Testing for Outlying Observations in a Sample Group[1]

CHARLES D. BONHAM

Associate Professor, Department of Range Science, Colorado University, Fort Collins, Colorado.

Highlight

The need to test for "outliers" is often overlooked both in statistical analyses of data, and in applied statistics courses. Instead of discarding an "odd" value from the sample data based on intuition, an objective approach should be used in handling spurious values found in a data group. An outlier testing procedure can be also useful in constructing future sampling designs.

Testing for homogeneity in statistical sampling is an old problem particularly when references are made to variances and goodness of fit tests. If one should be interested in homogeneity within a sample group collected for statistical analysis, consideration might be given to individual data values. Often one or more of these points present problems of interpreting the data. Apparent inconsistencies, one or more values, which are frequently found within a sample group contribute a large amount of variation to the total variance. In view of possible difficulties in making conclusions, should non-homogeneous data points be present, one might consider a test for "outliers" to check for possible diversions from the normal population.

Range researchers and technicians have frequently found one or more sample observations to be inconsistent with other values in the sample group. Even though the sample was randomly drawn from a population of the statistical universe, a critical examination of the data collected is still necessary for assurance that the results of sampling are representative of the thing under examination. Some research workers would accept any possible "outlier" as a "representative" of the population, even though it tends to distort the statistical results and any subsequent conclusions which are formulated from sample information. Other workers may discard a data point from the statistical analysis because "common sense" indicates such action would be appropriate. The latter effort may be completely justified while the former may be questionable. In either case, an objective method would be helpful in making a decision as to the disposition of the sample value or values in question.

Discussion of the Model

It is not proposed to discard data to allow for favorable statistical conclusions, but rather to use acceptable methods for probabilistic exclusion of observations that do not represent that portion of the population under investigation. Each value is to represent the same population parameter, μ, the true population mean, and a given observation may so depart from this popula-

[1] Received July 3, 1970.

tion parameter as to be of no value in estimating it. We would have reason for wanting to know if this has happened. Possible reasons for this departure can be seen and perhaps explained entirely by: (1) errors in measurements or recording and/or (2) the population may not be homogeneous in the unimodal sense (Grubbs, 1950). Knowledge of the first could lead to a correction of the data while the latter reason would be helpful in changing the sampling plans to obtain a more reliable estimate of the population parameters. Further, due consideration should be given to theoretical sampling models which require that each member of the population be given an equal chance at being a representative in the drawn sample. After each sample point is acquired statistically, further analysis then could be used to determine the relative position of a given point in relation to the remaining points and its worth in estimating the population mean.

Literature dealing with the philosophy behind testing outlying observations has recently increased due to renewed interest in spurious data values. Anscombe (1960) and Grubbs (1969) include a good discussion on the history and philosophy of rejecting outlying observations in a sample group. Most of the statistical developments on outliers or spurious values deal with the estimation of μ from a normal population (Anscombe, 1960). Inclusion or exclusion of suspect values, then, will be subject to certain risks which have been referred to as premium or protection. Such risks can be evaluated by observing the mean squared error and noting an inflation or alternately a decrease in its value depending on inclusion or exclusion of the suspect value (Tiao and Guttman, 1967). Rider (1933) and Grubbs (1950, 1969) present a review of the work that has been done in the statistical field dealing with outliers. The development by Grubbs (1950) resulted in the calculation of the probability that an outlying value would occur inconsistently with the remaining values in the sample. This technique should be used in situations where odd values occur and there appears to be a question of proper data treatment. The statistic,

$$S_n^2/S^2 = \sum_{i=1}^{n-1} (X_i - \overline{X}_n)^2 / \sum_{i=1}^{n} (X_i - \overline{X})^2$$

is calculated where S_n^2 is the sums of squares calculated excluding the suspect value and S^2 is the sums of squares calculated using all data points. The distribution of this ratio yields percentage points which are arranged in tabular form for acceptance or rejection of the suspect value. Another basic approach to the problem was taken by Thompson (1935). He showed that if the value,

$$T = (X_i - \overline{X})/s,$$

where

$$\overline{X} = (1/n) \sum_{i=1}^{n} X_i;$$

$$s^2 = (1/n) \sum_{i=1}^{n} (X_i - \overline{X})^2; \quad \text{and } X_i$$

is an observation within a given sample group of size n which was drawn from a normal population, is calculated, then the statistic,

$$t = T\sqrt{n-2}/\sqrt{n-1-T^2}$$

follows the "t-distribution" with n–2 degrees of freedom. This test is useful for testing the criterion that a single observation came from a population with divergent mean and for $n < 30$ (Pearson and Chandra Sekan, 1936). One shortcoming of Thompson's (1935) approach can be seen from the practical viewpoint. That is, the value selected arbitrarily is not necessarily the largest value or the smallest value and we are indeed more interested in one or both of these sample points. Grubbs' (1950) proposal is the more applicable for objective discarding of biological data points.

Application of the Test

Data used as an example are the amounts of K_2O (lb./acre) present in alpine soil collected on Independence Pass, Colorado, as follows:

70	130	145	400
90	130	185	
120	140	200	

It would be desirable to know if these sample values show uniformity over the entire area or if any of the values are inconsistent with the others. The value 400 is strongly suspected of being an "outlier." Therefore, the ratio S_n^2/S^2 is used for an evaluation of this data point. The ratio is equal to 0.174 and from Table 1 of Grubbs (1950) $P < 0.01$ so the observation is rejected as being an "outlier." Elimination of this value reduced the variance of the sample group by 80 percent and consequently any further tests involving the mean of this group would be more effective in detecting real differences. Having rejected the largest value, attention is focused on the lowest value 70 to check it for inconsistency as a further example even though it appears to be homogeneous with the remaining values. Now $S_n^2 = 8700.0$ and $S^2 = 13372.2$. Calculating the new ratio to be 0.651 and comparing the value with the tabular values, $P < 0.10$ so the value 70 is retained as a homogeneous value. It should be pointed out that from a practical standpoint, other probability levels may be chosen to fit the individual need, depending on the type of data and circumstances surrounding the sampling procedures.

For procedures involving the two lowest values and/or the two highest values, referral is made to Grubbs (1950, 1969). It should also be noted that the method is similar to one-way analysis of variance tests where one group contains only one sample while the other group contains the remaining n–1 samples.

The sample area from which the observation, 400, came was further observed to be different from the remaining nine. The area was located on a rocky ridge with a poorly developed soil whereas the other areas were in depressions having fairly well developed soils. Ecologically, a separate strata based on soil profile exists in the area and should be recognized in further analyses.

Conclusions

It has been noted that sample data drawn from supposed normal populations can have inconsistent or spurious data points present which offer difficulties in drawing conclusions from a statistical analysis. Yet, standard courses in applied statistics often do not include methods for handling such data. Probabilistic exclusion of such outlying points can be accomplished by a homogeneity test which is an easy and objective method for eliminating a value that could be the result of (1) a measurement error and/or is (2) from a population with a divergent mean (another strata). After statistical rejection of such data, possible errors may be detected and/or sampling plans could be objectively changed to:

(1) Sampling by stratification to include the suspected point within a stratum having a homogeneous group and thus obtain a more

reliable estimate of the population parameters or,

(2) Allow for critical examination of the data to assure that the results are representative of the thing that is sampled.

Literature Cited

ANSCOMBE, F. J. 1960. Rejection of outliers. Technometrics 2:123–147.

GRUBBS, FRANK E. 1950. Sample criteria for testing outlying observations. Ann. of Math. Stat. 21:27–58.

GRUBBS, FRANK E. 1969. Procedures for detecting outlying observations in samples. Technometrics 11:1–21.

PEARSON, E. S., AND C. CHANDRA SEKAN. 1936. The efficiency of statistical tools and a criterion for the rejection of outlying observations. Biometrika 28:308–320.

RIDER, P. R. 1933. Criteria for rejection of observations. Wash. Univ. Studies-New Series, Sci. and Tech. No. 8, St. Louis.

THOMPSON, W. R. 1935. On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. Ann. of Math. Stat. 6:214–219.

TIAO, G. C., AND IRWIN GUTTMAN. 1967. Analysis of outliers with adjusted residuals. Technometrics 9: 541–559.