P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.

P. McCullagh. *Celebrating Statistics: Papers in honour of Sir David Cox on his 80th birthday*, chapter Exchangeability and regression models, pages 89–113. Oxford, 2005.

P. McCullagh and D. Clifford. Evidence for conformal invariance of crop yields. Accepted at *Proceedings of the Royal Society A*, 2006.

J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. New York : Springer-Verlag, 2000.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer-Verlag New York, Inc., 4th edition, 2002.

G. Wahba. *Spline Models for Observational Data*. SIAM, Philadephia, 1990.

F. Yates. Complex experiments. *Journal of the Royal Statistical Society (Supplement)*, 2:181–247, 1935.

*David Clifford, CSIRO*
*Peter McCullagh, University of Chicago*
David.Clifford@csiro.au
pmcc@galton.uchicago.edu

# Processing data for outliers

*by Lukasz Komsta*

The results obtained by repeating the same measurement several times can be treated as a sample coming from an infinite, most often normally distributed population.

In many cases, for example quantitative chemical analysis, there is no possibility to repeat measurement many times due to very high costs of such validation. Therefore, all estimates and parameters of an experiment must be obtained from a small sample.

Some repeats of an experiment can be biased by crude error, resulting in values which do not match the other data. Such values, called outliers, are very easy to be identified in large samples. But in small samples, often less than 10 values, identifying outliers is more difficult, but even more important. A small sample contaminated with outlying values will result in estimates significantly different from real parameters of whole population (Barnett, 1994).

The problem of identifying outliers in small samples properly (and making a proper decision about removing or leaving suspicious data) is very old and the first papers discussing this problem were published in the 1920s. The problem remained unresolved until 1950s, due to lack of computing technology to perform valid Monte-Carlo simulations. Although current trends in outlier detection rely on robust estimates, the tests described below are still in use in many cases (especially chemical analysis) due to their simplicity.

## Dixon test

All concepts of outlier analysis were collected by Dixon (1950) :

1. chi-squared scores of data
2. score of extreme value
3. ratio of range to standard deviation (two opposite outliers)
4. ratio of variances without suspicious value(s) and with them
5. ratio of ranges and subranges

The last concept seemed to the author to have the best performance and Dixon introduced his famous test in the next part of the paper. He defined several coefficients which must be calculated on an **ordered** sample $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. The formulas below show these coefficients in two variants for each of them - when the suspicious value is lowest and highest.

$$r_{10} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \quad (1)$$

$$r_{11} = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \quad (2)$$

$$r_{12} = \frac{x_{(2)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(3)}} \quad (3)$$

$$r_{20} = \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} \quad (4)$$

$$r_{21} = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \quad (5)$$

$$r_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}} \quad (6)$$

The critical values of the above statistics were not given in the original paper, but only discussed. A year later (Dixon, 1951) the next paper with critical values appeared. Dixon discussed the distribution of his variables in a very comprehensive way, and concluded that a numerical procedure to obtain critical values is available only for sample with $n = 3$. Quantiles for other sample sizes were obtained in this paper by simulations and tabularized.

Dixon did not publish the 0.975 quantile for his test, which is needed to perform two-sided test at 95% confidence level. These values were obtained by interpolation later (Rorabacher, 1991) with some corrections of typographical errors in the original Dixon table.

## Grubbs test

Another criteria for outlying observations proposed by Dixon were discussed at the same time by Grubbs (1950). He proposed three variants of his test. For one outlier

$$G = \frac{|x_{outlying} - \bar{x}|}{s}, U = \frac{S_1^2}{S^2} \tag{7}$$

For two outliers on opposite tails

$$G = \frac{x_{max} - x_{min}}{s}, U = \frac{S_2^2}{S^2} \tag{8}$$

For two outliers on the same tail

$$U = G = \frac{S_2^2}{S^2} \tag{9}$$

where $S_1^2$ is the variance of the sample with one suspicious value excluded, and $S_2^2$ is the variance with two values excluded.

If the estimators in equation (7) are biased (with $n$ in denominator), then simple dependence occurs between $S$ and $U$:

$$U = 1 - \frac{1}{n-1}G^2 \tag{10}$$

and it makes $U$ and $G$ statistics equivalent in their testing power.

The $G$ distribution for one outlier test was discussed earlier (Pearson and Chekar, 1936) and the following formula can be used to approximate the critical value:

$$G = t_{\alpha/n,n-2}\sqrt{\frac{n-1}{n-2+t^2_{\alpha/n,n-2}}} \tag{11}$$

When discussing (8) statistics, Grubbs gives only $G$ values, because $U$ was too complicated to calculate. He writes *"…the limits of integration do not turn out to be functions of single variables and the task of computing the resulting multiple integral may be rather difficult."*

The simple formula for approximating critical values of $G$ value (range to standard deviation) was given by David, Hartley and Pearson (1954):

$$G = \sqrt{\frac{2(n-1)t^2_{\alpha/n(n-1),n-2}}{n-2+t^2_{\alpha/n(n-1),n-2}}} \tag{12}$$

The ratio of variances used to test for two outliers on the same tail (9) is not available to integrate in a numerical manner. Thus, critical values were simulated by Grubbs and must be approximated by interpolation from tabularized data.

## Cochran test

The other test used to detect crude errors in experiment is the Cochran test (Snedecor and Cochran, 1980). It is designed to detect outlying (or inlying) variance in a group of datasets. It is based on simple statistics - the ratio of maximum (or minimum) variance to the sum of all variances:

$$C = \frac{S^2_{max}}{\sum S^2} \tag{13}$$

The critical value approximation can be obtained using the following formula:

$$C = \frac{1}{1+(k-1)F_{\alpha/k,n(k-1),n}} \tag{14}$$

where $k$ is the number of groups, and $n$ is the number of observations in each group. If groups have different length, the arithmetic mean is used as $n$.

## The R implementation

Package **outliers**, available now on CRAN, contains a set of functions for performing these tests and calculating their critical values. The most difficult problem to solve was introducing a proper and good working algorithm to provide quantile and p-value calculation of tabularized distribution by interpolation.

After some experiments I have decided to use regression of 3rd degree orthogonal polynomial, fitted to four given points, taken from the original Dixon or Grubbs table, nearest to the argument. When an argument comes from outside of tabularized range, it is extrapolated by fitting a linear model to the first (or last) two known values. This method is implemented in the internal function `qtable`, used for calculation of Dixon values and Grubbs values for two outliers at the same tail.

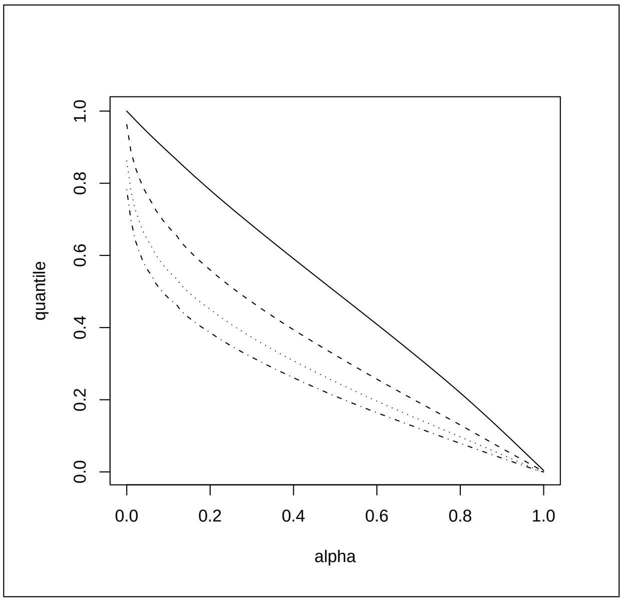

Figure 1: Interpolated quantile curves of Dixon test for $n = 3$ (solid), $n = 4$ (dashed), $n = 5$ (dotted) and $n = 6$ (dotdashed).

The proposed algorithm provides a good continuity-like effect of quantile calculation as shown on Figures (1) and (3). Critical values can be obtained by `qdixon, qgrubbs, qcochran`. The corresponding reverse routines (calculating p-values) are named `pdixon, pgrubbs, pcochran`. The continuity effect of reverse functions is depicted on Figure (2)

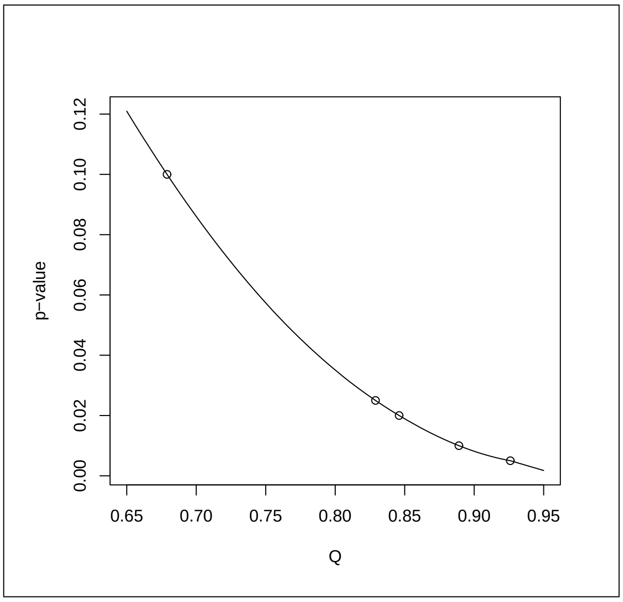

Figure 2: Interpolated p-value curve of Dixon test for $n = 4$, with percentage points given in original Dixon table

The most important functions implemented in the package are: `dixon.test`, `grubbs.test` and `cochran.test`. According to the given parameters, these functions can perform all of the tests mentioned above. Additionally, there is an easy way to obtain different kinds of data scores, by the `scores` function.

It is also possible to obtain the most suspicious value (the largest difference from the mean) by the `outlier` function. If this value is examined and proved to be an outlier, it can be easily removed or replaced by the arithmetic mean by `rm.outlier`.

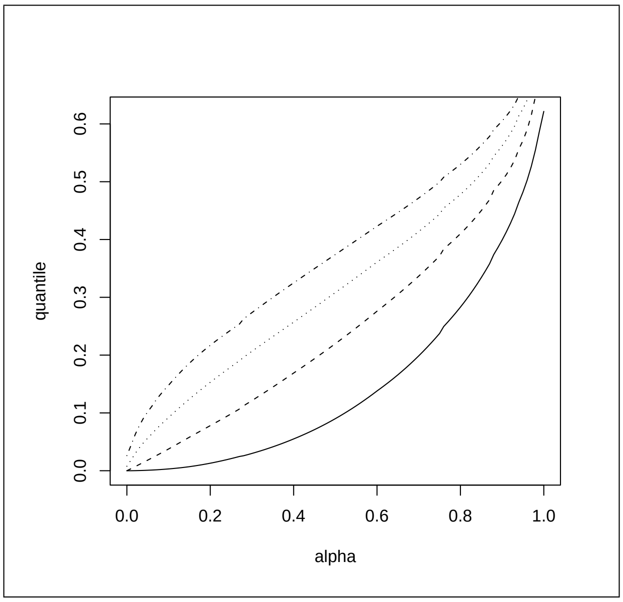

Figure 3: Interpolated quantile curves of Grubbs test for two outliers on one tail; $n = 4$ (solid), $n = 5$ (dashed), $n = 6$ (dotted) and $n = 7$ (dotdashed).

## Examples

Suppose we have six measurements, where the last one is visually larger than the others. We can now test, if it should be treated as an outlier, and removed from further calculations.

```
> x = c(56.5,55.1,57.2,55.3,57.4,60.5)
> dixon.test(x)

        Dixon test for outliers

data:  x
Q = 0.5741, p-value = 0.08689
alternative hypothesis: highest value 60.5 is an outlier

> grubbs.test(x)

        Grubbs test for one outlier

data:  x
G = 1.7861, U = 0.2344, p-value = 0.06738
alternative hypothesis: highest value 60.5 is an outlier

> scores(x) # this is only example, not recommended for
              small sample
[1] -0.2551552 -0.9695897  0.1020621 -0.8675276  0.2041241
[6] 1.7860863
> scores(x,prob=0.95)
[1] FALSE FALSE FALSE FALSE FALSE  TRUE
> scores(x,prob=0.975)
[1] FALSE FALSE FALSE FALSE FALSE FALSE
>
```

As we see, both tests did not reject the null hypothesis and the suspicious value should not be removed. Further calculations should be done on the full sample.

Another example is testing a simple dataset for two opposite outliers, and two outliers at one tail:

```
> x = c(41.3,44.5,44.7,45.9,46.8,49.1)
> grubbs.test(x,type=11)

        Grubbs test for two opposite outliers

data:  x
G = 2.9908, U = 0.1025, p-value = 0.06497
alternative hypothesis: 41.3 and 49.1 are outliers

> x = c(45.1,45.9,46.1,46.2,49.1,49.2)
> grubbs.test(x,type=20)

        Grubbs test for two outliers

data:  x
U = 0.0482, p-value = 0.03984
alternative hypothesis: highest values 49.1 , 49.2 are outliers

>
```

In the first dataset, the smallest and greatest value should not be rejected. The second example rejects the null hypothesis: 49.1 and 49.2 are outliers, and calculations should be made without them.

The last example is testing for outlying variance. We have calculated variance in 8 groups (5 measurements in each group) of results and obtained: 1.2, 2.5, 2.9, 3.5, 3.6, 3.9, 4.0, 7.9. We must check now if the smallest or largest variance should be considered in calculations:

```
> v = c(1.2, 2.5, 2.9, 3.5, 3.6, 3.9, 4.0, 7.9)
> n = rep(5,8)
> cochran.test(v,n)

        Cochran test for outlying variance

data:  v
C = 0.2678, df = 5, k = 8, p-value = 0.3579
alternative hypothesis: Group 8 has outlying variance
sample estimates:
  1   2   3   4   5   6   7   8
1.2 2.5 2.9 3.5 3.6 3.9 4.0 7.9

> cochran.test(v,n,inlying=TRUE)

        Cochran test for inlying variance

data:  v
C = 0.0407, df = 5, k = 8, p-value < 2.2e-16
alternative hypothesis: Group 1 has inlying variance
sample estimates:
  1   2   3   4   5   6   7   8
1.2 2.5 2.9 3.5 3.6 3.9 4.0 7.9

>
```

The tests show that first group has inlying variance, significantly smaller than the others.

## Bibliography

V. Barnett, T. Lewis. *Outliers in statistical data.* Wiley.

W.J. Dixon. Analysis of extreme values. *Ann. Math. Stat.*, 21(4):488-506, 1950.

W.J. Dixon. Ratios involving extreme values. *Ann. Math. Stat.*, 22(1):68-78, 1951.

D.B. Rorabacher. Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon Q Parameter and Related Subrange Ratios at the 95 percent Confidence Level. *Anal. Chem.* , 83(2):139-146, 1991.

F.E. Grubbs. Sample Criteria for testing outlying observations. *Ann. Math. Stat.* , 21(1):27-58, 1950.

E.S. Pearson, C.C. Chekar. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika* , 28(3):308-320, 1936.

H.A. David, H.O. Hartley, E.S. Pearson. The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika* , 41(3):482-493, 1954.

G.W. Snedecor, W.G. Cochran. Statistical Methods. *Iowa State University Press* , 1980.

*Lukasz Komsta*
*Department of Medicinal Chemistry*
*Skubiszewski Medical University of Lublin, Poland*
luke@novum.am.lublin.pl

# Analysing equity portfolios in R

**Using the portfolio package**

*by David Kane and Jeff Enos*

## Introduction

R is used by major financial institutions around the world to manage billions of dollars in equity (stock) portfolios. Unfortunately, there is no open source R package for facilitating this task. The **portfolio** package is meant to fill that gap. Using **portfolio**, an analyst can create an object of class `portfolioBasic` (weights only) or `portfolio` (weights and shares), examine its *exposures* to various factors, calculate its *performance* over time, and determine the *contributions* to performance from various categories of stocks. Exposures, performance and contributions are the basic building blocks of portfolio analysis.