datapoints
a statistics q&a

Sample Sizes

Considering the Number of Items to Include in a Sample

BY PETER FORTINI

Q: How many items from a lot should I sample to determine the average value for a property of the lot?

A: The number of items to sample is a compromise between the precision with which you must obtain the average and the cost of sampling and testing all the items in the sample.

ASTM E122, Practice for Calculating Sample Size to Estimate, With Specified Precision, the Average for a Characteristic of a Lot or Process, provides a sample size formula based on the standard deviation of values for items in the lot and the required precision of the average value:

$$n=\left(\frac{3\sigma}{\delta}\right)^2$$

in which σ is the standard deviation of the property in the lot and δ is the required precision of the average value. The objective is that the error in the average due to sampling should almost certainly be less than δ.

The principle behind the formula is the distribution of the average of *n* items from a random sample of a population (in this case, the property values of items in the lot). If the population has a mean μ and standard deviation σ, then the law of large numbers and its extension, the central limit theorem, tell us that when a random sample of *n* items is drawn from the population,

- The expected value of the average $\bar{x}$ is equal to the population mean μ,
- The standard deviation of the average $\bar{x}$ is equal to $\sigma/\sqrt{n}$, and
- The form of the distribution of the average is close to Gaussian.

The Gaussian, or normal, distribution is the well-known "bell-shaped curve" of probability and statistics. For a random quantity with a Gaussian distribution, values will be within one standard deviation approximately two-thirds of the time, within two standard deviations approximately 95 percent of the time, and within three standard deviations approximately 99.7 percent of the time. The last degree of assurance is the one aimed for by the factor 3 in the sample size formula.

Notably absent from influence on the required sample size for given accuracy is the size of the lot or population. The precision of the average does not depend on population size unless the lot is so small that the number to be sampled is a significant fraction of all the items.

Sampling strategies often do take larger samples of larger populations. For example, a traditional prescription is given by the square-root-of-*N*-plus-one rule. Taking larger samples for larger lots does not contradict the sample size formula. The larger the lot, the more important it might be to achieve a precise determination

of the mean value. In other words, the value δ in the sample size formula becomes smaller.

Also absent from significant influence on the required sample size is the form of the distribution of the property in the lot. The property does not have to have a normal distribution in order for the normal distribution to apply to the average. The property values may be skewed to one side or rectangular (box-shaped). The distribution of the average depends strongly only on the lot mean and standard deviation.

That the sample be a random representative sample is critical. The notion that the sample average has a statistical distribution at all depends on it. If, for example, you take a grab sample of five items, then those five give you only a snapshot of a small portion of the lot, and none of the distribution theory applies.

The lot standard deviation σ plays the key role. This is awkward, because the standard deviation may not be known at the time the sampling is planned. If similar material has been sampled before, a good projection is to pool standard deviations over the previous samplings. If it is practical to take the sample in two steps, then another effective strategy is take an initial sample of n_1 items, where n_1 is the required sample size based on a crude guess of the lot standard deviation, and may be as few as 5 to 10. Then calculate the standard deviation, s, of this initial sample, and use it with the required precision, δ, of the average to determine the size $n = (3s/\delta)^2$ for the combined sample. In the second sampling step, sample the additional $n_2 = n - n_1$ items required.

If you know more about the lot, for example, that it contains runs that differ in mean value or that units in the lot differ in size, then this information can be exploited to design a sampling plan that may be more accurate than a simple random sample. ASTM E1402-08, Guide for Sampling Design, describes types of sampling plans that use the additional information.

More often than not, the number of units determined using the sample size formula will be larger than you can afford to sample and measure. When this happens, remember that sample sizes are always a compromise between precision and cost. You can relax the precision requirement, increasing δ. Variations of the sample size formula also replace the 3 by a smaller number, 2, or a figure from the Student's t table. Doing so increases the chance that the sampling error will exceed δ.

The sample size may also be adjusted to fit a convenient number for sampling and testing. For example, you may estimate that you need 28 items from a lot, but testing is best done in batches of ten. Then 30 samples is the number of units to sample.

PETER FORTINI, *Wyeth Biotech, is a member of Committee E11 on Quality and Statistics.*

DEAN NEUBAUER *is the column coordinator and E11.90.03 publications chair.*

Statistics play an important role in the ASTM International standards you write, such as the development of precision and bias statements for test methods, running interlaboratory studies, knowing how to round numbers properly and determining sample size. A panel of experts is ready to answer your questions about how to use statistical principles in ASTM standards. Please send your questions to SN ***Editor in Chief Maryann Gorman at mgorman@astm.org or ASTM International, 100 Barr Harbor Drive, P.O. Box C700, W. Conshohocken, PA 19428-2959.***