The effectiveness of robust RMCD control chart as outliers' detector

IOP Conf. Series: Journal of Physics: Conf. Series **943** (2017) 012039 doi:10.1088/1742-6596/943/1/012039

The effectiveness of robust RMCD control chart as outliers' detector

Darmanto and Suci Astutik

Study Program of Statistics, Department of Mathematics, Brawijaya University, IDN

E-mail: darman_stat@ub.ac.id

Abstract. A well-known control chart to monitor a multivariate process is Hotelling's T^2 which its parameters are estimated classically, very sensitive and also marred by masking and swamping of outliers data effect. To overcome these situation, robust estimators are strongly recommended. One of robust estimators is re-weighted minimum covariance determinant (RMCD) which has robust characteristics as same as MCD. In this paper, the effectiveness term is accuracy of the RMCD control chart in detecting outliers as real outliers. In other word, how effectively this control chart can identify and remove masking and swamping effects of outliers. We assessed the effectiveness the robust control chart based on simulation by considering different scenarios: n sample sizes, proportion of outliers, number of p quality characteristics. We found that in some scenarios, this RMCD robust control chart works effectively.

1. Introduction

Application of statistics as a science and analysis data method is used widely in almost all fields. By statistics, following data collecting, the researchers do their data analysis and further sometimes the result is used in decision making. To get valid and correct result of data analysis, it requires in choosing and using statistical method correctly. In addition, the researchers also have to follow right step of data analysis. One of the steps towards obtaining an appropriate data analysis is testing and detecting an outlier data. The outliers may be often assumed as a mistake or error, but in some condition they bring important information about whole data. If this step is ignore, it may lead to wrong model construction, got biased parameter estimation and incorrect result of data analysis. Then, identification of outlaying observations is an important part to modelling and data analysis [1].

Hawkins [2] defines an outlier *as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.* Analogue with Hawkins, Johnson [3] defines an outlier *as an observation in a data set which appears to be inconsistent with the remainder of that set of data*. An outlier is divided into univariate and multivariate. Almost all of the univariate methods in outlier detecting based on the assumption of prior distribution of the data, which is assumed to be identically and independently distributed (i.i.d). In addition, there are inappropriate tests for detecting univariate outliers further assume that the distribution parameters and the kind of expected outliers are have to known [4].

In univariate outlier is detected separately to each by each variable. But, in multivariate cases, outlier should be detected to all variables simultaneously. For many cases multivariate data can not be detected as

outliers when variables do not have correlation. Multivariate outlier detection is could be worked when multivariate analysis is performed, and the interaction among different variables are compared within the class of data [5]. Figure 1 is a simple example which presents data point having two variables, x and y. From the figure can be seen that the lower left observation is clearly a multivariate outlier. If two these variables are measured separately with respect to the spread of values along the x and y axes, observations fall close to the centre of the univariate distribution. In a brief, the test for outliers should take into reason the correlation between the two variables, which in this case appear abnormal.

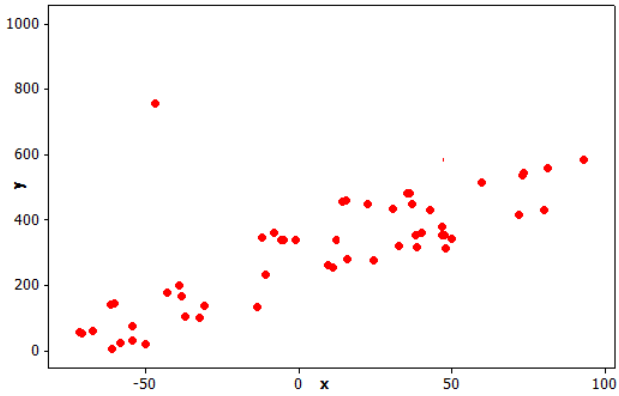


Figure 1. Multivariate Outlier

In many multivariable observations cases, there are data sets with multiple outliers. The multiple outliers are subject to *masking* and *swamping* effects. Even though not mathematically exact, Acuna and Rodriguez [6] give an illustration understanding for these two effects as follows: ***Masking effect***, it is said that one outlier masks a second outlier. If the second outlier can be detected as an outlier, but not in the presence of the first outlier. Thus, if the first outlier is eliminated from data the second instance is appeared as an outlier. ***Swamping effect*****,** it is said that one outlier swamps a second observation, if the next can be detected as an outlier because of the presence of the first one. It means that, after the first outlier is deleted from data sets, the following observation be a normal.

In industrial manufacturing, especially for multivariate process controlling, Hotelling-T^2 control chart is widely used [7]. The control chart is closely-related to multivariate outlier detection method. It assumes the event where the multivariate stream of measures represents a stochastic process, and the require procedure in detecting outlier is online. The Hotelling-T^2 control chart uses Mahalanobis distance method to calculate the statistic of T^2. The formula of Mahalanobis distance given in Eq. (1) below

$$MD_i^2 = \left(\mathbf{X}_i - T\left(\mathbf{X}\right)\right)' C\left(\mathbf{X}\right)^{-1}\left(\mathbf{X}_i - T\left(\mathbf{X}\right)\right) \quad (1)$$

Where $\mathbf{X}_i$ is an m observations random vector, $T(\mathbf{X})$ is a p-dimensional estimated location parameter and $C(\mathbf{X})$ is a $p \times p$ estimated scale parameter. The observation of $\mathbf{X}_i$ follows p-variates Normal distribution with population mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, MD_i^2 has χ^2 distribution by p degree of freedom [8]. Analogue with this formula of *mahalanobis* distance, statistic T^2 in the Hotelling-T^2 control chart we calculate

$$T_i^2 = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \mathbf{S}_X^{-1} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \quad (2)$$

Where $\mathbf{X}_i = \left(X_{i1},\ldots,X_{ip}\right)', i = 1,\ldots,m,$ are the m p-variate observations with the sample mean $\bar{\mathbf{X}} = \sum_{i=1}^{m} \mathbf{X}_i$ and variance-covariance matrix $\mathbf{S} = \frac{\sum_{i=1}^{m}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'}{m-1}$. A large value of T_i^2 (compare to specific distribution: *Beta* or *F* distribution) indicates that the process has shifted which is statistically sentenced by the identification of usual patterns in the control chart [9].

As shown in (2), the Hotelling-T^2 uses the classical sample vector mean and sample variance-covariance matrix to estimate the population vector mean and variance-covariance matrix. However, mean vector and covariance matrix which are estimated classically and used in T^2 are very sensitive to the presence of outliers data. In addition, these estimators also are marred by the effects of masking and swamping of the outliers. In 1998, Sullivan and Woodall [10] showed that the classical estimators are not effective in detecting shifts in the process mean and they proposed a control chart which uses the successive difference of observations in estimating a variance-covariance matrix, but the control chart is not effective in detecting process shift. To overcome the problem, robust estimators are strongly recommended. There are several robust estimators are previously had been proposed. Some of the robust estimators are minimum variance ellipsoid (MVE) and minimum covariance determinant (MCD) which are proposed by Roesseeuw [11], [12]. By developing MCD, Roesseeuw and Van Zomeren [13], Lopuhaa and Rousseew [14], and Willems et.al. [15] proposed re-weighted MCD (RMCD). The RMCD estimators have robustness characteristics as same as MCD estimators such as breakdown point, affine equivariance and asymptotic normality.

The following years, in 2003, Vargas [16] used robust estimators whether MVE or MCD in application of control chart for identifying the outliers. These robust control charts did not use the exact distribution of the classical T^2 control chart, so he estimated the control limits based on simulation. Jensen et al. (2007) also did simulation to estimated the control limits and tabulated these estimation for sample sizes m = 10,…, 100, dimensions p = 2,…, 10 and confidence level $1-\alpha = 0.95$. Chenouri, et.al [9] proposed robust control chart by RMCD in detecting the outliers. They did like Jensen et. al. to estimated the robust estimators by generating 10,000 observations in Phase I and then used it to calculate statistic T_i^2 in Phase II. They also tabulated these estimation for dimensions p = 2,…, 10 and confidence level $1-\alpha = 0.95; 0.99; 0.999$.

Our approach in detecting the outliers and monitoring the multivariate observations differ in three ways of Vargas [16], Jensen et.al. [17] and Chenouri et.al. [9]. We propose to use robust estimators of RMCD in application of the robust control chart. We assessed the effectiveness the robust control chart based on simulation by considering different scenarios: n sample sizes, proportion of outlier, number of p quality characteristics. The effectiveness term is accuracy of the robust control chart in detecting outlier as an outlier. In other word, how effectively this control chart can identify and also remove the masking and swamping effects of outliers.

2. Robust Estimation and Robust Control Chart of RMCD

2.1. Robust Estimation of RMCD

The robustness can be identified based on two properties are affine equivariance and breakdown point. An estimator that has affine equivariance characteristic means the estimator has a stable value even though the data has been transformed and rotated.

Definition: ***Affine Equivariance*** (Olive [18])

Assume that **X** is $m \times p$ observations data matrix. Let again assume that $\mathbf{B} = \mathbf{1}\mathbf{b}'$ where **1** is an $m \times 1$ vector of ones and **b** is $p \times 1$ constant vector then *i*th row of **B** is $\mathbf{b}_i' = \mathbf{b}'$ for *i* = 1, 2, ..., *m*. For such a matrix **B**, consider the affine transformation $\mathbf{Z} = \mathbf{XA} + \mathbf{B}$ where **A** is any non-singular $p \times p$, then the multivariate location and dispersion estimators of $\mathbf{t}$ and $\mathbf{S}$ would be said *affine equivariance* if:

$\mathbf{t(Z) = t(XA + B) = A't(X) + b,}$ and

$\mathbf{S(Z) = S(XA + B) = A'S}\left(\mathbf{X}\right)\mathbf{A.}$

Proposition: If $\left(\mathbf{t},\mathbf{S}\right)$ is *affine equivariance*, then:

$$d_i^2\left(\mathbf{X}\right) \equiv d_i^2\left[\mathbf{t}\left(\mathbf{X}\right),\mathbf{S}\left(\mathbf{X}\right)\right] = d_i^2\left[\mathbf{t}\left(\mathbf{Z}\right),\mathbf{S}\left(\mathbf{Z}\right)\right] \equiv d_i^2\left(\mathbf{Z}\right).$$

Based on the definition and the proposition, will be showed that the *mahalanobis* distance estimated are invariant under affine equivariance. Since $\mathbf{Z} = \mathbf{XA} + \mathbf{B}$ has *i*-th row, $\mathbf{z}_i' = \mathbf{x}_i'\mathbf{A} + \mathbf{b}'$, then

$$\begin{aligned}
d_i^2\left(\mathbf{Z}\right) &= \left[\mathbf{z}_i' - \mathbf{t}\left(\mathbf{Z}\right)\right]' \mathbf{S}^{-1}\left(\mathbf{Z}\right)\left[\mathbf{z}_i' - \mathbf{t}\left(\mathbf{Z}\right)\right] \\
&= \left[\left(\mathbf{x}_i\mathbf{A}' + \mathbf{b}\right) - \mathbf{t}\left(\mathbf{XA} + \mathbf{B}\right)\right]' \mathbf{S}^{-1}\left(\mathbf{XA} + \mathbf{B}\right)\left[\left(\mathbf{x}_i\mathbf{A}' + \mathbf{b}\right) - \mathbf{t}\left(\mathbf{XA} + \mathbf{B}\right)\right] \\
&= \left[\mathbf{A}'\mathbf{x}_i + \mathbf{b} - \mathbf{A}'\mathbf{t}\left(\mathbf{X}\right) - \mathbf{b}\right]' \left[\mathbf{A}'\mathbf{S}\left(\mathbf{X}\right)\mathbf{A}\right]^{-1}\left[\mathbf{A}'\mathbf{x}_i + \mathbf{b} - \mathbf{A}'\mathbf{t}\left(\mathbf{X}\right) - \mathbf{b}\right] \\
&= \left\{\mathbf{A}'\left[\mathbf{x}_i - \mathbf{t}\left(\mathbf{X}\right)\right]\right\}' \mathbf{S}^{-1}\left(\mathbf{X}\right)\left\{\mathbf{A}'\left[\mathbf{x}_i - \mathbf{t}\left(\mathbf{X}\right)\right]\right\} \\
&= \mathbf{A}\left[\mathbf{x}_i - \mathbf{t}\left(\mathbf{X}\right)\right]' \mathbf{S}^{-1}\left(\mathbf{X}\right)\mathbf{A}'\left[\mathbf{x}_i - \mathbf{t}\left(\mathbf{X}\right)\right] \\
&= \left[\mathbf{x}_i - \mathbf{t}\left(\mathbf{X}\right)\right]' \mathbf{S}^{-1}\left(\mathbf{X}\right)\left[\mathbf{x}_i - \mathbf{t}\left(\mathbf{X}\right)\right] \\
&= d_i^2\left(\mathbf{X}\right). \qquad (\textit{proofed}).
\end{aligned}$$

The breakdown point (BP) is a popular global measure of robustness which is introduced by Donoho and Huber [19]. The value of BP of estimators lays on the closed interval [0,1] [20]. The value of BP means proportion of outlier data can be covered before the estimators of data differ from the good one (no outliers). A higher BP leads to a more robust estimators and the highest attainable BP is 0.5 in the case of median in the univariate case.

Theorem: ***Breakdown Point*** **(BP) MCD** (Lopuhaa and Rosseeuw [13])

Assume that $\mathbf{X} = \left[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\right]'$ is a set of *m* observation data with dimension *p*-variate and $m \geq p+1$. The estimators $\bar{\mathbf{X}}_{MCD}$ and $\mathbf{S}_{MCD}$ are MCD estimators for mean vector and variance-covariance

matrix of **X**. If p = 1, then $\mathrm{BP}\left[\bar{\mathbf{X}}_{MCD}\right]=\frac{\frac{(m+1)}{2}}{m}$ and $\mathrm{BP}\left[\mathbf{S}_{MCD}\right]=\frac{\frac{m}{2}}{m}$. When $p \geq 2$,then $\mathrm{BP}\left[\bar{\mathbf{X}}_{MCD}\right]=\mathrm{BP}\left[\mathbf{S}_{MCD}\right]=\frac{\frac{(m-p+1)}{2}}{m}$.

The RMCD is extended of MCD with goal is efficiency and speed of computation. To get the RMCD estimators, C-Step algorithm is applied. There are two stages, the first stage is to sort the observations with the minimum distance, then the second stage is to get the result of a subset that has a determinant of the minimum variance-covariance matrix. Assume again that $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_m$ is a random sample Normal p-variate distribution with population mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, then the RMCD algorithm can be described as follows:

Stage I:

1. Determine subset H_1, with data sizes $h=\frac{m+p+1}{2}$ from the previous m sample size.
2. Based on all observations in H_1, calculate mean vector $\bar{\mathbf{x}}_{H_1}$ and variance-covariance matrix $\mathbf{S}_{H_1}$. Then using formula $RD_i^2=\left(\mathbf{X}_i-\bar{\mathbf{X}}_{RD}\right)'\mathbf{S}_{RD}^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}_{RD}\right)$, calculate distance for all $i=1,2,\ldots,m$ where $d_{H_1}^2(i)=d_{RD}^2(i)$ with $\bar{\mathbf{x}}_{RD}=\bar{\mathbf{x}}_{H_1}$ and $\mathbf{S}_{RD}=\mathbf{S}_{H_1}$.
3. Ascending sort of $d_{H_1}^2(i)$ to get $d_{H_1}^2(\pi_1)\leq d_{H_1}^2(\pi_2)\leq\ldots\leq d_{H_1}^2(\pi_m)$ where π is a combination in $\{1,2,\ldots,m\}$.
4. Define $H_2=\left\{X_{(\pi_1)},X_{(\pi_2)},\ldots,X_{(\pi_h)}\right\}$, then calculate mean vector $\bar{\mathbf{x}}_{H_2}$, variance-covariance matrix $\mathbf{S}_{H_2}$, and $d_{H_2}^2(i)$.
5. If $\det\left(\mathbf{S}_{H_2}\right)=0$, then subset $H_2 = H_1$.
6. Repeat step 1-5 to get different members of subset H (example: 500 times).
7. From 500 subset of H that are calculated the determinant from step 1-5, choose 10 subset which have the smallest determinants.
8. Continue to iteration process for the 10 smallest determinant subset until convergent (kth iteration). The convergent condition is attained for all subset if $\det(\mathbf{S}_{H_1})\geq\det(\mathbf{S}_{H_2})\geq\ldots\geq\det(\mathbf{S}_{H_k})=\det(\mathbf{S}_{H_{k+1}})$.
9. Compare the 10 convergent smallest determinant and choose one the smallest determinant. After get $\bar{\mathbf{x}}_{H_k}$ and $\mathbf{S}_{H_k}$, then calculate the distance by formula below

$$d_{H_k}^2(i)=\left(\mathbf{X}_i-\bar{\mathbf{X}}_{H_k}\right)'\mathbf{S}_{H_k}^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}_{H_k}\right) \quad (3)$$

10. To get data distribution consistence which is generated from multivariate normal distribution, then define:

$$\bar{\mathbf{x}}_0=\bar{\mathbf{x}}_{H_k}\text{ and }\mathbf{S}_0=\frac{med_{\left(d_{H_k}^2(i)\right)}}{\chi_{p,0.5}^2}\mathbf{S}_{H_k} \quad (4)$$

Stage II:

In the RMCD algorithm, we re-weighted the estimators of location and scale which is calculated from the first stage. The score 1 is given to $d_0^2(i)$ that is not unusual observations (not outlier) statistically, and score 0 for outliers.

1. The stage I estimators $\bar{\mathbf{x}}_0$ and $\mathbf{S}_0$ are used to determined the weight in calculating robust distance

$$w_i = \begin{cases} 1, \text{if } d_0^2(i) \le \chi^2_{p,0.975} \\ 0, \quad \text{others} \end{cases} \tag{5}$$

2. Based on Eq. (5) then we get [21]

$$\bar{\mathbf{x}}_{RMCD} = \frac{\sum_{i=1}^{m} w_i \mathbf{x}_i}{\sum_{i=1}^{m} w_i}$$
$$\mathbf{S}_{RMCD} = \frac{\sum_{i=1}^{m} w_i \left(\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD}\right)'}{\sum_{i=1}^{m} w_i - 1} \tag{6}$$

From Eq. (6), then

$$d^2_{RMCD}(i) = \left(\mathbf{X}_i - \bar{\mathbf{X}}_{RMCD}\right)' \mathbf{S}^{-1}_{RMCD} \left(\mathbf{X}_i - \bar{\mathbf{X}}_{RMCD}\right), i = 1, 2, \ldots, m. \tag{7}$$

2.2. RMCD Robust Control Chart

We propose to use the Hotelling-T^2 control chart with RMCD robust estimators of location and dispersion parameters for detecting and monitoring the outliers that appear in multivariable process. The RMCD estimators shown in Eq. (6) and Eq. (7) substitute location and dispersion parameters in Eq. (2) then we can rewrite below

$$T^2_{RMCD}(i) = \left(\mathbf{X}_i - \bar{\mathbf{X}}_{RMCD}\right)' \mathbf{S}^{-1}_{RMCD} \left(\mathbf{X}_i - \bar{\mathbf{X}}_{RMCD}\right) \tag{8}$$

Where $\bar{\mathbf{X}}_{RMCD}$ and $\mathbf{S}_{RMCD}$ are mean vector and variance-covariance matrix under the RMCD method based on *m* multivariate observations.

3. Computation and main result

3.1. Method

To evaluate robustness performance of the RMCD estimators and to know how effectively the T^2_{RMCD} control chart in detecting the outliers data, the first step is generate simulation data randomly by hierarchy model $f(\square) = (1-\varepsilon) f_0(\square) + \varepsilon f_1(\square)$, where $f_0(\square)$ is a base distribution of normal multivariate distribution $N_p(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$, for instance with p = 2, $\boldsymbol{\mu_0} = [0,0]'$, $\boldsymbol{\Sigma_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The $f_1(\square)$ is contaminant distribution which substitute as many as εm in $f_0(\square)$; ε is an outlier proportion data. The effect of outliers data using shift outlier that shift mean vector so the centre of data change with hierarchy model below (Hubert and Van Dreissen [22]) :

$$f(\square) = (1-\varepsilon) N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \varepsilon N_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}_0); \boldsymbol{\mu}^* = \boldsymbol{\mu_0} + \delta \tag{9}$$

The assessment to know how effectively the RMCD robust control chart by simulation considering sample sizes m = 50, 100, 250 and 500; p characteristics quality = 2, 5, 8, 10, 12 and 15; percentage of outliers data = 0%, 5%, 10%, 15%, 20%, 25% and 30% and all scenarios repeated 5.000 times. Then calculate statistic of T^2 classically by Eq. (2) and T^2_{RMCD} by Eq. (8) for each scenario. The last step is count outliers data for each simulation scenario, compare between T^2 and T^2_{RMCD}.

3.2. Manual Algorithm of RMCD

To simplify understanding of RMCD algorithm, in this sub-chapter will be given an example of estimator calculations generated based on RMCD algorithm. The first step is to generate multivariate random data with m = 15 and p = 3 by $\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. The result data is as follows:

$$\mathbf{X}' = \begin{bmatrix} 3.0578 & 3.4929 & 5.1135 & 2.3233 & 1.2720 & 1.4274 & 3.4649 & 3.3828 \\ 2.9334 & 2.6593 & 4.1785 & 2.7495 & 4.1988 & 3.0727 & 1.4649 & 2.7601 \\ 3.1134 & 4.0699 & 2.3903 & 2.7771 & 3.9704 & 3.8960 & 2.9456 & 3.4176 \end{bmatrix}$$

$$\begin{bmatrix} 1.4765 & 3.2052 & 4.4202 & 3.0558 & 2.5506 & 3.3772 & 2.1032 \\ 2.2610 & 4.7330 & 4.0700 & 1.5055 & 3.5060 & 3.0582 & 1.8471 \\ 0.6338 & 3.3305 & 3.2270 & 4.3926 & 2.1865 & 3.9412 & 2.2278 \end{bmatrix}$$

Step-:

1. Determine the subset of $H_{1\text{-}1}$ with as many members $h = \frac{m+p+1}{2} = \frac{15+3+1}{2} = 8.5 \cong 9$ as selected randomly from observation m. Suppose the selected observation to be a subset of $H_{1\text{-}1}$ are observations 1, 11, 14, 6, 5, 10, 13, 2, and 9 then the element $H_{1\text{-}1}$ is

$$H_{1-1} = \begin{bmatrix} 3.0578 & 2.9334 & 3.1134 \\ 4.4202 & 4.0700 & 3.2270 \\ 3.3772 & 3.0582 & 3.9412 \\ 1.4274 & 3.0727 & 3.8960 \\ 1.2720 & 4.1988 & 3.9704 \\ 3.2052 & 4.7330 & 3.3305 \\ 2.5506 & 3.5060 & 2.1865 \\ 3.4929 & 2.6593 & 4.0699 \\ 1.4765 & 2.2610 & 0.6338 \end{bmatrix}$$

and from $H_{1\text{-}1}$ the mean vector estimation and the variance-covariance matrix is obtained

$$\left(\bar{\mathbf{x}}_{H_{1-1}}, \mathbf{S}_{H_{1-1}}\right) = \left(\begin{bmatrix} 2.6978 \\ 3.3880 \\ 3.1521 \end{bmatrix}, \begin{bmatrix} 1.2009 & 0.1867 & 0.3222 \\ 0.1867 & 0.6452 & 0.3166 \\ 0.3222 & 0.3166 & 1.2502 \end{bmatrix} \right)$$

With $\det(\mathbf{S}_{H_{1-1}}) = 0.7758$, so as to obtain distance for each observation

$$d^2_{H_{1-1}}(i) = \begin{Bmatrix} 0.538;\ 2.8042;\ 7.5479;\ 0.6639;\ 3.965;\ 2.5964;\ 7.4721;\ 1.4183; \\ 5.7073;\ 3.0595;\ 3.0545;\ 9.9535;\ 0.9794;\ 1.2952;\ 3.7179 \end{Bmatrix}.$$

2. Create a subset of $H_{1\text{-}2}$ whose element is the observation as much as h which has the smallest of $d^2_{H_{1-1}}(i)$ observations 1, 4, 13, 14, 8, 6, 2, 11, and 10. Thus

$$H_{1-2} = \begin{bmatrix} 3.0578 & 2.9334 & 3.1134 \\ 2.3233 & 2.7495 & 2.7771 \\ 2.5506 & 3.5060 & 2.1865 \\ 3.3772 & 3.0582 & 3.9412 \\ 3.3828 & 2.7601 & 3.4176 \\ 1.4274 & 3.0727 & 3.8960 \\ 3.4929 & 2.6593 & 4.0699 \\ 4.4202 & 4.0700 & 3.2270 \\ 3.2052 & 4.7330 & 3.3305 \end{bmatrix}$$

$$\text{with } \left(\bar{\mathbf{x}}_{H_{1-2}}, \mathbf{S}_{H_{1-2}}\right) = \left(\begin{bmatrix} 3.0264 \\ 3.2825 \\ 3.3288 \end{bmatrix}, \begin{bmatrix} 0.7151 & 0.1743 & 0.0585 \\ 0.1743 & 0.4920 & -0.0910 \\ 0.0585 & -0.0910 & 0.3650 \end{bmatrix} \right).$$

obtain $\det(\mathbf{S}_{H_{1-2}}) = 0.1078$. So, it is known $\det(\mathbf{S}_{H_{1-1}}) > \det(\mathbf{S}_{H_{1-2}})$ that for the first subset the member used is $H_{1\text{-}2}$.

3. Repeat steps 1 and 2 in Phase I to form another H_j subset as much as h with j generally 500 times [21].
4. Select 10 H_j subset that yield the smallest determinant value and continue the iteration process until convergent. Table 1 shows the jth subset which gives the smallest determinant value and the end result after iteration until convergent.

Tabel 1. The jth Subset Generates the Smallest Determinant

j th-	Observation Elements $H_{j\text{-}2}$	$\det(\mathbf{S}_{H_{j-2}})$	Observation Elements $H_{j\text{-}k}$	$\det(\mathbf{S}_{H_{j-k}})$	Lots of iterations
24	1, 4, 8, 14, 2, 15, 13, 12, 9	0.02056	1, 2, 14, 15, 13, 8, 4, 12, 9	0.02056	3
95	8, 1, 2, 15, 14, 4, 13, 12, 9	0.02056	1, 2, 14, 15, 13, 8, 4, 12, 9	0.02056	3
238	1, 4, 8, 14, 2, 13, 15, 12, 9	0.02056	1, 2, 14, 15, 13, 8, 4, 12, 9	0.02056	3
281	8, 1, 2, 15, 14, 4, 13, 12, 9	0.02056	1, 2, 14, 15, 13, 8, 4, 12, 9	0.02056	3
19	1, 8, 14, 2, 13, 4, 15, 10, 12	0.02486	1, 14, 2, 13, 4, 8, 15, 12, 10	0.02486	3
250	1, 8, 14, 2, 13, 4, 15, 10, 12	0.02486	1, 14, 2, 13, 4, 8, 15, 12, 10	0.02486	3

263	4, 1, 14, 8, 13, 2, 15, 12, 10	0.02486	1, 14, 2, 13, 4, 8, 15, 12, 10	0.02486	3
280	4, 1, 14, 8, 13, 2, 15, 12, 10	0.02486	1, 14, 2, 13, 4, 8, 15, 12, 10	0.02486	3
296	4, 1, 14, 8, 13, 2, 15, 12, 10	0.02486	1, 14, 2, 13, 4, 8, 15, 12, 10	0.02486	3
376	4 , 1, 13, 15, 14, 8, 2, 10, 12	0.02486	1, 14, 2, 13, 4, 8, 15, 12, 10	0.02486	3

Based on Table 1 it can be seen that the subset H_j which gives the smallest determinant value is the 24th, 95th, 238th and 28th subset with the mean vector estimator and the variance-covariance matrix is

$$\left(\bar{\mathbf{x}}_{H_{24-3}}, \mathbf{S}_{H_{24-3}}\right) = \left(\begin{bmatrix} 2.7578 \\ 2.5867 \\ 2.9733 \end{bmatrix}, \begin{bmatrix} 0.4744 & 0.1095 & 0.7353 \\ 0.1095 & 0.3840 & -0.0394 \\ 0.7353 & -0.0394 & 1.3822 \end{bmatrix}\right)$$

With robust distance,

$$d^2_{24-3}(i) = \left\{ \begin{matrix} 0.6249;\ 1.2468;\ 123.3981;\ 3.8800;\ 178.4003; \\ 103.8074;\ 34.6788;\ 2.9567;\ 4.6033;\ 18.1007; \\ 32.6061;\ 4.5569;\ 2.7776;\ 1.4391;\ 1.9141 \end{matrix} \right\} i = 1,\ldots,15.$$

5. Define $\bar{\mathbf{x}}_0$ and $\mathbf{S}_0$ use Eq. (4) with $med_{\left(d^2_{H_k}(i)\right)} = 4.5569$ and $\chi^2_{p,0.5} = 2.3659$, thus

$$\left(\bar{\mathbf{x}}_0, \mathbf{S}_0\right) = \left(\begin{bmatrix} 2.7578 \\ 2.5867 \\ 2.9733 \end{bmatrix}, \begin{bmatrix} 0.9137 & 0.2109 & 1.4161 \\ 0.2109 & 0.7395 & -0.0758 \\ 1.4161 & -0.0758 & 2.6622 \end{bmatrix}\right)$$

With robust distance:

$$d^2_0(i) = \left\{ \begin{matrix} 0.3244;\ 0.6473;\ 64.0677;\ 2.0145;\ 92.6247; \\ 53.8963;\ 18.0051;\ 1.5351;\ 2.3900;\ 9.3978; \\ 16.9289;\ 2.3659;\ 1.4421;\ 0.7472;\ 0.9937 \end{matrix} \right\} i = 1,\ldots,15.$$

6. Giving weight to $\bar{\mathbf{x}}_0$ and $\mathbf{S}_0$ that $d^2_0(i)$ obtained in step 5 is used as a weighting basis. Based on the Eq. (5) $\left(\chi^2_{(3;0.975)} = 9.3484\right)$, weights are obtained:

$$w_i = \{\ 1,1,\ 0,\ 1,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1\}, i = 1,\ldots,15 \text{ and } \sum_{i=1}^{15} w_i = 9.$$

7. Calculate the RMCD robust estimator by Eq. (6) to obtain:

$$\left(\bar{\mathbf{x}}_{RMCD}, \mathbf{S}_{RMCD}\right) = \left(\begin{bmatrix} 2.7578 \\ 2.5867 \\ 2.9733 \end{bmatrix}, \begin{bmatrix} 0.4744 & 0.1095 & 0.7353 \\ 0.1095 & 0.3840 & -0.0394 \\ 0.7353 & -0.0394 & 1.3822 \end{bmatrix}\right)$$

With robust distance of RMCD as follow:

$$d^2_{RMCD}(i) = \begin{Bmatrix} 0.6249;\ 1.2468;\ 123.3981;\ 3.8800;\ 178.4003; \\ 103.8074;\ 34.6788;\ 2.9567;\ 4.6033;\ 18.1007; \\ 32.6061;\ 4.5569;\ 2.7776;\ 1.4391;\ 1.9141 \end{Bmatrix} i = 1,\ldots,15.$$

3.3. Robustness of RMCD

The robustness of the RMCD estimator can be seen from the consistency of statistical value calculated based on the RMCD method. The mean vector and variance-covariance matrix are said to be robust against the likelihood of emission if the value does not fluctuate greatly (relatively constant) than when there is no outliers. The robustness of the estimator can be determined based on the breakdown point (BP). Estimator RMCD has a maximum BP value of 0.5 which means RMCD estimator can accommodate the existence of up to 50% of total data. If the percentage of outliers data is more than 50%, then the mean vector and variance-covariance matrix will shift from the actual value. As the controls are mean, variance and covariance are calculated from the generated data $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, with $p = 2$, $\boldsymbol{\mu}_0 = [0,0]'$, $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $m = 20$, percentage of outliers = 0%, 5%, 10%, …, 50% and shifted (δ) = 0; 0,5; 1; 1,5; 2; 2,5; 3.

Figures 2 and 3 show the mean statistical values for the X_1 and X_2 variables respectively calculated by the classical method and the RMCD method. The mean parameters are based on the generated data for each variable X_1 and the variable X_2 is 0. The mean statistic X_1 computed by the classical method estimates the value of the parameter around 0.794 before the shift, while the mean X_1 is calculated with RMCD method is much lower that is around 0.605 value. Similarly, for the mean of X_2, there is no shift of 0.809 discharged value if calculated by the classical method and located around 0.619 by RMCD method.

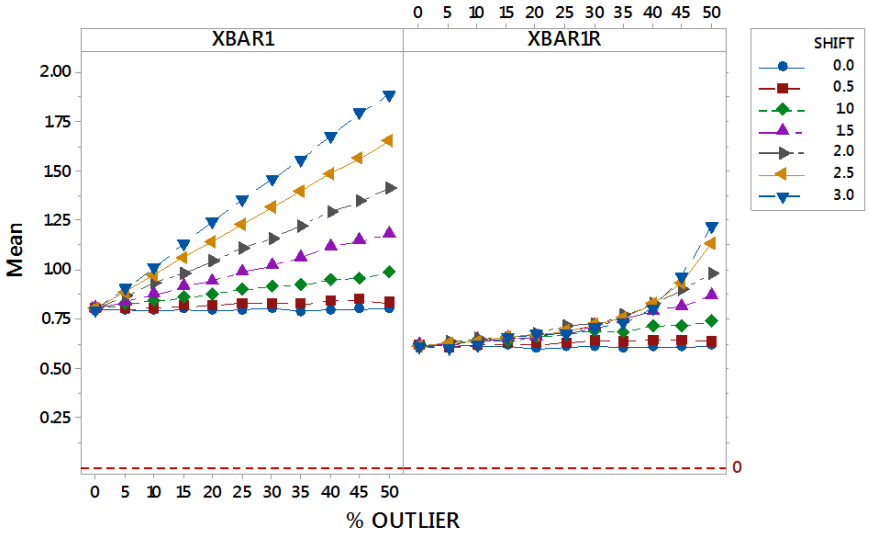


Figure 2. Comparison of Mean Values X_1 Classical Calculations and RMCD with Different Shifts and Percentage of Outliers

When the mean is shifted at a value of 1.5 using the hierarchical function in Eq. (9), the mean X_1 on the 0% percentage is at 0.801 if calculated by the classical method and 0.613 by the RMCD method. This

value increases with increasing percentage of data involved in data, eg in percentage of 30% statistically the mean classical X_1 is 1.017 and X_1 RMCD estimates much smaller that is 0.710. Similarly, on the X_2 variable, the mean statistic of classic X_2 estimates the parameters larger than the real, while the mean RMCD X_2 estimates are much smaller than the classical X_2. For example, when the mean is shifted to the number 2 with a 35% percentile percentage, the mean produced using the classical method is 1.221, while the mean calculated using the RMCD method is 0.769.

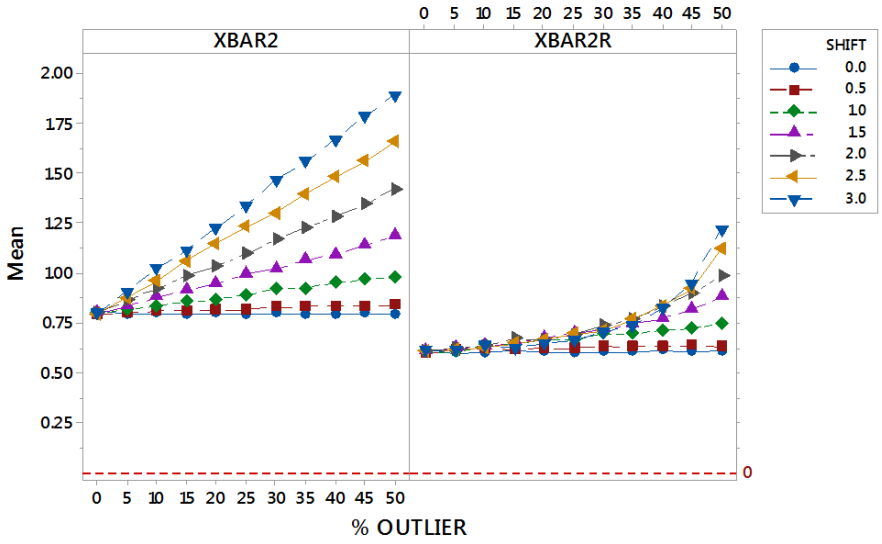


Figure 3. Comparison of Mean Values X_2 Classical Calculations and RMCD with Different Shifts and Percentage of Outliers

The estimated variance generated from the classical method also gives different results than the estimated variance resulting from the RMCD method. As a control, the variance of the generated data for variables X_1 and X_2 is 1 with covariance 0.

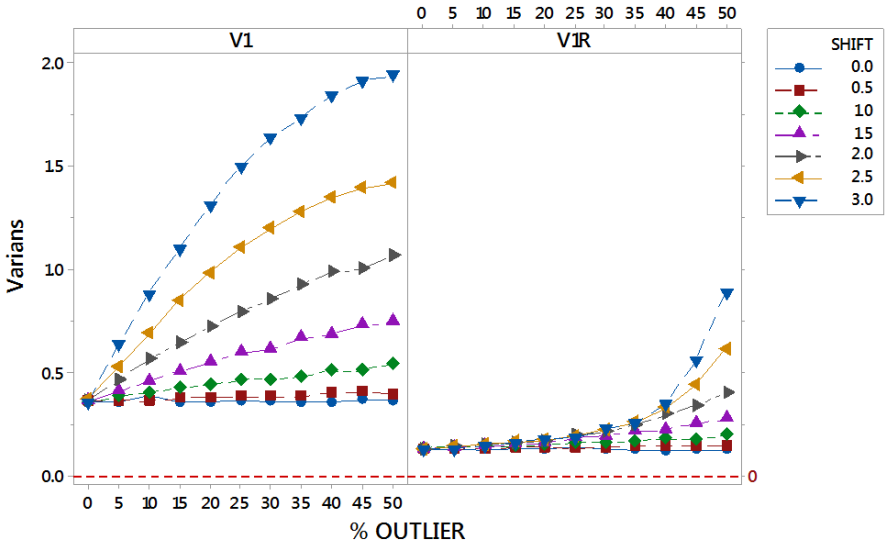


Figure 4. Comparison of Variance Values X_1 Classical Calculations and RMCD with Different Shifts and Percentage of Outliers

Based on the simulation results (Fig. 4 and Fig. 5), the variance is estimated to be less than the initial value, 1. In Figure 4 and Figure 5 it is known that the larger the mean is shifted and the greater the percentage of the outliers involved in the data, the variance estimate also the greater it is. However, the estimation of variance X_1 and X_2 produced by the classical method has a much greater value than that of the RMCD method.

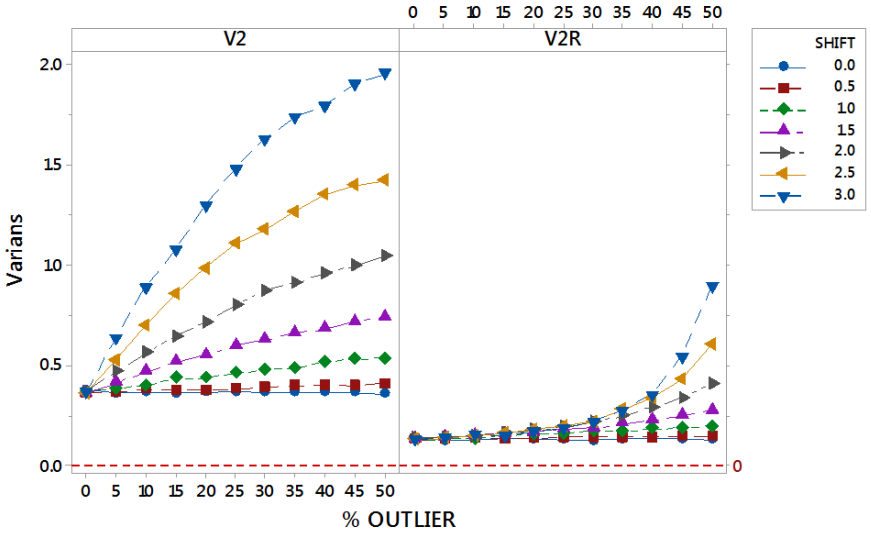


Figure 5. Comparison of Variance Values X_2 Classical Calculations and RMCD with Different Shifts and Percentage of Outliers

3.4. The Effectiveness of T^2_{RMCD} as An Outlier Detector

The effectiveness of the statistics T^2_{RMCD} in detecting the presence of a call can be determined based on statistical accuracy in defining the data of the channel as the data of the channel. For example, if the data is included in 3 (three) outliers, then it is said to have 100% effective if the statistic is capable of detecting all three of them as data.

Figures 5, 6, 7 and 8 show the ratio of sensitivity in various p to m = 50, m = 100, m = 250 and m = 500 respectively. Based on Figure 4.5 with m = 50, it is generally known that the percentage of statistical sensitivity in detecting the presence of sine is much greater than the statistics. For example at p = 2, with a 20% percentage of 50 data or in other words there are 10 out of 50 data kept away from the mean, statistics T^2_{RMCD} detect almost exact (103.26%) of the real, while statistics T^2 are only capable of detecting around 7.5% or in other words, statistics T^2 are only capable of detecting 1 of 10 data including the data of the channel. The sensitivity of the T^2_{RMCD} constant control chart to accurately detect the presence of the data is up to p = 5, while at p = 8, 10, 12 and 15 the T^2_{RMCD} statistics detect in excess the existence of outliers up to 3.5 times the actual number of outliers. As for statistics T^2, as the proportion of output and quality characteristics grow, this chart is incapable of detecting altogether presence.

It can also be seen from Figure 5 that the degree of statistic T^2 sensitivity in detecting the presence of outlier decreases as the variables are involved. At a 10% percentage point, the statistic T^2 sensitivity level decreased from 103.48% with p = 2 to 35.36% with p = 5. This decrease continued to 20.76% with the increase of the variable to p = 8, 14,32 % at p = 10, declined again to 9.32% at p = 12 and statistic T^2 only detected 5.68% when the variable involved was 15. Figure 6 also shows that there is over sensitive statistics T^2_{RMCD} in detecting the presence of the outliers at p is equal to 8 for all the proportion of output from the data, whereas at p less than 8 statistics T^2_{RMCD} it is almost appropriate to detect the number of outliers in the data.

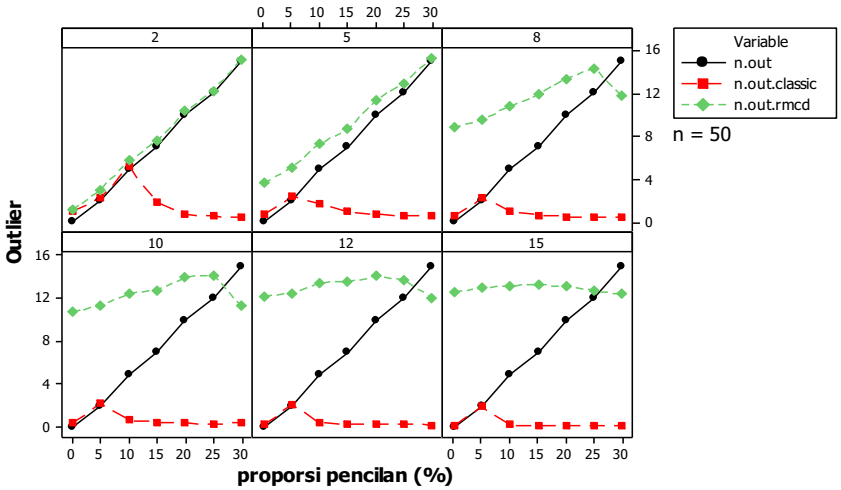


Figure 6. The Effectiveness Comparison Between Statistic T^2 and T^2_{RMCD} for m = 50

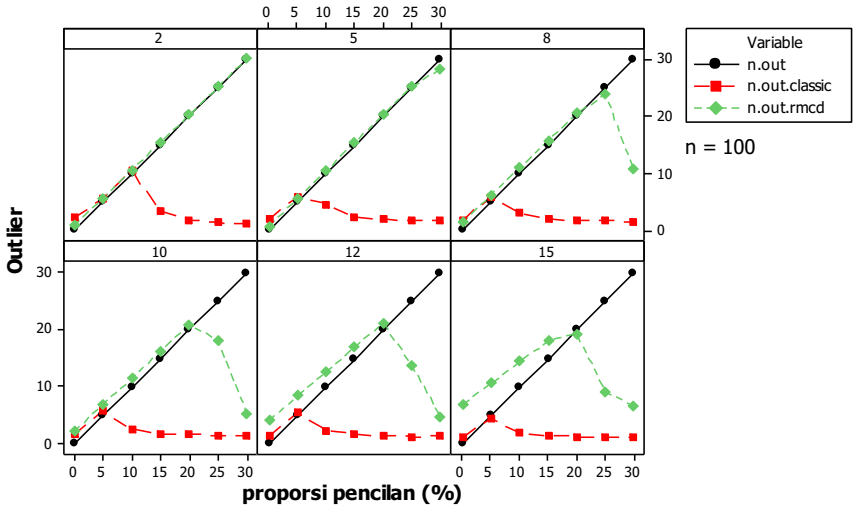
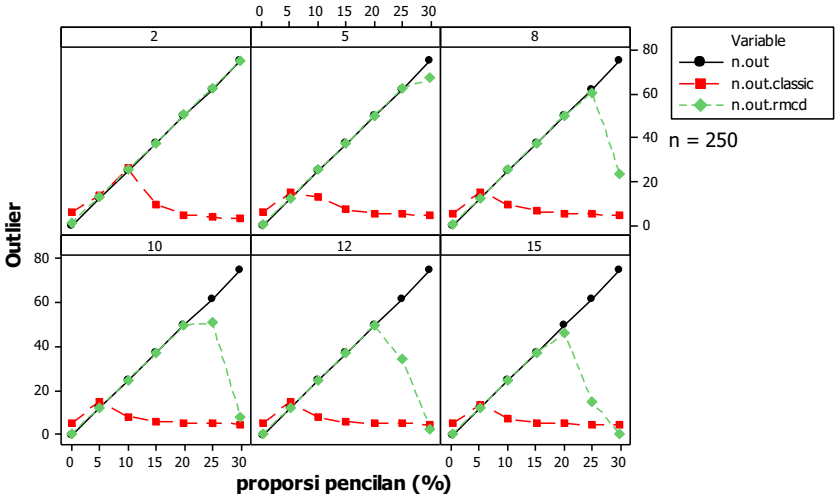


Figure 7. The Effectiveness Comparison Between Statistic T^2 and T^2_{RMCD} for m =100

As shown in Figure 5, Figure 6, Figure 7, Figure 8, and Figure 9 also show a relatively similar pattern of comparison rates of sensitivity. At a proportion of up to 30%, for p less than 8, the statistic T^2_{RMCD} almost accurately detect the presence of the data. As for p more than equal to 8, in the proportion of 0% up to between 20% -25%, the T^2_{RMCD} control chart is almost capable of precisely detecting the presence of outliers, but this sensitivity decreases as the proportion of sine increases by more than 25% (say 30%).

In contrast to statistics T^2_{RMCD}, statistics T^2 are less sensitive in detecting the presence of a call even when the percentage of earnings is 10%. The greater the percentage of enlarged audiences in the data, the lower the degree of statistical sensitivity in detecting its presence. Decrease in the level of statistic T^2 sensitivity also occurs along with the increase in the number of variables at the same percentage of the same. In addition, based on the results of this simulation, it can be seen that the larger the size of the data, then the consistency of the T^2_{RMCD} control chart in detecting the existence of the data with the right way is better.



.

Figure 8. The Effectiveness Comparison Between Statistic T^2 and T^2_{RMCD} for $m = 250$

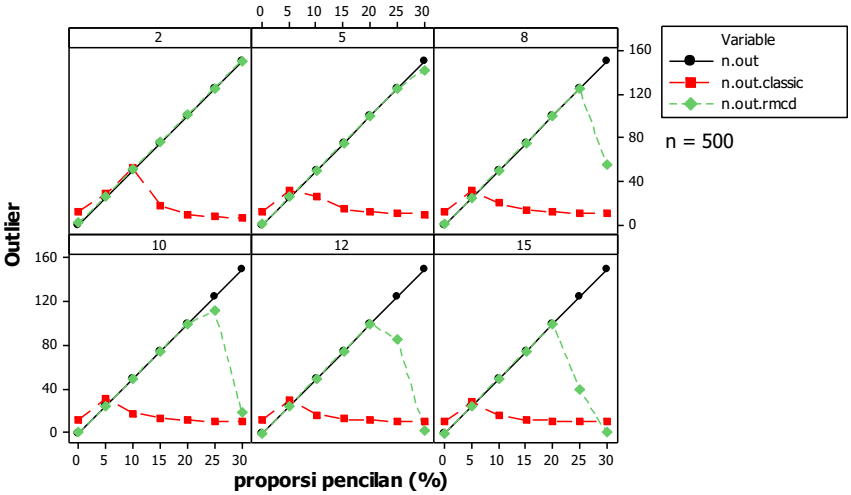


Figure 9. The Effectiveness Comparison Between Statistic T^2 and T^2_{RMCD} for $m = 500$

4. Conclusion

Based on the results, there are two conclusions that can be taken. Firstly, on the size of the data $m = 50$ with p less than equal to 5 for all proportions of outliers (less than equal to 30%), the control chart almost precisely detects the presence of the data of the call, whereas at p more than equal to 8 occurs over sensitive. Secondly, In the data size $m = 100$, 250 and 500 or greater, for all p with a proportion of less than 25%, the control chart almost accurately detects the presence of the data. As for the proportion of outliers more than 25%, occurs less sensitive.

Acknowledgement
We thank Faculty of Mathematics and Natural Sciences, Brawijaya University for funding our research by BOPTN 2017 and also deeply thank Ari Prasetyo, S.Si for many helping in programs and many discussing.

References
[1] Liu H, Shah S, and Jiang W 2004 On-line outlier detection and data cleaning *Computers and Chemical Engineering* **28** 1635
[2] Hawkins D 1980 *Identification of Outliers* (London: Chapman and Hall)
[3] Johnson R 1992 *Applied Multivariate Statistical Analysis* (New Jersey: Prentice Hall)
[4] Barnett V, and Lewis T 1994 *Outliers in Statistical Data* (New York: John Wiley)
[5] Ben-Gal I 2005 *Outlier Detection in: Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (Tel Aviv: Kluwe Academic Publishers)
[6] Acuna E, and Rodriguez C A 2004 Meta analysis study of outlier detection methods in classification *IPSI* (Venice: University of Poerto Rico)
[7] Montgomery, D C 2009 *Introduction to Statistical Quality Control* (New York: John Wiley & Sons)
[8] Midi H, and Shabbak 2011 A Robust Multivariate Control Chart to Detect Small Shift in Mean *Mathematical Problems in Engineering* **2011** 1
[9] Chenouri S, Stefan H S, and Asokan M V 2009 A Multivariate Robust Control Chart For Individual Observation *Journal of Quality Technology* **43** 259
[10] Sullivan J H, and Woodall W H 1998 Adapting Control Charts for the Preliminary Analysis of Multivariate Observations *Communications in Statistics, Simulation and Computation* **27** 953
[11] Rousseeuw P J 1985 Multivariate Estimation with High Breakdown Point. In *Mathematical Statistics and Applications, Section B* **1** 283
[12] Rousseeuw P J, and Yohai V 1984 Robust Regression by Means of S-Estimators *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics* **26** 256
[13] Rousseeuw P J, and Van Zomeren B C 1990 Unmasking Multivariate Outliers and Leverage points *Journal of American Statistical Association* **85** 633
[14] Lopuhaa H P, and Rousseeuw P J 1991 Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrix *The Annals of Statistics* **19** 229
[15] Willems G, Pison G, Rousseeuw P J, and Van Alest S 2002 A Robust Hotelling Test *Metrika* **55** 125
[16] Vargas J A 2003 Robust Estimation in Multivariate Control Charts for Individual Observations *Journal of Quality Technology* **35** 367
[17] Jensen W A, Birch J B, and Woodall W H 2007 High Breakdown Estimation Methods for Phase I Multivariate Control Chart *Quality and Reliability Engineering International* **23** 615
[18] Olive D J 2005 *Applied Robust Statistics* (Illinois: Southern Illinois University)
[19] Donoho D L, and Huber P J 1983 The Notion of Breakdown Point In *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday* (Belmont: Wadsworth)
[20] Davies P L, and Gather U 2007 The Breakdown Point – Examples and Counter Examples *REVSTAT-Statistics Journal* **5** 1
[21] Rousseeuw P J, and Van Driessen K 1999 Fast Algorithm for Minimum Covariance Determinant Estimator *Technometrics* **41** 212
[22] Hubert M, and Van Driessen K 2004 Fast and Robust Discriminant Analysis *Computational Statistics and Data Analysis* **45** 301