Statistical Analysis of Externally Calibrated Measurement Systems

Lynn Vanatta, Manager, Chromatography Research and Statistics, Air Liquide-Balazs™ Analytical Services, 13546 N. Central Expressway, Dallas TX 75243, 972-995-7541, lynn.vanatta@airliquide.com and **David Coleman**, Sr. Technical Specialist in Statistics, Alcoa Technical Center, Alcoa Center PA 15069

Abstract and Introduction

Measurement systems are at the heart of most all analytical laboratories, and generate data upon which chemical quality control and fab processes are based. To assess the quality of such results, various statistical procedures have been devised and implemented. However, the reliability and usefulness of these protocols depend not only on the care with which they are implemented, but also on the degree to which their underlying assumptions are understood and met.

This paper discusses several different statistical approaches, as they apply to qualifying data from externally calibrated measurement systems. (Such systems are ones whose raw data are not in the units desired by the customer. External standards of known concentrations must be analyzed and a regression-based calibration curve must be generated to transform the results into the desired units.) These statistical approaches are: 1) statistical process control, 2) gage repeatability and reproducibility studies, 3) traditional method detection limits and quantitation limits and 4) designed calibration studies. For each protocol, the underlying assumptions, strengths and weaknesses will be presented, and (where appropriate) example calculations will be discussed.

Statistical Process Control (SPC)

SPC was developed to be used primarily in a manufacturing setting. However, the technique also is used to monitor and evaluate analytical instruments. The goal is to be certain that the instrument and measurement process are in control. For this use of SPC, a check standard is analyzed routinely (ideally, these analyses are made at defined time intervals). The calculated concentrations are plotted on a graph known as an individuals (X) chart. A second plot, known as a moving-range (mR) chart is also constructed. (The difference between the first two successive check-standard results is calculated for the first point. For the second point, a new range is calculated by dropping the first point and adding the third. This pattern is followed with each new analysis.) Upper and lower control limits are calculated for each chart. Rules are established to decide when excursions outside these limits constitute an out-of-control situation.

A more robust version of this technique involves the analyses of a pair of check standards. One standard is at the low end of the working range and the other solution is at the high end. A pair of charts is generated for each standard. A pair is also constructed for the difference between the two results, a procedure that may reveal trends that are hard to spot in either of the other two pairs of charts.

The primary assumptions behind this SPC technique are that the concentration data are independent (i.e., not drifting) and exhibit a Gaussian distribution. The advantage SPC provides is that the charts provide a real-time, dynamic tool for spotting shifts and trends in the instrument's performance. However, the technique does not capture the uncertainty or possible bias (i.e., true concentration minus reported concentration) associated with the calibration process. Additionally, rates of false negatives and false positives are not incorporated into SPC. Finally, unless approximately 20 averages have been calculated, the control charts may simply be tracking noise. Thus, this technique is not a realistic alternative for evaluating a newly calibrated (or recalibrated) method.

Gage Repeatability & Reproducibility (GR&R) Studies

GR&R studies determine how much measurement-system variability can be ascribed to the operator (one definition of reproducibility) and how much can be ascribed to the measurement itself (repeatability). Typically, variability charts are used in the data-analysis process.

Several assumptions are associated with GR&R. First, the data are continuous, have not been rounded, and contain no outliers. Second, the only contributors to variation (other than different sample values) are the analyst/operator and the instrument/measurement system. Third, the data are in post-calibration units. Fourth, there is no interaction between any two factors (i.e., no factor influences any other). Fifth, measurement variation (i.e., standard deviation) is constant across any given factor (e.g., all analysts are equally consistent). Sixth, the chosen analysts and measurements are representative of a population.

It is not uncommon for at least some of the above assumptions to be false. The most frequent (and most damaging) violations occur with the first, second and fifth stipulations; it is actually highly likely that the standard deviation will change across at least one factor. Also, a factor that often contributes significantly to variation is the instrument-calibration procedure; yet GR&R does not address this source.

Further problems arise with GR&R because it is not strictly applicable to all measurement-analysis scenarios. For example, a single-instrument, single-operator method does not fall cleanly into this type of study. In these situations, a factor (e.g., the day) other than the operator must be chosen to assess reproducibility. Because the above assumptions and applicability requirements often are not met, GR&R is not an appropriate tool for evaluating an externally calibrated measurement system.

Traditional Detection Limits (DLs) and Quantitation Limits (QLs)

An almost universally utilized approach to evaluate methods is the calculation of a method detection limit (if the sensitivity of the procedure will be challenged) and a quantitation limit. The assumptions are that: 1) a blank or (low-level standard) is available, 2) results (converted to concentration units) from seven or eight replicate analyses are available and 3) the false-positive rate (FP; i.e., false detections when measuring blanks) is set at 1%. Once the data are obtained, the standard deviation of the responses is calculated and multiplied by Student's *t* (2.998 for seven degrees of freedom and FP = 1%). The result (known as 3σ) is the detection limit. (If a low-level standard was used instead of a blank, the calculated DL must be greater than [(1/5) *

the concentration] to be valid.) The quantitation level is set at a higher number of standard deviations, typically 10σ. An example of DL calculations is shown in Figure 1.

The advantages of this technique are: 1) the data are easy to generate (only seven or eight replicate analyses are required), 2) the calculations are simple and 3) false positives are controlled tightly (i.e., FP = 1%). However, four statistically important issues are ignored: 1) the false-negative rate (FN), 2) calibration uncertainty, 3) bias and 4) any change that may occur in response standard deviation as concentration changes.

Failure to address FNs has serious consequences. Assume that the 3σ DL coincides with the threshold (i.e., the concentration below which no response can be detected at all). Assume also that a sample with true concentration equal to the 3σ DL is analyzed in replicate. The Normal distribution of calculated concentrations will be centered at the threshold. Thus, half of the analyses will have responses that cannot be detected because they fall below the threshold. In other words, FN = 50%!

If the standard deviation of the response changes with concentration, then Ordinary Least Squares (OLS, the "default" fitting technique used with calibration models) is inappropriate; OLS assumes that this standard deviation is constant throughout the concentration range. In such non-constant cases, Weighted Least Squares (WLS) should be used. With this technique, weights are applied to the responses and are the reciprocal of the square of the standard deviation. The result is that the noisy data are not allowed to influence the curve as much as are the more well-behaved numbers. If OLS is used inappropriately, the estimates of calibration coefficients are noisier, the curve's prediction interval is incorrect (see below for more details on this latter topic), and the estimated detection limit will be incorrect. In the end, traditional DLs and QLs typically are not sound statistically and should be viewed as such.

Designed Calibration Studies

The preferred technique to evaluate methods for externally calibrated instruments is a carefully designed calibration study. Not only do the resulting data allow for proper assessment of the method, but also they provide a realistic quantification of the uncertainty inherent in any reported sample result.

The assumptions behind this approach are that: 1) reliable, known values are available for the calibration standards, 2) the instrument responses (e.g., peak area or absorption units) exhibit a Gaussian distribution and 3) the calibration study has been designed properly so as to include an adequate number and spacing of concentrations, plus an adequate number of replicates.

These three stipulations typically can be met. The third directive requires a little time and careful thought, but the process becomes almost intuitive after several studies have been designed and performed. A carefully constructed calibration study will allow the analyst to: 1) model the standard deviation of the response (i.e., decide if response variation changes with concentration), 2) obtain a low detection limit (if sensitivity is an issue), 3) detect curvature in the data, 4) obtain high precision in any critical concentration area, 5) construct a curve that will cover the concentration range expected for typical samples, and (possibly most importantly), 6)

construct the prediction interval (for any sample that is analyzed via this curve, this interval reveals the overall uncertainty associated with the result).

If recovery is an issue, a spiking study should be conducted after the calibration work is completed. Concentrations of the spikes are predicted via the calibration curve. These predictions then can be plotted vs. true concentration, and the associated prediction interval constructed. Once unknown samples have been analyzed via the calibration curve, this recovery plot can be used to correct the values for any recovery problems. The intercept of the line represents any arithmetic offset; the slope of the line represents the proportional recovery. For the chosen confidence level, the recovery curve's prediction interval gives the overall uncertainty associated with the entire method. An example plot is given in Figure 2.

Summary

Although there is no perfect technique for the statistical analysis of externally calibrated measurement systems, many of the typically used procedures are not adequate for this type of evaluation. In general, designed calibration studies are preferred (over, e.g., SPC, GR&R studies and traditional DL/QL calculations), since they can account for more of the variability that is typically encountered in a method.

Bios

Lynn Vanatta is the Manager of Chromatography Research and Statistics at Air Liquide - Balazs™ Analytical Services in Dallas, Texas. She is an analytical chemist who specializes in ion chromatography. She routinely develops statistically sound methods for trace anions in semiconductor-grade water and chemicals. In 1995, Lynn was a participant in "Statistically Valid Detection Limits & Quantitation Limits," a short course taught by David Coleman. They now co-instruct the short course; in addition, Lynn teaches her own related class at the International Ion Chromatography Symposium (IICS). They have published ten papers and are working on an eleventh. In 2002, they began writing a bi-monthly statistics column for *American Laboratory*.

Lynn is active in the ASTM Committee D19 on Water. She is chairman of the task groups on Precision and Bias and on Electronic-grade Water, and also is the D19 Definitions Advisor. She is on the Scientific Committee of the IICS, and has organized and chaired a conference session for the semiconductor industry every year since 1997; she was Program Chairman of the entire Symposium in 2001 and 2002, and was co-chairman in 2003.

David Coleman is a Sr. Technical Specialist in Statistics in the Applied Stat Group within the Mfg. Systems and Technology platform at Alcoa Labs. David regularly designs and analyzes experiments (especially inter-lab studies, calibration studies, and measurement-capability experiments to quantify the performance of measurement systems). He has authored and co-authored papers in *Journal of Chromatography A, Environmental Science & Technology, Chemometrics and Intelligent Lab Systems*, and *Technometrics*. David and Prof. Robert Gibbons co-authored, Statistical Methods for Detection and Quantification of Environmental Contamination (2001, Wiley). He is active in the American Statistical Association, and in the ASTM D19 task group on Detection & Quantitation.

References

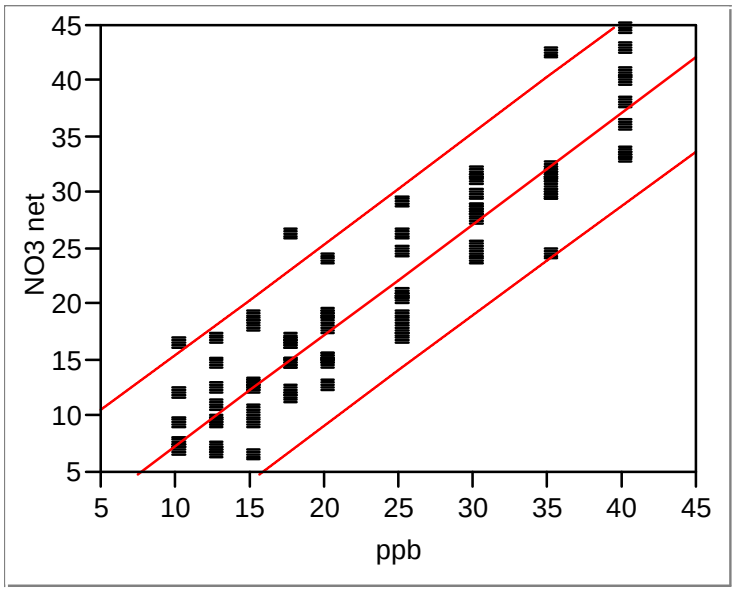
SPC for the Rest of Us, Hy Pitt, Addison-Wesley, Reading MA, 1994.

JMP 4.0 Statistics and Graphics Guide, SAS Institute, 2000, Chapter 26 (Variability Charts: Variability Chart and Gage R&R Analysis).

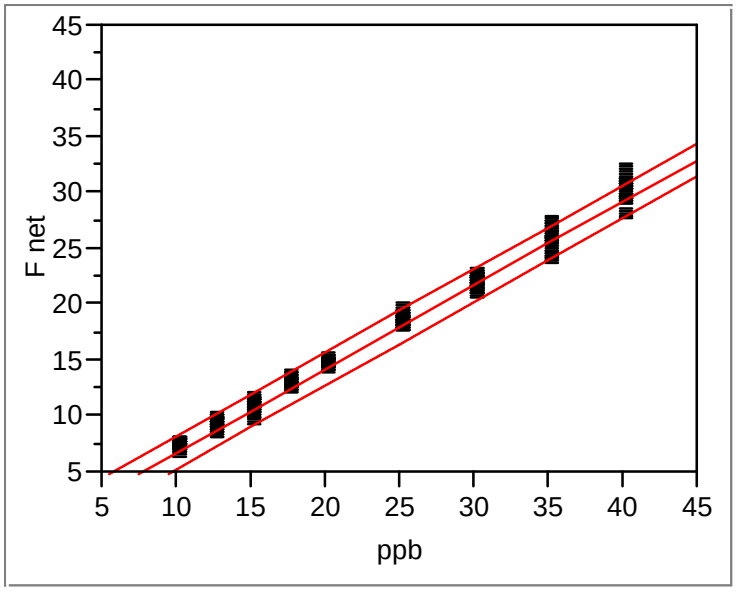
Figure 1. MDL calculations for nitrate in 30% hydrogen peroxide, using each of the three lowest spike concentrations. All MDL estimates are greater than 1/5 of the respective spike level. Note the dependence of the MDL on the spike concentration.

Spike conc. (ppb)	MDL - Nitrate	Spike ppb / 5
10.1	15.4	2.0
12.5	11.2	2.5
15.0	13.2	3.0

Figure 2. Recovery plots for nitrate and fluoride in 30% hydrogen peroxide. The bias portion of the recovery is given by the y-intercept. The proportional recovery is given by the slope. The uncertainty in the recovery estimates is given by the prediction interval (the dotted lines enveloping the recovery curve). Note that high slope does not guarantee high precision, and vice versa.



NO3 net = -2.473028 + 0.9914632 ppb



F net = -0.729979 + 0.7487938 ppb