A COMPARISON OF OUTLIER LABELING CRITERIA IN UNIVARIATE MEASUREMENTS

Menus Nkurunziza[1] and Lea Vermeire[2]
[1]Université du Burundi, Burundi
[2]KU Leuven Kulak, Belgium
menus.nkurunziza@ub.edu.bi, lea.vermeire@kuleuven-kulak.be

For the detection of potential outliers in univariate measurements, undergraduate statistics courses often refer to the boxplot. In the workfield, various other sector-linked criteria for outliers are also popular, e.g. Chauvenet's criterion in engineering. We compare statistical properties of five current criteria – the 3-sigma rule, the Z-score, Chauvenet's criterion, the M-score or median criterion, and the boxplot or Tukey's criterion. In particular, in case of a normal population, a joint structure of the five criteria is detecte,d and large sample asymptotic properties of their non-outlier intervals are derived. Pointing at these results should help students to match the statistics course and the lab practice during their education or in their future professional environment. Next, for mathematical statistics students, proving these results may be an instructive activity.

INTRODUCTION

The detection and treatment of outliers in data is of great concern, as reflected in the vast statistical literature on the subject. An outlier in univariate data is in concept a data point that is extreme with respect to the mass of the data and that can have a big influence on the results of the analysis. More precisely, an outlier in a sample is an extreme value that is significantly too extreme, e.g. the maximum is an outlier if it is statistically too large for the distribution of the maximum under the population model (Barnett & Lewis, 1996, Saporta, 2011). A potential outlier is an extreme data point that the researcher labels as unlikely on view or by some descriptive criterion. To label potential outliers in univariate measurements, undergraduate statistics courses often refer to the boxplot, while in the workfield various other criteria remain also popular, such as the 3-sigma criterion and the Z-score in chemometrics, Chauvenet's criterion in engineering, the M-score or median criterion to overcome some shortcomings of the Z-score. Such a criterion is used as a quick tool to detect data that should be looked at carefully for erroneous registration or for being a statistical outlier. But do these criteria provide the same results? This paper investigates how different these criteria behave, in particular in large samples or when the parent population of the data has a normal distribution, which is often an acceptable model for measurements.

These criteria are useful tools for outlier labeling, but they do not exempt the researcher from doing a statistical test for outlier detection, like Dixon's Q-test or Grubbs' ESD-test for one outlier, or Rosner's generalized ESD-test for multiple outliers (e.g. Rosner, 2011).

Notations

Consider an independent and identically distributed (i.i.d.) sample $\mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{n}$ from a model population X. The following sample statistics will be used:

- the sample mean, the sample variances and the sample standard deviations : $\bar{X} = \sum_i X_i/n$, $S^2 = \sum_i (X_i - \bar{X})^2/(n-1)$, $\tilde{S}^2 = \sum_i (X_i - \bar{X})^2/n$, S, $\tilde{S}$;
- the order statistics: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$;
- the extremes: $X_{\min} = \min\{X_1, \cdots, X_n\} = X_{(1)}$, $X_{\max} = \max\{X_1, \cdots, X_n\} = X_{(n)}$;
- the p-quantile ξ_p $(0 < p < 1)$: $\xi_p = X_{(np)}$ if np is an integer, $\xi_p = X_{(\lfloor np \rfloor + 1)}$ if np is not integer; $\lfloor a \rfloor$ denotes the largest integer which is less than or equal to the real number a;
- the median, the quartiles, the interquartile range and the median absolute deviation: $\tilde{X} = \mathrm{med}\{X_1, \cdots, X_n\}$, $Q_1 = \xi_{0.25}$, $Q_3 = \xi_{0.75}$, IQR= $Q_3 - Q_1$, MAD $= \mathrm{med}\{|X_i - \tilde{X}| : i = 1, \cdots, n\}$
- the most extreme point $X_\star$: $X_\star = X_i$ if $|X_i - \bar{X}|$ is maximal among all n data, and thus $X_\star \in \{X_{\min}, X_{\max}\}$;
- the reduced statistics, e.g. $\bar{X}_{n-1}$ is typically the mean computed on the $n-1$ data under exclusion of the most extreme value, similarly for S^2_{n-1} etc. In that context, $\bar{X} = \bar{X}_n$.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

FIVE CRITERIA FOR OUTLIER LABELING

Definitions. Five criteria for potential outliers are considered: 3-sigma, Z-score or normal criterion, Chauvenet's criterion, M-score or median criterion, boxplot or Tukey's criterion with coefficient c (c = 1.5 reports mild outliers, c = 3 reports extreme outliers). Their historical definition is given in table 1, first column.

Table 1. Five criteria for a potential outlier – the historical definition, the non-outlier interval, its large sample expression, and the latter for a normal parent distribution

Outlier criterion for X_i	Non-outlier interval	Limit interval ($n \to \infty$) under mild conditions ([6])	Limit interval in case of a normal population $X \sim \mathrm{N}(\mu, \sigma^2)$
3-sigma ([1]) $\lvert X_i - \bar{X}_{n-1}\rvert > 3S_{n-1}$, i.e. $\lvert Z_{i,n-1}\rvert > 3$	$\bar{X}_{n-1} \pm 3S_{n-1}$	No general result	$\mu \pm 3\sigma$
Z-score ([2]) $\lvert Z_i\rvert > 3$	$\bar{X} \pm 3S$	$\mu \pm 3\sigma$	$\mu \pm 3\sigma$
Chauvenet ([3]) $n\mathrm{P}\{X \in [\bar{X} \pm \lvert X_i - \bar{X}\rvert]\} > 1/2$ with $X \sim N(\bar{X}, S^2)$, i.e. $\lvert Z_i\rvert > z_{1-1/4n}$	$\bar{X} \pm z_{1-1/4n}S$	No general result	$\mathbb{R}$
M-score ([4]) $\lvert M_i\rvert > 3.5$ with $M_i = 0.6745\ \frac{X_i - \tilde{X}}{\mathrm{MAD}}$	$\tilde{X} \pm 3.5\frac{\mathrm{MAD}}{q}$	$X_{\mathrm{med}} \pm 3.5\frac{\mathrm{MAD}_X}{q}$	$\mu \pm 3.5\sigma$
Boxplot(c) ([5]) $X_i < Q_1 - c\mathrm{IQR}$ or $X_i > Q_3 + c\mathrm{IQR}$	$[Q_1 - c\mathrm{IQR}\ ,\ Q_3 + c\mathrm{IQR}]$ or $\frac{Q_1+Q_3}{2} \pm k\frac{\mathrm{IQR}}{2q}$ where $k = (1+2c)q$	$[Q_{1,X} - c\mathrm{IQR_X}\ ,\ Q_{3,X} + c\mathrm{IQR_X}]$ or $\frac{Q_{1,X}+Q_{3,X}}{2} \pm k\frac{\mathrm{IQR_X}}{2q}$	$\mu \pm \{(1+2c)q\}\sigma$
c = 1.5			$\mu \pm 2.7\sigma$
c = 3			$\mu \pm 4.7\sigma$

Notes

([1]) : $Z_{i,n-1}, \bar{X}_{n-1}, S_{n-1}$ are the reduced statistics, i.e. the statistics as in ([2]), but computed on $n-1$ data, after exclusion of the suspect point.

([2]) : $Z_i = (X_i - \bar{X})/S$, $\bar{X} = \sum_i X_i/n$, $S^2 = \sum_i (X_i - \bar{X})^2/(n-1)$.

([3]) : $z_{1-1/4n}$ is the (1-1/4n)-quantile, with right tail probability 1/4n, for the standard normal distribution N(0,1).

([4]) : $\tilde{X} = \mathrm{med}\{X_1, \cdots, X_n\}$, $\mathrm{MAD} = \mathrm{med}\{|X_i - \tilde{X}| : i = 1, \cdots, n\}$, $q = z_{0.75} = Q_{3,Z} \approx 0.6745$ is the right tail quartile of the standard normal distribution N(0,1). $X_{\mathrm{med}} = \mathrm{med}(X)$ and $\mathrm{MAD}_X = \mathrm{med}|X - X_{\mathrm{med}}|$ are the population median and the population MAD.

([5]) : Q_1 , Q_3 are the sample quartiles leaving respectively 25% in the left tail and 25% in the right tail, IQR= $Q_3 - Q_1$. Then $Q_{1,X}$, $Q_{3,X}$ and IQR_X are the corresponding population quantities.

([6]) : Conditions for the Z-score: μ = E(X), σ^2 = V(X) < ∞; for the M-score: X_{med} and MAD_X are unique, and the cumulative distribution function F_X is not flat in a right neighborhood of X_{med}; for the boxplot: the population quartiles are unique, and F_X is not flat in a right neighborhood of each.

The 3-sigma criterion is exclusive, as it excludes the suspect point from the computations on the presumed population data, while the other four criteria are inclusive. The Z-score is the inclusive form of the 3-sigma criterion; as such it follows the paradigm of a statistical test, giving the data the credit of no suspect value unless their test statistic imposes the opposite. Chauvenet's criterion (1863) is more severe as to outliers than the Z-score as long as the sample size $n \leq 185$. Iglewicz & Hoaglin (1993) noticed that the Z-score can be misleading as its maximum is

$(n-1)/\sqrt{n}$ (Seo, 2006, provides a proof), and thus up to $n \leq 10$ the Z-score will never find an outlier; for that reason they introduced the M-score. The classical boxplot, as presented in many textbooks (e.g., Ennos, 2007; Moore et al, 2012) and statistical software packages, is the one with coefficient c = 1.5 for mild outliers and c = 3 for extreme outliers. The boxplot is the only one of the five criteria that includes in its outlier decision making the possible asymmetry of the data distribution; moreover it offers a visual summary of the essential characteristics of the data distribution.

The second column in the table rewrites the definition from the first column in the form of the non-outlier interval. It is obtained by straightforward computations.

Example. The simulated data 2.46 1.01 0.17 2.56 1.55 -0.12 0.91 1.99 1.49 5.02 (n=10), were obtained as a normal N(1,1) sample contaminated by the extreme point from an equal size exponential sample under the same median. The sample statistics $\bar{X} = 1.704$, S = 1.462, $\tilde{X} = 1.520$, $Q_1 = 0.91$, $Q_3 = 2.46$, IQR = 1.55, $X_\star = 5.02 = X_{\max}$, $X_{\min} = -0.12$, MAD = 0.775, $z_{1-1/4n} = z_{0.975} = 1.960$, $\bar{X}_{n-1} = 1.336$, $S_{n-1} = 0.937$, lead to the non-outlier intervals: 3-sigma [-1.48;4.15], Z [-2.68;6.09], Chauvenet [-1.16;4.57], M [-2.50;5.54], boxplot(1.5) [-1.42;4.79] and boxplot(3) [-3.74;7.11]. Hence the most extreme point 5.02 is labeled as an outlier by the criteria 3-sigma, Chauvenet and boxplot(1.5), and not by the other three criteria.

LARGE SAMPLE STRONG CONVERGENCE

Theorem. For each of the five criteria, the non-outlier interval converges almost surely to the interval given in table 1, columns 3 and 4, respectively for a general parent distribution and for the normal parent ($n \to \infty$). Hereby an interval is treated as the vector of its border points.

Interpretation. The results for a general parent population lead to the following interpretation, under the mild conditions given in the table. The non-outlier interval of the Z-score, the M-score and the boxplot converges almost surely to the corresponding population interval. Absence of a similar property for the 3-sigma criterion and Chauvenet's criterion may be caused by deviation of normality of the parent, e.g. due to asymmetry.

The case when the parent population is normal, $X \sim \mathrm{N}(\mu, \sigma^2)$, allows the following three interpretations. The five non-outlier intervals have a joint form $\hat{\mu} \pm k\hat{\sigma}$, with strong consistent estimators $\hat{\mu} \to_{as} \mu$, $\hat{\sigma} \to_{as} \sigma$, and a criterion specific coefficient k, which is a constant, except for Chauvenet's criterion where it depends on the sample size n. The two criteria 3-sigma and Z-score are asymptotically equivalent; thus in large samples they provide almost always the same outliers. The large sample limit of the non-outlier interval is $\mu \pm k\sigma$, with growing k = 2.7, 3, 3.5, 4.7, respectively for boxplot(1.5), 3-sigma and Z-score, M-score, boxplot(3); for Chauvenet's criterion the limit is the whole real line. Thus in large samples, Chauvenet's criterion will almost never detect outliers, while the reported outliers tend to be more severe as they come from the criteria boxplot(1.5), 3-sigma or Z-score, M-score, boxplot(3), in this order.

OUTLINE OF PROOF

Most convergence results follow directly from: the strong consistency (almost sure convergence) of moment statistics such as $\bar{X} \to_{as} \mu$, $S \to_{as} \sigma$, the strong consistency of quantiles such as $Q_1 \to_{as} Q_{1,X}$ etc. (Serfling, 2002, p. 75), the strong consistency of MAD as $\mathrm{MAD} \to_{as} \mathrm{MAD}_X$ (with mild conditions in Serfling & Mazumder, 2009, building further on Hall & Welsh, 1985), and Slutsky's theorems on convergence preservation under transformations (e.g. Ferguson, 1996, p. 39-42). Less obvious are the proofs of the asymptotic properties in the normal case, for 3-sigma and Chauvenet's criterion.

For 3-sigma and a normal parent $X \sim \mathrm{N}(\mu,\sigma^2)$, we here consider only the case where $X_{\max} = X_{(n)}$ is the suspect value, and thus compute the reduced statistics under exclusion of $X_{\max}$. Careful calculations provide the following expressions for the reduced statistics as functions of the full statistics and $X_{\max}$:

$$\bar{X}_{n-1} = \frac{n}{n-1}\left(\bar{X} - \frac{X_{\max}}{n}\right), \quad \tilde{S}^2_{n-1} = \frac{n}{n-1}\left[\tilde{S}^2 - \frac{n}{n-1}\left(\frac{X_{\max}}{\sqrt{n}} - \frac{\bar{X}}{\sqrt{n}}\right)^2\right], \quad S^2_{n-1} = \frac{n-1}{n-2}\tilde{S}^2_{n-1}.$$

In the right sides we know that $\bar{X} \to_{as} \mu$, $\tilde{S}^2 \to_{as} \sigma^2$, $n/(n-1) \to 1$; then by a Slutsky theorem $\bar{X}/\sqrt{n} \to_{as} 0$. For a normal distribution, $X_{\max}/(2\log n)^{1/2} \to_{as} \sigma$ (Serfling, 2002, p. 91); then as a corollary $X_{\max}/n^k \to_{as} 0$ for all $k > 0$. Thus, by Slutsky's theorems, $\bar{X}_{n-1} \to_{as} \mu$, $S_{n-1} \to_{as} \sigma$. Hence, again by a Slutsky theorem, the vector $(\bar{X}_{n-1} - 3S_{n-1}, \bar{X}_{n-1} + 3S_{n-1}) \to_{as} (\mu - 3\sigma, \mu + 3\sigma)$.

For Chauvenet's criterion with a normal parent, it is sufficient to show that $\bar{X} + z_{1-1/4n}S \to_{as} \infty$, that is $\mathrm{P}(\bar{X} + z_{1-1/4n}S > A) \to 1$ for any $A > 0$. Appropriate standardisation in the left side provides $\mathrm{P}(\bar{X} + z_{1-1/4n}S > A) = \mathrm{P}\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \frac{\sqrt{2}}{z_{1-1/4n}} + \frac{S-\sigma}{\sigma/\sqrt{2n}} > \left(\frac{A-\mu}{z_{1-1/4n}\sigma} - 1\right)\sqrt{2n}\right]$. The latter is of the form $\mathrm{P}(Z_n > a_n)$, where in Z_n the asymptotic normality of the sample statistics $\bar{X}$ and S under a normal parent can be applied. Thus $Z_n \to_{\mathrm{d}} Z \sim N(0,1)$, $a_n \to -\infty$. It can be shown that then $\mathrm{P}(Z_n > a_n) \to 1$, which finishes the proof.

CONCLUSION

Five current criteria for outlier labeling, 3-sigma, Z-score, Chauvenet's criterion, M-score and boxplot, have been compared through their non-outlier intervals.

For a general population, under mild conditions, the interval from the Z-score, M-score or boxplot converges strongly to the corresponding population interval.

If the parent population is normal, then the non-outlier intervals of all five criteria have the same structure of a *k*-sigma interval, $\hat{\mu} \pm k\hat{\sigma}$, where $\hat{\mu}$ and $\hat{\sigma}$ are strong consistent estimators of μ and σ, and k is a criterion specific constant, except for Chauvenet's criterion where k depends on the sample size n. In the normal case, the two criteria 3-sigma and Z-score, which are the exclusive and the inclusive form of a 3σ- interval, are large sample equivalent. Chauvenet's criterion is large sample zero outlier-detecting, as its limit interval is the whole real line. The other criteria, in the large sample case, can be ordered for reporting from mild outliers on to only more extreme outliers, in the order boxplot(1.5), Z-score and 3-sigma, M-score and boxplot(3).

From an educational point the study may be of interest to two groups of students. The results will help science students to match the boxplot from their undergraduate course in statistics with criteria in the lab in their further training or their later professional environment. Students in mathematical statistics may experience a challenging activity in the proofs, as these proofs are achievable through a few powerful results from the literature and multiple applications of the statistical limit theorems from their advanced course.

REFERENCES

Barnett, V., & Lewis, T. (1996). *Outliers in statistical data* (3rd ed.). Chichester: Wiley.

Chauvenet, W. (1863). *A manual of spherical data and practical astronomy*, vol. II (5th ed., 1960). New York: Dover.

Ennos, R. (2007). *Statistical and data handling in biology* (2nd ed.). Harlow: Pearson.

Ferguson, T. S. (1996). *A course in large sample theory*. London: Chapman & Hall.

Hall, P., & Welsh, A.H. (1985). Limit theorems for the median deviation. *Annals of the Institute of Statistical Mathematics, 37,* 27-36.

Iglewicz, B., & Hoaglin, D. (1993). *How to detect and handle outliers*. Milwaukee: The ASQC Basic References in Quality Control: Statistical Techniques, 10.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2012). *Introduction to the practice of statistics* (7th ed.). New York: Freeman.

Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Boston: Brooks/Cole.

Saporta, G. (2011). *Probabilités, analyse des données et statistique* (3rd rev. ed.). Paris: Technip.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* (master's thesis). Pittsburgh: University of Pittsburgh, supervisor G.M. Marsh.

Serfling, R. (2002). *Approximation theorems of mathematical statistics.* New York: Wiley.

Serfling, R., & Mazumder, S. (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statistics and Probability Letters, 79,* 1767-1773.