL are calculated by entering a confidence level (e.g., 90 percent or 95 percent), the population size, the upper error limit (or tolerance level) and an "expected error rate." The expected error rate is a way to establish a margin of error, and is *not* the population proportion assumed in calculating the sample size. Rather, the sample size is calculated assuming that the population proportion equals the tolerance level, or "upper error limit." ACL returns a sample size large enough to produce a maximum margin of error of (tolerance level – expected error rate).

[24] A point made in Arkin, Herbert. 1982. *Sampling Methods for the Auditor: An Advanced Treatment.* New Jersey: Prentice Hall.

© 2006-2007 National Association of Insurance Commissioners

For example, the ACL manual describes the following scenario:[25]

Population = 40,000
Confidence = 95%
Upper Error Limit = 5%
Expected Error Rate = 2%

ACL returns a sample size of 184, with a margin of error of no more than three percent (5% – 2%). The maximum allowable errors in the sample is four, such that if the sample contains four or fewer errors, the hypothesis that the true error rate is greater than five percent can be rejected with at least 95 percent confidence.

The following table displays the probability distribution for these results in a form similar to the AST. The cumulative probability column displays the percentage of samples (if taken over time) that would contain the corresponding number of errors *or fewer, if the population error rate equaled the tolerable error, or five percent.* For example, six or fewer errors would be obtained in 18.2 percent of samples. The 95 percent confidence limit is the point at which this probability falls below five percent (100 – 95), which occurs when the number of errors is less than or equal to four. Thus, with four errors in the sample, it can be concluded with 95.6 percent (100 – 4.4) confidence that the true error rate is less than five percent, and does not exceed the critical threshold.

Probability Distribution if Population Proportion is 5%				
Population	Sample	Errors	% Errors	Cumulative Probability
40,000	184	0	0.0%	0.0%
40,000	184	1	0.5%	0.1%
40,000	184	2	1.1%	0.5%
40,000	184	3	1.6%	1.6%
40,000	184	4	2.2%	4.4%
40,000	184	5	2.7%	9.8%
40,000	184	6	3.3%	18.2%
40,000	184	7	3.8%	29.4%
40,000	184	8	4.3%	42.5%
40,000	184	9	4.9%	56.0%
40,000	184	10	5.4%	68.5%

To generate sample sizes roughly equivalent to those contained in the AST, enter a confidence level of 95 percent, and the upper error limit for claims or non-claims processes (7 percent or 10 percent). Since the AST was initially constructed using a 4.5 percent margin of error for claims, the "expected error rate" for ACL is 2.5 percent (upper error limit—margin of error) = (7% – 4.5%) = 2.5%. ACL's sample size will not exactly duplicate those listed in the AST, since the sample sizes in the AST were adjusted for Type I risks.

Because the ACL routine is a one-tailed test appropriate for acceptance sampling, confidence limits must be modified for the stage-two sample, which generally employs a two-tailed test. That is, during the second-stage sampling, the concern is generally not whether the population proportion exceeds some

[25] *ACL for Windows Command Reference.* 1996. Vancouver: ACL Software, page 451.

© 2006-2007 National Association of Insurance Commissioners

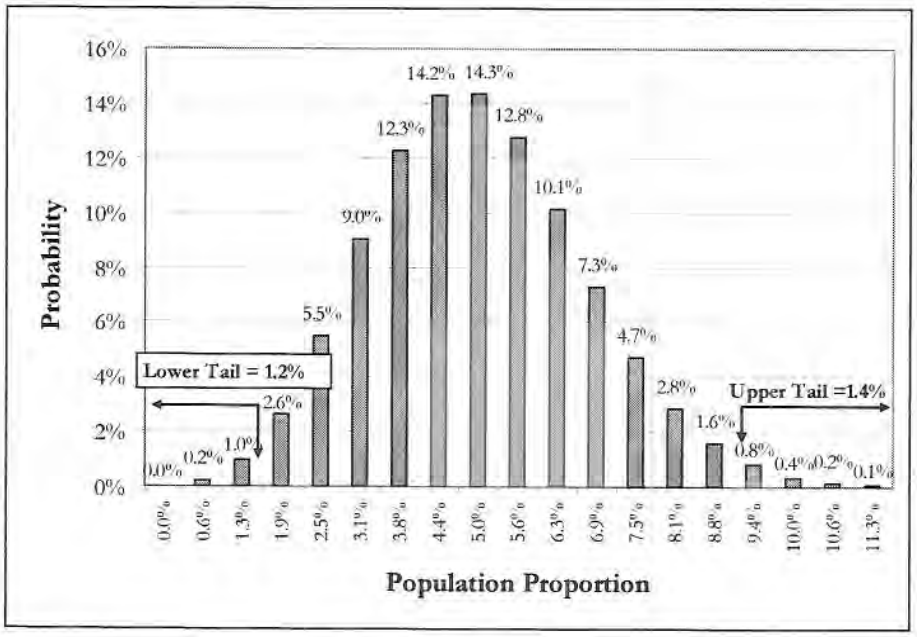
specified value. Instead, the second-stage sample is designed to establish *upper and lower* bounds within which the true value may lie.

For example, assume that eight errors are discovered in a sample size of 160. Substituting the sample error rate of five percent for the unknown population error rate, the following probability distribution is obtained. To obtain a confidence level of *at least* 95 percent that the true proportion lies within an interval around 5 percent, the sum of the probabilities of both upper and lower bounds cannot exceed 5 percent (100 – 95 percent confidence=5 percent). The minimum of a 95 percent confidence interval around 5 percent is, thus, 1.3 percent and 8.8 percent. The actual confidence limit is the sum of the probabilities of the upper and lower tails, or the areas that fall outside of the confidence interval: [100 – lower tail probability – upper tail probability] = [100 – 1.2 – (100 – 98.5)] = (100 – 1.2 – 1.5) = 97.3%.

Probability Distribution if Population Proportion is 5%				
Population	Sample	Errors	% Errors	Probability
40,000	160	0	0.0%	0.0%
40,000	160	1	0.6%	0.3%
40,000	160	2	1.3%	1.2%
40,000	160	3	1.9%	3.9%
40,000	160	4	2.5%	9.3%
40,000	160	5	3.1%	18.4%
40,000	160	6	3.8%	30.7%
40,000	160	7	4.4%	44.9%
40,000	160	8	5.0%	59.3%
40,000	160	9	5.6%	72.0%
40,000	160	10	6.3%	82.1%
40,000	160	11	6.9%	89.4%
40,000	160	12	7.5%	94.1%
40,000	160	13	8.1%	97.0%
40,000	160	14	8.8%	98.5%
40,000	160	15	9.4%	99.3%
40,000	160	16	10.0%	99.7%
40,000	160	17	10.6%	99.9%
40,000	160	18	11.3%	100%

The probabilities are displayed graphically below. In this chart, probabilities are non-cumulative, and represent the probability of a single proportion. The confidence interval is the area of the graph excluding the upper and lower tails, or between 1.9 percent and 8.8 percent inclusive. Alternatively, the normal approximation formula given in Section M yields a confidence interval of 1.6 percent and 8.4 percent, which is very close to the more exact hypergeometric limits. The normal approximation works well because, as reader will note, the shape of the distribution in the graph is approximates the normal or "bell-shaped" curve.

© 2006-2007 National Association of Insurance Commissioners



To estimate a sample size sufficient for a two-tailed test, the confidence level entered into ACL's one-tailed test must be increased. To calculate a sample for a 95 percent confidence interval for a two-tailed test, enter a confidence level of 97.5 percent, such that [2 * (100 – 97.5) = 5%].

The sample sizes produced by ACL are somewhat larger than those produced by the hypergeometric distribution, since the ACL algorithm utilizes a slightly different probability distribution.

M. Sampling Formulas

1. Formulas

A formula for approximating the required size of the second-stage sample to produce estimates with a given level of confidence and precision is:

$$S = \frac{Nz^2P(1-P)}{(N-1)e^2 + z^2P(1-P)}$$

where:

$N =$ Size of the population from which the sample will be taken
$S =$ Sample size
$Z =$ Standard normal deviate (or standard deviation). For a desired confidence level of 95 percent confidence, use z=1.96.
$P =$ The unknown population proportion to be estimated. Regulators should use their best judgment, as well as evidence from the initial sample, to select a value for P.
$E =$ Margin of error or degree of accuracy of the sample estimate expressed as a proportion (for example, use .05 instead of 5 percent)

The desired outcome is to produce a sample of sufficient size to support a conclusion with 95 percent confidence that the true proportion is within $\pm e$, or the margin of error, of the sample

© 2006-2007 National Association of Insurance Commissioners

proportion. The actual margin of error depends on the accuracy of the initial guess for the population proportion. As discussed above, the margin of error increases as the true proportion approaches 50 percent. If the initial guess was 10 percent, but the resulting sample proportion was 30 percent, the confidence limits will be significantly wider than initially anticipated. Confidence limits, therefore, must be calculated after the analysis of the sample is completed.

2. Confidence interval formula:

$$e = z\sqrt{\frac{P(1-P)}{S}}\sqrt{\frac{N-S}{N-1}}$$

where:

$e =$ Margin of error
$z =$ The standard deviation of the sampling estimate. Use z=1.96 for a 95 percent confidence interval.
$P =$ The sample proportion (as a substitute for the unknown population proportion)
$S =$ Sample size
$N =$ Population size

The formulas for sample size and confidence interval are for samples taken from small populations. The term $\sqrt{\frac{N-S}{N-1}}$ is the population correction factor. As populations increase in size, the term has less impact on the resulting estimate. For example, with a sample size of 300 and population size of 500, the term reduces to .63. If the population size is increased to 5,000, the term is .96, and it quickly approaches 1 at 50,000 (or .997). Thus, for populations greater than 50,000, the term can safely be dropped from the equation, since further increases in population have little impact on the margin of error, or the necessary sample size. In other words, a sample sufficient for sampling from a population of 50,000 will also be sufficient for sampling from a population of 50 million. In each case, the population is "large enough," and the formula for "large populations" can safely be used, which calculates a sample size without reference to the population size.

3. Skip interval formula:

$$I = (\frac{2N}{S}) - 1$$

where:

$I =$ Skip interval
$N =$ Population size
$S =$ Sample size

4. Procedure:

   a. Determine the sample size.

   b. Determine the skip interval.

© 2006-2007 National Association of Insurance Commissioners

c. Using a table of random numbers, or a random number generator, select the sample indicated.

d. Apply test(s) to sample and tabulate raw frequency expressing each frequency as proportion.

e. Using the 95 percent confidence interval formula, calculate an interval for each population proportion.

5. Calculation example:

From a population of 20,000 homeowners insurance policy files, investigate the accuracy of rating procedures. Using the sampling formula, find sample size as follows:

$$S = \frac{(20{,}000)(1.96^2)(.5)(1-.5)}{(20{,}000-1)(.05^2)+(1.96^2)(.5)(1-.5)}$$

S = 376.9, or 377

After the sample size is determined, the skip interval formula for a sample size of 400 yields:

$$I = \frac{2(20{,}000)}{400} - 1 = 99$$

From a random number table, select 400 numbers between 1 and 99 (ignore those numbers that fall outside the range of 1 to 99; e.g., disregard 138, 191, 295, 0, etc.) For sake of illustration, suppose the first 10 numbers are:

K1 – 03	K2 – 16
K3 – 12	K4 – 55
K5 - 56	K6 – 33
K7 – 57	K8 – 18
K9 – 25	K10 – 23

Begin the selection process by skipping the first three files and selecting the fourth file, then skip the next 16 files and select the 21st file and so on.

Suppose the following results were tabulated from the sample of 400 files.

	(a)	(b)	(c)
Category	Count	Percent (P) (Column a ÷ sample size)	95% Confidence Interval
Policies with rate overcharge	21	5.25	2.16
Polices with rate undercharge	14	3.50	1.78
Policies with non-premium error	10	2.50	1.51
Policies with insufficient information	10	2.50	1.51
Total rated incorrectly	55	13.75	3. 34

© 2006-2007 National Association of Insurance Commissioners

Using the 95 percent confidence interval formula, the confidence interval for 5.25 would be:

$$1.96\sqrt{\frac{(5.25)(94.75)}{400}}\sqrt{\frac{20{,}000-400}{20{,}000-1}}$$

1.96 x 1.115 x .99 = 2.16 (rounded)

Thus, the regulator can be 95 percent confident that the proportion of policies with a rate overcharge is between 3.09 (column b – 2.16) and 7.41 (column b + 2.16).

© 2006-2007 National Association of Insurance Commissioners