Outliers

Shengping Yang PhD, Gilbert Berdine MD

I have recently completed data collection for two of my research projects. Now, I am creating some histograms to visualize the distributions of important variables. To my surprise, there seems to be some extreme observations. For example, in one project, one ICU patient has a BMI value of 3.1; in another project, two patients who experienced stroke have CSF protein levels above 1,500 mg/dL. Based on my experience, these values seem to be unusual. Should I call them outliers? If so, how should I deal with such outliers?

It is always a good practice to "know" your data before performing data analysis. Visualizing your data by creating descriptive plots provides an initial assessment of the data distribution as well as possible outliers. However, there are controversies regarding how to identify outliers and how to handle outliers in data analysis. Therefore, before looking into these specific cases, let's start with the definition of an outlier.

1. What is an outlier?

The commonly used definition of an outlier is "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" (Hawkins, 1980). There are also several similar definitions, such as "an observation in a data set which appears to be inconsistent with the remainder of that set of data" (Johnson, 1992), and "an observation that appears to deviate markedly from other members of the sample in which it occurs" (Barnett and Lewis, 1994).

2. Why do we care about outliers?

Outliers can cause serious problems in data analysis. First, most parametric analysis methods require valid data distribution assumptions, and the existence of outliers very often results in the violation of such assumptions. Second, outliers increase data variation and thus reduce the power of statistical tests, which is not desirable. Third, if outliers reflect a mixture of observations from a population other than the target population, analyzing data with such outliers produces biased estimations of the target population parameters. In addition, outliers might be erroneous observations, e.g., errors occurred during data input. Therefore, to achieve meaningful and unbiased data analysis, outliers have to be appropriately identified and handled.

On other occasions, outliers might, however, be the observations of interest. For example, an abnormal diagnosis test result might indicate a potential health problem, and thus patients with abnormal results are a possible focus of inquiry.

3. How do we find outliers?

There are a number of criteria for identifying outliers, including visual inspection and analytic procedures.

3.1 Visual inspection

Using a boxplot to indicate outliers was initially introduced by Tukey in 1977. Specifically, any value below Q1 (the lower quartile)-1.5×IQR (Inter-Quartile Range, a measure of statistical dispersion being equal to the difference between the upper and lower quartiles) or above Q3(the upper quartile)+1.5×IQR is considered to be an outlier. For example, in Figure

Corresponding author: Shengping Yang PhD
Contact Information: Shengping.Yang@ttuhsc.edu
DOI: 10.12746/swrccc2016.0413.178

1A, the observation with the largest value is greater than Q3+1.5×IQR and thus considered to be an outlier.

Although straightforward, the above univariate approach might not be able to detect outliers in multivariate settings. For example, the observation in red in Figure 1B cannot be detected as an outlier based on either variable x or y alone because neither the x nor y value of that observation is extreme. In this case, a scatter plot works quite well, and the observation in red clearly deviates from all other observations in a scatter plot.

3.2 Analytic procedures

There are a number of statistical methods for identifying outliers. They can be generally categorized into parametric methods and model-free methods. We will focus primarily on discussing parametric methods in this article.

The general idea of the parameter methods is to compute the parameters assuming all data points come from a certain kind of statistical distribution, e.g., a normal distribution. The observations that have a low probability of coming from such a distribution are considered to be outliers.

3.2.1 Univariate methods

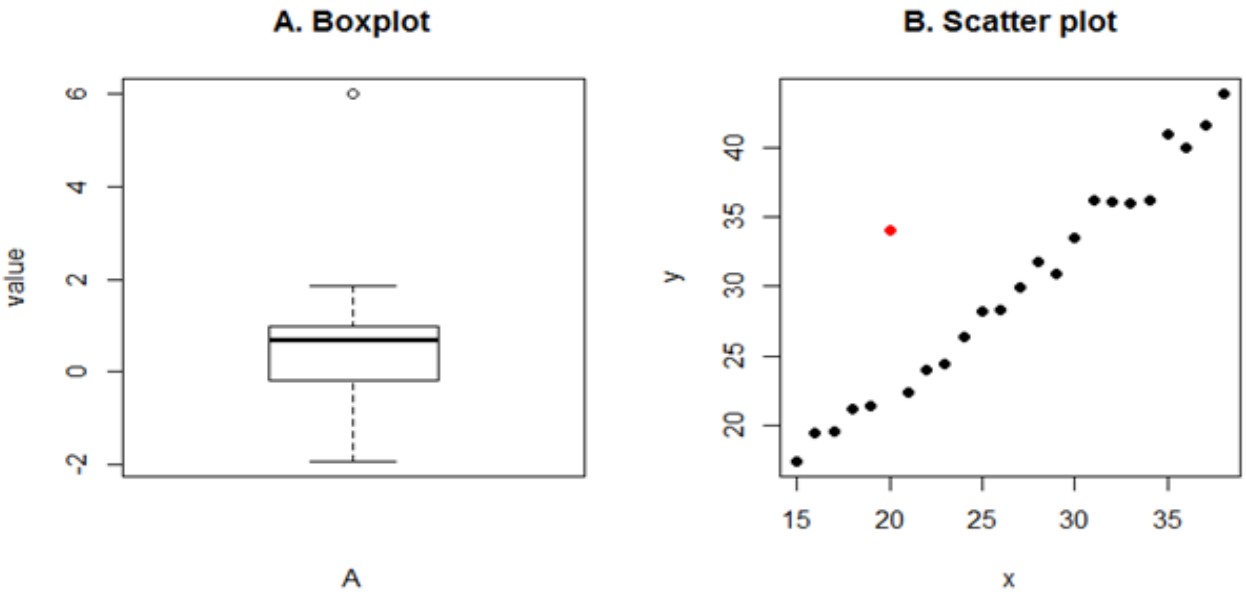
Grubbs's test is one of the parameter methods for detecting outliers in single samples. The test statistic is defined as:

$$G = \frac{\max_{i = 1,..., N} | Y_i - \bar{Y} |}{S} ,$$

where $\bar{Y}$ and S denote the sample mean and standard deviation, respectively.

In general, Grubbs' test detects one outlier at a time. The whole process is iterative, and it stops once no more outliers can be detected. The *grubbs.test* function in the R *outliers* package can be used for performing such a test. Note that Grubbs' test would not perform well for samples with ≤ 6 observations.

Figure 1 Graphical examination of outliers



Other univariate methods for outlier detection include the chi-squared test and the generalized extreme studentized deviate test (Rosner, 1983).

3.2.2 Multivariate methods

In regression analysis, abnormal observations in the response variable are called outliers, and abnormal observations in the predictors are called leverage points. Although a bad leverage point can substantially distort the effect estimate of the regression slope(s), here we will focus mainly on issues associated with detecting outliers.

Both standardized and studentized deleted residuals can be used for detecting potential outliers. Under model assumptions, the standardized residuals have a standard normal distribution. If an observation has an unusually large standardized residual, e.g., greater than 3, then it might potentially be an outlier. Similarly, the studentized deleted residuals follow a t distribution with n-p-1 degrees of freedom (Neter *et al.*, 1996). Observations with studentized deleted residuals greater than 3 or less than -3 should be considered as potential outliers.

Other multivariate methods for outlier detection use distance measures to indicate whether an observation is far away from the center of the data distribution. For example, observations with a large *Mahalanobis* distance are considered as outliers. Cook's D, which combines information on the residual and leverage, is another test statistic for detecting multivariate outliers. Observations with high Cook's D values (the conventional cut-off point is 4/n) are more likely to be problematic.

In both the univariate and multivariate methods discussed above, mean and variance-covariance were used for detecting outliers. However, mean and variance themselves are sensitive to outliers, and one "bad" observation might completely skew the mean and substantially inflate the variance, thus using robust estimates of the distribution parameters can improve the performance of outlier detection. However, those methods are beyond the scope of this discussion.

4. How do we deal with outliers?

There is a considerable amount of controversy regarding how to handle outliers, and different recommendations are made for outliers of different nature.

4.1 Illegitimate outliers

If it can be determined that an outlier is likely to be caused by a known error, then the best way to handle such an outlier is to remove or even correct it, if possible. For example, the ICU patient with a BMI of 3.1 seems to be an obvious error. The general recommendation is to review the patient database, check if weight and height of that specific patient were measured and recorded correctly, and if possible, recalculate the BMI. Note that if it is not feasible to correct the erroneous value, then the recommendation is to remove it.

4.2 Legitimate outliers/outliers with unknown causes

There is no general answer as to how to handle legitimate outliers or outliers with unknown causes. Some researchers suggest removing all detectable outliers so that the parameter estimates are more relevant to the target population; others suggest keeping outliers to avoid possible data manipulation since the outliers are legitimate. In fact, whether to keep or remove such outliers should be determined in a case by case manner. A danger exists that "outliers" are removed in a biased way in order to make the data fit the hypothesis.

4.2.1 Keep outliers

Many researchers recommend keeping all the outliers in data analysis. For example, sometimes outliers might be just valid extreme observations due to random variability and reflect the inherent property of random sampling. If this is the case, then they should be kept and treated in the same manner as all other observations in data analysis. For example, although

the CSF protein level in a normal person is usually substantially below 100mg/dL, it is possible that patients with disrupted CSF protein reabsorption have CSF protein levels as high as 3,500 mg/dL (Shah and Kelly, 1999). Therefore the two extreme CSF protein level values are very likely true. After confirming that they are not man-made errors, we should include them in the data analysis.

In situations in which outliers are associated with data skewness, certain transformations, e.g., log transformations, can mitigate the effect of outliers; meanwhile, the transformed data might also better meet the distribution assumption. For example, by including the two extreme values, the CSF protein levels have a skewed distribution, thus a log transformation would be recommended before any statistical testing is performed.

Other times, if outliers distort the upper and lower tails of the data distribution, then the data can be Winsorized. By setting all outliers to a specified percentile of the data, e.g., a 90% Winsorization means setting all data below 5th percentile to the 5th percentile, and all data above 95th percentile to the 95th percentile, the effect of outliers is mitigated, while the ranks of the observations are preserved. Analysis can be performed on the Winsorized data.

Additionally, robust methods can be used. “Robust” means less sensitive to outliers. For example, median is less affected by outliers compared to mean, and thus is a robust statistic. In terms of regression analysis, there are many forms of robust regressions, such as Least Absolute Deviations regression (absolute values of the residuals are less sensitive to outliers than the square of residuals), Huber regression, Schweppe regression, and Least Median of Square regressions, etc.

4.2.2 Remove outliers

Removing outliers even when they are legitimate is an entirely different opinion. Researchers who support outlier removal argue that meaningful statistical analysis should focus on modeling the majority of a population, so does the data interpretation. Meanwhile, data with outliers removed very often have less variation and better meet the assumptions of data analysis.

However, researchers who disagree with such an opinion argue that removing data points on the basis of statistical analysis without an assignable cause is not a good justification. In addition, removing “detectable” outliers introduces new problems. For example, due to masking (when a group of true outliers exist, they can pull the mean estimate toward them, thus, few or none of the true outliers appear to be extreme values, i.e., some of the true outliers were masked by other outliers) and swamping (a group of true outliers make one or more observations appear to be outliers) effects, some true outliers do not appear to be outliers, and thus removing one outlier introduces another one. Therefore, removing outliers might not be a straightforward solution.

4.2.3 Sensitivity analysis

Since there are pros and cons for both keeping and removing outliers, many times it is prudent to perform analyses with and without the suspected outliers to see if there is any difference. If there is no difference, then it does not matter whether the outliers are kept. Otherwise, more work might need to be done to investigate the causes of the outliers. Presentation of both sets of results may be the best choice.

5. Challenges in detecting and handling outliers

There are controversies about the definition of an outlier and the decision whether to remove or keep them. Visual inspection can provide an initial assessment of outliers. However, sometimes, the decision whether to keep or discard a data point is not clear cut. Other times, it might not be feasible to generate multi-dimensional plots, e. g., with dimensions greater than 3. For legitimate outliers or outliers with unknown causes, different opinions exist. Keeping them means having to deal with all the problems associated with outliers, and removing them might introduce new problems. Most of the outlier detection methods have

their limitations, and many would not work well if the number of observations is small. Robust methods are less sensitive to outliers; however, they usually suffer from reduced statistical power and are computational intensive. Novel statistical methods are needed for overcoming these limitations.

In general, outlier detection and handling is not solely a statistical issue. Instead, outliers should be addressed in a holistic way by considering the research objective, the logistic feasibility of detection and removal, and the statistical validity of the data as a whole.

Author affiliation : Shengping Yang is in the Department of Pathology at Texas Tech University Health Sciences Center in Lubbock, TX. Gilbert Berdine is in the Department of Internal Medicine at TTUHSC in Lubbock, TX.
Submitted: 12/17/2015
Accepted: 1/10/2016
Published electronically: 1/15/2016

References

1. Barnett V, Lewis T. Outliers in statistical data (3rd ed). 1994. New York: Wiley.
2. Grubbs FE. Sample criteria for testing outlying observations. Annals of Mathematical Statistics 1950; 21 (1): 27–58. doi:10.1214/aoms/1177729885
3. Hawkins DM. Identification of outliers. 1980. London: Chapman and Hall.
4. Johnson R. Applied Multivariate Statistical Analysis. 1992. Prentice Hall.
5. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied Linear Statistical Models (4th ed.). 1996. WCB McGraw-Hill.
6. Osborne JW, Overbay A. The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research, and Evaluation, 2004. p 9.
7. Rosner B. Percentage points for a generalized ESD many-outlier procedure, Technometrics, 1983. 25(2), 165-172.
8. Shah SM, Kelly KM. Emergency Neurology: Principles and Practice. 1999. Cambridge University Press.
9. Tukey JW. Exploratory data analysis. 1977. Addison-Wesley.