THE LAGRANGIAN MULTIPLIER TEST

By S. D. Silvey

University of Glasgow

1. Introduction. One of the problems which occurs most frequently in practical statistics is that of deciding, on the basis of a number of independent observations on a random variable, whether a finite dimensional parameter involved in the distribution function of the random variable belongs to a proper subset ω of the set Ω of possible parameters. Naturally this problem has received considerable attention and the main method which is currently applied in dealing with it is the well-known Neyman-Pearson likelihood ratio test. Direct application of this test involves finding the supremum of the likelihood function in the set ω and this in turn often involves the solution of restricted likelihood equations containing a Lagrangian multiplier. And the same set of of equations has to be solved if, irrespective of the likelihood ratio test, it is desired to obtain a maximum likelihood estimate in the set ω of the unknown parameter. Rather surprisingly, since the problem is of such frequent occurrence, little seems to have appeared in statistical literature on such restricted maximum likelihood estimates, the main results in this field being contained in a recent paper by Aitchison and Silvey [1].

In this paper the authors introduced, on an intuitive basis, a method of testing whether the true parameter does belong to ω, this method being based on the distribution of a random Lagrangian multiplier appearing in the restricted likelihood equations. It is the object of this present paper to discuss this Lagrangian multiplier test. In order to do so, it is necessary to consider how the results of the previous paper must be modified when the true parameter does not belong to the set ω, because only in this way can we obtain any notion of the power of the test. Discussion of this point forms the initial part of the present paper. We will then show the connection between the Lagrangian multiplier test and the likelihood ratio test. Finally, since often in practice situations arise where the information matrix is singular, we will consider how the Lagrangian multiplier test must be adapted to meet this contingency.

The approach adopted by Aitchison and Silvey [1] in the discussion of restricted estimates is essentially Cramér's approach [4] to maximum likelihood estimates, i.e., attention is concentrated on solutions of the likelihood equations rather than on genuine maximum likelihood estimates. Such an approach is really unsuitable in the present instance where we do not necessarily assume that the true parameter does belong to the subset ω. And we will use instead the method used by Wald [7] in his discussion of the consistency of maximum likelihood estimators. As has been pointed out by Kraft and Le Cam [5], Wald's approach to unrestricted maximum likelihood estimation is much more illumi-

Received March 14, 1958.

nating than that of Cramér and, not surprisingly, this is still true of restricted estimation. Unfortunately the change in viewpoint necessitates certain changes in the notation used by Aitchison and Silvey, and these we will now introduce in describing mathematically the situation to be discussed.

2. Notation. The basic situation in which we shall be interested is described mathematically as follows.

Corresponding to each point $\theta = (\theta_1, \theta_2, \cdots, \theta_s)$ in some subset Ω of s-dimensional Euclidean space, denoted by R^s, is a distribution function $F(\cdot, \theta)^*$ defined on R^α; where α is some given integer. A random variable X, taking values in R^α has distribution function $F(\cdot, \theta_0)$ where θ_0 is known to belong to Ω but is otherwise unknown; though it is suspected that θ_0 belongs to a subset $\omega = \Omega \cap \{\theta : h(\theta) = 0\}$ of Ω, where $h = (h_1, h_2, \cdots, h_r)$ is a well-behaved function from R^s into R^r, $r < s$.

We will assume, as is usual, that for all $\theta \varepsilon \Omega$, $F(\cdot, \theta)$ is either discrete or absolutely continuous, and admits an elementary probability law $f(\cdot, \theta)$. Then for a given sequence $x = (x_1, x_2, \cdots, x_n, \cdots)$ of independent observations on X, the log-likelihood function $\log L_n(x, \cdot)$ is defined on Ω by $\log L_n x, \theta) = \sum_{i=1}^n \log f(x_i, \theta)$. By a maximum likelihood estimate of θ_0 in any subset ω^* of Ω, we mean an element $\hat{\theta}_n(x, \omega^*)$ of ω^* which is such that

$$\log L_n(x, \hat{\theta}_n(x, \omega^*)) = \sup_{\theta \varepsilon \omega^*} \log L_n(x, \theta).$$

If a single-valued function $\hat{\theta}_n(\cdot, \omega^*)$ is thus defined for almost all x, then $\hat{\theta}_n(\cdot, \omega^*)$ is a random variable called a maximum likelihood estimator of θ_0 in ω^*. When we refer to "almost all x" we mean almost all with respect to the probability measure defined on the sequence space of points x by the consideration that the components of a sequence x are regarded as independent observations on a random variable X with distribution function $F(\cdot, \theta_0)$. Similarly "almost all $t \varepsilon R^\alpha$" means almost all with respect to the probability measure defined on R^α by $F(\cdot, \theta_0)$.

The matrix whose (i, j)th element is $\int_{R^\alpha} \partial \log f(t, \theta)/\partial\theta_i \cdot \partial \log f(t, \theta)/\partial\theta_j \, dF(t, \theta)$, we will denote by $\mathbf{B}_\theta$. Further, $\mathbf{H}_\theta$ will denote the $s \times r$ matrix $(\partial h_j(\theta)/\partial\theta_i)$. For any real function ζ defined on R^s, $\mathbf{D}\zeta(\theta)$ will denote the column vector whose ith component is $\partial\zeta(\theta)/\partial\theta_i$, while $\mathbf{D}^2\zeta(\theta)$ will denote the $s \times s$ matrix whose (i, j)th component is $\partial^2\zeta(\theta)/\partial\theta_i\partial\theta_j$. Generally column vectors corresponding to points in Euclidean space will be printed in the corresponding boldface type so that, for example, the column vector $\boldsymbol{\theta}$ corresponds to the point θ.

We will be interested initially in the emergence of $\hat{\theta}_n(x, \omega)$ as a solution of the equations

$$n^{-1}\mathbf{D} \log L_n(x, \theta) + \mathbf{H}_\theta \boldsymbol{\lambda} = \mathbf{0}$$

$$\mathbf{h}(\theta) = \mathbf{0},$$

where λ is a Lagrangian multiplier in R^r, and generally in the restricted maximum likelihood estimator $\hat{\theta}_n(\cdot, \omega)$.

3. $\hat{\theta}_n(x, \omega)$ and the likelihood equations. Naturally the discussion on which we have embarked will involve the introduction of various assumptions concerning F and h. The assumptions that we will introduce are not designed to achieve complete mathematical generality but are, we hope, of such a nature that they will not obscure the over-all mathematical picture and will be satisfied in many practical problems. The first of these assumptions is as follows.

Assumption 1. For every $\theta \varepsilon \Omega$, $z(\theta) = \int_{R^\alpha} \log f(t, \theta)\, dF(t, \theta_0)$ exists.

The whole problem of maximum likelihood estimation, restricted and unrestricted, is closely bound up with the behaviour of the function z, because the Law of Large Numbers ensures that, for each θ, the sequence $(n^{-1} \log L_n(x, \theta))$ converges, for almost all x, to $z(\theta)$. If, further, this convergence is uniform with respect to θ, then for large n and most x, $n^{-1} \log L_n(x, \cdot)$ will be uniformly near z and under suitable conditions will attain its supremum in ω near the point (if such exists) where z attains its supremum in ω. The assumptions which we will now introduce are designed to achieve this desirable situation.

Assumption 2. Ω is a convex compact subset of R^s.

Assumption 3. For almost all $t \varepsilon R^\alpha$, $\log f(t, \cdot)$ is continuous on Ω.

Assumption 4. For almost all $t \varepsilon R^\alpha$, and for every $\theta \varepsilon \Omega$, $\partial \log f(t, \theta)/\partial\theta_i$ $(i = 1, 2, \cdots, s)$ exists and $|\partial \log f(t, \theta)/\partial\theta_i| < g(t)(i = 1, 2, \cdots, s)$ where $\int_{R^\alpha} g(t)\, dF(t, \theta_0)$ is finite.

Assumption 5. The function h is continuous on Ω.

Assumption 6. There exists a point $\theta^* \varepsilon \omega$ such that $z(\theta^*) > z(\theta)$ when $\theta \varepsilon \omega$ and $\theta \neq \theta^*$.

Assumptions 2–4 ensure that for almost all x the sequence $(n^{-1} \log L_n(x, \theta))$ converges to $z(\theta)$ uniformly with respect to θ in the set Ω. Assumptions 2 and 5 ensure that ω is a compact subset of R^s and consequently that any continuous function on ω attains its supremum at some point of ω. In particular the function $\log L_n(x, \cdot)$, for almost all x, attains its supremum in ω at some point $\hat{\theta}_n(x, \omega)$ of ω. Assumption 6 then ensures that for almost all x the sequence $(\hat{\theta}_n(x, \omega))$ converges to θ^*. The proofs of these results are fairly straightforward and we omit them.

It is of some interest to note that if $\theta_0 \varepsilon \omega$ then usually θ_0 will satisfy the condition demanded of θ^*. This has been proved by Wald [7]. In fact, when interest is concentrated on the case where $\theta_0 \varepsilon \omega$, Assumption 6 may be replaced by the following

Assumption 6A. $\theta_0 \varepsilon \omega$ and if $\theta \neq \theta_0$ then for at least one $t \varepsilon R^\alpha$, $F(t, \theta) \neq F(t, \theta_0)$. This is sufficient to ensure that $z(\theta_0) > z(\theta)$ if $\theta \neq \theta_0$.

As stated above, Assumptions 1–6 ensure the existence of a maximum likelihood estimator in ω of θ_0 which converges with probability one to θ^*. If in addition we make the following Assumption 7 then for large n and most x, $\hat{\theta}_n(x, \omega)$ will be an interior point of ω and consequently will emerge as a solution of the restricted likelihood equations, when the function h is differentiable.

Assumption 7. θ^* is an interior point of ω. Now making assumptions 1–7, we will use these likelihood equations in discussing the asymptotic distribution of $\hat{\theta}_n(\cdot, \omega)$.

4. The asymptotic distribution of $\hat{\theta}_n(\cdot, \omega)$. The method by which the asymptotic distribution of maximum likelihood estimators is usually derived, for example by Cramér [4], involves expanding the likelihood function by Taylor's Theorem. In order that we may adopt this method in the present instance we now introduce the following assumptions, similar to those of Cramér.

Assumption 8. The functions h_i possess first and second order partial derivatives which are continuous (and so bounded) on Ω.

Assumption 9. For almost all $t \varepsilon R^\alpha$ the function $\log f(t, \cdot)$ possesses continuous second order partial derivatives in a neighborhood of θ^*. Also, if θ belongs to this neighborhood, then $|\partial^2 \log f(t, \theta)/\partial\theta_i\partial\theta_j| < G_1(t)$ $(i, j = 1, 2, \cdots, s)$ where $\int_{R^\alpha} G_1(t)\, dF(t, \theta_0)$ is finite.

Assumption 10. For almost all $t \varepsilon R^\alpha$ the function $\log f(t, \cdot)$ possesses third order partial derivatives in a neighborhood of θ^* and, if θ is in this neighborhood, then

$$|\partial^3 \log f(t, \theta)/\partial\theta_i\partial\theta_j\partial\theta_k| < G_2(t) \qquad (i, j, k = 1, 2, \cdots, s),$$

where $\int_{R^\alpha} G_2(t)\, dF(t, \theta_0)$ is finite.

(4.1) Important implications for our purposes of Assumptions 4, 9 and 10 are as follows.

(4.1.1) The vector $\mathbf{D}z(\theta)$ exists for every $\theta \varepsilon \Omega$ and the sequence $(\mathbf{D}n^{-1} \log L_n(x, \theta))$ of vectors converges for almost all x to $\mathbf{D}z(\theta)$ (Assumption 4).

(4.1.2) The matrix $\mathbf{D}^2z(\theta^*)$ exists and the sequence $(\mathbf{D}^2n^{-1} \log L_n(x, \theta^*))$ of matrices converges for almost all x to $\mathbf{D}^2z(\theta^*)$ (Assumption 9).

(4.1.3) For almost all x and $i, j, k = 1, 2, \cdots, s$ the sequence $(n^{-1}\partial^3 \log L_n(x, \theta)/\partial\theta_i\partial\theta_j\partial\theta_k)$ is bounded uniformly with respect to θ in a neighborhood of θ^* (Assumption 10).

Each of these three statements is almost a direct consequence of the Strong Law of Large Numbers.

We are now in a position to obtain the asymptotic distribution of $\hat{\theta}_n(\cdot, \omega)$. For brevity we will now write $\hat{\theta}$ instead of $\hat{\theta}_n(x, \omega)$. Since $\hat{\theta} \rightarrow \theta^*$ for almost all x, we find by applying Taylor's Theorem and using (4.1.2) and (4.1.3) that

(4.1.4) $\mathbf{D}n^{-1} \log L_n(x, \hat{\theta}) = \mathbf{D}n^{-1} \log L_n(x, \theta^*) + [\mathbf{D}^2z(\theta^*) + o(1)]\,[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*]$

for almost all x.

Also because of the continuity of the first partial derivatives of the functions h_i, for almost all x,

(4.1.5) $$\mathbf{H}_{\hat{\theta}} = \mathbf{H}_{\theta^*} + o(1)$$

and

$$\mathbf{h}(\hat{\theta}) = [\mathbf{H}'_{\theta*} + o(1)][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*]. \tag{4.1.6}$$

For almost any x, if n is sufficiently large, $\hat{\theta}$ will, with a certain Lagrangian multiplier $\hat{\lambda}_n(x)$, satisfy the restricted likelihood equations. So we have, writing $\hat{\lambda}$ in place of $\hat{\lambda}_n(x)$ for brevity,

$$\mathbf{D}n^{-1} \log L_n(x, \theta^*) + [\mathbf{D}^2 z(\theta^*) + o(1)][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*] + \mathbf{H}_{\hat{\theta}}\hat{\boldsymbol{\lambda}} = \mathbf{0}, \tag{4.1.7}$$

$$[\mathbf{H}'_{\theta*} + o(1)][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*] = \mathbf{0}. \tag{4.1.8}$$

Since $z(\theta^*)$ is a maximum in the set ω of the function z, there exists a Lagrangian multiplier $\lambda^* = (\lambda_1^*, \lambda_2^*, \cdots, \lambda_r^*)$ such that

$$\mathbf{D}z(\theta^*) + \mathbf{H}_{\theta*}\boldsymbol{\lambda}^* = \mathbf{0}, \tag{4.1.9}$$

and on subtracting (4.1.9) from (4.1.7), and using (4.1.5) we obtain

$$\begin{aligned} &[\mathbf{D}n^{-1} \log L_n(x, \theta^*) - \mathbf{D}z(\theta^*)] + [\mathbf{D}^2 z(\theta^*) + o(1)][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*] \\ &\qquad + [\mathbf{H}_{\theta*} + o(1)][\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*] + [\mathbf{H}_{\hat{\theta}} - \mathbf{H}_{\theta*}]\boldsymbol{\lambda}^* = \mathbf{0}. \end{aligned} \tag{4.1.10}$$

Now on expanding the elements of the matrix $\mathbf{H}_{\hat{\theta}}$ by Taylor's Theorem, we find that, because of the continuity of the second order partial derivatives of the functions h_i, for almost all x,

$$[\mathbf{H}_{\hat{\theta}} - \mathbf{H}_{\theta*}]\boldsymbol{\lambda}^* = \left[\sum_{i=1}^{r} \lambda_i^* \, \mathbf{D}^2 h_i(\theta^*) + o(1)\right][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*]. \tag{4.1.11}$$

We will denote by $-\mathbf{B}_{\theta*}^{\dagger}$ the matrix $\mathbf{D}^2 z(\theta^*) + \sum_{i=1}^{r} \lambda_i^* \mathbf{D}^2 h_i(\theta^*)$. Then on substituting in (4.1.10) the expression for $[\mathbf{H}_{\hat{\theta}} - \mathbf{H}_{\theta*}]\boldsymbol{\lambda}^*$ contained in (4.1.11) we have

$$\begin{aligned} &[\mathbf{B}_{\theta*}^{\dagger} + o(1)][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*] - [\mathbf{H}_{\theta*} + o(1)][\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*] \\ &\qquad = \mathbf{D}n^{-1} \log L_n(x, \theta^*) - \mathbf{D}z(\theta^*), \end{aligned} \tag{4.1.12}$$

and combining (4.1.12) and (4.1.8) we may write

$$\begin{aligned} &\begin{bmatrix} \mathbf{B}_{\theta*}^{\dagger} + o(1) & -\mathbf{H}_{\theta*} + o(1) \\ -\mathbf{H}'_{\theta*} + o(1) & \mathbf{O} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^* \\ \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^* \end{bmatrix} \\ &\qquad = \begin{bmatrix} \mathbf{D}n^{-1} \log L_n(x, \theta^*) - \mathbf{D}z(\theta^*) \\ \mathbf{O} \end{bmatrix}. \end{aligned} \tag{4.1.13}$$

We will now make the final assumptions which enable us to derive the asymptotic distribution of $\hat{\theta}_n(\cdot, \omega)$ and $\hat{\lambda}_n(\cdot)$.

Assumption 11. The matrix

$$\begin{bmatrix} \mathbf{B}_{\theta*}^{\dagger} & -\mathbf{H}_{\theta*} \\ -\mathbf{H}'_{\theta*} & \mathbf{O} \end{bmatrix}$$

is non-singular.

Assumption 12. For $i, j = 1, 2, \cdots, s$, $\beta_{ij}(\theta^*) = \int_{R^\alpha} \partial \log f(t, \theta^*)/\partial\theta_i \cdot \partial \log f(t, \theta^*)/\partial\theta_j \, dF(t, \theta_0)$ exists.

We now define

$$\begin{bmatrix} \mathbf{P}^\dagger_{\theta^*} & \mathbf{Q}^\dagger_{\theta^*} \\ \mathbf{Q}^{\dagger\prime}_{\theta^*} & \mathbf{R}^\dagger_{\theta^*} \end{bmatrix} = \begin{bmatrix} \mathbf{B}^\dagger_{\theta^*} & -\mathbf{H}_{\theta^*} \\ -\mathbf{H}'_{\theta^*} & \mathbf{O} \end{bmatrix}^{-1},$$

and $\mathbf{V}_{\theta^*} = (\beta_{ij}(\theta^*)) - [\mathbf{D}z(\theta^*)][\mathbf{D}z(\theta^*)]'$. By the multivariate form of a Central Limit Theorem (Cramér [3]) it follows from the existence of the matrix $\mathbf{V}_{\theta^*}$ that the distribution of $\sqrt{n}[\mathbf{D}n^{-1} \log L_n(\cdot, \theta^*) - \mathbf{D}z(\theta^*)]$ is asymptotically normal with mean $\mathbf{0}$ and variance matrix $\mathbf{V}_{\theta^*}$. Then from (4.1.13), by the multivariate extension of a theorem of Cramér [4] we have the results stated in the following lemma.

LEMMA 1. *Under Assumptions* 1–12 *the random vector*

$$\sqrt{n}\begin{bmatrix} \hat{\boldsymbol{\theta}}_n(\cdot, \omega) - \boldsymbol{\theta}^* \\ \hat{\boldsymbol{\lambda}}_n(\cdot) - \boldsymbol{\lambda}^* \end{bmatrix}$$

is asymptotically normally distributed with mean $\mathbf{0}$ *and variance matrix*

$$\begin{bmatrix} \mathbf{P}^\dagger_{\theta^*} \mathbf{V}_{\theta^*} \mathbf{P}^\dagger_{\theta^*} & \mathbf{P}^\dagger_{\theta^*} \mathbf{V}_{\theta^*} \mathbf{Q}^\dagger_{\theta^*} \\ \mathbf{Q}^{\dagger\prime}_{\theta^*} \mathbf{V}_{\theta^*} \mathbf{P}^\dagger_{\theta^*} & \mathbf{Q}^{\dagger\prime}_{\theta^*} \mathbf{V}_{\theta^*} \mathbf{Q}^\dagger_{\theta^*} \end{bmatrix}.$$

We have now obtained a formal result regarding the behavior for large n of the restricted maximum likelihood estimator, a result which might be used in most practical situations to determine the large sample power function of the test of the hypothesis that $\theta_0 \varepsilon \omega$, proposed by Aitchison and Silvey. (This might involve a considerable amount of computation). The extent to which the method of solving the likelihood equations which is proposed in the same paper can be used when $\theta_0 \not\varepsilon \omega$ remains obscure, as does any general picture of the power of the test. However some light is shed on these questions by considering how the results here obtained particularize in the case when $\theta_0 \varepsilon \omega$.

(4.2) Accordingly we consider what happens when we replace Assumption 6 by Assumption 6A. Then θ_0 replaces θ^* and $z(\theta_0)$, the maximum of z in the set ω, is also the maximum of z in the set Ω. Hence $\mathbf{D}z(\theta^*) = \mathbf{0}$ and $\boldsymbol{\lambda}^* = \mathbf{0}$. The matrix $\mathbf{V}_{\theta^*}$ becomes the matrix $\mathbf{B}_{\theta_0}$ and, with the mild additional assumption

Assumption 13. $\int_{R^\alpha} \partial^2 f(t, \theta_0)/\partial\theta_i\partial\theta_j \, dt = 0$ $(i, j = 1, 2, \cdots, s)$, the matrix $\mathbf{B}^\dagger_{\theta^*}$ also becomes $\mathbf{B}_{\theta_0}$. Consequently we have exactly the result of the previous paper [1] concerning the asymptotic distribution of the restricted estimator and the corresponding Lagrangian multiplier. The assumptions made here in deriving this distribution are, so far as comparison is possible, stronger than the assumptions of the previous paper, but we have now obtained a result concerning the genuine maximum likelihood estimator rather than merely a solution of the likelihood equations. (A greater degree of similarity between the two sets of assumptions is apparent if we note that in the case where $\theta_0 \varepsilon \omega$ we might replace Assumption 11 by the following

Assumption 11A. The matrix $\mathbf{B}_{\theta_0}$ is positive definite and the matrix $\mathbf{H}_{\theta_0}$ is of rank r).

(4.3) It is now possible to obtain a picture of the typical practical situation when n is large and θ_0, while not belonging to the set ω, is very near this set. Usually then $z(\theta_0)$ will be $\sup_{\theta \varepsilon \Omega} z(\theta)$ and θ^* will be near θ_0 so that $\mathbf{D}z(\theta^*)$ will be near $\mathbf{D}z(\theta_0) = \mathbf{0}$. Then λ^* also will be near 0, though, since n is large, $\sqrt{n}\lambda^*$ may be appreciably different from 0. Also the elements of $\mathbf{D}^2 z(\theta^*)$ will be near those of $\mathbf{D}^2 z(\theta_0) = -\mathbf{B}_{\theta_0}$. If in addition $\beta_{ij}(\theta^*)$ is near the corresponding element of $\mathbf{B}_{\theta_0}$, as will usually be the case, then we can say that approximately

$$\sqrt{n}\begin{bmatrix} \hat{\boldsymbol{\theta}}_n(\cdot, \omega) - \boldsymbol{\theta}^* \\ \hat{\boldsymbol{\lambda}}_n(\cdot) - \boldsymbol{\lambda}^* \end{bmatrix}$$

will have a multivariate normal distribution with mean $\mathbf{0}$ and variance matrix

$$\begin{bmatrix} \mathbf{P}_{\theta_0} & \mathbf{O} \\ \mathbf{O} & -\mathbf{R}_{\theta_0} \end{bmatrix},$$

this matrix being as defined in [1]. (It would be possible to give a rigorous mathematical derivation of this result by imagining the true parameter θ_0 to vary with n in such a way that the distance of θ_0 from the set ω tended to 0 as $n \to \infty$, and by imposing suitable restrictions on the functions f and h to ensure that what is here said to happen usually would in fact happen. But this does not seem particularly profitable).

(4.4) Finally in this connection, because of the remarks made in the previous paragraph and of the flexibility of Newton's method of solving equations, we might expect that, in the case where θ_0 is near the set ω and n is large, the iterative method of solving the restricted likelihood equations suggested in [1] will still apply.

5. Three tests of the hypothesis that $\theta_0 \varepsilon \omega$**.** We will now compare three intuitively reasonable tests of the hypothesis that $\theta_0 \varepsilon \omega$. These are as follows.

(i) *The likelihood ratio test.* We accept the hypothesis if $\mu(x) = \sup_{\theta \varepsilon \omega} L_n(x, \theta)/\sup_{\theta \varepsilon \Omega} L_n(x, \theta)$ is "sufficiently near" 1.

(ii) *The Wald test.* Assuming the existence of $\hat{\theta}_n(x, \Omega)$, we accept the hypothesis if $h(\hat{\theta}_n(x, \Omega))$ is "sufficiently near" 0. (Wald [8]).

(iii) *The Lagrangian multiplier test.* Assuming the existence of $\hat{\theta}_n(x, w)$ and $\hat{\lambda}_n(x)$ we accept the hypothesis if $\hat{\lambda}_n(x)$ is "sufficiently near" 0. (Aitchison and Silvey [1]).

For typographical brevity we will now write $\hat{\theta}$ for the unrestricted maximum likelihood estimator $\hat{\theta}_n(\cdot, \Omega)$, $\hat{\theta}$ for the restricted maximum likelihood estimator $\hat{\theta}_n(\cdot, \omega)$ and $\hat{\lambda}$ for the random variable $\hat{\lambda}_n(\cdot)$.

The measure of the distance from 0 of $h(\hat{\theta})$ used by Wald is, in our notation,

$-n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[\mathbf{h}(\dot{\theta})]$; his test is based on this random variable and he has shown that under general conditions the asymptotic distributions of $-2 \log \mu$ and $-n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[\mathbf{h}(\dot{\theta})]$ are the same. The measure of the distance from 0 of $\hat{\lambda}$ in the test proposed by Aitchison and Silvey is $-n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{-1}\hat{\boldsymbol{\lambda}}$. We will now show that subject to the following assumptions A we have

$$\text{plim } 2 \log \mu = \text{plim } n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[\mathbf{h}(\dot{\theta})] = \text{plim } n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{-1}\hat{\boldsymbol{\lambda}}.$$

Assumptions A. By assumptions A we mean the following set of assumptions:—1–5, 6A, 7–10, 11A, 13 and

Assumption 12A. The matrix $\mathbf{B}_\theta$ exists in a neighborhood of θ_0, and its elements are continuous functions of θ there. Of course when assumption 6A is made, θ^* is replaced by θ_0 in subsequent assumptions.

We have already seen that these assumptions imply that $\hat{\theta}$ exists and almost certainly converges to θ_0, and that for large n and most x, $\hat{\lambda}_n(x)$ exists. It is not difficult to use the particular form to which (4.1:13) reduces when assumption 6A replaces assumption 6 to obtain the results

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{-\frac{1}{2}}\mathbf{P}_{\theta_0}\mathbf{D} \log L_n(\cdot, \theta_0) + o_p(1), \tag{5.1}$$

$$\sqrt{n}\hat{\boldsymbol{\lambda}} = n^{-\frac{1}{2}}\mathbf{Q}_{\theta_0}'\mathbf{D} \log L_n(\cdot, \theta_0) + o_p(1). \tag{5.2}$$

Here o_p is used in the sense of Mann and Wald [6] and $\mathbf{P}_{\theta_0}$, $\mathbf{Q}_{\theta_0}$ are defined by

$$\begin{bmatrix} \mathbf{B}_{\theta_0} & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}_{\theta_0}' & \mathbf{O} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P}_{\theta_0} & \mathbf{Q}_{\theta_0} \\ \mathbf{Q}_{\theta_0}' & \mathbf{R}_{\theta_0} \end{bmatrix}. \tag{5.3}$$

Also it is easy to show by the same kind of argument as has been applied above that the assumptions A imply that $\dot{\theta}$ exists and almost certainly converges to θ_0 and that

(5.4) for almost any x and sufficiently large n,

$$\mathbf{D} \log L_n[x, \dot{\theta}(x)] = \mathbf{0},$$

$$\sqrt{n}(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{-\frac{1}{2}}\mathbf{B}_{\theta_0}^{-1}\mathbf{D} \log L_n(\cdot, \theta_0) + o_p(1). \tag{5.5}$$

We will now use these results to prove the following lemmas.

Lemma 2. *Subject to assumptions* A,

$$-2 \log \mu = n(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})'\mathbf{B}_{\theta_0}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + o_p(1).$$

Proof. Clearly from (5.1) and (5.5), $\|\hat{\theta} - \dot{\theta}\| = O_p(n^{-\frac{1}{2}})$. Hence on expanding $\log L_n(\cdot, \hat{\theta})$ by Taylor's Theorem, we have, in virtue of (4.1.3) and (5.4)

$$\log L_n(\cdot, \hat{\theta}) = \log L_n(\cdot, \dot{\theta}) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})'[\mathbf{D}^2 \log L_n(\cdot, \dot{\theta})][\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}] + o_p(1).$$

Again from Taylor's Theorem we have $n^{-1}\mathbf{D}^2 \log L_n(\cdot, \dot{\theta}) = n^{-1}\mathbf{D}^2 \log L_n(\cdot, \theta_0) + o_p(1)$, and from (4.1.2) and assumptions 9 and 13 (which imply $\mathbf{D}^2 z(\theta_0) = -\mathbf{B}_{\theta_0}$)

$$n^{-1}\mathbf{D}^2 \log L_n(\cdot, \theta_0) = -\mathbf{B}_{\theta_0} + o_p(1).$$

Hence

$$\log \mu = \log L_n(\cdot, \hat{\theta}) - \log L_n(\cdot, \dot{\theta})$$
$$= -\tfrac{1}{2}n(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})'[\mathbf{B}_{\theta_0} + o_p(1)](\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + o_p(1),$$

and the result follows because $\|\hat{\theta} - \dot{\theta}\| = O_p(n^{-\frac{1}{2}})$.

LEMMA 3. *Subject to assumptions A*, $2 \log \mu = n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{-1}\hat{\boldsymbol{\lambda}} + o_p(1)$.

PROOF. We have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) = n^{-\frac{1}{2}}(\mathbf{P}_{\theta_0} - \mathbf{B}_{\theta_0}^{-1})\mathbf{D} \log L_n(\cdot, \theta_0) + o_p(1).$$

Now

$$[\mathbf{P}_{\theta_0} - \mathbf{B}_{\theta_0}^{-1}]\mathbf{B}_{\theta_0}[\mathbf{P}_{\theta_0} - \mathbf{B}_{\theta_0}^{-1}] = \mathbf{B}_{\theta_0}^{-1} - \mathbf{P}_{\theta_0} = -\mathbf{Q}_{\theta_0}\mathbf{R}_{\theta_0}^{-1}\mathbf{Q}'_{\theta_0},$$

these matrix relationships following easily from the definition of $\mathbf{P}_{\theta_0}$, $\mathbf{Q}_{\theta_0}$ and $\mathbf{R}_{\theta_0}$ in (5.3). Hence

$$n(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})'\mathbf{B}_{\theta_0}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) = -n^{-1}[\mathbf{D} \log L_n(\cdot, \theta_0)]'\mathbf{Q}_{\theta_0}\mathbf{R}_{\theta_0}^{-1}\mathbf{Q}'_{\theta_0}[\mathbf{D} \log L_n(\cdot, \theta_0)] + o_p(1)$$
$$= -n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\theta_0}^{-1}\hat{\boldsymbol{\lambda}} + o_p(1), \qquad \text{by (5.2).}$$

Since, according to assumption 12A, the elements of the matrix $\mathbf{B}_\theta$ are continuous functions in a neighborhood of θ_0, and by 11A $\mathbf{B}_{\theta_0}$ is positive definite, $\mathbf{B}_\theta$ will also be positive definite in a neighborhood of θ_0. Similarly $\mathbf{H}_\theta$ is of rank r in a neighborhood of θ_0 and so the matrix $\mathbf{R}_\theta$ exists and its elements are continuous functions of θ in a neighborhood of θ_0. It follows from the strong convergence of $\hat{\theta}$ to θ_0 that $\mathbf{R}_{\hat{\theta}}^{-1} = \mathbf{R}_{\theta_0}^{-1} + o_p(1)$, and this completes the proof.

LEMMA 4. *Subject to assumptions* A, $2 \log \mu = n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[\mathbf{h}(\dot{\theta})] + o_p(1)$.

PROOF. Since the second derivatives of the functions h_i are bounded on Ω (assumption 9) and since $\|\dot{\theta} - \theta_0\|$ is $O_p(n^{-\frac{1}{2}})$, we have

$$\mathbf{h}(\dot{\theta}) = \mathbf{h}(\theta_0) + \mathbf{H}'_{\theta_0}(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O_p(n^{-1})$$
$$= \mathbf{H}'_{\theta_0}(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O_p(n^{-1}),$$

since by 6A, $\theta_0 \varepsilon \omega$. Hence $\sqrt{n}\mathbf{h}(\dot{\theta}) = n^{-\frac{1}{2}}\mathbf{H}'_{\theta_0}\mathbf{B}_{\theta_0}^{-1}\mathbf{D} \log L_n(\cdot, \theta_0) + o_p(1)$ and $n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\theta_0}[\mathbf{h}(\dot{\theta})] = n^{-1}[\mathbf{D} \log L_n(\cdot, \theta_0)]'\mathbf{B}_{\theta_0}^{-1}\mathbf{H}_{\theta_0}\mathbf{R}_{\theta_0}\mathbf{H}'_{\theta_0}\mathbf{B}_{\theta_0}^{-1}[\mathbf{D} \log L_n(\cdot, \theta_0)] + o_p(1)$. It is easy to show that $\mathbf{B}_{\theta_0}^{-1}\mathbf{H}_{\theta_0}\mathbf{R}_{\theta_0}\mathbf{H}'_{\theta_0}\mathbf{B}_{\theta_0}^{-1} = \mathbf{Q}_{\theta_0}\mathbf{R}_{\theta_0}^{-1}\mathbf{Q}'_{\theta_0}$, and it follows that $n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\theta_0}[\mathbf{h}(\dot{\theta})] = n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\theta_0}^{-1}\hat{\boldsymbol{\lambda}} + o_p(1)$. The proof is then completed by the remark that, as in Lemma 3, $\mathbf{R}_{\theta_0} = \mathbf{R}_{\dot{\theta}'} + o_p(1)$.

LEMMA 5. *Subject to assumptions* A, *each of the random variables* $-2 \log \mu$, $-n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[h(\dot{\theta})]$ *and* $-n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{-1}\hat{\boldsymbol{\lambda}}$ *is asymptotically distributed as* χ^2 *with r degrees of freedom.*

This follows from lemmas 3 and 4 and from the fact that $\sqrt{n}\hat{\boldsymbol{\lambda}}$ is asymptotically normally distributed with mean 0 and variance matrix $-\mathbf{R}_{\hat{\theta}}$.

In consequence of lemma 5, when n is large the natural choices of critical regions of size α for testing the hypothesis that $\theta_0 \varepsilon \omega$ on the bases (i), (ii) and (iii) are C_1, C_2 and C_3 respectively where

C_1 is the set of x on which $-2 \log \mu > k_\alpha$, .

C_2 is the set of x on which $-n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[\mathbf{h}(\dot{\theta})] > k_\alpha$, and

C_3 is the set of x on which $-n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\tilde{\theta}}^{-1}\hat{\boldsymbol{\lambda}} > k_\alpha$.

Here k_α is determined by $\Pr\{\chi^2_{[r]} > k_\alpha\} = \alpha$.

Wald [8] has shown that usually the tests based on the critical regions C_1 and C_2 have asymptotically the same power. His argument shows essentially that if n is large and θ_0 is not near $\dot{\omega}$, the power of each test is near 1, while if θ_0 is near ω each of the random variables $-2 \log \mu$ and $-n[\mathbf{h}(\dot{\theta})]'\mathbf{R}_{\dot{\theta}}[\mathbf{h}(\dot{\theta})]$ has approximately a non-central χ^2-distribution with the same parameters. We now inquire, without going into rigorous mathematical detail, whether this type of argument will usually hold when we compare the tests based on the critical regions C_1 and C_3.

We consider first what happens when n is large and θ_0 is near ω. Then as we have seen, θ^* will usually be near θ_0 and we suppose that θ_0 is near enough ω to ensure that $\theta^* - \theta_0$ is near 0, though $\sqrt{n}(\theta^* - \theta_0)$ may be appreciably different from 0. In virtue of the remarks made in (4.3) we will then have, in most practical situations,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim n^{-\frac{1}{2}}\mathbf{P}_{\theta_0}\mathbf{D} \log L_n(\cdot, \theta_0), \tag{5.6}$$

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*) \sim n^{-\frac{1}{2}}\mathbf{Q}'_{\theta_0}\mathbf{D} \log L_n(\cdot, \theta_0) \tag{5.7}$$

where $\sim$ denotes approximate equality with probability near 1, for large n. Also since $\mathbf{D}z(\theta^*) + \mathbf{H}_{\theta^*}\boldsymbol{\lambda}^* = \mathbf{0}$ and since usually $\mathbf{D}z(\theta_0) = \mathbf{0}$ and $\mathbf{D}^2z(\theta_0) = -\mathbf{B}_{\theta_0}$, we will have

$$-\mathbf{B}_{\theta_0}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) + \mathbf{H}_{\theta_0}\boldsymbol{\lambda}^* = \mathbf{0}, \tag{5.8}$$

approximately. Since the distribution of $\dot{\theta}$ does not depend on whether θ_0 is in ω or not, it will remain true (see (5.5)) that

$$\sqrt{n}(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim n^{-\frac{1}{2}}\mathbf{B}_{\theta_0}^{-1}\mathbf{D} \log L_n(\cdot, \theta_0). \tag{5.9}$$

Also examination of the details of the proof of lemma 2 shows that the result there obtained, namely

$$-2 \log \mu \sim n(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})\mathbf{B}_{\theta_0}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \tag{5.10}$$

still holds.

Now from (5.6) and (5.9) we have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) &\sim \sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0) + n^{-\frac{1}{2}}(\mathbf{P}_{\theta_0} - B_{\theta_0}^{-1})\mathbf{D} \log L_n(\cdot, \theta_0)\\ &= \sqrt{n}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) + n^{-\frac{1}{2}}\mathbf{Q}_{\theta_0}\mathbf{R}_{\theta_0}^{-1}\mathbf{Q}'_{\theta_0}\mathbf{D} \log L_n(\cdot, \theta_0)\\ &\sim \sqrt{n}\mathbf{B}_{\theta_0}^{-1}\mathbf{H}_{\theta_0}\boldsymbol{\lambda}^* + \sqrt{n}\mathbf{Q}_{\theta_0}\mathbf{R}_{\theta_0}^{-1}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*),\end{aligned}$$

by (5.8) and (5.7). It is not difficult to show that $\mathbf{Q}_{\theta_0}\mathbf{R}_{\theta_0}^{-1} = \mathbf{B}_{\theta_0}^{-1}\mathbf{H}_{\theta_0}$, and so

$\sqrt{n}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \sim \sqrt{n}\mathbf{B}_{\theta_0}^{-1}\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}}$. Hence, in the usual practical situation, when n is large and θ_0 is near enough ω to ensure that $\theta^* - \theta_0$ is near 0, we will have

$$-2 \log \mu \sim n(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})'\mathbf{B}_{\theta_0}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \sim n\hat{\boldsymbol{\lambda}}'\mathbf{H}_{\theta_0}'\mathbf{B}_{\theta_0}^{-1}\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}} = n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\theta_0}^{-1}\hat{\boldsymbol{\lambda}} \sim n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{-1}\hat{\boldsymbol{\lambda}},$$

and consequently the tests based on the critical regions C_1 and C_3 will have approximately the same power in these circumstances. Moreover it is easy to see that each of the random variables $-2 \log \mu$ and $-n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{-1}\hat{\boldsymbol{\lambda}}$ will then have approximately a non-central χ^2-distribution with r degrees of freedom and parameter $-n\boldsymbol{\lambda}^{*\prime}\mathbf{R}_{\hat{\theta}}^{-1}\boldsymbol{\lambda}^*$. (Again this argument could clearly be made rigorous by imagining θ_0 to vary with n in such a way that $\|\theta^* - \theta_0\| = O(n^{-\frac{1}{2}})$ and by imposing suitable conditions on the functions f and h).

We now consider the power of the Lagrangian multiplier test when n is large and θ_0 is not near ω. Then the asymptotic distribution of $\hat{\boldsymbol{\lambda}}_n$ will usually be as given in Lemma 1. Now, if λ^* is not near 0, then with a high probability $\sqrt{n}\hat{\lambda}$ will be far from 0 and since normally the matrix $-\mathbf{R}_{\hat{\theta}}$ will be positive definite, the power of the test based on C_3 will be near 1. However there is a possibility that θ_0 might be such that the function z has a stationary value at θ^*, in which case $\lambda^* = 0$. Then $-n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}\hat{\boldsymbol{\lambda}}$ would not necessarily be large with a high probability and consequently the power of the test based on C_3 would not be near 1 for such a θ_0. But this is a contingency which does not seem likely to arise often (the author has been unable to find an example of it) and we may conclude that in most practical situations the Lagrangian multiplier test is equivalent, for large samples, to the likelihood ratio test.

6. Singular information matrices. As we have said previously the whole problem of maximum likelihood estimation is closely bound up with the behavior of the function z. In particular, for unrestricted estimation it is important that z should have a maximum turning value in Ω at θ_0, for this condition plays an important part in ensuring consistency of $\hat{\theta}_n(\cdot, \Omega)$. Now the demands that $z(\theta_0)$ should be a maximum turning value of z in Ω and that $\mathbf{B}_{\theta_0}$ should be positive definite are not unrelated. For it is usually true that z has a stationary value at θ_0, i.e., that $\mathbf{D}z(\theta_0) = \mathbf{0}$ and also that $\mathbf{D}^2z(\theta_0) = -\mathbf{B}_{\theta_0}$: these results depend only on f being such that we can "differentiate under the integral sign." So that if θ is near θ_0 we will usually have

$$z(\theta) - z(\theta_0) = -\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{B}_{\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O(\|\theta - \theta_0\|^3). \tag{6.1}$$

Hence if $\mathbf{B}_{\theta_0}$ is not positive definite it may very well happen that $z(\theta_0)$ is not a maximum turning value of z in Ω and much of unrestricted estimation theory would then break down.

However, even if $\mathbf{B}_{\theta_0}$ is not positive definite and $z(\theta_0)$ is not a maximum turning value of z in Ω, it may still be the case that if θ_0 belongs to the subset ω of Ω, $z(\theta_0)$ is a maximum turning value of z in ω so that restricted estimation theory may not need drastic revision. And it is of some theoretical interest to consider just what revision is necessary in this case. Moreover this problem is of practical interest because it often happens that it is natural, either for reasons of symmetry or for some other reason, to describe the distribution of a random variable in terms of a parameter θ in such a way that neither is $\mathbf{B}_{\theta_0}$ positive definite nor is $z(\theta_0)$ a maximum of z in Ω. For instance if X has a multinomial distribution and describes an experiment in which an individual can fall into any one of s classes, it is natural for reasons of symmetry to denote the probabilities associated with the different classes by $\theta_i/\sum_{i=1}^{s}\theta_i$ $(i = 1, 2, \cdots, s)$. The set Ω of possible parameters is $\{\theta \varepsilon R^s : \theta_i > 0 \ (i = 1, 2, \cdots, s)\}$, and it is easy to verify that neither is $\mathbf{B}_\theta$ positive definite for any θ in Ω nor is $z(\theta_0)$ a maximum turning value of z in Ω. (In this case it is clear that this is so because we have set in s-dimensional space a parameter that is really $(s - 1)$-dimensional). However it is obvious that there is no difficulty about restricted estimation in the subset of Ω in which $\sum_{i=1}^{s}\theta_i = 1$.

We will now consider what revision is necessary of that part of the foregoing theory based on the assumptions A, if we drop the demand that $\mathbf{B}_{\theta_0}$ be positive definite (assumption 11A) and replace assumption 6A by the following assumption 6B, while maintaining the remainder of the assumptions A.

Assumption 6B. $\theta_0 \varepsilon \omega$ and for any other point θ of ω, $F(t, \theta) \neq F(t, \theta_0)$ for at least one t. Roughly speaking, we may explain the introduction of assumption 6B as follows. If assumption 6A is not satisfied, the parameter is not identifiable in the set Ω, i.e., there are different θ's in Ω which give the same distribution of X. However we wish θ_0 to be identifiable in the subset ω, in order that restricted estimation may still be possible. Hence we make assumption 6B.

It is easy to verify that these assumptions imply the existence of a consistent estimator $\hat{\theta}_n(\cdot, \omega)$ of θ_0, that for almost any x and sufficiently large n, $\hat{\theta}_n(x, \omega)$ with a Lagrangian multiplier $\lambda_n(x)$ satisfies the restricted likelihood equations and that

$$(6.2) \quad \begin{bmatrix} \mathbf{B}_{\theta_0} + o(1) & -\mathbf{H}_{\theta_0} + o(1) \\ -\mathbf{H}'_{\theta_0} + o(1) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_n(x, \omega) - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}}_n(x) \end{bmatrix} = \begin{bmatrix} \mathbf{D}n^{-1} \log L_n(x, \theta_0) \\ \mathbf{0} \end{bmatrix}$$

for almost any x. Now however, since we have dropped the requirement that $\mathbf{B}_{\theta_0}$ be positive definite and since subsequent theory concerning the asymptotic distributions of $\hat{\theta}_n(\cdot, \omega)$, λ_n and associated random variables makes considerable use of the inverse of $\mathbf{B}_{\theta_0}$, this theory no longer applies. To enable us to replace this theory we will now introduce assumption 11B which is associated with assumption 6B in the same manner as 11A was shown at the beginning of this section to be associated with 6A. This assumption will provide a natural connection between properties of the matrix $\mathbf{B}_{\theta_0}$, the subset ω and the facts that θ_0 is identifiable when it is known to belong to ω (assumption 6B), but unidentifiable in Ω.

Assumption 11B. The matrix $\mathbf{H}_{\theta_0}$ is of rank r. The matrix $\mathbf{B}_{\theta_0}$ is of rank $s - t$ where $t \leqq r$. There exists an $s \times t$ sub-matrix $\mathbf{H}_1$ of $\mathbf{H}_{\theta_0}$ such that $\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}_1'$ is positive definite. (Without any loss of generality we may assume that $\mathbf{H}_1$ is the matrix composed of the first t columns of $\mathbf{H}_{\theta_0}$ and we may write

$$\mathbf{H}_{\theta_0} = [\mathbf{H}_1\,\mathbf{H}_2]).$$

We will now define the set of assumptions B.

Assumptions B. By assumptions B we will mean the set of assumptions A with 6B and 11B replacing 6A and 11A respectively.

Now subject to assumptions B, if $\mathbf{y}$ denotes an s-dimensional random vector normally distributed with mean $\mathbf{0}$ and variance matrix $\mathbf{B}_{\theta_0}$ and if we write $\hat{\theta}$ in place of $\hat{\theta}_n(\cdot,\,\omega)$ and $\hat{\lambda}$ in place of $\hat{\lambda}_n(\cdot)$, then from (6.2) we have, as before,

$$\text{(6.3)} \qquad \sqrt{n}\begin{bmatrix} \mathbf{B}_{\theta_0} & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}'_{\theta_0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} \sim \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix},$$

and since $\sqrt{n}\,\mathbf{H}'_{\theta_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathbf{0}$ it follows that

$$\text{(6.4)} \qquad \sqrt{n}\begin{bmatrix} \mathbf{B}_{\theta_0} + \mathbf{H}_1\,\mathbf{H}'_1 & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}_{\theta_0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} \sim \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}.$$

Since $\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}'_1$ is positive definite and $\mathbf{H}_{\theta_0}$ is of rank r, the matrix

$$\begin{bmatrix} \mathbf{B}_{\theta_0} + \mathbf{H}_1\,\mathbf{H}'_1 & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}'_{\theta_0} & \mathbf{0} \end{bmatrix}$$

is non-singular and we define $\mathbf{P}^*_{\theta_0}$, $\mathbf{Q}^*_{\theta_0}$ and $\mathbf{R}^*_{\theta_0}$ by

$$\text{(6.5)} \qquad \begin{bmatrix} \mathbf{P}^*_{\theta_0} & \mathbf{Q}^*_{\theta_0} \\ \mathbf{Q}^{*\prime}_{\theta_0} & \mathbf{R}^*_{\theta_0} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{\theta_0} + \mathbf{H}_1\,\mathbf{H}'_1 & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}'_{\theta_0} & \mathbf{0} \end{bmatrix}^{-1}.$$

We will also define $\mathbf{S}_{\theta_0}$ by

$$\text{(6.6)} \qquad \mathbf{S}_{\theta_0} = -\mathbf{R}^*_{\theta_0} - \begin{bmatrix} \mathbf{I}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{I}_t$ denotes the unit $t \times t$ matrix.

We will now prove two lemmas concerning the distributions of statistics in which we are interested.

LEMMA 6. *Subject to assumptions B, the vector*

$$\sqrt{n}\begin{bmatrix} \hat{\boldsymbol{\theta}}' - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix}$$

is asymptotically normally distributed with mcan $\mathbf{0}$ *and variance matrix*

$$\begin{bmatrix} \mathbf{P}^*_{\theta_0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{\theta_0} \end{bmatrix}$$

Proof. From (6.4) we have, as previously, the result that

$$\sqrt{n}\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix}$$

is asymptotically normal with mean $\mathbf{0}$ and variance matrix

$$\begin{bmatrix} \mathbf{P}_{\theta_0}^* \mathbf{B}_{\theta_0} \mathbf{P}_{\theta_0}^* & \mathbf{P}_{\theta_0}^* \mathbf{B}_{\theta_0} \mathbf{Q}_{\theta_0}^* \\ \mathbf{Q}_{\theta_0}^{*\prime} \mathbf{B}_{\theta_0} \mathbf{P}_{\theta_0}^* & \mathbf{Q}_{\theta_0}^{*\prime} \mathbf{B}_{\theta_0} \mathbf{Q}_{\theta_0}^* \end{bmatrix}.$$

Now $\mathbf{P}_{\theta_0}^* \mathbf{B}_{\theta_0} \mathbf{P}_{\theta_0}^* = \mathbf{P}_{\theta_0}^*(\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}_1')\mathbf{P}_{\theta_0}^* - \mathbf{P}_{\theta_0}^*\mathbf{H}_1\mathbf{H}_1'\mathbf{P}_{\theta_0}^*$ and, as previously, the first term on the right hand side of this equation is $\mathbf{P}_{\theta_0}^*$. Also from (6.5)

$$\mathbf{P}_{\theta_0}^* \mathbf{H}_{\theta_0} = \mathbf{0}$$

and in particular $\mathbf{P}_{\theta_0}^* \mathbf{H}_1 = \mathbf{0}$. It follows that $\mathbf{P}_{\theta_0}^* \mathbf{B}_{\theta_0} \mathbf{P}_{\theta_0}^* = \mathbf{P}_{\theta_0}^*$; and in a similar manner it may be shown that $\mathbf{P}_{\theta_0}^* \mathbf{B}_{\theta_0} \mathbf{Q}_{\theta_0}^* = \mathbf{0}$. We also have

$$\mathbf{Q}_{\theta_0}^{*\prime}\mathbf{B}_{\theta_0}\mathbf{Q}_{\theta_0}^* = -\mathbf{R}_{\theta_0}^* - \mathbf{Q}_{\theta_0}^{*\prime}\mathbf{H}_1\mathbf{H}_1'\mathbf{Q}_{\theta_0}^*,$$

and from (6.5) $\mathbf{Q}_{\theta_0}^{*\prime}\mathbf{H}_{\theta_0} = -\mathbf{I}_r$, so that, in particular, $\mathbf{Q}_{\theta_0}^{*\prime}\mathbf{H}_1 = -[\mathbf{I}_t\ \mathbf{0}]'$. It follows that

$$\mathbf{Q}_{\theta_0}^{*\prime}\mathbf{B}_{\theta_0}\mathbf{Q}_{\theta_0}^* = -\mathbf{R}_{\theta_0}^* - \begin{bmatrix} \mathbf{I}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{S}_{\theta_0},$$

and this completes the proof.

Lemma 7. *Subject to assumptions* B, $-n\hat{\boldsymbol{\lambda}}'\mathbf{R}_{\hat{\theta}}^{*-1}\hat{\boldsymbol{\lambda}}$ *is asymptotically distributed as* χ^2 *with* $r - t$ *degrees of freedom.*

Proof. Since $\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}_1'$ is positive definite and $\mathbf{B}_{\theta_0}$ is of rank $s - t$, there exists a non-singular matrix $\mathbf{W}$ such that

$$\mathbf{W}'(\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}_1')\mathbf{W} = \mathbf{I}_s$$

and

$$\mathbf{W}'\mathbf{B}_{\theta_0}\mathbf{W} = \begin{bmatrix} \mathbf{\Lambda}_{s-t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{\Lambda}_{s-t}$ is a diagonal $s - t \times s - t$ matrix. Then

$$\mathbf{W}'\mathbf{H}_1\mathbf{H}_1'\mathbf{W} = \mathbf{I}_s - \begin{bmatrix} \mathbf{\Lambda}_{s-t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and since $\mathbf{H}_1\mathbf{H}_1'$ is of rank t, it follows that $\mathbf{\Lambda}_{s-t} = \mathbf{I}_{s-t}$ and that

$$\mathbf{W}'\mathbf{H}_1\mathbf{H}_1'\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_t \end{bmatrix}.$$

We now define an s-dimensional random variable $m = (m_1, m_2, \cdots, m_s)$ by $\mathbf{m} = \mathbf{W}'\mathbf{y}$. Then $\mathbf{m}$ is normally distributed with mean $\mathbf{0}$ and variance matrix

$$\mathbf{W}'\mathbf{B}_{\theta_0}\mathbf{W} = \begin{bmatrix} \mathbf{I}_{s-t} & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows that $m_1, m_2, \cdots, m_{s-t}$ are independent $N(0, 1)$ random variables, while $m_{s-t+1} = m_{s-t+2} = \cdots = m_s = 0$.

Now from (6.4) we have

$$\sqrt{n}\begin{bmatrix} (\mathbf{W}\mathbf{W}')^{-1} & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}_{\theta_0}' & 0 \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} \sim \begin{bmatrix} \mathbf{W}'^{-1}\mathbf{m} \\ 0 \end{bmatrix},$$

and so

$$(6.7) \qquad \sqrt{n}\begin{bmatrix} \mathbf{W}^{-1} & -\mathbf{W}'\mathbf{H}_{\theta_0} \\ -\mathbf{H}_{\theta_0}' & 0 \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} \sim \begin{bmatrix} \mathbf{m} \\ 0 \end{bmatrix}$$

Hence

$$\mathbf{m}'\mathbf{m} \sim n\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix}'\begin{bmatrix} (\mathbf{W}\mathbf{W}')^{-1} + \mathbf{H}_{\theta_0}\mathbf{H}_{\theta_0}' & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}_{\theta_0}' & \mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{bmatrix},$$

i.e., since $\mathbf{H}_{\theta_0}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathbf{0}$,

$$(6.8) \qquad \mathbf{m}'\mathbf{m} \sim n[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]'\mathbf{B}_{\theta_0}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0] + n\hat{\boldsymbol{\lambda}}'\mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}}$$

Now from (6.4) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathbf{P}_{\theta_0}^*(W')^{-1}\mathbf{m}$ and, as previously, $\mathbf{P}_{\theta_0}^*$ is of rank $s - r$. Hence asymptotically, when $n[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]'\mathbf{B}_{\theta_0}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]$ is expressed as a quadratic form in $m_1, m_2, \cdots, m_{s-t}$, its rank is at most $s - r$. We will now show that when $n\hat{\boldsymbol{\lambda}}'\mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}}$ is expressed as a quadratic form in $m_1, m_2, \cdots, m_{s-t}$, its rank is at most $r - t$.

From (6.7) we have, again since $\mathbf{H}_{\theta_0}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathbf{0}$,

$$-\mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{m} \sim \sqrt{n}\mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}}.$$

Now

$$\mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{m} = \begin{bmatrix} \mathbf{H}_1'\mathbf{W}\mathbf{m} \\ \mathbf{H}_2'\mathbf{W}\mathbf{m} \end{bmatrix}$$

and, since

$$\mathbf{m}'\mathbf{W}'\mathbf{H}_1\mathbf{H}_1'\mathbf{W}\mathbf{m} = \mathbf{m}'\begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I}_t \end{bmatrix}\mathbf{m} = 0,$$

we have $\mathbf{H}_1'\mathbf{W}\mathbf{m} = \mathbf{0}$. Hence

$$-\sqrt{n}\mathbf{H}_{\theta_0}'\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}} \sim \begin{bmatrix} 0 \\ \mathbf{H}_2' \end{bmatrix}\mathbf{W}\mathbf{m}.$$

Since the rank of $\mathbf{H}_2$ is at most $r - t$, it follows that, asymptotically, when $n\hat{\boldsymbol{\lambda}}'\mathbf{H}'_{\theta_0}\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}}$ is expressed as a quadratic form in $m_1, m_2, \cdots, m_{s-t}$, its rank is at most $r - t$. Now from (6.8) by applying Cochran's Theorem (Cramér [4]) we have the result that asymptotically $n[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]'\mathbf{B}_{\theta_0}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]$ and

$$n\hat{\boldsymbol{\lambda}}'\mathbf{H}'_{\theta_0}\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0}\hat{\boldsymbol{\lambda}}$$

are independently distributed as χ^2 with $s - r$ and $r - t$ degrees of freedom respectively.

The proof of Lemma 7 is completed by the remarks that

$$\mathbf{H}'_{\theta_0}\mathbf{W}\mathbf{W}'\mathbf{H}_{\theta_0} = \mathbf{H}'_{\theta_0}(\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}'_1)^{-1}\mathbf{H}_{\theta_0} = -\mathbf{R}^{*-1}_{\theta_0}$$

and that $\mathbf{R}^{*-1}_{\theta_0} \sim \mathbf{R}^{*-1}_{\hat{\theta}}$.

The results proved in this section, and the methods of proof, make it clear how the technique suggested by Aitchison and Silvey [1] for solving the restricted likelihood equations can usually be adapted, and how the Lagrangian multiplier test can usually be applied when the matrix $\mathbf{B}_{\theta_0}$ is singular and the function h is suitable. We will not amplify this point.

7. Different numbers of observations on several random variables. Experimental material being what it is, and experimenters being as they are, it is not often that the statistician is faced with an estimation problem in the ideal circumstances of being given a number of observations on a vector valued random variable. The more usual situation confronting him is that he is given n_1 observations on a random variable X_1 whose probability density function depends on s_1 parameters $\theta_1, \theta_2, \cdots, \theta_{s_1}$, n_2 observations on a random variable X_2 whose probability density function depends on s_2 parameters $\theta_{s_1+1}, \theta_{s_1+2}, \cdots, \theta_{s_1+s_2}, \cdots$ and n_k observations on a random variable X_k, whose probability density function depends on s_k parameters $\theta_{s_1+s_2+\cdots+s_{k-1}+1}, \cdots, \theta_s$, where $s = s_1 + s_2 + \cdots + s_k$. And he is presented with the problem of deciding whether the true parameter $\theta_0 = (\theta_1^0, \theta_2^0, \cdots, \theta_s^0)$ belongs to a set

$$\omega = \{\theta \,\varepsilon\, \Omega : h(\theta) = 0\},$$

Ω and h being as before. If $n_1 = n_2 = \cdots = n_k$ then we may interpret the observations as observations on a vector valued random variable and the foregoing theory applies. But if the n's are not all equal we cannot do this, and in order to enlarge the sphere of the Lagrangian multiplier test we have to consider this situation separately. In discussing it we will avoid all mathematical detail and will be content to indicate very briefly the modifications necessary in the test.

We will denote by x^* a given set of $n_1 + n_2 + \cdots + n_k$ observations on the random variables $X_1, X_2, \cdots, X_k$, and $\log L(x^*, \theta)$ will denote the value of the log-likelihood function at the point θ. Now if $\theta_0 \,\varepsilon\, \omega$ then the same kind of argument as we have used before may be used to show that it will usually be the case that $\hat{\theta}(x^*, \omega)$ exists, is near θ_0 when $n_1, n_2, \cdots, n_{k-1}$ and n_k are all large, and with a Lagrangian multiplier $\hat{\lambda}(x^*)$ satisfies the likelihood equations

$$\mathbf{D}\log L(x^*, \theta) + \mathbf{H}_\theta \boldsymbol{\lambda} = \mathbf{0}$$
$$\mathbf{h}(\theta) = \mathbf{0}.$$

We now introduce a matrix $\mathbf{N}$ defined by

$$\mathbf{N} = \begin{bmatrix} n_1 \mathbf{I}_{s_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & n_2 \mathbf{I}_{s_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & & n_k \mathbf{I}_{s_k} \end{bmatrix}$$

The information matrix $\mathbf{B}_\theta$ is defined in this case by

$$\mathbf{B}_\theta = -\mathbf{N}^{-1}[E_\theta \mathbf{D}^2 \log L(\cdot, \underline{\theta})],$$

where E_θ denotes expected value when θ is take as the true parameter. Then again by the type of argument used previously we may show that for most x^*, when $n_1, n_2, \cdots, n_k$ are large,

$$\begin{bmatrix} \mathbf{NB}_{\theta_0} & -\mathbf{H}_{\theta_0} \\ -\mathbf{H}'_{\theta_0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}(x^*, \omega) - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}}(x^*) \end{bmatrix} \sim \begin{bmatrix} \mathbf{D} \log L(x^*, \theta_0) \\ \mathbf{0} \end{bmatrix}. \quad (7.1)$$

Also it will usually be true that $\mathbf{D}\log L(x^*, \theta_0)$ can be regarded as an observation on a random variable which is approximately normal with mean $\mathbf{0}$ and variance matrix $\mathbf{NB}_{\theta_0}$.

Now in the case where $\mathbf{B}_{\theta_0}$ is positive definite we may use (7.1) in the same way as before to show that when $\theta_0 \,\varepsilon\, \omega$ and $n_1, n_2, \cdots, n_k$ are large,

$$\hat{\boldsymbol{\lambda}}'\mathbf{H}'_{\hat{\theta}}[\mathbf{NB}_{\hat{\theta}}]^{-1}\mathbf{H}_{\hat{\theta}}\hat{\boldsymbol{\lambda}}$$

will usually be distributed approximately as χ^2 with r degrees of freedom, and it is this statistic which we use in the modified form of the Lagrangian multiplier test. Alternatively when $\mathbf{B}_{\theta_0}$ is of rank $s - t$, when each of the functions $h_1, h_2, \cdots, h_t$ is a function of only the parameters involved in the distribution of one of the X's and $\mathbf{B}_{\theta_0} + \mathbf{H}_1\mathbf{H}_1'$ is positive definite, the statistic on which the test is based is $\hat{\boldsymbol{\lambda}}'\mathbf{H}'_{\hat{\theta}}[\mathbf{N}(\mathbf{B}_{\hat{\theta}} + \mathbf{H}_1\mathbf{H}_1')]^{-1}\mathbf{H}_{\hat{\theta}}\hat{\boldsymbol{\lambda}}$, which will usually be distributed as χ^2 with $r - t$ degrees of freedom when $n_1, n_2, \cdots, n_k$ are large.

We conclude by applying the Lagrangian multiplier test in a familiar situation.

Homogeneity in the 2 × 2 *contingency table.* One of the three situations (Cochran [2]) in which the 2 × 2 contingency table arises is as follows. We are given n_1 observations on a random variable X_1 whose distribution is defined by

$$\Pr\{X_1 = (1, 0)\} = \theta_1^0/(\theta_1^0 + \theta_2^0),$$
$$\Pr\{X_1 = (0, 1)\} = \theta_2^0/(\theta_1^0 + \theta_2^0),$$

and n_2 observations on an independent random variable X_2 whose distribution is defined similarly in terms of θ_3^0 and θ_4^0. These observations can be summarised in a 2×2 contingency table as follows.

Number of occurrences of different values of X_1 and X_2.

	(1, 0)	(0, 1)	Total
X_1	n_{11}	n_{12}	n_1
X_2	n_{21}	n_{22}	n_2
Total	m_1	m_2	n

We suppose that the point $\theta_0 = (\theta_1^0, \theta_2^0, \theta_3^0, \theta_4^0)$ is known to belong to the set $\Omega = \{\theta \varepsilon R^4 : \epsilon \leqq \theta_i \leqq 1/\epsilon \ (i = 1, 2, 3, 4)\}$ where ϵ is a small positive number. In this case we also have

$$\log L(x^*, \theta) = \text{constant} + n_{11} \log \theta_1 + n_{12} \log \theta_2 - n_1 \log (\theta_1 + \theta_2)$$
$$+ n_{21} \log \theta_3 + n_{22} \log \theta_4 - n_2 \log (\theta_3 + \theta_4).$$

The matrix

$$\mathbf{B}_\theta = \begin{bmatrix} \theta_1^{-1} - (\theta_1 + \theta_2)^{-1} & -(\theta_1 + \theta_2)^{-1} & 0 & 0 \\ -(\theta_1 + \theta_2)^{-1} & \theta_2^{-1} - (\theta_1 + \theta_2)^{-1} & 0 & 0 \\ 0 & 0 & \theta_3^{-1} - (\theta_3 + \theta_4)^{-1} & -(\theta_3 + \theta_4)^{-1} \\ 0 & 0 & -(\theta_3 + \theta_4)^{-1} & \theta_4^{-1} - (\theta_3 + \theta_4)^{-1} \end{bmatrix}$$

has rank 2. Homogeneity of X_1 and X_2 means that $\theta_1^0/(\theta_1^0 + \theta_2^0) = \theta_3^0/(\theta_3^0 + \theta_4^0)$ and we consider estimating θ_0 subject to the restrictions

$$\mathbf{h}(\theta) = \begin{bmatrix} \theta_1 + \theta_2 - 1 \\ \theta_3 + \theta_4 - 1 \\ \theta_1 - \theta_3 \end{bmatrix} = \mathbf{0},$$

so that

$$\mathbf{H}_\theta \equiv \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix}.$$

If $\mathbf{H}_1$ is the leading 4×2 sub-matrix of $\mathbf{H}_\theta$, then for any $\theta \varepsilon \omega$,

$$\mathbf{B}_\theta + \mathbf{H}_1 \mathbf{H}_1' = \begin{bmatrix} \theta_1^{-1} & 0 & 0 & 0 \\ 0 & \theta_2^{-1} & 0 & 0 \\ 0 & 0 & \theta_3^{-1} & 0 \\ 0 & 0 & 0 & \theta_4^{-1} \end{bmatrix}$$

which is positive definite.

The likelihood equations are easily solved in this case and we find that

$$\hat{\theta}_1(x^*, \omega) = \hat{\theta}_3(x^*, \omega) = m_1/n$$

while $\hat{\theta}_2(x^*, \omega) = \hat{\theta}_4(x^*, \omega) = m_2/n$. It is not difficult to verify that the statistic $\hat{\boldsymbol{\lambda}}'\mathbf{H}_{\hat{\theta}}'[\mathbf{N}(\mathbf{B}_{\hat{\theta}} + \mathbf{H}_1\mathbf{H}_1')]^{-1}\mathbf{H}_{\hat{\theta}}\hat{\boldsymbol{\lambda}}$ is the usual statistic used in the χ^2-test of homogeneity in a 2 × 2 table, so that this test is a particular case of the Lagrangian multiplier test. And it illustrates most aspects of the preceding theory. The computational procedure for applying the Lagrangian multiplier test in less familiar and more complicated situations will be set out in a subsequent paper.

REFERENCES

[1] J. Aitchison and S. D. Silvey, "Maximum likelihood estimation of parameters subject to restraints," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 813–828.

[2] W. G. Cochran, "The χ^2-test of goodness of fit," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 315–345.

[3] H. Cramér, *Random Variables and Probability Distributions*, Cambridge University Press, 1937.

[4] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[5] C. Kraft and L. LeCam, "A remark on the roots of the maximum likelihood equation," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1174–1177.

[6] H. B. Mann and A. Wald, "On stochastic limit and order relationships," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 217–226.

[7] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 595–601.

[8] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Am. Math. Soc.*, Vol. 54 (1943), pp. 426–482.