

Testing for Outliers in Small to Moderate Samples

By Kenneth C. Syracuse, Ph.D.

Keywords:

Outlier, Extreme Value

Abstract

The presence of outliers in a sample data set can have a great effect on subsequent analyses. It is important to determine if a causal relationship exists which may allow for the removal of the point as coming from a unit no longer homogeneous with the remaining units. For example, a unit improperly tested. If such a cause can not be determined or the effect of such a cause cannot be determined, it may be necessary to use a statistical test to determine the validity of the point¹.

Introduction / Discussion

The presence of outliers in a sample data set can have a great effect on subsequent analyses. It is important to determine if a causal relationship exists which may allow for the removal of the point as coming from a unit no longer homogeneous with the remaining units. When the sample size is large enough, most researchers assume normally distributed data and use the 3 sigma rule of thumb.² When the sample size is small to moderate a more specialized procedure is recommended. One such test is Dixon's Test and it is typically employed when the data are limited to the present sample.

Dixon's Test

Dixon (1951) provides a test for such a determination; it is as follows:

1. Order the observations.
 - if the point in question is the smallest, orient the data in ascending order
 - if the point in question is the largest, orient the data in descending order
2. Compute the following statistic:

$$r_{10} = (x_2 - x_1)/(x_n - x_1),$$

where x_1 is the first ordered value, x_2 is the second, and x_n is the nth ordered value.

3. Compare to the values in Table 1
4. If the value computed in step 2, above is larger than the table value, you may be 95% confident that the point is an outlier.

The methodology can be employed recursively to determine if more than one outlier exists.

¹ Outlier testing should be coupled with sound judgment.

² Calculate the mean and 3 standard deviation limits and remove any observation beyond the limit. More on testing for normality in latter submissions.

Example

The results from a Phase I clinical trial are presented below:

Unit No.	Count
1	2097.6
2	1974.1
3	1978.2
4	1975.5
5	1972.8
6	1973.4

Step 1: The questionable point is associated with unit number 1. Following the above procedure, the data are ordered and oriented in decreasing order:

2097.6, 1978.2, 1975.5, 1974.1, 1973.4, 1972.8

Step 2: Compute

$$r_{10} = (1978.2 - 2097.6)/(1972.8 - 2097.6) = -119.4/-124.8 = 0.9567$$

Step 3: An examination of Table 1 finds that for a sample of size 6, the critical value is 0.560.

Step 4: Since the value computed in step 2 is larger than the critical, we are 95% confident that the value is an outlier and can remove it from the data set.

Sample Size	Alpha = 0.05
3	0.941
4	0.765
5	0.642
6	0.560
7	0.507
8	0.554
9	0.512
10	0.477
11	0.576
12	0.546
13	0.521
14	0.546
15	0.525
16	0.507
17	0.490
18	0.475
19	0.462
20	0.450

Table 1

Reference:

Dixon W.J. (1951) "Ratios involving extreme values," *Annals of Mathematical Statistics*