AN EMPIRICAL EVALUATION OF MULTIVARIATE SEQUENTIAL PROCEDURE FOR TESTING MEANS

By R. H. Appleby and R. J. Freund

Viriginia Polytechnic Institute

0. Summary. Jackson and Bradley (1959, 1961a, 1961b) developed and studied a sequential (multivariate) T^2 test of hypotheses on a vector of means, and an analogous χ^2 test for known covariance structure. The present paper presents the results of Monte Carlo sampling on the operating characteristics and average sample numbers (ASN) of these tests. Consideration is restricted to the behavior of these tests at specific null and alternate hypotheses (H_0 and H_1) with nominal α and β errors of .05.

The empirical α and β errors are, in general, less than .05 and appear to decrease as the number of variables increases. The empirical ASN are appreciably smaller than the corresponding fixed sample sizes, and approximate the ASN that Jackson and Bradley obtained using Bhate's conjecture. The estimation of the fixed sample sizes were based on the nominal α and β errors of .05 while the sequential test ASN were, of course, associated with the resulting smaller error probabilities. Thus the true advantage of the sequential test is understated by the above sample size comparisons.

This study investigates the behavior of the sequential test at H_0 and H_1 only. A similar study involving points between H_0 and H_1 would be of definite values for (1) ascertaining whether the advantages of the sequential procedure hold under situations other than H_0 and H_1, and (2) suggesting methods to overcome the conservatism of the test as it now stands.

1. Introduction. Wald (1947), Rushton (1952), and others (see Jackson (1960)) have developed a sequential testing procedure which, under the assumption that a sample comes from a normally distributed population with mean μ and variance σ^2, is a test of hypothesis

$$H_0 : (\mu - \mu_0)/\sigma = 0, \qquad \text{versus}$$
$$H_1 : (\mu - \mu_0)/\sigma = \pm\lambda,$$

where α, the probability of erroneously rejecting H_0, and β, the probability of erroneously accepting H_0, are specified in advance. Jackson and Bradley (1959, 1961a, 1961b) extend this test to the multivariate normal case, specifically to the hypothesis:

$$H_0 : (\underline{\mu} - \underline{\mu}_0)'\Sigma^{-1}(\underline{\mu} - \underline{\mu}_0) = 0 \qquad \text{versus}$$
$$H_1 : (\underline{\mu} - \underline{\mu}_0)'\Sigma^{-1}(\underline{\mu} - \underline{\mu}_0) = \lambda^2,$$

Received May 5, 1961; revised July 14, 1962.

where

$\boldsymbol{\mu}$ is a vector of means,
$\boldsymbol{\mu}_0$ is a vector of hypothesized means,
$\boldsymbol{\Sigma}$ is the matrix of variances and covariances, and
λ^2 is a non-centrality parameter.

Jackson and Bradley consider two cases:

Case I. called the multivariate sequential χ^2 test, where $\boldsymbol{\Sigma}$ is known, and

Case II. called the multivariate sequential T^2 test, where $\boldsymbol{\Sigma}$ is not known, and is estimated at each step of the sequential procedure by S_n, the usual estimator of $\boldsymbol{\Sigma}$.

For both cases, it is shown that the test procedure terminates with probability one, and that the actual α and β errors are approximately as specified. In addition, approximations to the Average Sample Number (ASN) are obtained using Bhate's conjecture. Tables to facilitate the multivariate sequential tests are available in technical reports by Jackson and Bradley (1959) and Freund and Jackson (1960).

The procedure for the sequential χ^2 test is to compute, after a sample of n observations, $\chi_n^2 = n(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_0)$, where $\bar{\mathbf{x}}_n$ is the vector of sample means based on n observations. Then, if

$$\chi_n^2 \leqq \underline{\chi}_n^2 \qquad \text{accept } H_0,$$

$$\chi_n^2 \geqq \bar{\chi}_n^2 \qquad \text{reject } H_0, \text{ and if}$$

$$\underline{\chi}_n^2 < \chi_n^2 < \bar{\chi}_n^2 \qquad \text{continue sampling,}$$

where $\bar{\chi}_n^2$ and $\underline{\chi}_n^2$ are values found in the tables and depend on p (the number of variables), λ^2, α, and β. Similar procedures are used for the sequential T^2 test, using the statistic $T_n^2 = n(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_0)'\mathbf{S}_n^{-1}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_0)$.

2. The Monte Carlo study. In simulation or Monte Carlo studies, a probabilistic process is simulated or generated by appropriate use of random numbers. A large number of simulated samples can be generated on an electronic computer and inferences can be drawn from the analysis of these repeated samples. Such simulations are useful to (1) solve distribution problems too difficult to examine theoretically and (2) to verify approximations used in mathematical derivations.

The introduction to the *Tables to Facilitate Sequential t-Tests* by K. J. Arnold (1951) contains a Monte Carlo study of 500 samples. This study showed the actual α and β errors to be somewhat below the desired 5 per cent, and Jackson and Bradley (1961a) indicate that the resulting ASN are of about the right order of magnitude compared to Bhate's conjecture.

In the present study, the multivariate normal population for the null case ($\lambda^2 = 0$) from which samples were drawn was $N(\mathbf{0}, \mathbf{I})$, i.e., all individual observations are independently distributed $N(0, 1)$. Random samples from this population were obtained by generating uniformly distributed random numbers by the "Middle Square" method, and using a normalizing transformation to

obtain random normal deviates with range from -4 to $+4$. Appropriate constants were added as required to produce samples from populations for specific alternate hypotheses, e.g., $\lambda^2 = 1$.

The use of a multivariate population distributed $N(\mathbf{0}, \mathbf{I})$ presents no loss in generality since any population $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be obtained from the former by linear transformations which do not affect T^2 and χ^2 tests. It should be noted, further, that the resulting S_n matrices are, not in general, identity matrices.

The resulting samples of n sets of p-variate vectors were used to compute, for any given test (specified by p, λ^2, Σ known or unknown; $\alpha = \beta = .05$ for all tests), the appropriate statistic (χ^2_n or T^2_n). This was compared to the corresponding tabled value, resulting in a decision to accept H_0, reject H_0, or continue sampling. Testing was initiated at the minimum sample size needed for a decision (given in the tables) and terminated when a decision was made to accept or reject, or when the number of observations exceeded available table entries. The complete testing procedure was repeated a large number of times for various representative combinations of p and λ^2. There was no programmed cut-off after a given number of samples, resulting in slightly odd numbers of samples (Tables 1 and 3).

The frequency of cases where the sample testing procedure reached the upper limits of available tables before a decision was reached was small (see Tables 1 and 3). For this reason, special decision procedures, e.g., Wald (1947), were not used except in one case as an illustration. These samples were ignored in the calculation of the empirical ASN, α and β errors, and thus these values are conditional on termination within the scope of the tables. This procedure results in a downward bias in the ASN; however, the resulting frequency tables of empirical ASN should be of value for planning experiments.

The combinations of p and λ^2 for which samples were drawn can be found in the heading of Tables 1 and 3. These combinations reflect the fact that in many applications, "interesting" differences from a hypothetical population will produce values of λ^2 which increase with the number of variables. The upper limits for n (the number of samples) reflect the fact that larger p and smaller λ^2 require larger samples; these limits are roughly two to three times the Bhate's conjecture ASN's.

3. Results of the Monte Carlo study. In this study four characteristics of the sequential tests of H_0 against H_1 are of interest:

1. The correctness of Bhate's conjecture for ASN,
2. The correctness of the desired α and β errors,
3. The proportion of samples which did not terminate with a decision before exhausting available tables, and
4. The advantage of using the sequential test when H_0 or H_1, are in fact, the case.

The above characteristics are discussed below for the χ^2 test and the T^2 test respectively.

TABLE 1

Information on sample numbers, sequential χ^2 tests

p	2			3			5			9		
λ^2	0.5	1.0	2.0	0.5	1.0	2.0	1.0	2.0	5.0	2.0	5.0	10.0
	Upper Limit of Tables											
	60	45	30	60	45	30	50	25	10	30	12	6
	H_0											
Sample Sizes	Frequencies of Sample Numbers											
≦5	0	6	570	10	3	207	0	139	463	60	433	500
6–10	1	570	359	4	214	256	148	307	37	332	64	—
11–15	145	985	69	174	181	33	186	65	—	86	3	—
16–20	157	94	4	275	67	3	90	11	—	31	—	—
21–30	127	39	1	391	35	2	63	2	—	8	—	—
31–40	50	4	—	114	3	—	0	—	—	—	—	—
41–60	24	0	—	56	2	—	0	—	—	—	—	—
Not term.	3	0	0	18	0	0	0	1	0	0	0	0
Total	507	998	1003	1042	505	501	487	525	500	517	500	500
	Average Sample Numbers											
Observed	21.4	11.1	6.0	23.1	12.5	6.5	14.5	7.5	3.3	9.1	4.0	2.3
Bhate's	25	13	7	27	14	7	15	8	3	9	4	2
Fixed	32	16	8	35	18	9	20	10	4	12	5	3
	H_1											
Sample Sizes	Frequencies of Sample Numbers											
≦5	14	210	638	8	76	282	37	261	458	144	429	498
6–10	97	451	311	77	217	185	205	230	39	252	68	2
11–15	137	212	48	122	119	41	129	51	—	78	2	—
16–20	97	99	6	109	51	4	71	15	—	18	—	—
21–30	103	36	0	131	33	0	59	3	—	4	—	—
31–40	40	1	—	50	2	—	8	—	—	—	—	—
41–60	18	0	—	40	0	—	1	—	—	—	—	—
Not term.	2	0	0	6	0	0	0	1	3	0	1	0
Total	508	1000	1003	543	498	512	510	561	500	496	500	500
	Average Sample Numbers											
Observed	18.1	9.6	5.3	20.7	10.7	5.9	12.7	6.7	3.1	8.0	3.7	2.1
Bhate's	15	8	4	17	9	5	11	6	3	7	3	2
Fixed	32	16	8	35	18	9	20	10	4	12	5	3

3.1. *The multivariate sequential χ^2 test.* Table 1 summarizes, for all parameters used in the sampling study the observed ASN, Bhate's conjecture ASN, and the sample size needed for the equivalent fixed sample χ^2 test. Zeroes in the frequency tables indicate no sample sizes in that category; dashes indicate that the size category was beyond table limits. Conjectured ASN's were rounded to the next largest integer.

The fixed sample sizes were computed for $\alpha = \beta = .05$ and were determined using a variance stabilizing (square root) transformation of non-central χ^2 developed by Hofer (1960): let u be distributed as $\chi^2(n)$ with non-centrality parameter λ^2, then $z = (u - \frac{1}{2}n)^{\frac{1}{2}}$ will be approximately normal with expectation $(\frac{1}{2}n + \lambda^2)^{\frac{1}{2}} - \frac{1}{2}(\frac{1}{2}n + \lambda^2)^{-\frac{1}{2}}$, and variance 1. This variance stabilizing transformation was extensively studied by Hofer for degrees of freedom varying from 2 to 150. The maximum error of approximation to the exact power was .013 and occurred at 2 d.f. for a non-centrality parameter near 1; for all other cases the maximum error was well below .010. This transformation thus compares favorably with others available in the literature and recommends itself by its ease of computation.

It can be seen from this table that the conjectured ASN's are of the same order of magnitude as the actual ASN's, although in general they are too large under H_0 and too small under H_1. Nevertheless, the conjectured ASN's are sufficiently close to be of help in planning experiments. Further, as expected, a rather substantial savings in sample size is realized over the fixed sample procedure at H_0 and H_1.

Table 1 also shows the number of samples which were terminated without

TABLE 2

Observed α and β errors for the χ^2 test

p	λ^2	α	β
2	0.5	.030[1]	.036[1]
2	1.0	.020[1]	.038[2]
2	2.0	.022[2]	.019[2]
3	0.5	.035[2]	.035[1]
3	1.0	.031[1]	.030[1]
3	2.0	.016[1]	.031[1]
5	1.0	.024[1]	.036[1]
5	2.0	.011[1]	.034[1]
5	5.0	.022[1]	.018[1]
9	2.0	.031[1]	.030[1]
9	5.0	.012[1]	.026[1]
9	10.0	.018[1]	.014[1]

[1] Approximate 95 per cent confidence interval ± .020.

[2] Approximate 95 per cent confidence interval ± .014.

decision. The largest proportion of no-decision samples occurs for $p = 3, \lambda^2 = .5$, a result which could be expected, since for this case the ratio of $[n(\lambda^2, p)]$/ [conjectured ASN] is smaller than it is for the other cases sampled. Nevertheless, even in this case, the number of non-decision samples is quite small and it may be concluded that the available tables are sufficient. Wald (1947) gives a method for treating such samples; this was used for the non-decision samples for $p = 3$, $\lambda^2 = .5$, changing the α and β errors from .035 to .037 for both hypotheses. Further, assuming sample size as the upper limit (60), in these cases, the observed ASN's changed to 23.6 and 21.1 for H_0 and H_1, respectively.

Table 2 shows the empirical probabilities of the α and β errors. The desired probability was .05 for both; the confidence intervals show that most empirical probabilities are significantly below the desired levels. Further, the empirical α and β errors appear to drop off as λ^2 increases.

3.2. *The multivariate sequential T^2 test.* Table 3 provides information on the sample sizes for the sequential T^2 test. The fixed-sample sizes, based on the nominal .05 for α and β, were obtained using the following variance stabilizing transformation: let F be distributed as $F(m, n)$ with non-centrality parameter λ^2, then $Z = \cosh^{-1}(w/a)$ is approximately normal with expectation $[\rho - \coth \rho/$

TABLE 3

Information on sample sizes, sequential T^2 test

	H_0						H_1					
p	2			9			2			9		
λ^2	0.5	1.0	2.0	2.0	6.0	10.0	0.5	1.0	2.0	2.0	6.0	10.0
	Upper Limit of Tables											
	60	45	30	50	30	20	60	45	30	50	30	20
Sample Sizes	Frequencies of Sample Numbers											
<5	0	0	162	—	—	—	0	0	10	—	—	—
6–10	0	252	259	9	35	41	0	154	372	1	1	1
11–15	100	152	65	43	56	58	133	194	101	16	86	96
16–20	163	64	14	27	8	1	116	94	18	61	12	3
21–30	154	36	6	20	1	—	144	55	3	22	1	—
31–40	56	6	—	1	—	—	74	7	—	0	—	—
41–60	23	1	—	0	—	—	33	0	—	0	—	—
Not term.	4	1	0	0	0	0	0	0	0	0	0	0
Total	500	512	506	100	100	100	500	524	504	100	100	100
	Average Sample Numbers											
Actual	22.4	12.4	7.7	16.0	11.9	11.2	23.1	13.8	9.1	18.5	14.0	13.0
Bhate's	26	14	8	15	11	10	21	13	9	18	14	13
Fixed	35	21	14	36	18	16	35	21	14	36	18	16

TABLE 4

Observed α and β errors, T^2 Test

p	λ^2	α	β
2	0.5	.034[2]	.042[2]
2	1.0	.047[2]	.032[2]
2	2.0	.042[2]	.012[2]
9	2.0	.030[1]	.010[1]
9	6.0	.020[1]	.010[1]
9	10.0	.030[1]	.020[1]

[1] Approximate 95% confidence interval $\pm$ .04.
[2] Approximate 95% confidence interval $\pm$ .02.

$(n - 4)]$ and variance $[2/(n - 4)]$, where $w = 1 - mF/n$, $a = [(m + n - 2)/(n - 2)]^{\frac{1}{3}}$, and $\rho = \cosh^{-1}[\lambda^2/a(n - 2) + a]$.

This transformation was discussed by Bargmann (1958) and Laubscher (1960). Both authors use a Taylor series expansion; Bargmann uses the second derivative terms for a correction on the expected value of Z, while Laubscher uses these for a modification on the transformation itself. Both authors present limited tables to show the approach to normality; Bargmann considers the approximation satisfactory, while Laubscher, citing both poor agreement with exact probabilities and the fact that the transformation may give no sensible approximation for small values of F, does not. More investigations are needed before this transformation is either finally rejected or accepted.

The results in Table 3 are quite similar to those obtained for the sequential χ^2 test. The empirical ASN are in the neighborhood of Bhate's conjecture ASN and are about $\frac{2}{3}$ of the fixed sample size. Very few samples did not reach a decision before the upper limit of available tables, indicating that presently available tables are sufficiently long.

Table 4 shows the empirical α and β errors; the desired probabilities were .05 for both. Because of the smaller sample sizes, most of the individual empirical probabilities are not statistically significantly different from .05; the results, however, appear very similar to those obtained for the χ^2 test.

BIBLIOGRAPHY

ARNOLD, K. J. (1951). Tables to facilitate sequential t-tests. Applied Math. Ser. No. 7, National Bureau of Standards, Washington, D. C.

BAKER, A. G. (1950). Properties of some tests in sequential analysis. *Biometrika* **37** 334–340.

BARGMANN, ROLF E. (1958). Some interpretations in the analysis of transformed data. Tech. Report, Dept. of Statistics, Virginia Polytechnic Inst., Blacksburg, Virginia.

FREUND, R. J. and JACKSON, J. E. (1960). Tables to facilitate multivariate sequential testing for means. Tech. Report, Dept. of Statistics, Virginia Polytechnic Inst., Blacksburg, Virginia.

HOFER, GLADYS F. (1960).A variance stabilizing transformation of the non-central χ^2 distribution. Unpublished Masters Thesis, Virginia Polytechnic Inst., Blacksburg, Virginia.

JACKSON, J. E. (1960). Bibliography on sequential analysis. *J. Amer. Statist. Assoc.* **55** 561–580.

JACKSON, J. E. and BRADLEY, R. A. (1959). Multivariate sequential procedures for testing means. Tech. Report, Dept. of Statistics, Virginia Polytechnic Inst., Blacksburg, Virginia.

JACKSON, J. EDWARD and BRADLEY, RALPH A. (1961a). Sequential χ^2 and T^2 tests. *Ann. Math. Statist.* **32** 1063–1077.

JACKSON, J. EDWARD and BRADLEY, RALPH A. (1961b) Sequential T^2 and χ^2 tests and their application to an acceptance sampling problem. *Technometrics* **3** 519–534.

LAUBSCHER, NICO F. (1960). Normalizing the noncentral t and F distributions. *Ann. Math. Statist.* **31** 1105–1112.

RUSHTON, S. (1952). On a two sided sequential t-test, *Biometrika* **39** 302–308.

WALD, A. (1947). *Sequential Analysis*. Wiley, New York.