The Annals of Statistics
1973, Vol. 1, No. 1, 161-165

NONCENTRAL CONVERGENCE OF WALD'S LARGE-SAMPLE TEST STATISTIC IN EXPONENTIAL FAMILIES[1]

BY T. W. F. STROUD

Queen's University at Kingston

It is shown that under very mild assumptions Wald's large-sample test statistic (quadratic form based on unrestricted maximum likelihood estimators) converges to noncentral chi-square under a sequence of local alternatives of the order $n^{-\frac{1}{2}}$, when the family of distributions is assumed to be of exponential type. This eliminates, for these families, the necessity of invoking the strict regularity conditions of Wald for the purpose of justifying the asymptotic distribution.

1. Introduction and summary. Let $\hat{\theta}_n$ be the maximum likelihood estimator (MLE) of a vector parameter θ based on a sample of n independent vector observations from $f(x; \theta)$. To test the hypothesis $\beta(\theta) = 0$ versus $\beta(\theta) \neq 0$, where β is a vector-valued function, one may reject for large values of the quadratic form $n[\beta(\hat{\theta}_n)]'\Sigma_\beta^{-1}(\hat{\theta}_n)[\beta(\hat{\theta}_n)]$, where $\Sigma_\beta(\theta)$ is the asymptotic covariance matrix of $n^{\frac{1}{2}}[\beta(\hat{\theta}_n) - \beta(\theta)]$. It is readily seen that the asymptotic null distribution of this statistic, under weak conditions, is central chi-square. Under a sequence of local alternatives $\{\theta_n\}$ converging to the null hypothesis at the rate $n^{-\frac{1}{2}}$, Wald [8] showed the statistic to have a limiting noncentral chi-square distribution. In the same paper, Wald proved a number of asymptotic optimality properties, and to this end he assumed some rather severe regularity conditions which do not hold in many statistical problems. Similar results were also obtained for the likelihood-ratio test.

A realistic criterion for the convergence to noncentral chi-square of the above quadratic form (known here as "the Wald statistic") has been presented by the author [7]. Briefly, it is sufficient that $n^{\frac{1}{2}}(\hat{\theta}_n - \theta_n)$ converge to a normal law $\mathscr{N}(0, \Sigma_0)$ and that the matrix used in the quadratic form as estimating the covariance matrix of $n^{\frac{1}{2}}(\hat{\theta}_n - \theta_n)$ converge stochastically to Σ_0, where Σ_0 is nonsingular. $\hat{\theta}_n$ and the matrix estimator may be any estimators, not necessarily MLE's. It has been shown [7] that this criterion is satisfied when MLE's are used in one-sample and two-sample problems, where the sampling is from multivariate normal distributions with unrestricted mean vectors and (nonsingular) covariance matrices. The present paper extends this result from the multivariate normal to the general exponential family, subject to very mild assumptions. The stated results deal with one-sample problems; the extension to two-sample problems may be performed as in [7].

Received July 19, 1971; revised June 5, 1972.

[1] Supported in part by National Research Council of Canada Grant A7220. This work was completed while the author was a fellow of the Summer Research Institute of the Canadian Mathematical Congress at Université Laval.

Exponential families are closely linked with maximum likelihood estimation because of the simple form assumed in such families by the MLE, and also because these are the only "regular" families where for arbitrary sample size the MLE can be a sufficient statistic ([5] page 51), although, as LeCam [4] has pointed out, Wald's results [8] show that, under the conditions assumed by Wald, MLE's are "asymptotically sufficient." Thus the families of distributions studied in this paper comprise those for which the Wald statistic based on the MLE uses, for finite sample sizes, all the relevant information.

In the remaining section, the techniques of proof used in [7] are applied to show that the sufficient condition for convergence described above holds for the multiparameter exponential family, under conditions which usually hold in practice when the family is genuinely exponential (degenerate boundary cases such as $|\Sigma| = 0$ in the multivariate normal or $p_i = 0$ for some i in the case of the multinomial must be excluded).

2. Convergence to noncentral chi-square. A key step in the proof of the main result is contained in the following lemma, which is a multivariate double-sequence version of the weak law of large numbers and central limit theorem.

LEMMA. *For $n = 1, 2, \cdots$, suppose t_n is a random vector with the representation $nt_n = \sum_{\alpha=1}^{n} w_{n\alpha}$, where for each n the $w_{n\alpha}$ are independent random vectors with common mean vector τ_n and covariance matrix Σ_n. Suppose also that* (i) $\tau_n \to \tau_0$, (ii) $\Sigma_n \to \Sigma_0 > 0$, *and* (iii) *there exists $M < \infty$ such that the fourth moments of $w_{n\alpha}$ are all $< M$ for all n. Then* $\operatorname{plim} t_n = \tau_0$ *and* $\mathscr{L}[n^{\frac{1}{2}}(t_n - \tau_n)] \to \mathscr{N}(0, \Sigma_0)$.

PROOF. The reader is referred to [7] (pages 1419–1421), where the result is proved for the case where t_n is the sample covariance matrix from a multivariate normal. In fact the same proof may be applied directly to the lemma as stated above. □

The main result is formulated in terms of the *natural parametrization* of the exponential family (see Lehmann [5] page 51–53) for some of its properties. The exponential models used widely in practice (gamma, Poisson, multinomial, etc.) may be put into this parametrization fairly readily. Of course, because MLE's are invariant under any one-one reparametrization, the Wald statistic has the same value regardless of the parametrization, and may be calculated in whatever way is most convenient. In particular, the *proof* of the main result uses for the parameter the expected value of the function of the observation appearing in the exponent, because this renders the MLE unbiased and hence facilitates application of the above lemma. Dempster [2] calls this the *moment* parametrization and points out some interesting duality properties between it and the natural parametrization. The only assumptions we need make about the exponential family (other than measurability) are those which guarantee a one-one correspondence between the two parametrizations and the existence of the MLE.

The following theorem gives conditions for a limiting noncentral chi-square distribution for the Wald statistic for testing $H: \beta(\theta) = 0$ versus $A: \beta(\theta) \neq 0$, where the parameter space Θ is assumed to be an open subset of the *natural parameter space* Θ_N ([5] page 51). We consider a fixed $\theta_0 \in \Theta$ satisfying $\beta(\theta_0) = 0$, and a sequence $\{\theta_n\}$ converging to θ_0 such that $n^{\frac{1}{2}}(\theta_n - \theta_0) \to \delta$. For the special case $\delta = 0$, a limiting central chi-square distribution is obtained. In what follows, the first subscript n refers to the position in the sequence $\{\theta_n\}$, a second subscript denotes place in a sample of size n, and superscripts refer to components of vectors. The symbol $\mathscr{E}^p$ denotes p-dimensional Euclidean space.

THEOREM. *For $n = 1, 2, \cdots$, let the observations in the sample $(x_{n1}, \cdots, x_{nn})$ be independently and identically distributed according to the density*

$$f(x; \theta_n) = c(\theta_n)k(x) \exp\{\sum_{i=1}^{p} \theta_n{}^i t^i(x)\}$$

(where x denotes any $x_{n\alpha}$) with respect to the σ-finite measure μ on the measurable space $(\mathscr{X}, \mathscr{A})$. Assume $t^1, \cdots, t^p$ and k are measurable real-valued functions on $(\mathscr{X}, \mathscr{A})$, and that $\theta_n \in \Theta$ for all n, where Θ is a fixed open subset of the fully p-dimensional natural parameter space Θ_N. Assume also that:

ASSUMPTION 1. *The matrix C_θ whose (i, j)th component is $-\partial^2 \log c(\theta)/\partial\theta^i \partial\theta^j$ is strictly positive definite for all $\theta \in \Theta_N{}^0$, the interior of Θ_N.*

ASSUMPTION 2. *For sufficiently large n the equations*

$$\frac{\partial}{\partial\theta^i} \log c(\theta) + u^i = 0\,, \qquad i = 1, \cdots, p$$

where $u^i = n^{-1} \sum_{\alpha=1}^{n} t^i(x_{n\alpha})$, have a solution $\hat{\theta}_n = \psi(u)$ in $\Theta_N{}^0$ with probability 1.

Let $\beta: \Theta_N \to \mathscr{E}^q$ satisfy $\beta(\theta_0) = 0$, where $\theta_0 \in \Theta$, and assume β to have continuous second partial derivatives in some sphere about θ_0. Assume further that the sequence of parameter points $\{\theta_n\}$ is such that $n^{\frac{1}{2}}(\theta_n - \theta_0) \to \delta$, for some fixed $\delta \in \mathscr{E}^p$.

Let B_θ be the matrix $(\partial\beta^r/\partial\theta^s)$, $r = 1, \cdots, q$; $s = 1, \cdots, p$. Then the distribution of the statistic $W_n = n[\beta(\hat{\theta}_n)]'(B_{\hat{\theta}_n} C_{\hat{\theta}_n}^{-1} B'_{\hat{\theta}_n})^{-1}\beta(\hat{\theta}_n)$ converges, as $n \to \infty$, to the noncentral chi-square distribution with q degrees of freedom and noncentrality parameter equal to

$$\delta' B'_{\theta_0}(B_{\theta_0} C_{\theta_0}^{-1} B'_{\theta_0})^{-1} B_{\theta_0} \delta\,.$$

REMARKS. Note that C_θ^{-1} is the asymptotic covariance matrix of $n^{\frac{1}{2}}(\hat{\theta} - \theta)$ (provided there is sufficient regularity) for fixed θ where $\hat{\theta}$ is the MLE of θ based on sample size n. Hence $B_\theta C_\theta{}^{-1} B_\theta{}'$ is the corresponding asymptotic covariance matrix of $n^{\frac{1}{2}}[\beta(\hat{\theta}) - \beta(\theta)]$ (see Rao [6] page 322). The noncentrality parameter stated at the end of the theorem can be easily seen to be the limit as $n \to \infty$ of $n[\beta(\theta_n)]'(B_{\theta_n} C_{\theta_n}^{-1} B'_{\theta_n})^{-1}\beta(\theta_n)$.

In virtually all exponential families used in practice the natural parameter space is open. In this case Θ_N, $\Theta_N{}^0$ and the space Θ_E defined below coincide.

PROOF. The first step is to put the problem in the context of the moment parametrization, after which the conclusion of the theorem will follow easily.

Let $\Theta_E = \{\theta \in \Theta_N : E_\theta|t(x)| < \infty\}$. For $\theta \in \Theta_E$, denote $\tau = E_\theta[t(x)]$. Then $\tau^i = -(\partial/\partial\theta^i)\log c(\theta)$, $i = 1, \cdots, p$. Note that because of Assumption 1 $t(x)$ is not restricted to a proper affine subspace of $\mathscr{E}^p$. By Berk ([1], Lemma 2.2) this implies that the mapping taking θ into τ ($\Theta_E \to \tau(\Theta_E)$) is one-one. By Lehmann ([5] page 52, Theorem 9), $\Theta_N{}^0 \subset \Theta_E$ (and in fact τ is infinitely differentiable on $\Theta_N{}^0$). The restricted mapping taking θ into τ ($\Theta_N{}^0 \to \tau(\Theta_N{}^0)$) is also one-one. Dempster ([2] page 331) sketches an alternative argument for the one-one property of $\theta \to \tau$ based on Assumption 1, pointing out that the positive definite matrix C_θ is the matrix of partial derivatives $\partial\tau^i/\partial\theta^j$ (see Devinatz [3] page 344, exercise 10). C_θ is also $(-1/n)$ times the Hessian of the log likelihood function; the strict convexity of the log likelihood thus implied by Assumption 1 assures the uniqueness of $\hat{\theta}_n$ and that this value of θ produces a global maximum.

Since the transformation $\psi : \tau \to \theta$ is one-one on $\tau(\Theta_N{}^0)$, we reparametrize the problem in terms of τ, and write

$$W_n = n[\gamma(\hat{\tau}_n)]'(G_{\hat{\tau}_n} C_{\hat{\theta}_n} G'_{\hat{\tau}_n})^{-1}\gamma(\hat{\tau}_n) ,$$

where $\hat{\tau}_n = n^{-1}\sum_{\alpha=1}^n t(x_{n\alpha})$, $\gamma = \beta \circ \psi$, $G_\tau = B_\theta C_\theta{}^{-1} =$ the matrix of partial derivatives $\partial\gamma^r/\partial\tau^s$, and the Fisher information matrix C_θ is also the covariance matrix of $t(x_{n\alpha})$ and of $n^{\frac{1}{2}}\hat{\tau}_n$.

Let τ_0 be the value of τ when $\theta = \theta_0$. It is straightforward to verify, by expanding τ as a function of θ in Taylor series about θ_0, that the assumption $n^{\frac{1}{2}}(\theta_n - \theta_0) \to \delta$ implies $n^{\frac{1}{2}}(\tau_n - \tau_0) \to \eta$, where $\eta = C_{\theta_0}\delta$. Rewriting the noncentrality parameter stated in the theorem as $\eta' G'_{\tau_0}(G_{\tau_0} C_{\theta_0} G'_{\tau_0})^{-1} G_{\tau_0}\eta$, the conclusion of the theorem now follows from the previously mentioned theorem of the author ([7] page 1415), upon demonstration that $\mathscr{L}[n^{\frac{1}{2}}(\hat{\tau}_n - \tau_n)] \to \mathscr{N}(0, C_{\theta_0})$ and plim $C_{\hat{\theta}_n} = C_{\theta_0}$, since by that theorem these two conditions plus the assumed positive-definitenss of C_{θ_0} are sufficient to ensure the stated convergence to noncentral chi-square.

From the above Lemma, with $w_{n\alpha} = t(x_{n\alpha})$, $t_n = \hat{\tau}_n$ and $\Sigma_n = C_{\theta_n}$, it follows that $\mathscr{L}[n^{\frac{1}{2}}(\hat{\tau}_n - \tau_n)] \to \mathscr{N}(0, C_{\theta_0})$ and plim $\hat{\tau}_n = \tau_0$; plim $C_{\hat{\theta}_n} = C_{\theta_0}$ follows from the latter result since C_θ is a continuous function of τ. (The fourth moments of $t(x_{n\alpha})$ are polynomials in the first four orders of partial derivatives of $\log c(\theta_n)$ and hence bounded as $\tau_n \to \tau_0$; this is a straightforward extension of [5], page 58, problem 14.) □

In the special case $\tau_n \equiv \tau_0$, the above argument contains an elementary proof of the (weak) consistency and asymptotic normality of the MLE for exponential families. Berk [1] has exploited the convexity of $-\log c(\theta)$ to obtain strong consistency and asymptotic normality of the MLE for exponential models in the more general situation where the true distribution of the observations need not belong to the exponential family.

Acknowledgment. The author is grateful for the advice and comments of Professor Robert Berk.

REFERENCES

[1] BERK, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43** 193–204.

[2] DEMPSTER, A. P. (1971). An overview of multivariate data analysis. *J. Multivariate Anal.* **1** 316–346.

[3] DEVINATZ, A. (1968). *Advanced Calculus.* Holt, Rinehart and Winston, New York.

[4] LECAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symposium Math. Statist. Prob.* **1** 129–156. Univ. of California Press.

[5] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[6] RAO, C. R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, New York.

[7] STROUD, T. W. F. (1971). On obtaining large-sample tests from asymptotically normal estimators. *Ann. Math. Statist.* **42** 1412–1424.

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

DEPARTMENT OF MATHEMATICS
QUEEN'S UNIVERSITY
KINGSTON, ONTARIO, CANADA