

## STATEMENT ON ASTM E2548-16

**To:** OSAC Program Office

**From:** David Kaye

**Subject:** Statistical Terminology in ASTM E2578-16

**Date:** November 6, 2020

### Introduction

In 2016, six OSAC Legal Resource Committee members submitted public comments opposing the addition of ASTM E2548-11e1 (Standard Guide for Sampling Seized Drugs for Qualitative and Quantitative Analysis) to the OSAC Registry. Despite these comments and criticism from other OSAC units, OSAC added the standard to the Registry (effective April 3, 2017).

ASTM International made minor changes to the 2011 Standard Guide in ASTM E2548-16. This superseding version has been proposed for addition to the Registry. It is difficult to say whether the 2016 version should be added to the Registry. On the one hand, it contains several improvements, and it can be argued that an improved version—no matter how slight the improvement—should replace the 2011 version on the Registry. On the other hand, a negative outcome would underscore the need to finally arrive at a standard that adequately addresses the concerns that lawyers, statisticians, and psychologists have voiced.<sup>1</sup>

Without trying to resolve the Registry-approval question, these comments reiterate several of the technical objections to the standard and offer further observations so that they will remain clearly in sight. The major, interrelated, problems in ASTM E2548-16 are the failure: (1) to state clearly that probability sampling is *required* to permit valid inferences about population features (the technical term is “parameters”); (2) to define basic terms; (3) to use standard terms (in the statistics literature) for concepts; (4) to give some guidance on what values for the confidence coefficient or related criteria are scientifically acceptable in picking sample sizes using standard statistical methods; and (5) to offer guidance on how to express sampling uncertainty in either the frequentist and Bayesian frameworks to which the Standard Guide alludes.

Because the Standard Guide’s discussion of statistical reasoning is vague, laboratories will have to consult other sources to know what to do, making the Guide useful primarily as a bibliography. OSAC and ASTM should work toward having an easily understood document for drug chemists who want to develop sampling plans and present the results of tests on a sample of items correctly. Therefore, the best strategy might be to bring in a statistical expert on sampling theory and methods to reorganize and rewrite this Standard Guide rather than to continue making patchwork changes and revisions. Nevertheless, the remainder of this statement contains notes on some of the specific wording in ASTM E2548-16 that is problematic. Some of the remarks are editorial and minor. Others are editorial and major. Some are conceptual and substantive.

---

<sup>1</sup> A paragraph at the end of the Guide states that it “must be reviewed every five years and if not revised, either reapproved or withdrawn.” The five-year review should occur next year, making it unclear why it is urgent to act on a 2016 version whose withdrawal or replacement is imminent.

## Detailed Remarks

1.1 This guide covers minimum considerations for sampling of seized drugs for qualitative and quantitative analysis.

← This is not a precise statement of the scope of the Standard Guide. What is a “minimum consideration”? Depending on how one construes the phrase “minimum consideration for sampling,” the statement of scope is not complete. The Guide also has sections on the documentation (§ 7) and reporting (§ 8) of the sampling part of the laboratory’s work. These exceed the stated scope. At the same time, the Guide lacks a parallel section on testimony on the nature and outcomes of the laboratory’s sampling. The scope section should outline the full scope and the important topics that are left out.

3.2 The principal purpose of sampling in the context of this guide is to answer relevant questions about a population by examination of a portion of the population. For example: What is the net weight of the population? What portion of the units of a population can be said to contain a given drug at a given level of confidence?

← The first sentence is not quite correct. The purpose of *sampling* is merely to produce a representative sample for analysis. The combination of sampling, analysis (measurement), and interpretation gives an answer to (or input that helps to answer) a legally relevant question without having to analyze every unit in the population.

“Population” should be defined in a new section that defines technical terms. The definition can be taken from one of the references or a statistics text. Same for “unit.”

The question “What is the net weight of the population?” should be “What is the net weight of a proscribed substance in the population?”.

“Level of confidence” should be defined. Does it refer only to the value of a confidence coefficient in a frequentist interval estimate? To a posterior probability in a Bayesian credible region? Both?

3.3 By developing a sampling strategy and implementing appropriate sampling schemes, as illustrated in Fig. 1, a laboratory will minimize the total number of required analytical determinations, while ensuring that all relevant legal and scientific requirements are met.

← “Sampling strategy” and “sampling scheme” should be defined. Neither phrase is common in statistics texts. There may be better terms.

This subsection presumes that there is no conflict between scientific and legal requirements. In any case, if “appropriate” means a procedure that reduces the number of units sampled to the minimum that is reasonably likely to estimate a parameter with scientifically and legally satisfactory precision, then the sentence should be rewritten to express that objective.

4.1 An appropriate sampling strategy is highly dependent on the purpose of the investigation, a customer’s request, and the anticipated use of the results. Laws and legal practices form the foundation of most strategies and shall be taken into account when designing a sampling scheme. Therefore, specific sampling strategies are not defined in this guide.

← The paragraph concludes with a non sequitur. Specific sample designs can and should be defined and discussed in a Standard Guide. Various legal constraints and customer requests can be enumerated, and guidelines can be provided for the enumerated situations. Doing so would supply more useful guidance than self-evident remarks such as “legal practices ... shall be taken into account.”

4.2.1 Sampling may be statistical or non-statistical. NOTE 1—For the purpose of this guide, the use of the term statistical is meant to include the notion of an approach that is probability-based.

← Normally, the word “statistical” means “relating to the use of statistics,” and a statistic is a number derived from or characterizing data (such as a batting average). In general (and as in the example of a batting average), the data need not be acquired by a method that involves the application of the mathematical theory of probability at some stage (if that is what “an approach that is probability based” means). Of course, one can choose to redefine a term to depart from its normal usage in science and ordinary life, but there is no reason to do so when a perfectly good term (in this case, “probability sampling”) is available in the statistical literature.

4.2.1.1 In many cases, a non-statistical approach may suffice. The sampling plan shall provide an adequate basis for answering questions of applicable law. For example, “Is there a drug present in the population?” “Are statutory enhancement levels satisfied by the analysis of a specified number of units?”

← The term “sampling plan” should be defined. How does it differ from a “sampling scheme,” “sampling strategy,” and “sampling procedure”? Figure 1 has all these terms in it, but most of them are not defined explicitly and clearly. (“Sampling scheme” is defined in § 5.1.)

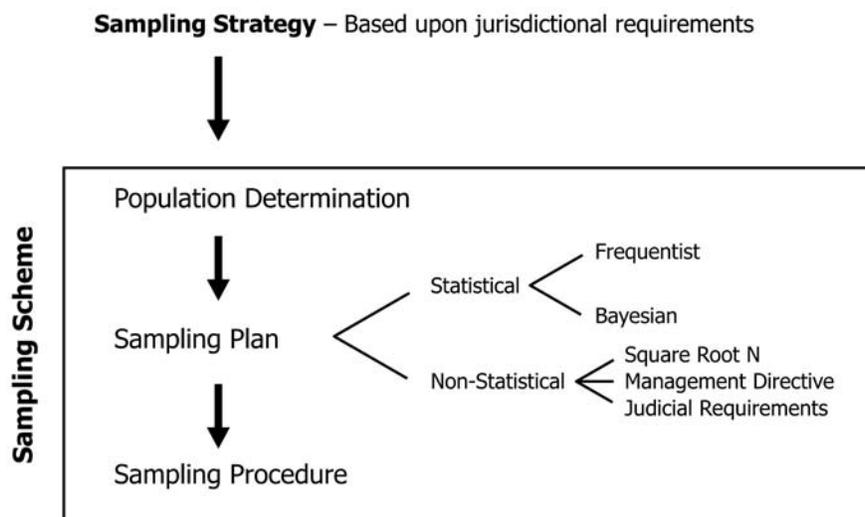
4.2.1.2 If an inference about the whole population is to be drawn from a sample, then the plan shall be either statistically based or have an appropriate statistical analysis completed and limits of the inference shall be documented.

← The phrase “or have an appropriate statistical analysis completed” has been added since 2011. But what “appropriate statistical analysis” will permit a statistical inference from a sample to the population in the absence of a probability sample? In some research, non-probability samples are treated as if they were probability samples, but it is not necessary to make that analogy if probability sampling is feasible, as it seems to be here. (If it is not always feasible, those situations should be identified.)

4.2.2 Selected units shall be analyzed to meet Practice E2329 if statistical inferences are to be made about the whole population.

← ASTM E2329 describes minimum requirements for qualitative analysis. It is not obvious why these requirements for identifying compounds do not apply to the analysis of every unit in a census of the population. Moreover, this provision seems beyond the scope of a standard on sampling. It concerns measurements or observations on the items in a sample.

**FIG. 1. Relationship of Various Levels Required in Sampling**



← Previously, the leaves in the top “statistical” branch were "Hypergeometric," "Bayesian," and "Other probability-based approaches." These have been trimmed to “Bayesian” and “Frequentist.” This is an improvement, since “Hypergeometric” is merely the name of a probability distribution, and not a form of sampling. (When sampling without replacement, the probability that a number  $X$  of units with a discrete feature will be picked in  $N$  draws follows the hypergeometric distribution. This is true for Bayesians and frequentists alike.)

The problem that remains in this branch is that neither “Bayesian” nor “frequentist” describes a sampling plan. For example, the results for a simple random sample (SRS) can be analyzed from either perspective, and the sample size can be pre-established according to computations with either methodology (or via likelihood computations, for that matter). Presumably, a sampling procedure that implements the SRS design would be a “sampling plan.” But the SRS design itself is neither Bayesian nor frequentist. In addition, there is a third school of statistical thought, likelihood theory, that could be used in analyzing sample data.

The leaves of the “Non-Statistical” branch are “Square Root N,” “Management Directive,” and “Judicial Requirements.” The first leaf refers to a procedure for establishing the size of a sample. The square root of  $N$  assuredly is a statistic, and unless the population size exceeds 9, the method is to draw an SRS (see United Nations Office on Drugs and Crime, Recommended Methods for Testing Opium, Morphine and Heroin: Manual for Use by National Drug Testing Laboratories, p.21). Consequently, the “square root method” can be used to obtain a sample statistic that is a statistically sound estimator of a population parameter.

The next two leaves can apply to any choice of a sampling design that comes from a manager or a court. The design could be statistically optimal or wasteful, but its statistical properties do not make it “non-statistical.” Perhaps the branch is supposed to be “not suitable for statistical inference,” but that does not lead to the three leaves either.

In sum, Figure 1 muddies the waters. It should be abandoned or redrawn after more appropriate and better defined terminology is in place.

5.2.1 The population determination shall take into account all typical forms and quantities in which exhibits may appear.

What is an “exhibit”? The term appears nowhere else in the Guide. Why should atypical forms and quantities (of what?) not be considered along with typical ones?

5.2.2 A population can consist of a single unit or multiple units.

5.2.3 A multiple unit population shall consist of items that are similar in relevant visual characteristics (for example, size, color, shape, etc.).

← The wording suggests that material packaged in containers of different sizes or color is not a population of legal interest. In this form, the material is neither a “single unit” nor a “multiple unit” population. The statistical question may concern a feature of all the material, and that collection would be the population. It appears that in this standard, a “population” is a sampling frame that is constructed from the population. The literature on attribute and variable sampling as well as multistage sampling should be consulted. In sampling theory, a “sampling unit” is not necessarily a single element of the population. There are smaller units of elements at each stage.

5.4 When a single unit or bulk population is to be analyzed, the issue of homogeneity shall be addressed within the sampling plan.

← Now it appears that there is a third kind of population—a “bulk population” that is neither a single unit nor a combination of units. The definitions of all these populations should be in a definitions section.

5.4.1 One sample is sufficient if the bulk material is homogeneous. Analysts can make bulk material homogeneous.

← Normally, a sample is the subset of elements drawn from the sampling frame that represents the population of interest. Of course, “sample” has another meaning—a single chunk of material for instrumental or visual analysis—but that is not necessarily a “sample” as the term is used in sampling theory. The sample from the population can consist of many items. Substituting “specimen” for “sample” in this section would avoid possible confusion.

5.5 For a multiple unit population, the sampling plan may be statistical or non-statistical.

← See supra comment on Figure 1.

5.5.1 Statistical approaches are applicable when inferences are made about the whole population. For example:

The probability that a given percentage of the population contains the drug of interest or is positive for a given characteristic.

The total net weight of the population is to be extrapolated from the weight of a sample.

← Normally one would speak of the population percentage as a parameter whose value is being estimated from sample data. The estimate is an inference. A hypothesis about the true value would be another inferential statement. The “probability that a given percentage of the population contains the drug of interest or is positive for a given characteristic,” if based on the sample data and a prior distribution, would be called a posterior probability for the inference. The posterior probability cannot

be computed with frequentist methods for inference, and this fact should be made clear in the Standard Guide (and included in reports that contain coverage probabilities).

The second example is not the inference; it is the objective of the study. An example of an inference would be “Extrapolating from the sample data, the total net weight of a controlled substance in the population is 500 grams.”

#### 5.5.1.1 Published examples of statistical approaches involving general considerations:

(1) Practice E105.

(2) Practice E141.

(3) Terminology E1732.

← A glossary or dictionary of terms is not an example of an approach.

(4) Guidelines on Representative Drug Sampling.<sup>4</sup>

<sup>4</sup> 2009 UNODC/ENFSI Guidelines on Representative Drug Sampling, European Network of Forensic Science Institutes (ENFSI), 2004, <http://www.ENFSI.org>.

← A 2009 document cannot be published in 2004.

(5) ISO 3534-1.

(6) ISO 3534-2.

← A glossary or dictionary of terms is not an example of an approach, but maybe some of the definitions contain such examples.

#### 5.5.2 Non-statistical approaches are appropriate if no inference is to be made about the whole population.

##### 5.5.2.1 Published examples:

(1) The “square root method.”<sup>7</sup>

(2) Methods listed in “Arbitrary Sampling” in *Guidelines on Representative Drug Sampling*.<sup>4</sup>

← The “square root method” of (1) is a form of “arbitrary sampling” according to the *Guidelines* cited in (2). Therefore, having the two categories of “non-statistical approaches” makes little sense (even if one can construe using an SRS with the various “square root methods” listed in the EU document for choosing the sample size as “non-statistical”).

5.5.2.3 A non-statistical sampling approach may allow an inference about the population. If a single population has been randomly sampled, the data may allow an inference to be drawn by (1) determining and reporting a confidence interval for an inferred population parameter (for example: weight or tablet count); or by (2) retrospectively using the results in a statistical model and determining the resulting probabilities and level of confidence.

← This paragraph is a real puzzler. The essence of probability sampling is that the probability that each and every element in the sampling frame is known. Knowing these probabilities allows one to use a statistical model to determine the sampling distribution of the estimator. That, in turn, leads to a standard error, a point estimate, and thus a confidence interval. If “non-statistical sampling” does not use probability sampling, how is one supposed to determine the standard error? The answer might be to analogize the approach actually used to a probability sampling method of some sort. But that interpretation of the paragraph is contradicted by “(2) retrospectively using the results in a statistical

model and determining the resulting probabilities and level of confidence.” (1) is already retrospective and uses a statistical model to produce the confidence interval, so what is (2) supposed to add? Furthermore, (2) is obscure at best. A “level of confidence” is a coverage probability. As such, what are “the resulting probabilities”?

#### 5.6.1 Establish the procedure for selecting the number of units that will comprise a sample.

← This sounds prescriptive, but the relative merits of different procedures have, strangely, not been discussed in this Standard Guide. Accordingly, the document supplies no guidance beyond a general admonition to read the literature mentioned in it and to do the right thing. What would that be? The UN report, for example, contends that even though the square-root sample-size determination lacks a sound theoretical foundation, it gives reliable results. The EU report, which describes this rule-of-thumb as “arbitrary,” acknowledges that it does “work well in many situations,” and explicates a number of variations on it. But statisticians have been known to dismiss it out of hand. E.g., Alan Julian Izenman, *Statistical and Legal Aspects of the Forensic Study of Illicit Drugs*, 16 *Stat. Sci.* 35–57, 47 (2001) (“The popularity of the square root rule, despite the lack of theoretical support for this rule, shows how an unfounded rule-of-thumb can be established in the practice of a particular field.”). What guidance does the Standard Guide have to offer?

##### 5.6.1.1 For non-statistical approaches, select a sample appropriate for the analytical objectives.

← There is no indication in the Standard Guide of how to go about doing this and no clear statement of what “analytical objectives” are. Usually, “analysis” in the Standard Guide refers to laboratory analysis (particularly § 6), but in this subsection the phrase “analytical objectives” seems to be a veiled reference to doing what management or the law dictates. Or, if that is not what the words point to, does not the same advice apply to “statistical approaches”? Surely, in all situations, one should select “a sample appropriate for the analytical objectives.” The reader is left uninformed by the Guide.

##### 5.6.1.2 For statistical approaches, random sampling shall be conducted.

← Given that probability sampling always is possible in this context (at least approximately), why would one ever want anything less? For example, if the sample size  $n$  is determined by the formula  $n = N/10$ , where  $N$  is the population size (based on an opinion in which some judge at some time said that he or she thought that samples of less than 10% could not be trusted), are we in the 5.6.1.1 realm of “non-statistical approaches”? The selection method should use a randomizing mechanism regardless of a “non-statistical” approach to determining the sample size.

##### 5.6.2.1 A random sample is one selected without bias and where each item has an equal chance of being selected.

← The words “without bias” seem like they represent a separate requirement, but they are merely superfluous. The reason to give an item in a sampling unit (which could be the sampling frame) an equal probability of being selected is to avoid bias (and give estimates of sampling error a solid foundation). The sentence could be broken into the three different thoughts: “A simple random sample is one in which each item in the sampling frame has an equal probability of being selected. Using a randomizing procedure to equalize the probability avoids selection bias and supplies the foundation for estimating sampling error.”

Why is simple random sampling the only acceptable form of probability sampling?

5.6.2.1 (cont'd) Computer generated random numbers or random number tables are commonly employed for such tasks and these should be included in the sampling plan.

5.6.2.2 Random sampling of items using random number tables may not be practical in all cases. In these instances, an alternate sampling plan shall be designed and documented to approach random selection. A practical solution involves a “black box” method, which refers to one that will prevent the sampler from consciously selecting a specific item from the population (that is, all units are placed in a box and the samples for testing are selected without bias). ...

← 5.6.2.1 requires a sampling plan that provides for selection according to random numbers even when their use is impractical. Then 5.6.2.2 says not to use that part of the plan in some cases. The final clause of 5.6.2.1 should be “and these should be part of the sampling plan unless a statement showing that they are impractical is included in the written plan.”

Why are “random number tables” but not pseudo-random number generators the subject of 5.6.2.2?

5.7 Sample Reduction—Sample reduction may be applied in cases where the weight or volume of the selected units is too large for laboratory analysis (Fig. 2, insert A).

← “Sample reduction” should be defined. Figure 2 does it to some extent (by listing the names of three methods for reducing the quantity), but a section with a general definition would improve readability. Again, in the figure, “sample” is used in the laboratory-analysis sense rather than the sampling-from-a-population sense. A word or phrase like “specimen” or “analytical unit” would be better for the former kind of sample.

## 6. Analysis

6.1 Statistically Selected Sample(s)—In accordance with 4.2.2, it is recommended that each unit comprising the sample be analyzed to meet Practice E2329 if statistical inferences are to be made about the whole population.

6.2 Non-statistically Selected Sample(s)—Practice E2329 shall be applied to at least one unit of the sample.

← See comment on § 4.2.2.

## 7. Documentation

7.1 Inferences drawn from the application of the sampling plan and subsequent analyses shall be documented.

← This section is exceedingly skimpy. It does not even cover the documentation mentioned in § 4.2.1.2 and § 5.6.2.2

## 8. Reporting

8.1 Sampling information shall be included in reports.

What “sampling information” should be reported? The subsections that follow (8.1.1 and 8.1.2) refer to nothing more than a statement of the results. They do not discuss reporting the parts of the sample design and selection mechanisms that led to the testing.

8.1.1 Statistically Selected Sample(s)—Reporting statistical inferences for a population is acceptable when testing is performed on the statistically selected units as stated in 6.1 above. The language in the report must make it clear to the reader that the results are based on a probability-based sampling plan.

8.1.2 Non-Statistically Selected Sample(s)—The language in the report must make it clear to the reader that the results apply to only the tested units. For example, 2 of 100 bags were analyzed and found to contain Cocaine.

← The Guide is confusing in its description of what qualifies as “a probability-based sampling plan.” For example, it appears from Figure 1 and parts of the text noted above that if the sample size is not chosen by using probability theory at that stage, the result is a “non-statistically selected sample.” The report therefore could not make any extrapolation to the population from an SRS with an ad hoc sample size. But using an unnecessarily large sample—which is what using a statistically optimal sample size guards against—does not undermine the validity of a statistical inference from the SRS. The laboratory should be able to make inferences from probability samples regardless of the method for choosing the sample size.

When the sample is not a probability sample—one drawn so as to give every element in the sampling frame a known probability of selection—an ampliative inference is of questionable validity, and the demand that the report “make it clear to the reader that the results apply to only the tested units” is reasonable. But merely stating that  $x$  out of  $N$  bags “were analyzed and found to contain Cocaine” does not make it clear that the sample proportion  $x/N$  might deviate greatly from the population proportion and that, without simple random sampling, statistical theory cannot be used to estimate the sampling error. Moreover, although the “ $x$  of  $N$ ” format for reporting is just as appropriate for estimates from probability samples, § 8.1.1 does not mention it. It only addresses the need to report that “a probability-based sampling plan” was used. It overlooks the issue of how to report the results themselves.

## 9. Keywords

9.1 analytical method; qualitative method; quantitative method; sampling; seized drug analytical method

← The only keyword that has to do with sampling is “sampling.” For a Standard Guide on sampling, that is shocking.