Robust Hotelling's T^2 Statistic for Test a Hypothesis about Mean Population based on M-Estimator

Mohanad N. Abdul Sayed
Department of Computer Systems Techniques, Technical Institute/Qurna, Southern Technical University, Basrah, Iraq
Mohanad87@stu.edu.iq

Article Info
Page Number: 8252 - 8260
Publication Issue:
Vol 71 No. 4 (2022)

Article History
***Article Received**: 25 March 2022*
***Revised**: 30 April 2022*
***Accepted**: 15 June 2022*
***Publication**: 19 August 2022*

Abstract

In this study, a robust estimator has been employed to replace the average vectors and correlation matrix in the traditional Hotelling's T^2 statistic. This modification makes use of the M-estimator. In order to demonstrate the modified Hotelling's T^2 statistic's advantage over the conventional Hotelling's T^2 statistic with regard to anomalies, the behavior of the changed Hotelling's T^2 statistic has indeed been compared to the standard Hotelling's T^2 statistic and explained. Whenever the number of samples, n, and the dimensions, p, are minimal, the modified Hotelling's T^2 statistic performs higher than the original Hotelling's T^2.

Keywords:M-estimator, Robust Estimation, T^2 data from Hotelling's.

1. Introduction

Statistical data one of techniques of multivariate statistical that is frequently was using to assess mean-related assumptions is Hotelling's T^2 [1]. In honour of Norman Hotelling, which initially discovered its distribution of the sample, it is known as Hotelling's T^2 [2,13]. The square of the unitary *t* is a multidimensional generalization known as Hotelling's T^2. Hotelling's T^2 analyses different organizations across many regression models concurrently, in contrast to multivariate T^2 [3].

Hotelling's T^2 can be utilized in a variety of contexts. The Hotelling's T^2 statistic, for instance, is employed to evaluate influence on all aspects between two data sets. The solitary Hotelling's T^2 is evaluated in this study, and every one of the studied variables corresponds to a statistical trait. In addition, matched comparisons, repeating measurements, and Hotelling's T^2 are utilized to evaluate mean vector over two different samples [2]. Several implementations' specifics are provided in [2]. In addition, Hotelling's T^2 are utilized for the flowchart [3].

The T^2 efficiency of Hotelling's was already assessed in this research. Furthermore, Hotelling's T^2 is susceptible to extremes [4], even a solitary extremely severe exception can significantly skew the results [5]. Several exceptions also create a "blurring effect," which lowers the effectiveness of the standard Hotelling's T^2 [6]. It is well recognized that all data must be utilized to determine the mean vector $\tilde{\lambda}$ and covariances Ψ. Therefore, the results of the mean vector $\tilde{\lambda}$and covariance Ψ will really be impacted by data with outliers. It might be difficult to keep away from such extremes in a multidimensional context.

incorporated in Hotelling's T^2 in this research. Huber [8] became the first to present the M-estimator for estimate of a simple position variable. Maronna eventually created the M- estimation for multidimensional position and variable with effectiveness [9]. The M-breakdown estimator's point can only be greater than $\frac{1}{(p+1)}$. According to the breaking point, the M- estimation becomes more reactive as the dimensionality rises [10]. A strong replacement for Hotelling's T^2 was already created utilizing this estimation in order to reduce the impact of outliers.

This investigation's goal is to assess how well T^2 by Hotelling, both as written and as changed, performs. There will be two distinct ways to modify the model. The first involves replace the covariance $\Psi$with the M-estimator Ψ_M. the next one involves using M-estimator, $\tilde{P}_M$and Ψ_M in place of both means vectors $\tilde{P}$ and covarianceΨ.

2. The Method

2.1T^2 from the Historic Hotelling

Let $P1, P2, \dots, Pn$ be just a representative sample of a community $Np(\mu, \Sigma)$. the following is indeed the traditional Hotelling's T^2 [2,15]

$$T^2 = k(\tilde{p} - \mu_0)'\Psi^{-1}(\tilde{p} - \mu_0) \quad (2.1)$$

where,

$\tilde{p}$is a mean samples matrix, $\Psi^{-1}$the antithesis of covariance matrices, $k$is sum of samples,

$\mu_0$is a reasonable approximation for vectors of mean.

To determine whether or not H_0; $\mu = \mu_0$ and $H_1: \mu \neq \mu_0$ are true. The essential valueof (2.1) established as (2.2)

$$CV_F = \frac{(k-1)\delta}{(k-\delta)} F_{\delta, k-\delta(\beta)} \qquad (2.2)$$

while δ is the number of dimensions, k is the quantity of samples, and β is the kind I errors. If T^2 exceeds the critical value within that situation, H_0 will be rejected indicating that there are variations in the mean vector.

2.2 Hotelling's T^2 Has Been Selected Depending on M-Estimator.

Let $P1, P2, \dots, Pn$ be just a representative sample of a community $Np(\mu, \Sigma)$.Hence, using M-estimator [11], both mean and covariance matrices is provided as

$$P_M = \sum_{i=1}^{k} \xi_i \, P_i \,/\, k \qquad (2.3)$$

and

$$\Psi_M = \frac{1}{\zeta k} \sum_{i=1}^{k} W_i^2 \,(P_i - \tilde{P})(P_i - \tilde{P})' \qquad (2.4)$$

here ξ_i is a functional, Ψ is indeed a fair estimation of covariances and ζ is picked. Essentially a downweight of a part of E data, the M- estimator. In a chi-squared distributed having δ levels of freedom, consider ϕ^2 become the $1 - \mathrm{E}$ quantile. Set $\xi_i = 1$ if $\psi_i \leq \phi$ and $\xi_i = \phi / \psi_i$in any other cases.

$$\psi_i^2 = k(P_i - \tilde{P})\Psi^{-1}(P_i - \tilde{P})' \quad (2.5)$$

The modified mean and covariance matrices estimation that results from the squared Clustering relationships adjustment using revised estimations is then used to produce new revised forecasts. The process is repeated after equilibrium has been reached.

In this work, the standard Hotelling'sT^2has also been changed using the M-estimator Ψ_M, in place of the covariance Ψ. A T^2 of the upgraded Hotelling is provided via.

$$T_M^2 = k^2(p_i - \tilde{P})\Psi_M^{-1}(p_i - \tilde{P})' \quad (2.6)$$

here $\tilde{P}$ denotes the sampling distribution vector, while Ψ_M^{-1} denotes the reverse covariance of M-estimator. While (2.6) examination is really the main objective, and examine additional adjusted Hotelling's T^2(2.7), where M-estimator was utilized to replace the sampling mean vector and covariance matrices.

$$T_{ME}^2 = k^2(pi - \tilde{P}_M)\Psi_M^{-1}(pi - \tilde{P}_M)' \qquad (2.7)$$

while k represents isnumber of samples, $\tilde{P}_M$ is just the M-mean estimator's vectors, and Ψ_M^{-1} is covariance of M-inverse estimator.

In addition to M-estimator, there are numerous additional reliable estimators. Applications are S-Estimators, the Minimal Volumes Ellipse,Constricted M-Estimators, andMinimal Covariance Predictor [11].

2.3 Designing a Model.

Because the dispersion of adjusted Hotelling's T^2 is unknown, the prevalence of both standard and adjusted Hotelling's T^2 has also been determined by modeling. From $Np(0, \Delta)$ at the magnitude of type 1 error, β = 0.05, we created 10000 sets of data. For every measurement, δ, and, the value is changed to 0, with 0 serving as the standard. Furthermore, δ the degree of deviation, and are all set to one for every dimension. We compute T^2 for traditional and adjusted Hotelling's T^2as provided by (2.1), (2.6), and utilizing these sets of data (2.7). The data types for ξ_i and are given. The equation determines how ξ_i functions; in this case, E has been set at 0.3. In order to make sure if Ψ is just a good estimate, a quantity of ζ is selected. The findings' 95th settings with different will be utilized create the CVs. Additionally, we compare the CVs using Formula (2.4) with the Chi-Squared distributions. For every one of the values of $\mathrm{k} = 40, 70, 120, \text{and } 300$ and $\delta = 6, \text{and } 7$, the predicted CVs were determined. Table 1 summarizes the findings.Pollution levels were $\Phi = 0, 0.1, and\ 0.2.$ 10000 sets of data were produced and calculated for every given criterion.

This simulation's structure was tainted by employing a combination of conventional simulations.

$$(1 - 2\Phi)Np(\mu_0, \Delta_0) + 2\Phi Np(\mu_1, \Delta_1) \qquad (2.8)$$

if Φ is the percentage of extremes, Δ_0 and μ_0 represents the covariance matrix and mean vector that are unsoiled, also$\mu_0$and $\mu_1$equals zero we obtain (2.9).

$$(1 - 2\Phi)Np(0, \Delta_0) + 2\Phi Np(0, \Delta_1) \qquad (2.9)$$

Standard deviation of$\Delta_0$to every parameter δ, $\sigma_0$is used to create a value of 0, with $\sigma_0$specified by being 1. In contrast way, Δ_1 to every parameter δ, $\sigma_1$is used to determine the value of 1. The values for σ_1 in this research are 6 and 7.

10000 simulations were used to evaluate the model using multiple sample quantities, quantity of parameters, and pollution levels. This model employs a kind 1 error of 0.05. The procedures in the model are as follows:

A collection was already produced.

Calculations (2.1), (2.6), and (2.2) were used to determine the values for T^2, T_M^2, and T_{ME}^2 (2.7).

3) Using Table 1, the quantities of T^2, T_M^2, and T_{ME}^2 larger than significance threshold were calculated.

4) The frequency of kind 1 error was employed to measure the effectiveness of T^2, T_{ME}^2, and T_{ME}^2.

3. The Discussions

Table 1 displays the parameter estimates for the modeled and typical distributions. Table 2 –3 displays the results of the original and adjusted Hotelling's T^2, for several examples.

Table 1shows typical and calculated parameter estimates

δ	kT_N^2		T_M^2	T_{ME}^2	CV_F	χ_δ^2
6	40	6.7436	6.7520	6.7921	6.5297	6.23
	70	7.1525	7.5603	7.6281	7.5328	6.23
	120	7.6926	7.7642	7.8704	7.6391	6.23
	300	7.7182	7.8265	7.8739	7.5799	6.23
7	40	8.8072	8.7634	8.7765	8.6072	8.47
	70	9.0223	9.2965	9.3107	9.2866	8.47
	120	9.5926	9.4651	9.6290	9.4824	8.47
	300	9.7267	9.6026	9.2716	9.3691	8.47
8	40	12.3852	11.8105	11.9214	11.7368	12.69

	70	12.7754	12.4626	12.7429	12.4022	12.69
	120	13.8628	12.8921	12.9210	12.7819	12.69
	300	13.2865	13.0178	13.4821	13.2782	12.69

Table 2 False detection rate for the original and adjusted versions of Hotelling's T^2 for $\delta = 6$ and $\beta = 0.05$.

K	Φ	σ	T_N^2	T_M^2	T_{ME}^2
40	0	(4,4,4,4,4,4)	0.066	0.041	0.065
	0.3	(6,6,6,6,6,6)	0.042	0.076	0.055
		(8,8,8,8,8,8)	0.051	0.139	0.063
	0.5	(6,6,6,6,6,6)	0.037	0.152	0.046
		(8,8,8,8,8,8)	0.048	0.287	0.055
70	0	(4,4,4,4,4,4)	0.061	0.088	0.069
	0.3	(6,6,6,6,6,6)	0.058	0.176	0.066
		(8,8,8,8,8,8)	0.034	0.292	0.044
	0.5	(6,6,6,6,6,6)	0.053	0.217	0.064
		(8,8,8,8,8,8)	0.036	0.283	0.047
120	0	(4,4,4,4,4,4)	0.061	0.068	0.066
	0.3	(6,6,6,6,6,6)	0.058	0.127	0.068
		(8,8,8,8,8,8)	0.046	0.185	0.057
	0.5	(6,6,6,6,6,6)	0.055	0.199	0.063
		(8,8,8,8,8,8)	0.038	0.218	0.046
300	0	(4,4,4,4,4,4)	0.044	0.077	0.054
	0.3	(6,6,6,6,6,6)	0.061	0.158	0.070

		(8,8,8,8,8,8)	0.057	0.168	0.063
	0.5	(6,6,6,6,6,6)	0.044	0.171	0.051
		(8,8,8,8,8,8)	0.045	0.219	0.058

Table 3 False detection rate for the original and adjusted versions of Hotelling's T^2 for $\delta = 7$ and $\beta = 0.05$.

K	Φ	σ	T_N^2	T_M^2	T_{ME}^2
40	0	(4,4,4,4,4,4,4)	0.052	0.048	0.050
	0.3	(6,6,6,6,6,6,6)	0.047	0.128	0.057
		(8,8,8,8,8,8,8)	0.063	0.179	0.070
	0.5	(6,6,6,6,6,6,6)	0.041	0.183	0.054
		(8,8,8,8,8,8,8)	0.038	0.215	0.051
70	0	(4,4,4,4,4,4,4)	0.049	0.046	0.048
	0.3	(6,6,6,6,6,6,6)	0.062	0.169	0.074
		(8,8,8,8,8,8,8)	0.051	0.211	0.069
	0.5	(6,6,6,6,6,6,6)	0.046	0.236	0.054
		(8,8,8,8,8,8,8)	0.039	0.247	0.045
120	0	(4,4,4,4,4,4,4)	0.057	0.053	0.055
	0.3	(6,6,6,6,6,6,6)	0.064	0.158	0.081
		(8,8,8,8,8,8,8)	0.047	0.179	0.058
	0.5	(6,6,6,6,6,6,6)	0.041	0.226	0.056
		(8,8,8,8,8,8,8)	0.049	0.267	0.051
300	0	(4,4,4,4,4,4,4)	0.051	0.050	0.049
	0.3	(6,6,6,6,6,6,6)	0.060	0.174	0.074
		(8,8,8,8,8,8,8)	0.054	0.205	0.068
	0.5	(6,6,6,6,6,6,6)	0.048	0.228	0.061
		(8,8,8,8,8,8,8)	0.051	0.231	0.066

The most crucial details is that, as table 1-3 illustrates, the percentage of false alarms of T_N^2 tends to fall even as value of outliers rises. In contrast side, as the quantity of outliers rises, the percentage of false alarms of T_{ME}^2 tends to rise, as demonstrated in Tables 1-3. The outcome is acceptable if T_N^2 and T_{ME}^2 produce same findings, which are H_0 rejection or failure. Initially, it is due to consistency of results, and secondly, due to the achievements' pattern, which shows that T_N^2tends to underperform and T_{ME}^2 tends to overstate. Therefore, the numerical simulations can be utilized as a guide by looking at the population dimension and size if the results vary between one another.It is advised to employ yet a reliable estimator.

4. The Conclusion

According to those results, T_{ME}^2 often performs better than T_N^2 if n is tiny. T_N^2, therefore, performs better than T_{ME}^2 as k grows. The efficiency of T_{ME}^2 degrades in comparison to T_N^2 as p rises. As a result, T_{ME}^2 performed best for k and δ are modest, the quantity of samples. T_N^2 is a preferable option if the dimensionsδ or the quantity of samples k are bigger. To improve T_M^2performancea tweak is required. It is advised to just assess T_M^2 effectiveness when $\widetilde{\Psi}$ is somewhat near to $\widetilde{\Psi}_M$.

References

[1] Raykov, T., &Marcoulides, G. A. (2008). An introduction to applied multivariate analysis. Routledge.

[2] Zelterman, D. (2015). Applied multivariate statistics with R (pp. 393-393). Basel, Switzerland: Springer International Publishing.

[3] Alfaro, J. L., & Ortega, J. F. (2009). A comparison of robust alternatives to Hotelling's T 2 control chart. Journal of Applied Statistics, 36(12), 1385-1396.

[4] Haddad, F., &Alsmadi, M. K. (2020). Improvement of The Hotelling's T 2 Charts Using Robust Location Winsorized One Step M-Estimator (WMOM). Punjab University Journal of Mathematics, 50(1).

[5] Maronna, R. A., Martin, R. D., &Yohai, V. J. (2006). Robust Statistics Theory and Methods John Wiley & Sons. Inc., USA.

[6]Maronna, R. A., Martin, R. D., &Yohai, V. J. (2006). Theory and methods. In Robust Statistics. Wiley.

[7] Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis—A review: Basic concepts. Chemometrics and intelligent laboratory systems, 85(2), 203-219.

[8] Huber, P. J. (1992). Robust estimation of a location parameter. In Breakthroughs in statistics (pp. 492-518). Springer, New York, NY.

[9] Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. The annals of statistics, 51-67.

[10] Lopuhaa, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. The Annals of Statistics, 1662-1683.

[11] Wilcox, R. R. (2011). Introduction to robust estimation and hypothesis testing. Academic press.

[12] Alfaro, J. L., & Ortega, J. F. (2008). A robust alternative to Hotelling's T2 control chart using trimmed estimators. Quality and Reliability Engineering International, 24(5), 601-611.

[13] Abu- Shawiesh, M. O. A., Golam Kibria, B. M., & George, F. (2014). A robust bivariate control chart alternative to the Hotelling's T2 control chart. Quality and Reliability Engineering International, 30(1), 25-35.

[14] Rousseeuw, P. J. (1991). Tutorial to robust statistics. Journal of chemometrics, 5(1), 1-20.

[15] Kaszuba, B. (2012). Applications of robust statistics in the portfolio theory. Mathematical Economics, (8 (15)), 63-82.