Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 69 (2006) 1068–1071
Talanta
www.elsevier.com/locate/talanta

Short communication

Estimation of type I error probability from experimental Dixon's "Q" parameter on testing for outliers within small size data sets

Constantinos E. Efstathiou *

Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, University Campus, Athens 15771, Greece

Received 18 September 2005; received in revised form 11 December 2005; accepted 15 December 2005
Available online 19 January 2006

Abstract

Common significance tests carried out using statistical software packages usually return to the user the probability p of type I error as the result. Based on p and the preset confidence level the user will decide on the acceptance or the rejection of the associated null hypothesis. Dixon's test (Q-test) is commonly used for the detection of an outlier within a set of N observations (typically: $N = 3$–12). Q-test can only be applied by comparing the experimental value of the statistic Q with tabulated critical Q-values corresponding to some standard values of p. Hence, for a given value of Q and a number of observations, N, the user knows only the range and not the value of the associated probability p of type I error (erroneous rejection). This is due to the lack of explicit expressions of the form $p = F(Q,N)$. In this work, a simple stochastic (Monte Carlo) approach is presented for the estimation of p corresponding to a given experimental value of Q and size N of the data set. In addition, based on Dixon's equations, explicit expressions of p are given for $N = 3$ and 4.

Keywords: Outliers; Q-test; Stochastic methods

1. Introduction

Significance tests are widely used in the statistical evaluation of experimental results and aim at an objective decision about the acceptance or rejection of the associated null hypothesis. Traditionally, these tests are carried out by calculating the experimental value of the corresponding statistic parameter followed by its comparison with a critical value located at the appropriate position of the relevant statistical table. The tabulated critical values have been calculated for several standard probabilities p of type I error (or α-risk), i.e. an erroneous rejection of the null hypothesis. Most tables found in analytical chemistry textbooks for the most commonly used statistical tests (e.g. Student's t-test, F-test, χ^2-test) contain critical values of the associated statistic for $p = 0.1$, 0.05, and 0.01, corresponding to confidence levels (CLs) of 90, 95 and 99%, since $\mathrm{CL} = (1 - p) \times 100$. A well-established and recommended practice, at least among analytical chemists, is to apply these tests on a CL of 95%, allowing thus a probability of an erroneous rejection of the null hypothesis not larger than 0.05. Further increase of the test CL, while diminishing the probability of erroneous rejection of the associated null hypothesis, increases perilously the probability of a type II error (or β-risk), i.e. an erroneous acceptance of the null hypothesis. More detailed tables containing critical values spread over a wider range of CLs (e.g. 80–99.9%) can be found in statistical treatises.

The extensive use of statistical software packages for performing significance tests have made obsolete the use of tabulated critical values; the usual outcome of these tests is directly the numerical value of the probability p, and the user then has to decide about the rejection or the acceptance of the corresponding null hypothesis. For example, a computer response of $p = 0.062$, indicates that the null hypothesis can be comfortably rejected at a CL = 90%, but it must be retained at CL = 95%. Similarly, a response of $p = 0.0002$, indicates that it is practically certain that the null hypothesis is invalid, whereas a test performed by using tabulated values would merely show that the null hypothesis has to be rejected at a CL of 99% if critical values at CLs higher than 99% are not available. Obviously, knowledge of p provides more information than a clear-cut decision of the "reject/accept" type using tabulated critical values.

* Tel.: +30 210 7274312; fax: +30 210 7274750.
E-mail address: cefstath@chem.uoa.gr.

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.12.031

Generally, the software calculates the aforementioned p-values using explicit expressions, i.e. of the type $p = F(S,N)$ (S: experimental value of the actual statistic, N: size of data set), of highly variable complexity, embedded in the corresponding statistical software package. Alternatively, in case of non-existing explicit expressions, approximation functions calculated by numerical techniques are used. In the latter case the output should normally be rounded to the appropriate number of significant figures reflecting the accuracy provided by the approximation function.

The detection of grossly deviant values (outliers) in data sets, accompanied by either rejection or accommodation is of utmost importance in measurements. Unconditional inclusion of outliers can distort the sought-for information, either this being the mean value and/or the standard deviation, or the model describing these data. General techniques for the rejection of outliers include α-trimming (α represents a fixed fraction of lower and upper data set values to be discarded); accommodation of them is usually performed after minimization of their influence, e.g. by employing data Winsorization [1]. Most Exploratory Data Analysis software packages include several graphical diagnostics showing how far apart and how evenly the data are distributed, such as the Box-and-Whisker graph, the QQ plot, the symmetry plot, the jittered dot plot, and the quantile-box plot. It should be stressed, however, that these techniques usually label as outliers all observations that fall outside of limits defined by the spreading of observations without providing an objective estimation of the probability of an erroneous rejection. Furthermore, their applicability and reliability diminishes as the size of the data set is reduced, e.g. to a single figure number, which is often the case when data represent hard-to-obtain replicated analytical results.

It seems that Dixon's Q-test, despite its numerous shortcomings and limitations can still be invoked for an objective "de-contamination" of small size data sets and most analytical chemistry textbooks have settled on its use. Historically, this test was an outcome of Dixon's thorough studies on the distribution density functions of various "subrange ratios" of ordered-according-to-value data sets drawn from populations of various distributions [2–4]. The popularity of Q-test over other more versatile outlier tests for deviant values can also be attributed to its simplicity [5]. A normal (Gaussian) distribution of data is assumed as in most significance tests and the test can be applied only once in a particular set of observations. This test is based on the calculation of the experimental Q-value (Q or r_{10} in Dixon's notation) defined as the ratio given by the distance of the suspect value from its nearest neighbor divided by the range of the values. If N is the sample size (i.e. the number of observations or replicates), the corresponding N-values are arranged in ascending order: $x_1 < x_2 < \ldots < x_N$, then for testing the smallest value (x_1) or the largest value (x_N) the following equations are used, respectively,

$$Q = \frac{x_2 - x_1}{x_N - x_1} \quad \text{or} \quad Q = \frac{x_N - x_{N-1}}{x_N - x_1} \tag{1}$$

If the obtained (experimental) Q-value exceeds the tabulated critical Q-test Q_{crit}-value for a given CL, e.g. 95%, then the suspect value can be rejected with a probability of erroneous rejection (type I error) of 5% or less (i.e. $p \leq 0.05$).

Unfortunately, the mathematical calculations used for the determination of Q_{crit}-values are arduous. Dixon gave a general multi-integral form of the equation providing the cumulative density function (cdf) of the distribution of Q for any distribution function $f(x)$ of data. For a normal distribution the evaluation of the multi-integral form is highly complicated, and the obtained analytical forms vary with each sample size N. Apart from the case of $N = 3$, no explicit expressions of the form $Q = F(p)_N$ can be obtained. In addition, Dixon gave an exact equation for $N = 4$ that cannot be solved to provide an explicit expression of Q but it can be readily solved for Q by using an iterative numerical technique. Finally, Dixon had to resort to tedious numerical integrations, using the multi-integral general form, for the calculation of Q_{crit}-values for $N = 5$, 7, 10, 15, 20, 25, 30 and interpolations for other values of N [3].

In his authoritative paper, Rorabacher discussed the overall issue of the rejection of deviant values. He performed a number of numerical corrections on some of the tabulated Q_{crit}-values, and by applying cubic regression interpolation on Dixon's critical values for various CLs, he calculated the set of Q_{crit}-values at CL = 95% which was missing from Dixon's original tables [6]. The corrected by Rorabacher Q_{crit}-values (for $N = 3, 4, \ldots 30$ and $p = 0.20, 0.10, 0.05, 0.04, 0.02$ and 0.01) are given with three decimal points, and they are generally believed to be accurate to within ± 0.002.

2. The stochastic approaches

In the past we have shown that a simple, non-deterministic or "stochastic" approach (i.e. based on the use of random numbers) can be used for the estimation of the critical values of Q-test [7]. In fact, the described approach is much more general, a kind of a "Gordian knot cut" solution, and it can be used for the estimation of critical values of any statistic used in significance tests, at any confidence level, no-matter how complex is the mathematics involved in the deterministic calculation, or even if this mathematics does not actually exist. Briefly, this estimation is based on the following algorithm: (i) a large as possible number N_{sim} (: number of simulations) of Q-values are collected from data sets of given size N randomly drawn from a normally distributed population; (ii) all N_{sim} Q-values are sorted in ascending order, i.e. $Q_1 < Q_2 < \ldots < Q_{N_{\text{sim}}}$; (iii) using the array of the sorted values, the value Q_L is taken as the critical value corresponding to a probability p of type I error, where L is the nearest integer to the product $(1 - p) \times N_{\text{sim}}$.

The reasoning behind this simple algorithm is that all N_{sim}-values of Q have been calculated from data sets drawn from the same population, i.e. they are all legitimate and "outliers-free". Therefore, the rejection of extreme observations with $Q > Q_L$, constitutes a type I error, with a probability of occurrence less or equal to p, hence Q_L acts as the critical value corresponding to p.

Obviously, the larger the N_{sim}-value the more accurate the critical value would be. Most computer languages can accommodate easily manipulated static arrays of real numbers of size

usually limited to few tens of thousands. This fact does not limit the accuracy of the calculation, since the calculation can be repeated as many times as necessary to obtain a mean critical value (along its standard deviation) with the required accuracy. The only price to be paid for improving the accuracy of the obtained critical values is computation time. The accuracy of the obtained results is also strongly dependent on the quality of the normally distributed (pseudo-)random number generator. Procedures for generating normally distributed pseudo-random numbers are included in many libraries of high-level programming languages, whereas readily implemented procedures have been described in literature [8,9].

Based on this stochastic approach, Muranaka has made available general Fortran programs (QTEST, QTEST2) for the computation of critical values for the typical Q-test along with critical values for a variety of other subranges for both normal and log-normal populations [10,11].

Plots of the ratio i/N_{sim} versus $Q_{,i}$ ($i = 1,2, \ldots N_{\text{sim}}$) are effectively the plots of the corresponding cumulative density functions (cdf) of the distributions of the Q-values. These cdf curves for sample sizes $N = 3, 4, \ldots 12$ are shown in Fig. 1a. The parallel to $Q_{,i}$ axis dashed line at $i/N_{\text{sim}} = 0.95$ crosses all cdf curves at points corresponding to the respective critical values at $p = 0.05$ (or CL = 95%). The most useful part of these plots is that at cdf ≥ 0.8 corresponding to $p \leq 0.20$ (or CL $\geq$ 80%). This part is shown in expanded and rearranged form in Fig. 1b, as a plot of Q versus p (the latter in logarithmic scale). By using this diagram, one can obtain a rough estimate of p for a given experimental Q-value and for a data set of given size ($N = 3, 4, \ldots, 12$).

Apart from using the cdf plots of Q-values for the graphical estimation of p, an even more simple stochastic approach can be used for this purpose. This estimation is based on the following algorithm: (i) One Q-value is determined from data sets (of given size N) randomly drawn from the same normally distributed population; (ii) if this Q-value is larger than the input (under examination) Q-value then a counter C is incremented; (iii) steps i and ii are repeated N_{sim} times; (iv) the ratio C/N_{sim} is taken as the estimate of p.

Obviously, a large as possible N_{sim} is required to obtain an accurate estimate of p. It should be noted that storing and sorting of the generated Q-values are not needed in this case. Typically, for $N_{\text{sim}} = 10^6$ the obtained p-values for $p \leq 0.1$ are accurate to ± 0.001, whereas a relative accuracy of 1–10% is expected for p-values in the range 0.01–0.0001.

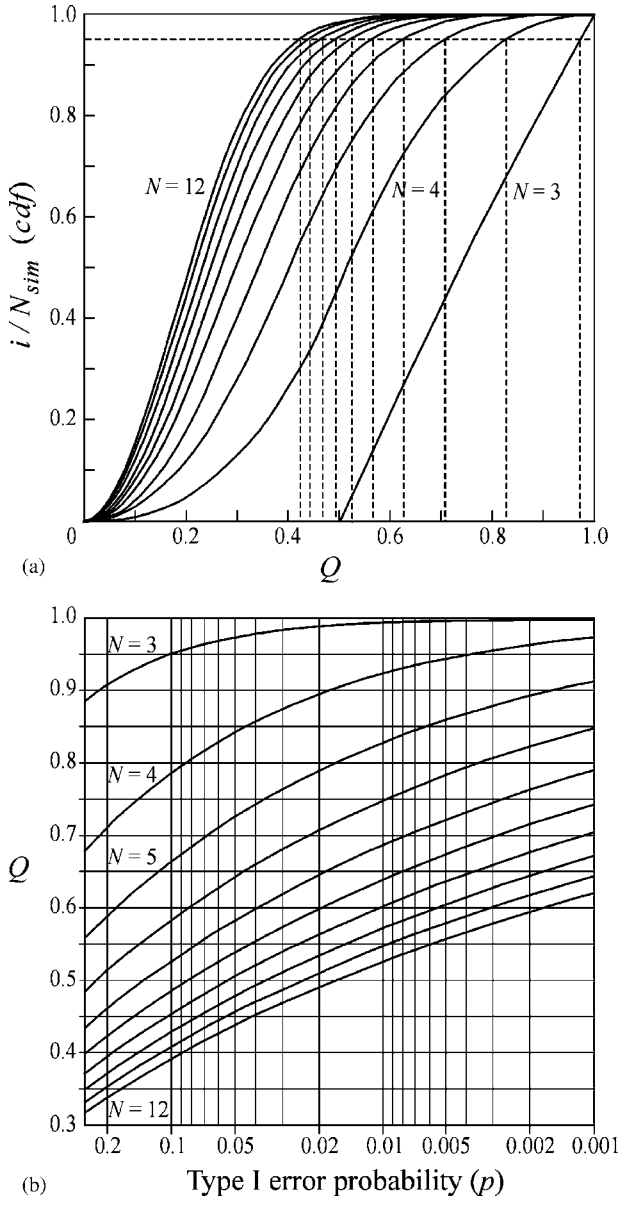

Fig. 1. (a) Cumulative distribution function diagrams of Q-values for $N = 3, 4 \ldots 12$; (b) semilogarithmic plot of Q versus p corresponding to the upper part (cdf ≥ 0.8) of diagram (a).

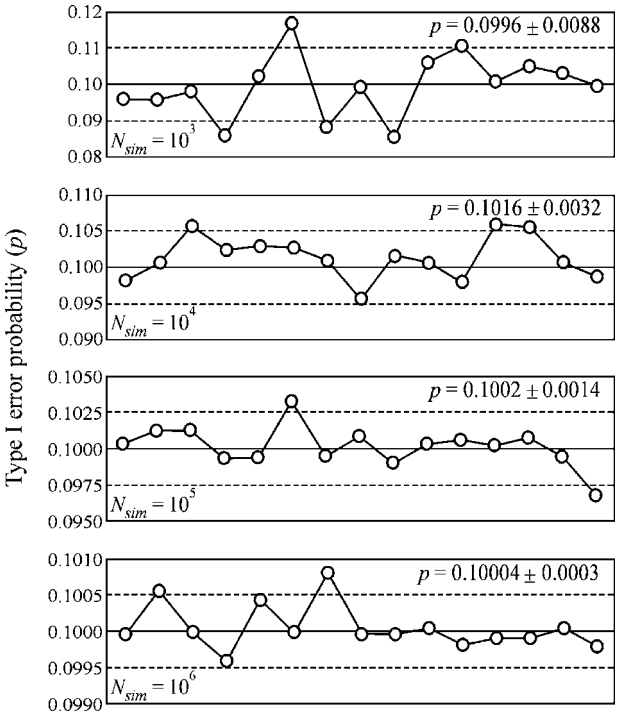

Fig. 2. Scatterplots of consecutively estimated p-values by the stochastic algorithm for $Q = 0.765533$ for data set size $N = 4$ (theoretically expected value: $p = 0.10000$), for $N_{\text{sim}} = 10^3$–10^6.

As in the case of the stochastic calculation of Q_{crit}-values, computation time is the only consideration. Using a Pentium-4 (at 1.60 GHz) personal computer and a program written in Borland's Delphi running under Windows XP, the time required to calculate a single p-value for a data set size N and by performing N_{sim} simulations, is approximated by the following equation: time (s) = $4.6 \times 10^{-7} \times N \times N_{sim}$.

3. Typical examples

Dixon obtained exact equations associating the cdf of Q-values with p-values (originally for one-tailed tests) only for data sets of size $N = 3$ or 4. By rearranging his equations and considering that a critical value corresponding to a probability p for a one-tailed test corresponds also to a probability $2 \times p$ for a two-tailed test, the following explicit forms of equations are obtained

$$p = 1 - \frac{6}{\pi}\arctan\left(\frac{2Q-1}{\sqrt{3}}\right) \quad (\text{for } N = 3) \tag{2}$$

$$p = -8 + \frac{12}{\pi}\left[\arctan(\sqrt{4Q^2 - 4Q + 3}) + \arctan\left(\frac{\sqrt{3Q^2 - 4Q + 4}}{Q}\right)\right] \quad (\text{for } N = 4) \tag{3}$$

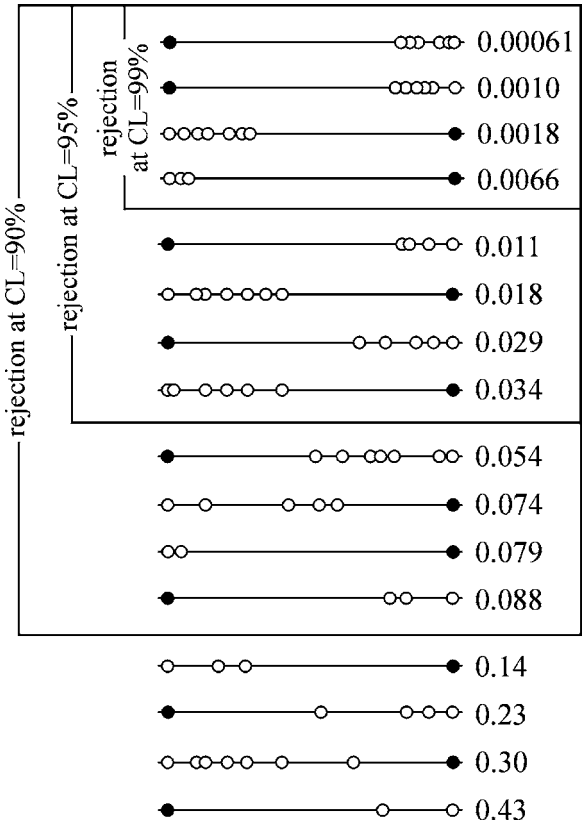

Fig. 3. Dotplots of data sets with a single suspect value (black dot). The corresponding probability p of type I error as it was generated by the stochastic procedure is shown at the right end of each dotplot.

The accuracy and the precision of the stochastic procedure can be tested by comparing the generated values of p with those obtained by using Eq. (2) or Eq. (3). For $Q = 0.765533$, Eq. (3) returns the value $p = 0.10000$. Consecutive results obtained by the stochastic procedure using this Q-value and $N = 4$ as input parameters (for N_{sim}: 10^3–10^6) are shown as scatterplots in Fig. 2.

As it was expected, the standard deviation (s_p) of the generated p-values appears practically proportional to $(N_{sim})^{-1/2}$ and more specifically is given by the following equation: $s_p = 0.27\,(N_{sim})^{-1/2} + 0.0003$ ($r = 0.998$).

Indicative examples of observations sets of size $N = 3$–8 with suspect values identified either as outliers at various confidence levels or as "legitimate" observations are shown as dotplots in Fig. 3. For each observations set, apart from the CL-based "reject/accept" type decision, the probability p of type I error is also given.

4. Conclusions

Although the use of Q-test is increasingly discouraged in favor of other robust methods, which also accommodate suspect values and take into account all data values (such as Huber method [12]), it remains a fact that Q-test is the simplest test for the objective rejection or acceptance of a grossly deviant value within a small set of observations. One of the shortcomings associated with its use was that it has to be carried out using tabulated critical values and no estimates of the probability p of an erroneous rejection could be obtained. This drawback can be overcome by using the explicit functions provided for $N = 3$ and 4, whereas the simple and accurate enough stochastic approach described here can be used in general. This approach can be readily extended for use on the other variants of Dixon's tests. A small utility program (Q_STOCH) running on the Microsoft Windows platform for the stochastic estimation of Q_{crit}-values for any p and for the estimation of p from any experimental Q-value can be sent by the author to any interested reader on request.

References

[1] V. Barnett, T. Lewis, Outliers in Statistical Data, third ed., John Wiley & Sons, Cichester, 1994, p. 41.
[2] W.J. Dixon, Ann. Math. Stat. 21 (1950) 488.
[3] W.J. Dixon, Ann. Math. Stat. 22 (1951) 68.
[4] R.B. Dean, W.J. Dixon, Anal. Chem. 23 (1951) 636.
[5] P.C. Meier, R.E. Zünd, Statistical Methods in Analytical Chemistry, John Wiley & Sons Inc., New York, 1993, p. 49.
[6] D.B. Rorabacher, Anal. Chem. 63 (1991) 139.
[7] C.E. Efstathiou, J. Chem. Educ. 69 (1992) 733.
[8] B.S. Gottfried, Programming with BASIC, Schaum's Outline Series, McGraw-Hill, New York, 1975, p. 229.
[9] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Verrerling, Numerical Recipes: The Art of Scientific Computing, Cambridge University, Cambridge, 1986, p. 202.
[10] K. Muranaka, QCPE Bull. 18 (1998) QCM 178.
[11] K. Muranaka, J. Chem. Educ. 76 (1999) 469.
[12] AMC Technical Briefs, 2001. No. 6, www.rsc.org/images/brief6_tcm18-25948.pdf.