教育研究學報 1992，第七卷， 110-114
Educational Research Journal 1992, Vol.7, pp. 110-114

An Overview of the Performance of Four Alternatives to Hotelling's T Square

LIN Wen-ying
The Chinese University of Hong Kong

Hotelling's T^2 is seriously nonrobust when the variance-covariance (v-c) matrices are heteroscedastic and when the sampled populations are skewed and have positive kurtosis. Based on some relevant empirical studies, four alternatives to T^2--James' first order, James' second order, Yao's, and Johansen's tests--, were reviewed. Under v-c heteroscedasticity and/or nonnormality, although James' first order test performs better than T^2, its estimated Type I error rates are not as good as those produced by the other three tests.

在做兩組間多變項分析(multivariate analysis)時，如果兩組間的變異數與共變數矩陣(variance － covariance matrices)不相同且兩組的分佈不是成對稱常態分配狀態時， Hotelling's T^2 則不是一個合適的統計分析方法。因此，有很多統計學者曾試着提出能替代 Hotelling' T^2 的一些統計分析方法。本篇文章旨在對四個能替代 Hotelling's T^2 的統計分析法： James' first order test, 及 James' second order test, Johansen's test, 及 Yao's test 做一介紹及文獻整理，以期能讓教育研究學者在做多變項分析欲選擇一合適的統計分析方法時，能有一個較明確的方向或原則去遵行。

In educational research, it is often necessary to compare two groups of subjects on the means of several response variables. That is, two independent random samples of sizes n_1 and n_2 on p dependent variables are collected from two populations, whose variance-covariance (v-c) matrices of size p x p are Σ_1 and Σ_2; then the hypthesis of equal mean vectors for the two populations is tested. Under the assumptions of normality and v-c homogeneity, Hotelling's (1931) T^2 is the uniformly most powerful test of the hypthesis. However, in practice, data are unlikely to meet these two assumptions. Stevens (1979) cited nine multivariate studies in which v-c heteroscedasticity was often a reality. Micceri (1989), on the other hand, conducted a survey of the skewness of 440 univariate distributions and found that only 6.8 percent of the distributions exhibited normality by virtue of their relative symmetry and both tail weight.

Investigations of the robustness[1] of Hotelling's T^2 with respect to violations of v-c homogeneity and/or normality have been conducted from both analytical (Ito, 1969; Ito & Schull, 1964) and empirical (Algina & Oshima, 1990; Algina & Tang, 1988; Everitt, 1979; Hakstian, Roed, & Lind, 1979; Holloway & Dunn, 1967; Hopkins & Clay, 1963; Nath & Duran, 1983; and Zwick, 1986) standpoints. All of the studies have indicated that under nonnormality T^2 is quite robust, but that under v-c heteroscedasticity T^2 may not be robust when $n_1 = n_2$ and is likely not robust when $n_1 \neq n_2$. Specifically, when the sample sizes are large and equal and when N/p (where $N = n_1 + n_2$) is large, T^2 is robust. Otherwise, it is less robust, being conservative[2] when the larger of n_1 and n_2 is drawn from the population with larger dispersion, and being liberal[3] the other way around. Furthermore, the discrepancy between the actual Type I error rate (α) and the nominal Type I error rate ($\hat{\tau}$) increases with the magnitude of the inequality of the two samples, with the degree of heteroscedasticity, and with p.

As a result, the Behrens-Fisher problems--a comparison of the means of two groups under v-c heteroscedasticity--has been a popular subject in multivariate research. Several alternatives to T^2, intended to produce better results for the Behrens-Fisher problem, have been proposed. Among the alternatives are James' (1954) first and second order tests, Yao's (1965) test, and Johansen's (1980) test. Distributed asymptotically as chi-square with p degrees of freedom, the statistic used by these four tests is

$$T_v^2 = (\bar{X}_1 - \bar{X}_2)'\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1}(\bar{X}_1 - \bar{X}_2)$$

where $\bar{X}_i$ and S_i are the mean vector and v-c matrix for the ith sample, i = 1, 2. The critical values of these four tests are as follows:

1. The critical value of James' test, either first or

second order, is composed of a series of terms in decreasing order of magnitude. Expressed in terms up to order $(n_i - 1)^{-1}$, the critical value of James' first order test is

$$h_1(S_1, S_2; \alpha) = \chi_\alpha^2(p)\left[1 + \frac{1}{2}\left(\frac{k_1}{p} + \frac{k_2 \chi_\alpha^2(p)}{p(p+2)}\right)\right],$$

where

$$k_1 = \sum_{i=1}^{2} \frac{[tr(I - W^{-1}W_i)]^2}{n_i - 1},$$

and

$$k_2 = 2\sum_{i=1}^{2} \frac{tr(I - W^{-1}W_i)^2}{n_i - 1} + \sum_{i=1}^{2} \frac{[tr(I - W^{-1}W_i)]^2}{n_i - 1}.$$

In these expressions, $W_i = (S_i/n_i)^{-1}$, $W = W_1 + W_2$, tr is the trace operator, and $X^2\alpha$ (p) is the $(1 - \alpha)$th percentile of the chi-square distribution with p degrees of freedom.

2. Expressed in terms up to order $(n_i - 1)^{-2}$, the critical value of James' second order test is $h_2(S_1, S_2; \alpha) = h_1(S_1, S_2; \alpha) + O[(n_i - 1)^{-2}]$ where $O[(n_i - 1)]^{-2}$ is given in formula 6.7 of James' (1954) paper.

3. The critical value of Yao's test is $T^2(p, f; \alpha)$, the $(1 - \alpha)$th percentile of the distribution of Hotelling's T^2. Let $E_i = S_i/n_i$ and $E = E_1 + E_2$. Define $V_i = (\overline{X}_1 - \overline{X}_2)' E^{-1} E_i E^{-1}(\overline{X}_1 - \overline{X}_2)$. The quantity f is defined by

$$\frac{1}{f} = \sum_{i=1}^{2} \frac{1}{(n_i - 1)} \left(\frac{V_i}{T_v^2}\right)^2.$$

4. The critical value of Johansen's test is $CF_\alpha(p,q)$, where $F_\alpha(p, q)$ is the $(1 - \alpha)$th percentile of the F distribution with p and q degrees of freedom, and
$C = p + 2A - 6A / (p + 2)$,
$q = p(p + 2) / 3A$,

$$A = \sum_{i=1}^{2} \frac{tr(I - W^{-1}W_i)^2}{2(n_i - 1)} + \sum_{i=1}^{2} \frac{[tr(I - W^{-1}W_i)]^2}{(n_i - 1)}.$$

Review of Literature

This review is based on some relevant empirical studies of the robustness of the four alternatives to Hotelling's T^2.

(1) Robustness of James' First Order Test

Yao (1965) conducted a simulation study for p = 2 to estimate Type I error rates of James' first order test and his own test. His results indicated that while both tests are quite robust under v-c heteroscedasticity, Yao's test performs on average better than James' test. Algina and Tang (1988) conducted a more extensive study than Yao's on the performance of T^2, Yao's test, and James' first order test under v-c heteroscedasticity. In their study, three of the factors and their levels were p = 2, 6, and 10; N/p = 6, 10, and 20; and $n_1:n_2$ = 1:5, 1:3, 1:4, 1:2, 1:1.5, 1:1.25, 1:1, 1.25:1, 1.5:1, 2:1, 3:1, 4:1, and 5:1. In terms of controlling the error rates, both James' and Yao's test perform better than T^2. Neither test tends to be conservative, with the estimated error rates of James' test being larger than those of Yao's.

Algina, Oshima, and Tang (1991) studied the effect of v-c heteroscedasticity and nonnormality upon James' first order test and the other three tests. Uniform, exponential, t(5), beta(5, 1.5)[4], Laplace, and lognormal distributions were used to generate nonnormal data. In comparison, the performance of the other three tests is slightly superior to that of James's first order test. That is, their estimated error rates are more closer to the nominal error rate than those produced by James' first order test.

(2) Robustness of James' Second Order Test and Johansen's Test

Algina, Oshima, and Tang (1991) investigated the robustness of James' second order and Johansen's tests under v-c heteroscedasticity and nonnormality. Using the six nonnormal distributions mentioned in (1), they found that both tests may not be robust when the v-c matrices are heteroscedastic and when the sampled populations are skewed and have positive kurtosis. Specifically, both tests are seriously nonrobust with exponential (skewness = 2 and kurtosis = 6) and lognormal (skewness = 6.18 and kurtosis = 110.93) distributions and slightly nonrobust with the beta distribution (skewness = -0.82 and kurtosis = 0.28). For example, in the negative relationship when $\alpha = .05$, p = 10, N/p = 20, and $n_1:n_2$ = 4:1, the estimated error rates for James' second order test and Johansen's test are respectively .095 and .101 under exponential distribution, .197 and .203

under lognormal distribution, and .063 and .069 under beta distribution. It is fairly robust with respect to the remaining three symmetric distributions.

Lin (1991) investigated the robustness of James' second order and Johansen's tests under v-c heteroscedasticity and nonnormality when N/p = 5, 7, and 9. Besides the normal distribution, the beta (2.5, 1) and exponential distributions were included in her study. The skewness and kurtosis are -0.73 and -0.24 for the beta distribution and 2.0 and 6.0 for the exponential distribution.

With α set at 0.05, Bradley's (1978) liberal criterion for robustness is a Type I error rate in the interval [0.025, 0.075]. Table 1 contains the percentages of $\hat{\tau}$s within this interval for the normal distributions with $\Sigma_2 \neq d^2\Sigma_1$, and with $\Sigma_2 = d^2\Sigma_1$. With the normal distribution and $\Sigma_2 \neq d^2\Sigma_1$, all $\hat{\tau}$s for both tests are within the interval for N/p = 7 or 9. For N/p = 5, $\hat{\tau}$s for both tests occur outside the interval in the negative[5] relationship when $n_1{:}n_2$ = 2:1. With the normal distribution and $\Sigma_2 = d^2\Sigma_1$, all but one $\hat{\tau}$s for both tests are within the interval in the equal-sample-size relationship, regardless of the level of N/p. On the other hand, in the negative relationship, the percentage of $\hat{\tau}$s outside the interval increases as N/p decreases. Specifically, for N/p = 9, only two of the $\hat{\tau}$s occur outside the interval in Johansen's test for $n_1{:}n_2$ = 2:1 and d = 3; for N/p = 7, $\hat{\tau}$s outside the interval occur in both tests for $n_1{:}n_2$ = 2:1 and d = 3; and for N/p = 5, $\hat{\tau}$s outside the interval occur in both tests for other conditions than $n_1{:}n_2$ = 1.5:1 and d = 1.5.

TABLE 1

Percentage of Estimated Type I Error Rates within Bradley's Criterion for the Normal Distributions with $\Sigma_2 \neq d^2\Sigma_1$ and with $\Sigma_2 = d^2\Sigma_1$ When α = .05

		$\Sigma_2 \neq d^2\Sigma_1$		$\Sigma_2 = d^2\Sigma_1$	
N/p	n_1 & n_2	Johansen	James	Johansen	James
5	$n_1 = n_2$	100%	100%	83%	100%
	$n_1 > n_2$[a]	50%	67%	17%	33%
	$n_1 < n_2$[b]	100%	100%		
7	$n_1 = n_2$	100%	100%	100%	100%
	$n_1 > n_2$	100%	100%	75%	75%
	$n_1 < n_2$	100%	100%		
9	$n_1 = n_2$	100%	100%	100%	100%
	n_1 > n2	100%	100%	83%	100%
	$n_1 < n_2$	100%	100%		

[a] negative relationship between sample sizes and v-c matrices.
[b] positive relationship between sample sizes and v-c matrices.

TABLE 2

Percentage of Estimated Type I Error Rates within Bradley's Criterion for the Beta. Distributions with $\Sigma_2 \neq d^2\Sigma_1$ and with $\Sigma_2 = d^2\Sigma_1$ When α = .05

		$\Sigma_2 \neq d^2\Sigma_1$		$\Sigma_2 = d^2\Sigma_1$	
N/p	n_1 & n_2	Johansen	James	Johansen	James
5	$n_1 = n_2$	67%	83%	50%	50%
	$n_1 > n_2$[a]	25%	33%	0%	8%
	$n_1 < n_2$[b]	100%	100%		
7	$n_1 = n_2$	100%	100%	83%	83%
	$n_1 > n_2$	75%	75%	17%	33%
	$n_1 < n_2$	100%	100%		
9	$n_1 = n_2$	100%	100%	100%	100%
	$n_1 > n_2$	92%	92%	67%	67%
	$n_1 < n_2$	100%	100%		

[a] negative relationship between sample sizes and v-c matrices.
[b] positive relationship between sample sizes and v-c matrices.

Table 2 contains the percentages of $\hat{t}$s within the interval for the beta distributions with $\Sigma_2 \neq d^2\Sigma_1$ and with $\Sigma_2 = d^2\Sigma_1$. With the beta distribution and $\Sigma_2 \neq d^2\Sigma_1$, for N/p = 9, only one $\hat{t}$ for each test occurs outside the interval in the negative relationship. The estimates are .0815 for James' test and .0855 for Johansen's test and occur when $n_1:n_2 = 2:1$ and d = 3. For N/p = 7, $\hat{t}$s outside the interval occur in the negative relationship when $n_1:n_2 = 2:1$ and d = 3. For N/p = 5, $\hat{t}$s outside the interval occur in the negative relationship when $n_1:n_2 \geq 1.5$, and occur in the equal-sample-size relationship when d = 3. With the beta distribution and $\Sigma_2 = d^2\Sigma_1$, for N/p = 9, $\hat{t}$s for both tests occur outside the interval in the negative relationship when $n_1:n_2 \geq 1.5$ and d = 3. For N/p = 7, $\hat{t}$s for both tests occur outside the interval in the negative relationship when $n_1:n_2 \geq 1.5$, and in the equal-sample-size relationship when d = 3. For N/p = 5, almost all $\hat{t}$s for both tests occur outside the interval in the negative relationship and half of the $\hat{t}$s in the equal-sample-size relationship when d = 3.

For the exponential distribution with $\Sigma_2 \neq d^2\Sigma_1$, the percentages of $\hat{t}$s within the interval are 42% for James' test and 33% for Johansen's test in the negative relationship and for N/p = 9. Most of them occur when d = 3, irrespective of the level of $n_1:n_2$.

In terms of the relationship between sample sizes and v-c matrices, the two tests perform similarly in the positive and equal-sample-size relationships, but James' test performs slightly better than Johansen's test in the negative relationship. For example, $\hat{t}$ is .0660 for James' test and .0715 for Johansen's test when N/p = 5, $n_1:n_2 = 1.5:1$, and d = 1.5. Furthermore, James' second order test tends to perform better for large values of p but Johansen's test tends to perform better for small values of p.

(3) Robustness of Yao's Test

As an extension of Welsh's (1938, 1951) solution to the Behrens-Fisher problem for two samples, Yao's (1965) simulation study showed that his test is slightly superior to James' first order test. Since only the bivariate case was considered, the generality of his results is limited.

Algina and Tang (1988) extended Yao's study to investigate the robustness to heteroscedasticity of Yao's test. They concluded that i) when $n_1 = n_2$, Yao's test generates appropriate error rates when $10 \leq N/p \leq 20$. When $n_1 < n_2$, Yao's test produces appropriate error rates except when $n_1:n_2 = 1:5$ and $N/p \leq 6$. When $n_1 > n_2$, Yao's test produces appropriate error rates provided that $N/p \geq 10$ and $n_1:n_2 < 2$. When $N/p \geq 20$, the test can be safely used when p = 6 and $n_1:n_2$ is as large as 3:1, and when p = 10 and $n_1:n_2$ is as large as 4:1.

Algina, Oshima, and Tang (1991) studied the robustness of Yao's test under v-c heteroscedasticity and nonnormality. Guided by the recommendations of Algina and Tang (1988) for the safe use of Yao's test, they used 10 and 20 for the ratio N/p. Their study indicated that when the sampled distributions are symmetric, Yao's test is quite robust under v-c heteroscedasticity and unequal sample sizes. However, when the distributions are skewed and have positive kurtosis, the test tends to be liberal. For example, in the negative relationship when $\alpha = 0.05$, p = 2, $n_1:n_2 = 2:1$, and N/p = 10, the estimated error rate of the test is 0.122 under the lognormal distribution. Their findings are consistent with those of Clinch and Keselman (1982), which indicated that the univariate counterpart to Yao's test is liberal with skewed distribution.

Conclusions and Recommendations

This review of the performance of the four alternatives to Hotelling's T^2 is intended to provide researchers with some guidelines when conducting multivariate tests for two samples under v-c heteroscedasticity and/or nonnormality.

Although James' first order test performs better than T^2 under v-c heteroscedasticity and nonnormality, its estimated Type I error rates are not as good as those produced by the other three tests. Hence, only James' second order, Yao's, and Johansen's tests will be included for further consideration.

The generalizability of the results for the three tests is limited by the range of values for the factors reviewed in this article. Furthermore, in practice, sample v-c matrices may not be a reliable guide to the direction of the relationship between sample sizes and v-c matrices and to the type of heteroscedasticity. Thus relationship (negative or positive) and type of heteroscedasticity will be suppressed for consideration. With these restrictions in mind, the following conclusions can be set forth:

1. Before collecting data, researchers need to consider the likely distributional characteristics of the data in order to select the appropriate sample sizes.
2. With symmetric distributions, N/p should be at least 9 if $1.5 < n_1:n_2 \leq 2.0$; N/p can be reduced to 7 if $n_1:n_2 \leq 1.5$; and N/p = 5 may be sufficient if $n_1 = n_2$.
3. With moderately skewed distributions (such as beta distribution), N/p = 7 may be sufficient if $n_1 = n_2$.

4. With skewed distributions (such as lognormal and exponential distributions), the performance of all the three tests is poor, even when N/p = 9. Additional research with larger N/p (i.e., N/p ≥ 20) ratios is required to determine the minimum ratio needed for the tests to perform adequately.
5. Both James' second order and Johansen's tests can be generalized to more than two groups, whereas Yao's test cannot. Therefore, when research involves two or more groups, James' and Johansen's tests are recommended.
6. By comparison, the performance of James' test is slightly better than that of Johansen's test, especially in the negative relationship. In addition, as p increases, the performance of James' test improves and that of Johansen's test declines. On the other hand, while James' test is quite involved computationally, Johansen's test is comparatively simple. In addition, Johansen's test can be used for factorial designs.

Notes

[1]Roughly speaking, a test is robust if its Type I error rates are nto affected drastically by v-c heteroscedasticity and/or nonnormality.
[2]A test is conservative if the actual error rate is smaller than the nominal error rate.
[3]A test is liberal if the actual error rate is larger than the nominal error rate.
[4]t(5) stands for Student's t distribution with 5 degrees of freedom and beta (5, 1.5) stands for beta distribution with parameters 5 and 1.5.
[5]Positive relationship means that the larger of n_1 and n_2 is drawn from the population with larger v-c matrix; negative relationship means the other way around; and equal-sample-size relationship means that since $n_1 = n_2$, it is irrelevant whether n_1 or n_2 is drawn from the population with either larger or smaller v-c matrix.

References

Algina, J., & Oshima, T. C. (1990). Robustness of the independent sample Hotelling's T^2 to variance-convariance heteroscedasticity when sample sizes are unequal or in small ratios. *Psychological Bulletin, 108,* 308-313.

Algina, J., Oshima, T. C., & Tang, K. L. (1991). Robustness of Yao's, James' and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. *Journal of Educational Statistics, 16,* 125-139.

Algina, J., & Tang, K. L. (1988). Type I error rates for Yao's and James' tests of equality of mean vectors under variance-covariance heteroscedasticity. *Journal of Educational Statistics, 13,* 281-290.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics, 7,* 207-214.

Everitt, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample T^2 tests. *Journal of the American Statistical Association, 74,* 48-51.

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T^2 procedure and the assumption of homogeneous covariance matrices. *Psychological Bulletin, 86,* 1255-1263.

Holloway, L. N., & Dunn, O. J. (1967). The robustness of Hotelling's T^2. *Journal of the American Statistical Association, 62,* 124-136.

Hopkins, J. W., & Clay, P. P. F. (1963). Some empirical distributions of bivariate T^2 and homoscedasticity criterion under unequal variance and leptokurtosis. *Journal of the American Statistical Association, 58,* 1048-1053.

Hotelling, H. (1931). The generalization of Student's ratio. *Annuals of Mathematical Statistics, 2,* 360-378.

Ito, K. (1969). On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures. In P. R. Krishnaiah (Ed.), *Multivariate Analysis-II* (pp. 87-120). New York: Academic Press.

Ito, K., & Schull, W. J. (1964). On the robustness of the T^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. *Biometrika, 38,* 324-329.

James, G. S. (1954). Tests of linear hypothesis in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika, 41,* 19-43.

Johansen, S. (1980). The Welch-James approxmination to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85-92.

Lin, W. Y. (1991). *Robustness of two multivariate tests to variance-covariance heteroscedasticity and nonnormality when total-sample-size-to-variable ratio is small.* Unpublished doctoral dissertation, University of Florida, Gainesville, FL.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Nath, R.,, & Duran, B. S. (1983). A robust test in the multivariate two-sample location problem. *American Journal of Mathematical and Management Sciences, 3,* 225-249.

Stevens, J. (1979). Comment on Olson: Choosing a test statistic in multivariate analysis. *Psychological Bulletin, 86,* 355-360.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika, 52,* 139-147.

Zwich, R. (1986). Rank and normal scores alternatives to Hotelling's T^2. *Multivariate Behavioral Research, 21,* 169-186.

Author

LIN Wen-ying, Lecturer, Department of Educational Psychology, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.