Some Procedures For Identifying Hotspots

Prof. R. G. Gurao*
Head, Department of Statistics
Brihan Maharashtra College of Commerce
Pune 411004, India.
(rggurao.bmcc@gmail.com)
Dr. Sharad Gore
Department of Statistics
University of Pune
Pune 411007, India.
(sharaddgore@gmail.com)

Abstract

The word **hotspot** was coined by British biologist Norman Myers in 1988. After general acceptance of the term hotspot it became common to use the term in different domains of investigation. Hotspot means something unusual or anomalous. Therefore, it is needed in monitoring a situation or a process, and can be used for early warning. In the statistical literature, hotspots are variously described as extreme observations, outliers, anomalous observations, and so on. Extreme values are excessively large or excessively small observations. Outliers are observations that are sufficiently away from the majority of observations to create a doubt about their belonging to the same population as all the other observations. In the present paper, some previous procedure of identifying outliers are discussed and some new procedures are suggested.

1 Introduction

Hotspots are super sensitive areas of problem cases. The term *hotspot* was coined by environmental scientists to describe areas of environmental issues such as biodiversity, deforestation, degeneration of soil quality, depleting ground water levels, and so on. Ecologists used this tern to refer to endangered natural habitats of exotic species of plants, animals, birds, reptiles, or insects. Presence of a hotspot leads to a significant disturbance in the routine activities and has the potential to cause major disasters. Hotspots remain point of attraction for researchers because of their utility in crime analysis, geo-informatics in climatology, early warning systems in case of epidemics and several other reasons. Identifying or detecting a hotspot is therefore of vital importance. In food and agricultural sector, the concept of a hotspot is used to review or identify a situation that, if left unattended, could create an unmanageable situation. In the statistical literature, hotspots are variously described extreme observations, outliers, anomalous observations, and so on. In the present paper, treating hotspot as an extreme value or an outlier, an effort are made to suggest some new procedures to identify it.

Hotspots do not have universal definitions and are usually identified by contrasting their behavior in the comparison with their neighbourhoods. In other words, hotspot are identified by analyzing field data rather than by understanding nature of the domain of study. It therefore becomes a statistical challenge to detect and locate hotspots so that appropriate remediation plan can be drawn. In the modern age of data flood, research generates and uses lots of data. Data must be inspected for strange elements and corrected if necessary before it is used. Detecting and treating outliers is one of the first activities to be performed on data. The literature on outliers is extensive and has contributions from other branches of science in addition to statistical methods. The literature survey may begin with the books by Hawkins (1980), Bartnett and Lewis(1994), and the chapters on outliers by Ben-Gal (2005). It may be interesting to note the variety of definitions of the term *outliers* in the statistical literature.

2 Some Definitions of Hotspot

- Grubbs (1969): An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it

occurs.

- Hawkins(1980): An outlier is the observation that deviates so much from other observations as to arose suspicion that it was generated by different mechanism.
- Johnson (1992): An outlier is an observation in a data set which appears to be inconsistent with the remainder of that set of data.
- Mendelhall et al. (1993): The term outlier applies to values that lie very far from the middle of the distribution in either direction.
- Barnett and Lewis (1994): An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.
- Pyle (1999): An outlier is a single or very low frequency occurrence of the value of variable that is far away from the bulk of the values of variable.

3 Statistical Tests of Extreme Values

Every sample has the largest value and the smallest value but, without an appropriate statistical test, these values cannot be called outliers. Some statistical tests have been proposed in the statistical literature to decide whether the sample maximum or the sample minimum is an extreme value. What may be interesting is to know that methods of detecting outliers were developed as early as 1852 and 1860 by Peirce and Chauvenet, respectively.

Chauvenet's criterion is to declare an observation to be an outlier if the probability of obtaining the particular deviation from the mean is less than $1/(2n)$, where n is the sample size.

Peirce's method is more rigorous and has the following steps.

1. Calculate the mean ($\bar{x}_n$) and standard deviation (σ) of the complete data set.
2. Obtain R from Peirce's table corresponding to the number of suspected outliers.
3. Calculate the maximum allowable deviation $d_{\max}(x_i, \bar{x}_n) = \sigma \cdot R$.

4. For any suspicious observation x_i, calculate $d(x_i, \bar{x}_n) = |x_i - \bar{x}_n|$.

5. Declare x_i to be an outlier if $d(x_i, \bar{x}_n) > d_{\max}(x_i, \bar{x}_n)$, that is, if $|x_i - \bar{x}_n| > \sigma \cdot R$.

6. Steps 2 to 5 are repeated by sequentially increasing the number of suspicious observations until no more outliers are detected.

The main problem with Peirce's method is the need of looking up the table and the limitation on the sample size arising out of this need. Peirce's table is available for sample sizes between 3 and 60 and the number of suspicious observations up to 9.

It is therefore found necessary to develop some more general and simple criteria for detection of outliers. Some methods are listed here.

1. Criteria based on probability or likelihood.

   A sample criterion for outlier detection may require only the mean and standard deviation of the data set. According to Chebyshev inequality, for any $k > 0$,

   $$P[|\mu| \geq k\sigma] \leq 1/k^2. \tag{1}$$

   Thus $(1 - \frac{1}{k^2})$ is the proportion of data within k standard deviations of the mean. Observations beyond these limits may be declared as outliers.

   Another criterion may further use the knowledge of the underlying probability distribution of observations and hence is expected to be better than the preceding. Obtain the probability of exceedance for a suspected observation X_n.

   $$P_F\left[X > \frac{X_n - \bar{X}_n}{\sigma}\right] = 1 - F\left(\frac{X_n - \bar{X}_n}{\sigma}\right). \tag{2}$$

   where F is the distribution function of the sample values. X_n is an outlier if this probability is sufficiently small. A similar method can be used if the sample minimum, X_1, is suspected to be an outlier. In that case, the probability of interest is

   $$P_F\left[X < \frac{X_1 - \bar{X}_n}{\sigma}\right] = F\left(\frac{X_1 - \bar{X}_n}{\sigma}\right). \tag{3}$$

One more possible criterion is to use the z-scores of sample values. Shiffler (1988) and Seo (2006) have shown that a possible maximum z score can be determined and is given by $(n-1)/\sqrt{n}$. One of the limitations of this criterion is that the standard deviation can be inflated by extreme values causing a masking effect on the criterion. As a consequence, moderate outliers may go undetected due to the presence of most extreme outliers.

One way of reducing the sensitivity of this criterion to presence of extreme values is to use the median instead of the mean. Let $\tilde{X}_n$ denote the median of sample values and define the median absolute deviation (MAD) by

$$\text{MAD} = \text{Median}(|X_i - \tilde{X}_n|) \tag{4}$$

Then the modified z score is given by

$$M_i = \frac{0.6745(X_i - \tilde{X}_n)}{\text{MAD}} \tag{5}$$

where $E(\text{MAD}) = 0.6745\sigma$ when observations follow a normal distribution.

2. Criteria based on Student's t test.

   Grubbs' test is used when the null hypothesis and the alternative are as follows.

   H_0: There are no outlier in the data set.

   H_1: There is at least one outlier in the data set.

   The test statistics is defined as

$$G = \frac{\underset{i=1,\ldots,n}{max}\,|X_i - \bar{X}_n|}{s} \tag{6}$$

   where $\bar{X}_n$ and s denote the sample mean and standard deviation, respectively. This test statistic is appropriate for a two sided alternatively. If the maximum value suspect, then the test statistics is

$$G = \frac{X_{max} - \bar{X}_n}{s}$$

and test statistic when the minimum value is to be tested is

$$G = \frac{\bar{X}_n - X_{min}}{s}.$$

The critical region for the two sided test is specified by the condition

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{\alpha}{2n},n-2}}{n-2+t^2_{\frac{\alpha}{2n},n-2}}} \tag{7}$$

For the one- sided test, $\frac{\alpha}{2n}$ is replaced with $\frac{\alpha}{n}$.

It is also possible to Cook's distance to detect outliers by treating influential observations as outlier.

4 Tests of Outliers with Examples

This section contains some tests of outliers. These can be used to decide whether the sample maximum or the sample minimum is an extreme value or outlier. All the tests are illustrated with examples. The data used for this purpose relate to daily precipitation amount (in inches) at Fort Collins, Colorado, USA, from January 1, 1900 to December 31, 1999. Yearly maximum precipitation amounts are computed and arranged in an ascending order of magnitude. The average yearly maximum precipitation is 1.7567 inches, the maximum yearly maximum precipitation is 4.63 inches, the second maximum is 4.43 inches, the minimum is 0.6 inches, and the second minimum is 0.71 inches. The sample mean square is 0.6916728, so that the root mean square is 0.8316687.

4.1 Test 1

Let $x_1, x_2, \cdots, x_n$ be the sample values. Arrange the sample values in an ascending order of magnitude, so that the ordered sample $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ is obtained. Let

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{8}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x}_n)^2 . \tag{9}$$

If $x_{(n)}$ is suspected to be an outlier, the test statistic is

$$T_n = \frac{x_{(n)} - \overline{x}_n}{s}. \tag{10}$$

If $x_{(1)}$ is suspected to be an outlier, the test statistic is

$$T_1 = \frac{\overline{x}_n - x_{(1)}}{s}. \tag{11}$$

The null hypothesis in both the cases is that all observations in the sample come from the same normal distribution.

For the Fort data,

$$\begin{aligned} T_n &= \frac{x_{(n)} - \overline{x}_n}{s} \\ &= \frac{4.63 - 1.7567}{0.8316687} \\ &= 3.454861. \end{aligned}$$

similarly,

$$\begin{aligned} T_1 &= \frac{\overline{x}_n - x_{(1)}}{s} \\ &= \frac{1.7567 - 0.6}{0.831687} \\ &= 1.390818. \end{aligned}$$

Using normal approximation, it can be concluded that $x_{(n)} = 4.63$ is an outlier while $x_{(1)} = 0.6$ is not an outlier. The normal approximation is justified in view of the sample size of 100.

4.2 Test 2

Suppose the sample values are arranged in an ascending order and the ordered sample values are $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. This test uses Dixon's Q statistic defined as follows.

$$Q_n = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}. \tag{12}$$

It is used for testing if the sample maximum $x_{(n)}$ is an outlier. Similarly, the Q statistic defined by

$$Q_1 = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \tag{13}$$

is used for testing if the sample minimum $x_{(1)}$ is an outlier.

For the Fort data,

$$\begin{aligned} Q_n &= \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} && (14) \\ &= \frac{4.63 - 4.43}{4.63 - 0.6} && (15) \\ &= 0.04962779, && (16) \end{aligned}$$

and

$$\begin{aligned} Q_1 &= \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} && (17) \\ &= \frac{0.71 - 0.6}{4.63 - 0.6} && (18) \\ &= 0.02729529. && (19) \end{aligned}$$

Since the exact probability distribution of Q_n is not known and critical values for Q_n are available in the literature for values of n up to 30 only, it is not possible to test significance of Q_n. If the values of annual maximum precipitation are assumed to be uniformly distributed, then an appropriate test is available and it states that the critical value for Q_n based on sample of size n is

$$Q_{n,\alpha} = 1 - \alpha^{1/(n-2)}. \quad (20)$$

By symmetry of the uniform distribution, the critical value for Q_1 based on sample of size of n is

$$Q_{1,\alpha} = 1 - \alpha^{1/(n-2)}. \quad (21)$$

The sample size $n = 100$ implies that the critical value is

$$Q_{n,0.05} = 0.0301. \quad (22)$$

It can therefore be concluded that $x_{(n)} = 4,63$ is an outlier, while $x_{(1)} = 0.60$ is not an outlier.

4.3 Test 3

This test is based on the variability of the entire sample and of the sample values after excluding the extreme value. More precisely, consider

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}, \tag{23}$$

$$\overline{x}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} x_{(i)}, \tag{24}$$

$$S_n^2 = \sum_{i=1}^{n} \left(x_{(i)} - \overline{x}_n\right)^2, \tag{25}$$

$$S_{n-1}^2 = \sum_{i=1}^{n-1} \left(x_{(i)} - \overline{x}_{n-1}\right)^2, \tag{26}$$

so that the test statistic for testing if the largest observation $x_{(n)}$ is an outlier can be defined as follows.

$$\frac{S_{n-1}^2}{S_n^2} = 1 - \frac{1}{n-1}\left(\frac{x_{(n)} - \overline{x}_n}{s}\right)^2, \tag{27}$$

where $s^2 = \frac{S^2}{n}$. It can be shown that

$$\frac{S_{n-1}^2}{S_n^2} = 1 - \frac{1}{n-1}T_n^2, \tag{28}$$

where T_n is defined in Equation (10). Further, defining

$$\overline{x}_{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} -2x_{(j)}, \tag{29}$$

it is possible to test whether the two largest observations are outliers with help of the test statistic

$$\frac{S_{n-2}^2}{S_n^2}. \tag{30}$$

Similar test statistics can be constructed if the smallest observation or the two smallest observations are suspected to be outliers.

5 References

1. Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data. 3rd ed. John Wiley and Sons, Chichester.

2. Chauvenet, W. (1860). A Treatise on Plane and Spherical Trigonometry. Philadelphia, pp. 256.

3. Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. Ann. Math. Statist., Vol. 21, No. 1, pp. 27-58.

4. Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics, Vol. 11, No. 1, pp. 1-21.

5. Hawkins D. (1980). Identification of Outliers, Chapman and Hall.

6. Johnson R. (1992). Applied Multivariate Statistical Analysis. Prentice Hall.

7. Mendenhall W., Reinmuth J.E. and Beaver R.J. (1993). Statistics for Management and Economics. Belmont, CA: Duxbury Press.

8. Myers, N. (1988). Threatened biotas: hot spots in tropical forests. The Environmentalist, Vol. 8, No. 3, pp. 187-208.

9. Peirce, B. (1852). Criterion for the Rejection of Doubtful Observations. The Astronomical Journal, Vol.II, No. 45, pp. 161-163.

10. Pyle, D. (1999). Data Preparation for Data Mining. San Francisco, CA: Morgan Kaufmann.

11. Seo, S. (2006). Thesis: A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets, University of Pittsburgh, Graduate School of Public Health.

12. Schiffler, R. E. (1988). Maximum Z Score and outliers. The American Statistician, Vol. 42, No.1, pp. 79-80.