

CHAPTER 2

STATISTICS AND MEASUREMENT

Good measurement is fundamental to quality assurance. That which cannot be measured cannot be guaranteed to a customer. If Brinell hardness 220 is needed for certain castings and one has no means of reliably measuring hardness, there is no way to provide the castings. So successful companies devote substantial resources to the development and maintenance of good measurement systems. In this chapter, we consider some basic concepts of measurement and discuss a variety of statistical tools aimed at quantifying and improving the effectiveness of measurement.

The chapter begins with an exposition of basic concepts and introduction to probability modeling of measurement error. Then elementary one- and two-sample statistical methods are applied to measurement problems in Sect. 2.2. Section 2.3 considers how slightly more complex statistical methods can be used to quantify the importance of sources of variability in measurement. Then Sect. 2.4 discusses studies conducted to evaluate the sizes of unavoidable measurement variation and variation in measurement chargeable to consistent differences between how operators use a measurement system. Section 2.5 considers statistical treatment of the measurement calibration problem. Finally, in Sect. 2.6 the chapter concludes with a brief section on contexts where “measurements” are go/no-go calls on individual items.

2.1 Basic Concepts in Metrology and Probability Modeling of Measurement

Metrology is the science of measurement. Measurement of many physical quantities (like lengths from inches to miles and weights from ounces to tons) is so commonplace that we think little about basic issues involved in metrology. But often engineers are forced by circumstances to leave the world of off-the-shelf measurement technology and devise their own instruments. And frequently because of externally imposed quality requirements for a product, one must ask “Can we even measure that?” Then the fundamental issues of **validity**, **precision**, and **accuracy** come into focus.

Definition 4 *A measurement or measuring method is said to be **valid** if it usefully or appropriately represents the feature of the measured object or phenomenon that is of interest.*

Definition 5 *A measurement system is said to be **precise** if it produces small variation in repeated measurement of the same object or phenomenon.*

Definition 6 *A measurement system is said to be **accurate** (or sometimes **unbiased**) if on average it produces the true or correct values of quantities of interest.*

Validity is the first concern when developing a measurement method. Without it, there is no point in proceeding to consider precision or accuracy. The issue is whether a method of measurement will faithfully portray the quantity of interest. When developing a new pH meter, one wants a device that will react to changes in acidity, not to changes in temperature of the solution being tested or to changes in the amount of light incident on the container holding the solution. When looking for a measure of customer satisfaction with a new model of automobile, one needs to consider those things that are important to customers. (For example, the number of warranty service calls per vehicle is probably a more valid measure of customer satisfaction or aggravation with a new car than warranty dollars spent per vehicle by the manufacturer.)

Precision of measurement has to do with getting similar values every time a particular measurement is done. A bathroom scale that can produce any number between 150 lb and 160 lb when one gets on it repeatedly is really not very useful. After establishing that a measurement system produces valid measurements, consistency of those measurements is needed. Figure 2.1 portrays some hardness measurements made by a group of students (Blad, Sobotka, and Zaugg) on a single metal specimen with three different hardness testers. The figure shows that the dial Rockwell tester produced the most consistent results and would therefore be termed the most precise.

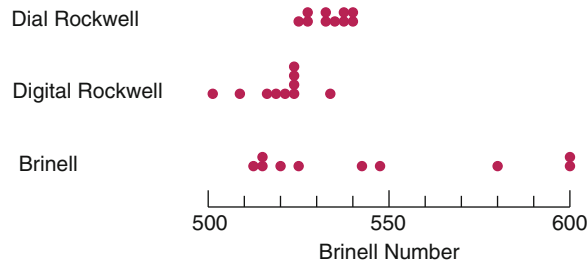


FIGURE 2.1. Brinell hardness measurements made on three different machines

Precision is largely an intrinsic property of a measurement method or device. There is not really any way to “adjust” for poor precision or to remedy it except to (1) overhaul or replace measurement technology or to (2) average multiple measurements. In this latter regard, the reader should be familiar with the fact from elementary statistics that if y_1, y_2, \dots, y_n can be thought of as independent measurements of the same quantity, each with some mean μ and standard deviation σ , then the sample mean, \bar{y} , has expected or average value μ and standard deviation σ/\sqrt{n} . So people sometimes rely on multiple measurements and averaging to reduce an unacceptable precision of individual measurement (quantified by σ) to an acceptable precision of average measurement (quantified by σ/\sqrt{n}).

But even validity and precision together don’t tell the whole story regarding the usefulness of real-world measurements. This can be illustrated by again considering Fig. 2.1. The dial Rockwell tester is apparently the most precise of the three testers. But it is *not* obvious from the figure what the truth is about “the” real Brinell hardness of the specimen. That is, the issue of accuracy remains. Whether any of the three testers produces essentially the “right” hardness value on average is not clear. In order to assess that, one needs to reference the testers to an accepted standard of hardness measurement.

The task of comparing a measurement method or device to a standard one and, if necessary, working out conversions that will allow the method to produce “correct” (converted) values on average is called **calibration**. In the United States, the National Institute of Standards and Technology (NIST) is responsible for maintaining and disseminating consistent standards for calibrating measurement equipment. One could imagine (if the problem were important enough) sending the students’ specimen to NIST for an authoritative hardness evaluation and using the result to calibrate the testers represented in Fig. 2.1. Or more likely, one might test some other specimens supplied by NIST as having known hardnesses and use those to assess the accuracy of the testers in question (and guide any recalibration that might be needed).

An analogy that is sometimes helpful in remembering the difference between accuracy and precision of measurement is that of target shooting. Accuracy in target shooting has to do with producing a pattern centered on the bull’s eye (the ideal). Precision has to do with producing a tight pattern (consistency). Figure 2.2 on page 36 illustrates four possibilities for accuracy and precision in target shooting.

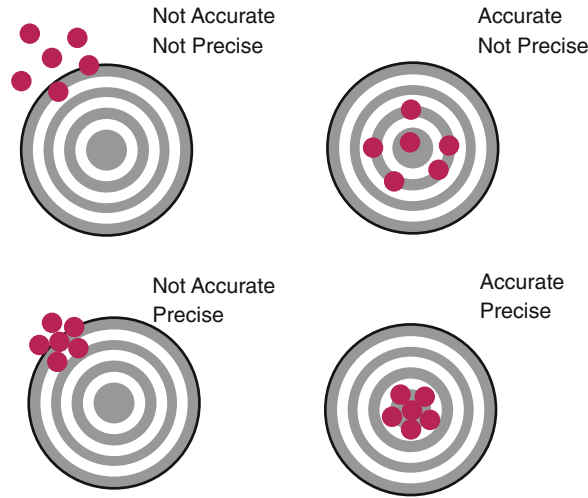


FIGURE 2.2. Measurement/target-shooting analogy

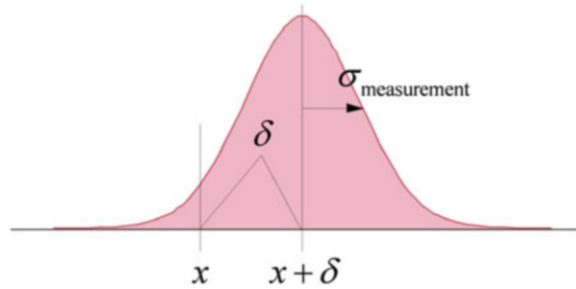


FIGURE 2.3. The distribution of a measurement y of a quantity x where measurement bias is δ and standard deviation of measurement is $\sigma_{\text{measurement}}$. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 47. © 2010 Taylor & Francis. Adapted with permission

Probability theory provides a helpful way to describe measurement error/variation. If a fixed quantity x called the **measurand** is to be measured with error (as all real-world quantities are), one might represent what is actually observed as

$$y = x + \epsilon \quad (2.1)$$

where ϵ is a random variable, say with mean δ and standard deviation $\sigma_{\text{measurement}}$. Model (2.1) says that the mean of what is observed is

$$\mu_y = x + \delta. \quad (2.2)$$

If $\delta = 0$, the measurement of x is accurate or unbiased. If δ is not 0, it is called the **measurement bias**. The standard deviation of y is (for fixed x) the standard

deviation of ϵ , $\sigma_{\text{measurement}}$. So $\sigma_{\text{measurement}}$ quantifies measurement precision in model (2.1). Figure 2.3 pictures the probability distribution of y and the elements x , δ , and $\sigma_{\text{measurement}}$.

Ideally, δ is 0 (and it is the work of calibration to attempt to make it 0). At a minimum, measurement devices are designed to have a **linearity** property. This means that over the range of measurands a device will normally be used to evaluate, if its bias is not 0, it is at least constant (i.e., δ does not depend upon x). This is illustrated in Fig. 2.4 (where we assume that the vertical and horizontal scales are the same).

Device
“Linearity”

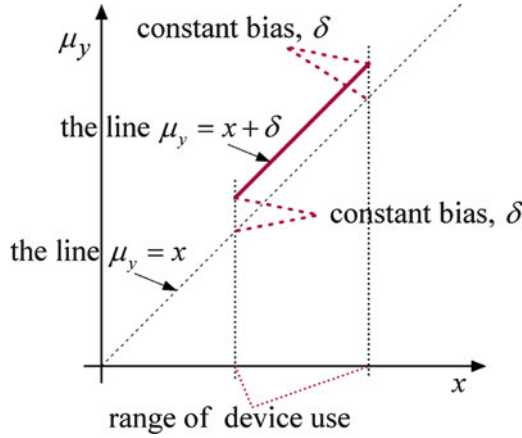


FIGURE 2.4. Measurement device “linearity” is bias constant in the measurand

Thinking in terms of model (2.1) is especially helpful when the measurand x itself is subject to variation. For instance, when parts produced on a machine have varying diameters x , one might think of model (2.1) as applying separately to each individual part diameter. But then in view of the reality of manufacturing variation, it makes sense to think of diameters as random, say with mean μ_x and standard deviation σ_x , independent of the measurement errors. This combination of assumptions then implies (for a linear device) that the mean of what is observed is

$$\mu_y = \mu_x + \delta \quad (2.3)$$

and the standard deviation of what is observed is

$$\sigma_y = \sqrt{\sigma_x^2 + \sigma_{\text{measurement}}^2} . \quad (2.4)$$

A nonzero δ is still a measurement bias, but now observed variation across parts is seen to include one component due to variation in x and another due to measurement error. The relationships (2.3) and (2.4) between the distributions of

Standard
Deviation of
Observations
Subject to
Measurement
Error

measurement error (ϵ) and item-to-item variation in the measurand (x) and the distribution of the observed measurements (y) are pictured in Fig. 2.5.

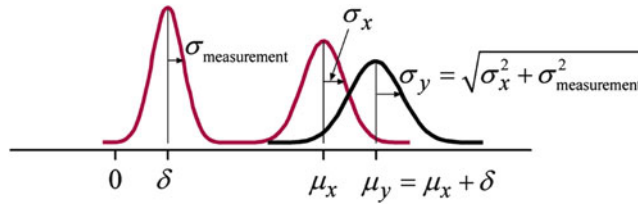


FIGURE 2.5. Random measurement error (maroon) and part variation (maroon) combine to produce observed variation (black). Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 47. © 2010 Taylor & Francis. Adapted with permission

Left to right on Fig. 2.5, the two distributions in maroon represent measurement error (with bias $\delta > 0$) and measurand variation that combine to produce variation in y represented by the distribution in black. It is the middle (maroon) distribution of x that is of fundamental interest, and the figure indicates that measurement error will both tend to shift location of that distribution and flatten it in the creation of the distribution of y . It is only this last distribution (the black one) that can be observed directly, and only when both δ and $\sigma_{\text{measurement}}$ are negligible (close to 0) are the distributions of x and y essentially the same.

Observe that Eq. (2.4) implies that

$$\sigma_x = \sqrt{\sigma_y^2 - \sigma_{\text{measurement}}^2}.$$

This suggests a way of estimating σ_x alone. If one has (single) measurements y for several parts that produce a sample standard deviation s_y and several measurements on a single part that produce a sample standard deviation s , then a plausible estimator of σ_x is

$$\hat{\sigma}_x = \sqrt{\max(0, s_y^2 - s^2)}. \quad (2.5)$$

In the next sections, we will explore the use of reasoning like this, formulas like (2.5), and the application of elementary confidence interval methods to quantify various aspects of measurement precision and bias.

Estimator of
Process or
Part Variation
Excluding
Measurement
Error

Section 2.1 Exercises

1. In a calibration study one compares outputs of a measurement device to “known” or “standard” values. What purpose does this serve?

2. **Pellet Densification.** Crocfer, Downey, Rixner, and Thomas studied the densification of Nd_2O_3 . Pellets of this material were fired at $1,400^\circ\text{C}$ for various lengths of time and the resulting densities measured (in g/cc). In this context, what are the measurand (x), y , ϵ , and δ ?
3. Suppose that in the context of problem 2, five pellets were each fired (for the same length of time), and their densities were each measured using a single device. Further, assume the measurement device has constant bias. How many measurands (xs), ys , ϵs , and δs are there in this setting?
4. In the study of problem 2, the purpose was to evaluate the *effect* of firing on pellet density. Each of the pellets fired had different original densities (that were not recorded). Does the measurement protocol described in problem 2 provide data that track what is of primary interest, i.e., does it produce a valid measure of firing effect? What additional data should have been collected? Why?
5. In the context of problem 2, the density of a single pellet was repeatedly measured five times using a single device. How many measurands (xs), ys , ϵs , and δs are there in this setting?
6. In the context of problem 2 suppose that the standard deviation of densities from repeated measurements of the same pellet with the same device is $\sqrt{2.0}$. Suppose further that the standard deviation of actual densities one pellet to the next (the standard deviation of measurands) is $\sqrt{1.4}$. What should one then expect for a standard deviation of measured density values pellet to pellet?
7. Consider the five pellets mentioned in problem 3. Density measurements similar to the following were obtained by a single operator using a single piece of equipment with a standard protocol under fixed physical circumstances:

6.5, 6.6, 4.9, 5.1, and 5.4 .

 - (a) What is the sample standard deviation of the $n = 5$ density measurements?
 - (b) In the notation of this section, which of σ_y , σ_x , or $\sigma_{\text{measurement}}$ is legitimately estimated by your sample standard deviation in (a)?
8. Again consider the five pellets of problem 3 and the five density values recorded in problem 7.
 - (a) Compute the average measured density.
 - (b) Assuming an appropriate model and using the notation of this section, what does your sample average estimate?

2.2 Elementary One- and Two-Sample Statistical Methods and Measurement

Elementary statistics courses provide basic inference methods for means and standard deviations based on one and two normal samples. (See, e.g., Sects. 6.3 and 6.4 of Vardeman and Jobe's *Basic Engineering Data Collection and Analysis*.) In this section we use elementary one- and two-sample confidence interval methods to study (in the simplest contexts possible) (1) how measurement error influences what can be learned from data and (2) how basic properties of that measurement error can be quantified. Subsequent sections will introduce more complicated data structures and statistical methods, but the basic modeling ideas and conceptual issues can most easily be understood by first addressing them without unnecessary (and tangential) complexity.

2.2.1 One-Sample Methods and Measurement Error

“Ordinary” confidence interval formulas based on a model that says that y_1, y_2, \dots, y_n are a sample from a normal distribution with mean μ and standard deviation σ are

$$\bar{y} \pm t \frac{s}{\sqrt{n}} \text{ for estimating } \mu \quad (2.6)$$

and

$$\left(s \sqrt{\frac{n-1}{\chi_{\text{upper}}^2}}, s \sqrt{\frac{n-1}{\chi_{\text{lower}}^2}} \right) \text{ for estimating } \sigma. \quad (2.7)$$

These are mathematically straightforward, but little is typically said in basic courses about the practical meaning of the parameters μ and σ . So a first point to make here is that sources of physical variation (and in particular, sources of measurement error and item-to-item variation) interact with data collection plans to give practical meaning to “ μ ” and “ σ .” This in turn governs what of practical importance can be learned from application of formulas like (2.6) and (2.7).

Two Initial Applications

Figures 2.6 and 2.7 are schematic representations of two different ways that a single “sample” of n observed values y might arise. These are:

1. Repeat measurements on a single measurand made using the same device, and

Confidence
Limits for a
Normal Mean

Confidence
Limits for a
Normal
Standard
Deviation

2. Single measurements made on multiple measurands coming from a stable process made using the same (linear) device.

Notice that henceforth we will use the language “device” as shorthand for a fixed combination of physical measurement equipment, operator identity, measurement procedure, and surrounding physical circumstances (like time of day, temperature, etc.). We will also use the shorthand “ y_i 's $\sim \text{ind}(\mu, \sigma)$ ” for the model statement that observations are independent with mean μ and standard deviation σ . And in schematics like Figs. 2.6 and 2.7, the rulers will represent generic measurement devices, the spheres generic measurands, and the factories generic processes.

This Book's
Use of the
Word “Device”

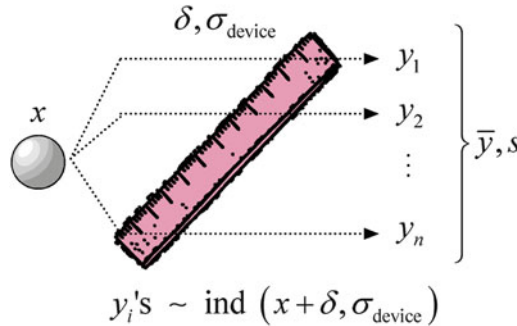


FIGURE 2.6. A single sample derived from n repeat measurements made with a fixed device on a single measurand. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 48. © 2010 Taylor & Francis. Adapted with permission

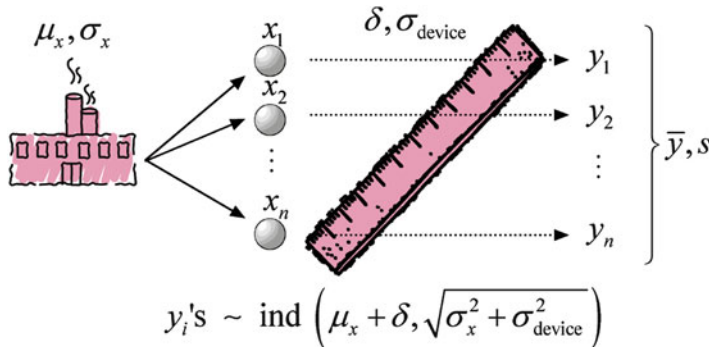


FIGURE 2.7. A single sample derived from single measurements made with a fixed (linear) device on each of n different measurands from a physically stable process. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 48. © 2010 Taylor & Francis. Adapted with permission

The case represented in Fig. 2.6 also corresponds to Fig. 2.3 (where “measurement” variation is simply that inherent in the reuse of the device to evaluate a given measurand). The case represented in Fig. 2.7 also corresponds to Fig. 2.5 (where again “measurement” variation is a variation inherent in the “device” and

now real part-to-part variation is represented by σ_x). Consider what formulas (2.6) and (2.7) provide in the two situations.

First, if as in Fig. 2.6 **n repeat measurements of a single measurand**, y_1, y_2, \dots, y_n , have sample mean \bar{y} and sample standard deviation s , applying the t confidence interval for a mean, one gets an inference for

$$x + \delta = \text{measurand plus bias} .$$

So:

1. In the event that the measurement device is known to be well-calibrated (one is sure that $\delta = 0$, there is no systematic error), the limits $\bar{y} \pm ts/\sqrt{n}$ based on $\nu = n - 1$ df are limits for x ,
2. In the event that what is being measured is a **standard for which x is known**, one may use the limits

$$(\bar{y} - x) \pm t \frac{s}{\sqrt{n}}$$

(once again based on $\nu = n - 1$ df) to estimate the device bias, δ .

Further, applying the χ^2 confidence interval for a standard deviation, one has an inference for the size of the device “noise,” σ_{device} .

Next consider what can be inferred from **single measurements made on n different measurands** y_1, y_2, \dots, y_n from a stable process with sample mean \bar{y} and sample standard deviation s as illustrated in Fig. 2.7. Here:

1. The limits $\bar{y} \pm ts/\sqrt{n}$ (for t based on $\nu = n - 1$ df) are limits for $\mu_x + \delta =$ the mean of the distribution of true values plus (the constant) bias,
2. The quantity s estimates $\sigma_y = \sqrt{\sigma_x^2 + \sigma_{\text{device}}^2}$ that we met first in display (2.4) and have noted really isn’t of fundamental interest. So there is little point in direct application of the χ^2 confidence limits (2.7) in this context.

Example 7 Measuring Concentration. Below are $n = 5$ consecutive concentration measurements made by a single analyst on a single physical specimen of material using a particular assay machine (the real units are not available, so for the sake of example, let’s call them “moles per liter,” mol/l):

$$1.0025, .9820, 1.0105, 1.0110, .9960$$

These have mean $\bar{y} = 1.0004$ mol/l and $s = .0120$ mol/l. Consulting an χ^2 table like Table A.3 using $\nu = 5 - 1 = 4$ df, we can find 95 % confidence limits for σ_{device} (the size of basic measurement variability) as

$$.0120 \sqrt{\frac{4}{11.143}} \text{ and } .0120 \sqrt{\frac{4}{.484}} \text{ i.e., } .0072 \text{ mol/l and } .0345 \text{ mol/l} .$$

(One moral here is that ordinary small sample sizes give very wide confidence limits for a standard deviation.) Consulting a t table like Table A.2 also using 4 df , we can find 95 % confidence limits for the measurand plus instrument bias ($x + \delta$) to be

$$1.0004 \pm 2.776 \frac{.0120}{\sqrt{5}}, \text{ i.e., } 1.0004 \text{ mol/l} \pm .0149 \text{ mol/l}.$$

Note that if the measurements in question were done on a standard material “known” to have actual concentration 1.0000 mol/l, these limits then correspond to limits for device bias of

$$0.0004 \text{ mol/l} \pm .0149 \text{ mol/l}.$$

Finally, suppose that subsequently samples from $n = 20$ different batches are analyzed and $\bar{y} = .9954$ and $s_y = .0300$. The 95 % t confidence limits

$$.9954 \pm 2.093 \frac{.0300}{\sqrt{20}}, \text{ i.e., } .9954 \text{ mol/l} \pm .0140 \text{ mol/l}.$$

are for $\mu_x + \delta$, the process mean plus any device bias/systematic error.

Application to a Sample Consisting of Single Measurements of a Single Measurand Made Using Multiple Devices (From a Large Population of Such Devices)

The two cases illustrated by Figs. 2.6 and 2.7 do not begin to exhaust the ways that the basic formulas (2.6) and (2.7) can be applied. We present two more applications of the one-sample formulas, beginning with an application where single measurements of a single measurand are made using multiple devices (from a large population of such devices).

There are contexts in which an organization has many “similar” measurement devices that could potentially be used to do measuring. In particular, a given piece of equipment might well be used by any of a large number of operators. Recall that we are using the word “device” to describe a particular combination of equipment, people, procedures, etc. used to produce a measurement. So, in this language, different operators with a fixed piece of equipment are different “devices.” A way to compare these devices would be to use some (say n of them) to measure a single measurand. This is illustrated in Fig. 2.8 on page 44.

In this context, a measurement is of the form

$$y = x + \epsilon,$$

where $\epsilon = \delta + \epsilon^*$, for δ the (randomly selected) bias of the device used and ϵ^* a measurement error with mean 0 and standard deviation σ_{device} (representing a repeat measurement variability for any one device). So one might write

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

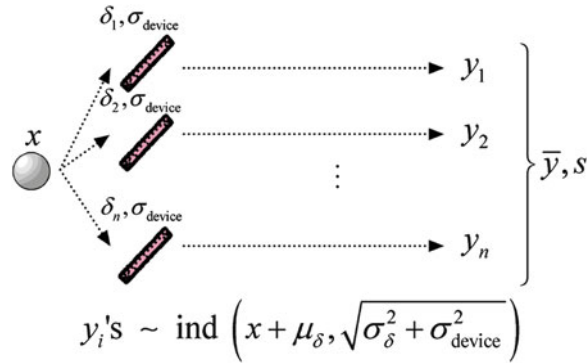


FIGURE 2.8. A single sample consisting of n single measurements of a fixed measurand made with each of n devices (from a large population of such devices with a common precision). Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 49. © 2010 Taylor & Francis. Adapted with permission

Thinking of x as fixed and δ and ϵ^* as independent random variables (δ with mean μ_δ , the average device bias, and standard deviation σ_δ measuring variability in device biases), the laws of mean and variance from elementary probability then imply that

$$\mu_y = x + \mu_\delta + 0 = x + \mu_\delta \quad (2.8)$$

and

$$\sigma_y = \sqrt{0 + \sigma_\delta^2 + \sigma_{\text{device}}^2} = \sqrt{\sigma_\delta^2 + \sigma_{\text{device}}^2} \quad (2.9)$$

as indicated on Fig. 2.8. The theoretical average measurement is the measurand plus the average bias, and the variability in measurements comes from both the variation in device biases and the intrinsic imprecision of any particular device.

In a context where a schematic like Fig. 2.8 represents a study where several operators each make a measurement on the same item using a fixed piece of equipment, the quantity

$$\sqrt{\sigma_\delta^2 + \sigma_{\text{device}}^2}$$

is a kind of overall measurement variation that is sometimes called “ $\sigma_{\text{R\&R}}$,” the first “R” standing for **repeatability** and referring to σ_{device} (a variability for fixed operator on the single item) and the second “R” standing for **reproducibility** and referring to σ_δ (a between-operator variability).

With μ_y and σ_y identified in displays (2.8) and (2.9), it is clear what the one-sample confidence limits (2.6) and (2.7) estimate in this context. Of the two, interval (2.7) for “ σ ” is probably the most important, since σ_y is interpretable in the context of an R&R study, while μ_y typically has little practical meaning. It is another question (that we will address in future sections with more complicated methods) how one might go about separating the two components of σ_y to assess the relative sizes of repeatability and reproducibility variation.

Repeatability
and
Reproducibility

Application to a Sample Consisting of Differences in Measurements on Multiple Measurands Made Using Two Linear Devices

Another way to create a single sample of numbers is this. With two devices available and n different measurands, one might measure each once with both devices and create n differences between device 1 and device 2 measurements. This is a way of potentially comparing the two devices and is illustrated in Fig. 2.9.

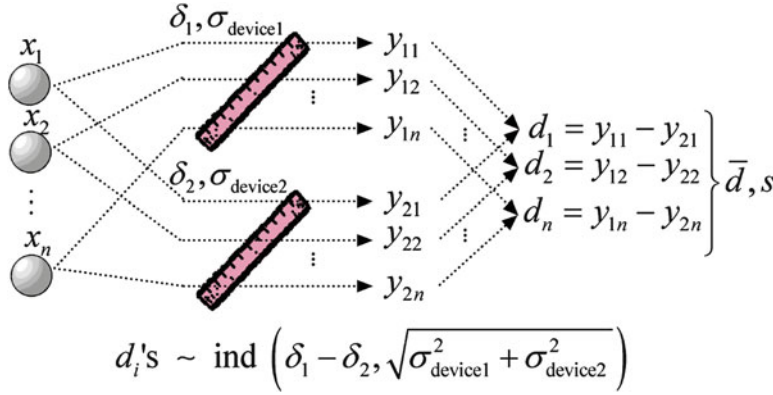


FIGURE 2.9. A single sample consisting of n differences of single measurements of n measurands made using two devices (assuming device linearity). Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 49. © 2010 Taylor & Francis. Adapted with permission

In this context, a difference is of the form

$$d = y_1 - y_2 = (x + \epsilon_1) - (x + \epsilon_2) = \epsilon_1 - \epsilon_2$$

and (again applying the laws of mean and variance from elementary probability) it follows that

$$\mu_d = \delta_1 - \delta_2 \quad \text{and} \quad \sigma_d = \sqrt{\sigma_{\text{device1}}^2 + \sigma_{\text{device2}}^2}$$

as indicated on Fig. 2.9. So applying the t interval for a mean (2.6), the limits

$$\bar{d} \pm t \frac{s}{\sqrt{n}} \quad (2.10)$$

provide a way to estimate $\delta_1 - \delta_2$, the difference in device biases.

Confidence
Limits for a
Mean
Difference

2.2.2 Two-Sample Methods and Measurement Error

Parallel to the one-sample formulas are the two-sample formulas of elementary statistics. These are based on a model that says that

$$y_{11}, y_{12}, \dots, y_{1n_1} \quad \text{and} \quad y_{21}, y_{22}, \dots, y_{2n_2}$$

are independent samples from normal distributions with respective means μ_1 and μ_2 and respective standard deviations σ_1 and σ_2 . In this context, the so-called Satterthwaite approximation gives limits

$$\bar{y}_1 - \bar{y}_2 \pm \hat{t} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ for estimating } \mu_1 - \mu_2, \quad (2.11)$$

where appropriate “approximate degrees of freedom” for \hat{t} are

$$\hat{\nu} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{(n_1 - 1)n_1^2} + \frac{s_2^4}{(n_2 - 1)n_2^2}}. \quad (2.12)$$

(This method is one that you may not have seen in an elementary statistics course, where often only methods valid when one assumes that $\sigma_1 = \sigma_2$ are presented. We use this method not only because it requires less in terms of model assumptions than the more common formula but also because we will have other uses for the Satterthwaite idea in this chapter, so it might as well be met first in this simple context.) It turns out that $\min((n_1 - 1), (n_2 - 1)) \leq \hat{\nu}$, so that a simple conservative version of this method uses degrees of freedom

$$\hat{\nu}^* = \min((n_1 - 1), (n_2 - 1)). \quad (2.13)$$

Further, in the two-sample context, there are elementary confidence limits

$$\frac{s_1}{s_2} \cdot \frac{1}{\sqrt{F_{(n_1-1), (n_2-1), \text{upper}}}}} \text{ and } \frac{s_1}{s_2} \cdot \frac{1}{\sqrt{F_{(n_1-1), (n_2-1), \text{lower}}}}} \text{ for } \frac{\sigma_1}{\sigma_2} \quad (2.14)$$

(and be reminded that $F_{(n_1-1), (n_2-1), \text{lower}} = 1/F_{(n_2-1), (n_1-1), \text{upper}}$ so that standard F tables giving only upper percentage points can be employed).

Application to Two Samples Consisting of Repeat Measurements of a Single Measurand Made Using Two Different Devices

One way to create “two samples” of measurements is to measure the same item repeatedly with two different devices. This possibility is illustrated in Fig. 2.10.

Direct application of the two-sample confidence interval formulas here shows that the two-sample Satterthwaite approximate t interval (2.11) provides limits for

$$\mu_1 - \mu_2 = (x + \delta_1) - (x + \delta_2) = \delta_1 - \delta_2$$

(the difference in device biases), while the F interval (2.14) provides a way of comparing device standard deviations σ_{device1} and σ_{device2} through direct estimation of

$$\frac{\sigma_{\text{device1}}}{\sigma_{\text{device2}}}.$$

Confidence
Limits for a
Difference in
Normal Means

Satterthwaite
Degrees of
Freedom for
Formula (2.11)

Conservative
Simplification
of Formula
(2.12)

Confidence
Limits for a
Ratio of Two
Normal
Standard
Deviations

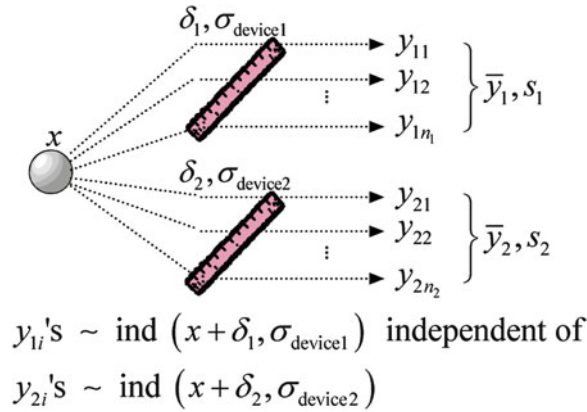


FIGURE 2.10. Two samples consisting of n_1 and n_2 measurements of a single measurand with two devices. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 49. © 2010 Taylor & Francis. Adapted with permission

This data collection plan thus provides for straightforward comparison of the basic characteristics of the two devices.

Example 8 Measuring Polystyrene “Packing Peanut” Size. In an in-class measurement exercise, two students used the same caliper to measure the “size” of a single polystyrene “packing peanut” according to a class-standard measurement protocol. Some summary statistics from their work follow.

<i>Student 1</i>	<i>Student 2</i>
$n_1 = 4$	$n_2 = 6$
$\bar{y}_1 = 1.42 \text{ cm}$	$\bar{y}_2 = 1.44 \text{ cm}$
$s_1 = .20 \text{ cm}$	$s_2 = .40 \text{ cm}$

In this context, the difference in the two measurement “devices” is the difference in “operators” making the measurements. Consider quantifying how this difference affects measurement.

To begin, note that from formula (2.12)

$$\hat{\nu} = \frac{\left(\frac{(.20)^2}{4} + \frac{(.40)^2}{6} \right)^2}{\frac{(.20)^4}{(4-1)(4)^2} + \frac{(.40)^4}{(6-1)(6)^2}} \approx 7.7,$$

or using the more conservative display (2.13), one gets

$$\hat{\nu}^* = \min((4-1), (6-1)) = 3.$$

So (rounding the first of these down to 7) one should use either 7 or 3 degrees of freedom with formula (2.11). For the sake of example, using $\hat{\nu}^* = 3$ degrees

of freedom, the upper 2.5 % point of the t distribution with 3 df is 3.182. So 95 % confidence limits for the difference in biases for the two operators using this caliper are

$$1.42 - 1.44 \pm 3.182 \sqrt{\frac{(.20)^2}{4} + \frac{(.40)^2}{6}},$$

i.e.,

$$-.02 \text{ cm} \pm .61 \text{ cm}.$$

The apparent difference in biases is small in comparison with the imprecision associated with that difference.

Then, since the upper 2.5 % point of the $F_{3,5}$ distribution is 7.764 and the upper 2.5 % point of the $F_{5,3}$ distribution is 14.885, 95 % confidence limits for the ratio of standard deviations of measurement for the two operators are

$$\frac{.20}{.40} \cdot \frac{1}{\sqrt{7.764}} \text{ and } \frac{.20}{.40} \cdot \sqrt{14.885},$$

i.e.,

$$.19 \text{ and } 1.93.$$

Since this interval covers values both smaller and larger than 1.00, there is in the limited information available here no clear indicator of which of these students is the most consistent in his or her use of the caliper in this measuring task.

Comparing
Devices When
Measurement
is Destructive

Application to Two Samples Consisting of Single Measurements Made with Two Devices on Multiple Measurands from a Stable Process (Only One Device Being Used for a Given Measurand)

There are quality assurance contexts in which measurement is **destructive** (and cannot be repeated for a single measurand) and nevertheless one needs to somehow compare two different devices. In such situations, the only thing that can be done is to take items from some large pool of items or from some stable process and (probably after randomly assigning them one at a time to one or the other of the devices) measure them and try to make comparisons based on the resulting samples. This possibility is illustrated in Fig. 2.11. This is a schematic for two samples consisting of single measurements made with two devices on multiple measurands from a stable process (only one device used for a given measurand).

Direct application of the two-sample Satterthwaite approximate t interval (2.11) provides limits for

$$\mu_1 - \mu_2 = (\mu_x + \delta_1) - (\mu_x + \delta_2) = \delta_1 - \delta_2$$

(the difference in device biases). So, even in contexts where measurement is destructive, it is possible to compare device biases. It's worth contemplating, however, the difference between the present scenario and the immediately preceding one (represented by Fig. 2.10).

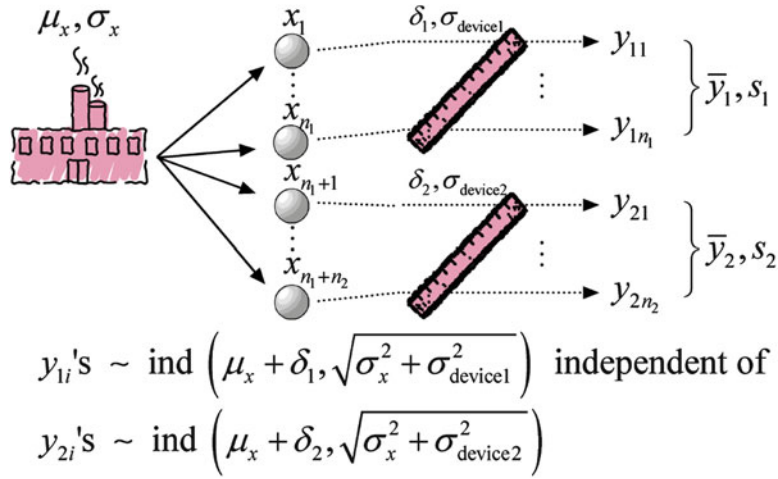


FIGURE 2.11. Two samples consisting of single measurements made on $n_1 + n_2$ measurands from a stable process, n_1 with device 1 and n_2 with device 2. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 50. © 2010 Taylor & Francis. Adapted with permission

The measurements y in Fig. 2.10 on page 47 are less variable than are the measurements y here in Fig. 2.11. This is evident in the standard deviations shown on the figures and follows from the fact that in the present case (unlike the previous one), measurements are affected by unit-to-unit/measurand-to-measurand variation. So all else being equal, one should expect limits (2.11) applied in the present context to be wider/less informative than when applied to data collected as in the last application. That should be in accord with intuition. One should expect to be able to learn more useful to comparing devices when the same item(s) can be remeasured than when it (they) cannot be remeasured.

Notice that if the F limits (2.14) are applied here, one winds up with only an indirect comparison of σ_{device1} and σ_{device2} , since all that can be easily estimated (using the limits (2.14)) is the ratio

$$\frac{\sqrt{\sigma_x^2 + \sigma_{\text{device1}}^2}}{\sqrt{\sigma_x^2 + \sigma_{\text{device2}}^2}}$$

and NOT the (more interesting) ratio $\sigma_{\text{device1}}/\sigma_{\text{device2}}$.

Application to Two Samples Consisting of Repeat Measurements Made with One Device on Two Measurands

A basic activity of quality assurance is the comparison of nominally identical items. Accordingly, another way to create two samples is to make repeated measurements on two measurands with a single device. This is illustrated in Fig. 2.12 on page 50.

In this context,

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

so that application of the two-sample Satterthwaite approximate t interval (2.11) provides limits for the *difference in the measurands* and a direct way of comparing the measurands. The device bias affects both samples in the same way and “washes out” when one takes a difference. (This, of course, assumes that the device is linear, i.e., that the bias is constant.)

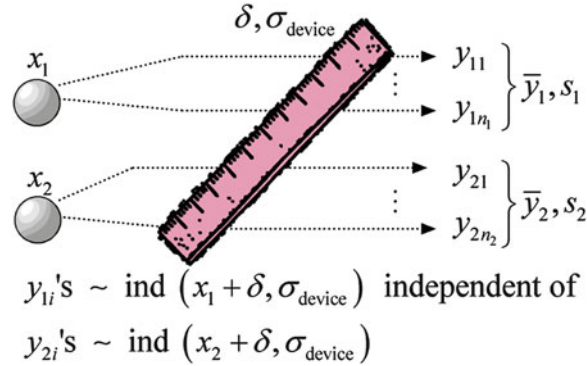


FIGURE 2.12. Two samples consisting of repeat measurements made with one device on two measurands. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 50. © 2010 Taylor & Francis. Adapted with permission

Application to Two Samples Consisting of Single Measurements Made Using a Single Linear Device on Multiple Measurands Produced by Two Stable Processes

Another basic activity of quality assurance is the comparison of nominally identical *processes*. Accordingly, another way to create two samples is to make single measurements on samples of measurands produced by two processes. This possibility is illustrated in Fig. 2.13.

In this context,

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

so that application of the two-sample Satterthwaite approximate t interval (2.11) provides limits for the *difference in the process mean measurands* and hence a direct way of comparing the processes. Again, the device bias affects both samples in the same way and “washes out” when one takes a difference (still assuming that the device is linear, i.e., that the bias is constant).

If the F limits (2.14) are applied here, one winds up with only an indirect comparison of σ_{x1} and σ_{x2} , since what can be easily estimated is the ratio

$$\frac{\sqrt{\sigma_{x1}^2 + \sigma_{\text{device}}^2}}{\sqrt{\sigma_{x2}^2 + \sigma_{\text{device}}^2}}$$

and not the practically more interesting σ_{x1}/σ_{x2} .

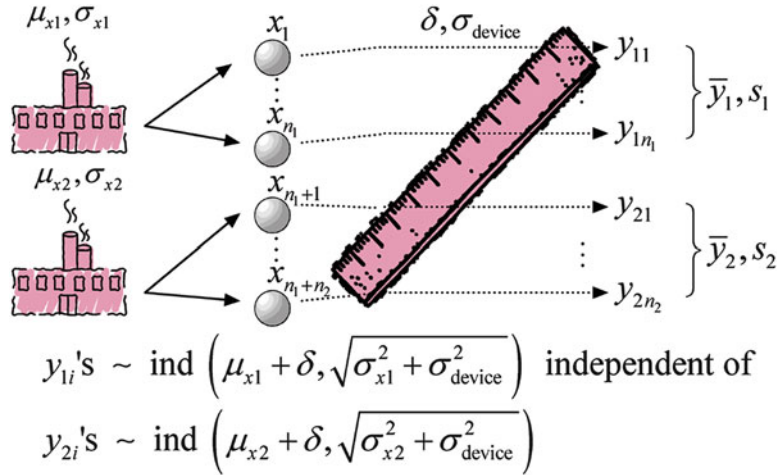


FIGURE 2.13. Two samples consisting of single measurements made using a single device on multiple measurands produced by two stable processes. Modified from “Elementary Statistical Methods and Measurement Error” by S.B. Vardeman et al., 2010, *The American Statistician*, 64(1), 50. © 2010 Taylor & Francis. Adapted with permission

Section 2.2 Exercises

- Consider again the **Pellet Densification** case of problem 7 in Sect. 2.1. Suppose the five data values 6.5, 6.6, 4.9, 5.1, and 5.4 were measured densities for a single pellet produced by five different operators using the same piece of measuring equipment (or by the same operator using five different pieces of equipment—the two scenarios are conceptually handled in the same way). Use the notation of this section (x , δ , μ_δ , σ_δ , and σ_{device}) below.
 - What does the sample average of these five data values estimate?
 - What does the sample standard deviation of these five data values estimate?
 - Which of the two estimates in (a) and (b) is probably more important? Why?
- Return again to the context of problem 7 of Sect. 2.1. Suppose the original set of five data values 6.5, 6.6, 4.9, 5.1, and 5.4 was obtained from five different pellets by operator 1 using piece of equipment 1. Using a second piece of equipment, operator 1 recorded densities 6.6, 5.7, 5.9, 6.2, and 6.3 for the same five pellets. So, for pellet 1, “device 1” produced measurement 6.5 and “device 2” produced 6.6; for pellet 2, “device 1” produced measurement 6.6 and “device 2” produced 5.7 and so on.

- (a) Give the five differences in measured densities (device 1 minus device 2). Calculate the sample average difference. What does this estimate? (Hint: Consider δs .)
 - (b) Calculate the sample standard deviation of the five differences (device 1 minus device 2). What does this estimate? (Hint: Consider the σ_{device} .)
 - (c) Find 90 % confidence limits for the average difference in measurements from the two devices.
3. Suppose the two sets of five measurements referred to in problems 1 and 2 actually came from one pellet, i.e., operator 1 measured the same pellet five times with piece of equipment 1 and then measured that same pellet five times with piece of equipment 2.
 - (a) Find a 95 % confidence interval for the ratio of the two device standard deviations ($\sigma_{\text{device1}}/\sigma_{\text{device2}}$). What do your limits indicate about the consistency of measurements from device 1 compared to that of measurements from device 2?
 - (b) Find a 95 % two-sample Satterthwaite approximate t interval for the difference in the two device averages (device 1 minus device 2). If your interval were to include 0, what would you conclude regarding device biases 1 and 2?
4. Consider now the same ten data values referred to in problems 2 and 3, but a different data collection plan. Suppose the first five data values were measurements on five different pellets by operator 1 using piece of equipment 1 and the second set of data values was for another set of pellets by operator 1 using piece of equipment 2. Assume both sets of pellets came from the same physically stable process.
 - (a) What does the sample standard deviation from the first set of five data values estimate?
 - (b) What does the sample standard deviation from the second set of five data values estimate?
 - (c) What does the difference in the two-sample average densities estimate?
5. Reflect on problems 3 and 4. Which data-taking approach is better for estimating the difference in device biases? Why?
6. In the same **Pellet Densification** context considered in problems 1 through 5, suppose one pellet was measured five times by operator 1 and a different pellet was measured five times by operator 1 (the same physical equipment was used for the entire set of ten observations). What is estimated by the difference in the two sample averages?

7. Once again in the context of problems 1 through 6, suppose the first five data values were measurements on five different pellets made by operator 1 using piece of equipment 1 and the second five were measurements of a different set of pellets by operator 1 using piece of equipment 1. Assume the two sets of pellets come from different firing methods (method 1 and method 2). Assume the two firing processes are physically stable.
 - (a) Find the two-sided 95 % two-sample Satterthwaite approximate t interval for the difference in the process mean measurands (method 1 minus method 2).
 - (b) In words, what does the interval in (a) estimate? In symbols, what does the interval in (a) estimate?
 - (c) With this approach to data taking, can either device bias be estimated directly? Why or why not?
8. Still in the context of problems 1 through 7, density measurements 6.5, 6.6, 4.9, 5.1, and 5.4 were obtained for five different pellets by a single operator using a single piece of measuring equipment under a standard protocol and fixed physical circumstances. Use the t confidence interval for a mean, and give 95 % confidence limits for the mean of the distribution of true densities plus measurement bias.
9. Suppose the five measurements in problem 8 are repeat measurements from only one pellet, not from five different pellets.
 - (a) Use the χ^2 confidence limits for a standard deviation (from elementary statistics), and give a 95 % confidence interval for $\sigma_{\text{measurement}}$.
 - (b) Use the t confidence interval formula for a mean from elementary statistics and give 95 % confidence limits for the (single) true pellet density plus measurement bias.

2.3 Some Intermediate Statistical Methods and Measurement

Through reference to familiar elementary one- and two-sample methods of statistical inference, Sect. 2.2 illustrated the basic insight that:

How sources of physical variation interact with a data collection plan governs what of practical importance can be learned from a data set, and in particular, how measurement error is reflected in the data set.

In this section we consider some computationally more complicated statistical methods and what they provide in terms of quantification of the impact of measurement variation on quality assurance data.

2.3.1 A Simple Method for Separating Process and Measurement Variation

In Sect. 2.1 we essentially observed that:

1. Repeated measurement of a single measurand with a single device allows one to estimate device variability,
2. Single measurements made on multiple measurands from a stable process allow one to estimate a combination of process and measurement variability,

and remarked that these facts suggest formula (2.5) as a way to estimate a process standard deviation alone. Our first objective in this section is to elaborate a bit on this thinking.

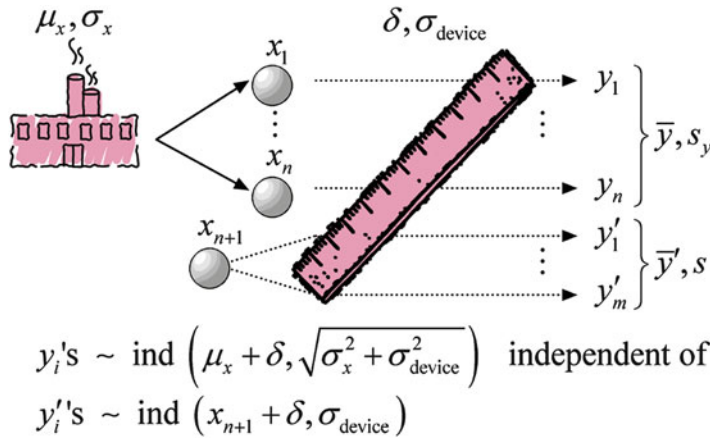


FIGURE 2.14. Schematic of a data collection plan that allows evaluation of σ_x without inflation by measurement variation

Figure 2.14 is a schematic of a data collection plan that combines elements 1 and 2 above. Here we use the notation y for the single measurements on n items from the process and the notation y' for the m repeat measurements on a single measurand. The sample standard deviation of the y s, s_y , is a natural empirical approximation for $\sigma_y = \sqrt{\sigma_x^2 + \sigma_{\text{device}}^2}$ and the sample standard deviation of the y' 's, s , is a natural empirical approximation for σ_{device} . That suggests that one estimates the process standard deviation with

Estimator of
Process
Standard
Deviation not
Inflated by
Measurement
Variability

$$\hat{\sigma}_x = \sqrt{\max(0, s_y^2 - s^2)} \quad (2.15)$$

as indicated in display (2.5). (The maximum of 0 and $s_y^2 - s^2$ under the root is there simply to ensure that one is not trying to take the square root of a negative number in the rare case that s exceeds s_y .) $\hat{\sigma}_x$ is not only a sensible single-number estimate of σ_x , but can also be used to make approximate confidence limits for the

process standard deviation. The so-called Satterthwaite approximation suggests that one uses

$$\hat{\sigma}_x \sqrt{\frac{\hat{\nu}}{\chi_{\text{upper}}^2}} \quad \text{and} \quad \hat{\sigma}_x \sqrt{\frac{\hat{\nu}}{\chi_{\text{lower}}^2}} \quad (2.16)$$

as limits for σ_x , where appropriate approximate degrees of freedom $\hat{\nu}$ to be used finding χ^2 percentage points are

$$\hat{\nu} = \frac{\hat{\sigma}_x^4}{\frac{s_y^4}{n-1} + \frac{s^4}{m-1}} \quad (2.17)$$

Example 9 (Example 7 Revisited.) In Example 7, we considered $m = 5$ measurements made by a single analyst on a single physical specimen of material using a particular assay machine that produced $s = .0120$ mol/l. Subsequently, specimens from $n = 20$ different batches were analyzed and $s_y = .0300$ mol/l. Using formula (2.15), an estimate of real process standard deviation uninflated by measurement variation is

$$\hat{\sigma}_x = \sqrt{\max\left(0, (.0300)^2 - (.0120)^2\right)} = .0275 \text{ mol/l}$$

and this value can be used to make confidence limits. By formula (2.17) approximate degrees of freedom are

$$\hat{\nu} = \frac{(.0275)^4}{\frac{(.0300)^4}{19} + \frac{(.0120)^4}{4}} = 11.96 .$$

So rounding down to $\hat{\nu} = 11$, since the upper 2.5 % point of the χ_{11}^2 distribution is 21.920 and the lower 2.5 % point is 3.816, by formula (2.16) approximate 95 % confidence limits for the real process standard deviation (σ_x) are

$$.0275 \sqrt{\frac{11}{21.920}} \quad \text{and} \quad .0275 \sqrt{\frac{11}{3.816}} ,$$

i.e.,

$$.0195 \text{ mol/l} \quad \text{and} \quad .0467 \text{ mol/l} .$$

Satterthwaite
Approximate
Confidence
Limits for a
Process
Standard
Deviation

Satterthwaite
Approximate df
for Use With
Limits (2.16)

2.3.2 One-Way Random Effects Models and Associated Inference

One of the basic models of intermediate statistical methods is the so-called one-way random effects model for I samples of observations

$$\begin{array}{c} y_{11}, y_{12}, \dots, y_{1n_1} \\ y_{21}, y_{22}, \dots, y_{2n_2} \\ \vdots \\ y_{I1}, y_{I2}, \dots, y_{In_I} \end{array}$$

This model says that the observations may be thought of as

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where the ϵ_{ij} are independent normal random variables with mean 0 and standard deviation σ , while the I values μ_i are independent normal random variables with mean μ and standard deviation σ_μ (independent of the ϵ s). (One can think of I means μ_i drawn at random from a normal distribution of μ_i s and subsequently observations y generated from I different normal populations with those means and a common standard deviation.) In this model, the three parameters are σ (the “within-group” standard deviation), σ_μ (the “between-group” standard deviation), and μ (the overall mean). The squares of the standard deviations are called “variance components” since for any particular observation, the laws of expectation and variance imply that

$$\mu_y = \mu + 0 = \mu \text{ and } \sigma_y^2 = \sigma_\mu^2 + \sigma^2$$

(i.e., σ_μ^2 and σ^2 are components of the variance of y).

Two quality assurance contexts where this model can be helpful are where

1. Multiple measurands from a stable process are each measured multiple times using the same device,
2. A single measurand is measured multiple times using multiple devices.

These two scenarios and the accompanying parameter values are illustrated in Figs. 2.15 and 2.16.

There are well-established (but not altogether simple) methods of inference associated with the one-way random effects model that can be applied to make confidence intervals for the model parameters (and inferences of practical interest in metrological applications). Some of these are based on so-called ANOVA methods and the one-way ANOVA identity that says that with

$$\bar{y}_i = \frac{1}{n_i} \sum_j y_{ij}, n = \sum_i n_i, \text{ and } \bar{y} = \frac{1}{n} \sum_i n_i \bar{y}_i,$$

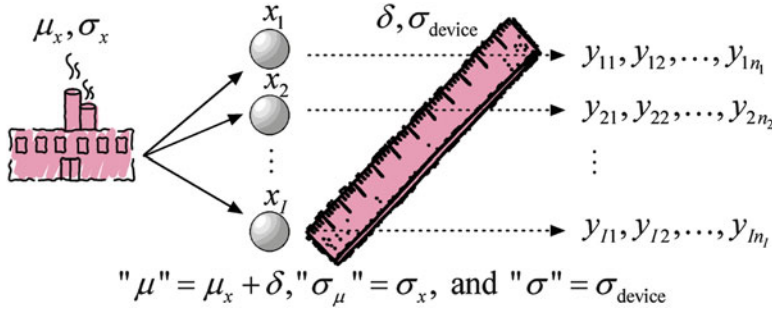


FIGURE 2.15. Multiple measurands from a stable process each measured multiple times using the same device

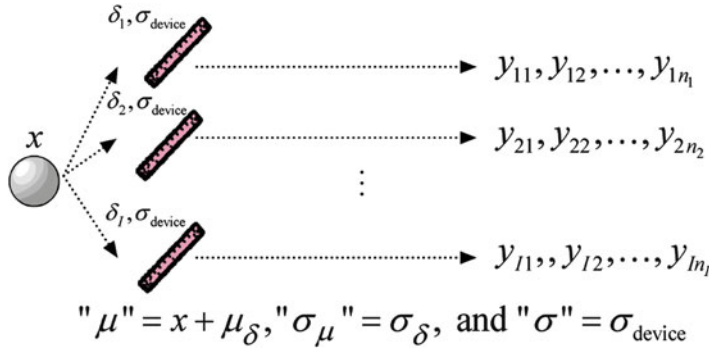


FIGURE 2.16. A single measurand measured multiple times using multiple devices

it is the case that

$$\sum_{i,j} (y_{ij} - \bar{y})^2 = \sum_i n_i (\bar{y}_i - \bar{y})^2 + \sum_{i,j} (y_{ij} - \bar{y}_i)^2 \quad (2.18)$$

or in shorthand “sum of squares” notation

$$SSTot = SSTr + SSE \quad (2.19)$$

$SSTot$ is a measure of overall raw variability in the whole data set. $SSTot$ is $n-1$ times the overall sample variance computed ignoring the boundaries between samples. SSE is a measure of variability left unaccounted for after taking account of the sample boundaries and is a multiple of a weighted average of the I sample variances. $SSTr$ is a measure of variation in the sample means \bar{y}_i , and is most simply thought of as the difference $SSTot - SSE$. The “sums of squares” SSE and $SSTr$ have respective associated degrees of freedom $n - I$ and $I - 1$. The ratios of sums of squares to their degrees of freedom are called “mean squares” and symbolized as

$$MSE = \frac{SSE}{n - I} \text{ and } MSTr = \frac{SSTr}{I - 1}. \quad (2.20)$$

One-Way
ANOVA
Identity

One-Way
ANOVA
Identity in Sum
of Squares
Notation

Confidence limits for the parameter σ^2 of the one-way random effects model can be built on the error mean square. A single-number estimate of σ is

$$\hat{\sigma} = \sqrt{MSE} \quad (2.21)$$

and confidence limits for σ are

$$\hat{\sigma} \sqrt{\frac{n-I}{\chi_{\text{upper}}^2}} \quad \text{and} \quad \hat{\sigma} \sqrt{\frac{n-I}{\chi_{\text{lower}}^2}} \quad (2.22)$$

where the appropriate degrees of freedom are $\nu = n - I$. Further, in the case that all n_i s are the same, i.e., $n_i = m$ for all i , the Satterthwaite approximation can be used to make fairly simple approximate confidence limits for σ_μ . That is, a single-number estimator of σ_μ is

$$\hat{\sigma}_\mu = \sqrt{\frac{1}{m} \max(0, MSTr - MSE)}, \quad (2.23)$$

and with approximate degrees of freedom

$$\hat{\nu} = \frac{m^2 \cdot \hat{\sigma}_\mu^4}{\frac{MSTr^2}{I-1} + \frac{MSE^2}{n-I}} \quad (2.24)$$

approximate confidence limits for σ_μ are

$$\hat{\sigma}_\mu \sqrt{\frac{\hat{\nu}}{\chi_{\text{upper}}^2}} \quad \text{and} \quad \hat{\sigma}_\mu \sqrt{\frac{\hat{\nu}}{\chi_{\text{lower}}^2}}. \quad (2.25)$$

Operationally, the mean squares implicitly defined in displays (2.18) through (2.20) are rarely computed “by hand.” And given that statistical software is going to be used, rather than employ the methods represented by formulas (2.21) through (2.25), more efficient methods of confidence interval estimation can be used. High-quality statistical software (like the open-source command line-driven R package or the commercial menu-driven JMP package) implements the best known methods of estimation of the parameters σ , σ_μ , and μ (based not on ANOVA methods, but instead on computationally more difficult REML methods) and prints out confidence limits directly.

Example 10 Part Hardness. Below are $m = 2$ hardness values (in mm) measured on each of $I = 9$ steel parts by a single operator at a farm implement manufacturer.

Part	1	2	3	4	5	6	7	8	9
	3.30	3.20	3.20	3.25	3.25	3.30	3.15	3.25	3.25
	3.30	3.25	3.30	3.30	3.30	3.30	3.20	3.20	3.30

This is a scenario of the type illustrated in Fig. 2.15. Either working “by hand” with formulas (2.18) through (2.20) or reading directly off a report from a statistical package,

$$MSE = .001389 \text{ and } MSTR = .003368$$

So using formulas (2.21) and (2.22) (here $n = mI = 18$ so that error degrees of freedom are $n - I = 18 - 9 = 9$), 95 % confidence limits for σ_{device} ($= \sigma$ here) are

$$\sqrt{.001389} \sqrt{\frac{9}{19.023}} \text{ and } \sqrt{.001389} \sqrt{\frac{9}{2.700}},$$

i.e.,

$$.026 \text{ mm and } .068 \text{ mm}.$$

Further, using formulas (2.23) through (2.25), Satterthwaite degrees of freedom for $\hat{\sigma}_\mu$ are

$$\hat{\nu} = \frac{(2^2) \left(\frac{1}{2} (.003368 - .001389) \right)^2}{\frac{(.003368)^2}{9-1} + \frac{(.001389)^2}{18-9}} \approx 2.4$$

and rounding down to 2 degrees of freedom, approximate 95 % confidence limits for σ_x ($= \sigma_\mu$ here) are

$$\sqrt{\frac{1}{2} (.003368 - .001389)} \sqrt{\frac{2}{7.378}} \text{ and } \sqrt{\frac{1}{2} (.003368 - .001389)} \sqrt{\frac{2}{.051}},$$

i.e.,

$$.016 \text{ mm and } .197 \text{ mm}.$$

The JMP package (using REML methods instead of the Satterthwaite approximation based on ANOVA mean squares) produces limits for σ_x :

$$0 \text{ mm and } \sqrt{.0027603} = .053 \text{ mm}.$$

These more reliable limits at least confirm that the simpler methods “get into the right ballpark” in this example.

What is clear from this analysis is that this is a case where part-to-part variation in hardness (measured by σ_x) is small enough and poorly determined enough in comparison with basic measurement noise (measured by σ_{device} estimated as $.03726 = \sqrt{.001389}$) that it is impossible to really tell its size.

Example 11 Paper Weighing. Below are $m = 3$ measurements of the weight (in g) of a single 20 cm \times 20 cm piece of 20 lb bond paper made by each of $I = 5$ different technicians using a single balance.

Operator	1	2	3	4	5
	3.481	3.448	3.485	3.475	3.472
	3.477	3.472	3.464	3.472	3.470
	3.470	3.470	3.477	3.473	3.474

This is a scenario of the type illustrated in Fig. 2.16 and further illustrates the concepts of repeatability (fixed device) variation and reproducibility (here, device-to-device, i.e., operator-to-operator) variation first discussed on page 44. Use of the JMP statistical package (and REML estimation) with these data produces 95 % confidence limits for the two standard deviations σ_δ ($= \sigma_\mu$ here) and σ_{device} ($= \sigma$ here). These place

$$0 < \sigma_\delta < \sqrt{4.5 \times 10^{-5}} = .0067 \text{ g}$$

and

$$.0057 \text{ g} = \sqrt{3.2 \times 10^{-5}} < \sigma_{device} < \sqrt{.0002014} = .0142 \text{ g}$$

with 95 % confidence. This is a case where repeatability variation is clearly larger than reproducibility (operator-to-operator) variation in weight measuring. If one doesn't like the overall size of measurement variation, it appears that some fundamental change in equipment or how it is used will be required. Simple training of the operators aimed at making how they use the equipment more uniform (and reduction of differences between their biases) has far less potential to improve measurement precision.

Section 2.3 Exercises

1. **Fiber Angle.** Grunig, Hamdorf, Herman, and Potthoff studied a carpet-like product. Fiber angles (to the backing) were of interest. Operator 1 obtained the values 19, 20, 20, and 23 (in degrees) from four measurements of fiber angle for a single specimen. This same operator then measured fiber angles once each for three other specimens of the “carpet” and obtained the values 20, 15, and 23.
 - (a) Using the methods of this section, give an estimate of the specimen-to-specimen standard deviation of fiber angle.
 - (b) Give the appropriate “approximate degrees of freedom” associated with your estimate from (a). Then find a 95 % confidence interval for the specimen-to-specimen fiber angle standard deviation.
2. Continue with the **Fiber Angle** case of problem 1. Operator 2 obtained the fiber angle measurements 20, 25, 17, and 22 from the first specimen mentioned in problem 1 and operator 3 obtained the values 20, 19, 15, and 16. (Fiber angle for the same specimen was measured four times by each of the three operators.) As before, all measurements were in degrees. The data summaries below are from the use of the JMP statistical package with these $n = 12$ measurements of fiber angle for this specimen. Use them to answer (a) through (c). (The estimates and confidence intervals in the

second table are for variances, not standard deviations. You will need to take square roots to get inferences for standard deviations.)

ANOVA table

Source	SS	df	MS
Operator	28.66	2	14.33
Error	60	9	6.66
Total	88.66	11	

REML variance component analysis

Random effect	VarComponent	95 % lower	95 % upper
Operator	1.92	-5.27	9.11
Error	6.66	3.15	22.22

- (a) Give an appropriate single-number estimate of $\sigma_{\text{repeatability}}$. Determine 95 % confidence limits for device (repeatability) standard deviation, $\sigma_{\text{repeatability}}$.
 - (b) From the computer output, give the appropriate estimate of $\sigma_{\text{reproducibility}}$. Give 95 % confidence limits for $\sigma_{\text{reproducibility}}$.
 - (c) Based on your answers to (a) and (b), where would you focus measurement improvement efforts?
3. Continuing with the **Fiber Angle** case, in addition to the repeat measurements 19, 20, 20, and 23 made by operator 1 on specimen 1, this person also measured angles on two other specimens. Four angle measurements on specimen 2 were 15, 17, 20, and 20, and four angle measurements on specimen 3 were 23, 20, 22, and 20. The data summaries below are from the use of the JMP statistical package with these $n = 12$ measurements for these three specimens. Use them to answer (a) through (c). (The estimates and confidence intervals in the second table are for variances, not standard deviations. You will need to take square roots to get inferences for standard deviations.)

ANOVA table

Source	SS	df	MS
Specimen	23.17	2	11.58
Error	33.75	9	3.75
Total	56.92	11	

REML variance component analysis

Random effect	VarComponent	95 % lower	95 % upper
Specimen	1.96	3.78	7.69
Error	3.75	1.77	12.5

- (a) Give an appropriate single-number estimate of σ_{device} . Determine 95 % confidence limits for device variation, σ_{device} .

- (b) From the computer output, give an appropriate estimate of σ_x . Give 95 % confidence limits for σ_x .
- (c) Based on your answers to (a) and (b), does it seem possible to determine fiber angle for a fixed specimen with acceptable precision? (Hint: Consider the sizes of the estimated σ_{device} and σ_x .)

2.4 Gauge R&R Studies

We have twice made some discussion of “gauge R&R,” first on page 44 in the context of comparison of two operators and then in Example 11, where three operators were involved. In both cases, only a single part (or measurand) was considered. In a typical industrial gauge R&R study, each of J operators uses the same gauge or measurement system to measure each of I parts (common to all operators) a total of m different times. Variation in measurement typical of that seen in the m measurements for a particular operator on a particular part is called the **repeatability** variation of the gauge. Variation which can be attributed to differences between the J operators is called **reproducibility** variation of the measurement system.

This section considers the analysis of such full-blown gauge R&R studies involving a total of mIJ measurements. We begin with a discussion of the two-way random effects model that is commonly used to support analyses of gauge R&R data. Then primarily for ease of exposition and making connections to common analyses of gauge R&R studies, we discuss some range-based statistical methods. Finally, we provide what are really superior analyses, based on ANOVA calculations.

2.4.1 Two-Way Random Effects Models and Gauge R&R Studies

Typical industrial gauge R&R data are conveniently thought of as laid out in the cells of a table with I rows corresponding to parts and J columns corresponding to operators.

Example 12 Gauge R&R for a 1-Inch Micrometer Caliper. Heyde, Kuebrick, and Swanson conducted a gauge R&R study on a certain micrometer caliper as part of a class project. Table 2.1 shows data that the $J = 3$ (student) operators obtained, each making $m = 3$ measurements of the heights of $I = 10$ steel punches.

Notice that even for a given punch/student combination, measured heights are not exactly the same. Further, it is possible to verify that averaging the 30 measurements made by student 1, a mean of about .49853 in is obtained, while corresponding means for students 2 and 3 are, respectively, about .49813 in and .49840 in. Student 1 may tend to measure slightly higher than students 2 and

TABLE 2.1. Measured heights of ten steel punches in 10^{-3} in

Punch	Student 1	Student 2	Student 3
1	496, 496, 499	497, 499, 497	497, 498, 496
2	498, 497, 499	498, 496, 499	497, 499, 500
3	498, 498, 498	497, 498, 497	496, 498, 497
4	497, 497, 498	496, 496, 499	498, 497, 497
5	499, 501, 500	499, 499, 499	499, 499, 500
6	499, 498, 499	500, 499, 497	498, 498, 498
7	503, 499, 502	498, 499, 499	500, 499, 502
8	500, 499, 499	501, 498, 499	500, 501, 499
9	499, 500, 499	500, 500, 498	500, 499, 500
10	497, 496, 496	500, 494, 496	496, 498, 496

3. That is, by these rough “eyeball” standards, there is some hint in these data of both repeatability and reproducibility components in the overall measurement imprecision.

To this point in our discussions of R&R, we have not involved more than a single measurand. Effectively, we have confined attention to a single row of a table like Table 2.1. Standard industrial gauge R&R studies treat multiple parts (partially as a way of making sure that reliability of measurement doesn’t obviously vary wildly across parts). So here we consider the kind of multiple-part case represented in Table 2.1.

The model most commonly used in this context is the so-called two-way random effects model that can be found in many intermediate-level statistical method texts. (See, e.g., Section 8.4 of Vardeman’s *Statistics for Engineering Problem Solving*.) Let

y_{ijk} = the k th measurement made by operator j on part i .

The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}, \quad (2.26)$$

Two-Way
Random
Effects Model

where the μ is an (unknown) constant, the α_i are normal random variables with mean 0 and variance σ_α^2 , the β_j are normal random variables with mean 0 and variance σ_β^2 , the $\alpha\beta_{ij}$ are normal random variables with mean 0 and variance $\sigma_{\alpha\beta}^2$, the ϵ_{ijk} are normal random variables with mean 0 and variance σ^2 , and all of the α s, β s, $\alpha\beta$ s, and ϵ s are independent. In this model, the unknown constant μ is an average (over all possible operators and all possible parts) measurement, the α s are (random) effects of different parts, the β s are (random) effects of different operators, the $\alpha\beta$ s are (random) joint effects peculiar to particular part \times operator combinations, and the ϵ s are (random) measurement errors. The variances σ_α^2 , σ_β^2 , $\sigma_{\alpha\beta}^2$, and σ^2 are called “variance components” and their sizes govern how much variability is seen in the measurements y_{ijk} .

Consider a hypothetical case with $I = 2$, $J = 2$, and $m = 2$. Model (2.26) says that there is a normal distribution with mean 0 and variance σ_α^2 from which α_1 and α_2 are drawn. And there is a normal distribution with mean 0 and variance σ_β^2 from which β_1 and β_2 are drawn. And there is a normal distribution with mean 0 and variance $\sigma_{\alpha\beta}^2$ from which $\alpha\beta_{11}$, $\alpha\beta_{12}$, $\alpha\beta_{21}$, and $\alpha\beta_{22}$ are drawn. And there is a normal distribution with mean 0 and variance σ^2 from which eight ϵ s are drawn. Then these realized values of the random effects are added to produce the eight measurements as indicated in Table 2.2.

TABLE 2.2. Measurements in a hypothetical gauge R&R study

	Operator 1	Operator 2
Part 1	$y_{111} = \mu + \alpha_1 + \beta_1 + \alpha\beta_{11} + \epsilon_{111}$ $y_{112} = \mu + \alpha_1 + \beta_1 + \alpha\beta_{11} + \epsilon_{112}$	$y_{121} = \mu + \alpha_1 + \beta_2 + \alpha\beta_{12} + \epsilon_{121}$ $y_{122} = \mu + \alpha_1 + \beta_2 + \alpha\beta_{12} + \epsilon_{122}$
Part 2	$y_{211} = \mu + \alpha_2 + \beta_1 + \alpha\beta_{21} + \epsilon_{211}$ $y_{212} = \mu + \alpha_2 + \beta_1 + \alpha\beta_{21} + \epsilon_{212}$	$y_{221} = \mu + \alpha_2 + \beta_2 + \alpha\beta_{22} + \epsilon_{221}$ $y_{222} = \mu + \alpha_2 + \beta_2 + \alpha\beta_{22} + \epsilon_{222}$

Either directly from Eq. (2.26) or as illustrated in Table 2.2, according to the two-way random effects model, the only differences between measurements for a fixed part \times operator combination are the measurement errors ϵ . And the variability of these is governed by the parameter σ . That is, σ is a measure of repeatability variation in this model, and one objective of an analysis of gauge R&R data is to estimate it.

Then, if one looks at a fixed “part i ” (row i), the quantity $\mu + \alpha_i$ is common across the row. In the context of a gauge R&R study this can be interpreted as the value of the i th measurand (these vary across parts/rows because the α_i vary). Then, still for a fixed part i , it is the values $\beta_j + \alpha\beta_{ij}$ that vary column/operator to column/operator. So in this gauge R&R context, this quantity functions as a kind of *part- i -specific operator bias*. (More on the qualifier “part i specific” in a bit.) According to model (2.26), the variance of $\beta_j + \alpha\beta_{ij}$ is $\sigma_\beta^2 + \sigma_{\alpha\beta}^2$, so an appropriate measure of reproducibility variation in this model is

$$\sigma_{\text{reproducibility}} = \sqrt{\sigma_\beta^2 + \sigma_{\alpha\beta}^2}. \quad (2.27)$$

According to the model, this is the standard deviation that would be experienced by many operators making a single measurement on the same part *assuming that there is no repeatability component to the overall variation*. Another way to say the same thing is to recognize this quantity as the standard deviation that would be experienced computing with long-run average measurements for many operators on the same part. That is, the quantity (2.27) is a measure of variability in operator bias for a fixed part in this model.

As long as one confines attention to a single row of a standard gauge R&R study, the one-way random effects model and analysis of Sect. 2.3 are relevant. The quantity $\sigma_{\text{reproducibility}}$ here is exactly σ_δ from application of the one-way model to a single-part gauge R&R study. (And the present σ is exactly σ_{device} .) What is new and at first perhaps a bit puzzling is that in the present context of multiple parts and display (2.27), the reproducibility variation has two components, σ_β and $\sigma_{\alpha\beta}$. This is because for a given part i , the model says that bias for operator

Repeatability
Standard
Deviation in the
Two-Way
Model

Reproducibility
Standard
Deviation in the
Two-Way
Model

j has both components β_j and $\alpha\beta_{ij}$. The model terms $\alpha\beta_{ij}$ allow “operator bias” to change part to part/measurand to measurand (an issue that simply doesn’t arise in the context of a single-part study). As such, they are *a measure of nonlinearity* (bias nonconstant in the measurand) in the overall measurement system. Two-way data like those in Table 2.1 allow one to estimate all of $\sigma_{\text{reproducibility}}$, σ_β and $\sigma_{\alpha\beta}$, and all else being equal, cases where the $\sigma_{\alpha\beta}$ component of $\sigma_{\text{reproducibility}}$ is small are preferable to those where it is large.

The quantity

$$\sigma_{\text{R\&R}} = \sqrt{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2} = \sqrt{\sigma_{\text{reproducibility}}^2 + \sigma^2} \quad (2.28)$$

Combined R&R
Standard
Deviation

is the standard deviation implied by the model (2.26) for many operators each making a single measurement on the same part. That is, quantity (2.28) is a measure of the combined imprecision in measurement attributable to *both* repeatability and reproducibility sources. And one might think of

$$\frac{\sigma^2}{\sigma_{\text{R\&R}}^2} = \frac{\sigma^2}{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2} \quad \text{and} \quad \frac{\sigma_{\text{reproducibility}}^2}{\sigma_{\text{R\&R}}^2} = \frac{\sigma_\beta^2 + \sigma_{\alpha\beta}^2}{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2} \quad (2.29)$$

as the fractions of total measurement variance due, respectively, to repeatability and reproducibility. If one can produce estimates of σ and $\sigma_{\text{reproducibility}}$, estimates of these quantities (2.28) and (2.29) follow in straightforward fashion.

It is common to treat some multiple of $\sigma_{\text{R\&R}}$ (often the multiplier is six, but sometimes 5.15 is used) as a kind of uncertainty associated with a measurement made using the gauge or measurement system in question. And when a gauge is being used to check conformance of a part dimension or other measured characteristics to engineering specifications (say, some lower specification L and some upper specification U), this multiple is compared to the spread in specifications. **Specifications** U and L are numbers set by product design engineers that are supposed to delineate what is required of a measured dimension in order that the item in question be functional. The hope is that measurement uncertainty is at least an order of magnitude smaller than the spread in specifications. Some organizations go so far as to call the quantity

Engineering
Specifications

$$GCR = \frac{6\sigma_{\text{R\&R}}}{U - L} \quad (2.30)$$

Gauge
Capability
Ratio

a **gauge capability (or precision-to-tolerance) ratio** and require that it be no larger than .1 (and preferably as small as .01) before using the gauge for checking conformance to such specifications. (In practice, one will only have an estimate of $\sigma_{\text{R\&R}}$ upon which to make an empirical approximation of a gauge capability ratio.)

2.4.2 Range-Based Estimation

Because range-based estimation (similar to, but not exactly the same as, what follows) is in common use for the analysis of gauge R&R studies and is easy to

describe, we will treat it here. In the next subsection, better methods based on ANOVA calculations (and REML methods) will be presented.

Consider first the estimation of σ . Restricting attention to any particular part \times operator combination, say part i and operator j , model (2.26) says that observations obtained for that combination differ only by independent normal random measurement error with mean 0 and variance σ^2 . That suggests that a measure of variability for the ij sample might be used as the basis of an estimator of σ . Historical precedent and ease of computation suggest measuring variability using a range (instead of a sample standard deviation or variance).

So let R_{ij} be the range of the m measurements on part i by operator j . The expected value of the range of a sample from a normal distribution is a constant (depending upon m) times the standard deviation of the distribution being sampled. The constants are well known and called d_2 . (We will write $d_2(m)$ to emphasize their dependence upon m and note that values of $d_2(m)$ are given in Table A.5.) It then follows that

$$ER_{ij} = d_2(m)\sigma,$$

which in turn suggests that the ratio

$$\frac{R_{ij}}{d_2(m)}$$

is a plausible estimator of σ . Better yet, one might average these over all $I \times J$ part \times operator combinations to produce the range-based estimator of σ :

Range-Based
Estimator for
Repeatability
Standard
Deviation

$$\hat{\sigma}_{\text{repeatability}} = \frac{\bar{R}}{d_2(m)}. \quad (2.31)$$

Example 13 (Example 12 continued.) Subtracting the smallest measurement for each part \times operator combination in Table 2.1 from the largest for that combination, one obtains the ranges in Table 2.3. These have mean $\bar{R} = 1.9$. From Table A.5, $d_2(3) = 1.693$. So using expression (2.31), an estimate of σ , the repeatability standard deviation for the caliper used by the students, is

$$\hat{\sigma}_{\text{repeatability}} = \frac{\bar{R}}{d_2(3)} = \frac{1.9}{1.693} = 1.12 \times 10^{-3} \text{ in.}$$

(Again, this is an estimate of the (long-run) standard deviation that would be experienced by any particular student measuring any particular punch many times.)

Consider now the standard deviation (2.27) representing the reproducibility portion of the gauge imprecision. It will be convenient to have some additional notation. Let

$$\bar{y}_{ij} = \text{the (sample) mean measurement made on part } i \text{ by operator } j \quad (2.32)$$

and

$$\begin{aligned}\Delta_i &= \max_j \bar{y}_{ij} - \min_j \bar{y}_{ij} \\ &= \text{the range of the mean measurements made on part } i .\end{aligned}$$

TABLE 2.3. Ranges of 30 part \times operator samples of measured punch Heights

Punch	Student 1	Student 2	Student 3
1	3	2	2
2	2	3	3
3	0	1	2
4	1	3	1
5	2	0	1
6	1	3	0
7	4	1	3
8	1	3	2
9	1	2	1
10	1	6	2

Notice that with the obvious notation for the sample average of the measurement errors ϵ , according to model (2.26),

$$\bar{y}_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \bar{\epsilon}_{ij} .$$

Thus, for a fixed part i these means \bar{y}_{ij} vary only according to independent normal random variables $\beta_j + \alpha\beta_{ij} + \bar{\epsilon}_{ij}$ that have mean 0 and variance $\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2/m$. Thus their range, Δ_i , has mean

$$E\Delta_i = d_2(J) \sqrt{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2/m} .$$

This suggests $\Delta_i/d_2(J)$ or, better yet, the average of these over all parts i , $\bar{\Delta}/d_2(J)$, as an estimator of $\sqrt{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2/m}$. This in turn suggests that one can estimate $\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2/m$ with $(\bar{\Delta}/d_2(J))^2$. Then remembering that $\bar{R}/d_2(m) = \hat{\sigma}_{\text{repeatability}}$ is an estimator of σ , an obvious estimator of $\sigma_\beta^2 + \sigma_{\alpha\beta}^2$ becomes

$$\left(\frac{\bar{\Delta}}{d_2(J)} \right)^2 - \frac{1}{m} \left(\frac{\bar{R}}{d_2(m)} \right)^2 . \quad (2.33)$$

The quantity (2.33) is meant to approximate $\sigma_\beta^2 + \sigma_{\alpha\beta}^2$, which is nonnegative. But the estimator (2.33) can on occasion give negative values. When this happens, it is sensible to replace the negative value by 0 and thus expression (2.33) by

$$\max \left(0, \left(\frac{\bar{\Delta}}{d_2(J)} \right)^2 - \frac{1}{m} \left(\frac{\bar{R}}{d_2(m)} \right)^2 \right) . \quad (2.34)$$

TABLE 2.4. Part \times operator means and ranges of such means for the punch height data

Punch(i)	\bar{y}_{i1}	\bar{y}_{i2}	\bar{y}_{i3}	Δ_i
1	497.00	497.67	497.00	.67
2	498.00	497.67	498.67	1.00
3	498.00	497.33	497.00	1.00
4	497.33	497.00	497.33	.33
5	500.00	499.00	499.33	1.00
6	498.67	498.67	498.00	.67
7	501.33	498.67	500.33	2.67
8	499.33	499.33	500.00	.67
9	499.33	499.33	499.67	.33
10	496.33	496.67	496.67	.33

So finally, an estimator of the reproducibility standard deviation can be had by taking the square root of expression (2.34). That is, one may estimate the quantity (2.27) with

$$\hat{\sigma}_{\text{reproducibility}} = \sqrt{\max\left(0, \left(\frac{\bar{\Delta}}{d_2(J)}\right)^2 - \frac{1}{m} \left(\frac{\bar{R}}{d_2(m)}\right)^2\right)}. \quad (2.35)$$

Example 14 (*Examples 12 and 13 continued.*) Table 2.4 organizes \bar{y}_{ij} and Δ_i values for the punch height measurements of Table 2.1. Then $\bar{\Delta} = 8.67/10 = .867$, and since $J = 3$, $d_2(J) = d_2(3) = 1.693$. So using Eq. (2.35),

$$\begin{aligned} \hat{\sigma}_{\text{reproducibility}} &= \sqrt{\max\left(0, \left(\frac{.867}{1.693}\right)^2 - \frac{1}{3} \left(\frac{1.9}{1.693}\right)^2\right)}, \\ &= \sqrt{\max(0, -.158)}, \\ &= 0. \end{aligned}$$

This calculation suggests that this is a problem where σ appears to be so large that the reproducibility standard deviation cannot be seen above the intrinsic “noise” in measurement conceptualized as the repeatability component of variation. Estimates of the ratios (2.29) based on $\hat{\sigma}_{\text{repeatability}}$ and $\hat{\sigma}_{\text{reproducibility}}$ would attribute fractions 1 and 0 of the overall variance in measurement to, respectively, repeatability and reproducibility.

2.4.3 ANOVA-Based Estimation

The formulas of the previous subsection are easy to discuss and use, but they are not at all the best available. Ranges are not the most effective tools for estimating normal standard deviations. And the range-based methods have no corresponding way for making confidence intervals. More effective (and computationally more demanding) statistical tools are available, and we proceed to discuss some of them.

TABLE 2.5. A generic gauge R&R two-way ANOVA table

Source	SS	df	MS
Part	SSA	$I - 1$	$MSA = SSA / (I - 1)$
Operator	SSB	$J - 1$	$MSB = SSB / (J - 1)$
Part \times Operator	$SSAB$	$(I - 1)(J - 1)$	$MSAB = SSAB / (I - 1)(J - 1)$
Error	SSE	$IJ(m - 1)$	$MSE = SSE / IJ(m - 1)$
Total	$SSTot$	$IJm - 1$	

An $I \times J \times m$ data set of y_{ijk} s like that produced in a typical gauge R&R study is often summarized in a so-called two-way ANOVA table. Table 2.5 is a generic version of such a summary. Any decent statistical package will process a gauge R&R data set and produce such a summary table. As in a one-way ANOVA, “mean squares” are essentially sample variances (squares of sample standard deviations). MSA is essentially a sample variance of part averages, MSB is essentially a sample variance of operator averages, MSE is an average of within-cell sample variances, and $MSTot$ isn’t typically calculated, but is a grand sample variance of all observations.

For purposes of being clear (and not because they are typically used for “hand calculation”) we provide formulas for sums of squares. With cell means \bar{y}_{ij} as in display (2.32), define row and column averages and the grand average of these

$$\bar{y}_{i.} = \frac{1}{J} \sum_j \bar{y}_{ij} \quad \text{and} \quad \bar{y}_{.j} = \frac{1}{I} \sum_i \bar{y}_{ij} \quad \text{and} \quad \bar{y}_{..} = \frac{1}{IJ} \sum_{ij} \bar{y}_{ij}.$$

Then the sums of squares are

$$\begin{aligned}
 SSTot &= \sum_{ijk} (y_{ijk} - \bar{y}_{..})^2, \\
 SSE &= \sum_{ijk} (y_{ijk} - \bar{y}_{ij})^2, \\
 SSA &= mJ \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2, \\
 SSB &= mI \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2, \text{ and} \\
 SSAB &= m \sum_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \\
 &= SSTot - SSE - SSA - SSB
 \end{aligned}$$

TABLE 2.6. Data from a small in-class gauge R&R study

Part	Operator 1	Operator 2	Operator 3
1	.52, .52	.54, .53	.55, .55
2	.56, .55	.54, .54	.55, .56
3	.57, .56	.55, .56	.57, .57
4	.55, .55	.54, .55	.56, .55

Corresponding degrees of freedom and mean squares are

$$\begin{aligned} df E &= (m - 1) IJ \text{ and } MSE = SSE / (m - 1) IJ , \\ df A &= I - 1 \text{ and } MSA = SSA / (I - 1) , \\ df B &= J - 1 \text{ and } MSB = SSB / (J - 1) , \text{ and} \\ df AB &= (I - 1)(J - 1) \text{ and } MSAB = SSAB / (I - 1)(J - 1) . \end{aligned}$$

Example 15 In-Class Gauge R&R Study. The data in Table 2.6 on page 69 were collected in an in-class gauge R&R exercise where $I = 4$ polystyrene packing peanuts were measured for size (in in) by $J = 3$ students $m = 2$ times apiece using the same inexpensive caliper. The JMP statistical package produces the sums of squares

$$\begin{aligned} SSA &= .00241250, SSB = .00080833, SSAB = .00072500, \\ SSE &= .00035000, \text{ and } SSTot = .00429583. \end{aligned}$$

for these data that can be used as raw material for making important inferences for the R&R study based on model (2.26). Corresponding mean squares are

$$\begin{aligned} MSE &= .00035000 / (2 - 1)(4)(3) = .00002917 , \\ MSA &= .00241250 / (4 - 1) = .00080417 , \\ MSB &= .00080833 / (3 - 1) = .00040417 , \text{ and} \\ MSAB &= .00072500 / (4 - 1)(3 - 1) = .00012083 . \end{aligned}$$

High-quality statistical software (like JMP or R) will automatically produce REML-based estimates and confidence intervals for the variance components $\sigma_\alpha^2, \sigma_\beta^2, \sigma_{\alpha\beta}^2$, and σ^2 . As the quantities $\sigma_{\text{reproducibility}}^2$ and $\sigma_{\text{R\&R}}^2$ are a bit specialized (being of interest in our R&R application of the two-way random effects model, but not in other common applications), inferences for them are not automatically available. It is possible, but usually not convenient, to use the output of REML analyses to make inferences for these more specialized quantities. So here we will provide formulas for ANOVA-based estimators of $\sigma, \sigma_{\text{reproducibility}}$, and $\sigma_{\text{R\&R}}$ and appropriate Satterthwaite approximate degrees of freedom for making confidence limits. (Where readers know how to obtain REML-based estimates and intervals, our recommendation is to use them in preference to ANOVA-based estimators that follow.)

Single-number estimators for the quantities of most interest in a gauge R&R study are

ANOVA-Based
Estimator for
Repeatability
Standard
Deviation

$$\hat{\sigma}_{\text{repeatability}} = \hat{\sigma} = \sqrt{MSE} , \quad (2.36)$$

$$\hat{\sigma}_{\text{reproducibility}} = \sqrt{\max \left(0, \frac{MSB}{mI} + \frac{(I-1)}{mI} MSAB - \frac{1}{m} MSE \right)}, \quad (2.37)$$

ANOVA-Based
Estimator for
Reproducibility
Standard
Deviation

and

$$\hat{\sigma}_{\text{R\&R}} = \sqrt{\frac{1}{mI} MSB + \frac{I-1}{mI} MSAB + \frac{m-1}{m} MSE}. \quad (2.38)$$

ANOVA-Based
Estimator
for $\sigma_{\text{R\&R}}$

Confidence limits based on any of these estimators are of the generic form (already used several times in this chapter)

$$“\hat{\sigma}” \sqrt{\frac{“\hat{\nu}”}{\chi_{\text{upper}}^2}} \quad \text{and} \quad “\hat{\sigma}” \sqrt{\frac{“\hat{\nu}”}{\chi_{\text{lower}}^2}} \quad (2.39)$$

Generic
Confidence
Limits for a
Standard
Deviation

where “ $\hat{\sigma}$ ” is one of the estimators, “ $\hat{\nu}$ ” is a corresponding (exact or “Satterthwaite approximate”) degrees of freedom, and the χ^2 percentage points are based on “ $\hat{\nu}$.” So it only remains to record formulas for appropriate degrees of freedom. These are

$$\nu_{\text{repeatability}} = IJ(m-1), \quad (2.40)$$

Degrees of
Freedom for
Use with
Formulas
(2.36) and
(2.39)

$$\begin{aligned} \hat{\nu}_{\text{reproducibility}} &= \frac{\hat{\sigma}_{\text{reproducibility}}^4}{\frac{\left(\frac{MSB}{mI}\right)^2}{J-1} + \frac{\left(\frac{(I-1)MSAB}{mI}\right)^2}{(I-1)(J-1)} + \frac{\left(\frac{MSE}{m}\right)^2}{IJ(m-1)}} \\ &= \frac{\hat{\sigma}_{\text{reproducibility}}^4}{\frac{1}{m^2} \left(\frac{MSB^2}{I^2(J-1)} + \frac{(I-1)MSAB^2}{I^2(J-1)} + \frac{MSE^2}{IJ(m-1)} \right)}, \end{aligned}$$

Degrees of
Freedom for
Use with
Formulas
(2.37) and
(2.39)

(2.41)

and

Degrees of
Freedom for
Use with
Formulas
(2.38) and
(2.39)

$$\begin{aligned}\hat{\nu}_{R\&R} &= \frac{\hat{\sigma}_{R\&R}^4}{\frac{\left(\frac{MSB}{mI}\right)^2}{J-1} + \frac{\left(\frac{(I-1)MSAB}{mI}\right)^2}{(I-1)(J-1)} + \frac{\left(\frac{(m-1)MSE}{m}\right)^2}{IJ(m-1)}} \\ &= \frac{\hat{\sigma}_{R\&R}^4}{\frac{1}{m^2} \left(\frac{MSB^2}{I^2(J-1)} + \frac{(I-1)MSAB^2}{I^2(J-1)} + \frac{(m-1)MSE^2}{IJ} \right)}\end{aligned}$$

(2.42)

Formulas (2.37), (2.41), (2.38), and (2.42) are tedious (but hardly impossible) to use with a pocket calculator. But a very small program, MathCAD worksheet, or spreadsheet template can be written to evaluate the estimates of standard deviations and approximate degrees of freedom from the sums of squares, m , I , and J .

Example 16 (Example 15 continued.) A two-way random effects analysis of the data of Table 2.6 made using the JMP statistical package produces REML-based confidence limits of

$$0 \text{ and } \sqrt{.0001359}, \text{ i.e., } 0 \text{ in and } .012 \text{ in for } \sigma_{\beta}$$

and

$$0 \text{ and } \sqrt{.0001152}, \text{ i.e., } 0 \text{ in and } .011 \text{ in for } \sigma_{\alpha\beta}.$$

There is thus at least the suggestion that a substantial part of the reproducibility variation in the data of Table 2.6 is a kind of nonconstant bias on the part of the student operators measuring the peanuts.

Using formulas (2.36), (2.37), and (2.38) it is possible to verify that in this problem

$$\begin{aligned}\hat{\sigma}_{\text{repeatability}} &= \hat{\sigma} = .005401 \text{ in}, \\ \hat{\sigma}_{\text{reproducibility}} &= .009014 \text{ in}, \text{ and} \\ \hat{\sigma}_{R\&R} &= .011 \text{ in}.\end{aligned}$$

Using formulas (2.40), (2.41), and (2.42), these have corresponding degrees of freedom

$$\begin{aligned}\nu_{\text{repeatability}} &= (4)(3)(2-1) = 12, \\ \hat{\nu}_{\text{reproducibility}} &= 4.04, \text{ and} \\ \hat{\nu}_{R\&R} &= 7.45.\end{aligned}$$

So (rounding degrees of freedom down in the last two cases) using the limits (2.39), 95 % confidence limits for $\sigma_{\text{repeatability}}$ are

$$.005401\sqrt{\frac{12}{23.337}} \text{ and } .005401\sqrt{\frac{12}{4.404}}$$

i.e.,

$$.0039 \text{ in and } .0089 \text{ in ,}$$

approximate 95 % confidence limits for $\sigma_{\text{reproducibility}}$ are

$$.009014\sqrt{\frac{4}{11.143}} \text{ and } .009014\sqrt{\frac{4}{.484}}$$

i.e.,

$$.0054 \text{ in and } .0259 \text{ in ,}$$

and approximate 95 % confidence limits for $\sigma_{\text{R\&R}}$ are

$$.011\sqrt{\frac{7}{16.013}} \text{ and } .011\sqrt{\frac{7}{1.690}}$$

i.e.,

$$.0073 \text{ in and } .0224 \text{ in .}$$

These intervals show that none of these standard deviations are terribly well determined (degrees of freedom are small and intervals are wide). If better information is needed, more data would have to be collected. But there is at least some indication that $\sigma_{\text{repeatability}}$ and $\sigma_{\text{reproducibility}}$ are roughly of the same order of magnitude. The caliper used to make the measurements was a fairly crude one, and there were detectable differences in the way the student operators used that caliper.

Suppose, for the sake of example, that engineering requirements on these polystyrene peanuts were that they be of size $.50 \text{ in} \pm .05 \text{ in}$. In such a context, the gauge capability ratio (2.30) could be estimated to be between

$$\frac{6(.0073)}{.55 - .45} = .44 \text{ and } \frac{6(.0224)}{.55 - .45} = 1.34.$$

These values are not small. (See again the discussion on page 65.) This measurement “system” is not really adequate to check conformance to even these crude $\pm .05 \text{ in}$ product requirements.

Some observations regarding the planning of a gauge R&R study are in order at this point. The precisions with which one can estimate σ , $\sigma_{\text{reproducibility}}$, and $\sigma_{\text{R\&R}}$ obviously depend upon I , J , and m . Roughly speaking, precision of estimation of σ is governed by the product $(m - 1)IJ$, so increasing any of the “dimensions” of the data array will improve estimation of repeatability. However, it is primarily J that governs the precision with which $\sigma_{\text{reproducibility}}$ and $\sigma_{\text{R\&R}}$ can be estimated. Only by increasing the number of operators in a gauge R&R study can one substantially improve the estimation of reproducibility variation.

While this fact about the estimation of reproducibility is perfectly plausible, its implications are not always fully appreciated (or at least not kept clearly in mind) by quality assurance practitioners. For example, many standard gauge R&R data collection forms allow for at most $J = 3$ operators. But 3 is a *very* small sample size when it comes to estimating a variance or standard deviation. So although the data in Table 2.1 are perhaps more or less typical of many R&R data sets, the small ($J = 3$) number of operators evident there should *not* be thought of as in any way ideal. To get a really good handle on the size of reproducibility variation, many more operators would be needed.

Section 2.4 Exercises

- Consider again the situation of problem 3 of the Sect. 2.3 exercises and the data from the **Fiber Angle** case used there. (Operator 1 measured fiber angle for three different specimens four times each.) Recast that scenario into the two-way framework of this section.
 - Give the values of I , J , and m .
 - Find a range-based estimate of σ_{device} .
 - Find a range-based estimate of σ_x .
- Based only on the data of problem 3 of the Sect. 2.3 exercises, can $\sigma_{\text{reproducibility}}$ be estimated? Why or why not?
- Consider again the situation of problems 2 and 1 of the Sect. 2.3 exercises and the data from the **Fiber Angle** case used there. (Fiber angle for specimen 1 was measured four times by each of operators 1, 2, and 3.) Recast that scenario into the two-way framework of this section.
 - Give the values of I , J , and m .
 - Find a range-based estimate of $\sigma_{\text{repeatability}}$.
 - Find a range-based estimate of $\sigma_{\text{reproducibility}}$.
 - Based only on the data considered here, can σ_x be estimated? Why or why not?
- Washer Assembly.** Sudam, Heimer, and Mueller studied a clothes washer base assembly. Two operators measured the distance from one edge of a washer base assembly to an attachment. For a single base assembly, the same distance was measured four times by each operator. This was repeated on three different base assemblies. The target distance was 13.320 with an upper specification of $U = 13.42$ and a lower specification of $L = 13.22$. A standard gauge R&R study was conducted and data like those below were obtained. (Units are 10^{-1} in.)

Part	Operator 1	Operator 2
1	13.285, 13.284, 13.283, 13.282	13.284, 13.288, 13.287, 13.283
2	13.298, 13.293, 13.291, 13.291	13.297, 13.292, 13.292, 13.293
3	13.357, 13.356, 13.354, 13.356	13.355, 13.354, 13.352, 13.357

- What were the values of I , J , and m in this study?
- Based on the ANOVA table for the data given below, find the estimates for $\sigma_{\text{repeatability}}$, $\sigma_{\text{reproducibility}}$, and $\sigma_{\text{R\&R}}$.
- Give 95 % confidence limits for $\sigma_{\text{repeatability}}$, $\sigma_{\text{reproducibility}}$, and $\sigma_{\text{R\&R}}$.
- Find 95 % confidence limits for the GCR. (Hint: Use the last of your answers to (c).)

ANOVA table			
Source	SS	df	MS
Part	.0236793	2	.0118396
Operator	.0000007	1	.0000007
Part \times Operator	.0000106	2	.0000053
Error	.0000895	18	.0000050
Total	.0237800	23	

2.5 Simple Linear Regression and Calibration Studies

Calibration is an essential activity in the qualification and maintenance of measurement devices. In a calibration study, one uses a measurement device to produce measurements on “standard” specimens with (relatively well-) “known” values of measurands and sees how the measurements compare to the known values. If there are systematic discrepancies between what is known to be true and what the device reads, a conversion scheme is created to (in future use of the device) adjust what is read to something that is hopefully closer to the (future) truth. A slight extension of “regression” analysis (curve fitting) as presented in an elementary statistics course is the relevant statistical methodology in making this conversion. (See, e.g., Section 9.1 of Vardeman and Jobe’s *Basic Engineering Data Collection and Analysis*.) This section discusses exactly how regression analysis is used in calibration.

Calibration studies employ true/gold-standard-measurement values of measurands x and “local” measurements y . (Strictly speaking, y need not even be in the same units as x .) Regression analysis can provide both “point conversions” and measures of uncertainty (the latter through inversion of “prediction limits”). The simplest version of this is where observed measurements are approximately

linearly related to measurands, i.e.,

$$y \approx \beta_0 + \beta_1 x$$

This is “linear calibration.” The standard statistical model for such a circumstance is

$$y = \beta_0 + \beta_1 x + \epsilon \quad (2.43)$$

for a normal error ϵ with mean 0 and standard deviation σ . (σ describes how much y s vary for a fixed x and in the present context typically amounts to a repeatability standard deviation.) This model can be pictured as in Fig. 2.17.

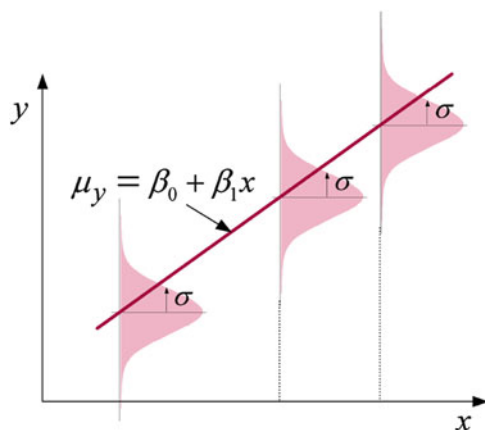


FIGURE 2.17. A schematic of the usual simple linear regression model (2.43)

For n data pairs (x_i, y_i) , simple linear regression methodology allows one to make confidence intervals and tests associated with the model and *prediction limits for a new measurement* y_{new} associated with a new measurand, x_{new} . These are of the form

$$(b_0 + b_1 x_{\text{new}}) \pm t s_{\text{LF}} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \quad (2.44)$$

where the least squares line is $\hat{y} = b_0 + b_1 x$ and s_{LF} (a “line-fitting” sample standard deviation) is an estimate of σ derived from the fit of the line to the data. Any good statistical package will compute and plot these limits as functions of x_{new} along with a least squares line through the data set.

Example 17 Measuring Cr^{6+} Concentration with a UV-Vis Spectrophotometer. The data below were taken from a web page of the School of Chemistry at the University of Witwatersrand developed and maintained by Dr. Dan Billing. They are measured absorbance values, y , for $n = 6$ solutions with “known” Cr^{6+} concentrations, x (in mg/l), from an analytical lab.

x	0	1	2	4	6	8
y	.002	.078	.163	.297	.464	.600

Figure 2.18 is a plot of these data, the corresponding least squares line, and the prediction limits (2.44).

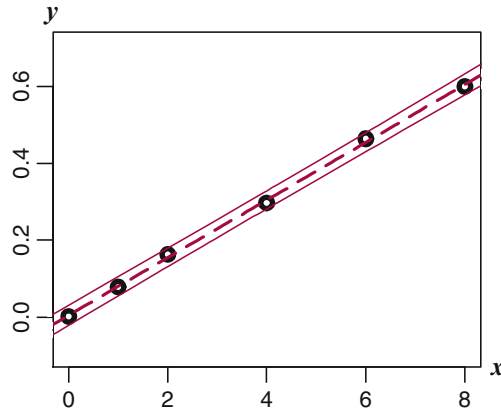


FIGURE 2.18. Scatterplot of the Cr^{6+} Concentration calibration data, least squares line, and prediction limits for y_{new}

What is here of most interest about simple linear regression technology is what it says about calibration and measurement in general. Some applications of inference methods based on the model (2.43) to metrology are the following.

1. From a simple linear regression output,

$$s_{\text{LF}} = \sqrt{MSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \text{“root mean square error”} \quad (2.45)$$

is an estimated repeatability standard deviation. One may make confidence intervals for $\sigma = \sigma_{\text{repeatability}}$ based on the estimate (2.45) using $\nu = n - 2$ degrees of freedom and limits

$$s_{\text{LF}} \sqrt{\frac{n-2}{\chi_{\text{upper}}^2}} \quad \text{and} \quad s_{\text{LF}} \sqrt{\frac{n-2}{\chi_{\text{lower}}^2}}. \quad (2.46)$$

Confidence
Limits for σ in
Model (2.43)

2. The least squares equation $\hat{y} = b_0 + b_1x$ can be solved for x , giving

$$\hat{x}_{\text{new}} = \frac{y_{\text{new}} - b_0}{b_1} \quad (2.47)$$

Conversion
Formula for
a Future
Measurement,
 y_{new}

as a way of estimating a new “gold-standard” value (a new measurand, x_{new}) from a measured local value, y_{new} .

3. One can take the prediction limits (2.44) for y_{new} and “turn them around” to get confidence limits for the x_{new} corresponding to a measured local y_{new} . This provides a defensible way to set “error bounds” on what y_{new} indicates about x_{new} .
4. In cases (unlike Example 17) where y and x are in the same units, confidence limits for the slope β_1 of the simple linear regression model

Confidence
Limits for β_1 in
Model (2.43)

$$b_1 \pm t \frac{s_{\text{LF}}}{\sqrt{\sum (x_i - \bar{x})^2}} \quad (2.48)$$

provide a way of investigating the constancy of bias (linearity of the measurement device in the sense introduced on page 37). That is, when x and y are in the same units, $\beta_1 = 1.0$ is the case of constant bias. If confidence limits for β_1 fail to include 1.0, there is clear evidence of device nonlinearity.

Example 18 (Example 17 continued.) The use of the JMP statistical package with the data of Example 17 produces

$$y = .0048702 + .0749895x \text{ with } s_{\text{LF}} = .007855 .$$

We might expect a local (y) repeatability standard deviation of around .008 (in the y absorbance units). In fact, 95 % confidence limits for σ can be made (using $n - 2 = 4$ degrees of freedom and formula (2.46)) as

$$.007855 \sqrt{\frac{4}{11.143}} \text{ and } .007855 \sqrt{\frac{4}{.484}} ,$$

i.e.,

$$.0047 \text{ and } .0226 .$$

Making use of the slope and intercept of the least squares line, a conversion formula for going from y_{new} to x_{new} is (as in display (2.47))

$$\hat{x}_{\text{new}} = \frac{y_{\text{new}} - .0048702}{.0749895} ,$$

So, for example, a future measured absorbance of $y_{\text{new}} = .20$ suggests a concentration of

$$\hat{x}_{\text{new}} = \frac{.20 - .0048702}{.0749895} = 2.60 \text{ mg/l} .$$

Finally, Fig. 2.19 on page 80 is a modification of Fig. 2.18 that illustrates how the plotted prediction limits (2.44) provide both 95 % predictions for a

new measurement on a fixed/known measurand and 95 % confidence limits on a new measurand, having observed a particular measurement. Reading from the figure, one is “95 % sure” that a future observed absorbance of .20 comes from a concentration between

$$2.28 \text{ mg/l and } 2.903 \text{ mg/l}.$$

Example 19 A Check on Device “Linearity.” A calibration data set due to John Mandel compared $n = 14$ measured values y for a single laboratory to corresponding consensus values x for the same specimens derived from multiple labs. (The units are not available, but were the same for x and y values.) A simple linear regression analysis of the data pairs produced

$$b_1 = .882 \text{ and } \frac{s_{\text{LF}}}{\sqrt{\sum (x_i - \bar{x})^2}} = .012$$

so that (using the upper 2.5 % point of the t_{12} distribution, 2.179, and formula (2.48)) 95 % confidence limits for β_1 are

$$.882 \pm 2.179 (.012)$$

or

$$.882 \pm .026.$$

A 95 % confidence interval for β_1 clearly does not include 1.0. So bias for the single laboratory was not constant. (The measurement “device” was not linear in the sense discussed on page 37.)

Section 2.5 Exercises

1. $n = 14$ polymer specimens of known weights, x , were weighed and the measured weights, y , recorded. The following table contains the data. (All units are gs.)

x	1	1	3	3	5	5	7
y	1.10	.95	2.98	3.01	5.02	4.99	6.97
x	7	10	10	12	12	14	14
y	7.10	10.03	9.99	12.00	11.98	14.10	14.00

- (a) Find the least squares line $\hat{y} = b_0 + b_1x$ for these data.

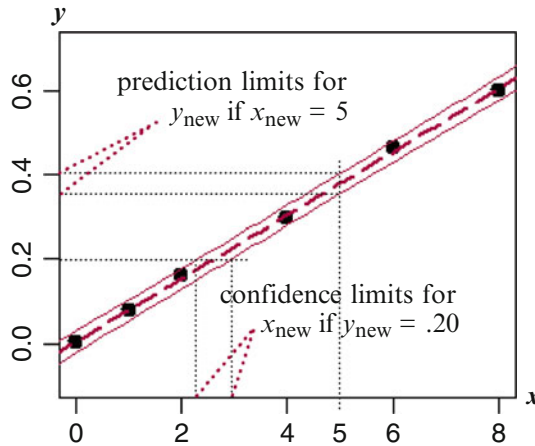


FIGURE 2.19. Confidence limits for x_{new} based on an observed y_{new} (and prediction limits (2.44))

- (b) Find the estimated repeatability standard deviation corresponding to your regression analysis.
 - (c) Find 95 % confidence limits for the y repeatability standard deviation based on your answer to (b).
2. In the context of problem 1, suppose a new specimen is measured as having a weight of 6.10 g.
 - (a) Find the “calibrated weight,” \hat{x} , corresponding to this new specimen based on your regression analysis.
 - (b) Find 95 % confidence limits for the slope of the relationship between measured and actual weight (β_1). Does the device used to produce the y measurements have constant bias (is it “linear”)? Why or why not?
3. Based on your regression analysis in problem 1, find 95 % prediction limits for the next measured weight for a new specimen with standard known weight of 8 g.
4. Would it be wise to use the above regression analyses to adjust a measured specimen weight of $y_{\text{new}} = .2$ g? Why or why not?

2.6 R&R Considerations for Go/No-Go Inspection

Ideally, observation of a process results in quantitative measurements. But there are some contexts in which all that is determined is whether an item or process condition is of one of two types, which we will for the present call “conforming”

and “nonconforming.” It is, for example, common to check the conformance of machined metal parts to some engineering requirements via the use of a “go/no-go gauge.” (A part is conforming if a critical dimension fits into the larger of two check fixtures and does not fit into the smaller of the two.) And it is common to task human beings with making visual inspections of manufactured items and producing an “ok/not-ok” call on each.

Engineers are sometimes then called upon to apply the qualitative “repeatability” and “reproducibility” concepts of metrology to such go/no-go or “0/1” contexts. One wants to separate some measure of overall inconsistency in 0/1 “calls” on items into pieces that can be mentally charged to inherent inconsistency in the equipment or method and the remainder that can be charged to differences between how operators use it. Exactly how to do this is presently not well established. The best available statistical methodology for this kind of problem is more complicated than can be presented here (involving so-called generalized linear models and random effects in these). What we *can* present is a rational way of making point estimates of what might be termed repeatability and reproducibility components of variation in 0/1 calls. (These are based on reasoning similar to that employed in Sect. 2.4.2 to find correct range-based estimates in usual measurement R&R contexts.) We then remind the reader of elementary methods of estimating differences in population proportions and in mean differences and point out their relevance in the present situation.

2.6.1 Some Simple Probability Modeling

To begin, think of coding a “nonconforming” call as “1” and a “conforming” call as “0” and having J operators each make m calls on a fixed part. Suppose that J operators have individual probabilities p_1, p_2, \dots, p_J of calling the part “nonconforming” on any single viewing and that across m viewings

X_j = the number of nonconforming calls among the m made by operator j

is binomial (m, p_j) . We’ll assume that the p_j are random draws from some population with mean π and variance v .

The quantity

$$p_j (1 - p_j)$$

is a kind of “per-call variance” associated with the declarations of operator j and might serve as a kind of repeatability variance for that operator. (Given the value of p_j , elementary probability theory says that the variance of X_j is $mp_j (1 - p_j)$.) The biggest problem here is that unlike what is true in the usual case of gauge R&R for measurements, this variance is not constant across operators. But its expected value, namely

$$\begin{aligned} E(p_j (1 - p_j)) &= \pi - E p_j^2 \\ &= \pi - (v + \pi^2) \\ &= \pi (1 - \pi) - v \end{aligned}$$

can be used as a sensible measure of variability in conforming/nonconforming classifications chargeable to repeatability sources. The variance v serves as a measure of reproducibility variance. This ultimately points to

$$\pi(1 - \pi)$$

as the “total R&R variance” here. That is, we make definitions for 0/1 contexts

R&R Variance
for One Part in
a 0/1 Context

$$\sigma_{\text{R\&R}}^2 = \pi(1 - \pi) \quad (2.49)$$

and

Repeatability
Variance for
One Part in a
0/1 Context

$$\sigma_{\text{repeatability}}^2 = \pi(1 - \pi) - v \quad (2.50)$$

and

Reproducibility
Variance for
One Part in a
0/1 Context

$$\sigma_{\text{reproducibility}}^2 = v \quad (2.51)$$

2.6.2 Simple R&R Point Estimates for 0/1 Contexts

Still thinking of a single fixed part, let

$$\hat{p}_j = \frac{\text{the number of “nonconforming” calls made by operator } j}{m} = \frac{X_j}{m}$$

and define the (sample) average of these:

$$\bar{\hat{p}} = \frac{1}{J} \sum_{j=1}^J \hat{p}_j .$$

It is possible to argue that

$$E\bar{\hat{p}} = \pi$$

so that a plausible estimate of $\sigma_{\text{R\&R}}^2$ is

Estimator of
R&R Variance
for a Single
Part in a 0/1
Context

$$\hat{\sigma}_{\text{R\&R}}^2 = \bar{\hat{p}}(1 - \bar{\hat{p}}) \quad (2.52)$$

Then, since $\hat{p}_j(1 - \hat{p}_j)$ is a plausible estimate of the “per-call variance” associated with the declarations of operator j , $p_j(1 - p_j)$, an estimate of $\sigma_{\text{repeatability}}^2$ is

Estimator of
Repeatability
Variance for a
Single Part in a
0/1 Context

$$\hat{\sigma}_{\text{repeatability}}^2 = \overline{\hat{p}(1 - \hat{p})} \quad (2.53)$$

(the sample average of the $\hat{p}_j(1 - \hat{p}_j)$). Finally, a simple estimate of $\sigma_{\text{reproducibility}}^2 = v$ is

$$\begin{aligned}\hat{\sigma}_{\text{reproducibility}}^2 &= \hat{\sigma}_{\text{R\&R}}^2 - \hat{\sigma}_{\text{repeatability}}^2 \\ &= \bar{\hat{p}}(1 - \bar{\hat{p}}) - \overline{\hat{p}(1 - \hat{p})}\end{aligned}\quad (2.54)$$

Estimator of
Reproducibility
Variance for a
Single Part in a
0/1 Context

Again, the estimators (2.52), (2.53), and (2.54) are based on a single part. Exactly what to do based on multiple parts (say I of them) is not completely obvious. But in order to produce a simple methodology, *we will simply average estimates made one part at a time across multiple parts*, presuming that parts in hand are sensibly thought of as a random sample of parts to be checked and that this averaging is a reasonable way to combine information across parts.

TABLE 2.7. Hypothetical results of visual inspection of five parts by three operators

	Operator 1		Operator 2		Operator 3	
	\hat{p}	$\hat{p}(1 - \hat{p})$	\hat{p}	$\hat{p}(1 - \hat{p})$	\hat{p}	$\hat{p}(1 - \hat{p})$
Part 1	.2	.16	.4	.24	.2	.16
Part 2	.6	.24	.6	.24	.7	.21
Part 3	1.0	0	.8	.16	.7	.21
Part 4	.1	.09	.1	.09	.1	.09
Part 5	.1	.09	.3	.21	.3	.21

In order for any of this to have a chance of working, m will need to be fairly large. The usual gauge R&R “ $m = 2$ or 3 ” just isn’t going to produce informative results in the present context. And in order for this to work in practice (so that an operator isn’t just repeatedly looking at the same few parts over and over and remembering how he or she has called them in the past), a large value of I may also be needed.

TABLE 2.8. R&R calculations for the hypothetical visual inspection data

	$\overline{\hat{p}(1 - \hat{p})} = \hat{\sigma}_{\text{repeatability}}^2$	$\bar{\hat{p}}$	$\bar{\hat{p}}(1 - \bar{\hat{p}}) = \hat{\sigma}_{\text{R\&R}}^2$	$\overline{\hat{p}(1 - \hat{p})} - \bar{\hat{p}}(1 - \bar{\hat{p}}) = \hat{\sigma}_{\text{reproducibility}}^2$
Part 1	.187	.2667	.1956	.0090
Part 2	.230	.6333	.2322	.0022
Part 3	.123	.8333	.1389	.0156
Part 4	.090	.1	.0900	0
Part 5	.170	.2333	.1789	.0098
Average	.160		.1671	.0071

Example 20 A Simple Numerical Example. For purposes of illustrating the formulas of this section, we will use a small numerical example due to Prof. Max Morris. Suppose that $I = 5$ parts are inspected by $J = 3$ operators, $m = 10$

times apiece, and that in Table 2.7 are sample fractions of “nonconforming” calls made by the operators and the corresponding estimates of per-call variance

Then the one-part-at-a-time and average-across-parts repeatability, R&R, and reproducibility estimates of variance are collected in Table 2.8 on page 83.

Then, for example, a fraction of only

$$\frac{.0071}{.1671} = 4.3\%$$

of the inconsistency in conforming/nonconforming calls seen in the original data seems to be attributable to clear differences in how the operators judge the parts (differences in the binomial “nonconforming call probabilities” p_j). Rather, the bulk of the variance seems to be attributable to unavoidable binomial variation. The p s are not close enough to either 0 or 1 to make the calls tend to be consistent. So the variation seen in the \hat{p} s in a given row is not clear evidence of large operator differences.

Of course, we need to remember that the computations above are on the variance (and not standard deviation) scale. On the (more natural) standard deviation scale, reproducibility variation

$$\sqrt{.0071} = .08$$

and repeatability variation

$$\sqrt{.160} = .40$$

are not quite so strikingly dissimilar.

2.6.3 Confidence Limits for Comparing Call Rates for Two Operators

It’s possible to use elementary confidence interval methods to compare call rates for two particular operators. This can be done for a particular fixed part or for “all” parts (supposing that the ones included in a study are a random sample from the universe of parts of interest). The first possibility can be viewed as the problem of estimating the difference in two binomial parameters, say p_1 and p_2 . The second can be approached as estimation of a mean difference in part-specific call rates, say μ_d .

A common elementary large-sample approximate confidence interval for $p_1 - p_2$ has end points

$$\hat{p}_1 - \hat{p}_2 \pm z \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

But, as it turns out, this formula can fail badly if either p is extreme or n is small. So we will use a slight modification that is more reliable, namely

$$\hat{p}_1 - \hat{p}_2 \pm z \sqrt{\frac{\tilde{p}_1 (1 - \tilde{p}_1)}{n_1} + \frac{\tilde{p}_2 (1 - \tilde{p}_2)}{n_2}} \quad (2.55)$$

Confidence
Limits for
 $p_1 - p_2$

where

$$\tilde{p}_i = \frac{n_i \hat{p}_i + 2}{n_i + 4} \quad (2.56)$$

Values to Use
in Formula
(2.55)

That is, under the square root of the usual formula, one essentially replaces the \hat{p} values with \tilde{p} values derived by adding two “successes” in four “additional trials” to the counts used to make up the \hat{p} values. (This has the effect of making the standard large-sample interval a bit wider and correcting the problem that without this modification for small sample sizes and extreme values of p , it can fail to hold its nominal confidence level.)

Example 21 (Example 20 continued.) Consider again part 1 from Example 20, and in particular consider the question of whether operator 1 and operator 2 have clearly different probabilities of calling that part nonconforming on a single call. With $\hat{p}_1 = .2$ and $\hat{p}_2 = .4$, formula (2.56) says that

$$\tilde{p}_1 = \frac{2 + 2}{10 + 4} = .2857 \text{ and } \tilde{p}_2 = \frac{4 + 2}{10 + 4} = .4286$$

so that using formula (2.55) approximate 95 % confidence limits for the difference $p_1 - p_2$ are

$$.2 - .4 \pm 1.96 \sqrt{\frac{.2857 (1 - .2857)}{10} + \frac{.4286 (1 - .4286)}{10}}$$

i.e.,

$$-.2 \pm .49$$

These limits cover 0 and there thus is no clear evidence in the $\hat{p}_1 = .2$ and $\hat{p}_2 = .4$ values (from the relatively small samples of sizes $m = 10$) that operators 1 and 2 have different probabilities of calling part 1 nonconforming.

The so-called “paired t ” confidence limits (2.10) for the mean of a difference $d = x_1 - x_2$ (say μ_d) based on a random sample of normally distributed values d_1, d_2, \dots, d_n are presented in most elementary statistics courses. While a difference in observed call rates for operators 1 and 2 on a particular part ($d = \hat{p}_1 - \hat{p}_2$) will typically not be normally distributed, for rough purposes it is adequate to appeal to the “robustness” of this inference method (the fact that it is widely believed to be effective even when normality is not wholly appropriate as a modeling assumption) and employ it to compare operator 1 and operator 2 in terms of average part-specific call rates.

Example 22 (Example 20 continued.) Consider again Example 20, and in particular the question of whether operator 1 and operator 2 have clearly different average (across parts) probabilities of calling parts nonconforming on a single call. The $n = 5$ differences in \hat{p} s for the two operators are

$$.2 - .4 = -.2, .6 - .6 = 0, 1.0 - .8 = .2, .1 - .1 = 0, \text{ and } .1 - .3 = -.2.$$

These numbers have sample mean $\bar{d} = -.04$ and sample standard deviation $s_d = .17$. Then using the fact that the upper 5% point of the t_4 distribution is 2.132, rough 90% two-sided confidence limits for the mean difference in call rates for the operators are

$$-.04 \pm 2.132 \frac{.17}{\sqrt{4}} \text{ that is, } -.04 \pm .18,$$

and there is not definitive evidence in Table 2.7 of a consistent difference in how operators 1 and 2 call parts on average.

Section 2.6 Exercises

- Suppose that ten parts are inspected by four operators 16 times apiece. Each inspection determines whether or not the item is conforming. The counts in the table below correspond to the numbers of “nonconforming” calls out of 16 inspections.

Part	Operator 1	Operator 2	Operator 3	Operator 4
1	10	11	11	10
2	11	9	12	10
3	8	8	9	7
4	15	14	14	16
5	12	14	11	12
6	15	15	16	15
7	14	11	14	12
8	16	16	15	15
9	13	15	14	15
10	16	15	16	16

- Using the data above, fill in the table below.

Part	\hat{p}	$\bar{p}(1-\bar{p})$	$\hat{p}(1-\hat{p})$
1			
2			
\vdots			
10			

- (b) What is the fraction of inconsistency in conforming/nonconforming calls that can be attributed to clear differences in how the operators judged the parts (differences in the binomial “nonconforming call probabilities” p_j)? (Make your answer on the variance scale.)
 - (c) What is the estimated reproducibility variation (on the standard deviation scale)?
 - (d) What is the estimated repeatability variation (on the standard deviation scale)?
 - (e) For part 10, give a 90 % confidence interval for the difference (operator 1 minus operator 3) in probabilities of a nonconforming call. Does it appear the operators 1 and 3 have different probabilities of a nonconforming call on *any* one of the parts? Why?
 - (f) Compare operator 1 and operator 3 average “nonconforming” call rates using 90 % two-sided confidence limits for a mean difference.
-

2.7 Chapter Summary

This chapter has been concerned with how measurement error impacts what can be learned from empirical data. It presented some ideas from the probability modeling of measurement variation and considered how the interpretation of elementary statistical inferences is affected by measurement error. Then a variety of more advanced statistical tools were discussed, because of their usefulness in quantifying, partitioning, and (in some cases) removing the effects of measurement variation in quality assurance and improvement projects.

2.8 Chapter 2 Exercises

1. Does a perfectly calibrated device return measurements of a measurand that are completely free of error? Explain.
2. Is a standard (an item with corresponding “known” measurand) needed in *both* device calibration and estimation of σ_{device} ? If not, which requires a standard? Explain.
3. A measurement device may have a bias as large as 1 unit (in absolute value) and a device standard deviation as large as 1 unit. You measure x and observe $y = 10$. If you believe in the simple (normal) measurement model and want to report an interval you are “at least 99 % sure” contains x , you should report what limits? (Hint: Before measurement, how far do you expect y to be from x with the indicated worst possible values of absolute

bias and standard deviation? Interpret “99 % sure” in “plus or minus three standard deviations” terms.)

4. The same axel diameter is measured $n_1 = 25$ times with device 1 and $n_2 = 25$ times with device 2, with resulting means and standard deviations $\bar{y}_1 = 2.001$ in, $\bar{y}_2 = 2.004$ in, $s_1 = .003$ in, and $s_2 = .004$ in. The upper 2.5 % point of the $F_{24,24}$ distribution is about 2.27.
 - (a) Give 95 % confidence limits for the difference in device biases.
 - (b) Give 95 % confidence limits for the ratio of the two device standard deviations.
 - (c) Is there a clear difference in device biases based on your interval in (a)? Why or why not?
 - (d) Is there a clear difference in device standard deviations based on your interval in (b)? Why or why not?
5. Two different (physically stable) production lines produce plastic pop bottles. Suppose $n_1 = 25$ bottles from line 1 and $n_2 = 25$ bottles from line 2 are burst tested on a single tester, with resulting means and standard deviations $\bar{y}_1 = 201$ psi, $\bar{y}_2 = 202$ psi, $s_1 = 3$ psi, and $s_2 = 4$ psi.
 - (a) Give a 95 % confidence interval for the difference between the mean burst strengths for lines 1 and 2 (line 1 minus line 2).
 - (b) Give a 95 % confidence interval for the ratio of burst strength standard deviations (line 1 divided by line 2). The upper 2.5 % point of the $F_{24,24}$ distribution is about 2.27.
 - (c) Is there a clear difference between mean burst strengths? Why or why not?
 - (d) Is there a clear difference between the consistencies of burst strengths? Why or why not?
6. Using a single tester, a single metal specimen was tested for Brinell hardness 20 times with resulting sample standard deviation of hardness 10 HB. Subsequently, 40 different specimens cut from the same ingot of steel have sample standard deviation of measured hardness 20 HB (using the same tester):
 - (a) Give 95 % confidence limits for a “test variability” standard deviation.
 - (b) Give approximate 95 % confidence limits for a specimen-to-specimen standard deviation of actual Brinell hardness.
7. An ANOVA analysis of a gauge R&R data set produced $\hat{\sigma}_{R\&R} = 53$ (in appropriate units) and $\hat{\nu}_{R\&R} = 3$. In these units, engineering specifications on a critical dimension of a machined steel part are *nominal* ± 200 . Give approximate 95 % confidence limits for a GCR (gauge capability ratio) for checking conformance to these specifications.

8. 95 % confidence limits for a particular gauge capability ratio are 6 to 8. What does this indicate about the usability of the gauge for checking conformance to the specifications under consideration?
9. Below is an analysis of variance table from a calibration study. The data were light intensities, y (in unspecified analyzer units), for specimens of known riboflavin concentration x (in $\mu\text{g}/\text{ml}$).

ANOVA table			
Source	SS	df	MS
Model	10946.445	1	10946.445
Error	27.155	8	3.4
Total	10973.6	9	

Parameter estimates for the simple linear regression model were $b_0 = 6.4634$ and $b_1 = 129.1768$.

- (a) Give a 95 % confidence interval for a repeatability standard deviation for this analyzer.
- (b) Suppose a new specimen with unknown concentration is analyzed and $y_{\text{new}} = 75$ is observed. Give a single-number estimate of the concentration in that specimen.
10. The final step in the production of some glass vials is a visual inspection presently carried out by human inspectors. A particular single vial (marked in an “invisible” ink that can be seen only under ultraviolet light) known to be defective is repeatedly run through the inspection process among a large number of newly produced vials. In fact, each of five company inspectors sees that vial ten times in a company study. Below are the rates at which that vial was identified as defective by the various operators (“1.0” means 100 %).

.6, .9, .9, 1.0, 1.0

- (a) In general, what two values of \hat{p} reflect perfect consistency of “defective/nondefective” calls made by a particular inspector?
- (b) What distribution models the number of correct “defective” calls (among ten calls) made by a particular inspector on the vial in question?
- (c) On the scale of (estimated) variances (not standard deviations), what is the fraction of overall variation seen in the “defective/nondefective” calls for this vial that should be attributed to operator-to-operator differences?
- (d) Give 95 % confidence limits for the long run difference in proportions of “defective” calls for the first operator (that made six out of ten “defective” calls) and the last operator (who made all “defective” calls).

11. **Laser Metal Cutting.** Davis, Martin and Poppinga used a ytterbium argon gas laser to make some cuts in 316 stainless steel. Using 95-MJ/pulse and 20-Hz settings on the laser and a 15.5-mm distance to the steel specimens (set at a 45° angle to the laser beam), the students made cuts in specimens using 100, 500, and 1000 pulses. The measured depths of four different cuts (in machine units) at each pulse level are given below (assume the same operator made all measurements and that repeatability variation is negligible here).

100 Pulses	500 Pulses	1000 Pulses
7.4, 8.6, 5.6, 8.0	24.2, 29.5, 26.5, 23.8	33.4, 37.5, 35.9, 34.8

- What is the response variable in this problem?
 - Give the sample average values for the 100, 500, and 1000 pulse levels. Calculate the sample range for the data at each pulse level. Give estimates of the standard deviation of cut depth for each level of pulse, first based on the sample range and then using the sample standard deviation. (You will have two estimates for each of the three population standard deviations.)
 - Assuming variability is the same for all three pulse levels, give an estimate of the common standard deviation based on the three sample ranges.
 - The concepts of measurement validity, precision, and accuracy are discussed in Sect. 2.1. The analysts decided to report the average cut depth for the different pulse levels. This averaging can be thought of in terms of improving which of (1) validity, (2) precision, or (3) accuracy (over the use of any single measurement)? The concept of calibration is most closely associated with which of the three?
12. **Fiber Angle.** Grunig, Hamdorf, Herman, and Potthoff studied a carpet-like product. They measured the angle at which fibers were glued to a sheet of base material. A piece of finished product was obtained and cut into five sections. Each of the four team members measured the fiber angle eight times for each section. The results of their measuring are given in Table 2.9 (in degrees above an undisclosed reference value). A corresponding ANOVA is also given in Table 2.10.
- Say what each term in the equation $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ means in this problem (including the subscripts i , j , and k).
 - Using ranges, estimate the repeatability and reproducibility standard deviations for angle measurement. Based on this analysis, what aspect of the measuring procedure seems to need the most attention? Explain.
 - Using ANOVA-based formulas, estimate the repeatability and reproducibility standard deviations for angle measurement. Is this analysis in essential agreement with that in part (b)? Explain.

TABLE 2.9. Data for problem 12

Angle	Analyst 1	Analyst 2	Analyst 3	Analyst 4
1	19, 20, 20, 23 20, 20, 20, 15	20, 25, 17, 22 23, 15, 23, 20	20, 19, 15, 16 20, 19, 12, 14	10, 10, 10, 5 5, 5, 5, 5
2	15, 17, 20, 20 10, 15, 15, 15	15, 13, 5, 10 8, 8, 10, 12	15, 20, 14, 16 13, 20, 15, 15	10, 10, 10, 10 10, 15, 15, 10
3	23, 20, 22, 20 25, 22, 20, 23	20, 23, 20, 20 23, 23, 22, 20	15, 20, 22, 18 15, 20, 16, 20	10, 10, 10, 15 15, 10, 10, 10
4	15, 16, 22, 15 15, 15, 22, 17	20, 22, 18, 23 23, 23, 24, 20	13, 13, 15, 20 11, 20, 13, 15	5, 10, 10, 10 10, 10, 10, 10
5	20, 20, 22, 20 27, 17, 20, 15	18, 20, 18, 23 20, 20, 18, 15	10, 14, 17, 12 11, 10, 15, 10	5, 10, 10, 10 10, 10, 10, 10

TABLE 2.10. ANOVA for problem 12

Source	SS	df	MS
Angle	390.913	4	97.728
Analyst	2217.15	3	739.05
Angle \times Analyst	797.788	12	66.482
Error	971.75	140	6.941
Total	4377.6	159	

- (d) Using your answer to (c), give an estimate of the standard deviation that would be experienced by many analysts making a single measurement on the same angle (in the same section) assuming there is no repeatability component to the overall variation.
- (e) Specifications on the fiber angle are *nominal* $\pm 5^\circ$. Estimate the gauge capability ratio using first the ranges and then ANOVA-based estimates. Does it appear this measurement method is adequate to check conformance to the specifications? Why or why not?

13. Refer to the **Fiber Angle** case in problem 12.

- (a) Is it preferable to have eight measurements on a given section by each analyst as opposed to, say, two measurements on a given section by each analyst? Why or why not?
- (b) For a given number of angle measurements per analyst \times section combination, is it preferable to have four analysts instead of two, six, or eight? Why or why not?

- (c) When making angle measurements for a given section, does it matter if the angle at a fixed location on the piece is repeatedly measured, or is it acceptable (or even preferable?) for each analyst to measure at eight different locations on the section? Discuss.
- (d) Continuing with (c), does it matter that the locations used on a given section varied analyst to analyst? Why or why not?

14. **Bolt Shanks.** A 1-in micrometer is used by an aircraft engine manufacturer to measure the diameter of a body-bound bolt shank. Specifications on this dimension have been set with a spread of .002 in. Three operators and ten body-bound bolt shanks were used in a gauge R&R study. Each bolt shank was measured twice by each operator (starting with part 1 and proceeding sequentially to part 10) to produce the data in Table 2.11 (in inches). A corresponding ANOVA is provided in Table 2.12 as well (SSs and MSs are in 10^{-6} in^2).

- (a) Plot the bolt shank diameter measurements versus part number using a different plotting symbol for each operator. (You may wish to also plot part \times operator means and connect consecutive ones for a given operator with line segments.) Discuss what your plot reveals about the measurement system.

TABLE 2.11. Data for problem 14

Operator				Operator			
Part	A	B	C	Part	A	B	C
1	.3473	.3467	.3472	6	.3472	.3463	.3471
	.3473	.3465	.3471		.3472	.3464	.3471
2	.3471	.3465	.3471	7	.3473	.3465	.3472
	.3471	.3464	.3471		.3473	.3469	.3471
3	.3472	.3467	.3471	8	.3474	.3470	.3473
	.3472	.3464	.3471		.3473	.3470	.3473
4	.3474	.3470	.3473	9	.3472	.3465	.3472
	.3475	.3470	.3474		.3472	.3466	.3471
5	.3474	.3470	.3473	10	.3474	.3470	.3474
	.3474	.3470	.3473		.3474	.3470	.3473

- (b) Find an ANOVA-based estimate of repeatability standard deviation.
- (c) Find an ANOVA-based estimated standard deviation for reproducibility assuming there is no repeatability component of variation.
- (d) Using your answers to (b) and (c), estimate the percent of total (R&R) measurement variance due to repeatability.
- (e) Using your answers to (b) and (c), estimate the percent of total measurement (R&R) variance due to reproducibility.

TABLE 2.12. ANOVA for problem 14

Source	SS	df	MS
Part	1.3	9	.145
Operator	3.78	2	1.89
Part×Operator	.321	18	.0178
Error	.195	30	.0065
Total	5.601	59	

- (f) Discuss the relationship of your plot in (a) to your answers to (b) through (e).
- (g) Find an ANOVA-based estimate of the gauge capability ratio. Is the measurement process acceptable for checking conformance to the specifications? Why or why not?
15. Refer to the **Bolt Shanks** case in problem 14. The data in Table 2.13 are from three new operators with a different set of ten body-bound bolt shanks (numbered as part 11 through part 20). An appropriate ANOVA is also provided for these new data in Table 2.14 (units for the SS's and MS's are 10^{-6} in^2).

TABLE 2.13. Data for problem 15

Operator				Operator			
Part	D	E	F	Part	D	E	F
11	.3694	.3693	.3693	16	.3692	.3692	.3692
	.3694	.3693	.3693		.3693	.3692	.3691
12	.3693	.3693	.3692	17	.3696	.3695	.3695
	.3693	.3692	.3692		.3696	.3695	.3695
13	.3698	.3697	.3697	18	.3697	.3696	.3696
	.3697	.3697	.3697		.3696	.3696	.3696
14	.3697	.3698	.3697	19	.3697	.3696	.3695
	.3696	.3697	.3697		.3696	.3695	.3696
15	.3694	.3695	.3695	20	.3697	.3697	.3698
	.3693	.3695	.3694		.3697	.3698	.3697

- (a) Answer (a) through (g) from problem 14 for these new data.
- (b) Are your answers to (a) qualitatively different than those for problem 14? If your answer is yes, in what ways do the results differ, and what might be the sources of the differences?
- (c) Do conclusions from this R&R study indicate a more consistent measurement process for body-bound bolt shanks than those in problem 14? Why or why not?

TABLE 2.14. ANOVA for problem 15

Source	SS	df	MS
Part	2.08	9	.231
Operator	.016	2	.008
Part \times Operator	.0873	18	.00485
Error	.07	30	.00233
Total	2.254	59	

16. **Transmission Gear Measurement.** Cummins, Rosario, and Vanek studied two gauges used to measure ring gear height and bevel gear height in the production of transmission differentials. (Ring gear height and bevel gear height determine the milling points for the customized transmission housings, creating the horizontal location in the housing and the “tightness” of the casing against the differential.) A test stand (hydraulically) puts a 1000 pound force on the differential. This force is used to keep the differential from free spinning while allowing spin with some force applied. A 3- in Mitutoyo digital depth micrometer and a 6- in Mitutoyo digital depth micrometer were used to make the measurements. Vanek used the 3- in micrometer and took two ring gear height measurements on differential 8D4. Using the same 3- in Mitutoyo micrometer, Cummins made two ring gear height measurements on the same part. Vanek then took two bevel gear height measurements with the 6- in Mitutoyo micrometer on the same differential. Cummins followed with the same 6- in micrometer and took two bevel gear height measurements on differential 8D4. This protocol was repeated two more times for the differential 8D4. The whole procedure was then applied to differential 31D4. The data are given in Table 2.15. ANOVAs are given for both the ring gear data (SS and MS units are 10^{-4} in²) and the bevel gear data (SS and MS units are 10^{-5} in²) in Tables 2.16 and 2.17, respectively.

- Consider the ring gear heights measured with the 3- in Mitutoyo micrometer. Give the values of m , I , and J .
- In the context of the ring gear height measurements, what do m , I , and J represent?
- Give an ANOVA-based estimated repeatability standard deviation for ring gear height measuring. Find a range-based estimate of this quantity.
- Give an ANOVA-based estimated reproducibility standard deviation for ring gear height measuring.
- The upper and lower specifications for ring gear heights are, respectively, 1.92 in and 1.88 in. If the company requires the gauge capability ratio to be no larger than .05, does the 3- in Mitutoyo micrometer, as currently used, seem to meet this requirement? Why or why not?

TABLE 2.15. Data for problem 16

Ring gear heights (inches) (3- in Mitutoyo micrometer)			Bevel gear heights (inches) (6- in Mitutoyo micrometer)		
	Vanek	Cummins		Vanek	Cummins
8D4	1.88515	1.88470	8D4	5.49950	5.49850
	1.88515	1.88470		5.49985	5.49945
	1.88540	1.88380		5.49975	5.49945
	1.88530	1.88510		5.50000	5.50005
	1.88485	1.88435		5.49930	5.50070
	1.88490	1.88450		5.49945	5.49945
31D4	1.88365	1.88270	31D4	5.49785	5.49700
	1.88370	1.88295		5.49775	5.49710
	1.88330	1.88235		5.49765	5.49615
	1.88325	1.88235		5.49750	5.49615
	1.88270	1.88280		5.49670	5.49595
	1.88265	1.88260		5.49680	5.49620

TABLE 2.16. ANOVA for problem 16 ring gear data

Source	SS	df	MS
Differential	.219	1	.219
Operator	.021	1	.021
Differential×Operator	.0000042	1	.0000042
Error	.0249	20	.00124
Total	.2644	23	

TABLE 2.17. ANOVA for problem 16 bevel gear data

Source	SS	df	MS
Differential	4.44	1	4.44
Operator	.148	1	.148
Differential×Operator	.124	1	.124
Error	.550	20	.02752
Total	5.262	23	

- (f) Repeat (a) through (e) for bevel gear heights measured with the 6-in Mitutoyo micrometer. Lower and upper specifications are, respectively, 5.50 in and 5.53 in for the bevel gear heights.

17. **Computer Locks.** Cheng, Lourits, Hugraha, and Sarief decided to study “tip diameter” for some computer safety locks produced by a campus machine shop. The team began its work with an evaluation of measurement precision for tip diameters. The data in Table 2.18 are in inches and represent two diameter measurements for each of two analysts made on all 25 locks machined on one day. An appropriate ANOVA is also given in Table 2.19. (The units for the SSs and MSs are 10^{-4} in^2 .)

TABLE 2.18. Data for problem 17

Part	Lourits	Cheng	Part	Lourits	Cheng
1	.375, .375	.374, .374	14	.373, .373	.379, .374
2	.375, .375	.377, .376	15	.372, .373	.374, .373
3	.375, .373	.374, .375	16	.373, .373	.374, .374
4	.375, .373	.375, .374	17	.373, .373	.374, .373
5	.374, .374	.374, .374	18	.373, .373	.373, .373
6	.374, .374	.374, .375	19	.373, .373	.376, .373
7	.374, .375	.375, .376	20	.373, .373	.373, .373
8	.374, .375	.374, .373	21	.374, .374	.374, .375
9	.374, .374	.375, .375	22	.375, .375	.374, .377
10	.374, .374	.374, .374	23	.375, .375	.376, .377
11	.375, .373	.374, .374	24	.376, .375	.376, .374
12	.375, .374	.376, .374	25	.374, .374	.374, .375
13	.376, .373	.373, .374			

TABLE 2.19. ANOVA for problem 17

Source	SS	df	MS
Part	.58	24	.0242
Operator	.0625	1	.0625
Part \times Operator	.22	24	.00917
Error	.445	50	.0089
Total	1.3075	99	

- (a) Organizations typically establish their own guidelines for interpreting the results of gauge R&R studies. One set of guidelines is shown below. ($6\hat{\sigma}_{\text{repeatability}} \div (U - L)$ expressed as a percentage is sometimes called the “% gauge” for repeatability. $6\hat{\sigma}_{\text{reproducibility}} \div (U - L)$ expressed as a percentage is sometimes called the “% gauge” for reproducibility.)

% gauge	Rating
33 %	Unacceptable
20 %	Marginal
10 %	Acceptable
2 %	Good
1 %	Excellent

Suppose that specifications for the lock tip diameters are $.375 \pm .002$ in. According to the guidelines above and using ANOVA-based estimates, how does the diameter measuring process “rate” (based on “% gauge” for repeatability and “% gauge” for reproducibility)? Why?

- (b) Find expressions for $\bar{y}_{\text{operator1}}$ and $\bar{y}_{\text{operator2}}$ as functions of the model terms used in the equation $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$.
- (c) Continuing with (b) and applying logic consistent with that used to develop Eq. (2.31), what does $|\bar{y}_{\text{operator1}} - \bar{y}_{\text{operator2}}|/d_2(2)$ estimate in terms of σ_α^2 , σ_β^2 , $\sigma_{\alpha\beta}^2$, and σ^2 ?
18. Refer to the **Computer Locks** case in problem 17. Consider the measurements made by Lourits. The sample average tip diameter for the i th randomly selected lock measured by Lourits can be written (holding only Lourits fixed) as

$$\bar{y}_{i\text{Lourits}} = \mu + \alpha_i + \beta_{\text{Lourits}} + \alpha\beta_{i\text{Lourits}} + \bar{\epsilon}_{i\text{Lourits}}.$$

- (a) What is the random portion of $\bar{y}_{i\text{Lourits}}$?
- (b) In terms of σ^2 , σ_α^2 , σ_β^2 , and $\sigma_{\alpha\beta}^2$, give the variance of your answer to part (a).
- (c) Letting Γ be the range of the 25 variables $\bar{y}_{i\text{Lourits}}$, what does $\Gamma/d_2(25)$ estimate?
- (d) Give the observed numerical value for $\Gamma/d_2(25)$ considered in part (c).
- (e) In terms of σ^2 , σ_α^2 , σ_β^2 , and $\sigma_{\alpha\beta}^2$, what is the variance of (different) lock tip diameters as measured by a single operator (say Lourits) assuming there is no repeatability variation?
- (f) In terms of σ^2 , σ_α^2 , σ_β^2 , and $\sigma_{\alpha\beta}^2$, what is the variance of (single-) diameter measurements made on (different) lock tips made by the same operator (say Lourits)? (Hint: This is your answer to (e) plus the repeatability variance, σ^2 .)
- (g) Using the Lourits data, find a range-based estimate of the repeatability variance.
- (h) Using the Lourits data, find a range-based estimate of your answer to (e). (Hint: Use your answers for (d) and (g) appropriately.)

- (i) Using the Lourits data, estimate your answer to (f). (Hint: Use your answers for (h) and (g) appropriately.)

19. **Implement Hardness.** Olsen, Hegstrom, and Casterton worked with a farm implement manufacturer on the hardness of a steel part. Before process monitoring and experimental design methodology were considered, the consistency of relevant hardness measurement was evaluated. Nine parts were obtained from a production line, and three operators agreed to participate in the measuring process evaluation. Each operator made two readings on each of nine parts. The data in Table 2.20 are in mm. An appropriate ANOVA is given in Table 2.21 (the units for the SSs and MSs are mm^2).

TABLE 2.20. Data for problem 19

Part	Operator			Part	Operator		
	A	B	C		A	B	C
1	3.30	3.25	3.30	6	3.30	3.30	3.25
	3.30	3.30	3.30		3.30	3.20	3.20
2	3.20	3.20	3.15	7	3.15	3.10	3.15
	3.25	3.30	3.30		3.20	3.20	3.20
3	3.20	3.20	3.25	8	3.25	3.20	3.20
	3.30	3.20	3.20		3.20	3.20	3.25
4	3.25	3.20	3.20	9	3.25	3.20	3.30
	3.30	3.25	3.20		3.30	3.30	3.40
5	3.25	3.10	3.20				
	3.30	3.10	3.15				

TABLE 2.21. ANOVA for problem 19

Source	SS	df	MS
Part	.08833	8	.01104
Operator	.01778	2	.00889
Part \times Operator	.04139	16	.00259
Error	.0575	27	.002130
Total	.205	59	

- (a) Say what each term in Eq. (2.26) means in the context of this problem.
- (b) What are the values of I , J , and m in this study?
- (c) Give an ANOVA-based estimate of the repeatability standard deviation, σ .
- (d) Give an ANOVA-based estimate of the reproducibility standard deviation, $\sigma_{\text{reproducibility}}$.

- (e) Estimate the gauge capability ratio using the ANOVA-based calculation if specifications on the hardness of this part are *nominal* ± 10 mm.
 - (f) Using the corporate gauge rating table given in problem 17, rate the repeatability and the reproducibility of the hardness measurement method.
 - (g) Does it appear the current measuring process is adequate to check conformance to *nominal* ± 10 mm hardness specifications? Why or why not?
20. Refer to the **Implement Hardness** case in problem 19.
- (a) Suppose each operator used a different gauge to measure hardness. How would this affect the interpretation of your calculations in problem 19?
 - (b) If it were known that measuring alters the part hardness in the vicinity of the point tested, how should this be addressed in a gauge R&R study?
 - (c) When an operator measures the same part two times in a row, it is likely the second measurement is “influenced” by the first in the sense that there is psychological pressure to produce a second measurement like the initial one. How might this affect results in a gauge R&R study? How could this problem be addressed/eliminated?
21. Is it important to include an evaluation of measuring processes early in a quality improvement effort? Why or why not?
22. Management tells engineers involved in a quality improvement project “We did a gauge R&R study last year and the estimated gauge capability ratio was .005. You don’t need to redo the study.” How should the engineers respond and why?
23. **Paper Weight.** Everingham, Hart, Hartong, Spears, and Jobe studied the top loading balance used by the Paper Science Department at Miami University, Oxford, Ohio. Two $20\text{ cm} \times 20\text{ cm}$ (400 cm^2) pieces of 20 lb bond paper were cut from several hundred feet of paper made in a departmental laboratory. Weights of the pieces obtained using the balance are given below in grams. The numbers in parentheses specify the order in which the measurements were made. (Piece 1 was measured 15 times, three times by each operator. That is, piece 1 was measured first by Spears, second by Spears, third by Hart, ..., 14th by Hartong, and lastly by Jobe.) Different orders were used for pieces 1 and 2, and both were determined using a random number generator. Usually, the upper specification minus the lower specification ($U - L$) is about 4 g/m^2 for the density of this type of paper. An appropriate ANOVA is given below (units for the SSs and MSs are g^2).

Piece	Hartong	Hart	Spears	Everingham	Jobe
1	(14) 3.481	(3) 3.448	(1) 3.485	(13) 3.475	(10) 3.472
	(12) 3.477	(9) 3.472	(2) 3.464	(4) 3.472	(5) 3.470
	(7) 3.470	(6) 3.470	(11) 3.477	(8) 3.473	(15) 3.474
2	(1) 3.258	(13) 3.245	(7) 3.256	(6) 3.249	(11) 3.241
	(2) 3.254	(12) 3.247	(5) 3.257	(15) 3.238	(8) 3.250
	(3) 3.258	(9) 3.239	(10) 3.245	(14) 3.240	(4) 3.254

ANOVA table for weight

Source	SS	df	MS
Piece	.37386	1	.37386
Operator	.00061	4	.000152
Piece \times Operator	.00013	4	.000032
Error	.00095	20	.000047
Total	.37555	29	

- What purpose is potentially served by randomizing the order of measurement as was done in this study?
- Give the table of operator \times piece ranges, R_{ij} .
- Give the table of operator \times piece averages, \bar{y}_{ij} .
- Give the ranges of the operator \times piece means, Δ_i .
- Express the observed weight range determined by Spears for piece 2 in g/m^2 . (Note: $10^4 \text{ cm}^2 = 1 \text{ m}^2$.)
- Find a gauge repeatability rating based on ranges. (See part (a) of problem 17.) Pay attention to units.
- Find a gauge reproducibility rating based on ranges. (Again see part (a) of problem 17 and pay attention to units.)
- Calculate an estimated gauge capability ratio. Pay attention to units.
- What minimum value for $(U - L)$ would guarantee an estimated gauge capability ratio of at most .1?
- Using ANOVA-based estimates, answer (f)–(h).
- Using ANOVA-based estimates, give an exact 95 % confidence interval for $\sigma_{\text{repeatability}}$. Your units should be g/m^2 .
- Using the ANOVA-based estimates, give 95 % approximate confidence limits for $\sigma_{\text{reproducibility}}$. Your units should be g/m^2 .

24. **Paper Thickness.** Everingham, Hart, Hartong, Spears, and Jobe continued their evaluation of the measuring equipment in the Paper Science Laboratory at Miami University by investigating the repeatability and reproducibility of the TMI automatic micrometer routinely used to measure

paper thickness. The same two $20\text{ cm} \times 20\text{ cm}$ pieces of 20 lb bond paper referred to in problem 23 were used in this study. But unlike measuring weight, measuring thickness alters the properties of the portion of the paper tested (by compressing it and thus changing the thickness). So an 8×8 grid was marked on each piece of paper. The corresponding squares were labeled 1, 2, . . . , 64 left to right, top to bottom. Ten squares from a given piece were randomly allocated to each operator (50 squares from each piece were measured). Because so many measurements were to be made, only the “turn” for each analyst was determined randomly, and each operator made all ten of his measurements on a given piece consecutively. A second randomization and corresponding order of measurement was made for piece 2. Hartong measured third on piece 1 and fifth on piece 2, Hart was first on piece 1 and third on piece 2, Spears was fifth and fourth, Everingham was second and second, and Jobe was fourth and first. The data are in Table 2.22 (in mm). The numbers in parenthesis identify the squares (from a given piece) measured. (Thus, for piece 1, Hart began the measurement procedure by recording thicknesses for squares 51, 54, 18, 63, . . . , 7; then Everingham measured squares 33, 38, . . . , 5, etc. After the data for piece 1 were obtained, measurement on piece 2 began. Jobe measured squares 9, 3, . . . , 22; then Everingham measured squares 43, 21, . . . , 57, etc.) An appropriate ANOVA is also given in Table 2.23 (units for the SSs and MSs are mm^2).

- (a) Say what each term in Eq. (2.26) means in the context of this problem.
- (b) How is this study different from a “garden-variety” gauge R&R study?
- (c) Will the nonstandard feature of this study tend to increase, decrease, or have no effect on the estimate of the repeatability standard deviation? Why?
- (d) Will the nonstandard feature of this study tend to increase, decrease, or have no effect on the estimated standard deviation of measurements from a given piece across many operators? Why?
- (e) Give the ANOVA-based estimated standard deviation of paper thickness measurements for a fixed piece \times operator combination, i.e., approximate the repeatability standard deviation assuming that square-to-square variation is negligible.
- (f) Give the ANOVA-based estimated standard deviation of thicknesses measured on a fixed piece across many operators. (The quantity being estimated should include but not be limited to variability for a fixed piece \times operator combination.) That is, approximate the reproducibility standard deviation assuming square-to-square variation is negligible.
- (g) What percent of the overall measurement variance is due to repeatability? What part is due to reproducibility?

TABLE 2.22. Data for problem 24

Piece	Hartong	Hart	Spears	Everingham	Jobe
1	(14) .201	(51) .195	(48) .192	(33) .183	(43) .185
	(25) .190	(54) .210	(58) .191	(38) .189	(40) .204
	(17) .190	(18) .200	(15) .198	(36) .196	(49) .194
	(21) .194	(63) .203	(55) .197	(3) .195	(12) .199
	(53) .212	(20) .196	(44) .207	(59) .192	(29) .192
	(16) .209	(50) .189	(23) .202	(45) .195	(13) .193
	(47) .208	(31) .205	(64) .196	(41) .185	(56) .190
	(42) .192	(37) .203	(57) .188	(9) .193	(2) .195
	(22) .198	(34) .195	(26) .201	(62) .194	(8) .199
	(35) .191	(7) .186	(1) .181	(5) .194	(6) .197
2	(5) .188	(14) .186	(55) .177	(43) .179	(9) .191
	(16) .173	(24) .171	(51) .174	(21) .194	(3) .180
	(11) .188	(62) .178	(36) .184	(18) .187	(42) .194
	(47) .180	(34) .175	(12) .180	(39) .175	(50) .183
	(25) .178	(29) .183	(38) .179	(6) .173	(53) .181
	(15) .188	(10) .185	(41) .186	(7) .179	(17) .188
	(56) .166	(30) .190	(63) .183	(64) .171	(33) .188
	(26) .173	(40) .177	(45) .172	(54) .184	(23) .173
	(8) .175	(58) .184	(31) .174	(59) .181	(60) .180
	(52) .183	(13) .186	(2) .178	(57) .187	(22) .176

TABLE 2.23. ANOVA for problem 24

Source	SS	df	MS
Piece	.00557	1	.00557
Operator	.00018	4	.000045
Piece×Operator	.00028	4	.00007
Error	.003986	90	.000044
Total	.010013	99	

25. **Paper Burst Strength.** An important property of finished paper is the force (lb/in^2) required to burst or break through it. Everingham, Hart, Hartong, Spears, and Jobe investigated the repeatability and reproducibility of existing measurement technology for this paper property. A Mullen tester in the Miami University Paper Science Department was studied. Since the same two $20\text{ cm} \times 20\text{ cm}$ pieces of paper referred to in problems 23 and 24 were available, the team used them in its gauge R&R study for burst strength measurement. The burst test destroys the portion of paper tested, so repeat measurement of exactly the same paper specimen is not possible. Hence, a grid of 10 approximately equal-sized rectangles, $10\text{ cm} \times 4\text{ cm}$ (each large

enough for the burst tester), was marked on each large paper piece. Each of the analysts was assigned to measure burst strength on two randomly selected rectangles from each piece. The measurement order was also randomized among the five operators for each paper piece. The data obtained are shown below. The ordered pairs specify the rectangle measured and the order of measurement. (For example, the ordered pair (2,9) in the top half of the table indicates that 8.8 lb/in^2 was obtained from rectangle number 2, the ninth rectangle measured from piece 1.) An ANOVA table for this study is also provided.

Piece	Hartong	Hart	Spears	Everingham	Jobe
1	(9,2) 13.5	(6,6) 10.5	(4,8) 12.9	(2,9) 8.8	(3,10) 12.4
	(7,5) 14.8	(5,1) 11.7	(1,4) 12.0	(8,3) 13.5	(10,7) 16.0
2	(3,9) 11.3	(1,8) 14.0	(5,6) 13.0	(6,7) 12.6	(2,1) 11.0
	(8,10) 12.0	(7,5) 12.5	(9,3) 13.1	(4,2) 12.7	(10,4) 10.6

ANOVA table for burst strength			
Source	SS	df	MS
Piece	.5445	1	.5445
Operator	2.692	4	.6730
Piece \times Operator	24.498	4	6.1245
Error	20.955	10	2.0955
Total	48.6895	19	

In the following, assume that specimen-to-specimen variation within a given piece of paper is negligible.

- To what set of operators can the conclusions of this study be applied?
- To what set of paper pieces can the conclusions of this study correctly be applied?
- What are the values of I , J , and m in this study?
- Give an ANOVA-based estimate of the repeatability standard deviation, σ .
- Give another estimate of the repeatability standard deviation, σ , this time based on ranges.
- Find an ANOVA-based estimate of $\sigma_{\text{reproducibility}}$.
- Find another estimate of $\sigma_{\text{reproducibility}}$, this one based on ranges
- Using the ANOVA, estimate the standard deviation of single burst measurements on a fixed piece of paper made by many operators, $\sigma_{\text{R\&R}}$.

26. **Paper Tensile Strength.** The final type of measurement method studied by Everingham, Hart, Hartong, Spears, and Jobe in the Paper Science Laboratory at Miami University was that for paper tensile strength. Since the burst tests discussed in problem 25 destroyed the $20\text{ cm} \times 20\text{ cm}$ pieces of 20 lb bond paper referred to there, two new $20\text{ cm} \times 20\text{ cm}$ pieces of paper were selected from the same run of paper. Ten $15\text{ cm} \times 20\text{ cm}$ strips were cut from each $20\text{ cm} \times 20\text{ cm}$ piece. Each set of ten strips was randomly allocated among the five operators (two strips per operator for each set of ten). The order of testing was randomized for the ten strips from each piece, and the same Thwing-Albert Intellect 500 tensile tester was used by each operator to measure the load required to pull apart the strips. The data appear below in kg. (Consider, e.g., the data given for piece 1, Hartong, (9,2) 4.95. A 4.95 - kg load was required to tear strip number 9 from piece 1 and the measurement was taken second in order among the ten strips measured for piece 1.) Since the testing destroyed the strip, the analysts had to assume strip-to-strip variation for a given piece to be negligible. An appropriate ANOVA is also given below (units for SSs and MSs are kg^2).

Piece	Everingham	Hart	Hartong	Spears	Jobe
1	(2,8) 4.34	(1,5) 4.34	(9,2) 4.95	(6,6) 4.03	(10,4) 4.51
	(8,10) 4.71	(4,3) 4.61	(7,7) 4.53	(3,9) 3.62	(5,1) 4.56
2	(4,7) 5.65	(6,6) 4.80	(1,1) 4.38	(2,2) 4.65	(9,5) 4.30
	(8,9) 4.51	(10,8) 4.75	(3,3) 3.89	(5,4) 5.06	(7,10) 3.87

ANOVA table for tensile strength			
Source	SS	df	MS
Piece	.13778	1	.1378
Operator	.69077	4	.17269
Piece \times Operator	1.88967	4	.47242
Error	1.226	10	.1226
Total	3.9442	19	

- Make a table of load averages, \bar{y}_{ij} , for the ten operator \times piece combinations.
- Plot the load averages \bar{y}_{ij} versus piece number for each of the operators (connect the two \bar{y}_{ij} s for each operator).
- Suppose the target tensile strength for strips of 20 lb bond paper is 4.8 kg. Typically, upper and lower specifications for paper properties are set 5% above and below a target. Estimate the gauge capability ratio under these conditions, using ANOVA-based calculations.

- (d) If upper and lower specifications for tensile strength of 20 lb bond paper are equal distances above and below a target of 4.8 kg, find the upper and lower limits such that the estimated gauge capability ratio is .01.
- (e) Redo part (d) for an estimated gauge capability ratio of .1.
- (f) Is it easier to make a gauge capability ratio better (smaller) by increasing its denominator or decreasing its numerator? Will your answer lead to a more consistent final product? Why or why not?

<http://www.springer.com/978-0-387-79105-0>

Statistical Methods for Quality Assurance
Basics, Measurement, Control, Capability, and
Improvement

Vardeman, S.B.; Jobe, J.M.

2016, XIV, 437 p. 104 illus., 99 illus. in color., Softcover

ISBN: 978-0-387-79105-0