An Elementary Introduction to Some Issues in Metrology and Some Implications of (and for) Probability and Statistics

Steve Vardeman

RTG/VIGRE

9/15/06, 9/22/06, and 9/29/06

- Metrology is the science of measurement

- If you can't measure, you *can't* collect data ... if your measurements are lousy, *you might as well not* collect data

- "Fun" statistics (like DOX, SPC, etc.) depends upon metrology

- Conversely, statistics can and does contribute to effective metrology ... (consider, for example, the presence of the SED Division at NIST)

What are some basic terms and fundamental concerns in metrology?

- A **measurand**, x, is a quantity for which one hopes to establish a correct or true value

- An axiom of life: measurement is never perfect, there is always **measurement error**, one doesn't see x but rather a **measurement**, y

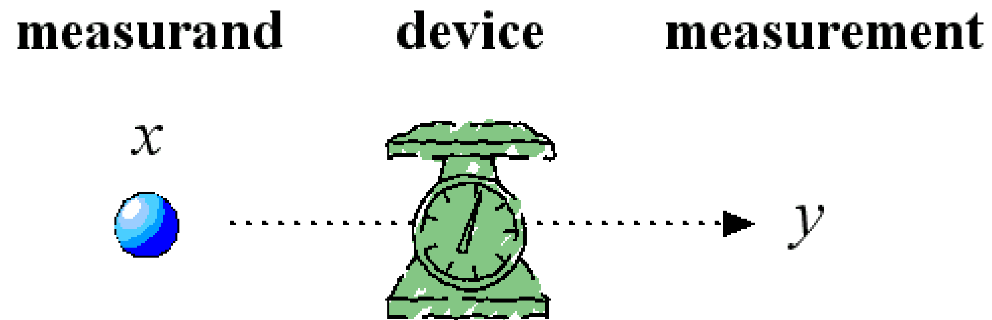

Figure 1: Measurement Cartoon

- Basic concerns in measurement are:
  - **Validity**
  - **Precision** (related to variance)
  - **Accuracy** (related to "bias")

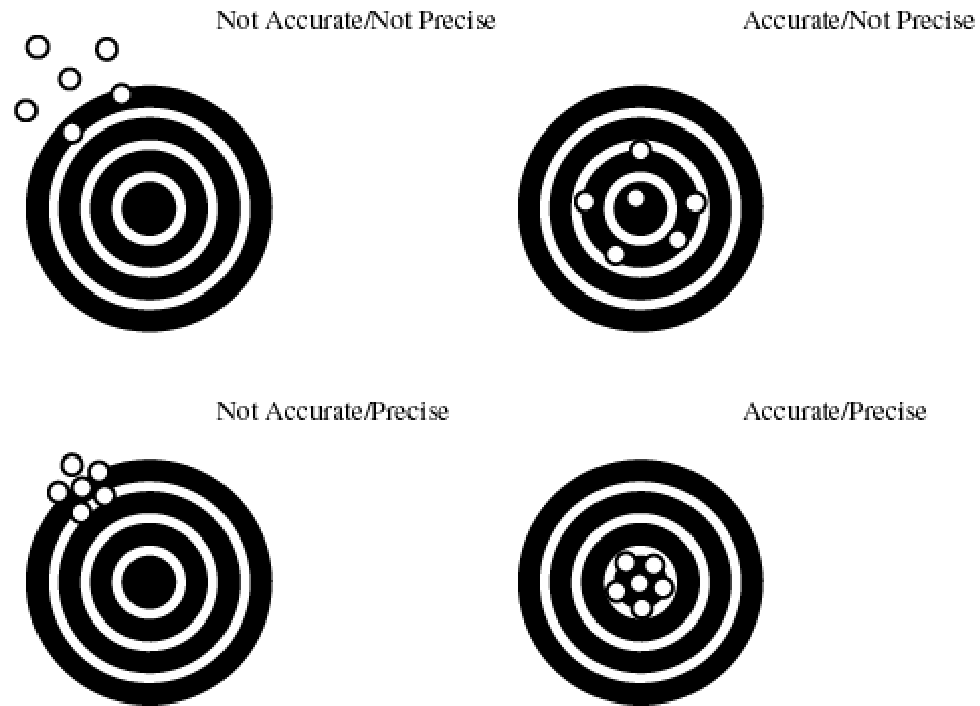

Figure 2: Measurement/Target Shooting Analogy

- Accuracy can only be assessed by comparison of observed y to a (mostly ;+}) "known" x ... this requires **standards** (hence the existence of NIST and related organizations)

- The business of measuring standards and inventing "corrections" to apply to observed y in an attempt to improve accuracy is called **calibration**

- Improvement of measurement precision typically involves

  - standardization of procedures for using particular measurement equipment

  - evolution of the sophistication of both the physical principles employed in equipment, and the manufacturing techniques used in the production of equipment of a given design

- It is often essential in both science and commerce to be able to assign an "**uncertainty**" to a measurement ... but exactly what that would mean is not altogether clear

The simplest possible probabilistic modeling of measurement error

Our knee-jerk (and often not too unreasonable) reaction as modelers/statisticians is to begin analysis by writing

$$\epsilon = y - x$$

or

$$y = x + \epsilon$$

and think, "Wouldn't it be nice if we could think of ϵ as normal ... say with mean β (for bias) and standard deviation σ_{device}?"

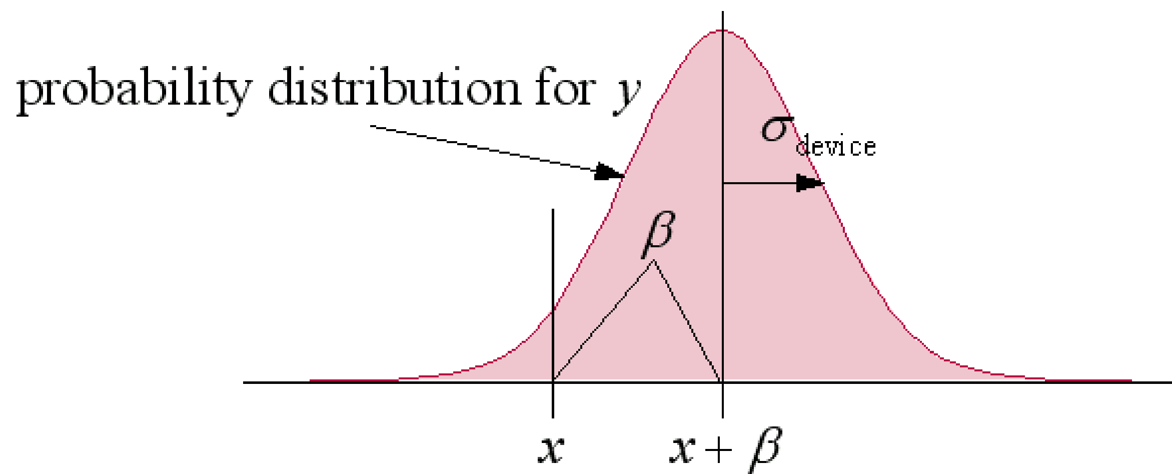

Figure 3: A Simple Measurement Error Model

There are some obvious questions about real-world interpretation of the mathematical elements of this model, and then specifics of the technical details of the model:

- Exactly what kinds of variation are described by this model? (e.g. does "device" mean a particular combination of physical equipment, protocol, and human user? ... or is it intended to cover some variation in these?)

- Why normal?

- Should bias depend upon the measurand? Should it be $\beta(x)$ instead of β?

- Should measurement precision depend upon measurand? Should it be $\sigma_{\text{device}}(x)$ instead of σ_{device}?

The last two of these are obviously important because one uses a device to try to evaluate different measurands. It is worth noting that metrologists seem to call devices with constant bias over the range of measurands a device will be used to evaluate "**linear**." (A better terminology might be "linear with slope 1," but the former language seems standard.) This is illustrated in the next figure, where we assume that the horizontal and vertical scales are the same.

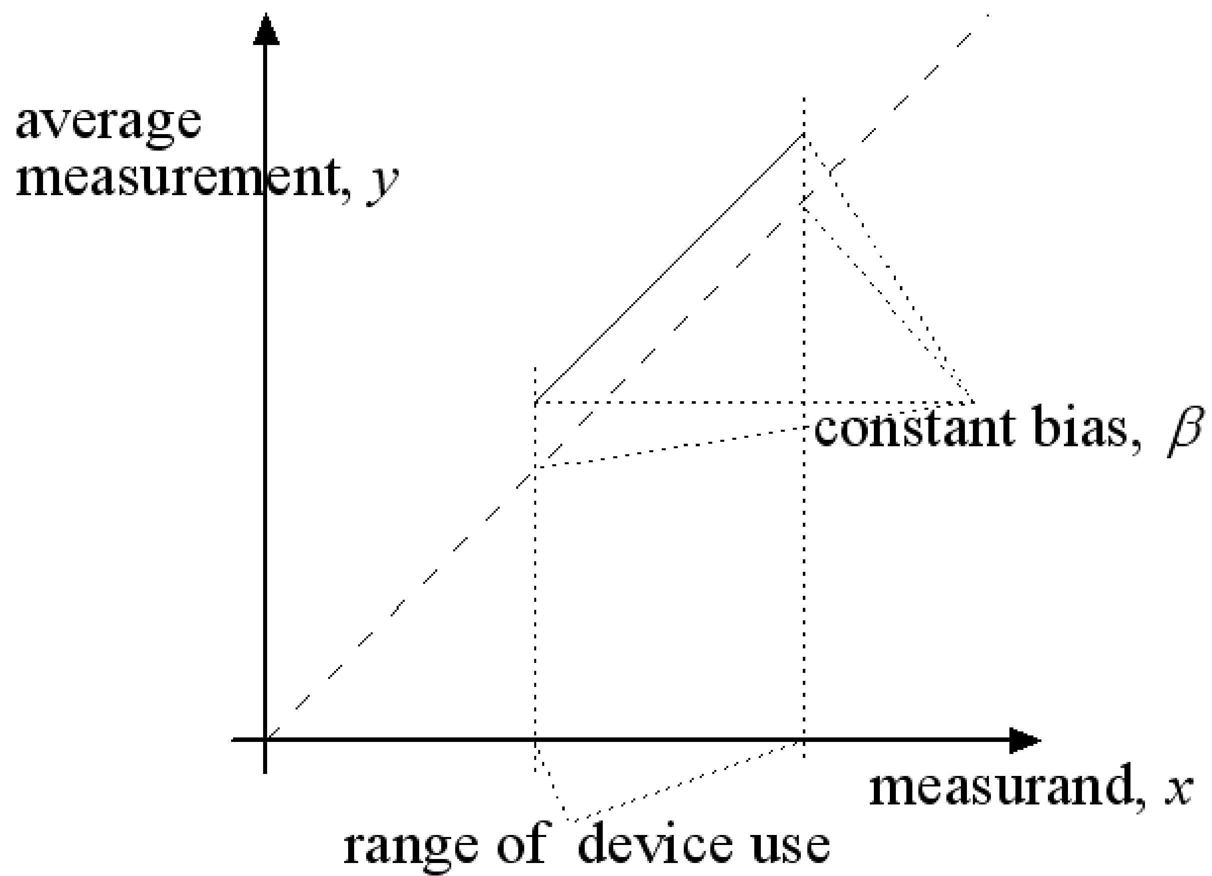

Figure 4: A Cartoon Illustrating Device *Linearity*

While there are indeed questions raised by these simplest of measurement model assumptions, even this simple model is instructive and really calls into question the interpretation of many of the simple inferences we teach beginning statistics students.

That is, we teach students that when sampling from a single normal (μ, σ) distribution, t confidence limits can be made for μ and χ^2 confidence limits can be made for σ. **But what, exactly, are μ and σ?** Consider a pair of situations. First, consider repeat observations on a single measurand, that we might perhaps model as iid observations. As illustrated in the next figure,

$$\text{"}\mu\text{"} = x + \beta \quad \text{and} \quad \text{"}\sigma\text{"} = \sigma_{\text{device}}$$

and it is *these* that can be estimated in elementary fashion.

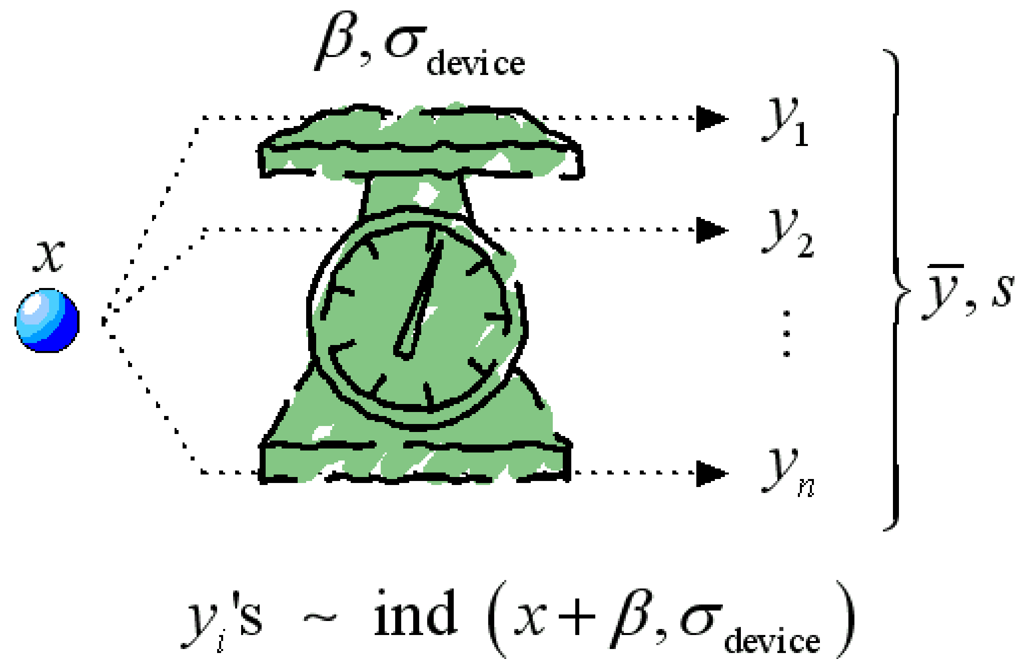

Figure 5: Multiple Measurements on a Single Measurand

Contrast this with a situation where there are multiple measurands "drawn at random from a normal population" or "generated by a stable normal process" with mean μ_x and standard deviation σ_x. If one assumes that measurement errors are (probabilistically) independent of these random measurands *and device bias and standard deviation are both constant*, the next figure is relevant, and it is obvious that what can be estimated in simple fashion are

$$"\mu" = \mu_x + \beta \text{ and } "\sigma" = \sqrt{\sigma_x^2 + \sigma_{\text{device}}^2}$$

Contrary to what even the very best of our students leaving elementary statistics courses believe, it is only when $\beta \approx 0$ and $\sigma_{\text{device}} \approx 0$ that elementary formulas really give inferences for "population parameters" (μ_x and σ_x).

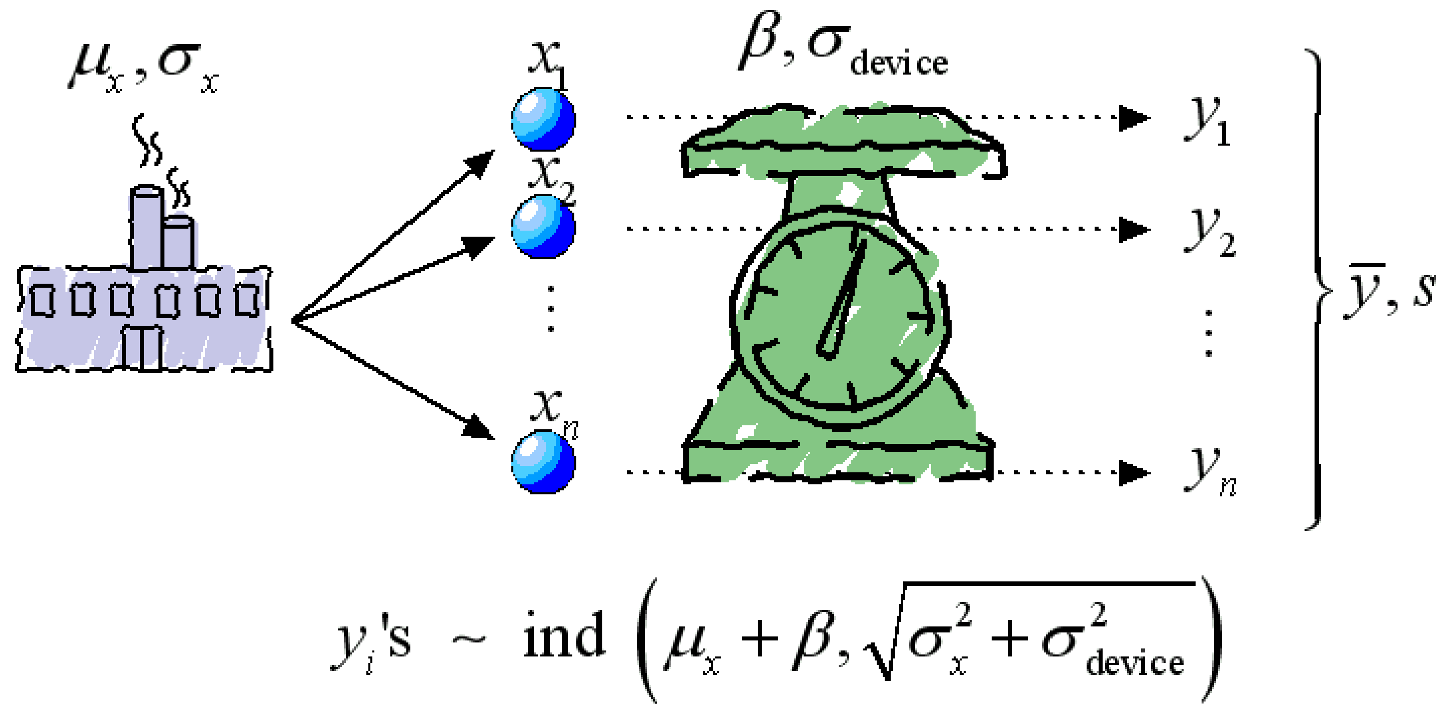

Figure 6: Cartoon Illustrating Single Measurements on Multiple Items From a Stable Process

Some fairly "straightforward" contributions of probability and statistics to metrology

Fairly simple normal theory statistical methods have been used to great advantage in assessing and improving the quality of a variety of physical measurements. Among these are:

- Random effects models and variance component estimation methods (in assessing and pointing out important sources of measurement imprecision)

- Regression models and analysis (in analyzing calibration studies and providing inferences about a future measurand x_{new} based on a future measurement y_{new})

- Tests of equality of variances and of equality of means in so-called "inter-laboratory tests"/"round robin tests" where a standard is sent to multiple analytical labs with the hope of comparing their precisions and biases

- "SPC" tools in the ongoing monitoring of performance of measurement devices

An example of the use of random effects models is in the common industrial "Gauge R&R" studies, where a single gauge/piece of equipment is used by J different operators to measure I different parts m times apiece. This results in a classical $I \times J$ factorial data set where the common cell sample size is m. Then for

$$y_{ijk} = \text{measurement } k \text{ on part } i \text{ by operator } j$$

it is common to use the two-way random effects model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

where the $\alpha_i, \beta_j, \alpha\beta_{ij}$, and ϵ_{ijk} are all independent mean 0 random variables, the α_i with variance σ^2_α, the β_j with variance σ^2_β, the $\alpha\beta_{ij}$ with variance $\sigma^2_{\alpha\beta}$, and the ϵ_{ijk} with variance σ^2.

Of central concern in a Gauge R&R context is the problem of separating "**repeatability**" variation in measurement (that variation that is characteristic of the same operator remeasuring the same part) from "**reproducibility**" variation in measurement (that variation that is chargeable to differences between operators). Different relative sizes for these have different implications for how one might attempt to reduce measurement variability.

In the two-way model, the parameter σ is a repeatability standard deviation and so standard methods are directly relevant to estimating

$$\sigma_{\text{repeatability}} = \sigma$$

On the other hand, the most natural interpretation of an operator-to-operator standard deviation is

$$\sigma_{\text{reproducibility}} = \sqrt{\sigma_{\beta}^{2} + \sigma_{\alpha\beta}^{2}}$$

(a long run standard deviation in operator biases) and of further potential interest is

$$\sigma_{\text{R\&R}} = \sqrt{\sigma_{\text{repeatability}}^{2} + \sigma_{\text{reproducibility}}^{2}}$$

(a long run standard deviation of many operators making a single measurement on the same part). These last two standard deviations are parametric functions that are not commonly estimated in other contexts where the two-way model is

employed. But substantial effort has been invested in finding reliable estimation methods for them.

The next figure shows a caliper and some "gauge blocks" (standards that can be used in a calibration study).

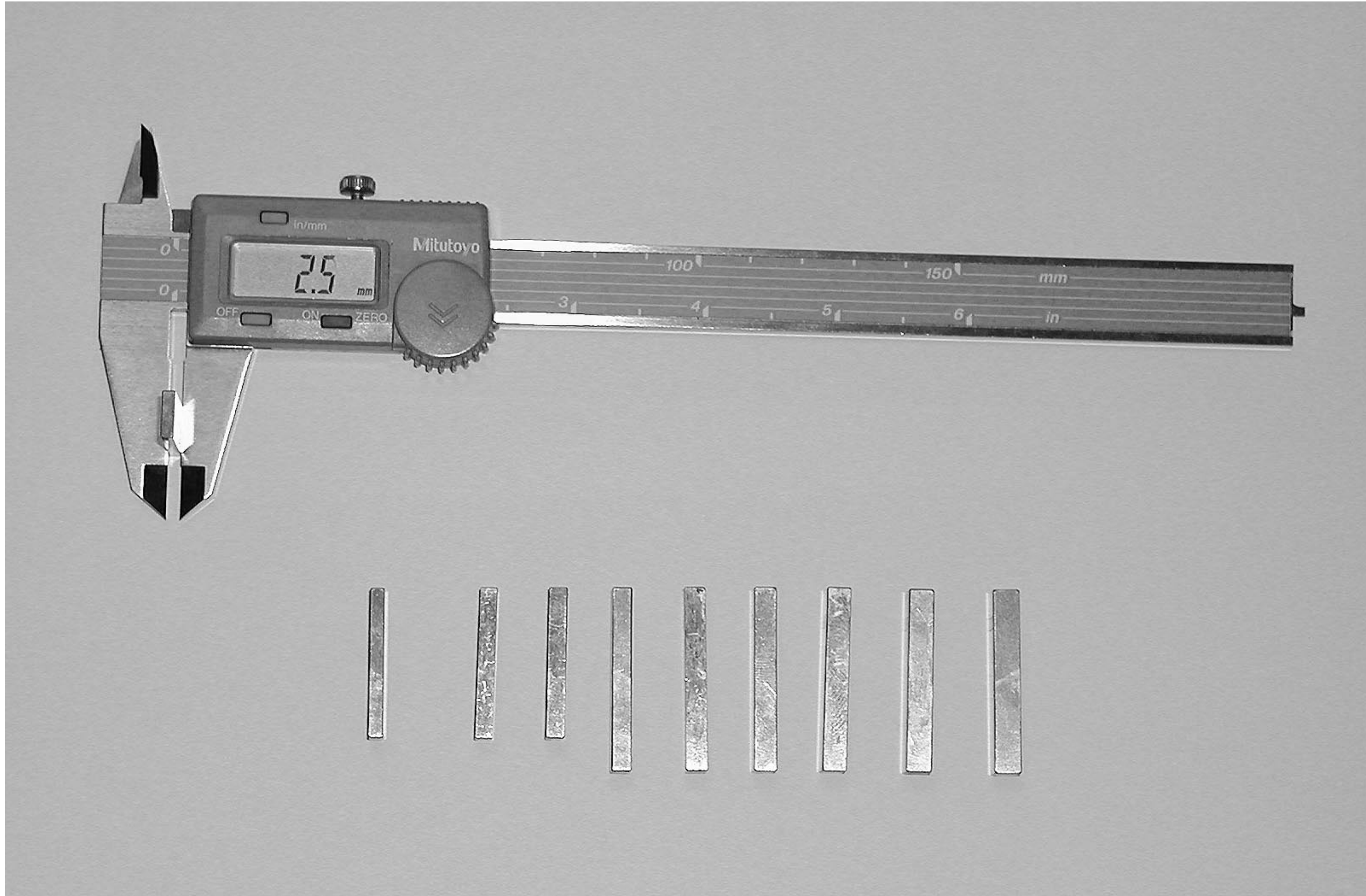

Figure 7: A Digital Caliper and Some "Gauge Blocks"

The primary statistical methodology that contributes to the practice of metrological calibration is regression analysis. Suppose, for example, that device standard deviation does not change with the measurand, but that bias does change with x. Then the simple model for measurement is that

$$\begin{aligned} y &= x + \epsilon \\ &= x + \beta(x) + (\epsilon - \beta(x)) \\ &= x + \beta(x) + \epsilon^* \end{aligned}$$

where ϵ^* is normal with mean 0 and standard deviation σ_{device}. If

$$\beta(x) = \beta_0 + \beta_1 x$$

we then have

$$\begin{aligned} y &= \beta_0 + (1 + \beta_1) x + \epsilon^* \\ &= \beta_0 + \beta_1^* x + \epsilon^* \end{aligned}$$

the usual simple linear regression model with slope $\beta_1^* = 1 + \beta_1$. Standard methods of inference in this model based on calibration data provide estimates of σ_{device} and tests of H_0:$\beta_1^* = 1$ (the hypothesis that the device is "linear"). Further, if the least squares equation is

$$\hat{y} = b_0 + b_1^* x$$

then solving for x,

$$\hat{x} = \frac{y - b_0}{b_1^*}$$

can be used to estimate a measurand x_{new} based on an observed measurement, y_{new}. In fact, a bit of careful probabilistic reasoning will establish that a 95% confidence set for x_{new} is

$$\{x | \text{ 95\% prediction limits for another } y \text{ given } x \text{ include } y_{\text{new}}\}$$

as illustrated below.

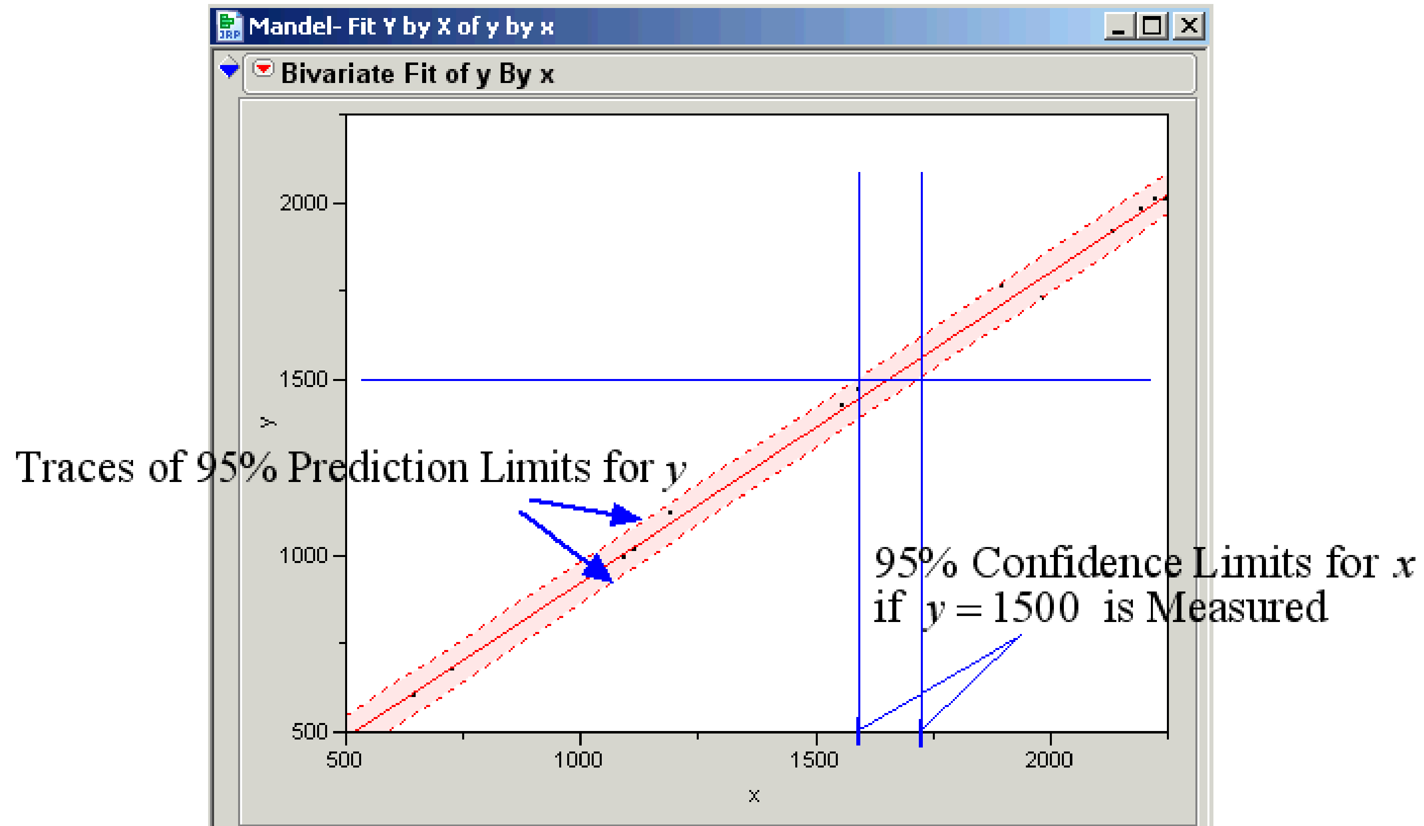

Figure 8: "Inversion" of Prediction Limits to Make Confidence Limits for x_{new} if $y_{\text{new}} = 1500$

"SPC" can contribute to effective measurement in the following way. If a device is periodically used to remeasure the same measurand, provided that measurand is physically stable, any changes in mean or standard deviation of measurements obtained can be attributed to changes in the device. So tools as simple as Shewhart $\bar{x}$ and s charts can be used to warn of changes in measurement device behavior.

Some difficulties

There are many ways in which the simple kinds of normal theory models (and associated inference methods) we like the most can fail to adequately describe real measurement. Some of these have more or less obvious "fixes," while others do not. All are worth considering/understanding.

Hysteresis

For one thing, the kind of simple statistical modeling discussed thus far is appropriate only when no previous measurement history is relevant when modeling any particular measurement made by a device. But it is not uncommon for some kinds of measurement devices to show so-called **hysteresis** effects. (The Meriam-Webster online dictionary defines this term as: *a retardation of an effect when the forces acting upon a body are changed*.) A resistance temperature measurement in a water bath that is being cooled may be too high, while a temperature measurement in that bath when it is being warmed may be too low.

This is a circumstance where in measuring $x(t)$ (the value of a measurand at time t), time t bias is somehow related to $x(s)$ for $s < t$. And if hysteresis effects are not ignorable, estimation of $x(t)$ based on $y(s)$ for $s \leq t$ becomes a time series/signal processing/filtering problem.

Instrument Dynamics

What may at first look much like the hysteresis possibility but is typically far less troublesome, is the possibility of measurement device **dynamics**. That is, where electronic circuitry is part of a measurement device there can be a time delay (characterized, for example, by instrument response that has some asymptote and or damped oscillatory behavior) before a measurement should be read as representing a measurand. Dealing with instrument dynamics is most often simply a matter of determining a period adequate to allow what the instrument reads to stabilize, and determining to only use "stabilized readings" as measured values.

Destructive Measurement

Much of the "obvious" statistical methodology potentially applicable to metrology requires remeasurement of the same measurand. But there are many contexts in the physical and engineering sciences where measurement is **destructive**, and remeasurement is not possible. In such cases, there is no magic (statistical or otherwise) that can be used to separate observed variation in measurement into measurand-to-measurand and basic measurement variation. All that can be done is to submit to a measurement device what one has reason to hope are "as similar as possible" measurands and recognize that what is observed in measurements y will be some combination of two types of variation. (If I am concerned about how consistent a "burst tester" is for measuring the properties of plastic soda bottles I and making, all I can do is try to keep all physical characteristics of the bottle making process "as close to constant as

possible," and submit bottles made under these constant process conditions for testing, knowing that what will be observed is bottle-to-bottle variation plus measurement error, and that there is no way to separate the two.)

Quantization/Digitalization

Digitalization/quantization of measurements is a reflection of the basic fact that any real measurement is read/reported only "to the nearest something." In contrast, our models for continuous random variables are for "real number" observations. While x might possibly be thought of as potentially any real number in some interval, the random variable y is discrete, and thus technically surely NOT normally distributed. That is, while one might entertain the

philosophical possibility that the gauge block in the caliper jaws in Figure 7 could actually be of *any* size from 0 mm to, say, 10 mm, what will be read from the caliper will be a number with a single decimal place.

This problem is arguably as old as metrology itself, but is now especially obvious in a time where gauges have digital readouts and data processing is done on explicitly digital computing equipment. It is mostly ignored by statisticians ... ?under the premise that it is typically unimportant? That may or may not be true.

Consider a model for measurement that is a modification of the earlier simple normal one, where

$$y = \text{round}\,(x + \epsilon)$$

round(·) is a rounding function and ϵ is normal with mean β and standard deviation σ_{device}. Whether this is a good description of a digital measurement

device like the caliper in Figure 7 is a question for an engineer. We will simply point out that in this model

- the characteristics of the distribution of y need not be at all like those of the distribution of $x + \epsilon$

- because of this, application of "ordinary" statistical methods can produce consistently misleading results

- development of methods based on interval-censored likelihoods fixes this problem

Use the notation

$$y^* = x + \epsilon$$

so that $y =$round(y^*) and consider the integer-rounded examples in the next two figures. In the first, the normal distribution of y^* and the discrete distribution of y are qualitatively similar, while in the second they are not.

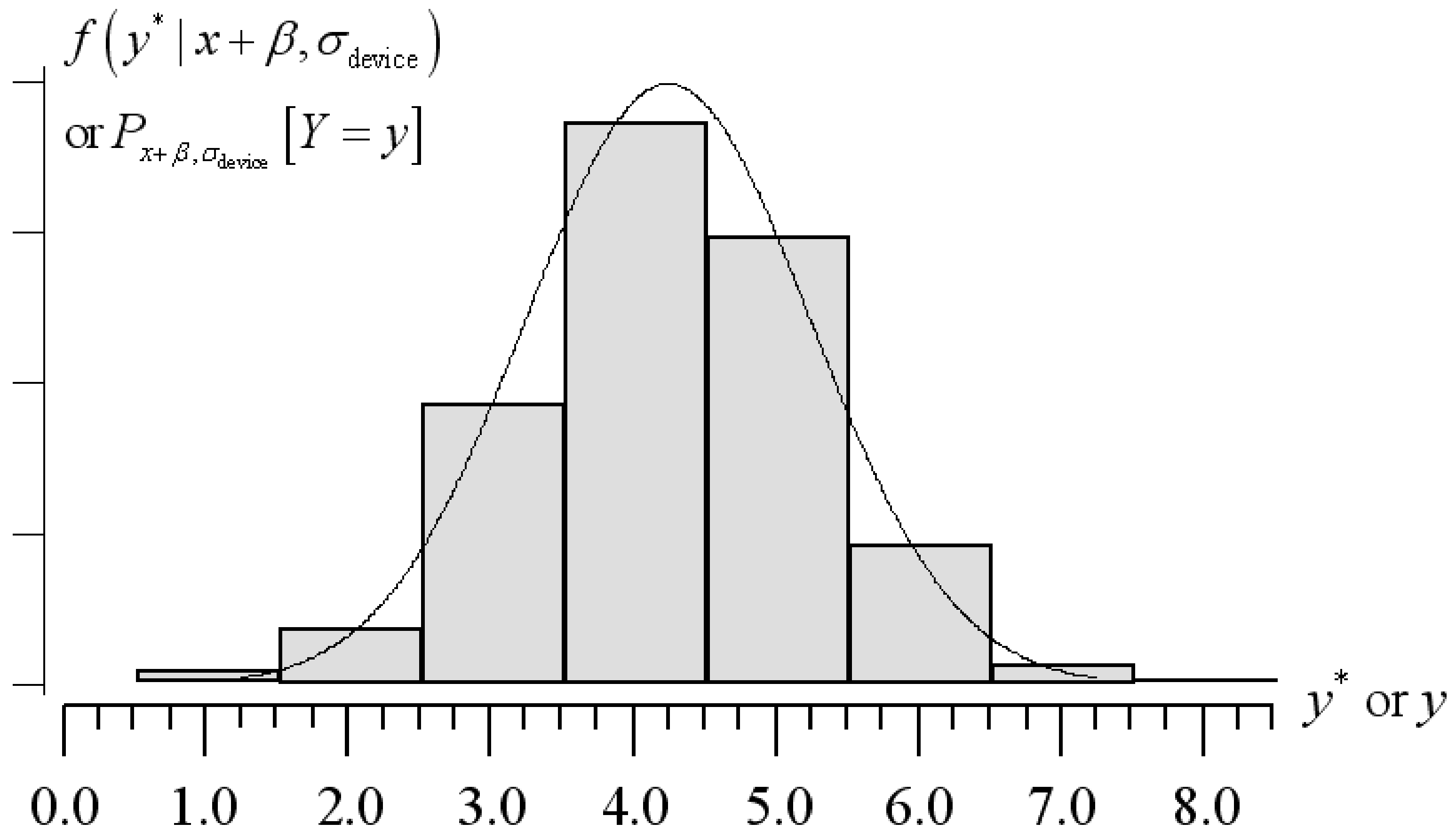

Figure 9: An Example Where the Effect of Rounding is Qualitatively Small

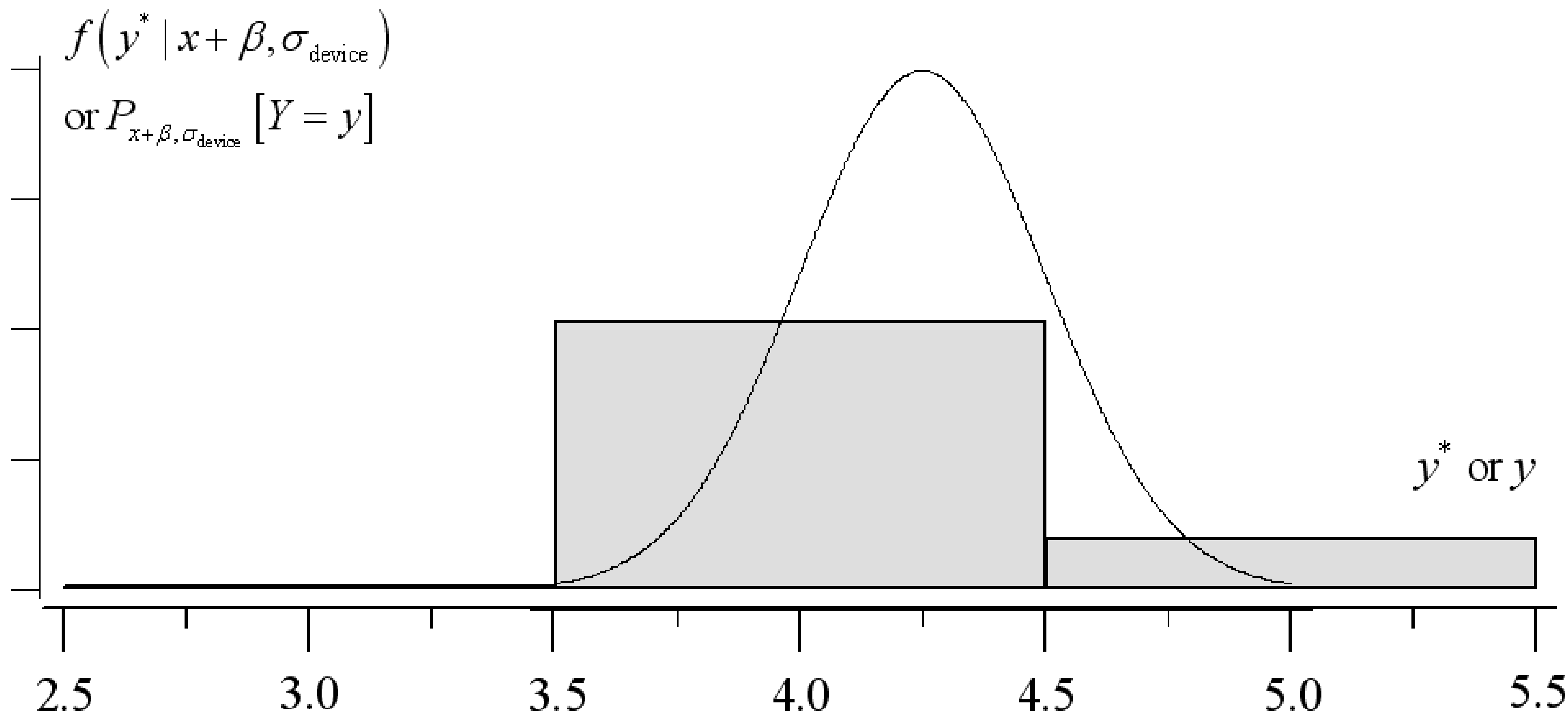

Figure 10: An Example Where the Effect of Rounding is Large

The moments for the two cases are

Figure	$\mu_{y^*} = x + \beta$	$\sigma_{y^*} = \sigma_{\text{device}}$	μ_y	σ_y
9	4.25	1.00	4.25	1.0809
10	4.25	.25	4.1573	.3678

"Ordinary" statistical methods will provide consistent estimation of characteristics of the distribution of what is observed (namely, y). But when σ_{y^*} is small in comparison to the granularity of the rounding, the characteristics of the distribution of y don't have to be at all like those of the distribution of y^*, which is already an imperfect representation of the measurand, x. A technical "fix" of this problem is to replace elementary inference methods with ones based on correct likelihoods (ones that recognize the **interval censoring** involved here). In a one-sample problem with integer rounding like that illustrated in Figures 9

and 10, this means replacing

$$L_{\mathsf{normal}}(\mu, \sigma) = \prod_{i=1}^{n} f(y_i | \mu, \sigma)$$

with

$$L_{\mathsf{rounded\ normal}}(\mu, \sigma) = \prod_{i=1}^{n} \left(\Phi\left(\frac{y_i + .5 - \mu}{\sigma}\right) - \Phi\left(\frac{y_i - .5 - \mu}{\sigma}\right) \right)$$

in the development of statistical methods. This at least allows consistent estimation of $\mu = x + \beta$ and $\sigma = \sigma_{\mathsf{device}}$.

Limits of Detection

As a final item on this list of potential difficulties there is the **limits of detection** problem in some sorts of measurements (like some chemical analyses). That is, in some contexts, rather than returning a number, an attempt to measure can return "below XXX." If the mechanism by which the **nondetect** response is produced can be thought of as

$$\text{nondetect} \Leftrightarrow y < L$$

and otherwise y is read, the notion of simple **left-censoring** can be used to produce appropriate likelihoods for statistical inference. But this simple model is a not always a good description of what really goes on. The mechanism by which a nondetect is produced is often neither independent of y, nor simple left censoring, nor well enough defined that effective statistical modeling is possible. (E.g., consider the difficulty of modeling the effects of psychological

and perceptual issues leading an individual analytical chemist to decide that there is so little of a trace element in a sample that any value he or she would report would be misleading, and that instead only "less than some threshold value I would take as meaningful result" should be reported.)

"Uncertainties" of measurement and probability and statistics

It is common to state that after measurement(s) and analysis one believes that a measurand, x, is bounded by

$$some\ value \pm U$$

where U is an "**uncertainty" of measurement**. Exactly what this should mean is, as far as I can tell, somewhat murky even in the minds of pretty good scientists. Qualitatively the intention is "clear" ... one is *pretty sure* that

$$x \in (some\ value - U, some\ value + U)$$

But a precise and consistent interpretation is another matter.

One possibility is that the meaning of a measurement uncertainty is an ordinary statistical meaning. That is, if one assumes that a device is well-calibrated/without bias, then if n measurements of x with the device produce values $y_1, y_2, \ldots, y_n$, ordinary statistical methods suggest that roughly 95% *confidence limits* for x are

$$\bar{y} \pm 2\frac{s}{\sqrt{n}}$$

and one might possibly take

$$some\ value = \bar{y} \ \text{ and } \ U = 2\frac{s}{\sqrt{n}}$$

and interpret the uncertainty bounds as confidence limits.

More complicated versions of this are, of course, quite possible. For example, if a device hasn't been calibrated, but n_1 measurements of x,

$$y_{11}, y_{12}, \cdots, y_{1n_1}$$

produce $\bar{y}_1$ and s_1, while n_2 measurements of a standard x_0,

$$y_{21}, y_{22}, \cdots, y_{2n_2}$$

produce $\bar{y}_2$ and s_2, then roughly 95% confidence limits for x are

$$\bar{y}_1 - (\bar{y}_2 - x_0) \pm 2\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and one might take

$$\mathit{some\ value} = \bar{y}_1 - (\bar{y}_2 - x_0) \quad \text{and} \quad U = 2\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and again interpret uncertainty bounds as confidence limits.

Where measurement uncertainty bounds are derived as and are completely interpretable as statistical confidence limits, it appears to be common to refer to measurement uncertainties as **Type A uncertainties** (or **statistical uncertainties**). But this is not the only way that scientists and engineers derive and try to interpret their measurement uncertainties.

Consider a simple case of length measurement, where someone wishes to measure the length of a nominally 5 km cross country running course with a wheel, as indicated in Figure 11.

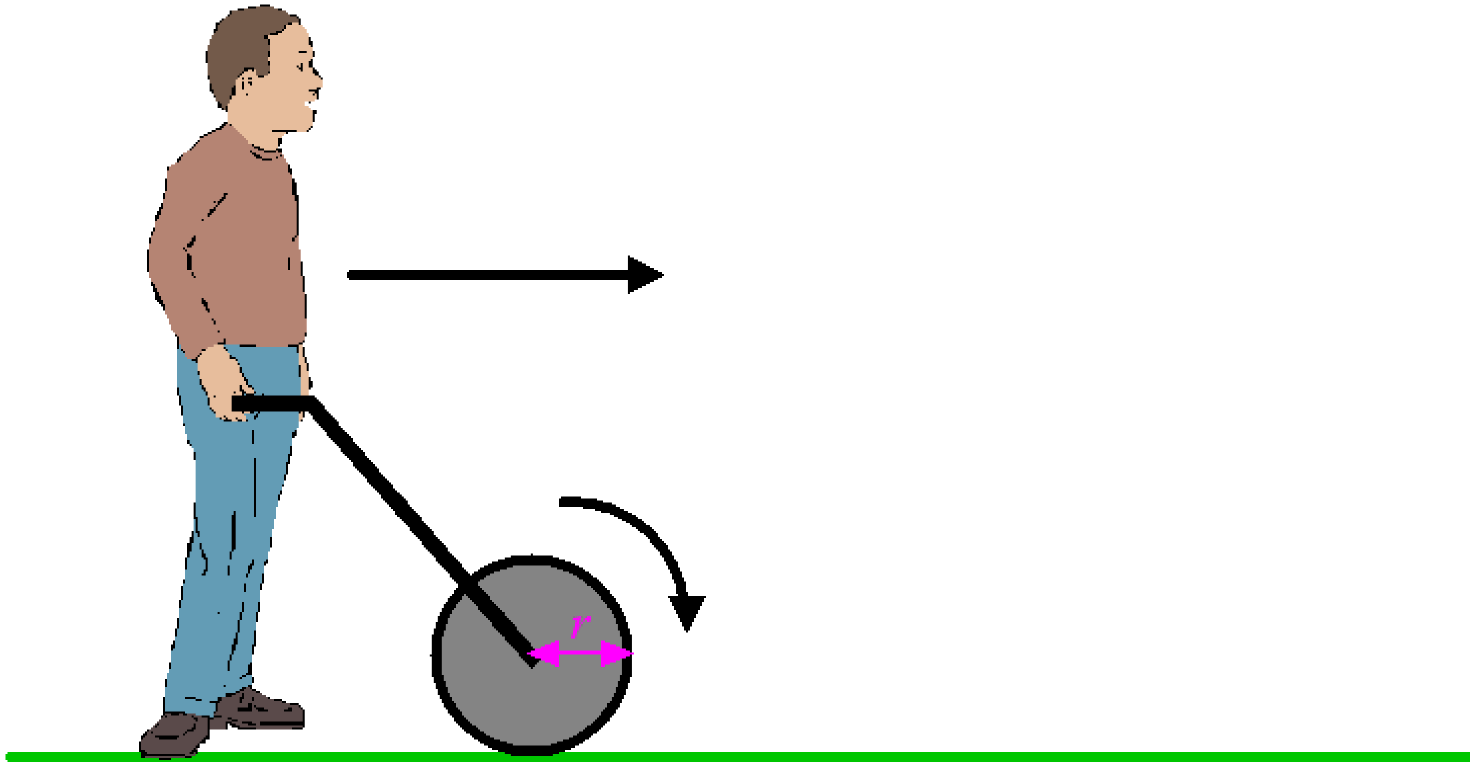

Figure 11: Cartoon Illustrating Use of a Measuring Wheeel

One might reason as follows. Suppose the radius of the wheel is nominally $r = .3\,\mathsf{m}$ and thus there should be about

$$\frac{5000}{2\pi\,(.3)} = 2,653.6$$

rotations of the wheel over a 5 km course. But perhaps manufacturing variability, wear, slight irregularity of shape, etc. could be expected to create as much as a .5% error in the actual radius and thus circumference of the wheel. Further, suppose that slipping of the wheel, irregularity of the terrain to be traversed, deviations from an exact course taken in the measurement process, reading error on the number of rotations made, etc. could be expected to create as much as a .5% error in the number of rotations made in measuring the course. Then in a worst case (where the wheel is too small by .5% and the realized number of revolutions for a wheel of that size is .5% higher than it

"should" be for a wheel of the actual size) a 5 km course will be measured as

$$\begin{aligned} number\ of\ rotations \times 2\pi\,(.3) &= (1.005) \times \frac{5000}{2\pi\,(.995)\,(.3)} \times 2\pi\,(.3) \\ &= 5050.3\,\text{m} \end{aligned}$$

Or, in the other worst case (where the wheel is too large by .5% and the realized number of revolutions for a wheel of that size is .5% lower than it "should" be for a wheel of the actual size) a 5 km course will be measured as

$$\begin{aligned} number\ of\ rotations \times 2\pi\,(.3) &= (.995) \times \frac{5000}{2\pi\,(1.005)\,(.3)} \times 2\pi\,(.3) \\ &= 4950.3\,\text{m} \end{aligned}$$

So perhaps, an "uncertainty" of about ± 50 m should be attached to a measurement made today of the length of the ISU cross country course with such a wheel.

Notice that this kind of calculation is not in any way a statistical inference. It is not directly based on data or statistical/probabilistic modeling, but rather some kind of "best guess" and "worst case" analyses. The $y \pm 50$ m values are something fundamentally different from confidence limits. This kind of "non-statistical" uncertainty is sometimes referred to as **Type B uncertainty**.

There are potentially many ways (some even less formal than what was used in our cross country course example) of arriving at non-statistical/Type B uncertainties. The NIST web site says "A Type B uncertainty is usually based on scientific judgment using all of the relevant information available, which may include:

- previous measurement data

- experience with, or general knowledge of, the behavior and property of relevant materials and instruments

- manufacturer's specifications

- data provided in calibration and other reports, and

- uncertainties assigned to reference data taken from other reports"

Now, precisely what to make of a Type B uncertainty statement is not immediately obvious. One may or may not really mean that he or she is ready to "take poison on" the measurand x being inside limits $some\ value \pm U$ in the

case of a Type B uncertainty statement. There is often some possibility that those limits fail to contain the measurand. How is this to be accounted for?

It is fairly standard practice to take the step of associating with a Type B uncertainty some kind of probability distribution that might describe what one believes about x. And it appears to be standard to use **normal**, **uniform**, or **triangular** distributions. For example, if one is really very "certain" that the limits $some\ value \pm U$ contain x, perhaps

a normal distribution with mean $\mu = some\ value$ and $\sigma = U/3$

or

a uniform $(somevalue - U, somevalue + U)$ distribution

or

a triangular distribution on $(somevalue - U, somevalue + U)$

as pictured in Figure 12 might be used to describe x.

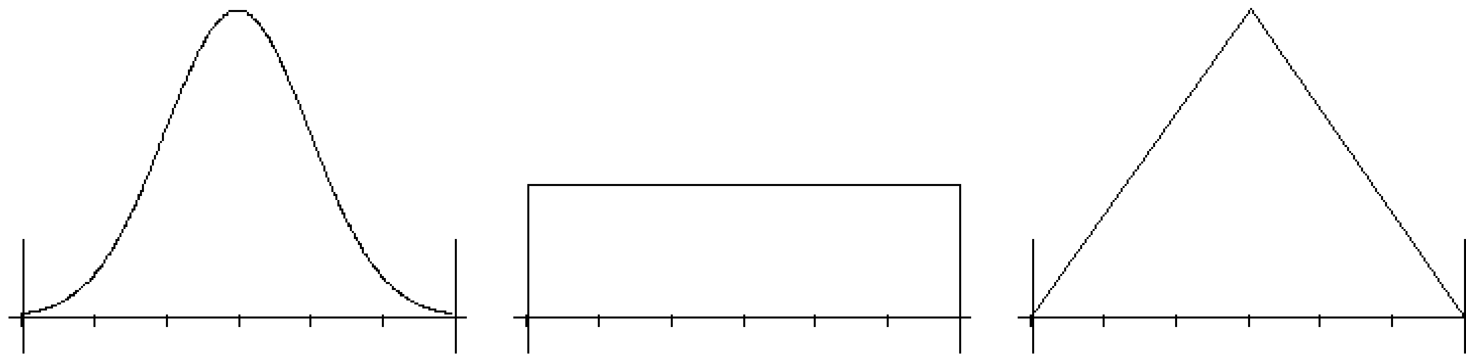

Figure 12: Three Possible "Distributions for x" Following Correpsonding to a Type B Uncertainty Statement

Or, perhaps U might be thought of as some number (often, the number is 2) of standard deviations of a probability distribution describing what one believes about x. That is, one might take "$some\ value \pm U$" as shorthand for

$$\text{a normal distribution with mean } \mu = some\ value \text{ and } \sigma = U/2$$

for

$$\text{a uniform distribution with mean } \mu = some\ value \text{ and } \sigma = U/\sqrt{3}$$

or for

$$\text{a triangular distribution with mean } \mu = some\ value \text{ and } \sigma = U/\sqrt{6}$$

to be employed as a description of ones beliefs about x. (These distributions are flatter than the ones in Figure 12 and place some probability outside the limits $some\ value \pm U$).

Now we have come to two fundamentally different types of thing potentially being called "measurement uncertainty"

- statistical uncertainty quantified in confidence statements, and

- "other" kinds of uncertainty being described via probability distributions.

There is reason to want to bring these apparently disparate meanings into one coherent intellectual framework. A way of potentially doing this is to replace classical inference methods with Bayesian inference methods. For example, once again assume that a device is well-calibrated/without bias, and that n measurements of x with the device produce values $y_1, y_2, \ldots, y_n$. Under model assumptions that what is observed are iid normal random variables with mean x and standard deviation σ_{device}, standard (improper) Bayes analysis using a "prior distribution" on $(x, \sigma_{\text{device}})$ that makes x and $\log \sigma_{\text{device}}$ independent

and each uniform on $\Re$, produces a t_{n-1} posterior (conditional distribution given the data) for

$$\frac{x - \bar{y}}{\frac{s}{\sqrt{n}}}$$

That is, the posterior distribution of x is a scaled t distribution recentered at $\bar{y}$, and thus the roughly 95% *confidence limits* for x

$$\bar{y} \pm 2\frac{s}{\sqrt{n}}$$

are also roughly 95% *credible limits for* x. These Type A uncertainty bounds for x have a probability interpretation like those typically applied to Type B bounds, at least if one accepts *conditional (posterior) probability* on the same grounds as probability.

Indirect measurement and combining uncertainties

For some measurands there is no obvious corresponding direct means of measurement. Consider, as a very elementary example, the problem of measuring the density of a liquid. One can typically evaluate density only by measuring out some volume of the liquid (using, say, a graduated cylinder) and then weighing the sample of liquid on a scale to determine a mass, and finally *calculating*

$$D = \frac{M}{V}$$

This is an example where **direct** measurements of M and V lead to an **indirect** measurement of D.

Actually, the measuring wheel example of the previous section can be thought of as another example of indirect measurement. The measured distance is

$$distance = number\ of\ rotations \times 2\pi r$$

and direct measurements of r and $number\ of\ rotations$ lead to an indirect measurement of $distance$.

A basic question of indirect measurement is "If x_i, $i = 1, 2, \ldots, k$ have uncertainty bounds

$$some\ value_i \pm U_i,$$

how should one develop uncertainty bounds for an indirect measurement

$$f(x_1, x_2, \ldots, x_k) \quad ?"$$

But once one has translated uncertainty bounds into probability distributions (in the the case of Type B uncertainty bounds) and/or posterior/conditional

distributions (in the case of Type A uncertainty bounds) this question can be rephrased as "Given distributions for x_i, $i = 1, 2, \ldots, k$ how does one find a distribution for $f(x_1, x_2, \ldots, x_k)$ ... and possibly use that distribution to identify sensible probability bounds that can be announced as uncertainty bounds for the indirect measurement?" But (at least if it makes sense to treat the direct measurands x_i as independent or otherwise identify a joint distribution for them from the marignals specified by the uncertainty bounds) this is "just" a Stat 542 problem. And while it is potentially unpleasant (to practically impossible) to do the calculus to handle this Stat 542 problem in closed form for anything but small k and simple f, simple Monte Carlo simulations provide an absolutely straightforward route through this problem.

As a final comment, it is worth noting here that an approximate analytical approach to the combining of uncertainties is to make a linearization of f and essentially apply the "delta method" or "propagation of errors" formula.

That is, under a model for the measurands that makes them uncorrelated, an approximate overall uncertainty might be stated as

$$somevalue \pm U$$

where $somevalue$ is f evaluated at the vector of mid-points of the individual uncertainty intervals and

$$U = \sqrt{\sum_{i=1}^{k} \left(\frac{\partial f}{\partial x_i} \right)^2 U_i^2}$$

where the partials are evaluated at the vector of mid-points of the individual uncertainty intervals.

Some elementary reading (in no particular order)

1. http://physics.nist.gov/cuu/Uncertainty/

2. Semyon Rabinovich (2005), *Measurement Errors and Uncertainty* (3rd Edition), Springer.

3. Gleser, Leon Jay (1998), Assessing uncertainty in measurement, *Statistical Science*, 13, 277-290.

4. D'Agostini, Giulio (2003), *Bayesian Reasoning in Data Analysis: A Critical Introduction*, World Scientific. (particularly the first 2 chapters)

5. Gertsbakh, Ilya (2003), *Measurement Theory for Engineers*, Springer.

6. Morris, Alan (2001), *Measurement & Instrumentation Principles*, Elsivier-Butterworth-Heinemann.

7. Vardeman, S.B. and Lee C.-S. (2005), Likelihood-based statistical estimation from quantized data, *IEEE Transactions on Instrumentation and Measurement*, 54(1),409-414.

8. Vardeman, S.B. and VanValkenburg, E.S. (1999), Two-way random-effects analyses and gauge R&R studies, *Technometrics*, 41, 202-211.