Asian Journal of Multidisciplinary Studies Available online at www.ajms.co.in
Volume 2, Issue 1, January 2014
ISSN: 2321-8819

The Effect of the Swamping Phenomenon on Several Block Procedures Samples

Ashok Rajak

Lecturer, Shivalik College of Engineering Sihniwala Shimlaroad PO,Sherpur Dehradun.(Uttrakhand),India

ABSTRACT In the statistical theory of the design of experiments, blocking is the arranging of experimental units in groups (blocks) that are similar to one another. Sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population. Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wish to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population .this paper reviews the general outlier problem, Two alternate approaches to the multiple outlier problem, consecutive and block testing, and their respective inherent weaknesses, masking and swamping, are discussed. In addition, the relative susceptibility of several tests for outliers in normal samples to the swamping phenomena is reported.

Keywords: Detection; Swamping; Masking

INTRODUCTION

“An outlier, or more descriptively an “outright liar” , is an observation (even a subset of observations) which appears to be out of line, that is, not consistent with the remainder of a data set”[1]. Such mavericks often either in-innocently missed in the blind transition between data col-election and computer or not so innocently slipped under the rug of “what isn’t seen can’t hurt”, should fire the curiosity and concern of researchers.

The Origin of Outliers

Outliers: The Story of Success is a non-fiction book written by Malcolm Glad well and published by Little, Brown and Company on November 18, 2008. In Outliers, Glad well examines the factors that contribute to high levels of success. To support his thesis, he examines the causes of why the majority of Canadian ice hockey players are born in the first few months of the calendar year, how Microsoft co-founder Bill Gates achieved his extreme wealth, how The Beatles became one of the most successful musical acts in human history, how Joseph Flom built Skadden, Arps, Slate, Meagher & Flom into one of the most successful law firms in the world, how cultural differences play a large part in perceived intelligence and rational decision making, and how two people with exceptional intelligence, Christopher **Langan and J. Robert** Oppenheimer, end up with such vastly different fortunes. Throughout the publication, **Gladwell** repeatedly mentions the "10,000-Hour Rule", “claiming that the key to success in any field is, to a large extent, a matter of practicing a specific task for a total of around 10,000 hours. As such, the outlier must be studied “relative to some initial model for the data generation” [2: p. 247].

Methods for Outlier Detection

“There are seven generally accepted “forms” of statistical tests of discordance [see, 3]: Excess/spread statistics are ratios of the differences between an outlier and its nearest neighbor to the range or spread of the sample “[4,5].

“Range/spread statistics replace the numerator of the excess/spread statistic with the sample range and contrast it with another measure of dispersion, often the sample standard deviation” [6,7].

“Deviation/spread statistics use in the numerator a measure of distance between an outlier and some measure of central location in the sample” [8].

“Sum of squares statistics are tests expressed as ratios of sums of squares for the reduced and total samples. Reduced sample sum of squares simply refers to the calculation based upon the total sample minus outliers” [8].

“Higher-order moment statistics are tests not specifically designed for assessing outliers, such as skewness and kurtosis, but which nevertheless are quite useful in this context” [9, 10].

“Extreme-location statistics take the form of ratios of extreme values (outliers to measures of central location), usually the sample mean” [11-14].

“ W-statistics for normal data are simply the ratio of the square of a linear combination of the ordered sample values to the sum of squares of the individual deviations about the mean” [15-17].

The Multiple Outlier Problems: Consecutive versus Block Testing

The simplest situation to imagine is a single upper or lower, potentially discordant, outlier. Suppose, however, that the researcher observes or anticipates a cluster (*k* 2, where *k* is the number of outliers) of outliers in an extreme upper and/or lower position relative to the reminder of the sample. Two contrasting approaches have been proposed in the literature for dealing with such multiple outlier problems.

Consecutive testing requires that a given single-outlier procedure be applied repeatedly to outliers, one at a time, beginning with the most extreme observation and proceeding "in order of decreasing degree of deviancy, until an observation "fails" the discordance test (., moving inward) and is declared to be consistent with the rest of the data set" [19: pp. 9-10]. "Alternatively, a single-outlier procedure can be applied consecutively in an outward direction until significance is reached" (20, 21]. This latter approach has an advantage that will be addressed below.

Block procedures, on the other hand, require the re-searcher to scrutinize suspected discordant observations as a unit. Tests have been devised for a single upper-and-lower outlier pair, blocks (*k* 2) of upper or lower outliers, or clusters (*k* 2) of both upper and lower suspected values. Block testing is an all-or-none proposition in that all observations in the unit are declared discordant, or none of them are.

In sample *A*, most observers would agree that the two upper values, 30 and 35, are suspicious. Where to draw the line to delineate the upper outliers becomes more of a challenge in sample *B*. And in sample *C* it is not even clear as to which end of the data set might, in fact, contain outliers.

Obviously, prespecification of the number of outliers is not an issue in consecutive testing. However, one problem facing the use of consecutive procedures is the effect of masking.

Masking

Barnett and Lewis define masking as "the tendency for the presence of extreme observations not declared as outliers to mask the discordance of more extreme observations under investigation as outliers" [22: p. 114]. Take, for example, data set *A* above. If one wished to check for the discordance of the values at the upper end of the sample using the consecutive testing approach, the first objective would be to apply the chosen single-outlier procedure to the observation $X_{(9)}$ = 35. (In this paper, $X_{(i)}$ refers to the ith ordered value in a given sample of elements.) If $X_{(9)}$ were to be declared discordant, the next move would be to test $X_{(8)}$ = 30 and so on until an observation failed to lead to a rejection. Presumably, $X_{(9)}$ and $X_{(8)}$ would be declared an outlier set. Masking can influence the outcome of this process, however, if the extreme nature of the observation $X_{(8)}$ (either a valid member of the sampled population or a discordant outlier itself) prevents the detection of $X_{(9)}$ as an outlier. Therefore, the masking phenomenon can halt the initiation of any consecutive testing procedure. However, consecutive application of single-outlier tests in an outward direction may avoid this problem [20,21].

Swamping

Drawing again on Sample *A* (previously discussed in connection with masking), a simple example will demonstrate block testing's analogue to masking: the swamping phenomenon. Suppose that, for whatever reason, a re-searcher wishes to apply a block procedure for outlier detection to sample *A* and specifies the number of suspicious observations (*k*) to be the upper 3. That is, the values $X_{(9)}$ = 35, $X_{(8)}$ = 30, and $X_{(7)}$ = 7 are to be tested as being inconsistent with the remainder of the sample against some prespecified underlying probability distribution. A block test applied to these three upper values may well declare them discordant as a unit; the extreme observations 30 and 35 have "carried" the otherwise un-exceptional value 7.

The above example cites a situation with an obvious upper outlier pair; however, the use of a block test for *k* = 3 upper outliers may declare 35, 30, and 7 discordant. The nearest neighbor to 30 and 35, *i.e.*, 7, may be a valid member of the population being sampled. The phenomena-non may easily be simplified to a single outlier swamping a single neighbor or generalized to a situation where a cluster of *k* outliers swamps the nearest (n - k)th neighbor(s).

Methods

The approach taken to study swamping in this study was similar to Fisher's in that it involved a simulation in which various sizes (*n* = 10, 30 and 50) of computer generated pseudo-random samples from a normal distribution (= 0, = 1) based upon a self-programmed four-parameter algorithm. Minitab (release 16.2.3) was used to generate assessments of the swamping phenomenon as well as all graphics. Outliers were placed at the upper end of each ordered sample according to specified criteria for each of the block tests being studied. Having all outliers placed, a unit-free swamping index was calculated for Special Case I (*k* = 1 outlier swamping its nearest neighbor), Special Case II (*k* = 2 outliers swamping a third value) and

Special Case III (k = 3 outliers swamping a fourth value, and the simplest example of the most generalized case of k outliers swamping the nearest $(n - k)^{\text{th}}$ neighbor).

Test Selection and the Swamping Consideration

This study focused on four specific block procedures for multiple outliers. Test T_2 was restricted in its usefulness due to its design (only appropriate for k = 2 outliers), as well as limitations in the availability of distributional information. It is recognized that any attempt at outlier test selection is influenced by many considerations. The researcher must contemplate the goals and objectives of the study, existing knowledge about the population.

Conclusion:

Findings in the swamping study indicated a similar phenomenon with the single outlier (or "centroid" of 2 or 3 outliers in cases II and III, respectively) being placed within the upper boundary set by $X_{n\ k}$ $_1$ more often with increasing sample size for T_1 and T_4, and yet T_2 and T_3 were inversely related to n in this respect. This could simply indicate, as suggested by Fisher's results, that those block tests exhibiting this enigma rather frequently are quite liberal in terms of outlier definition. However, this peculiarity deserves further thought, discussion, and investigation.

"The need to assess and rank single- and multiple-outlier tests in terms of their relative degree of "conservatism" concerning outlier definition, as well as "susceptibility" to such phenomena as masking and swamping may assist with developing the generalized outlier methodology" alluded to by Barnett and Lewis [3], Rosner [26], and others.

REFERENCES

1. F. Wilcoxon, "Personal Communication to J. K. Brewer in Class Presentation for Statistics 405," Florida State University, Tallahassee, 1962.
2. V. Barnett, "The Study of Outliers: Purpose and Model," *Applied Statistics*, Vol. 27, No. 3, 1978, pp. 242-250.
3. V. Barnett and T. Lewis, "Outliers in Statistical Data," 3rd Edition, Wiley, New York, 1994.
4. W. J. Dixon, "Ratios Involving Extreme Values," *Annals of Mathematical Statistics*, Vol. 22, No. 1, 1951, pp. 68-78.
5. J. O. Irwin, "On a Criterion for the Rejection of Outlying Observations," *Biometrika*, Vol. 17, No. 3-4, 1925, pp. 238-250.
6. H. A. David, H. O. Hartley and E. S. Pearson, "The Dis-tribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation," *Biometrika*, Vol. 41, No. 3-4, 1954, pp. 482-493.
7. E. S. Pearson and M. A. Stephens, "The Ratio of Range to Standard Deviation in the Same Normal Sample," *Bi-ometrika*, Vol. 51, No. 3-4, 1964, pp. 484-487.
8. F. E. Grubbs, "Sample Criteria for Testing Outlying Ob-servations," *Annals of Mathematical Statistics*, Vol. 21, No. 1, 1950, pp. 27-58.
9. T. S. Ferguson, "On the Rejection of Outliers," *Proceed-ings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 1961, pp. 377-381.
10. T. S. Ferguson, "Rules for the Rejection of Outliers," *Re-vue Institute Internationale de Statistica*, Vol. 29, No. 3, 1961, pp. 29-43.
11. B. Epstein, "Tests for the Validity of the Assumption That the Underlying Distribution of Life Is Exponential:
12. Part I," *Technometrics*, Vol. 2, No. 1, 1960, pp. 83-101.
13. B. Epstein, "Tests for the Validity of the Assumption That the Underlying Distribution of Life Is Exponential: Part II," *Technometrics*, Vol. 2, No. 1, 1960, pp. 167-183.
14. T. Lewis and N. R. J. Fieller, "A Recursive Algorithm for Null Distributions for Outliers: I, Gamma Samples," *Te-chnometrics*, Vol. 21, No. 3, 1978, pp. 371-376.
15. J. Likes, "Distribution of Dixon's Statistics in the Case of an Exponential Population," *Metrika*, Vol. 11, No. 1, 1960,

16. S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for Normality (Complete Samples),” *Biometrika*, Vol. 52, No. 3-4, 1966, pp. 591-611.

17. S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for the Exponential Distribution (Complete Sam-ples),” *Technometrics*, Vol. 14, No. 2, 1972, pp. 355-370.

18. S. S. Shapiro, M. B. Wilk and M. J. Chen, “A Compara-tive Study of Various Tests for Normality,” *Journal of the American Statistical Association*, Vol. 63, No. 324, 1968,1343-1372.

19. P. L. Fisher, “Comment on the Subjective Decisions Re-quired of the Researcher in the Selection of a Statistical Outlier Test,” *Florida Journal of Educational Research*, Vol. 12, 1980, pp. 27-41.

20. P. L. Fisher, “An Investigation of Outlier Definition and the Impact of the Masking Phenomenon on Several Sta-tistical Outlier Tests,” Unpublished Doctoral Dissertation, The Florida State University, Tallahassee, 1980.

21. T. J. Sweeting, “Independent Scale-Free Spacing for the Exponential and Uniform Distributions,” Statistics and Probability Letters, Vol. 1, No. 3, 1983, pp. 115-119

22. T. J. Sweeting, “Asymptotically Independent Scale-free Spacings with Applications to Discordancy Testing,” *An-nals of Statistics*, Vol. 14, 1986, pp. 1485-1496

23. V. Barnett and T. Lewis, “Outliers in Statistical Data,” 2nd Edition, Wiley, New York, 1984.

24. W. J. Dixon, “Analysis of Extreme Values,” *Annals of Mathematical Statistics*, Vol. 21, No. 4, 1950, pp. 488-506.

25. R. B. Murphy, “On Tests for Outlying Observations,” Un-published Doctoral Dissertation, Princeton University, Dis-sertation Abstracts International, 15/03, University Mi-crofilms No. A55-534, 1951.

26. G. L. Tietjen and R. H. Moore, “Some Grubbs-Type Sta-tistics for the Detection of Several Outliers,” *Technomet-rics*, Vol. 14, No. 3, 1972, pp. 583-597.

27. B. Rosner, “On the Detection of Many Outliers,” *Techno-metrics*, Vol. 17, No. 2, 1975, pp. 221-227.