Using Data for Laboratory Performance Improvement

Lynne B. Hare[1]
Statistical Strategies, LLC, Plymouth, MA, U.S.A.

ABSTRACT

The quality of a laboratory's most important product, information, can be assured through the proper use of statistical tools that support the quest for method accuracy and precision together with routine programs designed to assure their consistency over time. Methods for applying these tools are presented, along with examples and recommendations for routine applications. An additional topic, useful in assuring analytical accuracy, is the determination of sample sizes needed to detect crucial product defect and contamination levels.

"Healthy" analytical laboratories contribute to organizational success by delivering accurate and precise information to both internal and external clients of their services. A high-quality contribution does not happen by accident. It requires hard, continuous work and great attention to detail. Managers of analytical laboratories may entreat their staff members to pay close attention to protocol, engage in discussions of analytical procedures with colleagues, and compare occasional matched samples. These efforts, however, usually wane and lack sufficient impact to assure the long-term, consistent accuracy and precision that results in the client credibility necessary to maintain a truly productive laboratory that provides high-quality contributions.

For laboratory managers to lead their organizations to a healthier status, they must remember that all analytical values are estimates of the true state of nature and that the state of nature is inevitably contaminated by "error" associated with experimental variation. Error in this sense, is the difference between result and actuality. The estimate of the natural state is often referred to as a "signal," and the associated error (or variation) is referred to as "noise." Much focus should be directed to the signal-to-noise ratio. It has been said that there is never a signal without noise. More information on this subject can be found in Box (1) and Box et al. (2).

The immediate focus of a laboratory must be on attaining analytical accuracy and precision. An overall technology for obtaining the necessary data to measure both is measurement systems analysis (MSA). This is a divide-and-conquer process for first measuring accuracy in terms of linearity and bias against known values and then measuring precision by its component parts of reproducibility and repeatability. Reproducibility refers to consistency across equipment and analysts, whereas repeatability addresses uniformity among replicate readings by one analyst using the same equipment.

Accuracy can be assessed through the application of regression models as aids to calibration and detection of bias. More information on regression analysis can be found in Montgomery et al. (11).

Precision is measured by specially designed experiments tailored to the laboratory equipment and staff configuration. As a result, there are many varieties of these studies, commonly called Gage (or Gauge) repeatability and reproducibility (R&R) studies. Much of the literature on Six Sigma contains sections on MSA, including Gage R&R. Leadership aspects of Six Sigma, including MSA, are presented by Snee and Hoerl (12), whereas technical, statistical details are offered by Breyfogle (3) and Hare (5).

A laboratory's credibility is increased when its staff reports response estimates accompanied by intervals stating the uncertainty of the estimates. This can take the form of confidence intervals, standard deviation or standard error estimates, or any of a number of other expressions of uncertainty. More information on statistical intervals can be found in Meeker et al. (8).

As important as assessment of methodological accuracy and precision are, laboratory health does not end there; as with good human health, good laboratory health must be maintained. One maintenance device is the control chart. It provides a graphical representation of laboratory health by displaying the means and standard deviations among blind replicate samples over time. As such, it aids in the detection of unusual results and trends away from stability. Montgomery (10) provides details on many types of control charts and examples of their use.

Another maintenance device is the interlaboratory check sample program. As the name suggests, such a program is designed to compare the output of multiple laboratories in an effort to assure uniformity among laboratories within an association of laboratories commonly assessing the same analytes. As with Gage R&R studies, design-of-experiments technology is used to plan the allocation of samples to potential sources of variation. With interlaboratory check samples, however, the sources of variation include laboratory differences and may also include differences among technicians within laboratories, equipment differences, and even protocol differences. Designs can be quite complex, but they should include blindly submitted random samples and sample replication. Details of the statistical design of experiments are described in Box et al. (2) and Montgomery (9).

Laboratory managers and staff have opportunities to extend their reach by offering advice on the determination of the sample sizes necessary to detect differences that may be important to business success. Although this might typically be considered the domain of the statistician, questions of sample size might be better answered through multidisciplinary collaboration. Background information can be found in Hogg et al. (6), as well as other references on statistical inference.

In the following sections each of these topics is discussed and examples are presented.

Accuracy

The assessment of method accuracy usually is made against a known standard, and linear regression models are used to learn about the relationship between the unknown and the known. Some hypothetical data are shown in Table I. Vacuum-oven moisture data from 15 production samples are listed together with the corresponding readings from two competitive moisture

[1] Tel: +1.908.627.2309; E-mail: lynne.hare@comcast.net

https://doi.org/10.1094/ CFW-63-1-0006

meters. When known standards are not available, chemists must resort to other special methods, such as the use of known ingredient additions. These methods are not discussed here.

My first rule of data analysis is, "Always, always, always, without exception, plot the data—and look at the plot." One would think that the "look at the plot" part would not be necessary, but experience suggests otherwise. Moisture readings for each of the two meters plotted against their corresponding vacuum-oven moisture values are shown in Figure 1.

A close look shows that a linear fit for meter 1 might be reasonable, whereas the same fit for meter 2 might not. The graphical results of the two linear fits with confidence intervals shown in Figure 2 bear this out—notice the downward bias near the center of the vacuum-oven data for meter 2.

The linear, or first-order, model is

$$\eta = \beta_0 + \beta_1 x + \varepsilon \qquad (1)$$

and the second-order, or quadratic, model is

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad (2)$$

Table I. Hypothetical vacuum-oven moisture determinations with corresponding readings from two competitive moisture meters (data are sorted by vacuum-oven moisture)

Sample	Vacuum-Oven Moisture (%)	Moisture Meter 1 Reading (%)	Moisture Meter 2 Reading (%)
1	10.1	10.5	11.0
2	10.5	10.2	10.8
3	11.1	11.0	10.7
4	11.6	11.7	11.0
5	11.7	11.6	10.9
6	12.7	12.3	11.8
7	13.0	13.2	11.7
8	13.2	13.0	12.0
9	13.2	12.8	11.9
10	13.6	13.4	12.6
11	14.4	14.2	14.4
12	14.7	15.0	15.0
13	15.3	15.2	15.7
14	15.5	15.5	16.2
15	15.7	15.7	16.4

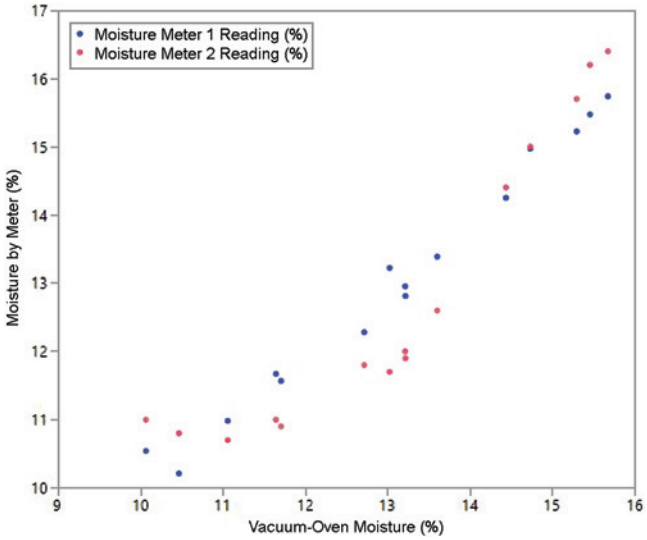


Fig. 1. A scatter plot of readings from two competitive moisture meters against their corresponding vacuum-oven moisture values.

In these models, η is the expected vacuum-oven moisture; the βs are the coefficients to be estimated from the data; x is the moisture resulting from the meter; and ε is the error or difference between actual and predicted values.

The variation of each of the two data sets partitioned into the total, that due to the model and that due to the errors, is shown in Table II. Note that for the linear fit, model 1, the error or root mean square error for meter 2 is much larger than that for meter 1. However, if the second-order model 2 is fit to the meter 2 data, the error matches that for meter 1.

When the second-order model is fit to the data representing meter 2, the coefficient of the squared term shows statistical significance (Table III), meaning that its high value relative to its standard error did not happen by chance alone. There is a specific cause.

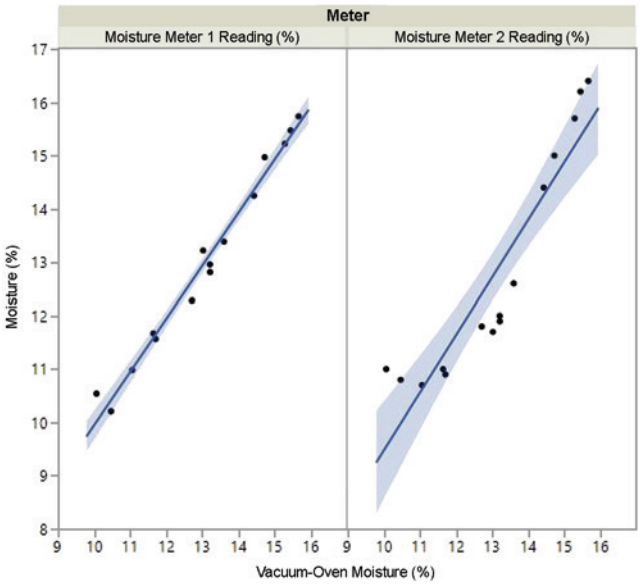


Fig. 2. Graphical summaries of linear regression model fits for two moisture meter readings against the corresponding vacuum-oven moisture readings. The line is the regression fit, and the shaded region surrounding the regression line represents the 95% confidence interval for the mean.

Table II. Regression analysis tables corresponding to first (linear) and second (quadratic) order models by meter[a]

Meter	Source	DF	Sum of Squares	Mean Square	Root Mean Square
Linear model					
1	Model	1	45.62	45.62	
	Error	13	0.85	0.07	0.26
	Total	14	46.47		
2	Model	1	53.69	53.69	
	Error	13	8.84	0.68	0.82
	Total	14	62.53		
Second-order model					
2	Model	2	61.91	30.96	
	Error	12	0.62	0.05	0.23
	Total	14	62.53		

[a] The root mean square error corresponding to the first-order model for meter 2 is much larger than that for meter 1. When the second-order model is fit to the data for meter 2, the root mean square error reduces to that for meter 1.

None of the above discussion is intended to rule out meter 2, so much as it is intended to show that the relationship between meter and truth can be approximated by regression analysis for calibration purposes. It may be that a linear relationship is more persuasive of veracity, but if a curved relationship, such as that of meter 2, can be confirmed, the corresponding equipment may be preferred, especially if it is associated with higher reliability or lower cost.

Precision

There is a hierarchy of variation inherent in all processes (Fig. 3). Total process and product variation consists of both the product variation and the variation associated with our ability to measure product attributes. Measurement variation, in turn, comprises both accuracy, as described above, and precision, which is comprised of repeatability and reproducibility.

Specially tailored studies, called Gage R&R studies, are designed to measure these latter two elements of measurement variation. Factors to be taken into account are analytical equipment differences, operator differences, and product sampling variation. In larger studies, differences among laboratories can also be taken into account. Part of the basic thinking focuses on equipment differences. If there are differences among devices, they can be resolved through adjustment or, if necessary, replacement. Another part of the thinking focuses on operator differences. They might be resolved through training to include benchmarking and reexamination of protocols.

The ultimate objective of a Gage R&R study is to quantify the variation due to reproducibility and repeatability. As described earlier, reproducibility is the variation experienced by multiple operators examining the same sample, perhaps with different instruments, whereas repeatability is the variation due to repeated measurement of the same sample.

Table III. Coefficient estimates of model 1 for meter 1 and model 2 for meter 2, with standard errors, *t* ratios, and probabilities[a]

Term	Estimate	SE	*t* Ratio	$P > \|t\|$
Linear model for meter 1				
Intercept	0.01	0.50	0.0	0.99
Linear term	0.99	0.04	26.4	<0.0001
Second-order model for meter 2				
Intercept	–2.94	0.46	–6.4	<0.0001
Linear term	1.14	0.03	33.9	<0.0001
Quadratic term	0.26	0.02	12.7	<0.0001

[a] The low probabilities confirm the significance of the model fits to the two data sets. The slope of the equation for meter 1, 0.99, is easily within the expected value of 1.0 given that its error estimate is 0.04. This suggests a likely one-to-one relationship between meter 1 and the vacuum oven.

Water activity data generated in a simple Gage R&R study are listed in Table IV. In this study, 10 production samples were homogenized and then divided among 3 operators whose duplicate samples were submitted blindly to them. For operators, it was business as usual. They knew neither the sample number nor the fact that they were analyzing replicates of the same sample for each of 10 samples as part of a designed study.

Typically, analysis of variance (ANOVA), together with its calculated variance components, is used for the analysis of Gage R&R studies. Although it is tempting to plunge into this analysis, it is very important to remember the first law of data analysis (described earlier) and plot the data. Plotting the data can help us avoid the wasted time involved in do-overs due to late discoveries of typos and outliers. We may see things in graphs, e.g., someone writes down 0.271 instead of 0.721, which we are not as likely to see in a table of numbers.

Replicate observations plotted for each of the three operators for each of the 10 samples are shown in Figure 4. Close inspection gives rise to the suspicion that operator 2 may have obtained lower water activities than the other two operators. To see this, it is necessary to spend some time carefully looking at the graph. After examining the graph, it may be concluded that, beyond the operator difference, there are no other obvious mes-

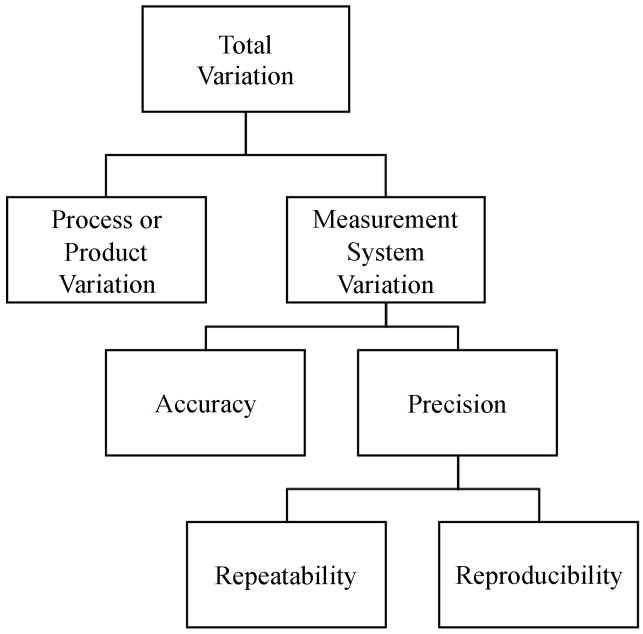


Fig. 3. The hierarchy of variation.

Table IV. Water activity data generated by a simple Gage repeatability and reproducibility (R&R) study

	Operator 1		Operator 2		Operator 3	
Sample	Replicate 1	Replicate 2	Replicate 1	Replicate 2	Replicate 1	Replicate 2
1	0.721	0.720	0.718	0.726	0.720	0.723
2	0.715	0.711	0.694	0.699	0.716	0.710
3	0.730	0.720	0.719	0.718	0.730	0.733
4	0.717	0.729	0.710	0.721	0.729	0.723
5	0.724	0.711	0.703	0.706	0.718	0.706
6	0.711	0.711	0.709	0.705	0.719	0.725
7	0.717	0.722	0.705	0.710	0.724	0.727
8	0.712	0.719	0.715	0.716	0.728	0.737
9	0.733	0.722	0.729	0.725	0.729	0.737
10	0.717	0.725	0.723	0.716	0.735	0.731

sages, and no typos or other quirks are evident. It is safe to proceed with the ANOVA.

The ANOVA partitions the total variation into separate, assignable sources of variation. It lists the amount of variation due to sample differences, operator differences, differences in operator results depending on operators (this is called an interaction), and the replicate variation that we take as random variation or random "error." Excellent statistical software packages are available to perform these calculations.

The source of variation, the corresponding degrees of freedom, sums of squares and mean squares, by line, are shown in Table V. Mean squares are sums of squares divided by degrees of freedom. The mean squares are divided by the error or "Reps(S×O)" (read replicates within the sample-by-operator interaction) mean square to create corresponding *F* ratios. *F* ratios have a known probability distribution that depends on numerator and denominator degrees of freedom, so it is possible to determine when an *F* ratio is larger than chance alone would allow. This study had no provision for repeated measures, but if it did, we would be able to determine from the ANOVA how much variation was due to repetitions as well.

The *F* ratios for operator and sample differences are so large that their probabilities are nearly zero. Differences among samples should come as no surprise. Samples taken from manufacturing processes over the long run will, in fact, show differences unless analytical method differences are insensitive. However, an opportunity for improvement is revealed by the low probability of the *F* ratio corresponding to operators. As illustrated in Figure 5, operators differ significantly from each other, and because they do, they contribute to the uncertainty of laboratory results. Laboratory health can be improved if differences among operators are removed.

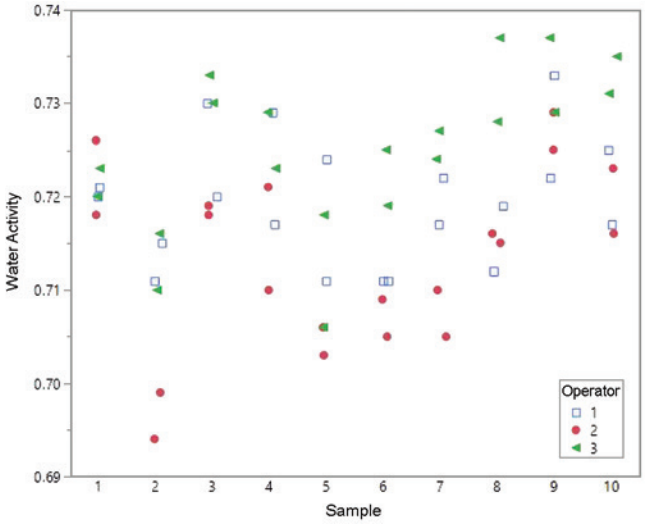


Fig. 4. Gage repeatability and reproducibility (R&R) data: water activity versus sample.

Table V. Analysis of variance (ANOVA) table for Gage repeatability and reproducibility (R&R) data

Source	DF	Sum of Squares	Mean Squares	*F* Ratio	*P* > *F*
Total	59	0.005241			
Operator	2	0.001358	0.000679	27.97	<0.0001
Sample	9	0.002465	0.000274	11.29	<0.0001
Sample × operator	18	0.000690	0.000038	1.58	0.1304
Reps(S×O)	30	0.000728	0.000024		

Maintaining Laboratory Health

Once established, accuracy and precision estimates will go adrift unless maintained. Routine monitoring and maintenance are essential. Two devices are useful for this purpose—one for within-laboratory control and the second for multilaboratory comparisons.

Within-Laboratory Control. The presence of a laboratory information management system (LIMS) and suitable statistical software greatly facilitate internal checks. A relatively simple device is a standard Shewhart chart for detecting errors and drift. Using it, the laboratory supervisor inserts blind replicate samples into the stream of routine samples to the analyst. It should be noted that the purpose in doing this is not so much to catch the analyst out as it is to aid in the detection of assignable causes of wayward results so they can be discovered and eliminated. Usually these samples are duplicates, but multiple blind replicates can be used if a procedure with greater power of detection is desired. Data from a stream of water activity data showing only those production samples that have blind duplicates are listed in Table VI.

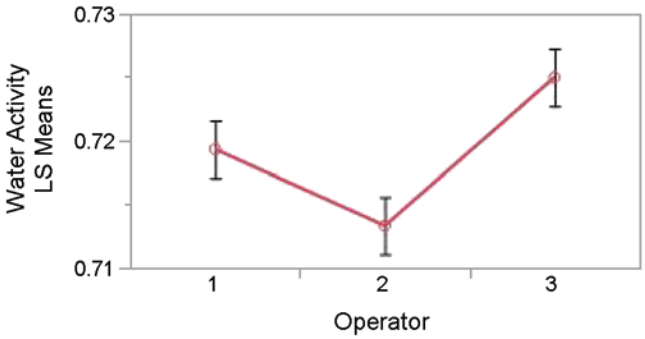


Fig. 5. Gage repeatability and reproducibility (R&R) example—operator differences are illustrated by software-drawn "least significant (LS) intervals," which are calculated such that if they fail to overlap differences among means may be declared significant. There is no overlap. All operators are different from each other.

Table VI. Blind replicate water activity determinations

Sample	Replicate 1	Replicate 2
1	0.731	0.741
2	0.714	0.706
3	0.727	0.737
4	0.716	0.718
5	0.719	0.716
6	0.704	0.72
7	0.713	0.723
8	0.723	0.711
9	0.721	0.716
10	0.714	0.726
11	0.706	0.711
12	0.727	0.725
13	0.718	0.731
14	0.711	0.703
15	0.723	0.721
16	0.712	0.705
17	0.727	0.728
18	0.724	0.782
19	0.728	0.744
20	0.718	0.703

The corresponding Shewhart control charts for the mean and range of the samples are shown in Figure 6. For purposes of analytical reproducibility, the mean chart might seem to be of little value. It simply shows the manufacturing variation, or so one might believe. Notice, however, that the range chart shows greater variation between duplicates at observation 18 than expected. At that same time point, the mean is out of control on the high side. Although the cause of the variation could be either analytical or production, there is greater wisdom in first checking the analysis before alerting the manufacturing department.

Multiple Laboratory Comparisons. Corporate and association laboratories should be in alignment with regard to reported analytical values. Often, interlaboratory test samples are circulated in round robin tests: several samples are created, subdivided, shuffled, and distributed to participating laboratories for analysis. Of course, it is best if blind duplicates are included. When all the data are gathered, they are subject to ANOVA modeling, partitioning sample, laboratory, interaction, and replicate variation so that laboratory differences may be identified. The data gathering and analysis process is similar to that described earlier for Gage R&R studies. The thinking is that once the outlying laboratories are identified, corrective action can be taken to bring their results into alignment with the other laboratories.

A useful graphical device for comparing laboratories when there are two replicates for each sample is the Youden plot (4). (Hare [4] describes the device in the context of a good statistical bedside manner; hence, the unusual title.) Jack Youden, a chemist and statistician at the National Institute for Standards and Technology (NIST), developed this plot in an effort to communicate the results of interlaboratory testing without the obfuscation many associate with statistical analyses.

The data in Table VII are used as an example. For each sample that appears in an interlaboratory test, Youden would plot the first replicate on the horizontal axis of a square graph and the corresponding second replicate on the vertical axis. The plotting symbol for each point is accompanied by a letter code designating the laboratory. A quick glance at Figure 7 shows that the two replicates for laboratory M are not in agreement. They are excluded from the calculation of the replication variation, which is then used to form a circle centered on the mean (or median) with a radius of 2.45 times the replication standard deviation. (Note, this is only an approximate radius. For information on how to obtain an exact radius visit the NIST web pages [www.nist.gov] on Youden plots.) The circle is intended to contain approximately 95% of laboratory means if laboratories do not differ.

Points outside the circle identify laboratories whose means differ from the mean (or median) of all results. Points that stray substantially from the diagonal line identify laboratories whose replicates do not agree with each other. An advantage of using a Youden plot is that the full diagnosis is communicated in one simple graph.

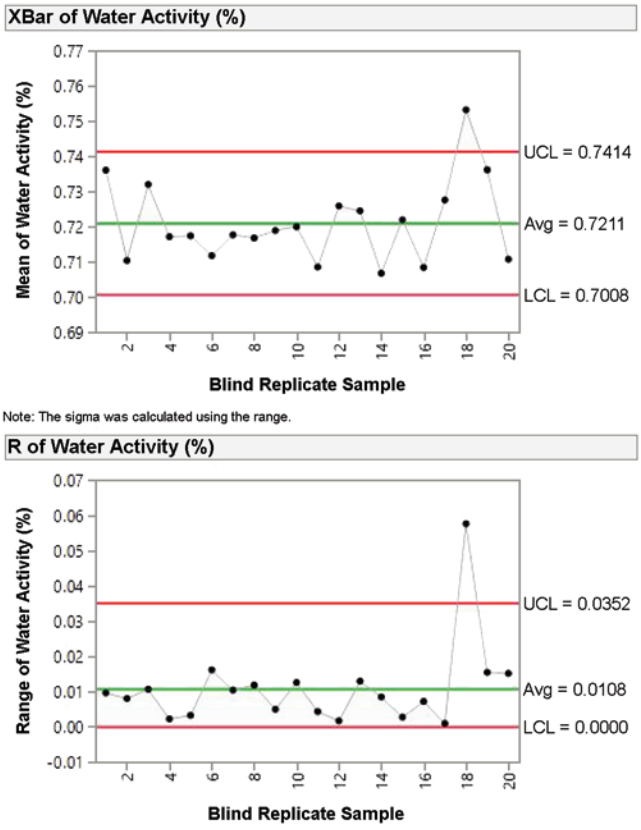


Fig. 6. Mean and range Shewhart control charts for blind analytical sampling assurance.

Table VII. Interlaboratory study data

Laboratory	Replicate 1	Replicate 2
A	57.9	70.9
B	57.3	71.1
C	67.9	77.9
D	84.0	78.9
E	84.0	55.3
F	58.1	57.6
G	61.9	55.3
H	39.2	33.9
I	56.4	58.8
J	45.3	51.6
K	63.0	62.7
L	60.9	75.6
M	35.4	112.7
N	81.7	77.6
O	58.9	60.5
P	51.9	55.8

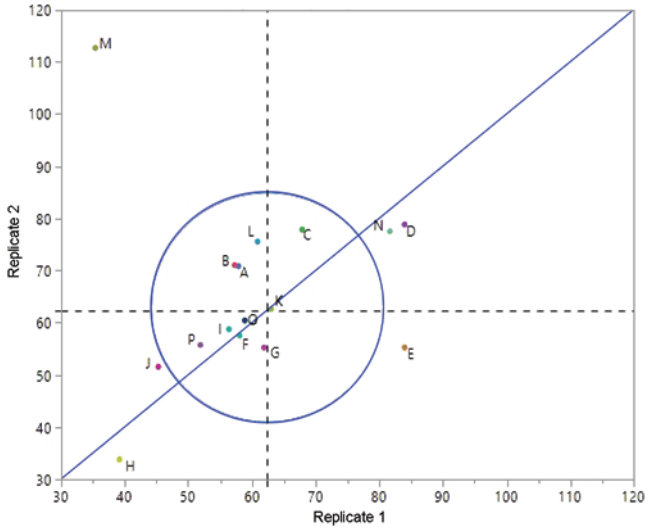


Fig. 7. Youden plot of vitamin A data: replicate 1 versus replicate 2.

Sample Size Determination

For most analytical measures, the determination of sample size depends on the error variation, the difference that might be considered important, and the risk of getting it wrong in both directions, i.e., the risk of declaring a difference when none exists and the risk of failing to detect a difference of a particular size. Most statistical software packages are able to calculate sample sizes given this input. First-time users are usually overly ambitious in the selection of tiny risks that drive sample sizes skyward, discrediting even remote use of statistical sampling. The result can be weakened opportunities for sound organizational decision making.

Many other means of sample size determination require special statistical attention. Two examples are provided.

Defect and Defective Item Detection. Seemingly inevitably, organizations will encounter the problem of suspect defects or defective items in otherwise normal production. The issue may arise as a result of customer complaints or chance encounters. For the laboratory, a first task may be to confirm their existence.

This might be considered a "needles in haystacks" problem, but for confirmation, a hypothetical proportion, p, must be assumed. The probability distribution of defects or defective items is assumed to be binomial and is given by

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where x is the number of defects or defective items in an individual sample; n is the number of units in the sample; and p is the hypothetical proportion of defects or defective items.

The expression $\binom{n}{x}$ is called the binomial coefficient and is equivalent to $\frac{n!}{x!(n-x)!}$. The "!" symbol is used to designate the product of the integers up to and including the letter preceding it. For example, $n! = 1 \cdot 2 \cdot 3 \cdots n$.

The probability of not finding a defect or defective unit in a sample of n units is

$$f(0) = \binom{n}{0} p^0 (1-p)^{n-0} = (1-p)^n$$

This means that the probability of detecting one or more defects or defective items in a sample of size n is $P = 1 - (1 - p)^n$.

If P is fixed at some specific value, such as 0.95 or 0.99, corresponding to 95% or 99% chance of detecting a defect or defective unit, then when the true proportion is p, the sample size, n, required is given by

$$n = \frac{\log(1-P)}{\log(1-p)}$$

For example, if it is speculated that the true defect or defective level is 1 in 1,000 or 0.001 and a 95% chance of detecting that level is desired, it will take

$$n = \frac{\log(1-0.95)}{\log(1-0.001)} = 2{,}994.2 \text{ or } 2{,}995 \text{ samples}$$

Accidental Inclusion. If, instead, the issue is one of accidental inclusion of an undesired substance, the probability distribution is often Poisson. The Poisson probability distribution function is

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2, \ldots$$

As an example of its application, consider International Commission on Microbiological Specifications for Foods sampling recommendations (7). For *Salmonella* sampling they recommend taking multiple 25 g samples to determine lot disposition. A standard is defined as 0 colony-forming units (CFU) in 100 g of product. In this case, the equation above reduces to

$$f(0) = e^{-\lambda}$$

The mean, λ, is expressed as the sample size, n, times the proportion (or probability) of a single CFU. If we seek detection of a single CFU, we might take 10 25 g samples for a total of 250 g, remembering that the standard is 0 CFU in 100 g of product. The probability function shows

$$f(0) = e^{-\frac{250(1)}{100}} = e^{-2.5} = 0.082$$

indicating an 8.2% chance of detection.

Concluding Comments

On first reading this article, and perhaps even beyond, the topics presented may seem daunting. Beginners and even those with some statistical experience should seek the assistance of a statistician who can collaborate with the laboratory team to tailor methods aimed at improving laboratory health. It is only by engaging in these practices that a laboratory can become a team fully engaged in continuous improvement and integrated with the vision and mission of the overall organization.

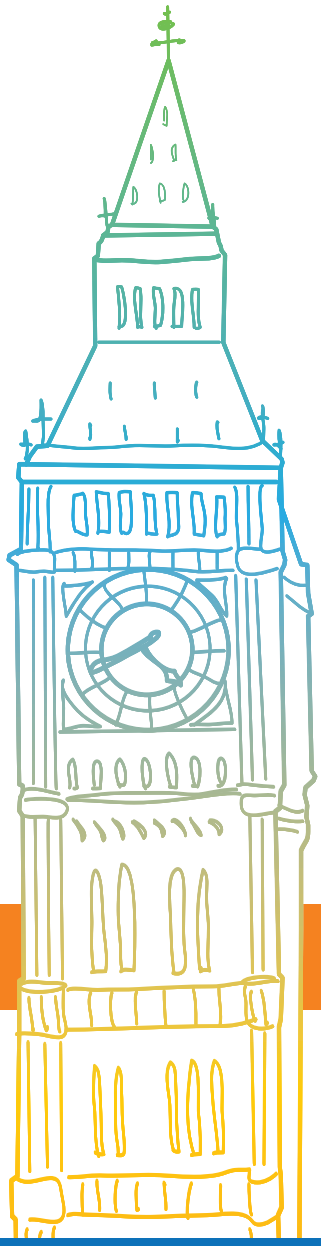
Author's Note

Graphs were produced using JMP 13 software (SAS Institute, Cary, NC).

References

1. Box, G. E. P. Signal-to-noise ratios, performance criteria and transformations. Technometrics 30:1, 1988.
2. Box, G. E. P., Hunter, J. S., and Hunter, W. G. *Statistics for Experimenters*. John Wiley and Sons, Hoboken, NJ, 2005.
3. Breyfogle, F. W. *Implementing Six Sigma*. John Wiley and Sons, Hoboken, NJ, 2003.
4. Hare, L. B. It's not always what you say, but how you say it. Qual. Prog., August, 2007.
5. Hare, L. B. Gage R&R reminders. Qual. Prog., January, 2012.
6. Hogg, R. V., Tanis, E. A., and Zimmerman, D. L. *Probability and Statistical Inference*, 9th ed. Pearson, Upper Saddle River, NJ, 2015.
7. International Commission on Microbiological Specifications for Foods. *Microorganisms in Foods 8: Use of Data for Assessing Process Control and Product Acceptance*. K. M. J. Swanson, ed. Springer, NY, 2011.
8. Meeker, W. Q., Hahn, G. J., and Escobar, L. A. *Statistical Intervals*, 2nd ed. John Wiley and Sons, Hoboken, NJ, 2017.
9. Montgomery, D. C. *Design and Analysis of Experiments*. John Wiley and Sons, Hoboken, NJ, 2012.
10. Montgomery, D. C. *Statistical Quality Control*. John Wiley and Sons, Hoboken, NJ, 2013.
11. Montgomery, D. C., Peck, E. A., and Vining, C. G. *Introduction to Linear Regression Analysis*, 5th ed. John Wiley and Sons, Hoboken, NJ, 2012.
12. Snee, R. D., and Hoerl, R. W. *Leading Six Sigma*. John Wiley and Sons, Hoboken, NJ, 2003.

CEREALS & GRAINS 18



The 2018 AACCI Annual Meeting is crossing the pond!

Join us in London to discuss the latest research and trends in cereal grain science.

SAVE THE DATE

October 21–23, 2018
Hilton London Metropole
London, United Kingdom

Get ALL the Latest Updates. Follow AACCI!

aaccnet.org/meet
#CerealsGrains18 #AACCI2018

AACC INTERNATIONAL