Multivariate Statistical Inference under Marginal Structure, II

by

Leon Jay Gleser*
Purdue University

Ingram Olkin
Stanford University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #325

July 1973

*Dr. Gleser's research sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Contract No. F44620-70-C-0066 at Johns Hopkins University and AFOSR Grant No. 72-2432 at Purdue University. The United States Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Multivariate Statistical Inference under Marginal Structure, II.

by

Leon Jay Gleser

Purdue University

and

Ingram Olkin

Stanford University

1. Introduction.

A previous paper [Gleser and Olkin (1972)] dealt with statistical inference problems in the context of an experimental design in which k randomly chosen groups of individuals from the same population of individuals are asked to take different psychological tests under identical testing conditions. The k tests $(T_0,T_1),(T_0,T_2),\ldots,(T_0,T_k)$ have one subtest T_0 in common. It is desired to test whether $(T_0,T_1),(T_0,T_2),\ldots,(T_0,T_k)$ are parallel forms of the same test.

The experimental design described above is a natural one for ongoing testing programs such as the Scholastic Aptitude Test (SAT), where forms must be changed from year to year (so that new items must be introduced and validated), yet experimentation with new forms and administration of old forms are done simultaneously. The design described above might be used in a given year to validate items for future years.

Another experimental design which has potential use in ongoing testing programs is one involving a certain hierarchical structure. Because this design is more complex than the design mentioned above, we confine our discussion to the case where 3 groups of individuals are randomly chosen from a given population of individuals and are tested under identical conditions. The three groups are given tests of the form

$$(1.1) \qquad (T_0, U_1, V_1),\ (T_0, U_1, V_2),\ (T_0, U_3, V_3) \ ,$$

respectively. Here T_0 is a subtest common to all three tests, U_1 is common to the first two groups, and V_1, V_2, V_3 are new subtests. The tests are characterized by $r = r_0 + r_1 + r_2$ scores, r_0 scores on subtest T_0, r_1 scores on subtest U_1 or U_3, and r_2 scores on subtest V_1, V_2, or V_3.

Let $x_0^{(g)}$, $x_1^{(g)}$, $x_2^{(g)}$ be the scores of a typical individual in the g-th group on subtests T_0, U_g, V_g respectively (where $U_g = U_1$, $g = 1,2$). By our assumptions about the subtests, $x_0^{(g)}$ is an r_0-dimensional (row) vector, $x_1^{(g)}$ is an r_1-dimensional (row) vector, and $x_2^{(g)}$ is an r_2-dimensional (row) vector. Thus,

$$x^{(g)} = (x_0^{(g)}, x_1^{(g)}, x_2^{(g)})$$

is an r-dimensional (row) vector, $r = r_0 + r_1 + r_2$. We assume that $x^{(g)}$ has a multivariate normal distribution with mean vector

$$(1.3) \qquad \mu^{(g)} = (\mu_0^{(g)}, \mu_1^{(g)}, \mu_2^{(g)}) \ ,$$

and covariance matrix

$$
(1.4) \qquad \Sigma^{(g)} = \begin{pmatrix} \Sigma_{00}^{(g)} & \Sigma_{01}^{(g)} & \Sigma_{02}^{(g)} \\ \Sigma_{10}^{(g)} & \Sigma_{11}^{(g)} & \Sigma_{12}^{(g)} \\ \Sigma_{20}^{(g)} & \Sigma_{21}^{(g)} & \Sigma_{22}^{(g)} \end{pmatrix} ,
$$

where the blocking of $\mu^{(g)}$ and $\Sigma^{(g)}$ conforms to the blocking of $x^{(g)}$. That is, $\mu_j^{(g)}$ is $1 \times r_j$, $j = 0,1,2$, and $\Sigma_{jk}^{(g)}$ is $r_j \times r_k$, $j,k = 0,1,2$.

With this background in mind, we are interested in using observations obtained from the individuals tested to determine whether the 3 forms (T_0,U_1,V_1), (T_0,U_1,V_2), and (T_0,U_3,V_3) are parallel forms of the same psychological test. If the 3 forms are parallel, then the parameters $\mu^{(1)}$, $\mu^{(2)}$, $\mu^{(3)}$, $\Sigma^{(1)}$, $\Sigma^{(2)}$, $\Sigma^{(3)}$ satisfy the null hypothesis:

$$
(1.5) \qquad H: \quad \mu^{(1)} = \mu^{(2)} = \mu^{(3)} \quad \underline{\text{and}} \quad \Sigma^{(1)} = \Sigma^{(2)} = \Sigma^{(3)} .
$$

If the 3 forms are not parallel, then the construction of our experimental design assures us that at least the following relationships hold among the parameters:

$$
(1.5) \qquad A: \begin{cases} \mu_0^{(1)} = \mu_0^{(2)} = \mu_0^{(3)}, \quad \mu_1^{(1)} = \mu_1^{(2)}, \\ \Sigma_{00}^{(1)} = \Sigma_{00}^{(2)} = \Sigma_{00}^{(3)}, \quad \Sigma_{01}^{(1)} = \Sigma_{01}^{(2)}, \quad \Sigma_{11}^{(1)} = \Sigma_{11}^{(2)} \end{cases} .
$$

Thus, statistical verification of the hypothesis that the three forms (T_0,U_1,V_1), (T_0,U_1,V_2), (T_0,U_3,V_3) are parallel takes the form of a test of the null hypothesis H against the alternative hypothesis A.

In Section 2, the likelihood ratio test statistic for our hypothesis testing problem is derived. To carry out the test we use an asymptotic chi-square test. In Section 4, we show how our approach to deriving a test of the null hypothesis H can be used to construct statistical tests of hypothesis for the parallelism of test forms under various experimental designs related to those considered here and in the previous paper [Gleser and Olkin (1972)]. In Section 3 we provide an example in which some of the results of Section 2 are applied.

2. The Likelihood Ratio Test Statistic.

Assume that N_g individuals take the psychological test (T_0,U_g,V_g), $g = 1,2,3$. Let $x_i^{(g)} = (x_{i0}^{(g)},x_{i1}^{(g)},x_{i2}^{(g)})$ be the vector of scores of the i-th individual who takes the test (T_0,U_g,V_g), $i = 1,2,\ldots,N_g$; $g = 1,2,3$. Under our experimental design, we may assume that the score vectors $x_i^{(g)}$ are mutually statistically independent, $i = 1,2,\ldots,N_g$; $g = 1,2,3$. In this case, we need not consider all of the data, but may reduce our consideration to the sufficient statistic

$$(\underset{\sim}{\bar{x}},\underset{\sim}{V}) = (\bar{x}^{(1)},\bar{x}^{(2)},\bar{x}^{(3)},V^{(1)},V^{(2)},V^{(3)}) \quad ,$$

where

$$(2.1)\qquad \bar{x}^{(g)} = \frac{1}{N_g}\sum_{i=1}^{N_g}(x_{i0}^{(g)},x_{i1}^{(g)},x_{i2}^{(g)}) = (\bar{x}_0^{(g)},\bar{x}_1^{(g)},\bar{x}_2^{(g)}) \quad ,$$

and

$$(2.2)\qquad V^{(g)} = \begin{pmatrix} V_{00}^{(g)} & V_{01}^{(g)} & V_{02}^{(g)} \\ V_{10}^{(g)} & V_{11}^{(g)} & V_{12}^{(g)} \\ V_{20}^{(g)} & V_{21}^{(g)} & V_{22}^{(g)} \end{pmatrix} \quad ,$$

with

$$V_{ab}^{(g)} = \sum_{i=1}^{N_g} (x_{ia}^{(g)} - \bar{x}_a^{(g)})' \, (x_{ib}^{(g)} - \bar{x}_{ib}^{(g)}) \ ,$$

$a,b = 0,1,2;$ $g = 1,2,3$. The statistics $\bar{x}^{(1)}, \bar{x}^{(2)}, \bar{x}^{(3)}, V^{(1)}, V^{(2)}, V^{(3)}$ are mutually statistically independent. Further, $\bar{x}^{(g)}$ has an r-variate normal distribution with mean vector $\mu^{(g)}$ and covariance matrix $(N_g)^{-1}\Sigma^{(g)}$, and $V^{(g)}$ has a Wishart distribution with degrees of freedom $n_g \equiv N_g - 1$ and parameter $E(n_g^{-1}V^{(g)}) = \Sigma^{(g)}$; $g = 1,2,3$. From the above facts the joint density function $p(\bar{\underset{\sim}{x}},\underset{\sim}{V}|\underset{\sim}{\mu},\underset{\sim}{\Sigma})$ of the sufficient statistics $(\bar{\underset{\sim}{x}},\underset{\sim}{V})$ can be directly obtained, and can be exhibited as a function of the unknown parameters

$$\underset{\sim}{\mu} = (\mu^{(1)},\mu^{(2)},\mu^{(3)}), \qquad \underset{\sim}{\Sigma} = (\Sigma^{(1)},\Sigma^{(2)},\Sigma^{(3)}) \quad .$$

To obtain the likelihood ratio test statistic (LRTS) of H versus A, we could proceed ab initio to obtain the maximum of $p(\underset{\sim}{\bar{x}},\underset{\sim}{V}|\underset{\sim}{\mu},\underset{\sim}{\Sigma})$ over the parameters $\underset{\sim}{\mu}$, $\underset{\sim}{\Sigma}$, both under the restrictions (1.4) on $\underset{\sim}{\mu}$, $\underset{\sim}{\Sigma}$ imposed by H, and under the restrictions (1.5) on $\underset{\sim}{\mu}$, $\underset{\sim}{\Sigma}$ imposed by A. The ratio of these maxima

$$\lambda = \frac{\max_H p(\underset{\sim}{\bar{x}},\underset{\sim}{V}|\underset{\sim}{\mu},\underset{\sim}{\Sigma})}{\max_A p(\underset{\sim}{\bar{x}},\underset{\sim}{V}|\underset{\sim}{\mu},\underset{\sim}{\Sigma})} \tag{2.3}$$

is then the LRTS for testing H versus A. The null hypothesis H is rejected when λ is smaller than a predetermined constant λ^*, where λ^* is chosen so as to give a desired level of significance α for the test.

However, the determination of the LRTS is more easily accomplished by using the fact that λ is the product of the LRTS λ_1 for testing H versus the alternative

$$H^*: \begin{cases} (\mu_0^{(1)},\mu_1^{(1)}) = (\mu_0^{(2)},\mu_1^{(2)}) = (\mu_0^{(3)},\mu_1^{(3)}) , \\ \begin{pmatrix} \Sigma_{00}^{(1)} & \Sigma_{01}^{(1)} \\ \Sigma_{10}^{(1)} & \Sigma_{11}^{(1)} \end{pmatrix} = \begin{pmatrix} \Sigma_{00}^{(2)} & \Sigma_{01}^{(2)} \\ \Sigma_{10}^{(2)} & \Sigma_{11}^{(2)} \end{pmatrix} = \begin{pmatrix} \Sigma_{00}^{(3)} & \Sigma_{01}^{(3)} \\ \Sigma_{10}^{(3)} & \Sigma_{11}^{(3)} \end{pmatrix} , \end{cases} \tag{2.4}$$

and the LRTS λ_2 for testing H* versus A (Anderson (1958), Lemma 10.3.1). The hypothesis testing problems that give rise to λ_1 and λ_2 are of a form for which the LRTS has already been obtained [Gleser and Olkin (1972)]; thus, putting together the solutions of these two component testing problems immediately yields the desired LRTS λ.

2.1. The likelihood ratio test statistic λ_1.

Turning first to the test of H versus the alternative H^*, we see by comparing (1.4) and (2.4) that under both hypotheses the subvectors $(x_0^{(g)}, x_1^{(g)})$ of scores on (T_0, U_g) have identical marginal distributions, $g = 1,2,3$. Thus, without loss of statistical generality, we can regard (T_0, U_1) and (T_0, U_3) as being identical subtests $Z = (T,U)$. Hence, we have three psychological tests (Z,V_1), (Z,V_2), (Z,V_3), and want to test whether these three tests are parallel forms of the same psychological test. In the notation of Gleser and Olkin (1972), henceforth abbreviated G-O, this last hypothesis testing problem is to test the null hypothesis H_{mvc} that all three tests (Z,V_1), (Z,V_2), (Z,V_3) have identically distributed score vectors against the alternative hypothesis $H_{m'vc'}$ that only the score subvectors on the subtest Z are identically distributed. From the results obtained in G-O, the LRTS for this problem is

$$\lambda_1 = \frac{\left(\prod_{g=1}^{3} \left|\frac{1}{N_g} V_{22.0,1}^{(g)}\right|^{N_g/2}\right)\left(\left|\frac{1}{N}(I_{r_0+r_1},0)(W+A)(I_{r_0+r_1},0)'\right|^{N/2}\right)}{\left|\frac{1}{N}(W+A)\right|^{N/2}} , \tag{2.5}$$

where $N = N_1+N_2+N_3$, I_p is the $p \times p$ identity matrix,

$$W = V^{(1)} + V^{(2)} + V^{(3)}, \quad \bar{\bar{x}}. = \frac{1}{N}\sum_{g=1}^{3} N_g \bar{x}^{(g)} ,$$

$$A = \sum_{g=1}^{3} N_g(\bar{x}^{(g)} - \bar{\bar{x}})'(\bar{x}^{(g)} - \bar{\bar{x}}) , \tag{2.6}$$

and

$$V_{22.0,1}^{(g)} = V_{22}^{(g)} - (V_{20}^{(g)},V_{21}^{(g)}) \begin{pmatrix} V_{00}^{(g)} & V_{01}^{(g)} \\ V_{10}^{(g)} & V_{11}^{(g)} \end{pmatrix}^{-1} (V_{20}^{(g)},V_{21}^{(g)})' .$$

Note that

$$\begin{pmatrix} V_{00}^{(g)} & V_{01}^{(g)} \\ V_{10}^{(g)} & V_{11}^{(g)} \end{pmatrix} = (I_{r_0+r_1},0)\, V^{(g)} (I_{r_0+r_1},0)' ,$$

so that

$$\left|\frac{1}{N_g} V^{(g)}\right| = \left|\frac{1}{N_g} V_{22.0,1}^{(g)}\right| \left|\frac{1}{N_g} (I_{r_0+r_1},0)\, V^{(g)} (I_{r_0+r_1},0)'\right| .$$

Thus, we may rewrite (2.5) in the form

$$(2.7) \qquad \lambda_1 = \frac{\left|\frac{1}{N} (I_{r_0+r_1},0)(W+A)(I_{r_0+r_1},0)'\right|^{N/2} \prod_{g=1}^{3} \left|\frac{1}{N_g} V^{(g)}\right|^{N_g/2}}{\left|\frac{1}{N} (W+A)\right|^{N/2} \prod_{g=1}^{3} \left|\frac{1}{N_g} (I_{r_0+r_1},0)\, V^{(g)} (I_{r_0+r_1},0)'\right|^{N_g/2}} .$$

2.2. The likelihood ratio test statistic λ_2.

Comparing (2.4) to (1.5), we see that both the hypothesis H^* and the alternative hypothesis A place restrictions only on the parameters of the marginal distributions of the test score subvectors $(x_0^{(g)},x_1^{(g)})$,

$g = 1,2,3$. Under these conditions, it is straightforwardly shown that the LRTS λ_2 for testing H^* versus A can be found by reducing consideration to that part of the sufficient statistic $\bar{x}_0^{(g)}, \bar{x}_1^{(g)}, V_{00}^{(g)}, V_{01}^{(g)}, V_{11}^{(g)}$ formed from the test score subvectors $(x_{i0}^{(g)}, x_{i1}^{(g)})$, $i = 1,2,\ldots,N_g$; $g = 1,2,3$. That is, to find λ_2 we can act as if only $(x_{i0}^{(g)}, x_{i1}^{(g)})$ are observed, $i = 1,2,\ldots,N_g$; $g = 1,2,3$. Under that assumption, we see from (1.5) and (2.4) that $(x_0^{(1)}, x_1^{(1)})$ and $(x_0^{(2)}, x_1^{(2)})$ have identical distributions under H^* and A, and thus for testing H^* versus A, groups 1 and 2 can be combined into one group without loss of statistical generality. The testing problem now becomes one of determining whether (T_0, U_1) and (T_0, U_3) are parallel forms of the same psychological test, where scores on (T_0, U_1) are obtained from $N_1 + N_2$ individuals, and scores on (T_0, U_3) are obtained from N_3 individuals. From G-O, the LRTS for this problem (which, in the notation of G-O, is a LRTS of the form $\lambda_{mvc,m'vc'}$) is

$$(2.8) \qquad \lambda_2 = \frac{\left|\frac{1}{N}(I_{r_0},0)(W+B)(I_{r_0},0)'\right|^{N/2} \left|\frac{1}{N_1+N_2}\left(V^{(1)}+V^{(2)}\right)_{11.0}\right|^{\frac{(N_1+N_2)}{2}} \left|\frac{1}{N_3} V_{11.0}^{(3)}\right|^{\frac{N_3}{2}}}{\left|\frac{1}{N}(I_{r_0+r_1},0)(W+B)(I_{r_0+r_1},0)'\right|^{N/2}}$$

where

$$(2.9) \quad B = (N_1+N_2)\left(\frac{N_1}{N_1+N_2}\bar{x}_1 + \frac{N_2}{N_1+N_2}\bar{x}_2 - \bar{\bar{x}}\right)\left(\frac{N_1}{N_1+N_2}\bar{x}_1 + \frac{N_2}{N_1+N_2}\bar{x}_2 - \bar{\bar{x}}\right)'$$
$$+ N_3(\bar{x}_3 - \bar{\bar{x}})(\bar{x}_3 - \bar{\bar{x}})' \quad ,$$

and where for any matrix C,

$$(2.10) \qquad C = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix} ,$$

blocked in the manner of $V^{(1)}, V^{(2)}$, etc., we have

$$(2.11) \qquad C_{11.0} = C_{11} - C_{10} C_{00}^{-1} C_{01} .$$

Recall that for C as in (2.10),

$$(2.12) \qquad |C_{11.0}| = \frac{\left|\begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix}\right|}{|C_{00}|} = \frac{|(I_{r_0+r_1},0)C(I_{r_0+r_1},0)'|}{|C_{00}|} .$$

Using this fact, we can modify (2.8) into a form more suitable for computation:

$$(2.13) \quad \lambda_2 = \frac{\left|\frac{1}{N}(I_{r_0},0)(W+B)(I_{r_0},0)'\right|^{N/2} \left|\frac{1}{N_3}(I_{r_0+r_1},0)V^{(3)}(I_{r_0+r_1},0)'\right|^{N_3/2}}{\left|\frac{1}{N}(I_{r_0+r_1},0)(W+B)(I_{r_0+r_1},0)'\right|^{N/2} \left|\frac{1}{N_3} V_{00}^{(3)}\right|^{N_3/2}}$$

$$\times \frac{\left|\frac{1}{N_1+N_2}(I_{r_0+r_1},0)(V^{(1)}+V^{(2)})(I_{r_0+r_1},0)'\right|^{(N_1+N_2)/2}}{\left|\frac{1}{N_1+N_2}(V_{00}^{(1)}+V_{00}^{(2)})\right|^{(N_1+N_2)/2}} .$$

2.3. Calculation of the LRTS λ.

A comparison of (2.7) and (2.13) reveals that in calculating $\lambda = \lambda_1 \lambda_2$, there is little cancellation between terms in λ_1 and terms in λ_2. Thus, one reasonable way to compute λ is to first compute λ_1 and λ_2 separately, and then multiply λ_1 and λ_2 to obtain λ. However, to take advantage of the one cancellation that does occur between terms in λ_1 and λ_2, we recommend first calculating $\lambda_1 u$ and (λ_2/u), where

$$u = \left|\frac{1}{N_3} (I_{r_0+r_1},0) V^{(3)} (I_{r_0+r_1},0)'\right|^{N_3/2} ;$$

the LRTS λ can then be computed as the product of $\lambda_1 u$ and (λ_2/u).

In setting up the matrices $V^{(1)}, V^{(2)}, V^{(3)}, W, A, V^{(1)}+V^{(2)}$, and B for computation of λ_1 and λ_2, it is worth noting that some computational effort can be saved by taking advantage of the relationship

$$A = B + (N_1 + N_2)^{-1} N_1 N_2 (\bar{x}^{(1)} - \bar{x}^{(2)})' (\bar{x}^{(1)} - \bar{x}^{(2)})$$

holding between A and B (compare (2.6) and (2.9)).

The rejection region for the null hypothesis H based upon the test statistic λ is of the form $\lambda < \lambda^*$, where λ^* is a constant chosen so that the test of H versus A based on λ has the desired level of significance α. If we use the asymptotic approximation $-2\log\lambda \sim \chi^2_f$, the degrees of freedom, f, for the approximation is equal to:

$$f = (r_0+2r_1+3r_2) + \frac{r_0(r_0+1)}{2} + 2r_0r_1 + 3r_0r_2 + \frac{2r_1(r_1+1)}{2} + 3r_1r_2 + \frac{3r_2(r_2+1)}{2} - r - \frac{r(r+1)}{2}$$

(2.14)

$$= \frac{3}{2} r_1 + 3r_2 + r_0r_1 + 2r_0r_2 + 2r_1r_2 + \frac{1}{2} r_1^2 + r_2^2 .$$

2.4. A Bartlett Modification.

In a somewhat different hypothesis testing context (that of testing homogeneity of variances), Bartlett (1937) suggested that the small sample behavior of the LRT might be improved by assuming in the calculation of the LRTS that the number of observations taken in each population equals the degrees of freedom left after estimating the various nuisance parameters. This idea was applied to broader testing contexts by Anderson (1958), and more recently by Gleser and Olkin (1972). Following the arguments in G-O, the Bartlett modification of λ_1 would replace N_g by $N_g-r_0-r_1-1$ wherever N_g explicitly appears in (2.7), $g = 1,2,3$. The Bartlett modification of λ_2 would replace N_1+N_2 by $N_1+N_2-r_0-1$ and N_3 by N_3-r_0-1 wherever N_1+N_2 and N_3 explicitly appear in (2.13). To find the appropriate modification of $\lambda = \lambda_1\lambda_2$, we follow a rule implicitly used by Anderson (1958) and G-O, and replace N_g in the formula for λ by the smaller of the substituted sample sizes (degrees of freedom) used in the modifications of λ_1 and λ_2, $g = 1,2,3$. In this particular problem, this rule means that N_g is replaced by $N_g-r_0-r_1-1$ wherever N_g explicitly appears in the formulas for λ_1 and λ_2, $g = 1,2,3$. Under this modification, λ_1 is replaced by

$$(2.15) \quad L_1 = \frac{\left|\frac{1}{m}(I_{r_0+r_1},0)(W+A)(I_{r_0+r_1},0)'\right|^{\frac{m}{2}} \prod_{g=1}^{3}\left|\frac{1}{m_g}V^{(g)}\right|^{\frac{m_g}{2}}}{\left|\frac{1}{m}(W+A)\right|^{\frac{m}{2}} \prod_{g=1}^{3}\left|\frac{1}{m_g}(I_{r_0+r_1},0)V^{(g)}(I_{r_0+r_1},0)'\right|^{\frac{m_g}{2}}},$$

where $m_g = N_g-r_0-r_1-1$, $g = 1,2,3$, and

$$m = m_1+m_2+m_3 = N - 3(r_0+r_1+1) \quad .$$

Similarly λ_2 is replaced by

$$(2.16) \qquad L_2 = \frac{\left|\frac{1}{m}(I_{r_0},0)(W+B)(I_{r_0},0)'\right|^{\frac{m}{2}} \left|\frac{1}{m_3}(I_{r_0+r_1},0)V^{(3)}(I_{r_0+r_1},0)'\right|^{\frac{m_3}{2}}}{\left|\frac{1}{m}(I_{r_0+r_1},0)(W+B)(I_{r_0+r_1},0)'\right|^{\frac{m}{2}} \left|\frac{1}{m_3}V_{00}^{(3)}\right|^{\frac{m_3}{2}}}$$

$$\times \frac{\left|\frac{1}{m_1+m_2}(I_{r_0+r_1},0)(V^{(1)}+V^{(2)})(I_{r_0+r_1},0)'\right|^{\frac{(m_1+m_2)}{2}}}{\left|\frac{1}{m_1+m_2}(V_{00}^{(1)}+V_{00}^{(2)})\right|^{\frac{(m_1+m_2)}{2}}} \quad ,$$

and the Bartlett modification of λ is

$$(2.17) \qquad L = L_1L_2 = (L_1u^*)(L_2/u^*) \quad ,$$

where

$$u^* = \left|\frac{1}{m_3}(I_{r_0+r_1},0)V^{(3)}(I_{r_0+r_1},0)'\right|^{m_3/2} \quad .$$

2.5. The rejection region for H based on L

The appropriate rejection region for the null hypothesis H based upon the test statistic L is of the form $L < L^*$, where L^* is a constant chosen so that the test of H has a desired level of significance α. The form of this rejection region parallels the form of the rejection region for the test of H versus A based on the LRTS λ. Indeed, when $N_1 = N_2 = N_3 = K$, then $L = \lambda^{(K-r_0-r_1-1)/K}$, and thus when L^* is set equal to $(\lambda^*)^{(K-r_0-r_1-1)/K}$, λ and L define equivalent α-level tests of H. On the other hand, when N_1, N_2, N_3 are not all equal to one another, L and λ are not monotonically related, and thus do not define equivalent α-level tests of H.

To carry out the test of H versus A based on L it is necessary to find the value of the cut-off point L^* that will provide the desired level of significance α. Unfortunately, the distributional computations necessary in small samples to determine L^* are extremely complicated, and the results are still in an incomplete state. A similar comment applies to computation of the cut-off point λ^* that makes the test with rejection region $\lambda < \lambda^*$ have level α in small samples.

When N_1, N_2, N_3 are all of reasonably large magnitude, a large-sample approximation can be used to find L^*, namely

$$L^* = \exp(-\tfrac{1}{2}\chi_f^2(\alpha)), \tag{2.18}$$

where $\chi_f^2(\alpha)$ is the $(1-\alpha)$th fractile ($100(1-\alpha)$th percentile) of the χ_f^2 distribution, and f is given by (2.14). Note that this large-sample approximation is identical to the large-sample approximation

$$\lambda^* = \exp\ (-\tfrac{1}{2}\chi_f^2(\alpha)) \tag{2.19}$$

for λ^* (see Section 2.3). This, of course, is not surprising since L and λ have the same limiting distribution as $N_g \to \infty$, $g = 1, 2, 3$. It should be noted, however, that in moderate samples the rejection regions $\lambda < c$ and $L < c$ are not the same, regardless of the constant c, $0 < c < 1$. For example, when $N_1 = N_2 = N_3 = K$, then $L = \lambda^{(K-r_0-r_1-1)/K}$, so that unless $(K-r_0-r_1-1)/K$ is close enough to 1, it is possible for the tests with the respective rejection regions $\lambda < c$ and $L < c$ to have a positive probability (under H and A) of reaching different conclusions.

3. An Illustrative Example

For the purposes of demonstrating application of the test statistic λ obtained in Section 2, data was obtained from 300 randomly selected answer sheets taken from the April, 1971 administration of the Scholastic Aptitude Test. By selecting various sections from the Scholastic Aptitude Test (SAT), three different test forms (actually test sub-forms, since not all sections were used) were constructed. Each form was equally represented (100 observations) in the sample. All constructed test forms had a common verbal section (subtest T_0). Two of the forms had a common mathematics section (subtest U_1), while the third form used a different mathematics section of the same SAT (subtest U_3). The equating items from the SAT were combined into a third section which we can call the "equating section", which was assumed to differ among the three forms (thus creating subtests V_1, V_2, V_3). Each section (T, U, V) on a given constructed form (subform) was summarized by a single score: T was summarized by x_0, U by x_1, and V by x_2. If the three forms (subforms) of the SAT are parallel, the parameters of the joint distributions of these subtest scores on each of the three forms must satisfy the null hypothesis H. To test this null hypothesis against the alternative A of non-parallelism, we use the LRTS λ (Section 2) and reject H when

$$\lambda < \exp\{-\tfrac{1}{2}\chi_f^2(\alpha)\},$$

or equivalently when

$$-2\log\lambda > \chi_f^2(\alpha), \tag{3.1}$$

where α is the desired level of significance, f is given by (2.14), and $\chi_f^2(\alpha)$ is the (1-α)th fractile of the χ_f^2 distribution.

In the context of the given problem, $r_0 = r_1 = r_2 = 1$, $r = 3$, $N_1 = N_2$

$= N_3 = 100$, and $N = 300$. The values of the sufficient statistic $(\bar{\underset{\sim}{x}}, \underset{\sim}{V})$ are given in Table 1.

Table 1. Summary of the Test Score Data

$$\bar{x}^{(1)} = (14.44, \quad 12.64, \quad 14.66),$$

$$\bar{x}^{(2)} = (13.78, \quad 12.86, \quad 14.54),$$

$$\bar{x}^{(3)} = (14.41, \quad 10.36, \quad 14.74),$$

$$V^{(1)} = \begin{pmatrix} 6622.64 & 5260.84 & 5992.96 \\ 5260.84 & 7525.04 & 5275.76 \\ 5992.96 & 5275.76 & 7126.44 \end{pmatrix},$$

$$V^{(2)} = \begin{pmatrix} 4563.16 & 2965.92 & 4383.88 \\ 2965.92 & 5248.04 & 2985.56 \\ 4383.88 & 2985.56 & 5746.84 \end{pmatrix},$$

$$V^{(3)} = \begin{pmatrix} 6086.19 & 3055.24 & 5199.66 \\ 3055.24 & 3545.04 & 2940.36 \\ 5199.66 & 2940.36 & 6073.24 \end{pmatrix}.$$

From the data in Table 1, we find that

$$(3.2) \qquad -2 \log \lambda = 34.02.$$

Since $r_0 = r_1 = r_2 = 1$, we find from Equation (2.14) that $f = 11$. Since $\chi^2_{11}(.005) = 26.8$, it follows from (3.1) and (3.2) that the null hypothesis of parallelism of the three constructed SAT subforms is rejected at the 0.5% level of significance.

The rejection of the parallelism hypothesis H in this particular example occurs largely because of the difference between the average scores of the examinees on the mathematics subtests U_1 and U_3 (see Table 1). Since in our construction of these subtests we used two different mathematics sections of the <u>same</u> SAT form, and since these two sections are not supposed to be interchangeable (they presumably are designed to test different abilities or levels of ability), the obtained differences are not surprising. The rejection of the parallelism hypothesis thus merely reflects the artificial way in which the three subforms in our example were constructed, and should not be taken as an indication of any lack of parallelism of the forms actually used in administering the SAT.

4. Generalizations

The techniques of Section 2 can straightforwardly be extended to treat hypothesis testing problems in which scores on G forms of a test are compared to one another in an attempt to determine if the G forms are parallel. To present the relevant likelihood ratio test theory, however, we need to change our notation, in order that the results may be given in a compact form.

We assume that each test form consists of G subtests: S(1), S(2),..., S(G). Subtest S(i) has i different versions: $S_1(i)$, $S_{G-i+2}(i)$,..., $S_{G-1}(i)$, $S_G(i)$, where $S_1(i)$ is common to test forms 1,..., G-i+1, and version $S_j(i)$ appears in test form j, j = G-i+2,..., G. Thus, subtest S(1) has only one version $S_1(1)$ which appears in all test forms. Subtest S(2) has 2 versions $S_1(2)$ and $S_G(2)$, with $S_1(2)$ appearing in test forms 1, 2,..., G-1 and $S_G(2)$ appearing only in test form G. Finally, S(G) has G different versions $S_1(G)$, $S_2(G)$,..., $S_G(G)$, each version appearing in one and only one test form. The assignment of subtest versions to test forms is illustrated in Figure 1.

	Subtest 1	Subtest 2	Subtest 3	...	Subtest G-1	Subtest G
Test Form 1	$S_1(1)$	$S_1(2)$	$S_1(3)$	...	$S_1(G-1)$	$S_1(G)$
Test Form 2	$S_1(1)$	$S_1(2)$	$S_1(3)$	...	$S_1(G-1)$	$S_2(G)$
Test Form 3	$S_1(1)$	$S_1(2)$	$S_1(3)$	...	$S_3(G-1)$	$S_3(G)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Test Form G-1	$S_1(1)$	$S_1(2)$	$S_{G-1}(3)$	...	$S_{G-1}(G-1)$	$S_{G-1}(G)$
Test Form G	$S_1(1)$	$S_G(2)$	$S_G(3)$	...	$S_G(G-1)$	$S_G(G)$

Figure 1. Hierarchical assignment of subtest forms to the G test forms.

Let each of the G test forms be characterized by $r = \sum_{g=1}^{G} r_g$ scores: r_1 on subtest S(1), r_2 on subtest S(2),..., r_G on subtest S(G). Let $x_1^{(g)}, x_2^{(g)},\ldots, x_G^{(g)}$ be the scores of a typical individual on subtests S(1), S(2),..., S(G), respectively, of test form g. Hence $x_i^{(g)}$ is an $1 \times r_i$ vector, i = 1,2,..., G. Let

$$x^{(g)} = (x_1^{(g)}, x_2^{(g)},\ldots, x_G^{(g)}) \tag{4.1}$$

be the $1 \times r$ vector of test scores by a typical individual who takes test form g. As before, we assume that $x^{(g)}$ has a multivariate normal distribution with mean vector

$$\mu^{(g)} = (\mu_1^{(g)}, \mu_2^{(g)},\ldots, \mu_G^{(g)}) \tag{4.2}$$

and covariance matrix

$$
(4.3) \qquad \Sigma^{(g)} = \begin{pmatrix} \Sigma_{11}^{(g)} & \Sigma_{12}^{(g)} & \cdots & \Sigma_{1G}^{(g)} \\ \Sigma_{21}^{(g)} & \Sigma_{22}^{(g)} & \cdots & \Sigma_{2G}^{(g)} \\ \vdots & \vdots & & \vdots \\ \Sigma_{G1}^{(g)} & \Sigma_{G2}^{(g)} & \cdots & \Sigma_{GG}^{(g)} \end{pmatrix},
$$

where the blocking of $\mu^{(g)}$ and $\Sigma^{(g)}$ conforms to the blocking of $x^{(g)}$. That is, $\mu_j^{(g)}$ is $1 \times r_j$ and $\Sigma_{jk}^{(g)}$ is $r_j \times r_k$; $j,k = 1,2,\ldots, G$.

We assume that individuals have been assigned to test forms at random in such a way that N_g individuals take test form g, $g = 1,2,\ldots,G$. We also assume that the conditions under which individuals are examined are identical, and that individuals work independently of one another. Let the score vector of the ith individual taking test form g be

$$
(4.4) \qquad x_i^{(g)} = (x_{1i}^{(g)}, x_{2i}^{(g)}, \ldots, x_{Gi}^{(g)}).
$$

Under our above-stated assumptions the $x_i^{(g)}$, $i = 1,2,\ldots, N_g$; $g = 1,2,\ldots, G$, are mutually statistically independent, and a sufficient statistic for the parameters of the distributions of test scores on the G test forms is $(\underset{\sim}{\bar{x}}, \underset{\sim}{V})$, where $\underset{\sim}{\bar{x}} = (\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(G)})$, $\underset{\sim}{V} = (V^{(1)}, V^{(2)}, \ldots, V^{(G)})$,

$$
(4.5) \qquad \bar{x}^{(g)} = \frac{1}{N_g} \sum_{i=1}^{N_g} x_i^{(g)} = (\bar{x}_1^{(g)}, \bar{x}_2^{(g)}, \ldots, \bar{x}_G^{(g)}),
$$

and

$$
(4.6) \qquad V^{(g)} = \sum_{i=1}^{N_g} (x_i^{(g)} - \bar{x}^{(g)})' (x_i^{(g)} - \bar{x}^{(g)}) = \begin{pmatrix} V_{11}^{(g)} & V_{12}^{(g)} & \cdots & V_{1G}^{(g)} \\ V_{21}^{(g)} & V_{22}^{(g)} & \cdots & V_{2G}^{(g)} \\ \vdots & \vdots & & \vdots \\ V_{G1}^{(g)} & V_{G2}^{(g)} & \cdots & V_{GG}^{(g)} \end{pmatrix},
$$

$g = 1,2,\ldots, G$.

If the G test forms are not parallel, the construction of our experimental design assures us that at least the following relationships hold among the parameters:

$$(4.7) \qquad A: \begin{cases} \mu_j^{(1)} = \mu_j^{(2)} = \dots = \mu_j^{(G-j+1)}, & j = 1,2,\dots, G, \\ \Sigma_{jk}^{(1)} = \Sigma_{jk}^{(2)} = \dots = \Sigma_{jk}^{(G-j+1)}, & k \le j,\ j = 1,2,\dots, G. \end{cases}$$

On the other hand, if the G forms are parallel, then the following hypothesis about the parameters holds:

$$(4.8) \qquad H: \quad \mu^{(1)} = \mu^{(2)} = \dots = \mu^{(G)}, \ \Sigma^{(1)} = \Sigma^{(2)} = \dots = \Sigma^{(G)}.$$

The likelihood ratio test for testing the null hypothesis H against the alternative A is constructed in terms of $\bar{x}^{(1)}, \bar{x}^{(2)}, \dots, \bar{x}^{(G)}, V^{(1)}, V^{(2)}, \dots, V^{(G)}$, and the following quantities:

$$(4.9) \qquad M_j = \sum_{g=1}^{j} N_g, \qquad j = 1,2,\dots, G,$$

$$(4.10) \qquad q_j = \sum_{g=1}^{j} r_g, \qquad j = 1,2,\dots, G,$$

$$(4.11) \qquad E_j = (I_{q_j}, 0) : q_j \times r, \qquad j = 1,2,\dots, G,$$

$$(4.12) \qquad W_j = \sum_{g=1}^{j} V^{(g)}, \qquad j = 1,2,\dots, G,$$

and $\quad B_1 \equiv 0,$

$$(4.13) \quad B_j = B_{j-1} + \frac{N_j M_{j-1}}{M_j} \left(\frac{1}{M_{j-1}} \sum_{g=1}^{j-1} N_g \bar{x}^{(g)} - \bar{x}^{(j)}\right) \left(\frac{1}{M_{j-1}} \sum_{g=1}^{j-1} N_g \bar{x}^{(g)} - \bar{x}^{(j)}\right)',$$

for $j = 2,3,\dots, G$. Note that

$$(4.14) \qquad r = q_G = \sum_{g=1}^{G} r_g, \qquad N \equiv M_G = \sum_{g=1}^{G} N_g \ .$$

The likelihood ratio test statistic λ for testing H versus A can now be derived by a simple extension of the method used in Section 2 (for the case G = 3). The resulting LRTS is

$$(4.15)\qquad \lambda = \left(\prod_{j=2}^{G} \frac{\left|\frac{1}{M_G} E_{j-1} (W_G+B_j) E'_{j-1}\right|^{M_G/2} \left|\frac{1}{N_j} V^{(j)}\right|^{N_j/2}}{\left|\frac{1}{M_G} E_j (W_G+B_j) E'_j\right|^{M_G/2} \left|\frac{1}{N_j} E_{G-j+1} V^{(j)} E'_{G-j+1}\right|^{N_j/2}} \right)$$

$$\times \left(\prod_{j=2}^{G-1} \frac{\left|\frac{1}{M_j} E_j W_{G-j+1} E'_j\right|^{M_j/2}}{\left|\frac{1}{M_j} E_{j-1} W_{G-j+1} E'_{j-1}\right|^{M_j/2}} \right)$$

$$\times \left(\frac{\left|\frac{1}{N_1} V^{(1)}\right|^{N_1/2}}{\left|\frac{1}{N_1} E_{G-1} V^{(1)} E'_{G-1}\right|^{N_1/2}} \right).$$

We reject the null hypothesis H if

$$(4.16)\qquad \lambda < \lambda^*$$

where λ^* is chosen so as to give the test a desired level of significance α. For $N_1, N_2, \ldots, N_G$ all moderately large, we can use the fact that under H

$$(4.17)\qquad -2 \log \lambda \xrightarrow{\text{law}} \chi^2_f,$$

where

$$(4.18)\qquad f = \sum_{g=2}^{G} g\, \frac{r_g(r_g+r_{g-1}+3)}{2} + \frac{r_1(r_1+3)}{2} - \frac{1}{2}\left(\sum_{g=1}^{G} r_g\right)\left(\sum_{g=1}^{G} r_g+3\right),$$

to find an approximate level-α test based on λ. This test rejects H when

$$(4.19)\qquad -2 \log \lambda > \chi^2_f(\alpha),$$

where λ is given by (4.15), and f is given by (4.18).

Various other hierarchical designs for testing the parallelism psychological tests can be analyzed using the methods of likelihood ratio testing developed here and in the earlier paper (Gleser and Olkin (1972)). Discussion of these designs and their analysis is planned for a future paper.

ACKNOWLEDGEMENTS

This work was supported in part by the Educational Testing Service, The Air Force Office of Scientific Research (contract F44620-70-C-0066 at Johns Hopkins University and grant AFOSR 73-2432 at Purdue University) and the National Science Foundation (Grant GP-32326 X) at Stanford University. We would like to express our thanks to William H. Angoff and June Stern for their help in obtaining the data of Table 1, and to Britt-Marie Alsén for her help in performing the computations.

REFERENCES

[1] Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York.

[2] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society A, 160, 268-282.

[3] Gleser, L. J. and Olkin, I. (1972) Multivariate statistical inference under marginal structure. Brit. J. Math. Statist. Psychol. (in press).