ON THE DISTRIBUTION OF THE LIKELIHOOD RATIO[1]

By Herman Chernoff

Stanford University

1. Summary and Introduction. A classical result due to Wilks [1] on the distribution of the likelihood ratio λ is the following. Under suitable regularity conditions, if the hypothesis that a parameter θ lies on an r-dimensional hyperplane of k-dimensional space is true, the distribution of $-2 \log \lambda$ is asymptotically that of χ^2 with $k - r$ degrees of freedom.

In many important problems it is desired to test hypotheses which are not quite of the above type. For example, one may wish to test whether θ is on one side of a hyperplane, or to test whether θ is in the positive quadrant of a two-dimensional space. The asymptotic distribution of $-2 \log \lambda$ is examined when the value of the parameter is a boundary point of both the set of θ corresponding to the hypothesis and the set of θ corresponding to the alternative.

First the case of a single observation from a multivariate normal distribution, with mean θ and known covariance matrix, is treated. The general case is then shown to reduce to this special case where the covariance matrix is replaced by the inverse of the information matrix. In particular, if one tests whether θ is on one side or the other of a smooth $(k - 1)$-dimensional surface in k-dimensional space and θ lies on the surface, the asymptotic distribution of λ is that of a chance variable which is zero half the time and which behaves like χ^2 with one degree of freedom the other half of the time.

2. Notation and background. We shall use some of the notation and results of Mann and Wald [2]. In particular, if $\{x_n\}$ is a sequence of k-dimensional chance variables and $\{f_n\}$ a sequence of positive numbers, we write

$$x_n = O_p(f_n) \tag{1}$$

if for each $\epsilon > 0$, there is an M_ϵ such that $P\{| x_n | > M_\epsilon f_n\} < 1 - \epsilon$. Similarly we write

$$x_n = o_p(f_n) \tag{2}$$

if $x_n/f_n \rightarrow 0$ in probability, or equivalently, if, for each $\epsilon > 0$, there is a sequence $M_{n\epsilon} \rightarrow 0$ such that $P\{| x_n | > M_{n\epsilon} f_n\} < 1 - \epsilon$.

Mann and Wald have shown that the calculus used with the usual O and o notation applies to O_p and o_p. For example, if $x_n = O_p(\sqrt{n})$ and $y_n = o_p(1)$, then $x_n y_n = o_p(\sqrt{n})$. We shall frequently drop the subscript n where there is no ambiguity.

We write $d(x)$ for the distribution of x and $d\infty(x_n) = d(x)$ if the limiting

Received 7/24/53.

[1] This work was prepared with the partial support of the Office of Naval Research.

distribution of x_n exists and is that of x. The following are included in results of Mann and Wald [2]:

If $d\infty(x_n) = d(x)$, then $d\infty[x_n + o_p(1)] = d(x)$.

If $d\infty(x_n) = d(x)$ and g is continuous, then $d\infty[g(x_n)] = d[g(x)]$.

If $x_n \to c$ in probability and $x_n - c = O_p(f_n)$ and g has continuous rth order derivatives at c, then

$$g(x_n) = T(x_n, c, r) + o_p(f_n^r) \tag{3}$$

where $T(x, c, r)$ is the rth order Taylor expansion of $g(x)$ about c.

Furthermore, we shall use some properties of likelihood functions and maximum likelihood estimates which are implied by the following regularity conditions $\mathfrak{R}$ [3].

CONDITIONS $\mathfrak{R}$. The data $X = (x_1, x_2, \cdots, x_n)$ consist of n independent observations with common density $f(x, \theta)$ satisfying

(a) For almost all x, the derivatives

$$\frac{\partial \log f}{\partial \theta_i}, \quad \frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j}, \quad \frac{\partial^3 \log f}{\partial \theta_i \partial \theta_j \partial \theta_m}$$

exist for every θ in the closure of a neighborhood N of $\theta = 0$.

(b) *If* $\theta \varepsilon N$,

$$\left|\frac{\partial f}{\partial \theta_i}\right| < F(x), \qquad \left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}\right| < F(x), \qquad \left|\frac{\partial^3 \log f}{\partial \theta_i \partial \theta_j \partial \theta_m}\right| < H(x),$$

where F is finitely integrable and $E\{H(x)\} < M$, with M independent of θ.

(c) *If* $\theta \varepsilon N$, $\left\| E\left\{\frac{\partial \log f}{\partial \theta_i} \frac{\partial \log f}{\partial \theta_j}\right\} \right\|$ is finite and positive definite.

Let the likelihood function be given by

$$L(X, \theta) = \prod_{\alpha=1}^{n} f(x_\alpha, \theta).$$

From the above conditions, it follows that for $\theta \varepsilon N$

$$\frac{1}{n} \log L(X, \theta) = \frac{1}{n} \log L(X, 0) + A'\theta + \tfrac{1}{2}\theta' B\theta + |\theta|^3 \cdot O_p(1), \tag{4}$$

where A is the vector whose ith component is

$$A_i = \frac{1}{n} \sum_{\alpha=1}^{n} \frac{\partial \log f(x_\alpha, 0)}{\partial \theta_i}$$

and B is the matrix whose (i, j) term is

$$B_{ij} = \frac{1}{n} \sum_{\alpha=1}^{n} \frac{\partial^2 \log f(x_\alpha, 0)}{\partial \theta_i \partial \theta_j}.$$

If the "true" value of the parameter is given by $\theta = 0$, that is, θ is at the origin, then the asymptotic distribution of $\sqrt{n}\, A$ is normal with mean 0 and covariance matrix J, where

$$J = \left\| E\left\{\frac{\partial \log f(x, 0)}{\partial \theta_i} \frac{\partial \log f(x, 0)}{\partial \theta_j}\right\} \right\| = \| J_{ij} \|$$

is the positive definite information matrix and, furthermore, $B \to -J$ in probability. Also, the maximum likelihood estimate $\hat{\theta}$ computed under the assumption $\theta \varepsilon N$ satisfies

$$\hat{\theta} = J^{-1}A + o_p(1/\sqrt{n}). \tag{5}$$

Let us consider the likelihood ratio for the test of a hypothesis $H:\theta \varepsilon \omega \subset N$ against the alternative $K:\theta \varepsilon \tau \subset N$. For a set $\varphi \subset N$ in k-dimensional space we define

$$P_\varphi(X) = \sup_{\theta \varepsilon \varphi} L(X, \theta), \tag{6}$$

$$\lambda(X) = P_\omega(X)/P_{\omega \cup \tau}(X), \tag{7}$$

$$\lambda^*(X) = P_\omega(X)/P_\tau(X). \tag{8}$$

Since λ^* is more expressive than λ (that is, $\lambda = \lambda^*$ if $\lambda^* \leqq 1$ and $\lambda = 1$ if $\lambda^* > 1$) it suffices to study the distribution of λ^*. We also define $\hat{\theta}_\varphi$ as that value of θ in the closure of φ which maximizes $L(X, \theta)$. Then $L(X, \hat{\theta}_\varphi) = P_\varphi(X)$.

3. Examples involving the normal distribution. We shall present a few examples where the observations x have a multivariate normal distribution with mean θ and known covariance matrix Σ. Since the sample mean is a sufficient statistic for the mean of a normal distribution, it suffices to treat the case where the sample size $n = 1$. For this case

$$P_\omega(x) = (2\pi)^{-k/2} \mid \Sigma \mid^{-\frac{1}{2}} e^{-Q_\omega(x)/2} \tag{9}$$

where $Q_\omega(x) = \inf_{\theta \varepsilon \omega} (x - \theta)' \Sigma^{-1}(x - \theta)$. Then

$$-2 \log \lambda^*(x) = Q_\omega(x) - Q_\tau(x). \tag{10}$$

In the special case where $\Sigma = I$, $Q_\omega(x)$ represents the squared distance from x to ω. In the more general case also, $Q_\omega(x)$ may be regarded as a squared distance, but the distance is measured with respect to a non-Euclidean metric. We treat the following examples of varying degrees of specialty.

EXAMPLE 1. Let $\omega = \{(\theta_1, \theta_2): a_1\theta_1 + a_2\theta_2 < 0 < b_1\theta_1 + b_2\theta_2\}$, and let the complement of ω be τ, and $\Sigma = I$. Because of the symmetric nature of λ^*, we may assume that the angle φ from the vector (a_1, a_2) to the vector (b_1, b_2) is no more than 180°. Then ω represents a cone with vertex at the origin and with angle φ between the boundary lines. For $x \varepsilon \omega$, $-2 \log \lambda^*(x)$ is the negative of the squared distance from x to the boundary of ω; for $x \varepsilon \tau$, $-2 \log \lambda^*(x)$ is the squared distance from x to the boundary of ω (Fig. 1). If $\theta = (0, 0)$, the distribution of $-2 \log \lambda^*(x)$ depends only on φ. Let us call this distribution G_φ.

We note that if $\varphi = 180°$, we may rotate the axes so that ω is the right half

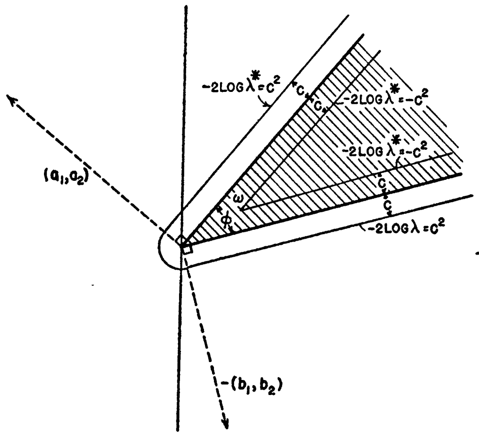

FIG. 1

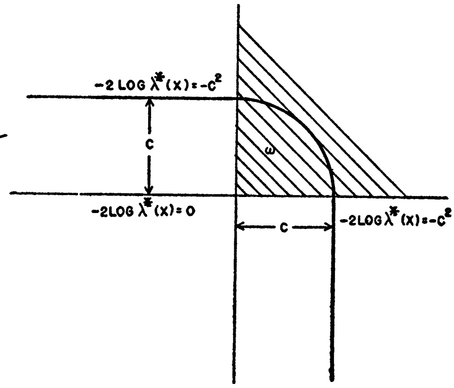

FIG. 2

plane. Then $-2 \log \lambda^*(x) = -x_1^2$ for $x_1 > 0$, and $= x_1^2$ for $x_1 \leqq 0$. Then $G_{180°}$ is characterized by the density

$$g(y) = \tfrac{1}{2}(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}|y|}\,|\,y\,|^{-\frac{1}{2}},$$

which is related to that of χ^2 in an obvious way.

EXAMPLE 2. We alter Example 1 to let Σ be a known positive definite symmetric matrix, not necessarily I. Then there is a nonsingular matrix T such that $\Sigma^{-1} = T'T$. Let $y = Tx$, $d = Ta$, $e = Tb$. Our problem then is reduced to that of the previous example where φ is now the angle from d to e. Since straight lines go into straight lines under a linear transformation, it follows that $\varphi \lesseqqgtr 180°$ if and only if the angle from a to b is $\lesseqqgtr 180°$. Also

$$\cos \varphi = \frac{d'e}{|\,d\,|\,|\,e\,|} = \frac{a'T'Tb}{\sqrt{(a'T'Ta)(b'T'Tb)}} = \frac{a'\Sigma^{-1}b}{\sqrt{(a'\Sigma^{-1}a)(b'\Sigma^{-1}b)}}.$$

EXAMPLE 3. Let ω be the first quadrant except for the origin and τ be the set consisting in the origin alone. Let $\Sigma = I$. Then $-2 \log \lambda^*(x)$ is

	$x_1 \leqq 0$	$0 \leqq x_1$
$0 \leqq x_2$	$-x_2^2$	$-(x_1^2 + x_2^2)$
$x_2 \leqq 0$	0	$-x_1^2$

It is easily seen (Fig. 2) that if $\theta = (0, 0)$,

$$P(2 \log \lambda^*(x) \leqq c) = \begin{cases} 0, & c < 0, \\ \tfrac{1}{2} + \tfrac{1}{2}P_1(c) + \tfrac{1}{4}P_2(c), & c \geqq 0, \end{cases}$$

where P_1 and P_2 are the c.d.f.'s for the χ^2 distributions with one and two degrees of freedom, respectively.

4. Main result. We sball now treat the case where θ is a k-dimensional parameter, and the density of $f(x, \theta)$ satisfies the regularity conditions $\mathfrak{R}$. We first prove

LEMMA 1. *If the origin is a limit point of* φ *and* $\hat{\theta}_\varphi \to 0$ *when* $\theta = 0$, *then* $\hat{\theta}_\varphi = O_p(1/\sqrt{n})$ *when* $\theta = 0$.

PROOF. Refer to equation 4 with θ replaced by $\hat{\theta}_\varphi$. For each $\epsilon > 0$ there is a sequence $c_{n\epsilon} \to 0$ and a K_ϵ such that with probability greater than $1 - \epsilon$

$$|\hat{\theta}_\varphi| < c_{n\epsilon}, \qquad |A| < \frac{K_\epsilon}{\sqrt{n}}, \qquad \sum_{i,j=1}^{k} (B_{ij} + J_{ij})^2 < c_{n\epsilon}$$

and the term represented by $|\hat{\theta}_\varphi|^3 O_p(1)$ is less than $K_\epsilon |\hat{\theta}_\varphi|^3$. When these inequalities are satisfied there is a K_ϵ^* such that

$$0 < A'\hat{\theta}_\varphi + \tfrac{1}{2}\hat{\theta}_\varphi' B\hat{\theta}_\varphi + |\hat{\theta}_\varphi|^3 O_p(1) < -\tfrac{1}{2}\hat{\theta}_\varphi' J\hat{\theta}_\varphi + K_\epsilon^* \left(\frac{|\hat{\theta}_\varphi|}{\sqrt{n}} + c_{n\epsilon} |\hat{\theta}_\varphi|^2\right).$$

But then there is a K_ϵ^{**} such that $\hat{\theta}_\varphi < K_\epsilon^{**}/\sqrt{n}$. The lemma follows.

DEFINITION 1. *A set C is positively homogeneous if* $\theta \varepsilon C$ *implies* $a\theta \varepsilon C$ *for* $a > 0$.

DEFINITION 2. *The set* φ *is approximated by the positively homogeneous set* C_φ *if*

$$\inf_{x \varepsilon C_\varphi} |x - y| = o(|y|) \text{ for } y \varepsilon \varphi \quad \text{and} \quad \inf_{y \varepsilon \varphi} |x - y| = o(|x|) \text{ for } x \varepsilon C_\varphi.$$

We may remark that a set bounded by smooth surface through the origin is approximated by the union of an open half-space with an optional positively homogeneous subset of the tangent hyperplane. It is also easy to see that if φ is approximated by a nonnull positively homogeneous set other than the whole space, then the origin is a boundary point of φ.

THEOREM 1. *If*

(1) *the regularity conditions* $\mathfrak{R}$ *are satisfied,*

(2) *the origin is a boundary point of* φ *implies that* $\hat{\theta}_\varphi \to 0$ *in probability when* $\theta = 0$, *and*

(3) *the sets* ω *and* τ *are approximated by nonnull and disjoint positively homogeneous sets* C_ω *and* C_τ,

then, when $\theta = 0$, *the asymptotic distribution of* λ^* *is the same as it would be for the test of* $\theta \varepsilon C_\omega$ *against* $\theta \varepsilon C_\tau$ *based on one observation from a population with distribution* $N(\theta, J^{-1})$.

PROOF. Throughout this proof we assume that the "true" value of the parameter is 0 and we use θ to represent the argument of the likelihood function. Since $\hat{\theta} = J^{-1}A + o_p(1/\sqrt{n})$,

$$\frac{1}{n} \log L(X, \hat{\theta}) = \frac{1}{n} \log L(X, 0) + \tfrac{1}{2}A'J^{-1}A + o_p(1/n).$$

Let $\theta = J^{-1}A + \eta$, with $\eta = O_p(1/\sqrt{n})$. Then

$$\frac{1}{n} \log L(X, \theta) = \frac{1}{n} \log L(X, 0) + \tfrac{1}{2}A'J^{-1}A - \tfrac{1}{2}\eta' J\eta + o_p(1/n).$$

Applying Lemma 1 to ω and τ,

$$-2 \log \lambda^*(X) = n[\inf_{\theta \varepsilon \omega} \eta' J \eta - \inf_{\theta \varepsilon \tau} \eta' J \eta] + o_p(1).$$

But

$$\inf_{\theta \varepsilon \varphi} (y - \theta)' J (y - \theta) = \inf_{\theta \varepsilon C_\varphi} (y - \theta)' J (y - \theta) + o(|y|^2)$$

and therefore

$$-2 \log \lambda^*(X) =$$

$$n[\inf_{\theta \varepsilon C_\omega} (J^{-1}A - \theta)' J (J^{-1}A - \theta) - \inf_{\theta \varepsilon C_\tau} (J^{-1}A - \theta)' J (J^{-1}A - \theta)] + o_p(1).$$

Since C_ω and C_τ are positively homogeneous,

$$-2 \log \lambda^*(X) = \inf_{\theta \varepsilon C_\omega} (z - \theta)' J (z - \theta) - \inf_{\theta \varepsilon C_\tau} (z - \theta)' J (z - \theta) + o_p(1),$$

where $z = \sqrt{n}\, J^{-1}A$ and $d\infty(z) = N(0, J^{-1})$. The function $\inf_{\theta \varepsilon C_\omega}(z - \theta)' J(z - \theta)$ is certainly a continuous function of z. From the results of Mann and Wald [2] it follows that the asymptotic distribution of $-2 \log \lambda^*(X)$ is precisely the distribution of

$$\inf_{\theta \varepsilon C_\omega} (z - \theta)' J (z - \theta) - \inf_{\theta \varepsilon C_\tau} (z - \theta)' J (z - \theta)$$

under the assumption that $d(z) = N(0, J^{-1})$. Referring to Section 3, we see that this is precisely the result we seek.

REMARK. The final sentence of the introduction is a simple consequence of this theorem, together with the obvious extension of Example 1 in Section 3, the fact that nonsingular linear transformations transform hyperplanes into hyperplanes, the nature of the positively homogeneous approximation of a set bounded by a smooth surface, and the relation of λ to λ^*.

REFERENCES

[1] S. S. WILKS, "The large sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Stat.*, Vol. 9 (1938), p. 60.

[2] H. B. MANN, AND A. WALD, "On stochastic limit and order relationships," *Ann. Math. Stat.*, Vol. 14 (1943), p. 217.

[3] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946.