# A Comment on the Subjective Decisions Required of the Researcher in the Selection of a Statistical Outlier Test

Patricia Fisher

University of Tennessee - Knoxville

Statistical outliers are extreme, unexpected, and seemingly unrepresentative elements in data sets. Their distinguishing characteristic is that they are located some distance away from the main body of data -- distant enough to cuase an investigator to react with surprise to their peculiar position.

The initial reaction of the researcher who encounters an outlier is a subjective response based on several factors, including his or her presumption of the population probability distribution. Subjective judgements are frequently unreliable as measurements, and apparently, researchers' judgements of outliers are no exception. Collett and Lewis have demonstrated that researchers' outlier identifications vary according to individual, occasion, mode of presentation, and measurement units (1976).

An ostensible alternative to outlier identification by adjudication is the application of an objective statistical outlier test. However, the actual selection of a suitable procedure also turns out to be a problem of subjective judgement. There are several dozen documented tests. For a given research situation, careful selection (or more accurately, careful elimination) based on commonly available guidelines probably reduces the list to less than half a dozen possible

appropriate procedures. Beyond that, the researcher must rely mainly on intuition in the selection of an outlier test. A comment from Costner, writing on a different subject, provides a splendid description of the situation:[1]

"We suffer an embarrassment of riches with regard to the measures...it is frequently difficult to decide which specific measure is suited to one's needs...Although several very thoughtful papers and textbook discussions have attempted to clarify selected measures and to suggest relevant criteria of choice, reasonable clarity remains to be achieved with regard to...the basis for choosing among them." (1965, p. 341) Costner was speaking of measures of association, rather than outlier tests; however, the problem is precisely the same.

The purpose of this paper is to discuss several of the decisions facing the researcher in the choice of an outlier test, and to demonstrate the impact of these decisions on the ultimate outcome of the tests.

## Types of Outlier Tests

Upon surveying the avilable outlier tests developed for application under the assumption of particular probability distributions, Barnett and Lewis were able to identify six distinct categories (1978). Each of these categories is listed below, and an example test of each type is provided. (Throughout this paper, the notation $x_{(i)}$ refers to the ith order value of the sample, and $\bar{x}$ and $s^2$ denote unbiased sample estimates of $\mu$ and $\sigma^2$ .)

1. Excess/spread statistics, e.g.,

   $T_1 = (x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(1)})$

   (Dixon, 1951; Irwin, 1925);

2. Range/spread statistics, e.g.,

$T_2 = (x_{(n)} - x_{(1)})/s$,

(David, Hartley, and Pearson, 1954; Pearson and Stephens, 1964);

3. Deviation/spread statistics, e.g.,

$T_3 = (x_{(n)} - \bar{x})/s$, (or $T_3' = (\bar{x} - x_{(1)})/s$), Grubbs, 1950);

4. Sums of squares statistics, e.g.,

$$T_4 = \frac{\sum_{i=1}^{n-k} (x_{(i)} - \bar{x}_{n,n-1,\ldots,n-k+1})^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2}$$

where $\bar{x}_{n,n-1,\ldots,n-k+1} = \frac{\sum_{i=1}^{n-k} x_{(i)}}{n-k}$

and k = number of suspected outliers (Grubbs, 1950);

5. Higher-order moment statistics, e.g.,

$$T_5 = \frac{\sum_{i=i}^{n} (x_i - \bar{x})^4}{ns^4}$$

(Ferguson, 1961a); and

6. Extreme/location statistics, e.g.,

$T_6 = x_{(n)}/\bar{x}$ (Epstein, 1960a, b; Likes, 1966).

A glance at the examples from the six categories reveals that different criteria are used for each test, so it comes as no surprise that the tests yield different results. However, when different tests are equally appropriate for a given situation, confusion prevails. For example, three of the six tests presented, the excess/spread, the deviation/spread, and

the higher-order moment tests, would be judged appropriate, according to currently available guidelines, for application in a situation where a single upper outlier was suspected in a sample randomly drawn from a normal population. (The range/spread statistic in example two is used to test for an upper-and-lower outlier-pair in a "normal" sample. The sums of squares test in example four is used to test for multiple upper outliers; actually, for k=1, the test statistic has an identical distribution to that of example three. The extreme/location example is appropriate for Gamma Distributions.)

## Nature of the Outlier Tests

The first three types of tests presented, the excess/spread, range/spread, and deviation/spread statistics, each compare an indicator of the deviancy of the suspected outlier to some measure of sample spread. In essence, they each use the direct criterion of distance of the suspected observation from some other location. In the three examples, this distance is measured from the next largest value, from the smallest value, and from the sample mean, respectively. The example test in category three, utilizing distance from the mean, might have some appeal for a researcher, since it expresses the location of the deviant in the familiar "standard deviation units," but otherwise, there would seem to be no particular practical reason to strongly favor any one of these three criteria. The researcher would probably rather select from among these tests based on some knowledge of the test's performance record in correctly identifying outliers, or based on data revealing which test is, in general, most or least "willing" to identify outliers. Partial information of this sort can be dredged in bits and pieces

from the more technical statistical journals, but in general, is not readily available to practicing social science researcher.

The above measures of "outlier" distance are compared to a measure of sample spread, specifically the range or the standard deviation. For those tests that utilize the standard deviation, the calculation may be the usual unbiased estimate, $s^2$, or may be a sample estimate of population variance based on the "reduced" sample size n-1, omitting the suspected observation. Again, the decision is the researcher's: do we allow the deviant to testify at its own trial, by including its value in the sample estimate of $\sigma^2$, or do we ignore its contribution to sample spread in advance, before we have even asked the question regarding the likelihood of its valid membership in the population? The former approach tends to reduce the probability that the observation will be declared an outlier, since its inclusion increases s, and thus lowers the test ratio (rightfully so, if the observation is not an outlier, but deceitfully so, if it is). The latter approach has the opposite impact.

Outlier tests which fall into category four also utilize the criterion of distance of the extreme observation from the "rest" of the data, but in a less direct manner than the tests described above. Sums of squares statistics are based on a ratio of sums of squared deviations from the mean. The denominator is the sum of squares for the entire sample, and the numerator is the sum of squared deviations about the mean of the reduced sample, omitting one or more suspected outliers

from the calculation. As mentioned before, when used to test for one upper outlier, the result of the example presented in category four is identical to the deviation/spread example in category three. The sums of squares test, however, is commonly used to check for k outliers, where $2 \leq k \leq n$, and (once more!) k is subjectively determined by the researcher.

Higher-order moment statistics are based on the higher-order deviations about the mean, specifically the characteristics of skewness and kurtosis. These tests were originally intended as checks for normality. The rationale for their use as outlier tests is that, under the circumstance of random sampling from a normal population, observations extreme enough to significantly impact the skewness or kurtosis are "unlikely" to occur, and such observations could reasonably be declared outliers. The characteristics of skewness and kurtosis, however, are more subtle concepts than variance and central locations. A researcher may be able to see at a glance the effect on the mean and variance of a single extreme value, but this intuitive insight may not necessarily extend to the characteristics based on the third and fourth central moments. It is not clear, therefore, just what research conditions might motivate the investigator to select an outlier test from the higher-moment category, nor what the advantages of such a test might be.

The class of outlier tests referred to as extreme/location statistics has the simplest form of all the tests. They are documented as appropriate for Gamma distributions, but do not seem to be in common usage.

In summary, there are many different outlier tests available to the researcher, but except for the researcher's expectations of the appropriate model for the data (which is often another subjective judgement), there are very few guidelines for their practical application. The fact that tests appropriate for the same research situation use different criteria for identification of outliers illustrates the basic problem of a lack of a single satisfactory definition of "outlier."

The following section of this paper demonstrates that there are, in fact, important differences among the values that would be identified as sample outliers by several tests deemed suitable for a given research situation, and furthermore, that the subjective judgements required of the researcher in this decision-making process have a substantial impact on the testing results.

## Example Application

In any outlier situation, the researcher is faced with the following questions: What observations are likely candidates for outlier testing? Which outlier tests should be used? On what basis should the sample statistics required by the tests be calculated?

Suppose an investigator drawing a random sample from a presumed normal population, with unknown mean and variance, obtained the following data: 23, 31, 34, 37, 41, 43, 52, 75. This data set has a mean of 42 and a standard deviation of

15.86. (An explanation of the origin of the data, and all calculations for the discussion which follows, appear in the Appendix.)

Using a standard deviation based on the total sample, application of tests $T_1$ and $T_5$ to the above data set to check the upper observation $x_{(8)}$ yields non-significant results, whereas application of test $T_3$ declares the value 75 an outlier. Thus, it is immediately apparent that the determination of the status of the uppermost observation is at least partly dependent on the (perhaps random) choice of an outlier test.

Using the reduced sample standard deviation, which is one of the options available to the researcher, reverses the decision of test $T_5$, and results in the identification of the value 75 as an outlier by this test. (The use of the reduced statistic would have no impact on the results of the other two tests; $T_1$ does not utilize this value, and for $T_3$, the reduced variance would serve to increase the $T_3$ ratio, which was already above the critical value.) Once more, the intuitive-based decision required of the researcher influences test findings.

Turning now to the other end of the distribution and utilizing test $T_3$, the value $x_{(1)}$ is not identified as an outlier, using either total or reduced sample variance in the test. However, we have some indication that the upper value 75 may be, in fact, an outlier; if so, is it fair to

judge the lower value 23 with 75 in the sample? Exercising the researcher option of eliminating 75 as an upper outlier, test $T_3$ still fails to identify the value 23 as a lower outlier, based on total sample variance, but yields significant results, based on the reduced sample statistic.

Resorting to test $T_2$ to check the values 23 and 75 simultaneously as an upper-and-lower outlier-pair, the test declares both discordant using reduced sample variance, and neither discordant using total sample variance.

Finally, using test $T_4$ to check for an upper outlier-pair, the test declares observations $x_{(8)}$ and $x_{(7)}$ to be outliers, raising questions regarding the probable validity of the second largest observation, 52. Obviously, this example could be continued at length.

## Summary

Allowing Costner to speak for us once again, "on what basis does one choose ...? It is hardly surprising that some researchers are reluctant to enter such an esoteric thicket..." (p. 341). At times, researchers have been criticized for their subjective decisions to eliminate extreme observations from samples; others have probably avoided open criticism by simply "failing to report" extreme observations. However, such persons might argue convincingly that their single act of "outlier rejection" on purely subjective grounds might be less likely to result in error than a series of such subjective and arbitrary decisions.

Barnett and Lewis comment, "when all is said and done, the major problem in outlier study remains the one that faced the very earliest workers on the subject--what is an outlier?" (p. 286) "Surely the professional statistician [can no longer] reasonably withhold his contribution..." [to the area of outlier research.] "What of the future?" (p. 288) Outlier research has come far in 200 years of effort, without an acceptable definition of "outlier." How much progress could we enjoy if we knew exactly what we were talking about? Until a satisfactory definition of an outlier is formulated, the social science researcher needing an outlier test would be well-advised to do some research on the test itself, and on the consequences of the subjective decisions required in its very execution.

## Appendix

The data set for this example was derived as follows. The Random Number Generator of the International Mathematical and Statistical Library of computer programs was utilized to generate a random sample of size eight from a normal distribution with zero mean and unit variance. To eliminate negative numbers, the value four was added to each element. Then each value was multiplied by ten and truncated at the decimal, resulting in the random sample 23, 31, 34, 37, 41, 43, 52, and 75, representative of a normal population with a mean of 40 and a standard deviation of ten.

The following sample statistics were used in the calculations for the example application in this paper:

| | Mean | Standard Deviation |
|---|---|---|
| Total sample | 42.00 | 15.86 |
| Reduced sample, omitting $x_{(8)}$ | 37.29 | 9.29 |
| Reduced sample, omitting $x_{(1)}$ | 44.71 | 15.00 |
| Reduced sample, omitting $x_{(1)}$ and $x_{(8)}$ | 39.67 | 7.38 |

The following calculations were made for the example application in this paper:

A. Application of $T_1$, $T_3$, and $T_5$ to test $x_{(8)}$, using total sample statistics:

$$T_1 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} = \frac{75 - 52}{75 - 23} = 0.442 \quad (T_1 \text{ critical} = 0.468)$$

$$T_3 = \frac{x_{(n)} - \bar{x}}{s} = \frac{75 - 42}{15.86} = 2.08 \quad (T_3 \text{ critical} = 2.03)$$

$$T_5 = \frac{\sum_{i=1}^{\mu} (x_{(i)} - \bar{x})^4}{ns^4} = \frac{1345606}{8\ (15.86)^4} = 2.65 \quad (T_5 \text{ crit} = 3.5)$$

B. Application of $T_5$ to test $x_{(8)}$, using reduced sample

$$T_5 = \frac{1345606}{8\ (9.29)^4} = 22.63 \quad (T_5 \text{ critical} = 3.5)$$

C. Application of $T_3$ to test $x_{(1)}$

i) using total sample statistics:

$$T_3 = \frac{\bar{x} - x_{(1)}}{s} = \frac{42 - 23}{15.86} = 1.19 \quad (T_3 \text{ critical} = 2.03)$$

ii) using reduced sample standard deviation, omitting $x_{(i)}$:

$$T_3 = \frac{42 - 23}{15} = 1.27 \quad (T_3 \text{ critical} = 2.03)$$

D. Application of $T_3$ to test $x_{(1)}$ after eliminating upper "outlier" $x_{(8)}$ from the sample

i) using total sample statistics:

$$T_3 = \frac{37.29 - 23}{9.29} = 1.53 \quad (T_3 \text{ critical} = 1.94)$$

ii) using reduced sample standard deviation, omitting $x_{(1)}$:

$$T_3 = \frac{39.67 - 23}{7.38} = 2.26 \quad (T_3 \text{ critical} = 1.94)$$

E. Application of $T_2$ to test $x_{(1)}$ and $x_{(n)}$ simultaneously an outlier pair

i) using total sample statistics:

$$T_2 = \frac{x_{(n)} - x_{(1)}}{s} = \frac{75 - 23}{15.86} = 3.28 \quad (T_2 \text{ critical} = 3.40)$$

ii) Using reduced sample standard deviation omitting both $x_{(1)}$ and $x_{(8)}$:

$$T_2 = \frac{75 - 23}{7.38} = 7.048 \quad (T_2 \text{ critical} = 3.40)$$

E. Application of $T_4$ to test $x_{(7)}$ and $x_{(8)}$ as an upper outlier pair:

$$T_4 = \frac{\sum_{i=i}^{\mu-2} (x_i - \bar{x}_{n,\, n-1})^2}{\sum_{i=i}^{\mu} (x_i - \bar{x})^2}, \text{ where } \bar{x}_{n,\, n-1} = \frac{\sum_{i=i}^{\mu-2} x_i}{n - 2}$$

$$T_4 = \frac{264.83}{1762} = 0.1503 \quad (T_4 \text{ critical} = 0.148)$$

## References


Barnett, V. & Lewis, T. Outliers in statistical data. New York: John Wiley & Sons, 1978.

Collett, D. & Lewis, T. The subjective nature of outlier rejection procedures. Applied Statistics, 1976, 25, 228,237.

Costner, H. L. Criteria for measures of association. American Sociological Review 1965, 30, 341.

David, H. A., Hartley, H. O., & Pearson, E. S. The distribution of the ratio, in a single normal sample, of range to standard deviation. Biometrika, 1954, 41, 482-493.

Dixon, W. J. Analysis of extreme values. Annals of Mathematical Statistics, 1950, 21, 488-506.

Dixon, W. J. Ratios involving extreme values. Annals of Mathematical Statistics, 1951, 22, 68-78.

Epstein, B. Tests for the validity of the assumption that the underlying distribution of life is exponential: part I. Technometrics, 1960a, 2, 83-101.

Epstein, B. Tests for the validity of the assumption that the underlying distribution of life is exponential: part II. Technometrics, 1960b, 2, 167-183.

Ferguson, T. S. On the rejection of outliers. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1961a, 1, 253-287.

Grubbs, F. E. Sample criteria for testing outlying observations. Biometrika, 1925, 17, 238-250.

Likes, J. Distribution of Dixon's statistics in the case of an exponential population. Metrika, 1966, 11, 46-54.

Pearson, E. S., & Stephens, M. A. The ratio of range to standard deviation in the same normal sample. Biometrika, 1964, 51, 484-487.