Using Check Sample Results

TERRY C. NEELSEN

AACC International’s Check Sample and Proficiency Testing Service has offered a check sample service since 1948. Laboratories use the service as a means to monitor their equipment, personnel, and analytical methods and to offer an independent assessment of quality control to their internal and external customers. Check samples are specially prepared test samples that are sent to subscribing laboratories at specified time intervals. Some are monthly, some bimonthly, and some are quarterly. The Analysis Series can be chemical constituents (i.e., moisture, ash, protein, fats and fatty acids, fiber, sugars, minerals, and vitamins), physical properties (alveograph, farinograph, amylograph, mixograph, and RVA), and food safety (microbiological and sanitation).

The subscribers analyze the samples and return the results to AACC Intl. For each series, the results of all the participating laboratories are combined and analyzed according to standard statistical principles. Results are returned to the subscribers and the results of the entire analysis are released in coded form so that each participant can compare its own results with all of the other laboratories.

Statistical evaluation is based on a z-value in order to standardize and allow comparisons to be made. If the actual data were used, a different average would be obtained for each month for a given measurement. Also, some measurements are more difficult to do and have more natural variation (i.e., noise) than others.

The z-value for a laboratory is the number of standard deviations by which that laboratory’s result differs from the overall average value of all the labs (an outlier detection and deletion procedure is run first). A normal distribution is assumed, which means that 67% of the labs should be within one standard deviation of the mean (i.e., a z-score between −1 and +1) and 95% of the labs should be within two (actually 1.96) standard deviations of the mean (a z-score between −2 and +2). The actual z-scores are adjusted by removing any minus signs. If raw scores were used then a lab that had a score of +1 one month and a score of −1 the next month would have an average score of zero when in fact it had missed the mean by one standard deviation each month and, thus, the average adjusted z-score should be 1. The adjusted z-value is the one used to make comparisons. An adjusted mean z-value of under 2.000 for a calendar year earns a rating of satisfactory accuracy and precision. An adjusted mean z-value of under 1.000 earns a rating of outstanding accuracy and precision.

Examples

Table I contains the adjusted z-scores for six fictional laboratories that participated in a particular check sample series for 12 months. Note that labs A, C, E, and F all missed the mean by an average of 0.67 standard deviations for the year and would be rated Outstanding. Labs B and D missed by an average of 1.25 standard deviations and would be rated satisfactory. (Remember that z-score is the number of standard deviations from the mean.)

Without going further we would assume that Labs A, C, E, and F had similar results as did Labs B and D. For the purposes of estimating precision, these are the correct assumptions. All of these results are based on adjusted z-scores. If we go back to the actual z-scores and use a graphical approach, we can learn much more about the labs. Table II contains the actual z-scores before adjustment. Figures 1A–F show the data from Table II. These figures are an application of control charts which are one of the basic tools of quality control. You will note that the vertical axes of the charts is plus and minus three standard deviations (six sigma).

Labs A and B had analytical estimates around the mean value all year with Lab B showing more variation, which we already knew from Table I. We can assume that Lab A will produce an accurate and precise analysis. With Lab B, we would expect no systematic biases, but if we want a more precise analysis, we should consider analyzing more samples and then averaging the results. We can go back to the original data (before z-score trans-

Table I. Adjusted z-scores

<table>
<thead>
<tr>
<th>Month</th>
<th>Lab A</th>
<th>Lab B</th>
<th>Lab C</th>
<th>Lab D</th>
<th>Lab E</th>
<th>Lab F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.88</td>
<td>0.66</td>
<td>0.72</td>
<td>1.68</td>
<td>1.17</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.53</td>
<td>2.15</td>
<td>0.24</td>
<td>0.82</td>
<td>1.25</td>
<td>0.11</td>
</tr>
<tr>
<td>3</td>
<td>1.03</td>
<td>0.65</td>
<td>1.04</td>
<td>2.18</td>
<td>1.13</td>
<td>0.16</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.98</td>
<td>0.20</td>
<td>0.45</td>
<td>0.91</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>1.42</td>
<td>1.87</td>
<td>0.94</td>
<td>1.08</td>
<td>0.62</td>
<td>0.16</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>1.11</td>
<td>0.36</td>
<td>0.32</td>
<td>0.75</td>
<td>0.33</td>
</tr>
<tr>
<td>7</td>
<td>0.84</td>
<td>1.10</td>
<td>1.57</td>
<td>1.95</td>
<td>0.36</td>
<td>0.64</td>
</tr>
<tr>
<td>8</td>
<td>0.22</td>
<td>1.53</td>
<td>0.25</td>
<td>0.14</td>
<td>0.07</td>
<td>0.77</td>
</tr>
<tr>
<td>9</td>
<td>0.96</td>
<td>2.02</td>
<td>0.55</td>
<td>2.30</td>
<td>0.42</td>
<td>1.11</td>
</tr>
<tr>
<td>10</td>
<td>0.55</td>
<td>1.75</td>
<td>0.11</td>
<td>1.81</td>
<td>0.26</td>
<td>1.33</td>
</tr>
<tr>
<td>11</td>
<td>0.04</td>
<td>0.87</td>
<td>0.78</td>
<td>0.95</td>
<td>0.44</td>
<td>1.61</td>
</tr>
<tr>
<td>12</td>
<td>1.11</td>
<td>0.31</td>
<td>1.28</td>
<td>1.32</td>
<td>0.66</td>
<td>1.55</td>
</tr>
<tr>
<td>Ave.</td>
<td>0.67</td>
<td>1.25</td>
<td>0.67</td>
<td>1.25</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

1 Member, Check Sample Committee, AACC International. USDA-ARS, Peoria, IL; Terry.Nelsen@ARS.USDA.GOV.
formation), and if the analyses were done in duplicate or tripli-
cate, we can estimate how much of the imprecision is due to dif-
ferences between analyses of the same sample on a common date
(a repeatability issue) or between different samples analyzed
on different dates (a reproducibility issue) or a combination. Before
any of these analyses, however, the end user of the data should
decide how much variation is acceptable. If only a rough estimate
of a concentration is required, then both Labs A and B would
produce acceptable estimates and time and money need not be
spent on improvements where improvements aren’t needed.

Lab C had a higher estimate than the overall average for all 12
months. Lab C did not show a great deal of variance but its results
apparently are biased. (In technical terms, Lab C was precise but
not accurate.) A consistent bias can be caused by many things, but
it means that Lab C consistently did something different or used
different chemicals or equipment than that specified in the meth-
method. These differences are usually inadvertent and can be as simple
as temperature or humidity differences. Analytical methods may
have a step or instruction that is open to different interpretations
and can lead to a systematic error. We hope that the collaborative
study that was used to evaluate the method would have uncovered
any hidden possibilities of bias. I have examined check sample
results for several years and I have been surprised at the number
of systematic biases found. I have come to believe that many of
the differences we call “batch effects” can be caused by biases in
analytical procedures. Batch effects are a problem for quality
control officers in the food, chemical, pharmaceutical, and even
the paint industries. A producer uses the same ingredients in
the same equipment in the same procedures and ends up with a differ-
ent result. As an example, let’s imagine that you and I are both
told to bake a product with a given procedure that calls for a flour
of exactly 12% protein. Suppose the flour we have on hand just
happens to be exactly 12%. When I measure it I underestimate the
protein to be 11.8% so I blend it with a higher protein flour until I
believe that my flour is at 12%. You measure the same original
flour and get an overestimate of 12.2% so you blend it with a
lower protein flour to achieve what you believe is exactly 12%.
When we then each bake our separate products, they are different.
When the quality control analyst tries to find why they are different,
the specifications state that both products were baked with
exactly 12% protein, when, in fact, they weren’t. The quality con-
trol analyst has records showing that we both used 12% flour and
no way of knowing that we actually used flours with different
protein content.

Lab D showed considerable variation and for 10 of the 12
months its estimates were above the overall average. Lab D was
neither precise nor accurate. We can do a statistical test and find
that 2 of the 12 months indicate a significant (P < .05) bias.

Lab E’s estimates showed a trend in bias. Lab E started the year
by underestimating the sample concentration and ended the year
by overestimating the sample concentration. The quality control
detective will look for wear on instruments, aging reagents, etc.

Lab F stayed somewhat accurate throughout the year but its
precision deteriorated. The quality control detective will look for
a reason.

<table>
<thead>
<tr>
<th>Month</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−0.88</td>
<td>−0.66</td>
<td>0.72</td>
<td>1.68</td>
<td>−1.17</td>
<td>−0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.53</td>
<td>2.15</td>
<td>0.24</td>
<td>0.82</td>
<td>−1.25</td>
<td>−0.11</td>
</tr>
<tr>
<td>3</td>
<td>1.03</td>
<td>−0.65</td>
<td>1.04</td>
<td>2.18</td>
<td>−1.13</td>
<td>0.16</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.98</td>
<td>0.20</td>
<td>−0.45</td>
<td>−0.91</td>
<td>−0.22</td>
</tr>
<tr>
<td>5</td>
<td>−1.42</td>
<td>−1.87</td>
<td>0.94</td>
<td>1.08</td>
<td>−0.62</td>
<td>0.16</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>1.11</td>
<td>0.36</td>
<td>0.32</td>
<td>−0.75</td>
<td>0.33</td>
</tr>
<tr>
<td>7</td>
<td>−0.84</td>
<td>−1.10</td>
<td>1.57</td>
<td>1.95</td>
<td>−0.36</td>
<td>0.64</td>
</tr>
<tr>
<td>8</td>
<td>−0.22</td>
<td>−1.53</td>
<td>0.25</td>
<td>−0.14</td>
<td>−0.07</td>
<td>−0.77</td>
</tr>
<tr>
<td>9</td>
<td>0.96</td>
<td>−2.02</td>
<td>0.55</td>
<td>2.30</td>
<td>0.42</td>
<td>−1.11</td>
</tr>
<tr>
<td>10</td>
<td>−0.55</td>
<td>1.75</td>
<td>0.11</td>
<td>1.81</td>
<td>0.26</td>
<td>1.33</td>
</tr>
<tr>
<td>11</td>
<td>−0.04</td>
<td>0.87</td>
<td>0.78</td>
<td>0.95</td>
<td>0.44</td>
<td>1.61</td>
</tr>
<tr>
<td>12</td>
<td>−1.11</td>
<td>0.31</td>
<td>1.28</td>
<td>1.32</td>
<td>0.66</td>
<td>−1.55</td>
</tr>
</tbody>
</table>

| Ave.  | −0.17 | −0.06 | 0.67 | 1.15 | −0.37 | 0.04 |

Table II. Actual z-scores
Conclusions

This version of a control chart can be very useful for an analytical laboratory or a user of laboratory services. The chart visually shows a history or a trend. I do advise some caution, however, before jumping to conclusions. Quality control specialists suggest that 7–10 points on the chart are needed to show true (i.e., significant) biases or trends. A series that is analyzed monthly can keep a running chart of the last 12 monthly estimates. Bimonthly series could use two years of data and quarterly series could use three years of data. Some check sample programs (other than AACC Intl.’s) will use data from the same sample analyzed on two different days for an estimate of bias and precision. These data can be useful as a “snapshot” estimate of bias, but even these data should be considered over a longer time span if the data are available.

Finally, the user of the data should determine how much variation is acceptable for different purposes. Often times a rough estimate is sufficient while some analytes may need to be estimated very precisely. A running control chart can be a useful tool for long-term quality control.

Edited by Michel Dubois, Arnaud Dubat, and Bernard Launay

This new edition provides an understanding of the technical data generated by the instrument and gives timely application examples. This is the first revision of this resource in 20 years and it explains major modifications and improvements of the Alveograph through new and completely revised chapters. This handbook is essential for alveograph users. It helps you interpret results and modify procedures to improve product quality and consistency. The troubleshooting section alone is worth the price of the book.

2008; 8.5" x 11" softcover; 86 pages; 23 black and white illustrations; ISBN 978-1-891127-56-4; (1 pound); Item No. 27564

ORDER TODAY!
www.aaccnet.org (click “Books”)

Toll-Free 1.800.328.7560
in the U.S. and most of Canada; +1.651.454.7250 elsewhere

Gain the Laboratory Testing Advantage with the AACC International Check Sample and Proficiency Testing Services

“...enables us to substantiate personnel accuracy... helps to assure customers that we are accountable for their product specifications.”
— Brenda Stow, Panhandle Milling Co

“...gives us confidence that our instruments are working within the norm of other similar instruments.”
— Irene Parisien, Agriculture and Agri-Food Canada

www.aaccnet.org/checksample