Keeping It Simple

Terry C. Nelsen[1]
T. Nelsen Consulting, Port Byron, IL, U.S.A.

Just as the modern electronic age has changed the way we study chemistry, it has also changed the way we perform statistical analysis. In the old "wet chemistry" days, we worked for hours to produce a few measurements. Today we use electronic instruments that spew out data. Statistical analysis was once limited by what we called calculation intensity, because it took hours or even days to perform all the calculations required. Computers are especially good at rapid calculations and can do the job in seconds. We have also been blessed with software packages that combine analytics with graphics and are generally user-friendly.

I probably sound like a grumpy old timer, but I believe some of the old tried-and-true statistical methods are still useful, especially when combined with modern rapid analysis and good graphics. I am not saying we should ignore newer methods. They are useful and necessary, especially when a chemist shows up with a huge data set and asks a statistician to conduct a "salvage job." However, problems can arise with interpretation of results. For example, do you know what the first principal component and a latent variable are? Do you know what assumptions you are making about your data? Are you concerned about multicollinearity or heteroscedasticity? Many of these modern multivariate methods are used for "dimension reduction," which is a statistician's way of saying simplification. If your multivariate analysis is successful in reducing the dimensions of your data set, then consider further experiments or trials on only those variables you have identified as important.

The Basics

Many problems can be avoided with careful planning. After 35+ years, I am convinced the two biggest problems we have in research and quality assurance are 1) sampling and 2) lack of a clear objective.

I am pleased that others in this issue of *Cereal Foods World* are addressing sampling. I will simply remind you that your results will pertain to the materials on which you performed your tests. In methods evaluation, we ask method developers to test their methods using the materials that are expected to be used in the method. Will a method developed to estimate a constituent in wheat flour provide good results for other types of flours? If you grab a sample of grain for a test and then carefully clean and sort the kernels before the tests are performed, can you assume the mill will produce the same results for grain that has not been cleaned and sorted as carefully? Sample preparation is a part of the sampling method. When considering numbers of samples, remember that analyzing five separate samples of a variety grown in five different locations is not the same as bulking the five samples into one bag and then grabbing five samples for analysis.

The lack of a clear objective is not always obvious. When I read an objective such as "determine which proteins affect quality," my first question is, "How exactly do you plan to measure quality?" My next question is, "Does this quality pertain to the grower, the shipper, the miller, the baker, the retailer, or the final consumer?" I have found communications are not always clear between management and the technical people doing the work or between two or more laboratories working on the same project. When you are told to test or compare varieties or optimize processes, be sure everyone agrees on exactly what is being measured before you start.

Initial Considerations

Sometimes you need to describe an entity. What is the range of its occurrence in specific materials? Does it occur alone or always in the company of another entity? Does its level of occurrence depend on the presence, absence, or level of the other entity? Statistics can be used to clarify descriptions. For example, technical people know that a mean and standard deviation tell us more about a measurement than a simple range of the data. A regression formula can be more useful than a simple correlation.

Using a graphics package to look at the data is helpful. If you plan to test hypotheses, then you will need to know how your data are distributed. If you use normal statistical tests, you will usually assume normal distributions. If you run the usual tests on non-normal data, however, it will be more difficult to find real differences. Many measurements are not normally distributed. For example, time intervals or microbial growth can be skewed. Censored distributions are common when measuring at lower concentrations, and some of the data are reported as less than LOD (limit of detection), BDL (below detectable limit), or ND (not detected). Multimodal distributions are common for multiple overlapping peaks in chromatography. Particle size distributions can be confusing if chemical or physical factors cause different types of particles.

Hypothesis testing includes a probability statement. Instead of saying "the difference was statistically significant," try saying "the difference was consistent." Decide how large a difference interests you—a difference can be statistically significant but practically trivial. Usually, the finer the difference you want to detect, the more data you will need to collect. Decide on your acceptable level of significance. Do you need to be 99% sure, or will you accept 95% or even 90%?

How much variation is acceptable? Will this treatment or additive change the final product and by how much? We usually compare means (averages) in hypothesis testing. Be sure to consider comparing variances as well. When comparing the qualities of plant varieties over several environments, they can have similar properties on average, but one of the varieties may be quite stable in several environments, whereas another may be more responsive, being either above or below average as environments differ. Box plots of percent protein content in four varieties of wheat (A, B, C, and D), each grown in several locations, are shown in Figure 1. Looking at this plot, I would describe varieties A and B as similar in average protein content, but variety A appears to be more susceptible to environmental

[1] E-mail: nelsent@gmail.com

https://doi.org/10.1094/ CFW-63-1-0029

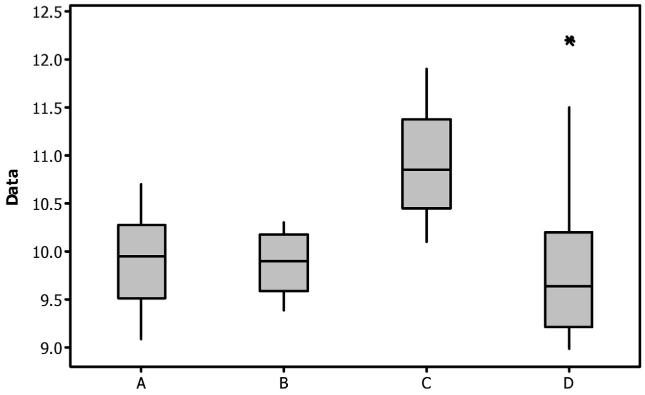


Fig. 1. Box plots.

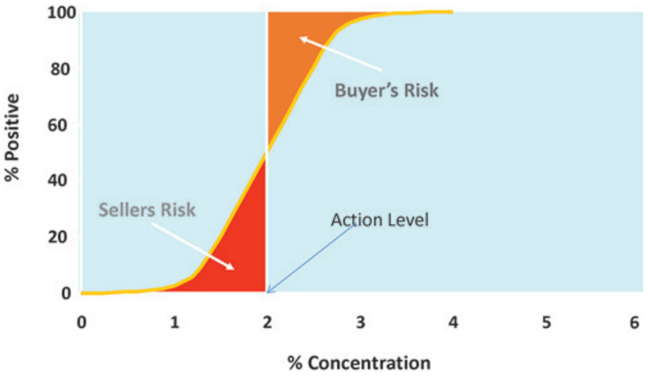


Fig. 2. Operating characteristic curve.

differences. Variety C has a higher average protein content and is similar in variance to variety A. Variety D has an overall lower protein content but appears to be very responsive to some environments. Your statistical software has box plot capability, and I highly recommend using it.

Identifying an Area of Interest

Determine your "area of interest." The area of interest can be those factors (e.g., proteins, additives, or treatments) that consistently affect your measurement. Or, it can be an adjustable range of a continuous variable, such as temperature, concentration, time, pressure, speed, and so on. In what range do you see the greatest effect? If you already know a lot about your subject matter, then concentrate on your expected area of interest. You can collect some data outside of your area of interest just to be sure you have not missed something important, but you can usually do these "checks" with fewer samples. If you are unsure about what you will find, then consider doing your investigation in stages.

A screening design can be used to efficiently determine which factors are important, and you can then concentrate on only those factors and avoid needlessly spending resources by collecting data on settings or combinations of factors that do not affect your measurements.

If you are looking for the effects of a continuous variable, then gather your initial data across a range of that variable. Your objective is to bracket the area of interest. Instead of low versus high or low/medium/high, plan to measure at points along a line. If you expect a straight-line response, then gather data at several points along that line. If you expect a curved response, such as the effects of temperature on some enzyme activity, then gather some data at the extremes but more data at the temperatures at which the peak or curve occurs. If you do not know what shape of curve to expect, then plan your project in stages: first gather some data across a wide range and then go back and gather more data where changes appear (i.e., the area of interest). If you already have a lot of data, then consider asking a statistician to use a dimension-reduction technique to find your area of interest.

An operating characteristic curve for a test kit detecting a particular toxin is shown in Figure 2. How could I have produced this curve most efficiently from the beginning if my objective was to evaluate the test kit at concentrations below 5%? My first trials could involve preparing concentrations at 1% intervals from 0 to 6% (bracketing the area of interest). Depending on the price of the kits, I could test three each at each percentage point. If we suppose the results show all three negative at 0 and 1%, all three positive at 4, 5, and 6%, and a mixture of positive and negative at 2 and 3%, that would narrow my area of interest to between 1 and 4%. Rather than continue to collect data below 1% or above 4%, I would concentrate my resources on gathering data in the area of interest. If the test kits were relatively inexpensive, I would double-check the results with a couple of tests outside the area, just for peace of mind. My advice is not to try to produce the curve all in one trial, because you could waste resources in gathering data outside the area of interest.

Also, consider what would happen if you conducted stage 1 of the trial and then got conflicting results in stage 2. Statisticians often recommend conducting a balanced, complete trial all at one time so that unknown factors do not pop up and bias the results. This is usually good advice for basic research, but your objective is to look for applied results. If an unknown factor causes your test kit or analytical method to produce different results at different times, then it is important for you to recognize this. If you can discover the unknown factor, you can make adjustments. If you cannot determine why you are obtaining different results under different circumstances, then maybe you should not trust the kit or method.

Refining the Design

If you are looking for effects of the presence or absence of several additives plus possible effects of time, temperature, or other variables all at once, then talk with a statistician about using a screening design such as a fractional factorial. Technical people complain that statisticians always want more data, but this is one area in which more information can be gathered with less data using efficient designs. Again, you can gather your data in stages: first a broad-based strategy to identify the area of interest and then a finer design to better describe responses. An efficient early-stage design can tell you which variables are important at which levels and in which combinations.

Chemists seem to love correlations, and statisticians love to tell them correlation does not imply causation. Think of correlations as hunting licenses. When you see a correlation between two measurements, then you get to hunt down and describe the specific relationship between those two measurements. Conversely, when you do not find a correlation where you expect to find one, then find out why not.

An example of an unexpected result (hidden variable) is shown in Figure 3. Data were gathered over three years to evaluate the effect of protein on loaf volume. In Figure 3 the graph on the left shows no relationship between protein and loaf volume—an unexpected result. The graph on the right is the same data looked at by year. It is obvious there is a relationship, but we also have to consider the effect of years. Another common cause of an unexpected nonrelationship is the result of restricted range. If I

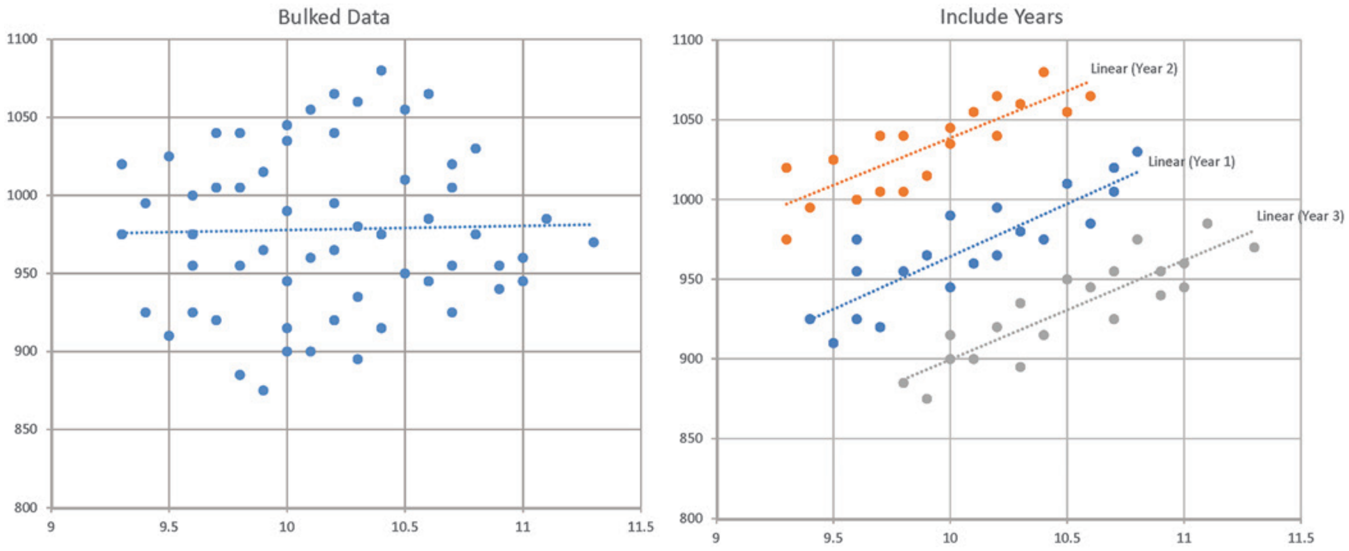


Fig. 3. Example of a hidden variable.

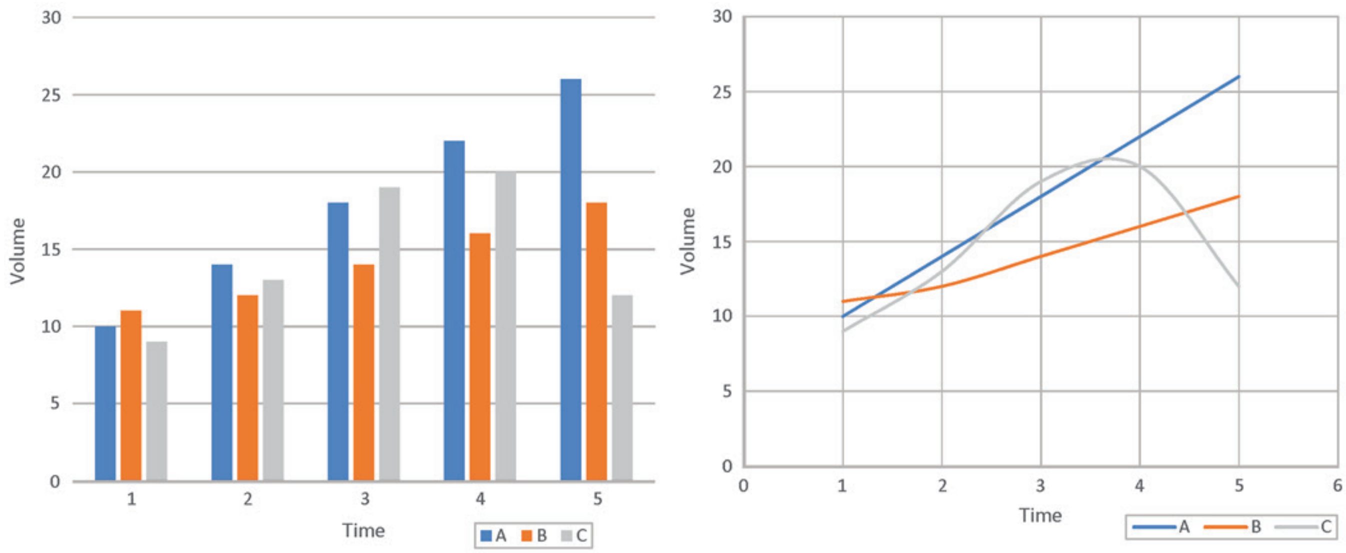


Fig. 4. Histogram versus line plot.

had looked at loaf volumes for a narrow range of proteins, I might not see the relationship.

Presenting Your Results

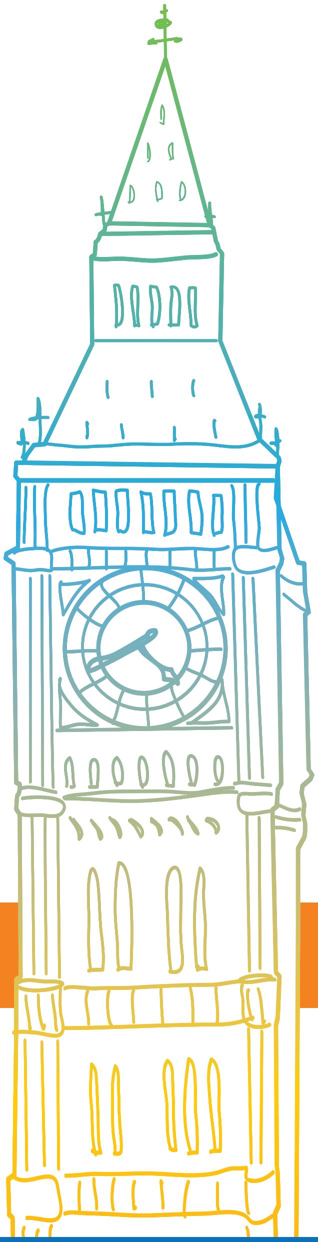
The best advice I can give someone who is planning to run an experiment or trial is first to consider how you want to present your results. I do not mean for you to anticipate your exact results but rather to consider if you want a table full of means with little letters attached to them, a series of curves, or something else. A statistician can then help you design your experiment to produce results in your preferred format.

The same results are presented in two different ways in Figure 4. The histogram is commonly used and easy to create (Fig. 4, left). You can even put little error bars on the top of the columns. A line plot of the same data is more informative, however (Fig. 4, right). Varieties A and B respond in a linear fashion to time. If their lines are not parallel, then you can assume there is an interaction between varieties A and B and time (i.e., they both respond positively to time but variety A more so than B). The response of variety C to time is more complicated, but it is different than the responses of the other two varieties. I always recommend using line plots and regression analysis when appropriate. Try to keep the values on the x-axis evenly spaced. The results will be easier to analyze and will look better on a graph. Use of a line graph would be inappropriate, for example, if the x-axis is not a continuous variable (e.g., five different locations or varieties). I sometimes cheat a bit in preliminary analysis and use a line graph just so I can spot interactions. Do not try to present your final results this way, however. If the values on the x-axis can be presented in any order, then a line graph is probably not appropriate.

Final Thoughts

In conclusion, I recommend you think through your experiment or trial before you start. Where, when, and how will you gather your samples? What will you do with them to prepare for analysis? What, exactly, do you plan to measure, and how would you like to present your results? Do not be afraid to run your trial in stages. Do not, however, use the term “pilot study.” For some reason, management seldom likes to fund pilot studies.

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