Sample size calculation in homogeneity assessment of certified reference materials

Juris Meija (NRC Metrology)

July 1, 2018

1 Background preparation

Statistics, Experimental Design, Optimization and Optimal Design, Data Analysis.

2 Overview: Measurement Standards

National Research Council Canada, Measurement Standards.

Certified reference materials (CRMs) are key to global comparability of measurement results as most laboratory measurements calibrate or verify their measurement results against the international CRMs. CRMs are typically created in a batch of several thousand units and it is inevitable that the property values (such as the mass fraction of arsenic) might vary between the units. This variability is captured as the uncertainty due to homogeneity of the material and is part of the overall combined uncertainty of the certified values. The question of how many samples need to be tested in order to get a reliable estimate of the homogeneity uncertainty is raised here. This question has financial as well as technical implications. If too few samples are analyzed, one might underestimate the homogeneity of the CRM which in turn might lead to diminished trust in future CRMs or a possible recall. Analyzing too many samples, on the other hand, is undesirable as it requires unnecessary expenses, labour, and depletion of the CRM stock.

3 Problem Description

There are two issues to consider.

First, uncertainties are parameters of probability distributions but we do not have a good statistical model on how impurities are typically distributed in materials. Most statistical methods, such as the one-way ANOVA, assume normal distribution of random effects but this cannot be the case here. The concentration levels of various industrial pollutants are often modeled probabilistically using Weibull, lognormal, or gamma distributions. A better understanding of how to model inhomogeneity distributions will, in turn, lead to better understanding of the meaning behind the uncertainties of CRMs.

Homogeneity of a CRM is typically assessed by performing replicate measurements from several CRM units using a balanced nested experimental design. Furthermore, one typically applies random effects statistical model to the data:

$$x_{i,j} = \mu + A_i + e_{i,j} \tag{1}$$

where $x_{i,j}$ represents the jth measurement result for a certain analyte from the ith unit of the CRM, μ is the population mean, A_i is the effect of unit i (due to inhomogeneity of the sample), and $e_{i,j}$ is random variable representing the measurement uncertainty. Typically these two effects (inhomogeneity of the sample and measurement uncertainty) are modeled as normal random variables $A_i \sim \mathrm{N}(0, u^2_{\text{hom}})$ and $e_{i,j} \sim \mathrm{N}(0, u^2_{\text{meas}})$ and the above statistical model is solved for μ, u^2_{hom}, and u^2_{meas} using one-way ANOVA by employing either traditional frequentist methods or by employing Bayesian methods.

We have a database of some 20,000 data containing measurement results of various chemicals in a variety of NRC chemical CRMs. Analysis of this dataset might be helpful when determining which probability distributions might be appropriate to model the underlying problem. In addition, it should be possible to evaluate the homogeneity estimates arising from a variety of statistical models which would approximate the probability density distribution of the analytes using a variety of non-gaussian distributions.

Second, we are seeking optimal parameters for the experimental design of homogeneity studies. International standards on this matter offer only anecdotal guidance. In addition, there are two schools of thought on this question: some argue that the optimal number of units to be analyzed depends on the lot size whereas others say that it does not. For example, the recommended number of units to be analyzed is $max(10, N^{1/3})$ (ISO Guide 35:2017, https://www.iso.org/standard/60281.html) or $min(15, 0.08N)$ (ASTM E826:1986), or simply that one should inspect a number of units between 10 and 30 (USP 905:2011). Strict application of the power analysis to this problem suggests that the sample size is independent from the total lot size. We are seeking to gain more insights on this matter. It is clear that the power analysis plays an important role.