ORNL/TM-2021/1906

# Essential Statistical Concepts

Stephen Croft
Tom L. Burr

**February 2021**

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

**ORNL/TM-2021/1906**

Nuclear Nonproliferation Division

# ESSENTIAL STATISTICAL CONCEPTS

Stephen Croft
Tom L. Burr

February 2021

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

# CONTENTS

# ACRONYMS

| | |
|---|---|
| AI | active inventory |
| df | degrees of freedom |
| CFR | Code of Federal Regulations |
| CI | confidence interval |
| IAEA | International Atomic Energy Agency |
| ID | inventory difference |
| LEID | limit of error for inventory difference |
| MBA | material balance area |
| MUF | material unaccounted for |
| NDA | nondestructive assay |
| NRC | Nuclear Regulatory Commission |
| PDF | probability density function |
| PoV | propagation of variance |
| SEID | standard error of inventory difference |
| SD | standard deviation |
| TMU | total measurement uncertainty |
| UQ | uncertainty quantification |

*Statistics is, or should be, about scientific investigation and how to do it better,*
*but many statisticians believe it is a branch of mathematics.*
*George E.P. Box (1919–2013)*

# 1. INTRODUCTION

Statistical methods are an essential element to making quality measurements, allowing the organization and interrogation of data in a structured way. This introductory chapter describes key concepts of measurement statistics that are used in many nondestructive assay (NDA) methods and is intended as a primer to the reading list provided in the reference section. (Croft and Burr 2016; Smith 1991; Smith 2013; Taylor 1997; Triola 2017). The textbooks by Triola on elementary statistics (Figure 1), Taylor, and Smith (2013) introduce uncertainty analysis in the physical sciences. Smith (1991) provides a rigorous introduction with examples to correlated variables. These texts collectively offer an excellent foundation for understanding essential statistical concepts.

The presence of a standardized system of measurements is so ingrained that it is easy to overlook. Even so, measurements that can be trusted by both domestic and international practitioners and stakeholders are the basis of trade, enable the development and application of technology, and underpin the scientific method; they are central to human civilization and culture. However, too often measurement practices and uncertainty assessment are neither well defined nor well understood across the different domains. This leads to a lack of consistency and different definitions of terminology, creating confusion, bias, and misunderstandings.

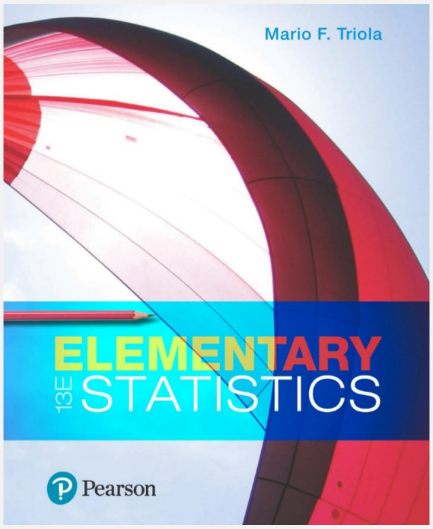

**Figure 1. Mario Triola's *Elementary Statistics* is considered a standard text within the discipline.**

By definition, metrology is the science of measurements. A measurement is a set of operations used to determine the value of a quantity where the object of a measurement is the **measurand**. A basic premise of measurement theory is that at the time of measurement, the quantity has a **definite value**. However, the presence of intrinsic measurement uncertainty means that although great care may be taken to replicate the measurement process, the outcomes across replicates will not match exactly—measurement uncertainty is always present. Therefore, whenever a measured value is presented, the possibility that it is wrong to some degree must be considered, and an appropriate design margin or contingency should be made. Because of this, no measurement result or scientific calculation has meaning unless the measurement uncertainty has also been defensibly assessed. The methodology of the measurement and the corresponding uncertainty assessment must be communicated in detail and with transparency and granularity so that the information can be reliably used for the intended purpose. Once a data set has been shown to be free of any confounding influences that do not need to be reported separately, the uncertainty associated with a measurement is typically reported at the summary level by stating the degree of confidence or degree of credibility (depending on the philosophical framework) that can be defensibly assigned to the measurement value. The measured value is the best estimate of the true, albeit unknown, value of the quantity of interest and is unlikely to be exact and so will almost always be higher or lower than the true value. One simple and common way to report the outcome of the uncertainty quantification process is to report a **confidence interval** (CI). This interval is a range of values that bracket the measured value, typically plus and minus a multiple, *k*, that has been calculated based on the uncertainty assessment. If a normal distribution applies to the uncertainty assessment, then *k* would represent a multiple of the standard deviation (SD). Another common way to quantify the measurement uncertainty is to estimate the so-called “random” and “systematic” **error variance** components, as will be explained later.

Most measurements are neither direct nor absolute but instead rely on calibration of instruments that must first be properly designed, manufactured, and adjusted. Calibration is the process that establishes the mathematical relationship between the response and reference standards. These reference standards are extremely important and are maintained by a worldwide network of national laboratories.
Reference standards are scientifically created values that are accepted, can be maintained, and for which a way exists to scale the standard value upward or downward. As an example, the former International Prototype of the Kilogram revised in May 2019 is shown in Figure 2. Physical reference standards are also used in the development and verification of measurement approaches and as part of quality assurance and performance demonstration programs. Measurement control is the process that ensures the calibration is within defined tolerance and that statistical process control methods can be applied throughout the process. Statistical methods help establish the calibration, monitor the health of the instrument, predict reliability, establish maintenance and recalibration intervals, identify process (performance) improvements, and quantify uncertainty. Statistical methods are used to manage resources (e.g., investment decisions in new equipment or total cost of ownership) and ensure best practices.

**Figure 2. The former International Prototype Kilogram. Source: Bureau International des Poids et Mesures (BIPM).**

Informally, statistics refers to a collection of numbers or facts (e.g., batting averages) but as used in science, statistics refers to quantities calculated from a sample and/or to inference. The **sample** is a subset of a population made by observations and measurements. For example, imagine that a large sack contains a mixture of blue and green marbles. One goal could be to infer or estimate which fraction of the marbles are blue without inspecting each marble. The entire contents of the sack (the mixture of blue and green marbles) are the entire population. If the total number of marbles in the sack is not too large, then it might be possible to remove the entire contents and count the number of blue and the number of green. However, when too many objects exist to reasonably count, one could pull marbles out of the sack to create a sample of size *n*. There are two ways to do this. One way would be to remove the *n* marbles without replacement. Another, sampling with replacement, would be to remove a single marble, record it, replace it, rerandomize the contents and make another selection and do this *n*-times If the number of marbles, *N*, in the sack is huge compared to the number sampled, *n*, then the difference between the information obtained by sampling with or without replacement will be small. However, when *N* is not much greater than *n*, the two approaches are different. In general, for radiometric applications, sampling with replacement is a great approximation. This example is analogous in some ways to making an $^{235}$U/U enrichment determination where the green marbles represent the $^{235}$U wt %, and the blue marbles represent the $^{238}$U wt % of a large sample.

**Descriptive statistics** summarize the properties of the sample of size *n*. **Inferential statistics** extends beyond just the sample to make estimates about the properties of the underlying but unobserved population. To do this, assumptions are made to varying degrees about the mathematical form of the population. The mathematical form can involve model parameters (parametric) or not (nonparametric). This difference is what separates nonparametric and parametric approaches; semiparametric approaches lie in between. Estimates of the population properties are not exact because the sample is an arbitrary subset of the population and is subject to random fluctuation. A pivotal quantity, or pivot, is a function of the sample observations and the (i.e., unobservable, and unknown true parameters of the population distribution, but it can be used to construct statistical tests and CIs. Pivotal quantities can be used to estimate estimator quality. In other words, in practice, fully understanding everything is impossible.

Real data sets, being finite in size (and hence of limited information compared to the population from which they are sampled), are always nuanced. Additionally, a single "correct," wholly objective way of assigning a defensible uncertainty may not exist.

Different users may also have quite different needs for the reported results, and therefore may take different approaches to assessing uncertainty under various circumstances. However, assessing and combining uncertainties for nuclear materials accountancy and control is best and most often approached through frequent and applied measurement statistics. This technique, which determines the probability of an outcome based on the relative frequency of observations, is a powerful tool for providing a useful quantitative uncertainty statement.

The overall or total measurement uncertainty (TMU) is the combination of many contributing influences. A list or a pie chart of fractional uncertainty contributions for a final result is conventionally called an **uncertainty budget**. When the uncertainty contributions (SDs) are independent and combined in quadrature, the information is often also presented in the form of the **fractional variance**. A few large uncertainty contributions typically dominate the TMU. If a lower TMU is needed to meet a particular data quality objective, the uncertainty budget identifies the best opportunities for improvement. These improvements are thus informed by the level of difficulty and resources required to deliver the greatest impact. However, when a particular method cannot meet a specific objective, a different technique may be needed.

A general requirement is that a measurement method should be both **accurate** (a qualitative term meaning close to the true but generally unknown value of the measurand) and **precise** (tightly grouped, small variance when the measurements are repeated). It must also meet the quantitative data quality objectives along with other relevant constraints. To this end, **uncertainty quantification** (UQ) is the process of quantifying the quality of a measurement result and is typically stated as a single numerical value. This value is the total measurement error SD and is a parameter that characterizes the spread or dispersion that can reasonably be attributed to the measurand (assuming that the result is subject to a normal distribution of errors).

The total error of the measurement, $e$, is usually abbreviated simply to "error" and defined as the signed quantity (measured value minus true). Algebraically this can be written as

$$e = x - \tau, \qquad (1)$$

where $x$ is the measured value of the measurand, and $\tau$ is the (true) value of the measurand. This definition and mathematical formulation are clearly the most natural in the linear (additive) model but can also apply to the multiplicative model case.

A perfect measurement would return the true value each time. However, the true value of the measurand is never known. Indeed, the usual goal of the measurement process is to estimate the true value. However, in some cases, an item can be prepared (e.g., for calibration purposes or for performance testing) using methods with accuracy superior to those of the in-situ NDA measurement technique. In these cases, the value of $\tau$ may be considered as well known. Such accepted values, which may be established by superior analytical techniques or by convention, are sometimes referred to as **conventional true values** or **nominal true values**. Sometimes a particular value is adopted as a matter of convenience or by comparisons and such values are known as **consensus values**. Conversely, the causes of error can be positive or negative. Table 1 describes examples of each.

**Table 1. Potential causes of total error**

| **Example potential causes of total error** | |
|---|---|
| Mistakes or transcription errors | Should be spotted and corrected during self-checking, working results in multiple ways and in peer review; ideally, these mistakes lead to outlier values that can be rejected due to assignable cause |
| Bad practice | Inadequate training may lead to unrecognized consequences; the use of inappropriate instrumentation or calibration items; extrapolating beyond a demonstrated dynamic range; assuming linearity; not checking for hysteresis effects |
| Poor assumptions | Acceptable knowledge and process knowledge should be confirmed and documented; when processes change, the acceptable knowledge (such as scaling "fingerprints" of difficult-to-measure nuclides from marker nuclides) needs to be reestablished; neglect of correlations needs to be justified |
| Interferences | May be recognized and correctable with an associated uncertainty or might not be recognized; can obscure signatures, making certain nuclides difficult to quantify |
| Model error | Arbitrary fits used outside the range of validity; the adoption of fixed consensus values such as specific gamma emission rates over best scientific values unless a suitable physics-based representation of the behavior is available, the functional description is only a convenient approximation and there will be regions where deviations will inevitably be larger than others; closely related to item-specific bias |
| Random fluctuations | Make measurements neither repeatable nor reproducible |
| Influences | Causes the measurement not to be a fair (unbiased) estimate of the (true) value of the measurand; an example is the use of a "black box" data analysis code, which might use hard-coded parameters, such as half-life and branching ratios that are out of date—although it may be of interest for consistency with historical results, it could be criticized from the perspective of transparency and best practice |

A direct determination of a quantity is rarely made. Instead, the measurement is usually the result of interpreting a response function. Uncertainty assessments of complex measurement systems and procedures often require subject matter expert professional knowledge, experience, and skill. The expertise of a professional statistician may also be of great value.

Two important classifications of error are recognized: **random** and **systematic**.

Random variability in principle can be quantified by empirical statistical methods and can also be reduced by making repeat measurements. Precise measurements exhibit good reproducibility. Random error, $e_r$, of the measurement is the difference between the measured value, $x$, and the mean, $\lambda$, that would result from an infinite number of measurements of the same quantity under repeatable conditions. Algebraically, this is represented as follows:

$$e_r = x - \lambda. \tag{2}$$

However, systematic effects (such as operator-specific error and item-specific error) are persistent, consistent, and reproducible and cannot be revealed by repeated measurements. Certain drift mechanisms and system aging may fall into this category, depending on the time scale of the data collection. Systematic effects can cause even precise measurements to be inaccurate [e.g., far from the (true) value of the measurand]. Mathematically, systematic error, $e_s$, is defined as the difference between the long-term average of the measured value, $\lambda$, and the (true) value, $\tau$, of the measurand. Algebraically, this can be expressed as:

$$e_s = \lambda - \tau. \tag{3}$$

Systematic error is used interchangeably with bias because it relates to a measuring instrument. Sometimes, systematic error is partitioned into short term, such as during one inspection period, or long term, such as during the entire data analysis period consisting of multiple inspection periods (Zhao 2010).

When a systematic error can be identified (e.g., through analysis or intercomparison of methods), good practice is to try to minimize it by design and apply a suitable correction factor when possible. The remaining or residual systematic error then comes from the remaining uncertainty in the correction factor.

Combining equations (2) and (3) leads to

$$e = x - \tau = (x - \lambda) + (\lambda - \tau) \tag{4}$$

or

$$e = e_r + e_s. \tag{5}$$

Although knowing either $e_r$ or $e_s$ (and therefore $e$) exactly is not possible, estimates of their typical magnitude can be made. A formal discussion on how to do this will be proposed later in the chapter but for now the results will be used. Let $\sigma_r$ and $\sigma_s$ denote the estimated random and systematic standard uncertainties (SDs), respectively. Assuming that the random and systematic effects are independent, the combined SD, $\sigma_c$, can be evaluated from the quadrature sum: $\sigma_c \approx \sqrt{\sigma_r^2 + \sigma_s^2}$. Note that a contribution to the combined standard uncertainty is either classified as random or systematic depending on (or conditioned by) the intended use for the measurement result. For example, to another person the random uncertainty assigned to a nominal value of a calibration item may become a systematic uncertainty for measurements that rely on the resulting calibration.

Detailed UQ is usually undertaken at the measurement process-design stage. Comparison between design performance and routine or achieved performance can identify reasons for significant differences.

The *Guide to the Expression of Uncertainty in Measurement* (Chunovkina and Chursin 2001) defines two general types of uncertainty evaluation: (1) Type A evaluation is based on the statistical analysis of a series of observations; and (2) Type B evaluation is based on any means other than the statistical evaluation of a series of measurements, such as the following:

- using data taken from handbooks,
- compilations and evaluated data files,
- vendor specifications,
- certificates and other reports, and
- prior experience including previous measurement data.

Describing and reporting these contributions separately is good practice, although when evaluating the TMU on an individual item, they are combined without distinction (i.e., they are treated on the same probabilistic footing as Type A contributions). To do this, Type B uncertainties must be associated with an assumed probability density function (PDF). Therefore, for example, suppose the temperature coefficient of an instrument from type test data is provided by the manufacturer. The manufacturer may have used statistical methods, but this instrument was not part of that study, although it is assumed to be typical of the instruments that were. Assume that the manufacturer's guidance is adopted and made use of in all assays. The value does not change, but it is not known perfectly. The manufacturer's range of values is assumed to apply to the instrument, and a Type B uncertainty is added to the TMU. In doing so, the state of belief might be represented as a rectangular distribution or as a normal distribution to interpret the adopted variance and to develop CIs accordingly. This effect can be seen in the earlier description of how to estimate the combined standard uncertainty. In general, variances can be added, but when the underlying distributions are not all normal, then CIs will not necessarily correspond to those of a standard Gaussian PDF. Figure 3 shows a normalized Gaussian distribution function for various values of μ and σ.

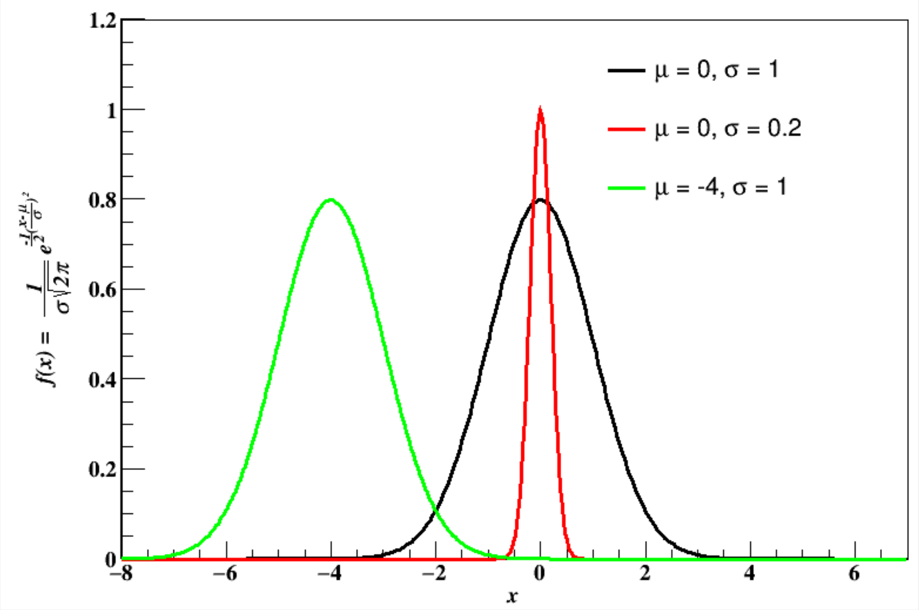

**Figure 3. Examples of a normalized Gaussian PDF for various values of μ and σ.**

The interpretation of measurement results is inherently **probabilistic**. Statistical methods and reasoning therefore underpin measurement science and UQ, even though UQ is neither a wholly mathematical nor wholly prescriptive undertaking.

The remaining sections of this chapter review key statistical concepts. Before moving on, take a moment to think of examples from your own experience.

## 2. DESCRIPTIVE STATISTICS

When an item is measured repeatedly (under the same conditions, which we usually take to mean in close succession) or when an entire experiment is replicated (using different hardware, different operators, etc.), often a spread of results is observed. However, there will also usually be a clear single clustering of results (unimodal) with the chance (relative frequency) of extreme values that fall steeply the further away the value is from the main group.

Even though there is some spread, the measurement results can be concisely summarized using just a few numbers without making any assumptions about the shape of the underlying or parent frequency or the probability distribution (i.e., nonparametrically). Describing the results using just a couple of numbers is a considerable simplification compared to having to use the full list of results. Important sample properties can be described using statistics. A sample statistic is just the name given to both the value and to the function used to calculate it from the set of data, subject to some mathematical formalities—such as the form of the function does not depend on the particular sample. For example, summary statistics concisely express what the sample implies about the underlying population (which is usually too big to know fully) without the need to assume an underlying mathematical model. This is the nonparametric approach.

In contrast, a parametric approach involves interpreting the data *within* an underlying mathematical model, which is described by model parameters. Whether a model is appropriate should be checked before relying on it. The observed sample data is the only direct connection to reality. If good care was taken to collect good data and disagreement exists between the model and the data, then the model is possibly naïve.

The main things to quantify for a sample are known as LDPOT:

- **Location**—the position of the data on a scale,
- **Dispersion**—a simple measure of how good the number is,
- **Probability distribution function**—a detailed description of all the possible outcomes,
- **Outliers**—the data set consistent with the model, and
- **Trends**—aspects measured by time, item category, operator, and so on that could confound interpretation.

## 3. MEASURE OF LOCATION

The **mode**, **median**, and **range** are used to describe the general **location** (with suitable units). In measurement science, the measure of location is the **mean**. This is also commonly referred to as a measure of the central tendency of the data.

Consider a list of *n* values of an **independent random variable**, *X*, the sample mean, $\bar{x}$, is calculated as follows (the uppercase *X* denotes a random variable; the lowercase *x* denotes a realized value of *X*):

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i. \tag{1}$$

Typically, the average is the best estimate (has the smallest average squared error) of the mean of the underlying population distribution. Every value is treated on the same footing and $n \cdot \bar{x}$ is the total amount "of stuff," as is to be expected. As $n \to \infty$, the computed value $\bar{x}$ tends to a constant, the mean of the underlying population distribution.[1]

## 4. MEASURE OF DISPERSION

The **sample variance** is the usual mathematical way to measure the spread, variability, or dispersion of the sample. The sample variance, $s^2$, is defined as follows.

$$s^2 = \frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{1}$$

The factor $\frac{1}{(n-1)}$ makes $s^2$ an unbiased estimator of the population variance, which means that in repeated samples of size n, the average value of $s^2 = \hat{\sigma}^2$ (the "hat" denotes an estimator) is the true population

[1] As an aside, the field of mathematical statistics distinguishes among several types of convergence, such as strong or weak convergence; such distinctions will not be needed here.

variance $\sigma^2$. It can also be shown that $s^2$ is the minimum variance unbiased estimator, which from a practical standpoint, means that it is efficient at approaching the population value. Again, as $n \to \infty$, intuitively it can be seen that $s^2$ converges to a constant value characteristic of the underlying complete population.

The sample SD is called $s\ (= +\sqrt{s^2} \geq 0)$.

The sample standard error is $se = s/\sqrt{n}$.

The term standard error was originally defined by the British statistician George Udny Yule (1871–1951), who laid the origins for his work in an 1899 paper on the causes of pauperism in England (Yule 1899). See *An Introduction to the Theory of Statistics* (1911) to gain insight into how statistical theory was approached at that time.

Standard error is important because it is a measure of the random error in a sample statistic, such as a mean. Such statements make sense because sample statistics behave randomly similar to the way individual measurements do. Whereas the sample SD is a measure of the dispersion of an individual repeat value, the standard error of the mean $se = s/\sqrt{n}$ is a measure of the dispersion on the sample mean of the reported estimate of the measurand. A simple and powerful fact is that the variability of the sample mean across hypothetical or real replicates of obtaining a sample of size $n$ can be predicted quite well by the sample $se$.

As discussed earlier, a pivotal quantity, or pivot, is a random variable defined by a function of (sample) observations and unobservable (population) parameters with the property that its probability distribution function does not depend on the unknown (population) parameters. We now introduce the **pivotal quantity**

$$t = \left(\frac{\bar{x} - \mu}{s/\sqrt{n}}\right), \tag{2}$$

which, for a **normal population**, is distributed with a student's *t*-distribution with $\nu = (n-1)$ degrees of freedom (df). This fact allows CIs to be placed around $\bar{x}$, having a given probability of containing the true but unknown value of $\mu$, the mean value of the underlying population that is being estimated. Notably, however, this is important when adopting results stemming from any parametric model that the conditions under which the mathematical model applies are appropriate for the given data set. This can often be difficult to do, and so may often just be a guess.

In inferential measurement science, what is reported as a form of shorthand, $\bar{x}$, is the best estimate result of the measurement and $s/\sqrt{n}$ as the optimal statement of the associated measurement uncertainty of the sample mean statistic. Extra information, such as the number of df, may be needed by a user of the reported values, $(\bar{x} \pm s/\sqrt{n})$, to properly interpret what it means in a way that is fit for their intended purpose.

**Descriptive statistics** describe the data. **Inferential statistics** make inferences or predictions from the data. This includes estimating parameters for the population, which is a generalization from the sample, and hypothesis testing.

This presentational form for $(\bar{x} \pm se)$ must be treated carefully, especially when results are reported with expanded standard uncertainties. CIs are revisited in the following section after discussion of the Central Limit Theorem. Often, a normal distribution can be used to describe the central region of the probability distribution of the measurement outcome, and a standard uncertainty value, $u$, is derived such that user-specified confidence exists that the true value will be within ±$u$ of the measured result.

The relative standard error, $rse = \frac{s/\sqrt{n}}{\bar{x}}$, can be quoted either as a fraction or as a percentage. Beware! Sometimes the context and traditional relative SD may be used to describe the same thing. For a population, $\delta = \sigma/\mu$ is referred to as coefficient of variation, although its use in NDA is not widespread.

**Example:** Suppose a sample of size $n$=3 comprises the numbers 1, 2, 3 with units of kg. The principal sample statistics are as shown in Table 2.

**Table 2. Sample statistics.**

| Index, $i$ | $x_i$, kg | $deviation = x_i - \bar{x}, kg$ | $deviation^2$, kg$^2$ |
|---|---|---|---|
| 1 | 1 | -1 | 1 |
| 2 | 2 | 0 | 0 |
| 3 | 3 | 1 | 1 |

$$sample\ mean, \bar{x} = \frac{1+2+3}{3} = 2.00 \text{ kg}.$$

Notably, the sum of deviations about the sample mean is zero, as it should be. (Exercise: Starting with (17.6), show that the sum of the deviations about the sample mean by definition is equal to zero.) The mean is intuitively gratifying because it is easy to appreciate. For example, potatoes to make a stew are sold by weight, not by piece. For making a stew, no difference exists between purchasing three average potatoes of 2 kg each or a sampling of three potatoes weighing 1, 2, and 3 kg at the grocery store.

$$sample\ variance, s^2 = \frac{sum\ of\ deviations\ squared}{n-1} = \frac{2}{2} = 1.00 \text{ kg}^2.$$

$$sample\ standard\ deviation, s = +\sqrt{s^2} = 1.00 \text{ kg}.$$

$$standard\ error, se = \frac{s}{\sqrt{n}} = \frac{1}{\sqrt{3}} \approx 0.58 \text{ kg}.$$

Other defensible ways to estimate the dispersion of the result exist, given the limited experimental data.

The premise of the bootstrap method is that each of the three results is equally likely. Thus, data sets can be constructed from the original data purely for the purpose of estimating variability through the process of sampling with replacement. The first of the three draws can be 1, 2, or 3, the second of the three draws can be 1, 2, or 3, and so on. This gives 27 ($3^3$) possibilities, and the SD of the mean for each of the 27 cases (which is 0.48) provides an estimate of the standard error of the original data set. For larger data sets, such as $n$ = 100 instead of $n$ = 3, far too many possibilities exist to enumerate, so one simply computes a reasonable number such as 1,000 bootstrap samples of size 100. The true variance is unknown in this case, so determining whether 0.58 is a better standard error than 0.48 is not possible.

The jackknife technique is a scheme based on rejecting each data in turn (Miller 1974).

### Sidebar 1. EXAMPLE

Fabrication techniques for a 16 × 16 pixelated array detector are being developed. A batch (sample size, $n$) of 256 sensors was tested and 17 were found to be defective. The estimated probability of a defective element is therefore $\hat{p} = \frac{17}{256} \approx 0.0664$. The theoretical SD of this estimate is $se(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$, but because $p$ is unknown, the expression is evaluated using the experimentally estimated (sample) value $\hat{p}$. Thus:

$$\widehat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\frac{17}{256}\cdot\frac{239}{256}}{256}} \approx 0.0156.$$

The estimated proportion of defective sensors for this production process based on this one sample is therefore (6.6±1.6)%, where the uncertainty indicates the approximate 68% CI (this would conventionally be called the margin of error in the context of a political poll intended to estimate a simple proportion statistic). The quantiles of the pivotal *t* statistic should provide a better-quality CI. However, in this case, *n* = 256 is so large that the quantiles of the *t* are essentially the same as the quantiles of the Gaussian. CIs for ratios is a separate topic in itself (Agresti 1998).

In reporting numerical values, the dilemma arises of knowing how many significant figures to use so that rounding errors (in either the mean or SD or both) do not introduce significant errors when the results are subsequently used in other calculations (for instance, in a weighted mean of values).This is especially true when the uncertainty in the SD can be large. Implying greater confidence in the results than can be justified would be bad practice. However, as noted earlier, reporting the SD to at least two digits and using it to define the least significant figures of the measurand are recommended. For example, (9.81±0.41) and (10.08±0.67) abide by this recommendation. Dean 2008 provides arguments for supporting this practice (Dean 2008). To summarize, usually giving the uncertainty to two significant figures and the results reported to match resolution are recommended, for example, $(1.953 \pm 0.028)$ kg, which is also commonly written as $1.953(28)$ kg.

Of course, if these results had been obtained from successive repeated measurements, the difference in the estimates in the SD might be attributable to chance alone. Taking a simple average rather than a weighted average is likely to be the more appropriate thing to do. In other words, some judgment may still be needed to interpret the data.

## 5. CENTRAL LIMIT THEOREM

In a scheme that is cleverly known as "the method of moments," the sample mean and sample variance are statistics used as estimates for the corresponding properties of the underlying population of events. Consequently, $\bar{x}$ and $s$ are themselves random variables and can also be sampled and studied. Suppose the underlying population of possible values has a mean and variance of $\mu_p$ and $\sigma_p^2$, respectively. Then, one form of the **Central Limit Theorem** states that if many samples of size $n$ are obtained from a much larger population, $n_p \gg n$, then the expectation value (long-term average) of the sample mean and the expectation value of sample variance are given by

$$\mu_{\bar{x}} = \mu_p, \tag{1}$$

and

$$\sigma_{\bar{x}} = \frac{\sigma_p}{\sqrt{n}}, \tag{2}$$

where $\sigma_{\bar{x}}$ is called the standard error of the sample mean and describes how the variation in the sample mean is less than the variation in individual values. Further, for large $n$ ($\gtrsim 25$),

$$\bar{x} \div N\left(\mu_p, \frac{\sigma_p^2}{n}\right), \tag{3}$$

which should be interpreted as the sample mean is approximately distributed as a normal (i.e., Gaussian or Laplace-Gauss) distribution (PDF) with mean $\mu_p$ and variance $\frac{\sigma_p^2}{n}$.

The Gaussian distribution has a characteristic symmetric bell shape with the following mathematical form, which proves to be especially convenient to work with shown mathematically:

$$g(x \mid \mu, \sigma^2\ ) \cdot dx = \frac{1}{\sqrt{\pi}} \cdot e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2} \cdot \frac{dx}{\sqrt{2}\sigma}, \tag{4}$$

where $g(x \mid \mu, \sigma^2\ ) \cdot dx$ is approximately the probability that the value of the random variable will be in the incremental interval of width $dx$ about $x$. The exact probability is the integral of $g(x \mid \mu, \sigma^2\ )$ from $x - dx$ to $x + dx$. See Figure 3 for a graphic representation.

Note that $g(x)$ is a two-parameter function of a real continuous variable. The mean is $\mu$ $(-\infty < \mu < \infty)$, and the variance is $\sigma^2$ $(\sigma^2 > 0)$. The integral under the curve is unity because the distribution is a true normalized PDF. The fractional area under the curve between two boundaries, $\Pr(a \leq x \leq b) = \int_a^b g(x) \cdot dx$, is the probability that the value will take on a value between the boundaries.

**Sidebar 2.** **PRACTICE**

**Noise that Can Cause Measurement Results to Scatter**

1. What are a few examples of noise that can cause results to scatter?
2. Generate a sample with each value being synthesized by adding independent random variables drawn from different distributions. For example, toss a coin 20 times. Assign 1 to heads and 0 to tails and record the totals for all 20 tosses. A histogram would show a somewhat bell-shaped spread even though the "things" that are summed came from non-normal distributions.

   This problem is treated similarly in the discussion on the binomial distribution for nuclear counting. The binomial distribution also describes the number of defectives in the example above involving fabrication techniques for a 16 × 16 pixelated array detector.

In another Central Limit Theorem example, the sum of random variables with finite variance leads to the emergence of a normal distribution. For making measurements, many sources of influence cannot be controlled, fluttering on a short timescale from one measurement to the next and contributing to the inherent variation in the measurement result. Therefore, treating measurement variability is quite common, assuming that a normal distribution adequately describes at least the central region (>95%) of outcomes.

## 6. NUCLEAR COUNTING EXPERIMENTS

Imagine that a number, $n$, of nuclei each has a fixed probability, $p > 0$, of decaying and being detected in a time period, $t$. Let the probability of detection $p$ be called the probability of success and the probability of not being detected, $q = (1 - p)$, be called the probability of failure. With $n$ fixed and $p$ fixed (each nucleus behaves independently) the complete probabilistic summary of possible outcomes is described by the discrete *binomial distribution*, $b(k \mid n, p), 0 \leq k \leq n$, stemming from the $n$-independent Bernoulli trials. Thus, the following can be written:

$$(p+q)^n = 1 = \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!} p^k q^{n-k} = \sum_{k=0}^{n} b(k \mid n, p)\,, 0 \leq k \leq n \tag{1}$$

There are various notations for the binomial coefficients $\binom{n}{k} = {}^{n}C_{k} = \frac{n!}{k!(n-k)!}$, which is to be read as "$n$ choose $k$," and gives the number of combinations or choices of $k$ successes from $n$ attempts with the order of the arrangement being unimportant. The values of the binomial coefficients are familiar from Pascal's triangle [Exercise: write out Pascal's triangle now]. The form of the binomial distribution and how to generalize it can be visualized by thinking about a coin-tossing experiment. The outcomes of a single coin toss are heads or tails (H or T, respectively). The outcomes of two coin tosses (H+T)(H+T) are HH, HT, TH, TT, which can be mathematically codified as

$$\text{“}1 \cdot H^2 \cdot T^0 + 2 \cdot H^1 \cdot T^1 + 1 \cdot H^0 T^2\text{,”}$$

and so on by repeated multiplication and collection of combinations. In this case, the power of $H$ gives the number of heads, and the coefficient gives the frequency. Setting H = T= 1/2 returns the probability (or alternatively one can normalize the outcomes) [Exercise: work through this example now and extend to three coin tosses. Note: There are some good videos on-line of the quincunx machine that also illustrates the point.] The mean and variance of the binomial distribution are $n \cdot p$ and $n \cdot p \cdot (1-p)$, respectively. Notably, as $p \to 0$, the numerical values converge.

### Sidebar 3. THE BINOMIAL, POISSON, AND GAUSSIAN DISTRIBUTIONS

The probability that off-site electrical power will be lost at a nuclear facility is estimated to be constant at 0.43/year. Over the 40-year operational life of the facility, what is the probability that off-site power will be lost at least once?

Hint: Work with the complementary event.

Answer: The probability that power will be lost one or more times after 40 trials is being sought. This is equal one minus the probability that power will never be lost and is given by

$$(1 - p(0)) = 1 - (1 - 0.043)^{40} \approx\ 0.172.$$

Under almost all conditions of practical interest, the binomial distribution can be mathematically recast by letting $n \to \infty$, $p \to 0$, $\mu = n \cdot p =$ constant, and $\sigma^2 = n \cdot p \cdot (1-p) \to \mu$, resulting in the *Poisson distribution.* The Poisson distribution can be derived as a basic distribution in its own right (e.g., to describe the sporadic annual number of deaths in the Prussian cavalry from horse kicks). For example,

$$p(k \mid \mu) = \frac{\mu^k \cdot e^{-\mu}}{k!}, 0 \geq k(an\ integer\ value) \leq \infty,$$

which is much simpler to deal with than the binomial. Note that the Poisson distribution is discrete and is fully specified by only a single parameter, the mean, $\mu$ a real number $> 0$, and $\sigma^2 = \mu$. Experimentally, this is an extremely important point because it means that from the result of *a single nuclear counting experiment* ("count"), one can obtain (through a mathematical process known as inversion which will not be covered here) both an estimate of the mean ($\hat{\mu} = N + 1$) and from it also of variance ($\hat{\sigma}^2 = \hat{\mu}$). Hence, one can make a quantitative estimate of the "counting statistics" reliability of the result. The Poisson distribution is the fundamental distribution of nuclear counting. When $\mu$ is small, the distribution is highly skewed toward small values of $k$. However, as $\mu$ becomes larger, the distribution becomes more symmetrical and gradually morphs into the shape of a Gaussian function $g(k \mid \mu, \mu\,)$ that is

$$\lim_{\mu\to\infty} \left(\frac{\mu^k \cdot e^{-\mu}}{k!}\right) \approx \int_{k-0.5}^{k+0.5} \frac{1}{\sqrt{\pi}} \cdot e^{-\left(\frac{x-\mu}{\sqrt{2\mu}}\right)^2} \cdot \frac{dx}{\sqrt{2\mu}} \approx g(k \mid \mu, \mu\,) \cdot 1 = \frac{e^{-\left(\frac{k-\mu}{\sqrt{2\mu}}\right)^2}}{\sqrt{2\pi\mu}}, k(integer) \geq 0.$$

For describing the central part (e.g., 95%) of the discrete Poisson distribution in nuclear counting experiments, the Gaussian approximation becomes "reasonably good" for $\mu \gtrsim 15$, with some flexibility, depending on the application.

Suppose 1 g of $^{235}$U exists that is $\sim \frac{1}{235.04\ \mathrm{g/mol}} \cdot 6.0221 \times 10^{23} \mathrm{atom/mol} \sim 2.56 \times 10^{21}$ atoms. The half-life of $^{235}$U is 703.8(5) × $10^6$ years (Chadwick et al. 2011), which corresponds to a probability for a given nucleus to decay per s [i.e., $\lambda = \ln(2)\,/(2.22 \times 10^{16}\ \mathrm{s})$] of approximately $3.121(2) \times 10^{-17}$. To illustrate the concept, suppose the probability of emitting a 185.7 keV photon is 0.570(6), the probability of the photon escaping the object is 0.6, the solid angle probability of striking a detector is 0.045, and the probability that the photon will fall into the set energy-deposition region of interest is 0.25. For a 1,000 s observation period, the probability of a successful detection is of the order of $1.2 \times 10^{-16}$. This verifies that the approximation of small probability is confirmed. Therefore, the variance is numerically equal to the mean, and the binomial distribution may be replaced by the Poisson distribution for nuclear counting examples.

Consequently, *almost all* NDA assessments of detection limits, CIs, and so on make use of the Gaussian approximation. This Gaussian approximation results in considerable technical simplification in combining and reporting measurement uncertainties but is not always a wholly satisfactory approach. However, for the rest of our discussion we shall assume that it is!

To illustrate using the Gaussian approximation in nuclear counting experiments, suppose 1,618 events are recorded over a 60 s interval in a region of interest—the pulse height spectrum. The best estimate of the count rate is $1{,}618/60 \sim 26.97$ counts per sec (cps). The associated SD is $\sqrt{1618}/60 \sim 0.67$ cps, and one would traditionally report the result as $(26.97 \pm 0.67)$ cps with a statement to the effect that the uncertainty is counting precision at the 1σ level. This is equivalent to specifying that the coverage factor is unity ("k = 1"). The assumption that the distribution is being approximated by a normal distribution (with infinite df) derived from integer values is usually implicit from the context and common use.

Suppose that no events ($N = 0$) are observed in a given Poisson experiment. Then, is it reasonable to assign an expected mean of zero with zero variance, which implies perfect knowledge? However, intuitively, this seems wrong. From a single observation of random behavior (it should be obvious) that full and complete knowledge cannot be attained. There are technical arguments in general (related to inversion on maximum likelihood with a flat *prior*) for using $\mathrm{mean} = \mathrm{variance} = (N + 1)$. This discussion has avoided the complication of small numbers by requiring $N$ to be sufficiently greater than zero.

## 7. CONFIDENCE INTERVALS

The most complete way to communicate the confidence in a measurement is to provide an estimate of the complete probability distribution function (PDF) along with the estimated value. In cases where the PDF can be approximated by a normal distribution one, common convention is to report the value along with an error bar of plus and minus one SD of the normal distribution. Recall that for a normal distribution, the mean and SD fully define the PDF. Whenever data are presented in this style, it is important to clearly state what convention (e.g., plus and minus one standard error) and other assumptions (e.g., normality) are being made and what other information is needed (e.g., sample size or effective degrees of freedom (df)) to interpret the uncertainty statement. The number of degrees of freedom is an important concept that will not be explored in detail here. But it is related to the fact that if one calculates the mean from a sample then there is only freedom to write (*n*-1) results because given these and the mean the *n*th result can be calculated. Thus, factors of (*n-p*), where $p$ is the number of derived parameters, often appear in statistical formulae.

The standard error calculated from a sample data set is itself a statistic or estimator, so it can legitimately be asked what is the "uncertainty in the uncertainty." This can be framed in a very general way, but here only the result for the normal distribution is quoted in a somewhat stylized way:

$$\bar{x} \pm \frac{s}{\sqrt{n}} \cdot \left(1 \pm \frac{1}{\sqrt{2(n-1)}}\right). \tag{1}$$

We see that the fractional uncertainty in the uncertainty is of the order of $\frac{1}{\sqrt{2(n-1)}}$. It takes a sample of size $n = 51$ before this factor reduces to 10%. (For a sample of $n = 6$, it is about 32%, a magnitude that in a different context could be thought of as a detection limit.) Although one is not usually interested in the value of $se = s/\sqrt{n}$ per se, only in how it helps express confidence in the estimated location result, this serves as a reminder that statistical estimates are not exact. Sometimes statistical estimates may be rather crude. The variance depends on the square of deviations, so large deviations contribute more. As a consequence, the variance will then scatter more, and so it requires more data points to locate it precisely.

Given the sample's statistics $\bar{x}$ and $s/\sqrt{n}$, it is only natural to ask how confident one is in the true but unknown population mean, $\mu$, that lies within some interval about the sample mean $\bar{x}$. Recall that for a normal distribution, the pivotal quantity $t = \left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right)$ is distributed according to a student *t*-distribution with $(n-1)$ df (here occurs a slight abuse of notation to use the lowercase "*t*," because lower case denotes a realized value and not a random variable; the right-hand side terms are lower case.) From this example, it can be shown that the $100 \cdot (1-\alpha)\%$ two-sided CI for $\mu$ can be expressed as

$$\mu = \left(\bar{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right). \tag{2}$$

Values of $t_{n-1,\alpha/2}$, the coverage factors for students *t*-distribution ("*t*-distribution table of two-sided critical t values"), can be generated in Microsoft Excel using the function call $TINV(\alpha, n-1)$.

For example, suppose a sample of six data points exists, and a 95% CI is desired (which is a common but arbitrary choice).

In this case, $\nu = (n-1) = 5$, $\alpha = 0.05$, and the **expanded uncertainty** becomes $\approx 2.57 \cdot \frac{s}{\sqrt{n}}$. In contrast, in the limit $(n-1) \to \infty$, and the sample become truly representative of a normal, the multiplier tends to $\approx 1.96$. The 95% confidence level is an arbitrary but commonly encountered choice. At this level, the probability (in a frequentist sense over many such CIs) of the true value falling outside the range is still $1/20$.

In other words, the actual meaning of the CI is somewhat different from the usual (mis)interpretation given above. In the usual interpretation, a CI of 95% is thought of as containing the true value with a probability of 95%. However, statistically what it means is that if the same CI construction method were applied many times, then 95% of the experimental CIs would include the true value. This is an important distinction when CIs are estimated through simulation, for example, by Monte Carlo sampling of the physical behaviors of a system. Strictly speaking, from a Bayesian perspective, one does not interpret a frequentist CI conditional on the data. Rather, a CI construction procedure is characterized by the coverage probability (whether the true values lies in the CI) over many similarly constructed intervals.

## 8. CONFIDENCE INTERVALS FOR A SINGLE POISSON OBSERVATION

For a single observation (sample size of $n = 1$) for which the result is $k$ counts collected from an assumed Poisson distribution with mean μ, an "exact" CI for μ with a confidence level (1-α) is given as

$$\frac{1}{2} Inv\_\chi^2(\alpha/2\,;2k) \geq \mu \leq \frac{1}{2} Inv_{\chi}{}^2(1-\alpha/2\,;2k+2), \tag{1}$$

where $Inv_\chi^2(x,\nu)$ is the inverse of the left-tailed probability of the chi-squared distribution.

Because the number of counts must be an integer, the conservative approach is to round the lower limit values down and to round the upper limit values up.

The chi-squared distribution, $\chi^2(x,\nu)$, with $\nu$ df for the variable $x$, has the form

$$\chi^2(x,\nu) = \frac{1}{2^{\nu/2} \cdot \Gamma(\nu/2)} \cdot x^{\nu/2-1} \cdot e^{-x/2}, \tag{2}$$

where $x$ is a real positive number, $\nu$ is a positive integer greater than or equal to one, and $\Gamma(z)$ is the gamma (factorial) function defined by Euler's Integral (Abramowitz 1968) [Note that, for $z = m$, and integer value, the form needed for this problem is $\Gamma(m) = (m-1)!$].

Although this result is well known in the statistical community, these confidence limits are rarely applied to nuclear counting in practice. For example, they are not applied to a curve-fitting algorithms used in gamma-ray spectroscopy. Instead, the normal approximation is typically invoked. The practical benefit of this simplification is even greater when the difference of two count distributions is considered (signal equals gross counts minus background). Formally, the difference of two independent Poisson random variables has a Skellam distribution.

## 9. TECHNIQUE SELECTION AND INSTRUMENT DESIGN

Several complementary methods and multiple physical realizations may be available to measure items of a given type and character. An appropriate selection of method and instrument that balances the conflicting objectives requires critical thinking and should involve the collective experience of the whole team.

## 10. CALIBRATION

Calibration is the procedure to establish the causal relationship between a measurand, the predictor variables, and other quantities (e.g., mass-deflection, energy-channel, and volume-level). In the simplest case, calibration establishes proportionality under controlled conditions. Thus, calibration can be established from the response to known reference items or standards or it can be established through comparison against an accepted standard instrument.

The items used for calibration must be traceable to internationally or nationally recognized standards through an unbroken chain of comparisons. International and national standards are the top tier of standards, but they are few and must be diligently maintained. As one progresses down the hierarchy of standards to the primary, secondary, and working levels, the uncertainty generally expands as the comparison uncertainty incurred at each stage contributes to the total uncertainty. In general, however, the accuracy of the calibration standards should still be small (1/3 is a common rule of thumb) compared to the overall calibration uncertainty goal.

Calibration requires both careful planning and careful execution by trained and experienced personnel who understand the measurement instrument, the basis of the assay technique, and what has to be done to obtain a calibration that meets the data quality objectives and why, including the consequences of taking liberties with the procedure. Known items that represent the unknowns to be assayed are commonly used. Special attention to blanks (background) and interferences is required. The following practices are crucial to establishing and maintaining a credible calibration free from unidentified systematic uncertainty that does not show up in repeat measurements on a reference item:

- Cross calibrations,
- Participation in interlaboratory comparisons,
- Round-robin exercises,
- A robust quality control program, and
- Regular performance demonstration measurements.

A written calibration procedure usually includes sections covering the following:

- Purpose, scope, definitions, and references;
- Attachments, equipment, and materials required;
- Safety, prerequisite conditions, test procedure; and
- Recording templates, acceptance criteria, approvals.

The frequency and accuracy objectives of a calibration depend on the importance of the data being generated and the consequence an error. Measurement control is also used to maintain tolerances and provide ongoing estimates of error SDs. Individual sensors as well as system-level performance can be subjected to calibration and measurement control. Initial factory calibrations are often replaced in whole or in part by field calibrations performed in situ to correctly incorporate the conditions of actual assays.

The calibration procedure may often be witnessed by independent experts to ensure honest execution, attention to detail, and integrity of reporting. Excellent documentation is crucial because, in addition to conveying quality to the client, it is the only evidence that calibration was done as intended. The calibration report also provides a way to record pertinent observations or changes occurring during field work.

The simplest calibration is that for a proportionate (linear through the origin), physics-based response function performed using a single calibration item. An example is when the Enrichment Meter Principle (EMP) is used to determine the attribute $^{235}$U enrichment ($^{235}$U:$^{tot}$U atom %) of a homogeneous compound under fixed geometry using a collimated high-resolution gamma-ray spectrometer. In this case, the net full-energy peak area counting rate $C$ of the combined 182.6 + 185.7 (both lines come directly from $^{235}$U) keV gamma-lines is obtained using a three region-of-interest algorithm, and the rate, $C$, varies in direct proportion to enrichment, $\alpha$. Calibration usually occurs within the causal relationship (rather than the inverse) and so

$$C = p \cdot \alpha, \tag{1}$$

where $p$ is the calibration model parameter (constant of proportionality) with units, in this example, of $cps/(atom\ \%)$. With a single well-known calibration item with nominal value $\alpha_o$ (e.g., mass spectrometry is far more accurate than the field application of the Enrichment Meter Principle when continuum, peaked background, rate, wall, and other corrections are considered), one can write

$$p = \frac{C_o \pm u_{C_o}}{\alpha_o}, \tag{2}$$

where the subscript refers to the calibration values, and $u_{C_o}$ is the estimated uncertainty in $C_o$, usually at the notional 68.26% confidence level (1σ-value for a normal distribution with an infinite number of df). A more conventional notation would be: $\hat{p} = \frac{C_o}{\alpha_o}$ with an approximate 68% CI given by $\hat{p} \pm \frac{u_{C_o}}{\alpha_o}$. Multiple conventions and traditions are often encountered in applied measurement statistics and physical scientists often rely heavily on context to clarify meaning, so be prepared to encounter a variety of styles in the literature.

In this special case, when an unknown item is measured, the assay value has the character of a direct relative determination. The expression for approximately 68% CI for $\alpha$ is given, according to the method of Propagation of Variance (PoV) which we'll describe in detail later, by (assuming $u_c$ is the estimated uncertainty in $C$)

$$\alpha = \left(\frac{C}{C_o}\right) \cdot \alpha_o \pm \alpha \cdot \sqrt{\left(\frac{u_C}{C}\right)^2 + \left(\frac{u_{C_o}}{C_o}\right)^2}, \tag{3}$$

where uncertainty in $\alpha_o$ is being neglected for the purposes of this illustration. If any other corrections are made, the previous comment about notation applies. The assigned uncertainty becomes clear after the discussion of how to combine uncertainties. However, the fractional uncertainty on $C_o$ is required to be sufficiently small so that zero or negative values are not credible.

Most calibrations are much more involved than this simple example because the measurement procedure involves many steps, the number of model parameters is larger, and more calibration items covering the full operational dynamic range are included. In addition to a slope, the response model may also require an intercept and nonlinear behavior (hysteresis is a special case because often a calibration is checked as calibration values increase and then decrease). A calibration is usually considered valid only between the lower and upper values of the calibration items used to avoid extrapolation. Items are included between these bounds to demonstrate smooth, predictable behavior or to establish some other interpolation scheme and associated tolerances.

All measurements performed using a given instrument over a given calibration period are correlated through the common estimated calibration parameters. If the same calibration reference items are used each time, a longer-term correlation exists. In evaluating aggregate values, this correlation needs to be recognized and included in the UQ assessment by including the covariance uncertainty structure of individual results or by writing the overall measurement equation explicitly in terms of the predictor variables. This is discussed more as in Section 12, Combining Uncertainties.

## 11. MEASUREMENT CONTROL

To ensure that the measurement process is maintained within an acceptable tolerance (i.e., is within measurement control) the general methods of statistical process control first introduced by Walter A. Shewhart can be applied. The basic method monitors performance using control charts maintained through check standards that are regularly measured by the process. A word of caution here is that such checks do not capture item-specific biases but only monitor that repeatability variance is under control and stable.

Each attribute chosen for tracking will typically be charted for value, range, and SD using the so-called X-, R-, and S-charts. For radiation measurements, random error variance is often estimated by dividing the acquisition time into a sequence of shorter intervals that can be analyzed statistically. For other kinds of sensors, repeat measurements or a short run of measurements might be used to update the S-chart. Observations over an extended setting to work period establish the initial variability. These are the baseline data. Later, if fluctuations occur outside of what seems reasonable given this history, alerts can be issued. A typical criterion are the Western Electric rules (Western Electric 1956). The following articles by Brian Lanning are also clear and accessible (Lanning 1995, 1998).

## 12. COMBINING UNCERTAINTIES

The starting point is the **measurement equation**. The most common approach is to linearize the functional dependence on each of the predictor variables about their mean values and to use a result of applied statistics called **propagation of variance** (PoV).

Recall when tractable, bootstrap and Monte Carlo sampling provide intuitive alternatives to estimate PDFs without having to know much about applied theoretical statistics. Monte Carlo sampling also allows discontinuous response functions (e.g., logic involving decision trees) to be easily studied.

Before discussing the PoV, two basic idealized error models are introduced— the additive and multiplicative error models. However, in practice, a mix of measurement models is common.

### 12.1 ADDITIVE ERROR MODEL

The additive error model can be defined as

$$x_i = x_t + \varepsilon_i + b_i, \tag{1}$$

where $i$ is the index of the datum, $x_i$ being the $i^{\text{th}}$ data point of a sample, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ is an independent random variable with an expectation value of zero but finite variance, and $b_i$ is an item-specific constant (deterministic) bias also commonly called the systematic error. Both $\varepsilon_i$ and $b_i$ have the same units as $x_t$. Note on the average $\bar{x}_i \rightarrow x_t + b_i$. The value of $\varepsilon_i$ can take on any real value positive or negative, although large deviations from zero are rare (as governed by the variance $\sigma_\varepsilon^2$).

### 12.2 MULTIPLICATIVE OR PROPORTIONAL ERROR MODEL

One form of the multiplicative error model can be defined as

$$x_i = \text{"}e^{(\delta_i + c_i)} \cdot x_t\text{"} \approx (1 + \delta_i + c_i) \cdot x_t, \tag{2}$$

where $x_i$ is the $i^{\text{th}}$ item, small changes are assumed, $\delta_i \sim N\left(0, \sigma_\delta^2\right)$ is an independent normal random variable, and $c_i$ is an item-specific constant (deterministic) bias or systematic factor. In this mode, both $\delta_i$ and $c_i$ are dimensionless numbers, just simple multiplicative factors. Note that on average $\bar{x}_i \rightarrow (1 + c_i) \cdot x_t$. The natural definition of error in the multiplicative model is the ratio measured to true.

**Sidebar 4. PRACTICE**

Evaluate the variance in the measured value for these two models.

The multiplicative model can be approximately transformed (provided that the total relative error SDs are approximately 10% or less) into a linear model in terms of transformed variables as follows: $\ln(x_i) = \ln\,(x_t) + \delta_i + c_i$. The reason why we chose $e^{(\delta_i + c_i)}$ in defining the multiplicative model is now clear – it leads to a linear simplification when the natural logarithm is taken. A more general multiplicative error model, $x_i = e^{(\delta_i + c_i)} x_t^{d_i}$, also has a simple logarithm transform: $\ln(x_i) = d_i \ln(x_t) + \delta_i + c_i$. It is always instructive to review data graphically and one way to identify whether the error model is additive or multiplicative in nature is to look at the calibration results $x_i\ vs.\ x_t$ in lin-lin and ln-ln space. For the calibration data reference or accepted values take the place of $x_t$.

## 12.3 PROPAGATION OF VARIANCE

The method to combine uncertainty in the case of a well-behaved relationship is reviewed in this section. Consider the measurement equation

$$y = f(x_1, x_2), \tag{3}$$

which expresses mathematically that $y$ is a function of the two variables $x_1$ and $x_2$. The measurement equation is the mapping relationship between the observables and other information into the quantity (or quantities) of interest. In general, the algorithm can also involve logic that introduces discontinuous threads, but here simple smooth behavior is assumed.

Over some small region about the point $(\bar{x}_1, \bar{x}_2)$, assume that one can linearize the relationship in the form of a first-order Taylor series approximation. That is, use this approximation:

$$y = f(\bar{x}_1, \bar{x}_2) + \left(\frac{\partial f}{\partial x_1}\right)_{(\bar{x}_1, \bar{x}_2)} \cdot (x_1 - \bar{x}_1) + \left(\frac{\partial f}{\partial x_2}\right)_{(\bar{x}_1, \bar{x}_2)} \cdot (x_2 - \bar{x}_2), \tag{4}$$

where the subscript on the partial derivatives (gradients, slopes, or sensitivity) terms emphasizes that they are to be evaluated at the point $(\bar{x}_1, \bar{x}_2)$ where each of the variables is set to its estimated mean value.

One can also estimate the partial derivatives numerically as follows.

$$\left(\frac{\partial f}{\partial x_1}\right)_{(\bar{x}_1, \bar{x}_2)} \approx \frac{f(\bar{x}_1 + \sigma_{x_1}, \bar{x}_2) - f(\bar{x}_1 - \sigma_{x_1}, \bar{x}_2)}{2 \cdot \sigma_{x_1}}, \tag{5}$$

with a similar expression for $\left(\frac{\partial f}{\partial x_2}\right)_{(\bar{x}_1, \bar{x}_2)}$.

The question of how to form the expectation value over all possibilities of the controlling input variables is straight forward because in the linear approximation, $E[y] = f(\mu_1, \mu_2)$, and the estimator $\bar{y} = f(\bar{x}_1, \bar{x}_2)$ is defensible because $E[\bar{x}_1] = \mu_1$ and $[\bar{x}_2] = \mu_2$. Thus, the deviation becomes

$$(y - \bar{y}) = \left(\frac{\partial f}{\partial x_1}\right) \cdot (x_1 - \bar{x}_1) + \left(\frac{\partial f}{\partial x_2}\right) \cdot (x_2 - \bar{x}_2), \tag{6}$$

where the subscripts on the partial derivatives are now implied.

Now suppose that knowledge of both $\bar{x}_1$ and $\bar{x}_2$ come from a sample data of equal size $n$. Then one can square and average the deviation to obtain the variance in the measurement results to obtain the fundamental PoV result, according to the linear approximation of the measurement equation,

$$\sigma_y^2 = \left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_{x_1}^2 + 2\left(\frac{\partial f}{\partial x_1}\right)\left(\frac{\partial f}{\partial x_1}\right) cov(x_1, x_2) + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_{x_2}^2, \tag{7}$$

where sample estimates of the standard errors in the means and the standard covariance of the means (i.e., the best estimates for the uncertainty structure of the underlying population) are given by:

$$\sigma_{x_1}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2$$

$$\sigma_{x_2}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2 \tag{8}$$

$$cov(x_1, x_2) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)\,(x_{2i} - \bar{x}_2) = \frac{r}{\sigma_{x_1}\sigma_{x_2}} = cov(x_2, x_1).$$

The linear correlation coefficient $r$ lies in the interval [-1, +1] and is a convenient measure of the strength of the linear correlation between the pair of variables. Whether $r$ differs from zero (no correlation) by a statistically significant amount requires a hypothesis test which will not be discussed here but note that visualization of the data, as well understanding any causal relationship within and between the data, is extremely important.

If a pair of random variables $x_1$ and $x_2$ are truly independent, then the expected value of $cov(x_1, x_2) = 0$ because a change in one of the values means nothing to the other. Put another way, for independent random variables, $E[(x_1 - \bar{x}_1) \cdot (x_2 - \bar{x}_2)] = E[x_1 - \bar{x}_1] \cdot E[x_2 - \bar{x}_2] = 0$. There could be good reason to assume $cov(x_1, x_2) = 0$ based on physics grounds. In cases where this is not clear, it is not uncommon to simply assume there are no significant covariances without proper analysis to justify it. This is bad practice and can lead to poorly expressed and misleading confidence estimates.

Although the PoV expression was developed for the case of sampled data, it can also be applied to the case of Type B uncertainties because conceptually Type B uncertainties can also be treated as following to some probabilistic distribution.

Moreover, even though one can formally write the sensitivity coefficients $S_{x_i} = \frac{\partial f}{\partial x_i}$ as a partial derivative, it might not be easy to express the derivatives analytically. Numerical differentiation can be used instead for well-behaved functions. In other cases, judgments may come into play, and one may need to poll several experts to get a distribution of views that can be propagated. When the assay involves logic trees, then one might consider performing many forward calculations by sampling the input variables according to their known uncertainty structure and correlations to construct the PDF of results.

## 12.4 COVARIANCE

Including covariance between random variables when combining uncertainties rather than using a treatment that relies on the variables also being independent is an important concept. Of course, the emphasis is then placed on recognizing that correlation exists and on how to estimate the value of the covariance (or linear correlation coefficient). Visualization of the data is often a great help. Working in terms of the known independent variables is also often very helpful. The covariance can then be found by analysis. For instance, suppose the thickness of a container wall has been measured close in time using two different and independent techniques other than for the fact that both require a correction for temperature, $\theta$.

Then, suppose the two thickness values are written as follows:

$$l_1 = y_1(1 - a_1(\theta - \theta_1)). \tag{9}$$

$$l_2 = y_2(1 - a_2(\theta - \theta_2)). \tag{10}$$

The two values are clearly linked through the common temperature measurement, which is itself subject to measurement uncertainty. Forming the product of deviations about $\theta = \bar{\theta}$, the estimated mean value, leads to

$$dl_1 dl_2 = \frac{\partial l_1}{\partial \theta} d\theta \frac{\partial l_2}{\partial \theta} d\theta = +y_1 a_1 y_2 a_2 d\theta^2, \tag{11}$$

which, after averaging, becomes

$$cov(l_1, l_2) = y_1 a_1 y_2 a_2 \sigma_\theta^2. \tag{12}$$

Another instance where correlation is often important is when calibration-model parameters (e.g., slope and intercept) are estimated from a data set of calibration data. Because all the parameters are computed from the same calibration data set, the values are naturally covariant. Results derived using combinations of the model parameters (for instance, the ratio of the slope to the intercept) must take into account the correlation between then. Alternatively, one could formulate the problem directly, so that the intermediate results do not need to be either calculated or reported, and one could dither all of the input data consistent with the known uncertainty structure and generate the PDF of the sought-after result empirically by brute force.

Data visualization (charts, diagrams, graphs, pictograms, animations) is a powerful way to communicate data and uncertainty. To avoid being misleading, it is also useful to keep in mind what a chart is not showing and that correlation is not causation. Aggregate data may conceal other factors that are at place such as operator differences, quieter mains power supply during the night shift, a gradually failing sensor, differences spectral analysis by software tool, and so on).

## 12.5 LINEAR ALGEBRA

A more compact and convenient way to write the PoV equations for larger problems is to use the power of linear algebra. If

$$y = f(x_1, x_2, \ldots . x_n), \tag{13}$$

where $n$ now denotes the number of variables, not a sample size, the variance $\sigma_y^2$ in $y$ can be expressed in matrix notation as

$$V_y = D^T V_x D, \tag{14}$$

where $V_y$ is the covariance matrix for $y$, which for this problem is a *1* × *1* array with element $\sigma_y^2$. The *covariance matrix, $V_x$*, describes the uncertainty structure in and between the *x*-values and is given by

$$V_x = \begin{pmatrix} \sigma_{x_1}^2 & cov(x_1, x_2) & cov(x_1, x_3) & \ldots & cov(x_1, x_n) \\ cov(x_2, x_1) & \sigma_{x_2}^2 & cov(x_2, x_3) & \ldots & cov(x_2, x_n) \\ cov(x_3, x_1) & cov(x_3, x_2) & \sigma_{x_3}^2 & \ldots & cov(x_3, x_n) \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ cov(x_n, x_1) & cov(x_n, x_2) & cov(x_n, x_3) & \ldots & \sigma_{x_n}^2 \end{pmatrix}. \tag{15}$$

The covariance matrix is symmetric and of size *n* × *n*.

$D$ is the column vector of partial derivatives

$$D = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}, \tag{16}$$

and $D^T$ is its transpose.

The previous result is easily generalized to the case where a collection of *y*s (*$y_1$, $y_2$, $y_3$, .... $y_n$)* are functions of the *x*s, ($x_1$, $x_2$, .... $x_n$). This is a very common situation in practice. For example, the same calibration data set is used to determine several model (fit) parameters, or various nuclear data "constants" may be collectively evaluated from the same set of differential and integral experimental data. In this case, redefine $D$ as follows:

$$D = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}, \tag{17}$$

where the elements may be populated by algebraic differentiation or numerically by finite-difference differentiation.

The matrix expression $V_y = D^T V_x D$ now returns the $m \times m$ symmetric covariance matrix describing the covariance structure between the $m$ y-values.

## 13. THE MATERIAL BALANCE EQUATION

The material balance equation over an accounting period or interval of time, $t$, for the amount of material present in a material balance area (MBA), is a statement of the conservation of mass

$$Ending = Starting + (In - Out) - MUF\ (or\ ID), \tag{1}$$

where MUF is the so-called "material unaccounted For" amount, which is also commonly called inventory difference (ID), and (In – Out) accounts for all transfers across the MBA boundary and includes (in some instances) radioactive decay (e.g., for $^{241}$Pu).

Assuming normally distributed experimental quantities, the PDF for mass, $m$, is also approximately Gaussian

$$p(m) \cdot dm = \frac{1}{\sqrt{\pi}} \cdot exp\left(-\left[\frac{m - m_{true}}{\sqrt{2} \cdot \sigma}\right]^2\right) \cdot \frac{dm}{\sqrt{2} \cdot \sigma}. \tag{2}$$

The safeguards objective is to make a timely detection of a significant quantity (1 significant quantity or more) of missing material with a given probability while maintaining some permissible small probability of false alarms. It follows familiar logic from Lloyd Currie's work on detection and quantification limits. For an extension to Currie, see (Agboraw 2017; Kirkpatrick, Venkataraman, and Young 2013).

Rearranging and being a little more formal in notation, the ID may be expressed as follows:

$$ID = [PB + (Receipts,\ X - Removals,\ Y)] - PE = BE\text{-}PE,$$

where

*PB* and *PE* are the beginning (or opening) and ending (or closing) physical materials inventories based on locating the material and performing measurements, sampling, weighing, and analysis,

and

$BE=BI+X\text{-}Y$ is the ending (or closing) accountancy book value based on the initial book inventory accounting records of *(X-Y)*.

When $BE\text{-}PE>0$, the ID is positive, which could be interpreted as a loss of material. A large positive inventory difference outside the estimated error limits could indicate the following:

- Accidental loss of material,
- Accumulation of holdup in difficult-to-measure items of equipment,
- A process change or operational problem,
- Theft, or
- Measurement bias that is not accounted for in the known random and systematic error sources.

Both the opening book inventory and the physical inventory are based on measured values (except for verification by item counting when feasible). The limits of error can be large for complex processing facilities. Therefore, an ID within the limit of error (for example, a negative value) does not exclude the possibility of loss. Thus, the information provided by the material accounting system is also supplemented by information from the internal control system, the physical protection system, inspections, and evaluation of various kinds (including special investigations) to resolve any issues, and so on.

In the United States, all Nuclear Regulatory Commission- (NRC-) licensed fuel facilities authorized to possess more than one effective kilogram of special nuclear material fall under the NRC's graded approach to safeguards and pursuant to 10 Code of Federal Regulations (CFR) 74.17. Operators must report the results of each physical inventory to the NRC. The frequency of physical inventory (6, 9, or 12 months) depends on the strategic significance (type and amounts) of material.

One measure of material balance closure over the period is determining whether the ID is consistent with zero within three times the standard error of inventory difference (SEID), "3σ." If it is, and the ID does not exceed the facility/site specific regulatory (license) mass limits, no compelling reason exists to think a diversion has taken place or that an investigation is called for.

The SEID is used to describe total SD (random plus systematic for all assay methods) associated with an ID value. It is the nominal 68% (1-σ) confidence level. From the one-sided normal probability table (one sided if the statistical test is for loss only, not for gain), an ID corresponding to a mass-loss of approximately $(Q - 1.3 \cdot SEID)$ would be detected with 90% probability. This can be seen by centering the measurement distribution on $Q$ and stepping back.

However, SEID is defined by US Department of Energy rules, not measurement science, as follows:

For Category III licensees subject to 10 CFR 74.31 or 74.33, the SEID is defined to be equal to quadrature sum of both the measurement and nonmeasurement variances associated with an ID, i.e., $SEID = \sqrt{var(meas) + var(non - meas)}$.

For both Category I licensees subject to 10 CFR 74.59 and Category II licensees subject to 10 CFR 10 74.43, the SEID is defined to be equal to the square root of the measurement variance (only) associated with an ID, i.e., $SEID = \sqrt{var(meas)}$.

In some instances, some parts of the inventory may not have changed so that the exact same inventory value gets used in both the beginning and closing values, such as a piece of equipment or an item that has remained intact and unused throughout the period. In such cases, defining a new quantity is useful. This quantity is the active inventory $(AI)$, which is a measure of throughput and the only part of the current inventory that is subject to new measurement uncertainty. Certain regulatory limit of error for inventory differences (LEID*)* are expressed in terms of the $AI$. For example, in criteria such as (US Department of Energy 2003), the ID cannot exceed 2% of the $AI$ ("throughput") up to 2 kg with 90% confidence. For complex and high-throughput facilities maintaining 2% accuracy, though physical measurements alone, are usually extremely challenging or not practical. In this example, meeting the 2 kg quantity becomes the goal and might drive the overall accountancy strategy, which may include:

- Process optimization and control and use of near-real-time monitoring,
- Emphasis on high-accuracy instrument selection, calibration, and acceptable knowledge,
- Definition of MBA boundaries, key measurement points, and the role of subMBAs, and
- Frequency of material balance closure so that amounts are kept small.

Because taking a facility down to perform wall-to-wall physical inventory is both time consuming and costly, designing the measurement strategy to be fully compliant in an efficient way should receive appropriate attention from the onset.

Often the NDA measurement program may support several needs, including operational, safety (criticality), materials control and accountancy, and waste management. The performance, uncertainty targets and reporting requirements of each consumer needs to be considered because retrofitting a solution can often be expensive and present a variety of issues.

## 14. TOP-DOWN VS. BOTTOM-UP UNCERTAINTY QUANTIFICATION FOR NONDESTRUCTIVE ASSAY IN SUPPORT OF THE MATERIAL BALANCE EQUATION

Recall that the bottom-up approach propagates or combines error variances from all identified sources of measurement variation. In contrast, the alternative top-down approach does not concern itself with creating a complete uncertainty budget by each contributing factor. Instead, the precision and accuracy (inverse of the random error SD and the systematic error SD, respectively) are evaluated by comparing against known values (e.g., performance demonstration plan items), or against other reference methods, or by using round-robin comparisons which represent independent experiments. The emphasis is on the analysis of paired data (measured—assigned true), and the overall uncertainty is evaluated by statistical methods by looking at the empirically observed scatter. The top-down approach quantifies performance but without the insights provided by the bottom-up analysis. Typically, the top-down uncertainty exceeds the bottom-up uncertainty, suggesting that the bottom-up approach may be incomplete or biased low (overly optimistic). The difference is referred to as **dark uncertainty** because it is hidden or unrecognized, sometimes simply because fielded NDA methods have error sources that are not accounted for in NDA laboratory bottom-up evaluations.

### Sidebar 5. PRACTICE

Sketch these two basic approaches to performance assessment and performance demonstration.

The International Atomic Energy Agency's (IAEA's) material balance equation uses the Inspector's

$$MUF \ = Facility\ MUF - D, \tag{1}$$

where *D* is a difference statistic between paired operator and inspector measurements. Better bottom-up UQ for NDA is needed in support of the material balance equation and to identify and manage dark uncertainty. Dark uncertainty is more commonly positive, which suggests that something has been overlooked in the bottom-up analysis or that the measurement process is not as well understood as believed. An experimenter is perhaps understandably proud of their technique and confident in their abilities to be overly optimistic and perhaps is not be aware of all uncertainty sources in fielded instruments.

It is important to remember in brainstorming the set of things that can influence the result that some of the most important variables may not be simple physical quantities, such as cross sections, energy spectra, mass compositions, geometry, and the like. Rather, variables can be implicit, such as assumptions and analysis procedures that may seem natural and obvious but should be challenged, nonetheless. For example, the add-a-source correction for passive neutron coincidence counting of drummed waste requires a volume-weighted-average response. However, typically experimental calibration data are available only on a crude spatial grid. This may point to a planning weakness.

For high moderator content, the mathematical procedure for how the volume-weighted-average response is defined (e.g., fit and integrate, create iso-contours and sum by ring, define volume elements around each point, the data into a tool such as AutoCAD and use splines and the built-in analysis tool) can have a large impact—50% or more relative difference between definitions. Some of the most critical dependencies may have to do with things that cannot be easily changed. For instance, the detector may exist already and not optimised to the current task, source tailoring to reduce $^{238}$U response in active neutron interrogation systems may be difficult to account for, wholly objective assessments might not be possible about the accuracy of simulation libraries used for calibration, and so forth. Many dependencies are intertwined even though they are often treated then as separate. For instance, if the source distribution is shifted, the counting precision and rate loss corrections will change even though a sensitivity analysis assumes the geometrical change in the detection efficiency is the main aspect of the measurement that was changed.

## CHAPTER 1 SUMMARY

- A scientific measurement or calculation is incomplete without a statement, supported by additional information, about the associated estimated uncertainty.

- “Uncertainty” is a useful qualitative term that often means the SD of a measurement error component. Sometimes “uncertainty” or “expanded uncertainty” refers to a CI width of half-width, which is often a multiple of the error SD.

- We have seen how a variety of nonparametric and parametric statistical methods can be used to organize and summarize sample data and to make inferences.

- Special mention was made of the Poisson distribution in connection to nuclear counting. It holds a special place in radiometric, and of the Central Limit Theorem, which is commonly invoked to justify using the normal distribution to describe a variety measurement uncertainties.

- Statistical methods help optimize and select measurement procedures to meet a given task and to maintain the measurement program within control.

- Different ways to assess uncertainty contributions exist. The top-down and bottom-up approaches were introduced.

- There are different ways to combine various uncertainty contributions to form the total measurement uncertainty. This was illustrated by taking a worked example that used the PoV. We emphasized the importance of providing the consumer of the analysis with a sufficient understanding of the measurement process, the measurement equation (or algorithm), and the mathematical techniques of statistical analysis applied. General classes of measurement error models (additive, multiplicative, mixed), were discussed. The example of the material balance equation was provided, and the chapter concluded with the example of a density measurement.

- Because the intended audience for this book is measurement experts, the chapter emphasized bottom-up UQ. Approximately every 10 years, the IAEA publishes relative standard deviation estimates for many measurement methods commonly used in nuclear safeguards (Zhao 2010); these relative standard deviation estimates are used at the IAEA to estimate the SEID and to design sampling plans to detect data falsification. Many of the published relative standard deviation estimates are based on top-down UQ, using specialized analysis of variance. Bonner et al. (2016) and Burr (Burr, Croft, et al. 2016) provide further discussion on how statistical methods are used to verify nuclear material inventories (Bonner 2016; Burr, Croft, et al. 2016; Burr, Krieger, et al. 2016). The American Society for Testing and Materials maintains a number of useful standards and guides in NDA instruments and methods and the National Institute of Standards and Technology has a very good on-line handbook on statistical methods (National Institute of Standards and Technology 2012).

## EXERCISES

Solutions for Exercises 3, 4, and 6 are included at the end of this chapter.

### Exercise 1

Consider the following contrived example in which the number of counts in channel two is twice that of channel one. Consider the linear combination $y = 2x_1 + 3x_2$ and, for instance, if the units of $x_1$ and $x_2$ are units of activity $y$ that might be intended to be a measure of radiation damage.

1. Using the data in Table 3, calculate the mean, variance of $x_1$ and $x_2$, and the covariance and linear correlation coefficient between them. Plot $x_2$ against $x_1$ to get a visual sense of whether the correlation is meaningful.

2. Combine the uncertainties by PoV. What is the effect of neglecting covariance?

3. Show that the combined uncertainty (in this case) is the same as in using only the independent variable $x_1$ and writing $y = 8x_1$.

**Table 3. Numerical data for use in the PoV example.**

| Reading | $x_1$ [Bq] | $x_2$ [Bq] |
|---|---|---|
| 1 | 10 | 20 |
| 2 | 9 | 18 |
| 3 | 11 | 22 |
| 4 | 12 | 24 |
| 5 | 8 | 16 |

In this case, five $y$-values could be generated from the paired data and the results computed directly. However, in general, this leads to combining uncertainties for which no simple table exists. Yet the idea of generating a distribution of $y$-values by Monte Carlo sampling of all the input variables from distribution (including bootstrapping of finite samples) can be an attractive alternative way of evaluating the overall uncertainty.

### Exercise 2

Letting $y = f(x)$, find the relative SD $\frac{\sigma_y}{y}$ for $\frac{\sigma_x}{x} = 0.01$ when: (1) $f(x) = ax^{-1}$; (2) $f(x) = ax^{-1/2}$; (3) $f(x) = ax^0$; (4) $f(x) = ax^{1/2}$; (5) $f(x) = ax^1$; (6) $f(x) = ax^{3/2}$; (7) $f(x) = ax^2$; (8) $f(x) = ln(ax)$.

### Exercise 3

Let $f = x - a$ and $g = y - b$. Find $cov(f, g)$ given $a$ and $b$ are simple constants and $x$ and $y$ are measured values with a finite covariance.

## Exercise 4

If the correction factor $\theta = \frac{\left(1+\frac{v_d}{v_1}\right)}{\left(1+r\frac{v_d}{v_1}\right)^2}$, what is the fraction standard uncertainty $\frac{\sigma_f}{f}$ in $f$ due to the fractional standard uncertainty in $v_d$?

Hint: For the function $f(x) = \frac{T(x)}{L^2(x)}$, show that the derivative $f'$ of $f(x)$ with respect to $x$ is $f' = \left(\frac{T\prime}{T} - 2\frac{L\prime}{L}\right)$.

## Exercise 5

Let $y_1 = \frac{x_1}{x_1+x_2}$ and $y_2 = \frac{x_2}{x_1+x_2}$ with $\sigma_{x_1}^2 = 1$, $\sigma_{x_2}^2 = 1$, $cov(x_1, x_2) = 0$.

Show $y_1 = 0.4 \pm 0.072$, $y_2 = 0.6 \pm 0.072$, and $cov(y_1, y_2) = -0.0052$.

Define a new relationship, z=$y_1 + y_2$. Using $V_z = D^T V_y D$, show $z = 1 \pm 0$, which is correct because by definition $y_1 + y_2 = \frac{x_1+x_2}{x_1+x_2}$. (See Section 12).

## Exercise 6

Let $y_1 = x$ and $y_2 = x^2$; $y_1$ and $y_2$ are clearly correlated. For $z = y_2/y_1$, show $\sigma_z^2 = \sigma_x^2$.

## AUTHOR BIOS

Stephen Croft
PhD Nuclear Physics, MSc Reactor Technology, BSc Physics
Nondestructive assay measurement techniques, algorithms and in situ applications

Dr. Croft has diverse experience in basic and applied radiometric science in both the private and public sectors (international and domestic). He has worked extensively on security (for example, burst detection), waste assay (including for geological repositories), and international safeguards and material accounting. Stephen and his physics team at Canberra designed and calibrated numerous systems for European Atomic Energy Community (Euratom), the International Atomic Energy Agency, and multiple facilities. He is a fellow of both the Institute of Physics and the Institute of Nuclear Material Management (INMM). He chaired INMM's Nondestructive Assay Working Group, and he is active in the American Society for Testing and Materials C26 and the Euratom Nondestructive Assay user group. His research interests include radiation metrology, neutron correlation counting, statistical methods for measurement science, and active interrogation.

Tom L. Burr
PhD Statistics, BS Mathematics and Physics
Uncertainty quantification for nuclear material assay

Dr Burr received BS degrees in mathematics and physics and a Ph.D. in statistics. Since 1992, Dr. Burr has worked as a statistician at Los Alamos National Laboratory, with a four-year change of station to the International Atomic Energy Agency to work on uncertainty quantification for measurements. His research interests include model uncertainty, multivariate time series, multivariate calibration, classical and molecular epidemiology, and exploratory data analysis methods including clustering and classification. Nonproliferation applications include remote monitoring, uncertainty quantification for nuclear material assay, process monitoring, image analysis, multivariate sequential analysis, pattern recognition, and nuclear materials accounting. Tom has coauthored over 250 technical reports, including more than 100 conference proceedings and 185 refereed articles.

# REFERENCES

Abramowitz, M., Stegun, I.A. 1968. *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (National Bureau of Standards Applied Mathematics: Washington, D.C.).

Agboraw, E., Bonner, E., Burr, T., Santi, P., Wals, S., Norman, C., Croft, S., Kirkpatrick, J. M., Krieger, T. 2017. 'Revisiting Currie's Minimum Detectable Activity for Non-Destructive Assay By Gamma Detection Using Tolerance Intervals', *ESARDA Bulletin*: 14-22.

Agresti, A., Coull, B. 1998. 'Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions', *American Statistician*, 52: 119-26.

Bonner, E., Burr, T., Krieger, and Norman, C. 2016. "Statistical Issues in Nuclear Safeguards." In *Encyclopedia of Statistical Science*. Wiley.

Burr, T., S. Croft, K. Jarman, A. Nicholson, C. Norman, and S. Walsh. 2016. 'Improved uncertainty quantification in nondestructive assay for nonproliferation', *Chemometrics and Intelligent Laboratory Systems*, 159: 164-73.

Burr, Tom, Thomas Krieger, Claude Norman, and Ke Zhao. 2016. 'The impact of metrology study sample size on uncertainty in IAEA safeguards calculations', *Epj Nuclear Sciences & Technologies*, 2.

Chadwick, M. B., M. Herman, P. Oblozinsky, M. E. Dunn, Y. Danon, A. C. Kahler, D. L. Smith, B. Pritychenko, G. Arbanas, R. Arcilla, R. Brewer, D. A. Brown, R. Capote, A. D. Carlson, Y. S. Cho, H. Derrien, K. Guber, G. M. Hale, S. Hoblit, S. Holloway, T. D. Johnson, T. Kawano, B. C. Kiedrowski, H. Kim, S. Kunieda, N. M. Larson, L. Leal, J. P. Lestone, R. C. Little, E. A. McCutchan, R. E. MacFarlane, M. MacInnes, C. M. Mattoon, R. D. McKnight, S. F. Mughabghab, G. P. A. Nobre, G. Palmiotti, A. Palumbo, M. T. Pigni, V. G. Pronyaev, R. O. Sayer, A. A. Sonzogni, N. C. Summers, P. Talou, I. J. Thompson, A. Trkov, R. L. Vogt, S. C. van der Marck, A. Wallner, M. C. White, D. Wiarda, and P. C. Young. 2011. 'ENDF/B-VII.1 Nuclear Data for Science and Technology: Cross Sections, Covariances, Fission Product Yields and Decay Data', *Nuclear Data Sheets*, 112: 2887-996.

Chunovkina, A., and A. Chursin. 2001. '"Guide to the Expression of Uncertainty in Measurement" (GUM) and "mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes" (MRA): Some problems of data processing and measurement uncertainty evaluation', *Advanced Mathematical and Computational Tools in Metrology V*, 57: 55-66.

Croft, S., and T. Burr. 2016. "A Brief Re-Introduction to Measurement Science Statistics." In. unpublished

Dean, V.F. (ed). 2008. "ICSBEP Guide to the Expression of Uncertainties." In.: Nuclear Energy Agency.

Kirkpatrick, J. M., R. Venkataraman, and B. M. Young. 2013. 'Minimum detectable activity, systematic uncertainties, and the ISO 11929 standard', *Journal of Radioanalytical and Nuclear Chemistry*, 296: 1005-10.

Lanning, B.M. 1995. "Making SPC easier with zone control charts." In *Institute for Nuclear Material Management*, 1027-31. Palm Desert, CA: Journal of Nuclear Material Management.

———. 1998. 'A Computer Program for Real-time Statistical Process Control Using Zone Control Charts', *Journal of Nuclear Material Management*, Fall: 32-35.

Miller, Rupert G. 1974. 'The jackknife-a review', *Biometrika*, 61: 1-15.

National Institute of Standards and Technology. 2012. "NIST/SEMATECH e-handbook of statistical methods." In.: US Department of Commerce.

Smith, D.L. 1991. *Probability, Statistics, and Data Uncertainties in Nuclear Science and Technology* (American Nuclear Society).

Smith, R.C. 2013. *Uncertainty Quantification: Theory, Implementation, and Applications* (SIAM).

Taylor, J.R. 1997. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (University Science Books).

Triola, M.F. 2017. *Elementary Statistics* (Pearson).

US Department of Energy. 2003. "Manual for Control and Accountability of Nuclear Materials." In.

Western Electric, Company. 1956. *Statistical quality control handbook* (Western Electric Co.: Indianapolis).

Yule, G. Udny. 1899. 'An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.)', *Journal of the Royal Statistical Society*, 62: 249-95.

Zhao, K., Penkin, M., Norman, C., Balsley, S., Mayer, K., Peerani, P., Pietri, C., Tapodi, S., Tsutaki, Y., Boella, M., Renha, Jr, G., Kuhn, E. 2010. "International Target Values 2010 for Measurement Uncertainties in Safeguarding Nuclear Materials." In, Medium: X; Size: 44 page(s). Vienna, Austria: International Atomic Energy Agency

## SOLUTIONS

### Exercise 3 Solution

$cov(f,g) = \langle df dg \rangle = \langle dx dy \rangle = cov(x,y)$. Note that we have used the shorthand $df = f - \bar{f} = (x-a) - (\bar{x} - a) = x - \bar{x}$, $dg = y - \bar{y}$ and $cov(f,g) = \langle (x-\bar{x})(y-\bar{y}) \rangle$. This also yields the useful sample result for paired data: $cov(x,y) = \overline{xy} - \bar{x}\bar{y}$. For the special case $x = y$, this reduces to $cov(x,x) = var(x) = \left(\overline{x^2} - \bar{x}^2\right)$.

### Discussion

Notably, because the PoV method linearizes the relationship between predictor and response and requires the function to be well behaved over the region of interest, some of the results obtained by the mechanical application of the PoV formula may require additional scrutiny for validity. For example, consider the case where the measurand $y$ is obtained from $\frac{1}{x}$, where $x$ is a random variable. Suppose $x$ is distributed according to a flat (uniform or rectangular) distribution between the limits $a$ and $b$, where $b > a > 0$. The $x$-distribution is symmetric about the mean $\mu_x = \frac{b+a}{2}$ and has a SD $\sigma_x = \frac{b-a}{\sqrt{12}}$. The $y$-distribution can be obtained by invoking 1:1 correspondence between $y$ and $x$, and hence incremental probabilities, i.e., $p(y)dy = p(x)dx$, which results in $p(y) = -\frac{1}{b-a} \cdot \frac{1}{y^2}$ between the lower limit $\frac{1}{b}$ and the upper limit $\frac{1}{a}$. As seen immediately, the distribution is not symmetric, nor is it centered on $\frac{1}{\mu_x}$. In fact, the expectation value of $y$ is $\mu_y = \frac{1}{b-a}\int_a^b \frac{dx}{x} = \frac{1}{b-a} \cdot \ln\left(\frac{b}{a}\right) = ln\left(\frac{\frac{b+a}{2}+\frac{b-a}{2}}{\frac{b+a}{2}-\frac{b-a}{2}}\right)$, which can be shown to tend to $\frac{1}{\left(\frac{b+a}{2}\right)} = \frac{1}{\mu_x}$ in the limit $\frac{\frac{b-a}{2}}{\frac{b+a}{2}} = \sqrt{3}\frac{\sigma_x}{\mu_x} \to 0$, i.e., when the $x$-distribution is narrow. As an exercise, show under what conditions $\frac{\sigma_y}{\mu_y} \to \frac{\sigma_x}{\mu_x}$.

### Comment

1/X does not have any finite moments if *X* ~ uniform on (0,1) (or even if *X* ~ normal))
Turning now to the case where the $x$-distribution is normal, it is again found that the $y$-distribution is not normal and that the uncertainty propagation is inherently nonlinear. In this case, we must also confront the theoretical possibility that $x$ can be arbitrarily close to zero so that $\frac{1}{x}$ can become extremely large and, in fact, $\frac{1}{x}$ has a Cauchy distribution and so all moments are infinite. If $|\mu_x| \geq 3\sigma_x$, then as a practical matter $\frac{1}{x}$ can be truncated so that all moments are finite, and the truncation has any effect with less than 1% relative frequency. However, if $|\mu_x| \leq 3\sigma_x$, then truncation might not be acceptable, and the variance is infinite As a rule of thumb, if $\left|\frac{\sigma_x}{\mu_x}\right| < 0.1$, the 68.3% CI will be well approximated by PoV with about 10% or better.

The takeaway message is that when applying PoV to reciprocal quantities, always check whether PoV leads to a valid approximation for the task or use an alternative method to estimate CIs, such as Monte Carlo sampling. However, even Monte Carlo sampling might be misleading, such as in the case where *X* has a normal distribution with large or moderate relative SD as just described.

### Exercise 4 Solution

$$\left|\frac{\frac{v_d}{v_1}}{\left(1+\frac{v_d}{v_1}\right)} - 2\frac{r\frac{v_d}{v_1}}{\left(1+r\frac{v_d}{v_1}\right)}\right|\left(\frac{\sigma_{v_d}}{v_d}\right).$$

### Exercise 6 Solution

Ignoring covariance, the result is $\sigma_z^2 = 5\sigma_x^2$.