

Outlier Management

Mavericks, rogues, flyers, wild values, spurious response, unrepresentative data and contaminated data are some of the terms that have been used to describe data that are disturbing to the observer. In fact, Barnett and Lewis, defined outliers as "...observations 'surprisingly far away from the main group'." Other more formal definitions have been proposed. "In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is at fault. Such values are called outliers. Test are available to ascertain whether they can be accepted as homogeneous with the rest of the sample." Marriott.

"We shall define an outliers in a set of data to be an observation or (subset of observations) which appears to be inconsistent with the remainder of that set of data." Barnett and Uwis.

Historically, our attitude toward outliers has been that they are mistakes that must be corrected or removed. Ad hoc and very informal procedures have been used to reject potential outliers. Outliers are always a concern in the pharmaceutical, biotechnology and medical devices areas, because of the implications for the health and safety of patients who trustingly put their lives in the hands of the doctors, pharmacists and manufacturers.

However, outliers have become an even greater interest and concern for all, since the Barr Case. We need scientifically and statistically sound methods that are defensible to each other, the FDA and in a court of law. The issue of unusual data in practice cannot be evaded. Outliers have been an issue with practitioners since before 1852 when B. Pierce commented in his paper *Criterion for the rejection of doubtful observations*:

"In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions,

and the introduction of which into the investigation can only serve . . . to perplex and mislead the inquirer." Barnett and Lewis, p 3.

Handling Outliers in the Past

Two extremes were adopted for outlier management. First, the overly pragmatic advocate, "When in doubt, toss it out." Secondly, the purest who keeps everything in and will not recognize an outlier unless absolute physical evidence is produced. Both approaches may be counterproductive in actual practice.

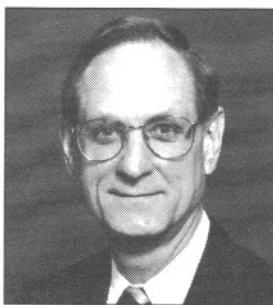
Some outliers are scorned and rejected, some are praised and accepted, depending the impact to the observer. Outliers may be an indication of poor measurement, typing errors, uncontrolled events, excessive variability and outright failures. Outliers may be a blessing. True outliers may indicate new positive breakthroughs, patentable ideas and competitive advantages. Outliers are

not good or bad but are valuable information that must be interpreted, managed and used to advantage. Conscience rational decisions must be made to maximize the information gained and minimize the regulatory risks.

Reasons for Occurrence of Outliers

There are many possible reasons for outliers; these may be known or unknown. Known reasons could include clear, obvious uncontested human errors and events, as in "I dropped the beaker," found calculation mistake, a recording error, a transcription error, reversed or misread digits, had a power "brown out," recalibrated the instrument, used a different raw material, or found that the room temperature or humidity changed. In these cases, the data value(s) are omitted or replaced and the analysis and reporting proceeds. (Note that in Europe a 1 is often written similar to a USA 7 and commas are used in place of periods). Where data is collected to assess homogeneity, content uniformity or similar characteristics, data cannot be rejected,

Continued on page 16



Lynn D. Tobeck

omitted or adjusted unless a clear, obvious uncontested physical reason is found.

However, many, if not most, outliers occur without obvious known reasons. In these cases we need to use all available background knowledge and statistical theory to find a course of action. Unknown reasons can include some from the known category as follows:

1. A single deterministic special cause such as a recording error, transcription error, reversed digits, misreading digits, error in the method of measurement, mis-calibrated instrument, incorrect machine setting and so on may be the reason, but can't be positively identified. (Note that special causes can lead to positive events as well. A different setting on a machine may give improved yields).
2. There could be multiple minor common causes that accumulate to produce an unusually large or small result. The probability of tossing a coin and getting eight heads in a row is less than a half of a percent, an unlikely but still possible event. Likewise, by random chance and as a result of many factors, a machine will produce an unusually large or small part. No error or mistake has been made.
Another example is the crash of USAir flight 427, outside of Pittsburgh on a clear day in September, 1994. "In the end, sources said, investigators expect to find a series of causes—a 'long thin chain' of perhaps improbable events that combined to bring down the plane." *The Washington Post*, Sunday, December 4, 1994, p A10, col 1.
3. The assumptions in the mind of the observer may not match the actual physical mechanisms. The observer may be assuming that the data are from a normal distribution when in fact it is from a log normal or other distribution. The apparently large value is quickly seen as very reasonable in another context.
4. The bulk of data may be from one distribution and the potential outlier may be from another distribution. The data set is contaminated by the

Mr. Torbeck is an international trainer and statistician consultant specializing in applied statistics and experimental design for pharmaceutical and medical device research and development, validation, quality control/assurance and production. Previously, in his over 20 years of experience in industry, he held various manager and director positions at G.D. Searle. In 1997, he won the PDA's Excellence in Teaching Award. Torbeck is a member of PDA, AOAC, American Statistics Association (ASA) and is past president of Chicago chapter of ASA. He holds B.S. and M.S. degrees in Statistics with minors in Computer Science and Operations Research. Contact him at 7812 Kenneth, Skokie, IL 60076-3506.

results of the second distribution. For example, 35 dies and punches on a tablet press are set correctly, but one is misaligned. 97% of the data is reasonable, 3% are "outliers."

How Can You Manage Outliers?

The first step in outlier management is recognition or labeling that potential outliers exist. This usually comes from a visual inspection of the data. The inspection may be done on the raw data or with a graph of some type. The form of presentation is critical to recog-

nition. How the data is presented, influences how we think about the data. Large collections of numbers are difficult to inspect in table form. Graphics such as dot plots, stem and leaf plots, histograms, scatter plots, box plots and time plots show unusual values vividly.

The second step is examination of the data set and all background information. Classify the data set into one variable, two variable or multiple variables. Further classify two variables into, for example, trend data, calibration curves, or paired data

such as heights and weights of ten year old children.

The third step is to propose a defensible statistical model for the data set. For one variable this could be a normal or log normal distribution. For two variables this could be a regression model, a time series or two variable normal distribution. These reference models can be theoretical or actual from previously collected data.

Fallacy 1

An experienced person can make a good decision about an outlier without a formal test. On the contrary, humans are notoriously bad at making these decisions, otherwise there would not be a requirement for double blind clinical trials.

A fourth step is confirmation or identification of outliers. This uses formal statistical significance tests to confirm that surprising values are inconsistent with the known or assumed statistical model. (Typically, mis-called outlier rejection.) An outlier is

Continued from page 16

only an outlier in the context of what we know or believe about the process that is generating the data, i.e. the model. An outlier in one model may be perfectly reasonable in another model.

1. *Accept the value(s)* as valid and as a valuable source of information about the science model being studied. Proceed with the statistical analysis.
2. *Statistically adjust* the outlier value(s) and proceed with the statistical analysis. i.e., replace it with the average.
3. Reject the outlier value(s) and proceed with the statistical analysis.
4. *Reject the current statistical model* for another more correct model. This may be done to include the outlier(s). i.e., use a log-normal model in place of a normal model.
5. Modify procedures to accommodate *future outliers* or contamination by using robust estimators and statistics. i.e., use the median in place of the average.
6. Physically *change the way data is collected* and recorded.
7. *Add the outlier data* to the historical data set for future analysis. Use it to create an updated reference distribution. This should be part of the Annual Product Review.

Fallacy 2

We should reject or accommodate outliers only when a "known" explanation or physical reason is available. This position ignores the possible prior information available from previous data collection in the same or similar circumstances.

Consider These Explanations

Concept 1. Outliers are not necessarily bad, or necessarily good, but information to be managed.

Concept 2. Actions on outliers depend on the known situation and the statistical model that can be defended.

Concept 3. An apparent outlier in one model may be quite unremarkable in another model.

Concept 4. Not all extreme values, large or small, are outliers, but all outliers are extreme values.

Concept 5. Humans seem to have built in "models" that are truncated normals. Anything in the lower 10% or upper 10% is seen as unusual.

Concept 6. Whenever we have more than one data point, unless they are exactly equal, one will be the largest and one will be the smallest. This may or may not have any practical relevance or meaning.

Concept 7. Outliers are more difficult to recognize in two variable and several variable models. The element of surprise is diminished in the complexity of the model.

Concept 8. An outlier in a regression model may not appear as an outlier in either X or Y, if viewed as one variable model.

Concept 9. Data that appears as an outlier today, may upon collection of more data, appear as part of the main group. i.e., 3 DNA assay values vs 300.

Concept 10. "All models are incorrect, but some are useful" G.E.P. Box.

Summary

Rule 1. Data must always be inspected for potential outliers.

Rule 2. Never take action on a value without an investigation.

Rule 3 Don't go fishing for outliers. Look for practical significance first. With $\alpha = 0.05$, 5% of significance tests will reject when the hypothesis is true. That is, if you look long enough you will find something to be significant just by random chance.

References

Dixon, W. J., "Processing Data for Outliers," *Biometrics*, BIOMA, March 1953, Vol. 9, No. 1, pp. 74-89.

ASTM, "Standard Practice for Dealing with Outlying Observations," ASTM E 178, 1980.

Marriott, F. H. C.; *A Dictionary of Statistical Terms*, 5 Ed., John Wiley & Sons, NY, NY, 1990, p. 149.

Iglewicz, B. and Hoaglin, D.; *How to Detect and Handle Outliers*, ASQC Quality Press, Milwaukee, WI, 1993.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, John Wiley, NY, NY.

© Copyright 1997 Lynn D. Torbeck ■