Elementary Statistical Methods and Measurement Error

(To Appear in *The American Statistician*)

Steve Vardeman, Joanne Wendelberger, Tom Burr, Mike Hamada, Marcus Jobe, Max Morris, Huaiqing Wu

(Supported in part by NSF grant DMS #0502347 EMSW21-RTG awarded to the Department of Statistics, Iowa State University.)

November 2009

Introduction

- Good measurement is essential to effective empirical learning. Good measurement and good statistics go hand-in-hand.
- How sources of physical variation interact with a data collection plan determine what can be learned in a statistical study ... and in particular, how measurement error is reflected in the resulting data.
- Even the most elementary statistical methods have their practical effectiveness limited by measurement variation.
- Even the most elementary statistical methods are helpful in quantifying the impact of measurement variation.

These things can and should be taught *earlier* rather than later.

Basic One and Two-Sample Methods

One sample inference methods like

$$\overline{y} \pm t\frac{s}{\sqrt{n}} \quad \text{and like} \quad s\sqrt{\frac{n-1}{\chi^2_{\text{upper}}}} \quad \text{and} \quad s\sqrt{\frac{n-1}{\chi^2_{\text{lower}}}}$$

are completely standard in elementary statistics teaching. Two-sample methods like

$$\overline{y}_1 - \overline{y}_2 \pm \hat{t}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and like

$$\frac{s_1}{s_2} \cdot \frac{1}{\sqrt{F_{(n_1-1),(n_2-1),\text{upper}}}} \quad \text{and} \quad \frac{s_1}{s_2} \cdot \frac{1}{\sqrt{F_{(n_1-1),(n_2-1),\text{lower}}}}$$

are only slight less common.

Basic One and Two-Sample Methods

Even these simple tools

- are limited in their practical usefulness by measurement error,
- are useful in quantifying measurement error, and
- provide a rich basis for raising/teaching important issues in measurement.

The "population mean(s)" and "population standard deviation(s)" addressed by the one- and two-sample methods are those of the **relevant data generating mechanism(s) ... including measurement.** Early careful and clear discussion of how these reflect measurement error is desperately needed in our teaching.

Basic Terminology

- A **measurand** is a numerical characteristic of an object being measured, x.
- A (measurement) **device** is a fixed combination of physical measurement equipment, operator/data-gatherer identity, measurement procedure, and surrounding physical circumstances (such as time of day, temperature, etc.) used to produce a **measurement**, y, intended to represent the measurand.
- A **measurement error** is the difference

$$\varepsilon = y - x \quad .$$

Simple Probability Modeling and Terminology (One Measurand)

If one adopts a probability model for y, the mean measurement error is the measurement **bias**

$$\delta(x) \equiv \mathsf{E}(\varepsilon) = \mathsf{E}(y - x)$$

(quantifying **accuracy**) and the standard deviation of the measurement error is

$$\sqrt{\mathsf{Var}(\varepsilon)} = \sigma_{\mathsf{meas}}(x)$$

(quantifying **precision**). And with this notation

$$\mathsf{E}(y) = x + \delta(x) \quad \text{and} \quad \sqrt{\mathsf{Var}(y)} = \sigma_{\mathsf{meas}}(x) \quad .$$

Calibration studies aim to reduce measurement bias (by finding "conversions" for raw measurements). At a minimum one hopes that bias is constant, $\delta(x) \equiv \delta$. Metrologists call such devices "**linear**."

Simple Probability Modeling and Terminology (One Measurand)

A simple representation of this modeling and terminology is:

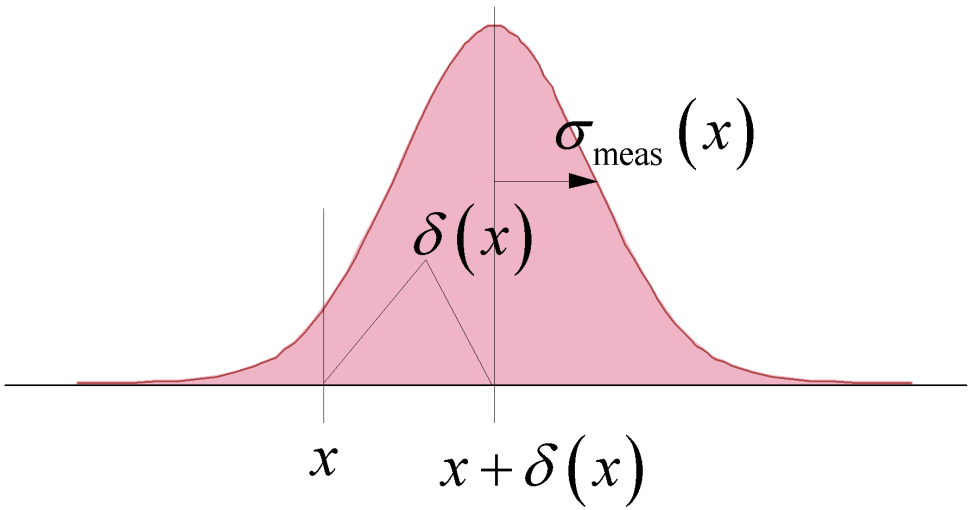


Figure: Distribution of a measurement y for a measurand x.

Simple Probability Modeling (Multiple Measurands)

Treating x as random (each realization representing a draw from an underlying population of measurands), with constant bias and precision of measurement, it may be sensible to model x and ε as independent (with $\mathrm{E}(x) = \mu_x$ and $\mathrm{Var}(x) = \sigma_x^2$) and get

$$\mathrm{E}(y) = \mu_x + \delta \quad \text{and} \quad \sqrt{\mathrm{Var}(y)} = \sqrt{\sigma_x^2 + \sigma_{\text{meas}}^2} \quad .$$

This can be represented to elementary audiences graphically:

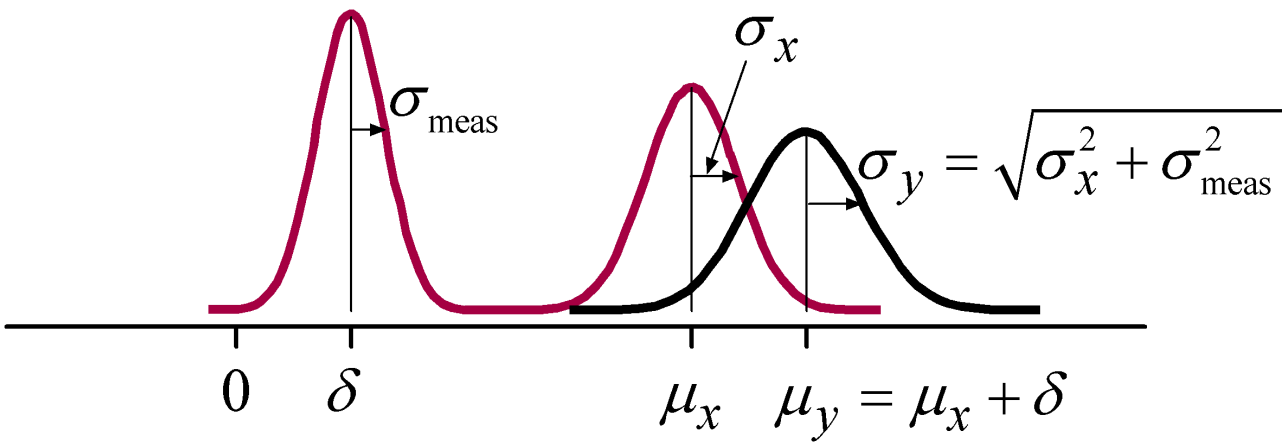


Figure: Random measurement error and measurand variation combine to produce observed variation.

One Sample and the "Box of Tickets Model"

Exactly what are μ and σ anyway???

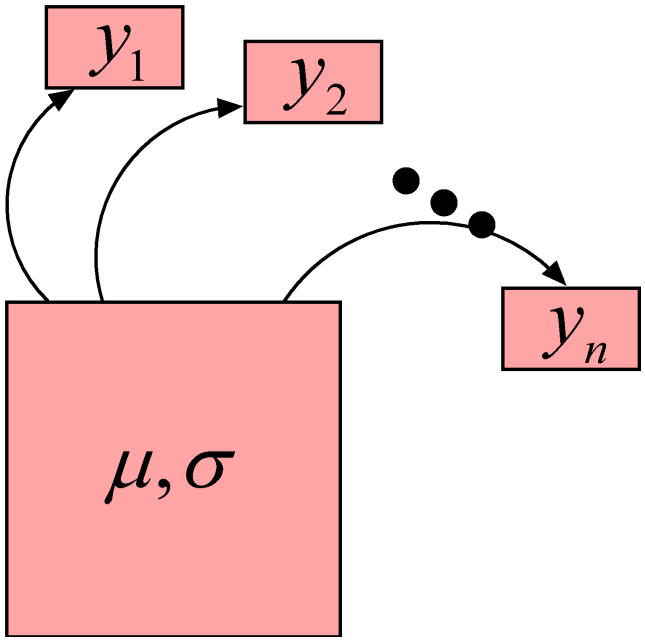


Figure: The standard elementary statistics "sampling from a box of tickets" model.

One-Sample Inference and Measurement Error (Case 1)

Consider applications of simple one-sample inference formulas recognizing the reality of measurement error. There are a number of ways a single sample of n measurements can arise. Here is a schematic for a first.

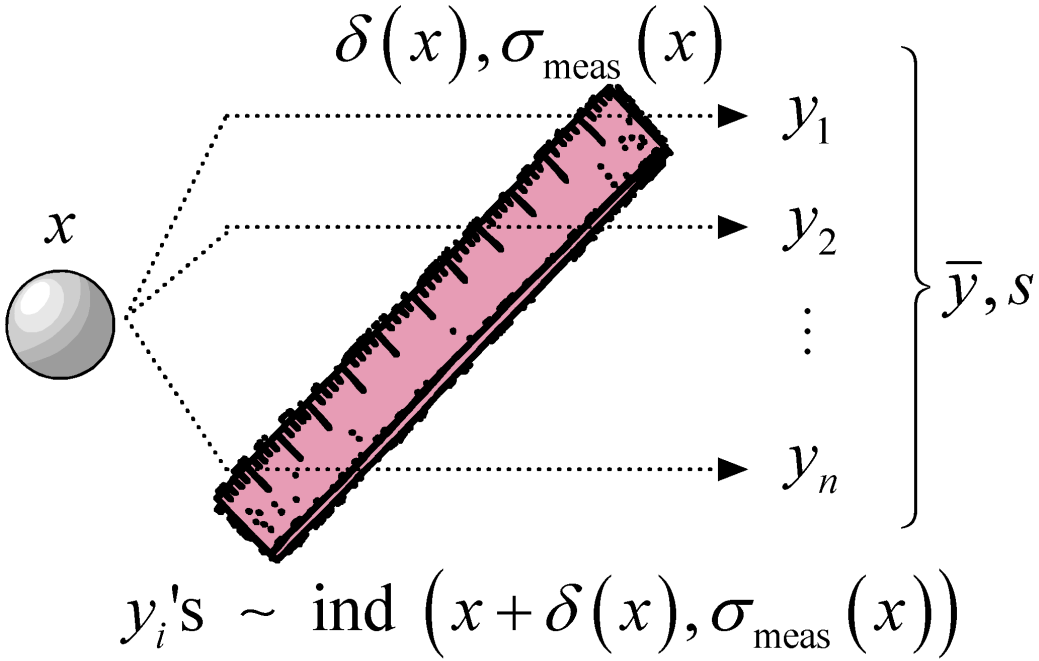
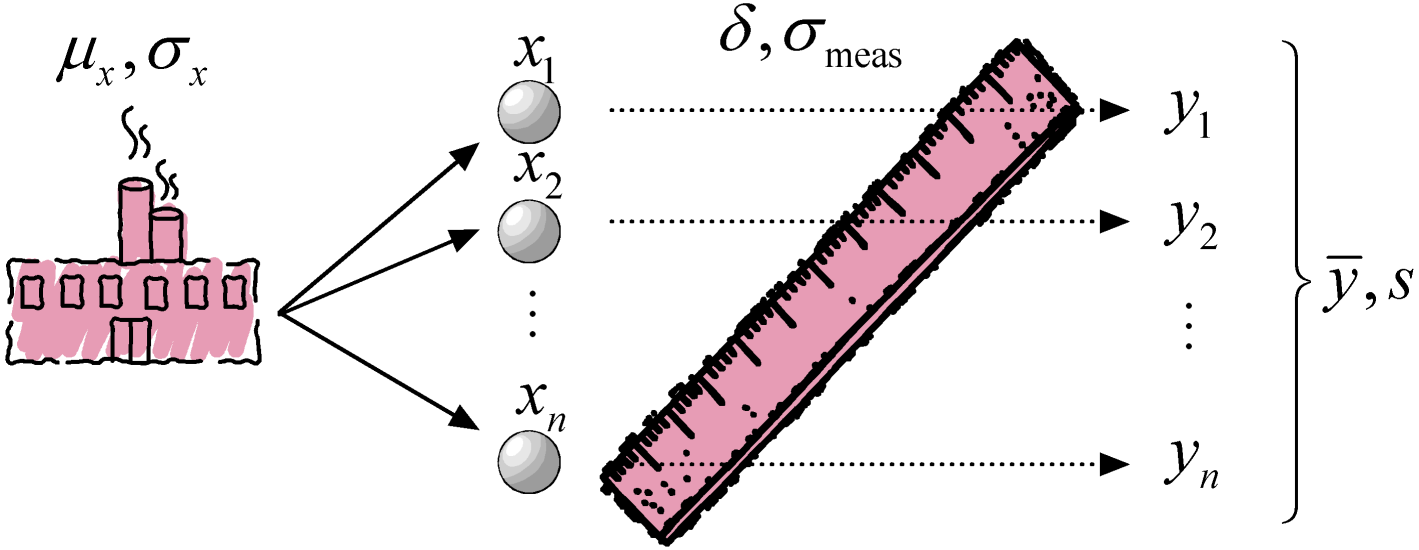


Figure: (Case 1) A single sample derived from n repeat measurements made on a single measurand with a fixed device.

One-Sample Inference and Measurement Error (Case 2)

Here is a schematic for a second (and quite different) way a single sample of n measurements can arise.



$$y_i\text{'s} \sim \text{ind}\left(\mu_x + \delta, \sqrt{\sigma_x^2 + \sigma_{\text{meas}}^2}\right)$$

Figure: (Case 2) A single sample derived from single measurements made of n different measurands from a physically stable process with a fixed device (assuming device linearity and constant measurement precision).

One-Sample Inference and Measurement Error (Case 1)

In Case 1 the t interval estimates $x + \delta(x)$ and the χ^2 interval estimates $\sigma_{\text{meas}}(x)$.

- The first means that
  - only a well-calibrated device provides the truth about x even in the long run, and
  - when a **standard** is available (so x is "known") $\delta(x)$ may be estimated ... and a **calibration adjustment** derived.
- The second means that device precision may be directly inferred.

One-Sample Inference and Measurement Error (Case 2)

In general, if the measurand is random

$$\mathsf{E}(y) = \mathsf{E}(x) + \mathsf{E}(\delta(x))$$

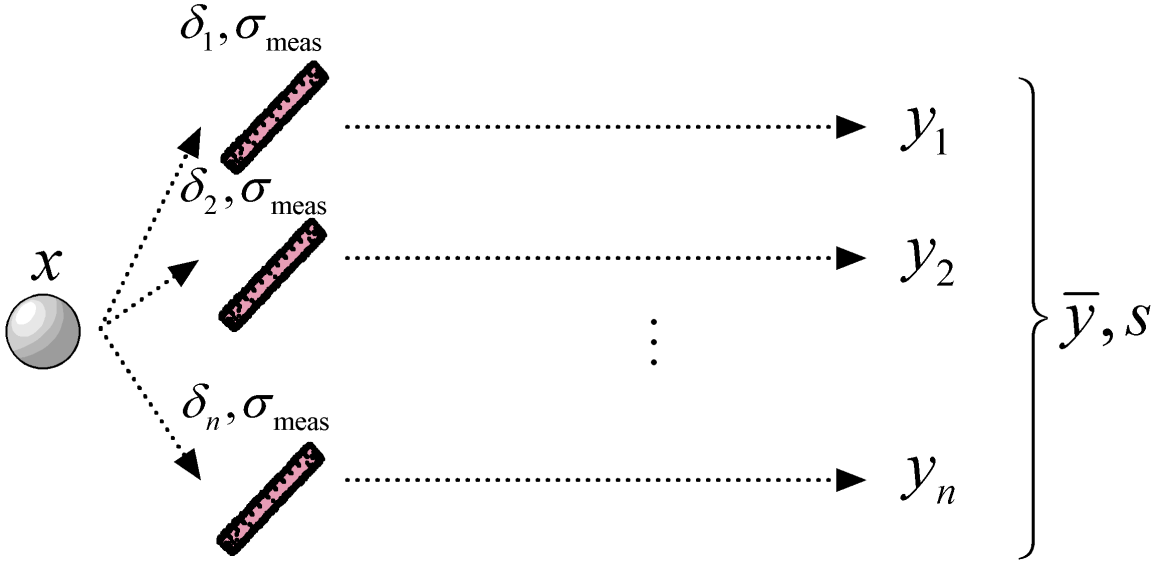
and

$$\mathsf{Var}(y) = \mathsf{E}\left(\sigma^2_{\text{meas}}(x)\right) + \mathsf{Var}(x + \delta(x))$$

which reduce to the simple forms $\mu_x + \delta$ and $\sigma^2_x + \sigma^2_{\text{meas}}$ on the figure only when the device is linear and precision is constant. Even then, in Case 2 the elementary statistics implicit assumption that a sample of measurements directly informs one about the distribution of the measurand is correct *only when bias and measurement standard deviation are negligible.*

One-Sample Inference and Measurement Error (Case 3)

Here is a schematic for a third way a single sample of n measurements can arise.



$$y_i\text{'s} \sim \text{ind}\left(x + \mu_\delta, \sqrt{\sigma_\delta^2 + \sigma_{\text{meas}}^2}\right)$$

Figure: (Case 3) A single sample obtained from a single measurement of a fixed measurand made with each of n different devices from a large population of such devices with a common measurement precision σ_{meas}.

One-Sample Inference and Measurement Error (Case 3)

We may suppress dependence upon the (fixed) x. Random choice of device makes δ and σ_{meas} random. When σ_{meas} is constant across devices one gets the simple mean $x + \mu_\delta$ and standard deviation $\sqrt{\sigma_\delta^2 + \sigma_{\text{meas}}^2}$ on the figure.

Case 3 can be applied to the important situation of use of the same measurement equipment by different operators. In this context σ_{meas}^2 is a **repeatability** variance and σ_δ^2 is a **reproducibility** variance. In this situation the t interval estimates the measurand plus average bias and the χ^2 interval estimates an "R&R standard deviation."

One-Sample Inference and Measurement Error (Cases 1 through 3)

- These cases provide motivation for elementary qualitative discussion of components of variance and data collection plans and calculation methods to separate them.
  - Cases 1 and 2 suggest that some combination of them could separate σ_x and σ_{meas}.
  - Cases 1 and 3 suggest that some combination of them could separate σ_δ and σ_{meas}.
- These cases already suggest the vital necessity of repeat measurement of a fixed measurand for a given measurement device for purposes of ascertaining the size of basic measurement noise.

One-Sample Inference and Measurement Error (Case 4)

Here is a schematic for a fourth way a single sample of n (differences of) measurements can arise.

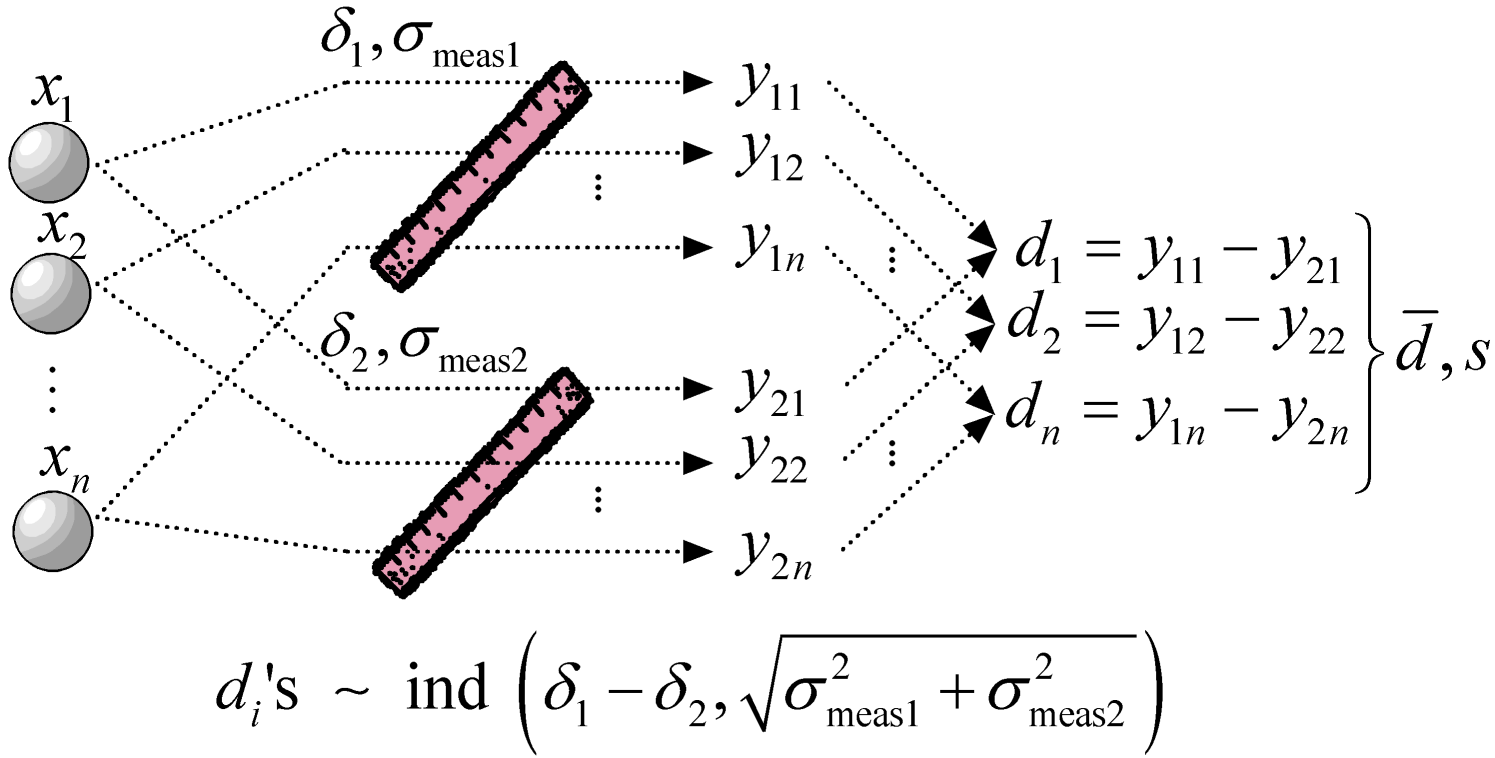


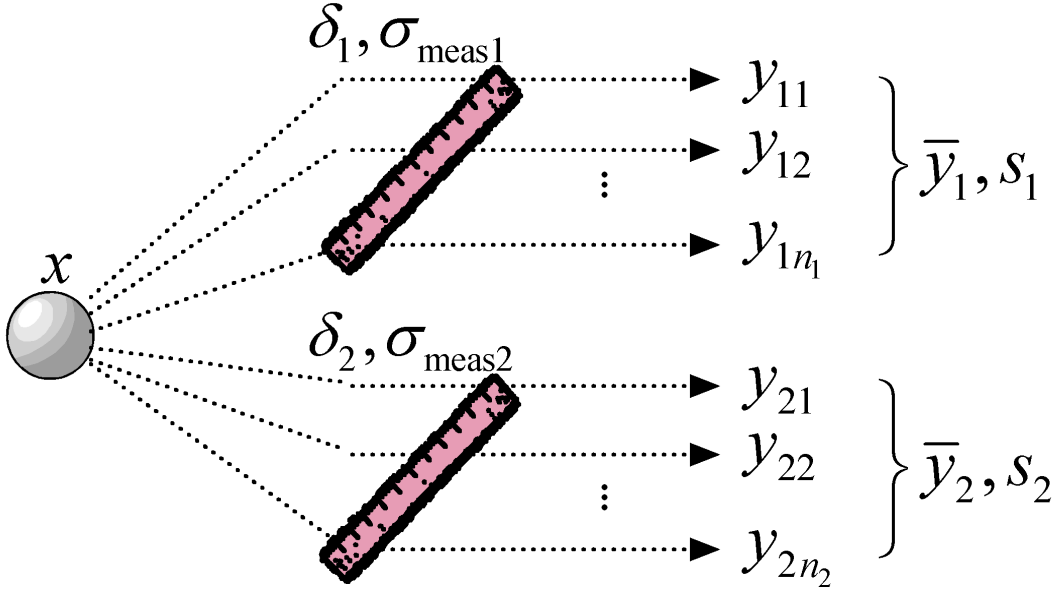
Figure: (Case 4) A single sample consisting of differences of measurements on n different measurands using two linear devices ($\delta_1(x) = \delta_1, \delta_2(x) = \delta_2$) with constant measurement precisions ($\sigma_{\text{meas1}}(x) = \sigma_{\text{meas1}}, \sigma_{\text{meas2}}(x) = \sigma_{\text{meas2}}$).

One-Sample Inference and Measurement Error (Case 4)

In Case 4, under the assumptions of device linearity and constant measurement precision, the one-sample t interval estimates $\delta_1 - \delta_2$. That is, one can use the method to compare device biases. In the event that the second device is known to be well-calibrated (so that δ_2 is known to be 0) this obviously provides a way to estimate the bias of the first device.

Two-Sample Inference and Measurement Error (Case 5)

A schematic for a two-sample version of Case 1 is next.



$$y_{1i}\text{'s} \sim \text{ind}\left(x+\delta_1, \sigma_{\text{meas1}}\right) \text{ independent of}$$

$$y_{2i}\text{'s} \sim \text{ind}\left(x+\delta_2, \sigma_{\text{meas2}}\right)$$

Figure: (Case 5) Two samples of n_1 and n_2 measurements of a single measurand with two devices.

Two-Sample Inference and Measurement Error (Case 5)

In Case 5, the difference in "population means" is

$$(x + \delta_1) - (x + \delta_2) = \delta_1 - \delta_2$$

and the ratio of "population standard deviations" is

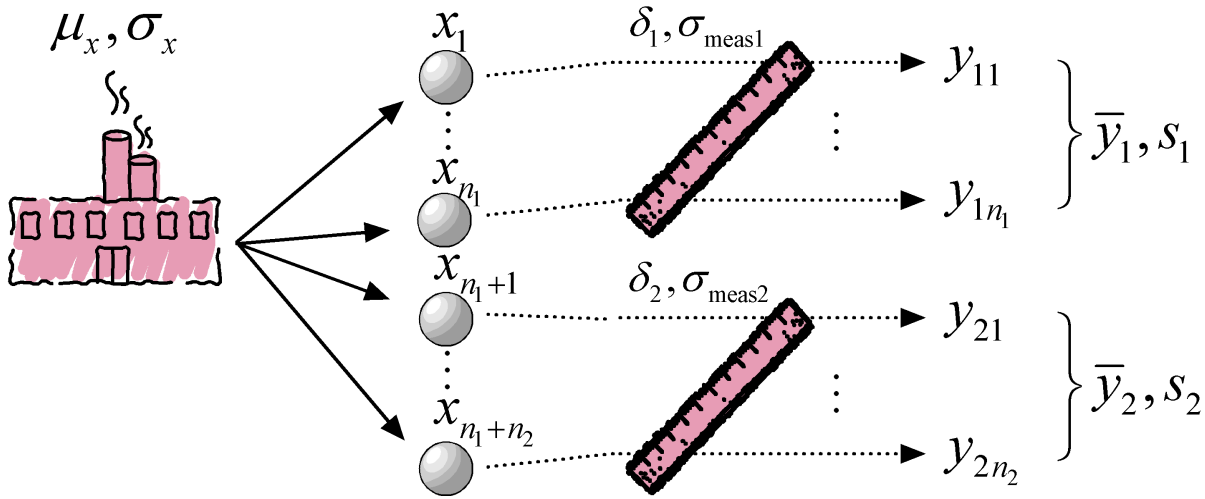
$$\sigma_{\text{meas1}} = \sigma_{\text{meas2}}$$

So the t interval can be used to compare device biases and the F interval can be used to compare device precisions (at x). If the devices are linear and with constant precisions these comparisons may be extended to other values of x.

One practically important situation where the data collection design of Case 5 is not available is that where measurement is **destructive**.

Two-Sample Inference and Measurement Error (Case 6)

A schematic for a two-sample design available even where measurement is destructive is below.



$$y_{1i}\text{'s} \sim \text{ind}\left(\mu_x + \delta_1, \sqrt{\sigma_x^2 + \sigma_{\text{meas1}}^2}\right) \text{ independent of}$$

$$y_{2i}\text{'s} \sim \text{ind}\left(\mu_x + \delta_2, \sqrt{\sigma_x^2 + \sigma_{\text{meas2}}^2}\right)$$

Figure: (Case 6) Two samples consisting of single measurements made on $n_1 + n_2$ measurands from a stable process, assuming devices are linear and have constant precisions.

Two-Sample Inference and Measurement Error (Case 6)

Arguing as for Case 2, under the data collection design of Case 6

$$\mathsf{E}(y_1) = \mathsf{E}(x) + \mathsf{E}(\delta_1(x)) \quad \text{and} \quad \mathsf{E}(y_2) = \mathsf{E}(x) + \mathsf{E}(\delta_2(x))$$

and

$$\begin{aligned} \mathsf{Var}(y_1) &= \mathsf{E}\left(\sigma^2_{\text{meas1}}(x)\right) + \mathsf{Var}(x + \delta_1(x)) \quad \text{and} \\ \mathsf{Var}(y_2) &= \mathsf{E}\left(\sigma^2_{\text{meas2}}(x)\right) + \mathsf{Var}(x + \delta_2(x)) \end{aligned}$$

These reduce to the forms on the schematic when the devices are linear and have constant precision.

So, for well-behaved devices, the t interval for the difference in population means allows comparison of biases.

Two-Sample Inference and Measurement Error (Cases 5 and 6)

Comparison of the "population standard deviations" specified in Cases 5 and 6 leads to the important conclusions that

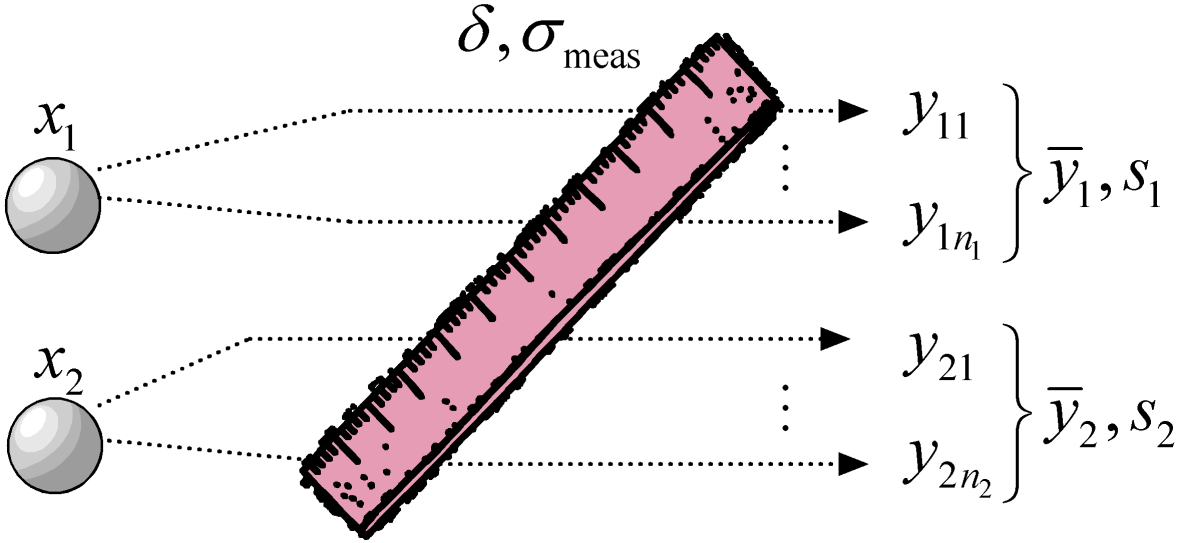
- t inferences for

$$\delta_1 - \delta_2$$

  based on the design of Case 5 are likely to be more informative than ones based on the design of Case 6 (where both are possible), and
- F inferences based on Case 6 provide an indirect way to compare device precisions, but they will be clouded by measurand variation.

Two-Sample Inference and Measurement Error (Case 7)

A final pair of schematics illustrating two-sample scenarios involving a single measurement device are next.

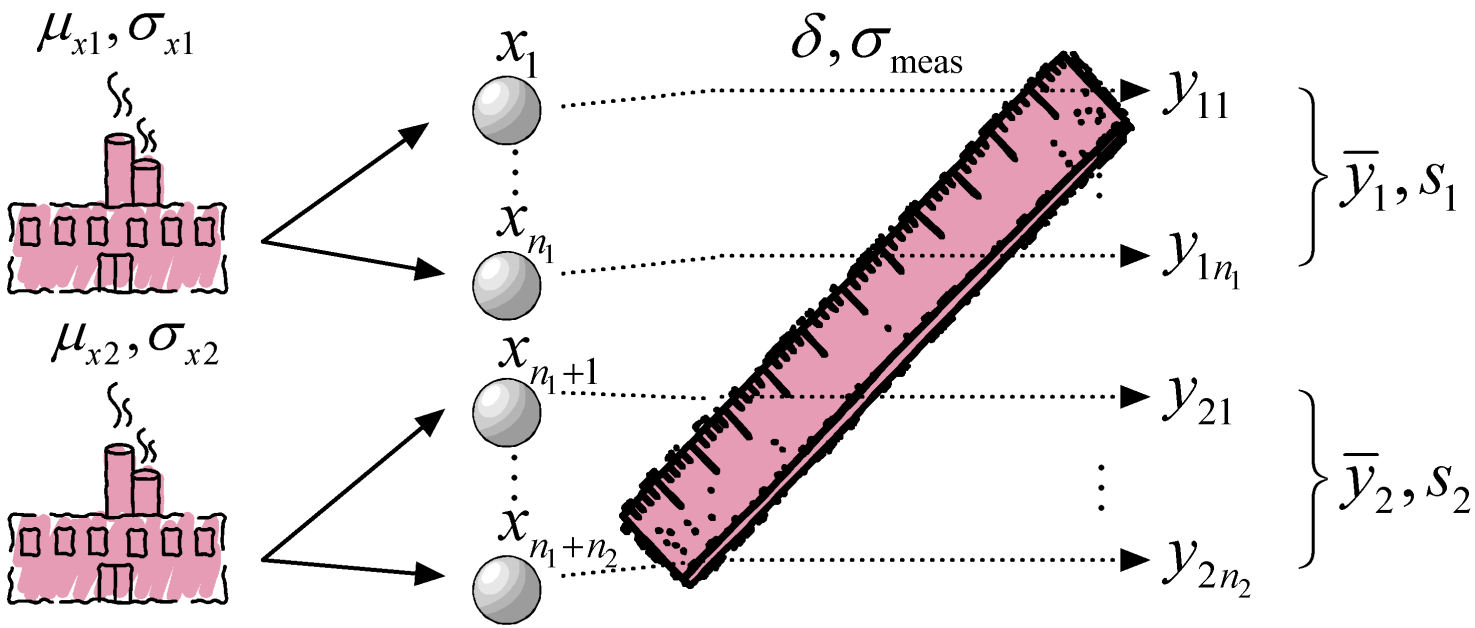


$$y_{1i}\text{'s} \sim \text{ind}\left(x_1 + \delta, \sigma_{\text{meas}}\right) \text{ independent of}$$

$$y_{2i}\text{'s} \sim \text{ind}\left(x_2 + \delta, \sigma_{\text{meas}}\right)$$

Figure: (Case 7) Two samples consisting of repeat measurements of two different measurands made with one device (assuming linearity and constant precision).

Two-Sample Inference and Measurement Error (Case 8)



$$y_{1i}\text{'s} \sim \text{ind}\left(\mu_{x1}+\delta, \sqrt{\sigma_{x1}^2+\sigma_{\text{meas}}^2}\right) \text{ independent of}$$

$$y_{2i}\text{'s} \sim \text{ind}\left(\mu_{x2}+\delta, \sqrt{\sigma_{x2}^2+\sigma_{\text{meas}}^2}\right)$$

Figure: (Case 8) Two samples consisting of single measurements made using a single device on multiple measurands produced by two stable processes (assuming device linearity and constant measurement precision).

Two-Sample Inference and Measurement Error (Case 7)

In Case 7 the difference of "population means" is

$$\mu_1 - \mu_2 = (x_1 + \delta(x_1)) - (x_2 + \delta(x_2))$$

If the device is linear (i.e., $\delta(x) = \delta$), this is simply

$$x_1 - x_2$$

and the t limits apply to the difference in measurands. (This actually holds even if the device does not have the constant precision indicated by the schematic.) There is no practical interest in comparing "population standard deviations" via the F limits in Case 7 (unless one allows the possibility that $\sigma_{\text{meas}}(x_1) \neq \sigma_{\text{meas}}(x_2)$ and wishes to investigate whether measurement precision varies with measurand).

Two-Sample Inference and Measurement Error (Case 8)

Case 8 enables the comparison of process characteristics. If the device is linear the t interval provides estimation of

$$\mu_1 - \mu_2 = (\mu_{x1} + \delta) - (\mu_{x2} + \delta) = \mu_{x1} - \mu_{x2}$$

the difference in process means. However, even assuming constant measurement precision, a direct comparison of process standard deviations (σ_{x1} and σ_{x2}) is not available. This data collection scheme is probably the most commonly used example of "tickets from two boxes" presented in elementary statistics courses. Failure to make the point that the F confidence limits in are essentially for comparing σ_{x1} and σ_{x2} only when measurement noise is completely negligible invites misinterpretation of statistical analysis results.

Instruction

- These concepts (and versions of these schematics) are part of a thorough introduction to statistics and measurement in an SQC course at ISU for IE juniors. Anecdotal evidence says these ideas can be grasped at this level.
- At ISU, we use simple hands-on in-class measurement exercises involving simple/cheap plastic dial calipers and (hard to size) Styrofoam ^{TM} packing peanuts to give students a better understanding of some of the data collection plans and measurement concepts illustrated here.
- A small experiment introducing the one-sample ideas to students in a first Business Stat course at Miami of Ohio had mixed results.

Summary/Reiteration

- Good measurement is essential to effective empirical learning. Good measurement and good statistics go hand-in-hand.
- How sources of physical variation interact with a data collection plan determine what can be learned in a statistical study ... and in particular, how measurement error is reflected in the resulting data.
- Even the most elementary statistical methods have their practical effectiveness limited by measurement variation.
- Even the most elementary statistical methods are helpful in quantifying the impact of measurement variation.

These things can and should be taught *earlier* rather than later.

Thanks

Thanks for your attention!