Outliers: An Evaluation of Methodologies

Dhiren Ghosh * Andrew Vogt †

Abstract

Given a random sample $x_1, x_2, \ldots, x_n$ from a population, we examine circumstances that lead to some of the values being deemed outliers and the methodologies proposed to analyze the data set in the presence of these outliers. We review methods of determining outliers and propose general principles for how to proceed. Often a mixture approach is appropriate: most observations seem to follow a pattern or to satisfy a model, while the outliers remain outside the pattern or model and require further investigation.

Key Words: bias, extreme value theory, Grubbs' test, heavy-tail phenomena, mixture models, Winsorization

1. Introduction

An outlier in a sample survey is an observation far away from most or all other observations. What do we do with it when we are trying to draw conclusions from the sample about the population?

There is of course the possibility that the outlier is an error. There are several kinds of errors. A measurement error is due to some mistake in the process of measuring - it may be a flaw in an instrument, an answer based on a misinterpretation of a question asked or a poorly worded question, or an ambiguous answer misinterpreted by the interviewer. A clerical error is an error that results when data are transcribed or copied either manually or by computer and a mistake in transcription takes place. Another kind of error is a sampling frame error: a unit not in the target population somehow is included in the sample. Numerous ways exist to detect such errors, but we will not consider them here. Instead we assume that they have been ruled out.

We are concerned with values that are genuine but extreme. Resnick (2007) points out that extreme values occur in network theory (file sizes, transmission rates and durations) and finance (return rates from risky investments, insurance claim sizes and frequencies).

Genuine outliers are typically treated in one of the following ways:

1) keep the outlier and treat it like any other data point; or

2) winsorize it (i.e., assign it lesser weight or modify its value so it is closer to the other sample values); or

3) eliminate it (drop it from the sample).

The danger of each of these methods is that they may produce poor estimates of parameters of interest. Methods 2) and 3) introduce statistical bias and may undervalue the outlier, while keeping it and treating it like the other points may overvalue it and cause the estimate to vary drastically from the true population value.

Below we investigate these methods, and make some proposals that we think are sensible about what an outlier really is and how to treat it.

*Synectics for Management Decisions, Inc., Arlington, Virginia 22209

†Department of Mathematics and Statistics, Georgetown University, Washington, D. C. 20057-1233

2. Winsorizing and Trimming

Treatments 2) and 3) above, winsorizing or eliminating, used to be standard ways of treating outliers. The desire for robust statistics and for measures insensitive to outliers was satisfied by dropping outliers or by modifying their values.

A common procedure has been to replace any data value above the ninety-fifth percentile of the sample data by the ninety-fifth percentile and any value below the fifth percentile by the fifth percentile. The assumption seems to be that the outlier does not look right and estimates will be improved if the outlier is made to look like other data. This suggests that the outlier value must be incorrect, an exaggeration of the truth (e.g., old persons sometimes overstate their age, especially in rural or underdeveloped regions). The value is replaced by a more plausible value. The new value is a compromise - it is not thrown out but it is watered down. The danger of bias is alleviated by retaining an attenuated version of the datum. Consider another example. Acreage devoted to a certain crop may make sense by itself but when the total acreage devoted to all crops is smaller than the reported total for one crop, something is amiss. It can also happen that the observation is entirely correct but is not part of the target population. This may be because the target population has not been defined as precisely as necessary. For example, land ownership by farmers may include several large parcels in dispute between erstwhile owners and the state.

However, these circumstances assume some kind of error has been made. In the current relatively advanced stage of statistical theory and practice, errors should be analyzed and specific methods are available to take them into account depending on the error type.

Equivalent to winsorizing is the use of various estimators for population parameters that assign lesser weight to outliers.

The most extreme way to lower the weight of an outlier is to trim it from the sample, i.e., eliminate it. If the outlier is a legitimate value, there is no justification for this and it is counter to the basic principle of random sampling. Of course, the statistician really seeks a representative sample, one that looks like the population in miniature. Under random sampling the proper way to achieve representativeness is to stratify the population by a relevant variable or variables for which the statistician has good prior information about relative stratum sizes. Stratification may lead to lesser weights (or greater weights) for outliers but must be based on reliable information.

3. Definitions of an Outlier

Underlying the question of how to treat outliers is the issue of whether a particular observation is an outlier. Morris Hansen many years ago (see, for example, p. 787 of Hansen et al. (1983)) proposed the rule of thumb that an outlier is any observation whose removal from the sample changes the estimate of a parameter of interest by 10 percent or more. For parameters with values near zero this is obviously suspect. It is also dependent on which parameter is of interest.

A natural method of determining outliers has long been available in the literature when the variable under study has a known mathematical distribution. For a symmetric distribution with mean μ it makes sense to compute:

$$P = P(|x_1 - \mu| < c, |x_2 - \mu| < c, ..., |x_n - \mu| < c) = P(|x - \mu| < c)^n.$$

where $x_1, x_2, ..., x_n$ is an independent sample from the distribution, and c is the distance of the farthest outlier in the actual sample of size n from μ. If the difference between P and 1 is very small, say 0.05, then there is only a five percent probability that one or more values would lie that far out or farther.

Consider the case where the distribution in question is the standard normal distribution. In that case we can look at several stand-ins for c. They are:

$$E_n = E(\max\{|Z_1|, |Z_2|, \ldots, |Z_n|\})$$

and

$$c = c_n(\alpha) \text{ where } P(|Z| < c) = (1-\alpha)^{1/n}.$$

		two-tailed		
n	E_n	$c_n(0.05)$	$c_n(0.01)$	$c_n(0.001)$
1	.80	1.96	2.58	3.29
2	1.13	2.24	2.81	3.48
3	1.32	2.39	2.93	3.59
4	1.48	2.49	3.02	3.66
5	1.55	2.57	3.09	3.72
6	1.67	2.63	3.14	3.76
7	1.74	2.69	3.19	3.80
10	1.87	2.80	3.29	3.89
15	2.05	2.93	3.40	3.99
20	2.17	3.02	3.48	4.06
30	2.32	3.14	3.59	4.15
50	2.52	3.29	3.72	4.26
100	2.75	3.48	3.89	4.42
200	2.97	3.66	4.06	4.56
500	3.25	3.89	4.26	4.75
1000	3.44	4.05	4.42	4.89

The normal distribution is very thin-tailed. As the sample size increases, plausible outliers get farther and farther out but their magnitude increases slowly. A table like this can be applied to any normal distribution provided we pass to z scores: $z = \frac{x-\mu}{\sigma}$.

This general method is associated historically with Benjamin Peirce (1809-1880) and his son Charles Sanders Peirce (1839-1914), William Chauvenet (1820-1870), and Frank Grubbs (1925?-2000). These individuals started with the assumption that the sample data are drawn from a normal population. If the mean and standard deviation of the population are known, then the z-score of the most extreme candidate outlier is computed and an estimate is made for the likelihood of the most extreme value being in the two tails associated with that z-score. If this is low, as in hypothesis testing, we declare the value an outlier. As in hypothesis testing, the decision about what is a low value is subjective.

Various subtleties have been addressed with this approach. One of them is that since the mean and standard deviation are usually estimated from the sample, T-distributions, which depend on the value of n, should be used rather than the standard normal distribution. Another is that the presence of multiple outliers may make it harder to reject any of them. At the NIST website (http://www.itl.nist.gov/

div898/handbook/eda/section3/eda35h1.htm) the latest version of Grubbs' test is described, along with methods to handle masking effects related to multiple outliers.

A similar analysis can be done with the standard exponential distribution e^{-x}, $x > 0$. In this case we look at:

$$E_n = E(\max\{X_1, X_2, \ldots, X_n\}),$$

and

$$c = c_n(\alpha) \text{ where } P(X < c) = (1-\alpha)^{1/n}.$$

The table we obtain is:

		one-tailed		
n	E_n	$c_n(0.05)$	$c_n(0.01)$	$c_n(0.001)$
1	1.00	3.00	4.61	6.91
2	1.48	3.68	5.30	7.60
3	1.85	4.08	5.70	8.01
4	2.07	4.36	5.99	8.29
5	2.28	4.58	6.21	8.52
6	2.46	4.77	6.39	8.70
7	2.58	4.92	6.55	8.85
10	2.93	5.28	6.90	9.21
15	3.36	5.68	7.31	9.62
20	3.62	5.97	7.60	9.90
30	3.98	6.37	8.00	10.31
50	4.45	6.88	8.51	10.82
100	5.15	7.58	9.21	11.51
200	5.87	8.27	9.90	12.21
500	6.82	9.18	10.81	13.12
1000	7.49	9.88	11.51	13.82

The exponential has a thin tail and plausible outlier values grow slowly with sample size. An exponential with mean μ can be normalized to this table by the substitution $z = \frac{x}{\mu}$. It is possible to refine this methodology when the mean μ must be estimated from the sample.

Another family of distributions commonly used for nonnegative variables is the gamma distribution. This distribution has two positive parameters, a shape parameter α and a scale parameter λ. When $0 \leq \alpha \leq 1$, the mode (most frequent value) is zero, and when $1 < \alpha$, the mode is positive. The mean of a gamma variable is $\frac{\alpha}{\lambda}$ and its standard deviation is $\frac{\sqrt{\alpha}}{\lambda}$. When $\alpha = 1$ we get an exponential distribution.

In the table below we take $\lambda = 1$ and consider two values of α, namely, 5 and 10.

	$\alpha = 5$		$\alpha = 10$	
n	E_n	$c_n(0.01)$	E_n	$c_n(0.01)$
1	5.00	11.60	10.00	18.78
2	6.17	12.59	11.78	20.00
3	6.96	13.16	12.71	20.68
4	7.41	13.55	13.44	21.16
5	7.82	13.86	13.92	21.53
6	8.16	14.10	14.26	21.83
7	8.38	14.31	14.62	22.08
10	8.95	14.79	15.35	22.65
15	9.58	15.33	16.19	23.29
20	10.04	15.70	16.78	23.74
30	10.62	16.23	17.57	24.37
50	11.37	16.89	18.50	25.15
100	12.36	17.78	19.63	26.19
200	13.34	18.65	20.89	27.21
500	14.63	19.79	22.46	28.53
1000	15.52	20.64	23.55	29.52

Gamma distributions are thin-tailed despite the flexibility offered by the additional parameter. To pass from a gamma variable X with arbitrary α and λ to one with λ normalized to 1, we take $Z = \lambda X$.

Other distributions can be treated similarly. An outlier for a given sample size is an extreme value (largest or smallest value) such that the a priori probability that the most extreme value is in the tail region determined by the actual value is less than five per cent or some similar small percentage.

4. The Real Questions

What has been said leaves several unanswered questions, the first of which has already been noted:

1) How small should the tail probability be before we declare a value an outlier?

2) Where do we get the distribution from?

3) Where do we get the family of distributions from?

4) What do we do with the outliers?

Answer to 1): As with the significance level in hypothesis testing there is no hard and fast answer. Outlier treatment is an art.

Answer to 2): If we have a family of distributions that plausibly describe most observations, we use prior knowledge, or the maximum likelihood method, or another related method to estimate the parameters of the distribution, including the outliers as input data. If an outlier is eliminated by the test, we reestimate the parameters without using the outlier.

Answer to 3): This requires judgment and experience and typically is based on graphical inspection of the entire data set. If most of the data seem to follow a particular family of distributions, this family is a candidate family.

Answer to 4): With only one or two outliers there is little that can be done but to preserve them for future study. However, if there are four or five outliers or more, we can employ extreme value theory and the theory of heavy tailed distributions.

One possibility in this case is to fit the outliers to a model of the following type:

$$P(X > x) = Cx^{-\alpha} \text{ for } x > x_0.$$

Here the constants C, α, and x_0 are estimated from the data, and $P(X > x_0)$ is taken to equal the fraction of outliers in the original sample. This leads typically to a mixture model where $P(a < X < b)$ is given by:

$$P(a < X < b/ \text{ basic distribution})(1-w)+P(a < X < b/X \text{ belongs to outlier set})w.$$

Here w can be taken to equal the fraction of data values determined to be outliers, and $P(a < X < b/X$ belongs to the outlier set) vanishes unless $b \geq x_0$ and equals $(x_0/x)^\alpha - (x_0/b)^\alpha$ otherwise, where $x = \max\{x_0, a\}$. The quantity x_0 can be taken to be the smallest member of the outlier set $\{x_1, \ldots, x_k\}$ and α is given by:

$$\frac{1}{\alpha} = \frac{1}{k}\Sigma_{i=1}^{k} \log(\frac{x_i}{x_0}).$$

The parameters of the population such as the mean and standard deviation are no longer items of interest in the mixture model. In a true mixture they distort our representation of the population. We are really interested in estimating parameters for the presumed subpopulations represented separately by the non-outliers and the outliers. It is quite appropriate to eliminate the outliers in estimating parameters for the first and more numerous subpopulation. In a preliminary way we can also study the outlier subpopulation. We are of course interested in determining the relative sizes of these two subpopulations. The quantity $\frac{w}{1-w}$, where w is as above, can be taken to be a rough estimate of the relative size.

REFERENCES

Hansen, M., Madow, W., and Tepping, B. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *J. Amer. Stat. Assoc.*, 78, 776-793.

Hawkins, D. M. (1980), "Identification of outliers," Chapman and Hall, London.

Resnick, S. (2007), "Heavy-Tail Phenomena: Probability and Statistical Modeling, " Springer, New York.

Sarhan, A., and Greenberg, B., ed's. (1962), "Contributions to Order Statistics," John Wiley & Sons, Inc., New York.