Attribute Measurement Error Analysis

Attribute data consist of classifications rather than measurements. Attribute inspection involves determining the classification of an item, for example, is it “good” or “bad”? The principles of good measurement for attribute inspection are the same as for measurement inspection (Table 9.7). Thus, it is possible to evaluate attribute measurement systems in

<table>
<thead>
<tr>
<th>Measurement Concept</th>
<th>Interpretation for Attribute Data</th>
<th>Suggested Metrics and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Items are correctly categorized.</td>
<td>Number of times correctly classified by all inspectors. Requires knowledge of the “true” value. Total number of evaluations by all inspectors.</td>
</tr>
<tr>
<td>Bias</td>
<td>The proportion of items in a given category is correct.</td>
<td>Overall average proportion in a given category for all inspectors minus correct proportion in a given category. Averaged over all categories. Requires knowledge of the “true” value.</td>
</tr>
<tr>
<td>Repeatability</td>
<td>When an inspector evaluates the same item multiple times in a short time interval, she assigns it to the same category every time.</td>
<td>For a given inspector: Total number of times repeat classifications agree. Total number of repeat classifications. Overall: Average of repeatabilities.</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>When all inspectors evaluate the same item, they all assign it to the same category.</td>
<td>Total number of times classifications for all concur. Total number of classifications.</td>
</tr>
<tr>
<td>Linearity</td>
<td>When an inspector evaluates items covering the full set of categories, her classifications are consistent across the categories.</td>
<td>Range of inaccuracy and bias across all categories. Requires knowledge of the “true” value. Note: Because there is no natural ordering for nominal data, the concept of linearity doesn’t really have a precise analog for attribute data on this scale. However, the suggested metrics will highlight interactions between inspectors and specific categories.</td>
</tr>
</tbody>
</table>

Table 9.7 Attribute Measurement Concepts
much the same way as we evaluate variable measurement systems. Much less work has been done on evaluating attribute measurement systems. The proposals provided in this book are those I’ve found to be useful for my employers and clients. The ideas are not part of any standard and you are encouraged to think about them critically before adopting them. I also include an example of Minitab’s attribute gage R&R analysis.

Operational Definitions
An operational definition is defined as a requirement that includes a means of measurement. “High quality solder” is a requirement that must be operationalized by a clear definition of what “high quality solder” means. This might include verbal descriptions, magnification power, photographs, physical comparison specimens, and many more criteria.

Examples of Operational Definitions
1. Operational definition of the Ozone Transport Assessment Group’s (OTAG) goal
   **Goal:** To identify reductions and recommend transported ozone and its precursors which, in combination with other measures, will enable attainment and maintenance of the ozone standard in the OTAG region.
   Suggested operational definition of the goal:
   1. A general modeled reduction in ozone and ozone precursors aloft throughout the OTAG region; and
   2. A reduction of ozone and ozone precursors both aloft and at ground level at the boundaries of non-attainment area modeling domains in the OTAG region; and
   3. A minimization of increases in peak ground level ozone concentrations in the OTAG region (This component of the operational definition is in review.).

2. Wellesley College Child Care Policy Research Partnership operational definition of unmet need
   1. Standard of comparison to judge the adequacy of neighborhood services: the median availability of services in the larger region (Hampden County).
   2. Thus, our definition of unmet need: The difference between the care available in the neighborhood and the median level of care in the surrounding region (stated in terms of child care slots indexed to the age-appropriate child population—“slots-per-tots”).

3. Operational definitions of acids and bases
   1. An *acid* is any substance that increases the concentration of the H$^+$ ion when it dissolves in water.
   2. A *base* is any substance that increases the concentration of the OH$^-$ ion when it dissolves in water.

4. Operational definition of “intelligence”
   1. Administer the Stanford-Binet IQ test to a person and score the result. The person’s intelligence is the score on the test.

5. Operational definition of “dark blue carpet”
A carpet will be deemed to be dark blue if

1. Judged by an inspector medically certified as having passed the U.S. Air Force test for color-blindness
   1.1. It matches the PANTONE color card 7462 C when both carpet and card are illuminated by GE “cool white” fluorescent tubes;
   1.2. Card and carpet are viewed at a distance between 16 and 24 inches.

**How to Conduct Attribute Inspection Studies**

Some commonly used approaches to attribute inspection analysis are shown in Table 9.8.

**Example of Attribute Inspection Error Analysis**

Two sheets with identical lithographed patterns are to be inspected under carefully controlled conditions by each of the three inspectors. Each sheet has been carefully examined multiple times by journeymen lithographers and they have determined that one of the sheets should be classified as acceptable, the other as unacceptable. The inspectors sit on a stool at a large table where the sheet will be mounted for inspection. The inspector can adjust the height of the stool and the angle of the table. A lighted magnifying glass is mounted to the table with an adjustable arm that lets the inspector move it to any part of the sheet (see Fig. 9.20).

Each inspector checks each sheet once in the morning and again in the afternoon. After each inspection, the inspector classifies the sheet as either acceptable or unacceptable. The entire study is repeated the following week. The results are shown in Table 9.9.

In Table 9.9 the part column identifies which sheet is being inspected, and the standard column is the classification for the sheet based on the journeymen’s evaluations. A 1 indicates that the sheet is acceptable, a 0 that it is unacceptable. The columns labeled InspA, InspB, and InspC show the classifications assigned by the three inspectors respectively. The reproducible column is a 1 if all three inspectors agree on the classification, whether their classification agrees with the standard or not. The accurate column is a 1 if all three inspectors classify the sheet correctly as shown in the standard column.

**Individual Inspector Accuracy**

Individual inspector accuracy is determined by comparing each inspector’s classification with the standard. For example, in cell C2 of Table 9.9 inspector A classified the unit as acceptable, and the standard column in the same row indicates that the classification is correct. However, in cell C3 the unit is classified as unacceptable when it actually is acceptable. Continuing this evaluation shows that inspector A made the correct assessment 7 out of 8 times, for an accuracy of 0.875 or 87.5%. The results for all inspectors are given in Table 9.10.

**Repeatability and Pairwise Reproducibility**

Repeatability is defined in Table 9.7 as the same inspector getting the same result when evaluating the same item more than once within a short time interval. Looking at InspA
<table>
<thead>
<tr>
<th>True Value</th>
<th>Method of Evaluation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expert Judgment</strong></td>
<td>An expert looks at the classifications after the operator makes normal classifications and decides which are correct and which are incorrect.</td>
<td>□ <strong>Metrics:</strong> 1. Percent correct 2. Quantifies the accuracy of the classifications 3. Simple to evaluate 4. Who says the expert is correct? 5. Care must be taken to include all types of attributes 6. Difficult to compare operators since different units are classified by different people 7. Acceptable level of performance must be decided upon. Consider cost, impact on customers, etc</td>
</tr>
<tr>
<td><strong>Known</strong></td>
<td><strong>Round Robin Study:</strong> A set of carefully identified objects is chosen to represent the full range of attributes. 1. Each item is evaluated by an expert and its condition recorded. 2. Each item is evaluated by every inspector at least twice.</td>
<td>□ <strong>Metrics:</strong> 1. Percent correct by inspector 2. Inspector repeatability 3. Inspector reproducibility 4. Stability 5. Inspector “linearity” 6. Full range of attributes included 7. All aspects of measurement error quantified 8. People know they’re being watched, may affect performance 9. Not routine conditions 10. Special care must be taken to insure rigor 11. Acceptable level of performance must be decided upon for each type of error. Consider cost, impact on customers, etc</td>
</tr>
<tr>
<td><strong>Unknown</strong></td>
<td><strong>Concurrence Study:</strong> A set of carefully identified objects is chosen to represent the full range of attributes, to the extent possible. 1. Each item is evaluated by every inspector at least twice.</td>
<td>□ Like a round robin, except true value isn’t known 2. No measures of accuracy or bias are possible. Can only measure agreement between equally qualified people 3. Full range of attributes included 4. People know they’re being watched, may affect performance 5. Not routine conditions 6. Special care must be taken to insure rigor 7. Acceptable level of performance must be decided upon for each type of error. Consider cost, impact on customers, etc</td>
</tr>
</tbody>
</table>

**Table 9.8** Methods of Evaluating Attribute Inspection
we see that when she evaluated Part 1 in the morning of “Today” she classified it as acceptable (1), but in the afternoon she said it was unacceptable (0). The other three morning/afternoon classifications matched each other. Thus, her repeatability is 3/4 or 75%.

Table 9.9 Results of Lithography Attribute Inspection Study

<table>
<thead>
<tr>
<th>Part</th>
<th>Standard</th>
<th>InspA</th>
<th>InspB</th>
<th>InspC</th>
<th>Date</th>
<th>Time</th>
<th>Reproducible</th>
<th>Accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Today</td>
<td>Morning</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Today</td>
<td>Afternoon</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Today</td>
<td>Morning</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Today</td>
<td>Afternoon</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Last Week</td>
<td>Morning</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Last Week</td>
<td>Afternoon</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Last Week</td>
<td>Morning</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Last Week</td>
<td>Afternoon</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 9.10 Inspector Accuracies

<table>
<thead>
<tr>
<th>Inspector</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>87.5%</td>
<td>100.0%</td>
<td>62.5%</td>
</tr>
</tbody>
</table>
Pairwise reproducibility is the comparison of each inspector with every other inspector when checking the same part at the same time on the same day. For example, on Part 1/Morning/Today, InspA’s classification matched that of InspB. However, for Part 1/Afternoon/Today InspA’s classification was different than that of InspB. There are eight such comparisons for each pair of inspectors. Looking at InspA versus InspB we see that they agreed 7 of the 8 times, for a pairwise repeatability of $7/8 = 0.875$.

In Table 9.11 the diagonal values are the repeatability scores and the off-diagonal elements are the pairwise reproducibility scores. The results are shown for “Today,” “Last Week” and both combined.

**Overall Repeatability, Reproducibility, Accuracy and Bias**

Information is always lost when summary statistics are used, but the data reduction often makes the tradeoff worthwhile. The calculations for the overall statistics are operationally defined as follows:

- **Repeatability** is the average of the repeatability scores for the 2 days combined; that is, $(0.75 + 1.00 + 0.25)/3 = 0.67$.
- **Reproducibility** is the average of the reproducibility scores for the 2 days combined (see Table 9.9); that is,

$$\left( \frac{1+0+0+0}{4} + \frac{1+0+0+1}{4} \right) / 2 = 0.50$$

- **Accuracy** is the average of the accuracy scores for the 2 days combined (see Table 9.9); that is,

$$\left( \frac{1+0+0+0}{4} + \frac{1+0+0+0}{4} \right) / 2 = 0.25$$

- **Bias** is the estimated proportion in a category minus the true proportion in the category. In this example the true percent defective is 50% (1 part in 2). Of the 24 evaluations, 12 evaluations classified the item as defective. Thus, the bias is $0.5 - 0.5 = 0$.

**Overall Stability**

Stability is calculated for each of the above metrics separately, as shown in Table 9.12.
Interpretation of Results

1. The system overall appears to be unbiased and accurate. However, the evaluation of individual inspectors indicates that there is room for improvement.

2. The results of the individual accuracy analysis indicate that inspector C has a problem with accuracy, see Table 9.10.

3. The results of the R&R (pairwise) indicate that inspector C has a problem with both repeatability and reproducibility, see Table 9.11.

4. The repeatability numbers are not very stable (Table 9.12). Comparing the diagonal elements for Today with those of Last Week in Table 9.11, we see that inspectors A and C tended to get different results for the different weeks. Otherwise the system appears to be relatively stable.

5. Reproducibility of inspectors A and B is not perfect. Some benefit might be obtained from looking at reasons for the difference.

6. Since inspector B’s results are more accurate and repeatable, studying her might lead to the discovery of best practices.

Minitab Attribute Gage R&R Example

Minitab includes a built-in capability to analyze attribute measurement systems, known as “attribute gage R&R.” We will repeat the above analysis using Minitab.

Minitab can’t work with the data as shown in Table 9.9, it must be rearranged. Once the data are in a format acceptable to Minitab, we enter the Attribute Gage R&R Study dialog box by choosing Stat > Quality Tools > Attribute Gage R&R Study (see Fig. 9.21). Note the checkbox “Categories of the attribute data are ordered.” Check this box if the data are ordinal and have more than two levels. Ordinal data means, for example, a 1 is in some sense “bigger” or “better” than a 0. For example, if we ask raters in a taste test a question like the following: “Rate the flavor as 0 (awful), 1 (OK), or 2 (delicious).” Our data are ordinal (acceptable is better than unacceptable), but there are only two levels, so we will not check this box.

Minitab evaluates the repeatability of appraisers by examining how often the appraiser “agrees with him/herself across trials.” It does this by looking at all of the classifications for each part and counting the number of parts where all classifications agreed. For our example each appraiser looked at two parts four times each. Minitab’s output, shown in Fig. 9.22, indicates that InspA rated 50% of the parts consistently,
Inspected Matched Percent (%) 95.0% CI
InspectA 2 1 50.0 (1.3, 98.7)
InspectB 2 2 100.0 (22.4, 100.0)
InspectC 2 0 0.0 (0.0, 77.6)

# Matched: Appraiser agrees with him/herself across trials.

Figure 9.22 Minitab within appraiser assessment agreement.

InspecA 100.0%, and InspecC 0%. The 95% confidence interval on the percentage agreement is also shown. The results are displayed graphically in Fig. 9.23.

Accuracy Analysis
Minitab evaluates accuracy by looking at how often all of an appraiser’s classifications for a given part agree with the standard. Figure 9.24 shows the results for our example. As before, Minitab combines the results for both days. The plot of these results is shown in Fig. 9.25.

Minitab also looks at whether or not there is a distinct pattern in the disagreements with the standard. It does this by counting the number of times the appraiser classified an item as a 1 when the standard said it was a 0 (the 1/0 Percent column), how often the appraiser classified an item as a 0 when it was a 1 (the 0/1 Percent column), and how often the...
**Figure 9.23** Plot of within appraiser assessment agreement.

**Each Appraiser vs Standard**

Assessment Agreement

<table>
<thead>
<tr>
<th>Appraiser</th>
<th># Inspected</th>
<th># Matched</th>
<th>Percent (%)</th>
<th>95.0% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>InspA</td>
<td>2</td>
<td>1</td>
<td>50.0</td>
<td>(1.3, 98.7)</td>
</tr>
<tr>
<td>InspB</td>
<td>2</td>
<td>2</td>
<td>100.0</td>
<td>(22.4, 100.0)</td>
</tr>
<tr>
<td>InspC</td>
<td>2</td>
<td>0</td>
<td>0.0</td>
<td>(0.0, 77.6)</td>
</tr>
</tbody>
</table>

# Matched: Appraiser's assessment across trials agrees with standard

**Figure 9.24** Minitab appraiser versus standard agreement.

**Figure 9.25** Plot of appraiser versus standard assessment agreement.
appraiser’s classifications were mixed, i.e., is not repeatable (the # Mixed Percent column). The results are shown in Fig. 9.26. The results indicate that there is no consistent bias, defined as consistently putting a unit into the same wrong category. The problem, as was shown in the previous analysis, is that appraisers A and C are not repeatable.

**Between Appraiser Assessments**
Next, Minitab looks at all of the appraiser assessments for each part and counts how often every appraiser agrees on the classification of the part. The results, shown in Fig. 9.27, indicate that this never happened during our experiment. The 95% confidence interval is also shown.

**All Appraisers versus Standard**
Finally, Minitab looks at all of the appraiser assessments for each part and counts how often every appraiser agrees on the classification of the part and their classification agrees with the standard. This can’t be any better than the between appraiser assessment agreement shown in Fig. 9.27. Unsurprisingly, the results, shown in Fig. 9.28, indicate that this never happened during our experiment. The 95% confidence interval is also shown.