RALIZATIONS OF BARTLETT'S AND HARTLEY'S TESTS OF HOMOGENEITY USING "OVERALL VARIABILITY"

Ashis SenGupta

Department of Statistics, Stanford University, California 94305 U.S.A.

AMS Subject Classification and Phrases: *likelihood ratio tests; standardized generalized variance.*

ABSTRACT

The generalized variance plays an important and useful role as a measure to compare overall variability of different populations in biological sciences (Goodman, 1968; Kocherlakota and Kocherlakota, 1983; Sokal, 1965). Here we present simple and elegant multivariate analogues to Bartlett's and Hartley's tests of homogeneity. Large sample distributions of the test statistics are presented and the practicefulness of the tests are demonstrated through several examples.

1. INTRODUCTION

Goodman (1968) has pointed out that the generalized variance "merits further investigation" and related statistical inference to be developed. This paper provides a step to meet that need. Suggested also by Sokal (1965), Wilks (1967) etc., the GV serves as useful measure to compare "overall variabililty" of different populations with regard to multiple characters as encountered in biological sciences. An excellent recent review on GV is given in Kocherlakota and Kocherlakota (1983).

Let X be a p-dimensional random vector variable with Cov(X) $= \Sigma$. Det $(\Sigma) = |\Sigma|$ is termed the GV. A further generalization of GV is felt necessary as seen, for example, from the following situations. (1) Reduction of dimensionality plays an important role in statistical analysis of biological data. Like Gnanadesikan (1977, p.77), SenGupta (1983) etc., one may be interested in making a choice between different sets of generalized canonical variables and that too of possibly different dimensions. (2) In case of vector observations, for some of which information on certain components are missing, one might have to restrict to those "only ... for which complete data were available" (Goodman, 1968, p. 191). This results in loss of data. Alternatively, in many cases, it may be reasonable to retain incomplete observations and compare data of different dimensions. For such situations where comparison of overall variability for populations of different dimensions are necessary, we propose as a generalization of GV, the standardized GV (SGV), $|\Sigma|^{1/p}$. We note that the SGV is a measure so scaled as to render it comparable to scatter for a scalar random variable and hence its magnitude is easier to comprehend.

We present generalizations of Bartlett's (Result 3, Section 2) and Hartley's shortcut (Section 4) tests of homogeneity of variances of several populations and the large sample distributions of the corresponding test statistics. Several examples from biological sciences are given to illustrate the usefulness and the simplicity of the proposed tests.

2. LIKELIHOOD RATIO TESTS FOR SGVs

Let $X \sim N_p(\mu, \Sigma)$. Throughout our discussion, unless otherwise stated, assume Σ to be non-singular. Denote the population SGV of X, $|\Sigma|^{1/p}$, by Δ^2 and that of the sample, $|S/N|^{1/p}$ by d^2 where S is the sample sums of products matrix based on a sample of size N. Also denote $|S|^{1/p}$ by s^2. [Note that, Anderson (1984) defines GV with the divisor N-1 instead of N].

Tests for SGVs of One and Two Independent Multivariate Normal Populations.

For the sake of completeness we present below Results 1 and

Result 1. The LRT for $H_o : \Delta^2 = \sigma_o^2$ against $H_1 : \Delta^2 \neq \sigma_o^2$ can equivalently given by,

Reject H_o iff $d^{2p}/\sigma_o^{2p} > a_o$ or $< a_1$,

a_o and a_1 are constants to be determined from the specified of the test.

Result 2. The LRT for $H_o : \Delta_1^2 = \Delta_2^2$ against $H_1 : \Delta_1^2 \neq \Delta_2^2$ can be equivalently given by,

Reject H_o iff $R = d_1^2/d_2^2 < r_1$ or $> r_2$

r_1 and r_2 are constants to be determined from the specified level of the test.

Proofs of the Results 1 and 2 and the exact distributions of the test statistics in terms of Special Functions are presented computable forms using the theory of Calculus of Residues (e.g. and Katiyar, 1979) in SenGupta (1981).

Test for the Equality of SGVs of k (>2) Independent Multivariate Normal Populations.

Let x_{it}, $t = 1, \ldots, N_i$, $i = 1, \ldots, k$ denote k random samples k independent populations $N_{p_i}(\mu_i, \Sigma_i)$, $i = 1, \ldots, k$ respectively. We are interested in testing $H_o : \Delta_i^2$, $i = 1, \ldots, k$ all equal, against the alternative H_1, that at least one of them differ. The MLE, under both H_o and H_1, of μ_i is $\bar{x}_i$, $i = 1, \ldots, k$. Let θ_{ij}, $= 1, \ldots, k$, $j = 1, \ldots, p_i$ be the characteristic roots of $\Sigma_i^{-1} S_i$ spectively where S_i, $i = 1, \ldots, k$ are the sample sums of products trices for X_i, $i = 1, \ldots, k$ respectively. For finding the MLEs Σ_i, $i = 1, \ldots, k$, under H_o, it suffices to consider,

$$\phi = C + \sum_{i=1}^{k} \sum_{j=1}^{p_i} \left[\frac{N_i}{2} \ell n\, \theta_{ij} - \frac{1}{2} \theta_{ij} \right] +$$

$$\sum_{i=1}^{k} \lambda_{ii+1} \left[\left(\sum_j \frac{1}{p_i} \ell n\, \theta_{ij} - \ell n\, s_i^2 \right) - \left(\sum_j \frac{1}{p_{i+1}} \ell n\, \theta_{i+1j} - \ell n\, s_{i+1}^2 \right) \right]$$

where C is a constant and λ_{ii+1} are undetermined Lagrange multipliers with $k+1$ being replaced by 1 in the suffixes. Differentiating ϕ with respect to θ_{ij}s and equating to zeros we have,

$$p_i N_i + (\lambda_{ii+1} - \lambda_{i-1i}) = p_i \theta_{ij} \; ; \; i = 1, \ldots, k \; , \; j = 1, \ldots, p_i \; , \; \lambda_{o1} \equiv \lambda_{k1}$$

$$\Rightarrow \theta_{ij} = \theta_{ij'} \Rightarrow \theta_{ij} = s_i^2 / \hat{\sigma}_o^2 \; , \; i = 1, \ldots, k \; ,$$

where $\hat{\sigma}_o^2$ is the MLE of σ_o^2, the common unknown value of Δ_i^2, $i = 1, \ldots, k$. So,

$$\hat{\sigma}_o^2 \sum_{i=1}^{k} p_i N_i + \hat{\sigma}_o^2 \left[\sum_{i=1}^{k} (\lambda_{ii+1} - \lambda_{i-1i}) \right] = \sum^{k} p_i s_i^2$$

$$\hat{\sigma}_o^2 = \sum_i p_i s_i^2 \Big/ \sum_i p_i N_i \; .$$

Note that $\hat{\sigma}_o^2$ agrees with the MLE for σ^2_o of the univariate case. Hence, we get,

Result 3. The LRT for $H_o : \Delta_i^2$, all equal, against H_1 : at least one of the Δ_i^2, $i = 1, \ldots, k$, differ is given by

$$\text{Reject } H_o \text{ if and only if , } \eta = \prod_{i=1}^{k} (d_i^2 / \hat{\sigma}_o^2)^{N_i p_i / 2} < \eta_o \; ,$$

where $\hat{\sigma}_o^2 = \Sigma \, p_i \, s_i^2 / \Sigma p_i N_i$ and η_o is a constant to be determined from the specified level of the test.

For the univariate case, with $p_i = 1$ and N_i replaced by $n_i = N_i - 1$, $i = 1, \ldots, k$, η reduces to the well-known Bartlett's statistic for testing homogeneity of variances of several independent normal populations. For the multivariate case, we propose below η_B^2 as a (Bartlett-type) modification of η^2,

$$\eta_B^2 = \prod_{i=1}^{k} (u_i^2)^{n_i p_i / \Sigma n_i p_i} \Big/ \left(\sum_{i=1}^{k} n_i p_i u_i^2 \Big/ \sum_{i=1}^{k} n_i p_i \right), \; n_i u_i^2 = N_i d_i^2 \; , \; i = 1, \ldots, k \; .$$

3. LARGE-SAMPLE APPROXIMATIONS

Some large sample approximations to the exact distributions of the test criteria considered above are now suggested. Existing approximations are also reviewed for the distributions of GV and SGV.

Asymptotic Distributions of GV and SGV.

Letting $nu^2 = Nd^2$, $n = N - 1$, we have from Anderson (1984), Theorem 7.5.4, that for large N,

$$\sqrt{n}(d^{2p}/\Delta^{2p} - 1) \xrightarrow{L} N(0, 2p).$$

It is known that $\eta = |S|/|\Sigma|$ is distributed as $\prod_{i=1}^{p} \chi^2_{N-i}$, where the χ^2's are all independent. Hoel (1937) suggested approximating the distribution of $w = \eta^{1/p}$ by the distribution with the density function

$$g(w) = C^{\frac{1}{2}p(N-p)} w^{\frac{1}{2}p(N-p)-1} e^{-Cw} / \Gamma[\,p(N-p)/2\,]$$

where

$$C \equiv C(p,N) = (p/2)[1 - \{(p-1)(p-2)/(2N)\}]^{1/p}$$

This turns out to be exact for $p = 1$ and $p = 2$.

Ghanadesikan and Gupta (1970) have suggested approximating the distribution of $\ln w = (1/p)\sum_{1}^{p} \ln \chi^2_{N-i}$, using the Central Limit Theorem, by the normal distribution.

We now propose a new approximation to the distribution of SGV. Application of the general result of Madansky and Olkin (1969) that $g(V) = |V|^{1/p}$ shows that, for large N,

$$\sqrt{N}(d^2/\Delta^2 - 1) \xrightarrow{L} N(0, 2/p).$$

In the light of this approximation to the distribution of the SGV, it is interesting to note the approximation to the distribution of GV that Anderson stated at the beginning.

Asymptotic Distributions of R and η

Denoting by C_i the $C(N_i, p_i)$ of Hoel's approximation, the density of R for large N_1 and N_2, can be approximated by that of,

$$[C_2 N_2 p_1 (N_1 - p_1)/C_1 N_1 p_2 (N_2 - p_2)]\, \delta^2 F_{p_1(N_1-p_1),\, p_2(N_2-p_2)},$$

where $\delta^2 = \Delta_1^2/\Delta_2^2$. The null and non-null distributions are obtained by setting $\delta^2 = 1$ and the specified value under the alternative hypothesis, respectively.

In addition to the usual χ^2 approximation to the likelihood ratio criterion η, another approximation is presented here. If N_i is large compared to p_i^2, $i = 1, \ldots, k$ then in the same lines of Hoel's approximation, we get $X_i = p_i n_i u_i^2 / \sigma_o^2$ can be approximated by a χ^2 variable with d.f. $p_i(N_i - p_i)$, $i = 1, \ldots, k$. Hence,

Lemma 3.2.1. If N_i is large compared to p_i^2, $i = 1, \ldots, k$, then the density of η_B^2 under H_o, can be approximated by $f(t)$ defined in Theorem 3.4.1 of SenGupta (1981) (where p_i's now can be any integers, not necessarily 1s or 2s only).

Similar result is seen to hold for η^2 also.

4. A MULTIVARIATE F_{max} CRITERION

A simpler statistic than η is now suggested for the special case when we have an equal number of observations, N, from k populations, each of equal dimension p. We propose the statistic

$$F_{p,max} = d^2_{max} / d^2_{min} .$$

For $p = 1$, this coincides with the F_{max} proposed by Hartley (1950) as a shortcut method for the univariate case. It is known that $\ln \chi^2_\nu$, for large ν, is approximately normal with variance $2/(\nu-1)$. Hence, $\ln d^2$ is approximately normal, for large N, with variance $\sum_{j=1}^{p} [2/p^2(N-j-1)]$. Thus the approximate percentage points of $F_{p,max}$ can be determined from

$$F_{p,max}(\alpha) = \exp\left[r_k(\alpha) \frac{1}{p} \left\{ \sum_{j=1}^{p} 2/(N-j-1) \right\}^{1/2}\right]$$

where $r_k(\alpha)$ is the $100\alpha\%$ point of the range r, in independent normal samples of size k. Tabulated values of $r_k(\alpha)$ are available from Pearson and Hartley (1956).

5. EXAMPLES

Sokal (1965) had suggested the use of GV to compare the overall variability of different populations in biological sciences. Following this suggestion, Goodman (1968) used the value of the sample GV as a descriptive measure for that purpose. Obviously, statistical tests

[illegible] for the required comparison and Examples 1 and 2 are [illegible] Goodman's data to which statistical tests of homogeneity [illegible]nces are now employed. Efforts for ranking and selection [illegible] be worthwhile if it is known that the populations do indeed [illegible]. Hence it seems reasonable that a preliminary test for homoge- [illegible] should precede the analysis for ranking and selection. This [illegible]rated through Example 3.

[illegible]mple 1. Based on different varieties of rice, Goodman (1968) [illegible]posed a grouping according to their sample GVs. This was [illegible]und to be consistent with the geographical and the other [illegible]nomic considerations. However, the need for a statistical [illegible] for such grouping is felt. Here, one may require that the popula- [illegible]s be the same for two varieties to belong to the same group. [illegible] observations each on X = (ear length, ear breadth), for the [illegible]eties Cateto Suline and Avanti Piching Ihu, we have

[illegible] $= d_1^2/d_2^2 = (.8686/.0961)^{1/2} = 3.01$.

[illegible] $H_o : \Delta_1^2 = \Delta_2^2$, $R \sim F_{2(45-2),\, 2(45-2)}$. Using this result and [illegible]sing equal tails, H_o is rejected at .01 level of significance. [illegible]he two varieties should belong to different classifications [illegible] concluded by Goodman (1968) using just the magnitudes of [illegible]s for the purpose. The d_i^2 values were made available at [illegible] State University.

Example 2. Consider again Goodman's data, now on cotton. [illegible] observed that the ". . . cotton populations indicate even [illegible] clearly that the generalized variance is a useful measure of [illegible] variability that merits further investigation." To compare [illegible]pulations statistically on the basis of their GVs, we proceed, [illegible] Example 1, to test the hypothesis of homogeneity of their [illegible] Eight characters: bract length, bract index, floral index, petal [illegible] androecium length, staminal index, boll length and branch angle [illegible] studied. From each of the three varieties of cotton: Gossypium [illegible]erse, Gossypium hirsutum and F_2, a sample of size 90 was [illegible]. Since the common sample size is quite large, it will be valid

to use the large-sample distribution of the $F_{p,max}$ test proposed in Section 4. Here, k=3, p=8 and N=90. Goodman (1968) has given the values of the ln GVs as -2.48, -2.79 and 12.93, respectively. Then,

$$F_{8,max} = \exp.[(12.93 + 2.79)/8] = \exp.(1.965) \text{ and}$$

$$F_{8,max}(.01) = \exp.\left[r_3(.01)\,\frac{1}{8}\left\{\sum_{j=1}^{8} 2/(90-j-1)\right\}^{1/2}\right] = \exp.(0.224).$$

Hence the three poplations differ, in terms of their GVs.

Example 3. Gnanadesikan and Gupta (1970) were interested in a ranking and selection procedure based on generalized variance. They considered 5(=p) - dimensional summaries of speech spectrographic data from a talker identification problem. The data consisted of 7(=N) replicate utterances of 10(=k) words for one particular speaker. Then,

$$F_{p,max} = (720616.4465/1.5411)^{1/5} = 13.6137 \text{ and}$$

$$F_{p,max}(.01) = 9.0737$$

Hence, the hypothesis of equal multidimensional scatter, as measured by SGV, is to be rejected.

6. REMARKS

As with any multidimensional measure, the SGV cannot be expected to be the unique measure best for all situations of multidimensional scatter. However, if we are interested in 'overall' scatter and where magnitude of individual variances separately are not of great concern, the SGV can be expected to perform adequately.

7. ACKNOWLEDGEMENTS

It's a pleasure to acknowledge the encouragements and helpful comments of Professors T. W. Anderson, C. R. Rao, and R.C. Srivastava. This research was partially supported by NSF Grant MCS 78-07736 and ONR Contract N00014-75-C-0442 at the Institute of Mathematical Studies in the Social Sciences, Stanford University.

[illegible] this paper was presented by the author, as an invited speaker, [illegible] Conference on Applications of Numerical Analysis and Special [illegible]s in Statistics, University of Maryland, October 6, 1980, [illegible]search was completed at the Indian Statistical Institute.

BIBLIOGRAPHY

[illegible]son, T.W. (1984). An Introduction to Multivariate Statistical [illegible]nalysis, John Wiley, New York.

[illegible]esikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations, John Wiley, New York.

[illegible]esikan, R. and Gupta, S.S. (1970). A selection procedure for multivariate normal distributions in terms of generalized variances, Technometrics, 12, 103-117.

[illegible]ar, M. (1968). A measure of 'overall variability' in populations, Biometrics, 24, 189-192.

[illegible]ley, H.O. (1950). The maximum F-ratio as a shortcut test for heterogeneity of variances, Biometrika, 37, 308-312.

[illegible], P.G. (1937). A significance test for component analysis, Annals of Mathematical Statistics, 8, 149-158.

[illegible]akota, S. and Kocherlakota, K. (1983). Generalized Variance, in Encyclopedia of Statistical Sciences, Vol. 3, (Johnson, N. L. and Kotz, S. eds.) John Wiley, New York, 354-357.

[illegible]sky, A. and Olkin, I. (1969). Approximate confidence regions for constraint parameters, Multivariate Analysis, I (Krishnaiah, P.R. ed.) Academic Press, New York.

[illegible] A.M. and Katiyar, R.S. (1979). Exact percentage points for testing independence, Biometrika, 66, 353-356.

[illegible]son, E.S. and Hartley, H.O. (1956). Biometrika Tables for Statisticians, The Syndics of the Cambridge University Press, Cambridge.

Gupta, A. (1981). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions, Tech. Rep. 50, Dept. of Statistcs, Stanford University, also to appear in The Journal of Multivariate Analysis.

Gupta, A. (1983). Generalized Canonical Variables, Encyclopedia of Statistical Sciences, Vol. 3, (Johnson, N.L. and Kotz, S. eds.) John Wiley, New York, 123-126.

Sokal, R.R. (1965). Statistical methods in systematics, Biological Review of the Cambridge Philosophical Society, 40, 337-391.

Wilks, S.S. (1967). Multidimensional statistical scatter, In Collected Papers; Contributions to Mathematical Statistics, (Anderson, T.W. ed) 597-614.

Received by Editorial Board member.

Recommended by S. Kocherlakota, University of Manitoba, Winnipeg, Canada.

Refereed Anonymously.