Outliers and Their Effect on Distribution Assessment
Topics of Discussion

- What is an Outlier? (Definition)
  - How can we use outliers
- Analysis of Outlying Observation
  - The Standards
  - Tests for Outliers in Minitab
- Importance of Data Distributions
- How Outliers Affect Distribution Assessment
  - Examples
What are Outlying Observations?

Do they have value?
What are Outliers?

- An outlier is an observation point that is distant from other observations.
- An outlier may be due to variability in the measurement or it may indicate experimental error: the latter are sometimes excluded from the data set.
- An outlier may be caused by a defective unit or a problem in the process.
- ISO 16269 defines it as “member of a small subset of observations that appears to be inconsistent with the remainder of a given sample.”
Anomaly detection (also known as outlier detection) is the search for items or events which do not conform to an expected pattern.

Anomalies are also referred to as outliers, change, deviation, surprise, aberrant, intrusion, etc.

An outlier may be caused by a defective unit or a problem in the process.
There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise.

Every effort must be made to determine what is causing the outlier to exist. Testing, equipment, operator error, materials, etc. all must be reviewed and assessment made.
What to do with outliers

• Outliers must be investigated

• Use the tools found in root cause analysis
  • Interview operators
  • Evaluate the measuring system
  • Try to duplicate the reading

• Don’t make assumptions or jump to conclusions

• Do not remove outliers simply because they are outliers
The Importance of Outliers

Note that outliers are not necessarily "bad" data-points; indeed they may well be the most important, most information rich, part of the dataset. Under no circumstances should they be automatically removed from the dataset. Outliers may deserve special consideration: they may be the key to the phenomenon under study or the result of human blunders.
1.1.1 An outlying observation may be merely an extreme manifestation of the random variability inherent in the data. If this is true, the value should be retained and processed in the same manner as the other observations in the sample.

1.1.2 On the other hand, an outlying observation may be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value. In such cases, it may be desirable to institute an investigation to ascertain the reason for the aberrant value.
Outliers  Extreme values in a data set that should be discarded before analysis.

WHAT???
Considerations before collecting data
What distribution is expected for continuous data?

- 3 distributions cover most circumstances
  - Random variables which are **added** together follow a **normal** distribution
    - Human height, weight, many dimensions
  - Random variables which are **multiplied** together follow a **lognormal** distribution
    - Also useful for data which are bounded at zero like tensile strength, burst strength. These are skewed right because there are no negative values. Cycle time and microbiologic data often follow the lognormal distribution.
  - **Weibull** distribution is versatile for skewed data.
    - Time to failure, fatigue life.
Analysis of Outliers

The Standards and Minitab
Standards Relating to Outliers

- ASTM E178-08 “Standard Practice for Dealing with Outlying Observations”
18 page document that focuses on Dixon Test.
Includes table for Tietjen-Moore Critical Values (One-Sided Test) for Lk
Offers an example with aas laboratory analysis
References Wilk-Shapiro W Statistic as a possible test for outliers
- 54 Page document that offers a myriad of methods for detecting and treating outliers.
- Methods include Box plot, Cochran’s test, Greenwood’s test.
- Full explanation of GESD (Generalized Extreme Studentized Deviate) method
- Demonstrates various graphical representations of data (Box plot, histogram, stem-and-leaf and probability plot)
- Much emphasis put on distribution understanding including Box-Cox plot and analysis.
Testing for outliers

- Other methods flag observations based on measures such as the interquartile range. For example, if $Q_1$ and $Q_3$ are the lowest and upper quartiles respectively, then one could define an outlier to be any observation outside the range:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$
The Box Plot

http://www.physics.csbsju.edu/stats/box2.html
The boxplot is interpreted as follows:

- The box itself contains the middle 50% of the data. The upper edge (hinge) of the box indicates the 75th percentile of the data set, and the lower hinge indicates the 25th percentile. The range of the middle two quartiles is known as the inter-quartile range.

- The line in the box indicates the median value of the data.

- If the median line within the box is not equidistant from the hinges, then the data is skewed.

- The ends of the vertical lines or "whiskers" indicate the minimum and maximum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range.

- The points outside the ends of the whiskers are outliers or suspected outliers.
Testing for outliers

- Model-based methods which are commonly used for identification assume that the data are from a normal distribution, and identify observations which are deemed “unlikely” based on mean and standard deviation.
Three good tests for outliers found in Minitab:

- Box plot
- Grubbs Test
- Dixon’s Test
  - Dixon’s Q ratio
  - Dixon’s 11 ratio
  - Dixon’s 12 ratio
  - Dixon’s 20 ratio
  - Dixon’s 21 ratio
  - Dixon’s 22 ratio
Outlier Testing in Minitab

Stat> Basic Statistics> Outlier Test

First determine the test you will use. The Dixon’s numbered ratios are based on sample size.

Next select if you want to evaluate the smallest, largest or smallest or largest value to evaluate.
Grubbs test

Dixon’s Q Ratio test for the same data set.

**Outlier Test: Outlier**

**Method**

Null hypothesis: All data values come from the same normal population

Alternative hypothesis: Largest data value is an outlier

Significance level: \( \alpha = 0.05 \)

**Dixon's Q Test**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier</td>
<td>25</td>
<td>18.26</td>
<td>18.638</td>
<td>22.761</td>
<td>26.10</td>
<td>0.43</td>
<td>0.002</td>
</tr>
</tbody>
</table>

x[i] denotes the ith smallest observation.

**Outlier**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Row</th>
<th>Outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier</td>
<td>15</td>
<td>26.1</td>
</tr>
</tbody>
</table>

**Dixon’s Q Ratio**
Simple Boxplot

Boxplot of Outlier
Other Tests for Outliers

- Chauvenet’s criterion
- Cochran’s C Test – Variance Outlier Test
- Peircece’s criterion
- The Jackknife Method
- Greenwood’s Test
The Importance of Distribution Fitting

Impact on Statistical Tests
Why test for normality?

- Reliability predictions are based on model (distribution)
- If model is inappropriate, misleading conclusions can be drawn.
Sensitive to normality

**Non-Sensitive**
- Confidence intervals on means
- T-tests
- ANOVA, including DOE and Regression
- X-bar Charts

**Sensitive**
- Confidence Intervals on standard deviations
- Tolerance Intervals, Reliability Intervals
- Variance tests
- $C_{pk}$, $C_p$, $P_{pk}$, $P_p$
- I-Charts
Reasons for failing normality

- Measurement resolution
- Data shift
- Multiple sources of data
- Truncated data
- Outliers
- Too much data
- The underlying distribution is not normal
The Effect of Outliers on Distribution

Some Examples
Summary for EW Data 2

Anderson-Darling Normality Test

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-Squared</td>
<td>0.70</td>
</tr>
<tr>
<td>P-Value</td>
<td>0.060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>110.71</td>
</tr>
<tr>
<td>StDev</td>
<td>7.99</td>
</tr>
<tr>
<td>Variance</td>
<td>63.87</td>
</tr>
<tr>
<td>Skewness</td>
<td>-1.17564</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1.99926</td>
</tr>
<tr>
<td>N</td>
<td>26</td>
</tr>
<tr>
<td>Minimum</td>
<td>86.20</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>104.40</td>
</tr>
<tr>
<td>Median</td>
<td>112.90</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>116.80</td>
</tr>
<tr>
<td>Maximum</td>
<td>122.40</td>
</tr>
</tbody>
</table>

95% Confidence Interval for Mean
107.48 to 113.94

95% Confidence Interval for Median
107.73 to 116.08

95% Confidence Interval for StDev
6.27 to 11.03
Summary for EW Data 2 No Outlier

<table>
<thead>
<tr>
<th>Anderson-Darling Normality Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-Squared</td>
</tr>
<tr>
<td>P-Value</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>111.69</td>
</tr>
<tr>
<td>StDev</td>
<td>6.36</td>
</tr>
<tr>
<td>Variance</td>
<td>40.51</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.353437</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.992428</td>
</tr>
<tr>
<td>N</td>
<td>25</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>100.40</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>104.70</td>
</tr>
<tr>
<td>Median</td>
<td>113.00</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>116.80</td>
</tr>
<tr>
<td>Maximum</td>
<td>122.40</td>
</tr>
</tbody>
</table>

95% Confidence Interval for Mean
109.06   114.32

95% Confidence Interval for Median
109.24   116.44

95% Confidence Interval for StDev
4.97     8.85

Project: Normality Examples 2011-06-22.MPJ
Summary Statistics Output

Summary for C2

Anderson-Darling Normality Test
- A-Squared: 2.77
- P-Value: 0.005

Summary Statistics:
- Mean: 3.9883
- StDev: 4.1086
- Variance: 16.8803
- Skewness: 1.90297
- Kurtosis: 3.02385
- N: 30

Confidence Intervals:
- 95% Confidence Interval for Mean: 2.4541, 5.5224
- 95% Confidence Interval for Median: 1.8680, 3.7050
- 95% Confidence Interval for StDev: 3.2721, 5.5232

Project: NORMALITY EXAMPLES 2011-06-22.MPJ
Investigate Non-Normal Distributions

Probability Plot for C2

Goodness of Fit Test

Normal
AD = 2.774
P-Value < 0.005

Lognormal
AD = 0.208
P-Value = 0.852

Project: NORMALITY EXAMPLES 2011-06-22.MPJ
Outlier or Not?

Data set – 200, 440, 220, 1100, 80, 55, 60, 1700, 100, 200, 160, 110, 170, 210, 3600, 90, 85, 35, 430, 180, 5, 15, 1, 30, 10, 10, 20, 15, 20, 15.
Dixon’s Q Ratio

Boxplot of Results

Outlier Test: Results

Method

Null hypothesis: All data values come from the same normal population
Alternative hypothesis: Largest data value is an outlier
Significance level: $\alpha = 0.05$

Dixon's Q Test

Results 30 1 5 1700 3600 0.53 0.000

$x[i]$ denotes the $i$th smallest observation.

Outlier

Variable Row Outlier
Results 15 3600

Outlier Plot of Results
Grubbs Test

Outlier Test: Results

Method
Null hypothesis: All data values come from the same normal population
Alternative hypothesis: Largest data value is an outlier
Significance level: $\alpha = 0.05$

Grubbs' Test

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Min</th>
<th>Max</th>
<th>G</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>30</td>
<td>312</td>
<td>715</td>
<td>1</td>
<td>3600</td>
<td>4.60</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Outlier

<table>
<thead>
<tr>
<th>Variable</th>
<th>Row</th>
<th>Outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>15</td>
<td>3600</td>
</tr>
</tbody>
</table>
Outlier or Not?

Data set – 200, 440, 220, 1100, 80, 55, 60, 1700, 100, 200, 160, 110, 170, 210, 3600, 90, 85, 35, 430, 180, 5, 15, 1, 30, 10, 10, 20, 15, 20, 15.
Review the Probability Plot

Probability Plot of Results
Normal

Mean 312.2
StDev 715.1
N 30
AD 6.234
P-Value <0.005
Conclusion
The sample with a reading of 3600 is expected to occur approximately 1 out of every 67 readings on an average and should not be considered extreme based on the distribution of the data. The sample with this count of Aerobes should be identified through a gram stain to determine the nature of its source, but unless abnormal concerns are raised based on the type of microorganism, no additional action should be taken.
Probability Plot with and without Outliers

The plot on the left shows a typical normal distribution. The one on the right contains two data points as outlying observations.

Probability plots can give us great insight into the distribution from our process.
Conclusion

- We make conclusions about processes based on the distribution models we use. Know your process distribution.
- Outliers can be a useful source of data, so investigate causes of outliers.
- Slow down on your assumptions. Look at the data several ways including a probability plot. Try to understand why your data is the way it is.
- When writing up your results, reference standards or technical literature.
References

- ASTM E178-08 “Standard Practice for Dealing With Outlying Observations”
- Minitab 17 Training Material “Analysis of Nonnormal Data for Quality”
Questions?