# Revisions to
# Section D. Standards of Chapter 14—Sampling
# of the *Market Regulation Handbook*

## D. Standards

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

1. **Pre-selection and Statistical Bias.** Pre-selection 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 pre-selection. 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 pre-selection 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 pre-selection 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 pre-selection 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 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.

   Pre-selection 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 pre-selection can be permitted. Generally, sample selection by the examinee should be avoided due to the difficulty in demonstrating that pre-selection has not occurred. Pre-selection 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, pre-selection 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 a 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 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.

Pre-selection can also occur due to the use of “pull lists” developed by the company’s computers/computer programmers. 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 pre-selection. Examples of this might be where no files appear in ZIP code XXXXX, or in the timeframe from 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 pre-selection 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 pre-selection of files occurs.

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 files 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. From a regulatory compliance standpoint, however, the tolerance level utilized can have an additional meaning beyond its use as an indicator of the size of sample needed to establish an error rate with a sufficient confidence level. Under the *Unfair Trade Practices Act* (#880) and *Unfair Claims Settlement Practices Act* (#900), one standard for establishing a violation of these laws is that a company commits a practice “with such frequency to indicate a general business practice.” Many states have included this general business practice standard (or a similar standard involving frequency) when enacting one or both of these models.

   Historically, a benchmark error rate of 7 percent has been established for auditing claim practices and 10 percent for other trade practices. Error rates exceeding these benchmarks are presumed to indicate a general business practice contrary to these laws. For uniformity in the application of these laws, and absent state case law that may apply an alternative standard, states that have the general business practice standard are strongly encouraged to utilize the 7 percent and 10 percent standards both as tolerance levels for statistical sampling purposes and as benchmarks for evaluating when violations of the state’s unfair claim and trade practices statutes have occurred. [1]

   On the other hand, many other state laws are not dependent upon the frequency of commission of an act in order to constitute a violation of the law—each instance of commission of the act constitutes a separate and distinct violation. For example, conducting business in a state without a license may constitute a violation of law each time it occurs, whether it is done once or one hundred times. This may also be true for the unfair claim and/or trade practices statutes in those states that have not adopted the general business practice standard of the NAIC models. The sampling error rate relative to such laws represents the probable number of violations within the

[1] With respect to sampling, readers are strongly cautioned not to confuse the two quite distinct meanings associated with the terms “tolerance level” and “benchmark error rate.” The former is a statistical construct with meaning only in terms of making probabilistic inferences, while the latter is a threshold used to establish the legal presumption of a general business practice. Important in this respect, the first stage sample cannot be used to establish with confidence that the true rate of noncompliance exceeds 7 percent or 10 percent. The small sample sizes only support the inference that one cannot confidently rule out such a possibility. The larger second stage sample is required to infer the actual rate of noncompliance, and determine whether this true rate exceeds some specified threshold.

total population rather than a benchmark for evaluating whether or not a violation has occurred. While it is not strictly necessary to use the 7 percent and 10 percent tolerance levels in these circumstances, states are still encouraged to do so when calculating appropriate sample sizes for consistency in both application and presentation. For this reason, all calculations in this chapter utilize the 7 percent and 10 percent tolerance levels.

In conclusion, and because it has been a source of some confusion, it is worth stressing that while the tolerable errors may reasonably be used as a benchmark to establish a "general business practice," the converse is not true. They do not establish a safe harbor such that violations below these levels are considered "tolerable" or legally permissible. They are merely tools embedded in the sampling process and are selected at levels to minimize the probability that a high volume of infractions will be missed by the smaller initial acceptance sample. The tolerable error levels discussed in this handbook simply are not relevant with respect to creating any maximum acceptable level of market misconduct, and should not be used as such.

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 a Chicago branch underwriting office, results cannot logically be extrapolated to a 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.

G:\MKTREG\DATA\D Working Groups\D WG 2014 MCES (PCW)\Docs_WG Calls\Sampling\Exposure Docs\Section D Standards from Chapter 14 Sampling 9-18-14.docx