Martin–Luther–Universität
Halle–Wittenberg
Institut für Mathematik

Outliers in Uniform Distribution

Sadaf Manzoor and Salahuddin

Report No. 08 (2009)

Editors:
Professors of the Institute for Mathematics, Martin-Luther-University Halle-Wittenberg.

Electronic version: see `http://www2.mathematik.uni-halle.de/institut/reports/`

Outliers in Uniform Distribution

Sadaf Manzoor and Salahuddin

Report No. 08 (2009)

Sadaf Manzoor
Salahuddin
University of Peshawar
Department of Statistics
25120 Peshawar
Pakistan
Email: sadaf505@yahoo.com
Email: salahuddin_90@hotmail.com

Outliers in Uniform Distribution

Sadaf Manzoor[1], Salahuddin[1]

(sadaf505@yahoo.com)

(salahuddin_90@hotmail.com)

April 23, 2009

Abstract

Observations which deviate strongly from the main part of the data, usually lebeled as 'outliers' are troublesome and may cause completely misleading results. Therefore, the solitariness of outliers is important for quality assurance. It is desirable to derive such approaches in a systematic manner from general principles and guidlines rather than human decision making or simply plotting the data.

A test for outliers of normally distributed data has been developed by J. W. Dixon [1]. The present study uses the same idea by arranging the data set in ascending order for the particular case where samples are comming out of Uniform distribution. Percentage points are tabulated for testing hypothesis and constructing confidence intervals for different significance levels.

1 Introduction

Outlier refers to an observation which appears to be inconsistent with the rest of data, relative to an assumed model. Such extreme observations may be reflecting some abnormality in the measured characteristic of a subject, or they may result from an error in the measurement or recording.

[1]Department of Statistics, University of Peshawar, 25120 Peshawar, Pakistan.

The observation that has a disproportionate influence on one or more aspects of estimate is more precisely called influential observation rather than an outlier.

There has been much debate in the literature regarding what to do with these extreme data points. Whether the decision is removal of outliers or some adjustment procedures are adopted, the first and basic step needed is its detection.

A very common statistical problem is that of comparing two samples and determining whether or not there is a significant difference between them. Procedures are available for comparing both paired and unpaired samples e.g $t-test$, $F-test$ and non-parametric signed rank test.

Iglewicz and Hoaglin provide a comprehensive test about labeling, accomodation and identification of outliers. There are various other approaches to single and multiple outlier detection depending on the application and number of observations in the data set, out of which Dixon's test is the simpler test which allows us to examine its least and the most outcomes.

Dixon(1950) [1] has proposed a different class of criteria, according to which the sample values $x_1, x_2, \cdots x_n$ are arranged in ascending order such as $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Based on the sample size, Dixon(1951) [2] has designed range ratios of the form $(x_n - x_{n-j})/(x_n - x_i)$ for small values of i and j and $n = 3, \cdots, 30$ taken from Gaussian distribution. These functions were designed for testing the consistency of suspected values with the sample as a whole.

The same distributions of Dixon's statistics for the case of sample from an exponential population are found by Likes and Praha(1967) [4].

Verma and Quiroz(2006) [8] have mentioned that although these tests were designed more then 50 years ago but still widely used in science and enginnering. They have reported the simulation procedure for Dixon's statistic along with new precise critical values for normal samples. Here we have applied the idea of Dixon in the case of standard uniform distribution.

2 Dixon's Statistic for Uniform Distribution

It is well known that an arbitrary random variable Y is uniformly distributed if it is having constant probability over an interval. Specifically given by

$$g(y) = \frac{1}{a-b} \qquad for \qquad a < y < b \tag{1}$$

Consequently it holds for the distribution function,

$$G(y) = \begin{cases} 0 & \text{if } y \leq a \\ \frac{y-a}{b-a} & \text{if } y \in]a, b[\\ 1 & \text{if } y > b \end{cases} \tag{2}$$

The Uniform distribution defines 'equal probability' over a given range for a continuous distribution. The main interest of this study is to consider the special case of Uniform distribution (with $a = 0$ and $b = 1$) known as standard uniform distribution, i.e.

$$f(y) = \begin{cases} 1 & : \quad y \notin (0,1) \\ 0 & : \quad y \in (0,1) \end{cases} \tag{3}$$

Let X be an arbitrary continuous random variable with density function f. We consider an ordered sample $(x_1, x_2, \ldots, x_n)$ for X , that is $x_1, x_2, \ldots, x_n$ are realizations of independent and identical random variables which are arranged in ascending order as $x_1 < x_2 < \ldots < x_n$ and the corresponding random variables are denoted by $X_1, X_2, \ldots, X_n$. The joint density for $x_1 < x_2 < \ldots < x_n$ is given by

$$f(x_1, x_2, \ldots, x_n) = n! f_X(x_1) f_X(x_2) \cdot \ldots \cdot f_X(x_n) \tag{4}$$

To obtain the density for x_1, x_{n-1} and x_n integrating all other terms except these three.

$$f(x_1, x_{n-1}, x_n) = n! \int_{x_1}^{x_3} \int_{x_1}^{x_4} \cdots \int_{x_1}^{x_{n-1}} f(x_1) f(x_2) \cdot \ldots \cdot f(x_n)\, dx_2 dx_3 \cdots dx_{n-2}.$$

$$= n! f(x_1) f(x_{n-1}) f(x_n) \int_{x_1}^{x_3} \int_{x_1}^{x_4} \cdots \int_{x_1}^{x_{n-1}} f(x_2) \cdot \ldots \cdot f(x_{n-2})\, dx_2 dx_3 \cdots dx_{n-2}.$$

Final result where we get the probability density function of x_1, x_{n-1} and x_n is:

$$= \frac{n!}{(n-3)!} f(x_1)\, dx_1 \left[\int_{x_1}^{x_{n-1}} f(t)\, dt \right]^{n-3} f(x_{n-1})\, dx_{n-1} f(x_n)\, dx_n \tag{5}$$

The density of $R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}$ can be obtained by setting $v = x_n - x_1$, $rv = x_n - x_{n-1}$ and $x = x_n$. Then integrating x and v over their corresponding ranges we have

$$f_{R_{10}}(r) = \frac{n!}{(n-3)!} \int_{\infty}^{-\infty} \int_{0}^{\infty} \left[\int_{x-v}^{x-rv} f(t)\, dt \right]^{n-3} f(x-v) f(x-rv) f(x) v\, dv\, dx. \tag{6}$$

For (3) with $v \leq x \leq 1$ we have following

$$\left[\int_{x-v}^{x-rv} f(t)\, dt \right]^{n-3} = (v - rv)^{n-3} = v^{n-3}(1-r)^{n-3} \tag{7}$$

Now applying formula (6) for the case of standard uniform distribution using (7) and $0 < r < 1$, to obtain

$$\begin{aligned} f_{R_{10}}(r) &= \frac{n!}{(n-3)!}(1-r)^{n-3} \int_0^1 \int_0^1 v^{n-2}\, dv\, dx \\ &= \frac{n!}{(n-3)!}(1-r)^{n-3} \int_0^1 v^{n-2}(1-v)\, dv \\ &= \frac{n!}{(n-3)!}(1-r)^{n-3} \beta(n-1, 2) \\ &= \frac{n!}{(n-3)!}(1-r)^{n-3} \frac{(n-2)!}{n!} \end{aligned}$$

and therefore,

$$f_{R_{10}}(r) = (n-2)(1-r)^{n-3}. \tag{8}$$

The distribution function is given by

$$\begin{aligned} F_{R_{10}}(r) &= \int_0^r f_{R_{10}}(s)\, ds \\ &= 1 - (1-r)^{n-2} \end{aligned}$$

for $r \in (0, 1)$ and

$$F_{R_{10}}(r) = \begin{cases} 0 & for \quad r \leq 0 \\ 1 & for \quad r > 1 \end{cases}$$

To obtain the percentage points for R we have to solve the following equation for given α:

$$1-(1-R)^{n-2}=1-\alpha$$

$$R=1-\alpha^{\frac{1}{n-2}} \tag{9}$$

The solutions are denoted by $r_{10}(\alpha, n)$. Some percentage points for different values of α and n are tabulated below. The problem is solved analytically with the help of Maple programming for the above distribution to obtain the critical values.

Table 1: *Percentage values for* r_{10}

$n \| \alpha$	0.005	0.01	0.02	0.05	0.1	0.5
3	0.9950	0.9900	0.9800	0.9500	0.9000	0.5000
4	0.9293	0.9000	0.8586	0.7763	0.6838	0.2929
5	0.8290	0.7845	0.7285	0.6316	0.5358	0.2063
6	0.7340	0.6837	0.6239	0.5271	0.4377	0.1591
7	0.6534	0.6019	0.5427	0.4507	0.3690	0.1295
8	0.5865	0.5358	0.4789	0.3930	0.3187	0.1091
9	0.5309	0.4821	0.4281	0.3481	0.2803	0.0942
10	0.4843	0.4377	0.3867	0.3123	0.2501	0.0829
11	0.4449	0.4005	0.3525	0.2831	0.2257	0.0741
12	0.4113	0.3690	0.3238	0.2589	0.2056	0.0669
13	0.3822	0.3421	0.2993	0.2384	0.1889	0.0611
14	0.3569	0.3187	0.2781	0.2209	0.1745	0.0561
15	0.3347	0.2983	0.2587	0.2058	0.1623	0.0519

3 Applications

3.1 Testing of Hypothesis

In general, any measurement has imperfactions that give rise to an error in the measurement result. A measurement result is complete only when a quantitative statement of its uncertainty is also conducted. Uncertainty is required to decide about results whether is sufficient for its main purpose and for the assurance that it is consistent with other similar results. A

statistical hypothesis test is an algorithm used to evaluate decision procedures for choosing between the alternatives which minimizes certain risks.

Let us apply hypothesis testing technique for our test by supposing that x_n is an extreme value i.e., it appears usually far from the rest of sample and we wish to check the hypothesis that either it is a real outlier or not.

- The n values comprising the set of observations under examination are arranged in ascending order $x_1 < x_2 < \cdots < x_n$.
- The test statistic $r = (x_n - x_{n-1})/(x_n - x_1)$ is calculated.
- This calculated value of r is then compared to the critical value of $r_{10}(1-\alpha)$ which are tabulated for different significance levels (α).
- Then decision will be made by following conclusion rules about x_n.
- If $r \geq r_{10}(1-\alpha)$ reject the hypothesis and conclude that x_n is not an outier.
- If $r < r_{10}(1-\alpha)$ then x_n is an outlier.

3.2 Confidence Intervals

Another measure "Confidence Interval" is relatively more informative than testing procedure. Instead of a single estimate, a confidence interval generates lower and upper boundaries. These interval estimates give an indication of how much uncertainty there is in our estimate. The narrower the interval, the more precise is our estimate. Calculated values of $r_{10}(1-\alpha)$ can be used for the construction of such boundaries.

$$r_{10} \pm r_t \sigma/\sqrt{n}$$

where r_t is taken from table for specific value of n and α, σ is the standard deviation and n is sample size.

3.3 Uniformity

It should be noted that for symmetric population, the distribution of $(x_n - x_{n-1})/(x_n - x_1)$ and $(x_2 - x_{n-1})/(x_n - x_1)$ will be same. The probability

that a uniformly distributed randon variable falls within any interval of fixed length is independent of the location of the interval itself, (but it is dependent on the interval size) so long as the interval is contained in the distribution's support, which is called 'uniformity'.

3.4 General Uniform Case

Let us consider the (1) & (2) in the light of above study where we have

$$g(y) = \frac{1}{b-a} \qquad for \qquad a < y < b$$

$a \neq 0$ and $b \neq 1$ we have two possibilities for above distribution when parameters (a and b) are known we can standerdize the distribution using their values.

3.5 Unknown Parameters

In case where parameters are unknown they can be determined through appropriate techniques and then can be standerdized.

3.5.1 MLE

Maximum Likelihood Estimation Method for parameter estimation is a totally analytic maximization procedure which provide efficient methods for quantifying uncertainty through connfidence bounds. The idea behind MLE method is to determine the parameters that maximize the probability of the sample data. In statistical point of view, this method is considered to be more robust and yields estimators with good statistical properties.

$$f(x) = \frac{1}{b-a}$$

The likelyhood function for $U(a, b)$ is

$$L(x_1, \cdots, x_n/a, b) = (\frac{1}{b-a})^n.$$

To maximize this we must minimize the value of $(b-a)$ (the interval length), yet we must keep all samples with in the range, i.e., $\forall x_i, x_i \in (a, b)$. An MLE for a and b would then be $\hat{a} = min(X_i), \hat{b} = max(X_i)$.

3.5.2 Method of Moments

Parameters can also be estimated through Method of Moments. Although Maximum Likelihood Estimators have higher probability of being close to the quantities to be estimated, Method of Moments is also not negligible due to its simplicity and easy implementation in most statistical software programs. This method generates moments using moment generating function and give solution for parameters of Uniform Distribution as :

$$a = \bar{x} - \sqrt{3}s$$

$$b = \bar{x} + \sqrt{3}s$$

A Considerable question rises here,

Will these estimates be reliable in the presence of outliers? (as $\bar{x}$ and s both are influenced by extreme values). To reply this question we can choose a heuristic approach. Use Median instead of $\bar{x}$ and Median Deviation instead of Standard Deviation can be an open way, as they do not depend on extreme observations.

3.6 Rational Debate

Let X be a random variable with distribution function F , so that F^{-1} exists. If Z is a random variable uniformly distributed then $F^{-1}(Z)$ has the distribution F. Consequently

$$F(X) = P\{F^{-1}(Z) \leq X\}$$

$$F(X) = P\{Z \leq F(X)\}$$

We now assume that the random variable X has values between 0 and 1 with $(X_1, X_2, \cdots, X_n)$ be an ordered sample for X then we get a sample $(F(X_1), F(X_2), \cdots, F(X_n))$. The significance test that X_n is an outlier can be corresponded to the significance test that $F(X_n)$ is outlier. It gives a line of thought to choose the better one out of following.

1. Follow the outlier hypothesis in above sense.

2. Calculate the percentage points of R_{10} for the given density and compare.

Acknowlegdement.

Great appreciation goes to Prof. Dr. Wilfried Grecksch. I highly acknowledge his kindness and support.

References

[1] Dixon, W. J. *Analysis of extreme values.* Annals of Mathematical Statistics, 21, 488–506, (1950).

[2] Dixon, W. J. *Ratio involving extreme values.* Annals of Mathematical Statistics, 22, 68-78, (1951).

[3] Everitt, B. S. The Cambridge Dictionary of Statistics, 2^{nd} Edition, Cambridge University Press, (2003).

[4] Likes, J. *Distribution of Dixon's Statistics in the case of an Exponential Population.* Niedersachsische Staats-und Universitatsbibliothek Göttingen, 46-54, (1967).

[5] Osborne, Jason, W. & Amy, O. *The Power of outliers (and why researchers should always check for them).* Practical Assessment, Research & Evaluation., 9(6), (2004).

[6] Ping, G., Young, J. D. and Wang, Y. X. *Outlier Detection in High Dimension Based on Projection.* Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 (2006).

[7] Steven, W. *A Review of Statistical Outlier Methods.* Pharmaceutical Technology, 1-5, (2006).

[8] Verma, S. P and Quiroz-Ruiz. *Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering.* Revista Mexicana de Ciencias Geologicas, 2, 133-161 2006.

Reports of the Institutes 2009

01-09. Jan Prüss, Vicente Vergara, Rico Zacher, *Well-Posedness and Long-Time Behaviour for the Non-Isothermal Cahn-Hilliard Equation with Memory*

02-09. Siegfried Carl, Patrick Winkert, *General comparison principle for variational-hemivariational inequalities*

03-09. Patrick Winkert, *$L^\infty-$Estimates for Nonlinear Elliptic Neumann Boundary Value Problems*

04-09. Wilma Di Palma, *John Neper's rods: Calculations are boring and tiring! Leon Battista Alberti's Cipher Wheel*

05-09. Stefan Sperlich, *An integration calculus for stochastic processes with stationary increments and spectral density with applications to parabolic Volterra equations*

06-09. Stefan Sperlich, *A Regularity Theory for Stochastic Processes with Stationary Increments and Spectral Density with Applications to Anomalous Diffusion*

07-09. Andreas Löhne, Thomas Riedrich, Christiane Tammer, *Festschrift zum 75. Geburtstag von Prof. Dr. Alfred Göpfert*

Reports are available via WWW: http://www2.mathematik.uni-halle.de/institut/reports/