Departures from Handbook Standards

The standards prescribed in this chapter adhere to generally accepted statistical principles. For example, it is difficult to find published works in peer-reviewed journals in any field in which statistical significance (or confidence) levels fall below 90 percent. As such, estimates of rates of noncompliance that fall below this minimal standard may be effectively challenged in judicial proceedings.

In the regulatory context, it is important to remember that a significance level represent meaningful probability. Accepting a confidence level of 90 percent constitutes a *de facto* tolerance that the regulatory findings will simply be wrong 10 percent of the time. In addition, estimates of noncompliance that are too imprecise (i.e. in which the margin of error is too wide) may simply frustrate the ability to draw any meaningful conclusions in some contexts, particularly when levels of non-compliance are low. For example, an estimate of a noncompliance rate of 7 percent with a margin of error of ± 6 percentage points may be so imprecise as to be of little or no use in many contexts. Such an estimate will not permit a regulator to reasonably discern very low levels of noncompliance (7 – 6 = 1 percent) from noncompliance rates that are much higher (7 + = 13 percent).

Smaller sample sizes than those recommended in this chapter will result in less robust statistical results – estimates will be less precise, and the probability that findings are simply wrong will increase. However, there are instances in which sample sizes necessary to achieve a high degree of precision and confidence are prohibitively costly in terms of both regulatory resources and to the examinees. In some cases, regulators and examinees may find it mutually beneficial to relax sampling standards to some degree, and accept more general findings that can be supported by smaller sample sizes (and perhaps also supported by other non-statistical information). For example, findings that can be expressed as general statements of the type that "an area of noncompliance exists and requires remediation" may be entirely reasonable even if the statistical evidence does not support a precise estimate of the *rate* of noncompliance.

Even if a finding is recognized by mutual consent, regulators should still proceed in full knowledge of the manner in which small sample sizes will impact the credibility of results. While the possibility of legal challenge may have been avoided by stipulation, there are still practical consequences relevant to regulatory oversight more generally. It is recommended that regulators recognize the loss of statistical precision associated with smaller samples, and balance such losses against efficiency gains in relation to regulatory objectives. The following tables are intended to serve as a reference for the use of small samples.

The table below displays the sample sizes necessary to support inferences at varying levels of confidence and precision. The example uses a population sized 5,000, and estimated noncompliance or "error" rates of 10 percent and 20 percent.[1] Required sample sizes can be reduced by accepting a higher margin of error and / or lower confidence levels.

Reading across the first row, it is apparent that producing a precise estimate of ± 2 percentage points for a rate of noncompliance of 10 percent requires a rather significant sample size of 738. That is, if one were to sample 738 files, and find that 10 percent of the sample files were noncompliant, one can be 95 percent

[1] Readers will recall that samples sizes vary with the rate of noncompliance being estimated. Error rates that approach 0% or 100% require the smallest samples, with maximum sample sizes necessary at an error rate of 50%.

confident that the true noncompliance rate is between 8 and 12 percent. In five percent of cases, the true noncompliance rate will fall outside of the lower and upper bounds.

However, such a precise estimate may not be required for many purposes. The required sample size can be significantly reduced by accepting greater imprecision. At a margin of error of 4 percentage points, the sample size drops to 208. The required sample can be reduced still further by relaxing the confidence levels. At a margin of error of 4 percentage points, the sample size to support conclusions with 90 percent confidence is reduced to 148.

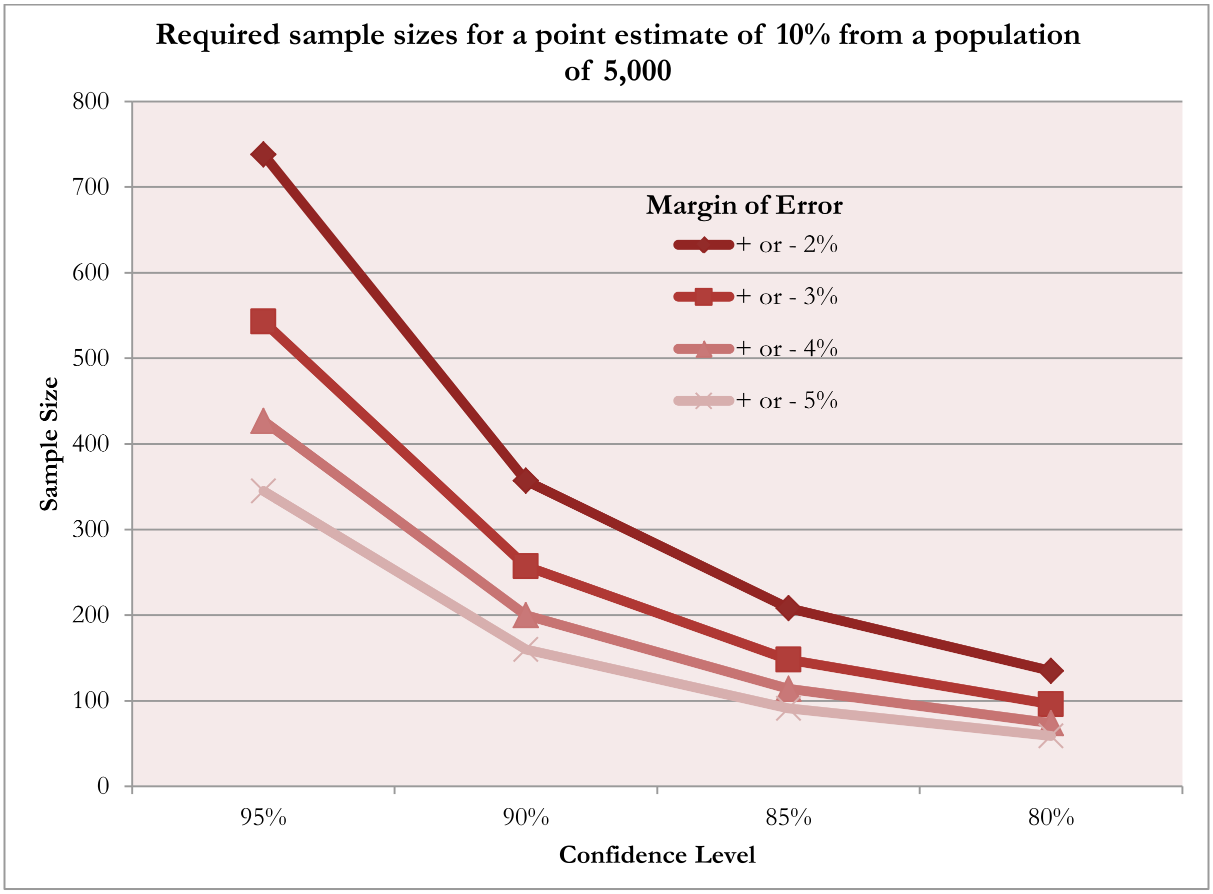
The procedure here results in significant savings of labor and still adheres to generally accepted statistical methods (retaining at least 90 percent confidence). If the sample sizes still prove prohibitively expensive for a given examination, the parameters can be further relaxed. Indeed, a very small sample of 59 will support some limited conclusions, albeit at a confidence of only 80 percent and a rather wide margin of error of five percentage points. In accepting such a low confidence, regulators should weight the objectives of the examination and the areas of noncompliance with the known risk that in twenty percent of cases the true rate of non-compliance with either fall below 5 percent or exceed 15 percent.

The responsiveness of sample size to altering parameters is displayed graphically in the figure following the table.

Required Sample Sizes at Different Parameter Levels

Population Size	Estimated Error Rate	Margin of Error (Confidence Interval)	Confidence Level	Required sample size
5,000	10%	± 2	95%	738
			90%	543
			85%	427
			80%	345
		± 3	95%	357
			90%	257
			85%	200
			80%	160
		± 4	95%	208
			90%	148
			85%	114
			80%	91
		± 5	95%	135
			90%	96
			85%	74
			80%	59
		± 2	95%	1,176
			90%	890
			85%	712
			80%	582
			95%	601

Required Sample Sizes at Different Parameter Levels				
Population Size	**Estimated Error Rate**	**Margin of Error (Confidence Interval)**	**Confidence Level**	**Required sample size**
	20%	± 3	90%	439
			85%	344
			80%	276
		± 4	95%	357
			90%	257
			85%	200
			80%	160
		± 5	95%	235
			90%	168
			85%	130
			80%	103



A second alternative exists that will permit reductions in sample sizes beyond relaxing standards governing confidence levels and precision. In some instances, an examination may not be concerned with whether the true noncompliance rate lies within a given numeric interval with a known probability. As with the initial acceptance sample, a regulator may only be concerned with whether the true error rate exceeds some threshold.[2] Since this method entails calculating the likelihood of events at only one end of a probability distribution, statisticians often refer to it as a "one-tailed" test. Essentially, the one-tailed test doubles the statistical confidence level, since one is not concerned with the probability of outcomes when the true noncompliance rate is below the defined threshold.

For example, assume that a sample sized 59 is drawn from a population of 5,000, as per the example in the preceding table. If a noncompliance rate of 10 percent is found in the sample, the conclusion that the true noncompliance rate is between 5 and 15 percent can be supported with 80 percent confidence. In 10 percent of samples, a result of 10 percent noncompliant files will appear *even if the true noncompliance rate in the population is below 5 percent.* Conversely, the same result of 10 percent noncompliant files will occur in 10 percent of samples when the true noncompliance rate exceeds 15 percent. The 10 percent probabilities at either "tail" together equal the 20 percent probability that the true noncompliance rate fall outside of the margin of error of five percentage point.

However, a one-tailed test is concerned only with the upper end of the probability – the 10 percent probability that the true noncompliance rate exceeds the upper bound of the confidence interval or 15 percent. Assume that in the same sample of 59 files, it is determined that 16 percent of the files are noncompliant. Halving the 20 percent uncertainty level of the two-tailed confidence level of 80 percent from the preceding table, one can be 90 percent confident, based on the sample results, that the true rate of noncompliance exceeds 10 percent. Thus, even very small samples can yield results at a level of confidence that conforms to statistical standards.

[2] Note that while the method is the same as acceptance sampling, the question is the opposite. The acceptance sample is designed to answer the question "can I be confidence that the true noncompliance rate falls *below* some threshold." Here, the more relevant question is whether the noncompliance rate *exceeds* a defined threshold.

Acceptance Sampling Using Different Tolerable Error Rates

There may be occasions in which an examiner might wish to use tolerable error thresholds other than the 7 and 10 percent values described in the preceding section. Generally, lower thresholds require *smaller* sample sizes to produce inferences at comparable levels of *statistical power*. Recall that there are two types of errors that the initial sampling methodology is designed to minimize. If the test is assessing whether the true error rate is below 7 percent, two possible incorrect inferences are:

1. False negative: an incorrect inference that the true noncompliance rate is less than 7 percent when it is in fact greater than 7 percent, and

2. False positive: an incorrect inference that the true noncompliance rate may exceed 7 percent when in fact it does not.[3]

Of the two types of errors, a *false negative* is considered the more serious, and the probability of it occurring is always set to less than or equal to 5 percent. However, a *false positive* can result in the unnecessary expenditure of additional resources required to further investigate the rate of noncompliance when the rate is below the acceptable threshold. For that reason, the risk of a *false positive* is also of some concern. In fact, failure to control for this risk defeats the purpose of the two-stage sampling approach.

The following table displays, for each level of tolerable error (4-10 percent) the sample sizes necessary to produce a false-negative risk of less than five percent. For example, if the tolerable error is fixed at 6 percent, and the decision rule is "no more than 2 deviations in the sample," then a sample size of 103 is necessary to ensure that the risk of a false negative is no more than 5 percent. That is, the probability that a population with a rate of noncompliance in excess of 6 percent will produce no more than 2 noncompliant files in a sample of 103 files is less than 5 percent.

Required sample sizes necessary to produce a false-negative risk of less than 5%						
	Maximum number of errors in sample					
Threshold (tolerable error)	**0**	**1**	**2**	**3**	**4**	**5**
4%	74	117	156	192	226	260
5%	**59**	**93**	**124**	153	181	208
6%	49	**77**	103	127	150	173
7%	**42**	**66**	**88**	109	128	148
8%	36	58	77	**95**	112	129
9%	32	51	68	84	**99**	114
10%	**29**	**46**	**61**	**75**	**89**	**103**

[3] The language used in point #2 is somewhat cumbersome, but recall from the earlier discussion that the acceptance sample method *does not* permit inferences that the true noncompliant rate *exceeds* the threshold or tolerable error. The method permits one to either conclude that the possibility that the true noncompliance rate exceeds the threshold 1. can be ruled out or 2. cannot be ruled out (with 95 percent certainty).

Given the sample sizes displayed in the preceding table, the following table displays the risk of a *false positive* when the true rate of noncompliance if three percentage points below the tolerable error. For example, as discussed above, a tolerable error of 6 percent and maximum of two 2 permissible deviations in the sample requires a sample sized 103 to ensure the risk of a *false negative* is below 5 percent. This same sample size and decision rule produces a 60 percent *false positive* risk when the true rate of noncompliance is (6 – 3) = 3%.

The highlighted cells in both tables represent the selected sample size / decision rule combination for each tolerable error. Note that at the selected sample sizes, the risk of the two types of inference errors is relatively constant across different tolerable errors. Note too that reducing the tolerable error generally requires *smaller* samples. For example, reducing the tolerable error from 7 to 5 percent reduces the required sample size from 88 to 59. Thus, adopting lower thresholds entails no additional use of regulatory resources, *at least with respect to initial acceptance sample.*[4]

False Positive Risk When the True Noncompliance Rate is Three Percentage Points Below the Tolerable Error						
	Maximum number of errors in sample					
Threshold (Tolerable Error)	**0**	**1**	**2**	**3**	**4**	**5**
4%	52.5%	32.7%	20.6%	12.8%	7.8%	4.8%
5%	69.6%	55.7%	45.2%	36.6%	29.7%	23.9%
6%	77.5%	67.6%	60.0%	53.1%	46.9%	41.8%
7%	**82.0%**	**74.7%**	68.9%	63.9%	58.4%	54.5%
8%	84.2%	79.3%	74.7%	70.5%	66.4%	62.9%
9%	86.2%	81.9%	78.2%	74.9%	71.5%	68.6%
10%	87.8%	84.2%	80.9%	77.9%	75.5%	73.5%

[4] Of course, on the whole, policing lower levels of noncompliance may on whole require greater time and effort, depending on how often noncompliance rates of between 5 and 7 percent occur.