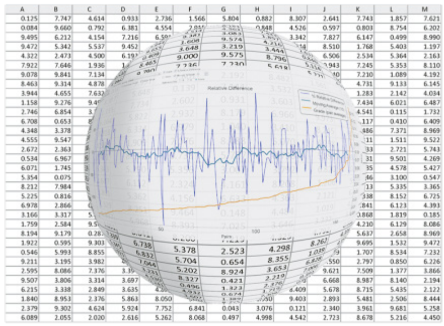
Agreement analysis—testing the boundaries between producers and consumers

Maxime Pitard
HonuaTek LLC



Introduction

The most serious mistakes are not being made as a result of wrong answers. The truly dangerous thing is asking the wrong questions. These words of wisdom from the late Peter Drucker underscore the importance of being well informed when making business decisions. This is especially true of producer–consumer boundaries in industrial processes, where the quality of a lot, or series of lots, can have significant financial consequences.

In order to mitigate any disagreement at these boundaries, concerned parties typically establish common protocols to monitor the quality of the lots produced. These monitoring protocols—often implemented with agents from a combination of consumers, producers and/or commonly agreed-to third parties—consist of splitting samples taken from the lots of interest and processing them separately to generate two, or more, sets of estimates which are then compared for differences (Figure 1).

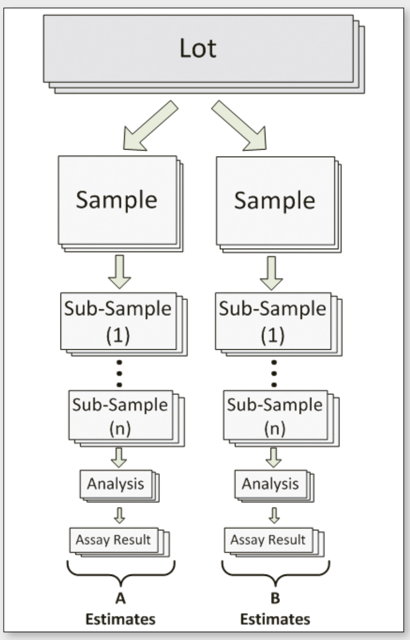


Figure 1. Lot estimation.

These monitoring protocols typically define:

- the party(ies) responsible for taking the samples;
- the party(ies) responsible for processing and analysing the samples;
- threshold(s) for disagreement between the estimates;
- the steps to take when estimates disagree.

This last point is of particular interest, as many organisations will make use of an umpire sample (also known as a referee or reserve sample) as the arbiter of any disagreement found in the estimates. This sample is often kept to the side, and used on an as-needed basis to resolve contractual disputes as per the protocol that defines its use. Regrettably, if the goal is to infer a deeper understanding of the source of disagreement and take corrective action on the offending process, then using the umpire sample for this purpose may very well provide the right answer to the wrong question.

Pierre Gy, in his ground-breaking book on the Theory of Sampling (TOS), dedicated two chapters to the problem of producer–consumer boundaries, where he presented an approach for testing for agreement between two series of independent estimates of a sample characteristic.[1] In these chapters he cautioned his readers on how typical bias testing is more focused on controlling risk from a seller's point of view, at an unknown and potentially significant risk to the buyers. He provided guidance in addressing this issue, and developed a simple and systematic approach to help monitor for the presence of statistically significant bias in a process: what we call the Gy bias test, or continuous bias test[2] (for its control chart-like properties).

FPSC Sampling Consultants in collaboration with HonuaTek saw the value in Pierre Gy's approach and built a software package to automate its use. As the software was developed and used with internal clients, it became clear that Gy's approach, although powerful, did not provide a complete view of the processes that generated the estimates; it was not enough to visualise the progression of bias over time, we needed to better understand the possible sources of this bias as well. Building upon the work of Pierre Gy and others,[3] the software was extended with additional capabilities to provide a more complete set of tools.

The resulting "Agreement Analysis" software incorporates a complementary set of techniques, originating from the fields of statistical process control and sampling statistics. It enables users to study the same data through different lenses, providing as complete a view as possible, while keeping the complexity down to a minimum.

Agreement Analysis currently supports five techniques (Figure 2):

- the Gy Bias Test to identify statistically significant bias between the estimates;
- the Scatter Plot to examine correlation between the estimates;
- the Relative Difference plot to investigate differences between estimates and their potential source;

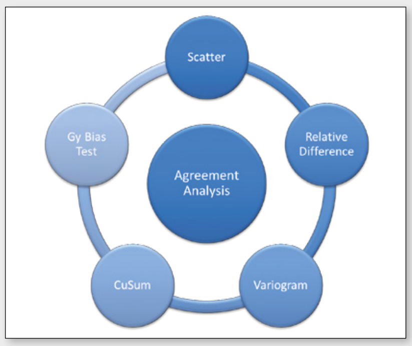


Figure 2. Five techniques.

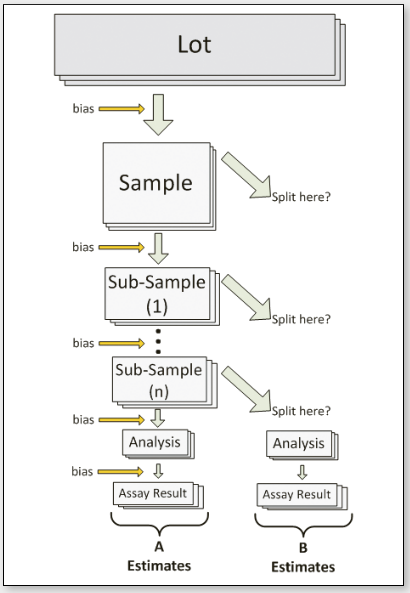


Figure 3. Source of bias.

- the Cumulative Sum plot to compare short-term trends in the estimates;
- the Variogram to compare the variability of the estimates over time.

Should the estimates be found to disagree, results from the different techniques can be combined to provide an overall synthesis of the possible sources of the disagreement.

By seamlessly integrating these five techniques, the Agreement Analysis software provides a collection of tools to analyse differences between sets of estimates, helping those responsible for quality control focus on answering the following key questions:

- Is the level of risk taken by both producer and consumer acceptable?
- Are the differences between the sets of estimates tolerable?
- If the estimates disagree, which set is most likely incorrect and why?

Background

TOS shows us that bias generating errors are present at every step where samples are handled in a lot estimation protocol—the introduction of bias cannot be avoided. The best that can be done is to minimise the amount of bias introduced at every step by following appropriate sampling practices.

There are a several stages in a lot estimation protocol where a sample can be split to generate the two sets of estimates needed for analysis (Figure 3). It is important to understand that any bias introduced prior to the sample split will be present in both sets of estimates and will be difficult to differentiate as coming from the protocol or the lot itself. To this end, the sooner the sample is split in the overall protocol, the more information is available to identify the sources of bias. This of course has to be weighed against the added cost of carrying out the full protocol on each sample.

Once the estimates are available, getting them into the Agreement Analysis software and applying the five techniques described in the following sections is a straight forward process.

Scatter

The scatter plot is a good starting point for analysis since this technique presents a simple visual indicator of how well one set of estimates can predict the values of the second set of estimates (Figure 4).[4]

This approach tests one of the main assumptions for agreement analysis; that the two sets of estimates are good linear predictors of each other, given they are different estimates of the same initial samples or sub-samples.

The plot consists of pairing up the estimates and plotting them on the graph, with the expectation that these will form a narrow dispersion ellipse along the identity line $x = y$. The plot can be inspected for a linear relationship, and the transverse of the dispersion ellipse estimated by a best fit line calculated using least-squares linear regression.

How well the estimates agree can be in-part measured by the correlation coefficient. Any random error introduced in the estimation protocol has the effect of widening the ellipse. Any systematic error in the estimates, depending on its nature, can shift the transverse of the ellipse away from the identity line as well as change its slope.

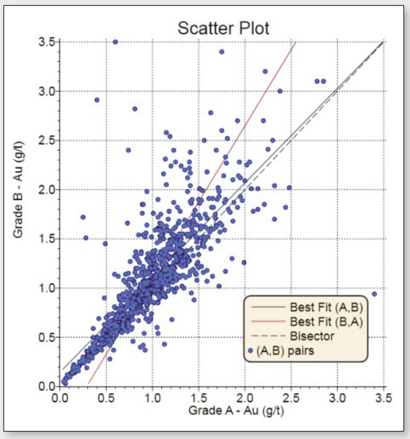


Figure 4. Scatter plot.

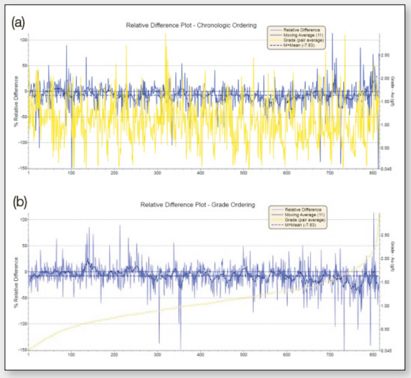


Figure 5. (a) Pairs ordered chronologically. (b) Pairs ordered by their average values from low to high.

Relative difference

The second technique used in the analysis is the relative difference plot, which provides a convenient control chart focused on detecting bias between the sets of estimates. This technique relies on several key plots: the average of the estimates, the per cent relative difference for each pair of estimates, the moving average of the per cent relative difference. This technique supports ordering the pairs of estimates in one of two ways:

- Pairs ordered chronologically; useful in detecting problems such as slow drift in one of the estimated sets (Figure 5a).
- Pairs ordered by their average values from low to high; providing information on how the differences fluctuate with the change in the estimates (Figure 5b).

Variogram

Adapting an approach introduced by Pentti Minkkinen,[3] the variographic technique is utilised to compare the process variability between the sets of estimates.

$$V_j = \frac{1}{2(N-j)}\sum_{i=1}^{N-j}\left[h_{i+j} - h_i\right]^2, \; j = 1,2,\ldots,\frac{N}{2}$$

The absolute variogram (V) of each set of estimates is calculated for the data increments (h) and lag (j), which are then compared for differences.

Items of particular interest for comparison from the variographic study include the short range, long range and cyclic terms,[5] (Figures 6a and 6b) which can also be used as input to other techniques such as the cumulative sum plot and Gy's bias test.

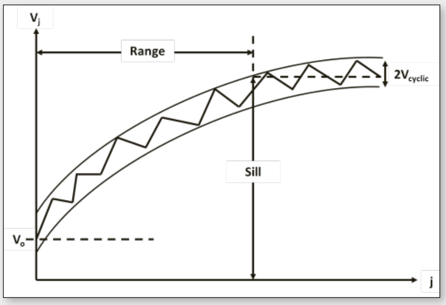


Figure 6a. Variogram.

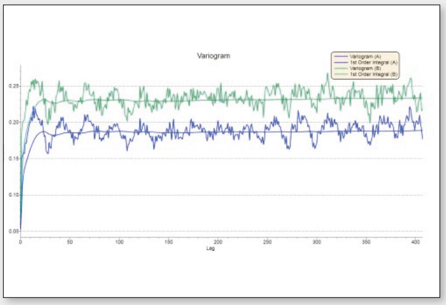


Figure 6b. Comparing variograms.

Where D_i is the estimate of the systematic differences, s_i the estimate of the standard deviation of the systematic differences and D_a the tolerated systematic difference.

Normalising the results of the tests over the t-distribution, we are able to construct an easy to interpret control chart with statistically significant conclusions between the random variable values of –1 and +1.

The control chart allows for a rich set of interpretations described in detail in Table 1, where W and W' are the results of the preliminary and secondary tests respectively.

Cumulative sum

The cumulative sum plot is a tool commonly used in statistical process control to detect small changes in mean level.[6]

For agreement analysis, a variant of this plot, the Tabular Cusum,[7] has been adapted to create control charts to compare the effect of the accumulation of small, persistent, non-random changes between the sets of estimates (Figure 7). This technique requires identifying both a target value (μ_0) separating the plot into upper and lower portions; and a slack value (K) to filter out small random fluctuations. The upper portion of the cusum (C^+) shows the cumulative effect of positive deviations from the target value with a minimum possible value of zero, and is defined by the recursive function:

$$C_0^+ = 0$$

$$C_i^+ = max[0, C_{i-1}^+ + x_i - (\mu_0 + K)]$$

The lower portion of the cusum (C^-) shows the cumulative effect of negative deviations from the target value with a maximum possible value of zero, and is defined by the recursive function:

$$C_0^- = 0$$

$$C_i^- = max[0, C_{i-1}^- - x_i + (\mu_0 - K)]$$

Typically, the construction of the cumulative sum plot requires a statistical model derived from historical data to establish values for parameters such as K. This is of lesser concern when applied to agreement analysis since the focus is not on identifying out of control situations, but instead on comparing the two sets of estimates.

Pierre Gy's Bias Test

This adaptation of classic hypothesis testing is used as the final technique of agreement analysis, and helps establish the statistical significance of observed differences.

Based on Pierre Gy's approach, this technique helps users visualise the evolution of bias over time when comparing two sets of estimates (Figure 8). The test involves taking each pair of estimates, and given a certain risk α, defining a random variable for a preliminary test W (testing for no difference in the pairs),

$$W_i = \frac{D_i\sqrt{i}}{s_i}, i = 1, 2, \ldots, N$$

and one or more random variables for the complementary tests W′ (testing for a tolerated systematic difference in the pairs).

$$W_i' = \frac{\left(|D_i| - D_a\right)\sqrt{i}}{s_i}, i = 1, 2, \ldots, N$$

Software features

Designed for ease of use, the Agreement Analysis software features a rich set of capabilities, including the ability to create an unlimited number of views into the data. Each view can be constructed from one of the five supported techniques and then uniquely configured for a user's specific requirements.

Comparative analyses can be performed by seamlessly transitioning amongst the different views, which can in turn be stored, along with the data, to project files for future retrieval.

With the ability to import data from various sources such as Microsoft Excel, the software provides complete customisation (styles, sizes and colours) of the displayed graphics and fonts (Figure 9).

Users have the option to save the graphics from the views to JPEG format, and reports containing all the views, along with their high resolution graphics and configuration parameters, can be generated and exported to Microsoft Word. These reports are stand-alone documents which can be fully edited by the users.

FPSC and HonuaTek will continue to collaborate and improve upon the capabilities of the Agreement Analysis software, driven

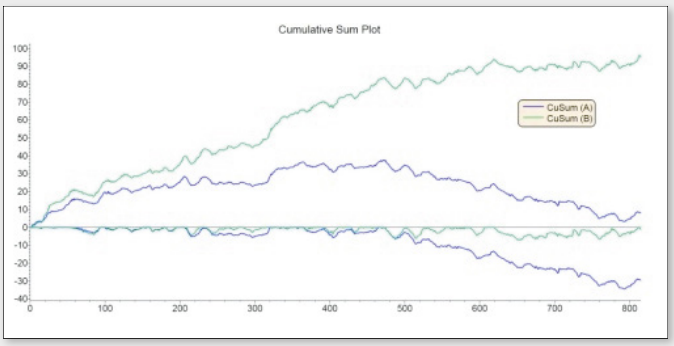


Figure 7. Cumulative sum plot.

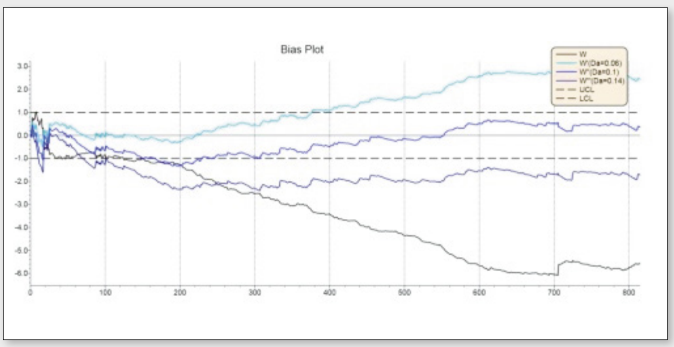


Figure 8. Bias plot.

Table 1.

W	W'	Interpretation
$>+1$	$>+1$	With the given risk we can conclude that set A is systematically higher than set B, and beyond the tolerated systematic difference.
$>+1$	$[-1,+1]$	With the given risk we can conclude that set A is systematically higher than set B—but we cannot determine if it is beyond the tolerated systematic difference.
$>+1$	<-1	With the given risk we can conclude that set A is systematically higher than set B, but within the tolerated systematic difference.
<-1	$>+1$	With the given risk we can conclude that set A is systematically lower than set B, and beyond the tolerated systematic difference.
<-1	$[-1,+1]$	With the given risk we can conclude that set A is systematically lower than set B—but we cannot determine if it is beyond the tolerated systematic difference.
<-1	<-1	With the given risk we can conclude that set A is systematically lower than set B, but within the tolerated systematic difference.
$[-1,+1]$	$[-1,+1]$	With the given risk, we do not have enough samples to draw a conclusion.
$[-1,+1]$	<-1	With the given risk we cannot conclude the presence of a systematic difference between set A and set B—if it does exist, it is less than the tolerated systematic difference.
$[-1,+1]$	$>+1$	The results are erroneous and no conclusion can be drawn from them.

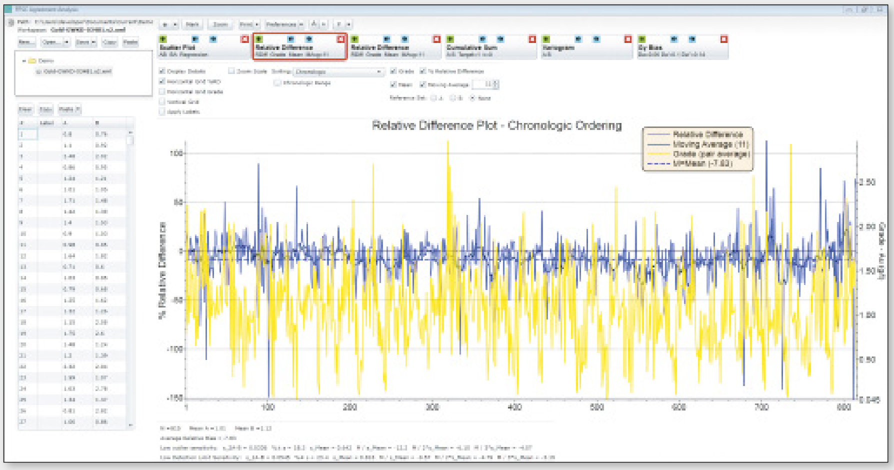


Figure 9. Software.

by experience and continuous feedback from our growing user community—helping provide the right answers to the right questions.

References

1. P.M. Gy, *Sampling of Particulate Materials—Theory and Practice, Developments in Geomathematics 4,* 2nd Edn. Elsevier Scientific Publishing Company, Amsterdam, pp. 379–403 (1982).
2. M. Pitard and R.C.A. Minnitt, "Use of the continuous bias test to monitor comparisons of independent measurements", *Proceedings of the 6th World Conference on Sampling and Blending,* Lima, Peru, pp. 409–417 (2013).
3. P. Minkkinen, "Comparison of independent process analytical measurements–a variographic study", *Proceedings of the 5th World Conference on Sampling and Blending,* Santiago, Chile, pp. 151–160 (2011).
4. D.C. Montgomery, *Introduction to Statistical Quality Control,* 6th Edn. John Wiley & Sons, New York, pp. 204–205 (2009).
5. F.F. Pitard, *Pierre Gy's Sampling Theory and Sampling Practice, Heterogeneity, Sampling Correctness, and Statistical Process Control*, 2nd Edn. CRC Press LLC, Boca Raton, pp. 89–121 (1993).
6. G.E.P. Box, J.S. Hunter and W.G. Hunter, *Statistics for Experimenters Design, Innovation and Discovery,* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, pp. 562–563 (1978).
7. D.M. Hawkins and D.H. Olwell, *Cumulative Sum Charts and Charting for Quality Improvement.* Springer-Verlag, New York, pp. 24–30 (1998).