JOURNAL OF CLINICAL RESEARCH BEST PRACTICES

Vol. 8, No. 6, June 2012 "Happy Trials to You"

Making Sense of Biostatistics: Outliers in Clinical Data

By Carleton Southworth

Outliers are data points that are unexpectedly far from the mean, and they can present problems and challenges when analyzing clinical data. They are found in nearly all clinical trials and are due to a variety of causes. They can result from unusual subjects and subject responses, but poor CRF design, inadequate site training, laboratory errors, and data entry and transcription errors can also cause them. Thus, outliers can be classified into two broad categories: (1) outliers that represent true but unusual values, and (2) outliers that represent some kind of measurement error. Either way, outliers are usually a signal that something is wrong, and an investigation is required.[1]

When outliers represent accurate measurements and are included in analyses, they can cause misleading estimates of central tendency. Central tendency is typically provided as the average or mean. For example, because most towns and cities have a few extremely expensive houses (i.e., outliers), housing prices are usually described as the "median house value." Most people want to know the price of a "typical" house, not the mean, which is distorted by the high-priced outliers. The median is the midpoint — half the values are above it and half below it — so a long tail on the distribution does not affect the median value.

Most variables have frequency distributions that form a bell-shaped curve — the "normal" distribution. Many statistical tests rely on the assumption that data are normally distributed. However, if the data are not normally distributed (e.g., there are outliers), then some statistical tests might yield an invalid result, especially with small samples. It is thus important to test for normality and, especially where it is absent, to recognize that outliers may be expected to occur and to prepare to deal with them.

The Standard Deviation

The standard deviation is calculated by determining the distance from the mean for every point in the distribution, disregarding whether the difference is positive or negative, and then averaging the result. With a normal distribution, the average ("mean") and the standard deviation determine approximately how many data points should lie above a particular value and how many should lie below a particular value. For example, if a value is one standard deviation from the mean, about 31.7% of data points should be further from the mean than the value being considered, either in the plus or minus direction. For a value that is two standard deviations from the mean, the percentage drops to about 4.5%; and for a value that is three standard deviations from the mean, the percentage drops further to about 0.26% to be further from the mean (fewer than 3 in 1,000).

How Outliers Are Defined

There is no universally accepted definition for outliers.[2] A common definition is a value three or more standard deviations from the mean, but another definition might be appropriate for very large datasets.

By definition, extreme data points are rare in a normal distribution, but they are not impossible — the tallest living person, Sultan Kösen, is 8' 3" tall. On the other hand, this is the most extreme data point in a population of 7 billion. Mr. Kösen is very tall, but he is only a true outlier if there is a significant gap between his height and the tapered end of the

normal distribution. As it turns out, 13 people (including one woman) have been recorded with an undisputed height of 8 feet tall or taller.[3]

Outliers can happen merely by chance, but they often indicate a measurement, recording or transcription error. For example, an outlying laboratory value might represent a quality control problem with the lab that is also distorting apparently normal values.[4] Or, it might represent a rare, unexpected and dangerous side effect caused by the study drug, such as rhabdomyolysis, the breakdown of muscle fibers sometimes seen with statin drugs. Either way, an investigation into causation is in order. The investigation should be conducted thoroughly and with an open mind, especially if the outlier represents a potentially dangerous result. There might even be multiple causes for the outlier.

Dealing with Outliers

When a new site starts enrolling, examine data from the first five or so subjects with a very discerning eye. Apparent outliers can occur quickly when entries are made in the wrong place or with the wrong unit of measure (e.g., pounds instead of kilograms). Addressing these outliers early by retraining site personnel can reduce error repetition and the need to correct errors in the future. Repeat this check when the personnel entering data at a site change. If outliers representing true and accurate values are expected in a clinical trial, then the protocol and statistical plan should define them unambiguously and specify how to handle them. For example, sometimes a measurement system is known to yield an occasional erroneous result due to sample contamination or overheating during shipment.

In most instances, the data capture system should check for outliers automatically so obvious mistakes can be corrected prior to entering the values in the database. Alternatively or in addition to this, the database can be checked automatically at a later time. Staying on top of outliers throughout a study reduces the time needed to clean the data for database lock and improves the chance of understanding and possibly correcting problems.

It may be legitimate to exclude outliers from the analysis when the atypical result can be definitively attributed to a measurement or other error and it is not possible to capture the correct value. However, when there is doubt about the cause, it is reasonable to run and report the analysis with and without the outlier(s). If outliers do not affect study results, then they can be left in the database as is (with a note of explanation). Statistical analyses with and without the outliers can determine whether such is the case.

The presence of outliers might indicate that the data are not normally distributed. In this case, non-parametric tests can be used. When non-normally distributed data are expected in a study, the analysis plan should define a method for addressing them prior to beginning enrollment.

When outliers representing true values are not anticipated, they might still occur, in which case it may be necessary to address them with a post-hoc deviation from the analysis plan. This deviation should be explained and presented along with the original analysis that was initially planned.

When unanticipated outliers are suspected, raise an alert immediately.[5] Next, assemble a team that includes a biostatistician, a medical reviewer, and perhaps an expert in the measurement system that yielded the suspected outlier. Without delay, the team should assess the outlier, especially if it presents a potential risk to subjects. The team might call for actions like sending a query to the investigator, repeating the measure, or taking another sample; in no instance should outliers be ignored. The team should document how to identify and handle similar outliers that occur subsequently so they can be handled without reconvening the team.

References

1. Barnett B, Lewis V. Outliers in statistical data. 1994, John Wiley and Sons: West Sussex, England.
2. Penny K, Jolliffe I. Journal of the Royal Statistical Society: Series D (The Statistician). September 2001: Volume 50, Issue 3, pages 295–307, September 2001.
3. Wikipedia, http://en.wikipedia.org/wiki/List_of_tallest_people
4. Witte D, Van Ness S, Angstadt D, Pennell B. Errors, mistakes, blunders, outliers, or unacceptable results: how many? Clinical Chemistry. August 1997 vol. 43 no. 8 1352-1356.
5. Ingelfinger J, Mosteller F, Thibodeau L. Ware J. Biostatistics in clinical medicine. MacMillan Publishing Co., Inc: New York, 1987.

Author

Carleton Southworth, AB MS RAC, is Director, Biostatistics at American Research Partners. Contact him at 1.574.367.8076 or csouthworth@americanresearchpartners.com.