The Annals of Mathematical Statistics
1968, Vol. 39, No. 5, 1402–1424

LARGE DEVIATIONS THEORY IN EXPONENTIAL FAMILIES[1]

BY BRADLEY EFRON AND DONALD TRUAX[2]

Stanford University and the University of Oregon

0. Summary. We consider repeated independent sampling from one member of an exponential family of probability distributions. The probability that the sample mean of n such observations falls into some set S_n is, by definition, a "large deviations", "small deviations", or "medium deviations" problem depending on the location of the set S_n relative to the expectation of the distribution.

We present a theorem which allows the accurate approximation of all such probabilities under a wide variety of circumstances. These approximations are shown to yield simple and numerically accurate expressions for the small sample power functions of hypothesis tests in the exponential family. Various large sample properties of exponential families are presented, many of which are seen to be extensions and refinements of familiar large deviations results.

The method employed is to replace the given exponential family by a suitably modified normal translation family, which is shown to approximate the original family uniformly well over any bounded subset of the parameter space. The simple and tractable nature of normal translation families then provides our results.

1. Introduction and an outline of the paper. Exponential families of probability distributions play a dominant role in parametric statistical analysis, embracing almost all of the common univariate and multivariate distributions. In this paper we represent a d parameter, or d dimensional, exponential family by

$$P_\theta(A) = \int_A e^{\theta' x - \psi(\theta)}\, d\mu(x)$$

for Borel sets A in Euclidean d-space E^d, or more simply by

$$dP_\theta(x) = e^{\theta' x - \psi(\theta)}\, d\mu(x)$$

for $x \varepsilon E^d$. Here $\mu(x)$ is a probability distribution on E^d, and θ takes values in Θ, that subset of E^d for which $\int_{E^d} e^{\theta' x}\, d\mu(x)$ is finite. The function $\psi(\theta)$ is the log moment generating function of $\mu(x)$, and yields the moments of P_θ by differentiation, for θ in the interior of Θ. Letting $\lambda(\theta)$ and $\Sigma(\theta)$ be the mean vector and covariance matrix of an observation X from P_θ,

$$\lambda(\theta) \equiv E_\theta X, \qquad \Sigma(\theta) = \mathrm{Cov}_\theta\, X,$$

Received 11 September 1967.

[1] Supported by Public Health Service Grant USPHS-GM-14554-01. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] The research of this author was sponsored by NSF Grant 4865.

we have

$$\lambda(\theta) = (\partial\psi(\theta)/\partial\theta^{(j)}) \qquad \text{and} \qquad \Sigma(\theta) = (\partial^2\psi(\theta)/\partial\theta^{(j)}\partial\theta^{(k)}),$$

superscripts indicating components of the θ vector.

It is well-known [13], Section 2.7, that Θ is a convex set in E^d, and we will assume that it contains an open set in that space. (If not, it is always possible to reparameterize to a lower dimensional exponential family where the condition does hold). Likewise, there is no loss of generality in assuming that μ does not put all of its probability mass on any $d - 1$ dimensional hyperplane in E^d, which is equivalent to assuming that $\Sigma(\theta)$ is non-singular for all $\theta \varepsilon \Theta$. We will occasionally perform linear transformations on θ and X in order to simplify our calculations. For example if θ_0 is a point of particular interest to us, we may write

$$dP_\theta(x) = e^{(\theta-\theta_0)'x-(\psi(\theta)-\psi(\theta_0))}\, dP_{\theta_0}(x).$$

Letting $\tilde{\theta} = \theta - \theta_0$, we see that we can take our special point θ_0 to be the origin without loss of generality, in which case P_{θ_0} becomes the distribution μ. Section 2, in particular, is written from this point of view, and then regeneralized in note 3. If we transform to $\tilde{\theta} = [\Sigma(\theta_0)]^{\frac{1}{2}}(\theta - \theta_0)$ and $\tilde{X} = [\Sigma(\theta_0)]^{-\frac{1}{2}}(X - \lambda(\theta_0))$ our exponential family has mean zero and covariance the identity matrix at the origin (assuming $\lambda(\theta_0)$ and $\Sigma(\theta_0)$ exist).

If $X_1, X_2, \cdots, X_n$ are independent observations from P_{θ_0}, the normalized vector $X = (\sum_1^n X_i - n\lambda(\theta_0))/n^{\frac{1}{2}}$ will fall into a set C in E^d with probability $P_{\theta_0,n}(C)$, say, which for large n approaches the probability of C under the normal distribution $\mathfrak{N}(0, \Sigma(\theta_0))$ (assuming that the boundary of C has measure zero). If the X_i are chosen from P_θ, $\theta \neq \theta_0$, then X will have mean $n^{\frac{1}{2}}(\lambda(\theta) - \lambda(\theta_0))$, and $P_{\theta,n}(C)$ will approach zero at an exponential rate. This fact, of course, underlies most tests of the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, and it is known [4], [14), that a complete class of such tests is those based on *convex* acceptance regions C.

The main result presented in this paper is a theorem comparing the probabilities $P_{\theta,n}(C)$—equivalently the power function of the test for H_0 based on C—for different exponential families. Let us say that $\tilde{P}_\theta$ *agrees* with P_θ at a point θ_0 interior to both parameter spaces Θ and $\tilde{\Theta}$ if $\lambda(\theta_0) = \tilde{\lambda}(\theta_0)$ and $\Sigma(\theta) = \tilde{\Sigma}(\theta_0)$. Then our comparison theorem can be stated in the following way (ignoring regularity conditions): If P_θ and $\tilde{P}_\theta$ agree at θ_0, then

$$P_{\theta,n}(C)/\tilde{P}_{\theta,n}(C) = e^{-n\{[\psi(\theta)-\psi(\theta_0)]-[\tilde{\psi}(\theta)-\tilde{\psi}(\theta_0)]\}}[1 + o_n(1)],$$

with the term $o_n(1)$ approaching zero uniformly for θ in any bounded subset of $\Theta \cap \tilde{\Theta}$, and C convex and contained within any bounded subset of E^d. In particular, if we take as our agreeing family the normal translation family $\tilde{P}_\theta \sim \mathfrak{N}(\lambda(\theta_0) + \Sigma(\theta_0)[\theta - \theta_0], \Sigma(\theta_0))$ (the form of the mean being necessary to write $d\tilde{P}_\theta$ in the form $e^{\theta'x-\tilde{\psi}(\theta)}\, d\tilde{\mu}(x)$), the bracketed term in the exponential becomes

$I - \frac{1}{2}J$, where I is the Kullback-Liebler information $E_{\theta_0} \log (dP_{\theta_0}/dP_\theta)(X)$ and $J = \text{Variance}_{\theta_0} (\log (dP_{\theta_0}/dP_\theta)(X))$.

This theorem is derived under various regularity conditions in Sections 2 and 3, and stated in a number of useful equivalent forms. In Section 4 it is employed to discuss two "large deviation" theorems. A sharpened version of the Chernoff bound on the Bayes risk of the likelihood ratio test is given, and a theorem of Hoeffding for large deviations of multinomial observations is generalized to exponential families.

Section 5 is concerned with asymptotic properties of the fixed level α likelihood ratio test of H_0 versus H_1. An optimality property for this test is given which is the large deviations equivalent of a familiar result of Wald, along with a derivation of the large deviations power function.

The comparison theorem is a useful computational tool for investigating the power function of likelihood ratio tests for both simple and composite hypotheses in our exponential family. Section 6 discusses numerical approximation of power functions in some detail, and offers several examples, including that of testing for independence in a multivariate normal distribution.

We have chosen to present these results in the context of exponential families for reasons of statistical relevance. An alternative presentation could be made in terms of large deviations theory, which is to say the probability that an average $n^{-1} \sum_1^n X_i$ of independent, identically distributed random variables falls a large distance away from its expectation. In a paper quite closely related to this one, [5], Borovkov and Rogozin show clearly the relation between these two points of view. Other papers of direct relevance are Bahadur and Rao [2], Efron [8], and Hoeffding [11], [12].

2. The main results. The technical basis of our results is the local central limit theorems of Rvačeva [16] and Stone [18], which we combine below as Lemma 1. These theorems yield a useful approximation (Theorem 1) for convolutions of the distribution μ, which is in turn the basis of our comparison theory, Theorems 2 (and the following notes), 3, and 5.

To increase the generality of our results, we drop the assumption that μ has a finite mean and variance,[3] and assume instead that μ belongs to the domain of attraction of some non-degenerate stable distribution ν in E^d. That is, we assume the existence of a sequence of vectors A_n and positive constants B_n such that if $X_1, X_2, \cdots, X_n$ are independent random vectors from μ, the distribution μ_n of the normalized sum $X = (\sum_1^n X_i - A_n)/B_n$ satisfies

$$\lim_{n\to\infty} \mu_n(D) = \lim_{n\to\infty} \text{Prob}_\mu (X \varepsilon D) = \nu(D) \tag{1}$$

for every ν-continuity set D. (For convenience we will assume $B_n \geqq 1$. This involves no loss of generality.)

NOTE. Throughout this paper we use X and the corresponding realized value x to represent both the normalized random variable above, and a single prototype

[3] In the language of the introduction, this allows θ_0 to be on the boundary of Θ.

observation $X \sim P_\theta$ from our exponential family, with the correct interpretation being clear from the context.

Following Stone [18], we say μ is *non-lattice* if $|f(\theta| < 1$ for all $\theta \neq 0$, where $f(\theta) = \int e^{i\theta' x}\, d\mu(x)$. If $\limsup_{\|\theta\|\to\infty} |f(\theta)| < 1$, μ is called strongly non-lattice. If the lattice generated by all differences $x - y$ of points of E^d having positive probability under the discrete distribution μ is a subset of the lattice of all points of E^d having integer coordinates, we say μ is a 1-*lattice* distribution. Finally, let $K(x, h)$ be the cube $\{y \in E^d : x_i \leqq y_i < x_i + h, i = 1, 2, \cdots, d\}$.

LEMMA 1. (Rvačeva and Stone). (i) *If μ is non-lattice, then*

$$\mu_n(K(x, h)) = \nu(K(x, h)) + \theta_n(x, h)(h^d + B_n^{-d}),$$

where $\lim_{n\to\infty} \theta_n(x, h) = 0$ *uniformly in x and h;*

(ii) *If μ is strongly non-lattice and $c < c_1$, where c_1 is defined by $e^{-c_1 d} = \limsup_{\|\theta\|\to\infty} |f(\theta)|$, then*

$$\mu_n(K(x, h)) = \nu(K(x, h)) + \theta_n(x, h)(h^d + e^{-nc}),$$

where $\lim_{n\to\infty} \theta_n(x, h) = 0$ *uniformly in x and h;*

(iii) *If μ is a 1-lattice distribution*

$$\mu_n(K(x, B_n^{-1})) = \nu(K(x, B_n^{-1})) + \theta_n(x) B_n^{-\alpha},$$

for $B_n x + A_n + (\frac{1}{2}, \frac{1}{2}, \cdots, \frac{1}{2})'$ in the lattice, where $\lim_{n\to\infty} \theta_n(x) = 0$ uniformly in such x.

Next, we will need a result on convex sets in E^d. In Eggleston [9] one may find an exposition on the mixed volumes of a linear array of convex sets. In the form that we will need, the result says that if C_1 and C_2 are compact convex sets in E^d and C is the linear combination $C = \lambda_1 C_1 + \lambda_2 C_2$, where $\lambda_1 \geqq 0$, $\lambda_2 \geqq 0$, then the volume of C, $V(C)$, is a polynomial of degree d in λ_1, λ_2 of the form

$$V(C) = \sum_{s=0}^{d} \binom{d}{s} V(C_1, s; C_2, d - s) \lambda_1^s \lambda_2^{d-s},$$

where the coefficients $V(C_1, s; C_2, d - s)$ are called the mixed volumes. It is also shown in [9] that if C_1 and C_2 are both contained in a convex set D, then $V(C_1, s; C_2, d - s) \leqq 2^d V(D)$ for $s = 0, 1, \cdots, d$.

In the following, the diameter of a set K in E^d will mean $\sup\{\|x - y\| : x, y \in K\}$, and the distance from a point x to K will be $d(x, K) = \inf_{y\in K} \|x - y\|$. Let C be a compact convex set in E^d which contains a sphere of radius $\epsilon > 0$, and consider a grid of cubes of diameter $t < \epsilon$ covering C. (To be a covering the cubes are non-overlapping, the union contains C, and each cube intersects C.) Denote by Q the number of cubes in the covering, and by q the number of cubes in the covering which meet the boundary of C.

LEMMA 2. $$q/Q \leqq 2^{2d+1} t/\epsilon.$$

PROOF. If a cube intersects the boundary of C, then that cube must lie in a neighborhood of radius t of the boundary of C. Thus, q/Q cannot exceed the ratio of the volume of this neighborhood to the volume of C, which in turn

implies

$$q/Q \leqq 2(V(N(C, t)) - V(C))/V(C),$$

where the t-neighborhood of C is $N(C, t) = \{y \varepsilon E^d : d(y, C) \leqq t\}$.

Now, the right hand side is unchanged by a translation of C, so we may as well assume that the sphere $S \subset C$ is centered at 0. Then, we can write

$$N(C, t) = C + (t/\epsilon)S.$$

By the formula for mixed volumes we have

$$V(N(C, t)) = V(C + t\epsilon^{-1}S) = \sum_{s=0}^{d} \binom{d}{s} V(C, d - s; S, s)(t\epsilon^{-1})^s.$$

Also, since $S \subset C$, $V(C, d - s; S, s) \leqq 2^d V(c)$. Thus, taking note of the fact that $V(C, d; S, 0) = V(C)$,

$$\begin{aligned}(V(N(C, t)) - V(C))/V(C) &= \sum_{s=1}^{d} \binom{d}{s} V(C, d - s; S, s)(V(C))^{-1}(t\epsilon^{-1})^s \\ &\leqq 2^d \sum_{s=1}^{d} \binom{d}{s} (t\epsilon^{-1})^s \leqq 2^d(t\epsilon^{-1}) \sum_{s=0}^{d} \binom{d}{s} = 2^{2d}(t\epsilon^{-1}).\end{aligned}$$

Hence, $q/Q \leqq 2^{2d+1}(t\epsilon^{-1})$.

We are now in a position to state and prove

THEOREM 1. (i) *If μ is non-lattice and C is a compact convex set containing a sphere of radius $\rho 2^{2d+1} d^{\frac{1}{2}}/B_n$, then for all $0 < \delta < \min[\rho, 1]$,*

$$|\mu_n(C)/\nu(C) - 1| \leqq (M\delta\rho^{-1} + 2o_n(1)/\delta^d)/m(1 - \delta\rho^{-1}) \tag{2}$$

where M denotes the maximum of the density $p(x)$ corresponding to ν over a closed neighborhood of C of radius $d^{\frac{1}{2}}\delta/B_n$, and m the minimum over C,

$$m = m(C) \equiv \inf_{x \varepsilon C} p(x).$$

The term $o_n(1) = \sup_{x,h} |\theta_n(x, h)|$, where $\theta_n(x, h)$ is as in Lemma 1(i). Note that $o_n(1)$ is independent of C and δ. (One should also note that every non-degenerate stable distribution has a bounded density $p(x)$ with respect to Lebesgue measure, so that M can be uniformly bounded by $\sup_{E^d} p(x)$.)

(ii) *If μ is strongly non-lattice and C contains a sphere of radius $\rho 2^{2d+1} d^{\frac{1}{2}}/e^{nc/d}$, where $c < c_1$ (see the definition of strongly nonlattice for a definition of c_1), then (2) holds with $o_n(1) = \sup_{x,h} |\theta_n(x, h)|$, where $\theta_n(x, h)$ is as in Lemma 1(ii), and M can be taken as the maximum of the density of ν over a closed neighborhood of C of radius $d^{\frac{1}{2}}\delta/e^{nc/d}$.*

(iii) *If μ is a 1-lattice distribution, and C contains a sphere of radius $\rho\, d^{\frac{1}{2}} 2^{2d+1}/B_n$ where $\rho > 1$, then*

$$|\mu_n(C)/\nu(C) - 1| \leqq (M + \rho o_n(1))/m(\rho - 1).$$

Here $o_n(1) = \sup_x |\theta_n(x)|$, where $\theta_n(x)$ is as given in Lemma 1(iii), and the supremum is taken over only those x such that $B_n x + A_n + (\frac{1}{2}, \frac{1}{2}, \cdots, \frac{1}{2})'$ is in the lattice. M and m are as in (i) with $\delta = 1$.

PROOF. First suppose μ is non-lattice. Let $K_1, K_2, \cdots$, denote cubes of side δ/B_n in a covering of C. Then

$$(\sum - \sum^*)\mu_n(K_i) \leqq \mu_n(C) \leqq \sum \mu_n(K_i),$$

where $\sum$ and $\sum^*$ denote respectively the sum over all the cubes in the covering, and the sum over only those cubes in the covering which intersect the boundary of C. According to Lemma 1

$$(\sum - \sum^*)\nu(K_i) - Q[(\delta/B_n)^d + (1/B_n)^d]o_n(1)$$
$$\leqq \mu_n(C) \leqq \sum \nu(K_i) + Q[(\delta/B_n)^d + (1/B_n)^d]o_n(1)$$

where Q denotes the number of terms in the sum $\sum$. This implies

$$\nu(C) - \sum^* \nu(K_i) - Qo_n'(1)/B_n^d \leqq \mu_n(C) \leqq \nu(C) + \sum^* \nu(K_i) + Qo_n'(1)/B_n^d,$$

where $o_n'(1) = (1 + \delta^d)o_n(1) \leqq 2o_n(1)$. Thus,

$$|\mu_n(C)/\nu(C) - 1| \leqq \sum^* \nu(K_i)/\nu(C) + (2Qo_n(1)/B_n^d)/\nu(C).$$

Now, $\sum^* \nu(K_i) \leqq Mq(\delta/B_n)^d$, and

$$\nu(C) \geqq (\sum - \sum^*)\nu(K_i) \geqq m(Q - q)(\delta/B_n)^d,$$

so,

$$|\mu_n(C)/\nu(C) - 1| \leqq (Mq + 2Qo_n(1)/\delta^d)/m(Q - q).$$

In Lemma 2, set $t = d^{\frac{1}{2}}\delta/B_n$, $\epsilon = \rho d^{\frac{1}{2}}2^{2d+1}/B_n$, so $q/Q \leqq \delta/\rho$, and we have

$$|\mu_n(C)/\nu(C) - 1| \leqq (M\delta/p + 2o_n(1)/\delta^d)/m(1 - \delta/\rho).$$

In the strongly non-lattice case everything is as above except that the cubes in the covering are taken with side $\delta e^{-cn/d}$.

In the case μ has a 1-lattice distribution, we restrict ourselves to cubes of side $1/B_n$ centered at points of the lattice. If we insist that C contain a sphere of radius $\rho d^{\frac{1}{2}}2^{2d+1}/B_n$, then again we have

$$|\mu_n(C)/\nu(C) - 1| \leqq (Mq + 2Qo_n(1))/m(Q - q) \leqq (M + 2\rho o_n(1))/m(\rho - 1).$$

COROLLARY. *If μ is non-lattice (strongly non-lattice), and $\Lambda_n(K)$ denotes the class of convex sets C which*

(i) *are contained within a fixed set K not depending on n and*

(ii) *contain a sphere of radius B_n^{-1} $(e^{-cn/d})$ then*

$$\lim_{n\to\infty} \sup_{C\in\Lambda_n(K)} |\mu_n(C)/\nu(C) - 1| = 0$$

provided that the density function $p(x)$ of ν has a positive infinum over K, $m(K) \equiv \inf_{x\in K} p(x) > 0$.

PROOF. The corollary follows immediately from Theorem 1 by letting δ go to zero with n sufficiently slowly in (2), say $\delta_n = [o_n(1)]^{1/(d+1)}$.

Using $m(C) \geqq m(K)$ for C contained in K, we then have

$$|\mu_n(C)/\nu(C) - 1| \leqq \frac{(M+2)[o_n(1)]^{1/(d+1)}}{m(K)(1-[o_n(1)]^{1/(d+1)})}$$

approaching zero at a rate independent of C in K.

The set K may be allowed to expand with n instead of remaining fixed, provided $m(K_n)$ does not converge to zero too rapidly. With δ_n chosen as above, we could take any sequence of sets K_n with $m(K_n) \geqq [o_n(1)]^{1/(d+2)}$ and still obtain $\lim_{n\to\infty} \sup_{C\varepsilon\Lambda_n} |\mu_n(C)/\nu(C) - 1| = 0$.

The corollary will also hold for μ a l-lattice distribution provided that we strengthen condition (ii) to state that each set C in $\Lambda_n(K)$ must contain a sphere of radius ρ_n/B_n, where ρ_n increases to infinity.

Another point worth noting is that if one were to impose stronger conditions on μ, for example that μ has a characteristic function whose pth power is integrable for some $p \geqq 1$, and second moments exist, then one could conclude that the density of the normalized sum converges uniformly to the normal density. In this case the local limit theorem could be made much stronger. The convexity condition is not needed, and the sets C can be taken as small as we please, i.e., we need not insist that they contain spheres of a given size. For an interesting discussion of local limit theorems under these more restrictive conditions the reader is referred to [5].

As in Section 1, we will let the probability measure μ generate an exponential family of distributions $dP_\theta(x) = e^{\theta' x - \psi(\theta)}\, d\mu(x)$. It is easily seen that the entire family P_θ is non-lattice (1-lattice) if μ is non-lattice (1-lattice). Suppose that μ belongs to the domain of attraction of ν, a stable distribution with density $p(x)$, and that A_n and B_n are chosen to satisfy (1). Denote by $P_{\theta,n}$ the distribution of $(\sum_1^n X_i - A_n)/B_n)$ when the X_i have distribution P_θ.

Define the "support function" $h(\theta)$ of a convex set C by

$$h(\theta) = \sup_{x\varepsilon C} \theta' x/\|\theta\|,$$

and for a given $\theta \varepsilon \Theta$, $\theta \neq 0$, and $y > 0$ let

$$C_y = \{x\varepsilon C: \theta' x/\|\theta\| \geqq h(\theta) - y/\|\theta\|\}.$$

THEOREM 2. *For a given convex set C, suppose that C_{1/B_n} contains a sphere of radius k/B_n, $0 < k < 1$. Then in the case where μ is a non-lattice distribution,*

$$P_{\theta,n}(C) = \exp(-n\psi(\theta) + \theta' A_n + B_n\|\theta\|h(\theta)) \cdot \int_0^\infty \nu(C_y) B_n e^{-B_n y}\, dy[1 + o_n^*(1)/mk^d],$$

where $o_n^*(1)$ tends to zero uniformly in θ, C, and k. (As before, $m = \inf_{x\varepsilon C} p(x)$, where $p(x)$ is the density of ν. The asterisk on $o_n^*(1)$ is intended to distinguish the symbol from the specific quantity in Lemma 1.)

PROOF. We have

$$\prod_{i=1}^n dP_\theta(x_i)/\prod_{i=1}^n d\mu(x_i) = e^{-n\psi(\theta)+\theta'(\sum n/i=1)(x_i)}$$

so that

$$dP_{\theta,n}(x)/d\mu_n(x) = \exp(-n\psi(\theta) + \theta' A_n + B_n\theta' x)$$

and

$$P_{\theta,n}(C) = \exp(-n\psi(\theta) + \theta' A_n + B_n\|\theta\|h(\theta)) \cdot \int_C \exp(B_n(\theta' x - h(\theta)\|\theta\|))\, d\mu_n(x).$$

Making the substitution $y = \|\theta\|h(\theta) - \theta' x$, and using integration by parts it is easy to see that the integral may also be expressed as the Laplace transform $\int_0^\infty \mu_n(C_y)B_n e^{-B_n y}\, dy$. We must therefore show that

$$\int_0^\infty \mu_n(C_y)B_n e^{-B_n y}\, dy / \int_0^\infty \nu(C_y)B_n e^{-B_n y}\, dy = 1 + o_n^*(1)/mk^d.$$

It is here that the local limit theorem plays its role, since the integrating density $B_n e^{-B_n y}$ approaches a delta function at the origin as B_n tends toward infinity.

Define $\epsilon_n = [o_n(1)]^{1/(2d+2)}$, where $o_n(1)$ is the sequence appearing in part (i) of Theorem 1. [We assume, without loss of generality, that $o_n(1) < 1$.] Then by convexity and the hypothesis of the theorem, for $y \geqq \epsilon_n/B_n$ the set C_y contains a sphere of radius $k\epsilon_n/B_n$. Letting $\delta = k2^{-(2d+1)}d^{-\frac{1}{2}}\epsilon_n^2$ in part (i) of Theorem 1 yields

$$|\mu_n(C_y)/\nu(C_y) - 1| \leqq M'(mk^d)^{-1}\epsilon_n(1 - \epsilon_n)^{-1}$$

for $y \geqq \epsilon_n/B_n$, where $M' = M + 2^{2d^2+d+1}d^{d/2}$, M being the supremum of the bounded density $p(x)$ over all of E^d.

From this inequality we infer that

$$\left|\int_{\epsilon_n/B_n}^\infty \mu_n(C_y)B_n e^{-B_n y}\, dy / \int_{\epsilon_n/B_n}^\infty \nu(C_y)B_n e^{-B_n y}\, dy - 1\right| \leqq o_n'(1)/mk^d,$$

where $o_n'(1) = M'\epsilon_n(1 - \epsilon_n)^{-1}$. In addition, note that

$$\int_0^{\epsilon_n/B_n} \mu_n(C_y)B_n e^{-B_n y}\, dy / \int_{\epsilon_n/B_n}^\infty \nu(C_y)B_n e^{-B_n y}\, dy$$
$$\leqq \mu_n(C_{\epsilon_n/B_n})(\nu(C_{\epsilon_n/B_n}))^{-1}(1 - e^{-\epsilon_n})(e^{-\epsilon_n})^{-1}$$
$$\leqq [1 + o_n'(1)/mk^d](1 - e^{-\epsilon_n})(e^{-\epsilon_n})^{-1}$$

and

$$\int_0^{\epsilon_n/B_n} \nu(C_y)B_n e^{-B_n y}\, dy / \int_{\epsilon_n/B_n}^\infty \nu(C_y)B_n e^{-B_n y}\, dy \leqq (1 - e^{-\epsilon_n})(e^{-\epsilon_n})^{-1}.$$

The elementary inequality

$$|(a_1 + a_2)/(b_1 + b_2) - 1| \leqq |a_1/b_1 - 1| + |a_1/b_1||b_2/b_1| + |a_2/b_1|$$

yields (with $a_1 = \int_{\epsilon_n/B_n}^\infty \mu_n(C_y)B_n e^{-B_n y}\, dy$, $a_2 = \int_0^{\epsilon_n/B_n} \mu_n(C_y)B_n e^{-B_n y}\, dy$, etc.)

$$\left|\int_0^\infty \mu_n(C_y)B_n e^{-B_n y}\, dy / \int_0^\infty \nu(C_y)B_n e^{-B_n y}\, dy - 1\right|$$
$$\leqq o_n'(1)/mk^d + 2[1 + o_n'(1)/mk^d](1 - e^{-\epsilon_n})/e^{-\epsilon_n}.$$

This completes the proof, and shows that we can take $o_n^*(1)$ to be proportional to $\epsilon_n = [o_n(1)]^{1/(2d+2)}$.

NOTE 1. The class $\Lambda(K, \tau)$ of convex sets C which (i) each contain a sphere of fixed radius $\tau > 0$, and (ii) are contained in a fixed bounded set K with $m(K) \equiv \inf_{x \varepsilon K} p(x) > 0$, all satisfy the hypothesis of Theorem 2 with $k \geqq k_0$, k_0 being a positive constant depending only on τ, $\|\theta\|$, and the diameter of the set K. The approximation factor can therefore be expressed as $o_n^*(1)/m(K)k_0^d$, which approaches zero uniformly fast within the class $\Lambda(K, \tau)$, and uniformly fast for all θ within any fixed bounded subset of Θ.

NOTE 2. $\int_0^\infty \nu(C_y) B_n e^{-B_n y}\, dy \geqq m \int_0^{1/B_n} c_d (ky)^d B_n e^{-B_n y}\, dy$ where $c_d r^d$ is the volume of a d-dimensional sphere of radius r. From this we conclude that there exists a positive constant c' such that $\int_0^\infty \nu(C_y) B_n e^{-B_n y}\, dy \geqq mc'(k/B_n)^d$. Letting $\phi_n = (d + 1) \log B_n$, we observe that

$$\int_{\phi_n/B_n}^\infty \mu_n(C_y) B_n e^{-B_n y}\, dy \leqq e^{-\phi_n} = (1/B_n)^{d+1}$$

for any set C_y. It follows that the conditions of Theorem 2 can be relaxed: it is sufficient that C_{ϕ_n/B_n} be convex for $\phi_n = (d + 1) \log B_n$, and that C_{1/B_n} contain a sphere of radius k/B_n. Then the conclusion of the theorem holds with $m = \inf_{x \varepsilon C_{\phi_n/B_n}} p(x)$, the new approximation term being proportional to $(mk^d)^{-1}(o_n^*(1) + 1/B_n)$.

NOTE 3. For an arbitrary point $\theta_0 \varepsilon \Theta$ we have

$$dP_\theta(x) = \exp\left((\theta - \theta_0)'x - (\psi(\theta) - \psi(\theta_0))\right) dP_{\theta_0}(x).$$

Suppose that for constants A_n and B_n, the distribution of $(\sum_1^n X_i - A_n)/B_n$ approaches the stable law ν, for independent observations X_i from P_{θ_0}. Then the statement of Theorem 2 becomes

$$P_{\theta,n}(C) = \exp\left(-n(\psi(\theta) - \psi(\theta_0)) + (\theta - \theta_0)'A_n + B_n\|\theta - \theta_0\| h(\theta - \theta_0)\right) \cdot \int_0^\infty \nu(C_y) B_n e^{-B_n y}\, dy [1 + o_n^*(1)/mk^d].$$

C_y now being defined by $C_y = \{x \varepsilon C,\ (\theta - \theta_0)'x/\|\theta - \theta_0\| \geqq h(\theta - \theta_0) - y/\|\theta - \theta_0\|\}$.

If we suppose that θ_0 is in the interior of Θ, then all the moments of the distribution P_{θ_0} exist, and ν must be a normal law. Using the notation

$$E_\theta X \equiv \lambda(\theta)$$

for the expectation of a random variable X from P_θ, we take $A_n = n\lambda(\theta_0)$ and $B_n = n^{\frac{1}{2}}$. Noting that the quantity $(\psi(\theta) - \psi(\theta_0)) - (\theta - \theta_0)'\lambda(\theta_0)$ is the Kullback-Leibler information number between P_{θ_0} and P_θ,

$$I(\theta_0, \theta) = E_{\theta_0} \log dP_{\theta_0}(X)/dP_\theta(X),$$

the expression of Theorem 2 becomes, in this case,

$$P_{\theta,n}(C) = \exp\left(-nI(\theta_0, \theta) + n^{\frac{1}{2}}\|\theta - \theta_0\| h(\theta - \theta_0)\right) \cdot \int_0^\infty \nu(C_y) B_n e^{-B_n y}\, dy [1 + o^*(1)/mk^d].$$

NOTE 4. In the strongly non-lattice case it is sufficient to demand that $C_{\epsilon-nc}$ contain a sphere of radius ke^{-nc}.

In the 1-lattice case a less precise result is obtained as follows:

THEOREM 3. *For a given convex set C, suppose there exists $\phi > 0$ and $\rho > Mm^{-1} + 1$ such that C_{ϕ/B_n} contains a sphere of radius $\rho d^{\frac{1}{2}}2^{2d+1}/B_n$. Then in the case where μ is a 1-lattice distribution,*

$$P_{\theta,n}(C) = \exp(-n\psi(\theta) + \theta' A_n + B_n\|\theta\|h(\theta))\int_0^\infty \nu(C_y)B_n e^{-B_n y}\,dy[e^{o_n(1)}]$$

where $e^{o_n(1)}$ is a factor bounded away from zero and infinity by

$$[1 - (M + \rho o_n(1))/m(\rho - 1)]e^{-\phi} \leqq e^{o_n(1)} \leqq [1 + (M + \rho o_n(1))/m(\rho - 1)]e^{\phi},$$

$o_n(1)$ being the infinitesimal sequence of part (iii), *Theorem* 1.

PROOF. Define

$$a_1 = \int_{\phi/B_n}^\infty \mu_n(C_y)B_n e^{-B_n y}\,dy, \qquad a_2 = \int_0^{\phi/B_n} \mu_n(C_y)B_n e^{-B_n y}\,dy,$$

and b_1 and b_2 the corresponding integrals with $\nu(C_y)$ replacing $\mu_n(C_y)$. Theorem 1, part (iii) yields

$$|a_1/b_1 - 1| \leqq (M + \rho o_n(1))/m(\rho - 1),$$

while from elementary calculations we have

$$a_2/a_1 \leqq (1 - e^{-\phi})/e^{-\phi} \quad \text{and} \quad b_2/b_1 \leqq (1 - e^{-\phi})/e^{-\phi}.$$

The result follows as in the proof of Theorem 2.

Theorems 2 and 3 allows us to approximate $P_{\theta,n}(C)$, which depends on μ_n, with an integral involving only the limiting distribution ν. Since our knowledge of ν is more precise than our knowledge of the distributions μ_n, a large class of accurate asymptotic expressions can be inferred from these theorems. A crude but useful and suggestive example is the following:

THEOREM 4. *Let $\Lambda(K, \tau)$ be the class of sets defined in Note* 1 *above, and let Θ' be any bounded subset of Θ. Then there exists positive constants a and b such that for all $C \varepsilon \Lambda(K, \tau)$ and $\theta \varepsilon \Theta'$,*

$$P_{\theta,n}(C) = \exp(-n\psi(\theta) + \theta' A_n + B_n\|\theta\|h(\theta))\, f_{\theta,n}(C),$$

where the factor $f_{\theta,n}(C)$ is bounded by

$$a/B_n^{\,d} \leqq f_{\theta,n}(C) \leqq b/B_n.$$

(This theorem applies to both the non-lattice and 1-lattice cases.)

PROOF. Simple geometric considerations show that there will exist a constant k_0, $0 < k_0 < 1$, depending on τ, the radius of K, and the radius of Θ', such that C_y will contain a sphere of radius $k_0 y$ for $0 < y < \tau$. It is clear then that the approximations for $P_{\theta,n}(C)$ expressed in Theorems 2 and 3 can be made to hold uniformly. We complete the proof with a suitable approximation for $\int_0^\infty \nu(C_y)B_n e^{-B_n y}\,dy$.

We note that by the comment above and the boundedness of the containing

set K, there exist positive constants a' and b' such that the volume of C_y satisfies

$$b'y^d \leqq \text{Vol } (C_y) \leqq a'y$$

for $\theta < y < \min(\tau, 1)$ and $\theta \varepsilon \Theta'$.

Letting $\phi_n = (d+1)\log B_n$ as in Note 2, we have

$$\int_0^{\phi_n/B_n} \nu(C_y) B_n e^{-B_n y}\, dy \geqq mb' B_n^{-d} \int_0^{\phi_n/B_n} B_n^{d+1} y^d e^{-B_n y}\, dy,$$

and the right-hand integral approaches the gamma function $\Gamma(d+1)$ as n goes to infinity. This establishes the lower bound of the theorem, and the upper bound follows in a similar way, making use of Note 2.

It is obvious from the above argument that with more specific knowledge of the set C it is a simple matter to obtain more precise estimates of $P_{\theta,n}(C)$. Such calculations are carried through in Sections 4 and 5 for the sets C corresponding to acceptance regions of likelihood ratio tests.

3. Agreeing exponential families. Suppose now that we have two exponential families,

$$dP_\theta(x) = e^{\theta' x - \psi(\theta)}\, d\mu(x) \quad \text{and} \quad d\tilde{P}_\theta(x) = e^{\theta' x - \tilde{\psi}(\theta)}\, d\tilde{\mu}(x),$$

and we consider a point θ_0 in the intersection of the two parameter spaces, $\theta_0 \varepsilon \Theta \cap \tilde{\Theta}$.

DEFINITION. P_θ and $\tilde{P}_\theta$ are said to *agree* at the point $\theta_0 \varepsilon \tilde{\Theta} \cap \Theta$ if there exists vectors $A_n \varepsilon E^d$ and positive constants B_n such that under both P_{θ_0} and $\tilde{P}_{\theta_0}$ the normalized sum of independent observations $X = (\sum_1^n X_i - A_n)/B_n$ converges to the same stable law ν, (i.e., P_{θ_0} and $\tilde{P}_{\theta_0}$ are attracted to the same stable law by the same normalizing constants). As before, we let $P_{\theta,n}(\tilde{P}_{\theta,n})$ represent the distribution of X under θ; in general, for $\theta \neq \theta_0$ the $P_{\theta,n}$ distribution will "move to infinity" as n grows large. Following the notation of Theorem 2, we have:

THEOREM 5. *For a given convex set C, suppose that C_{1/B_n} contains a sphere of radius k/B_n, $0 < k < 1$. Then if P_θ and $\tilde{P}_\theta$ agree at θ_0, and are both non-lattice distributions,*

$$P_{\theta,n}(C)/\tilde{P}_{\theta,n}(C) = \exp\left(-n[(\psi(\theta) - \psi(\theta_0)) - (\tilde{\psi}(\theta) - \tilde{\psi}(\theta_0))]\right) \cdot [1 + o_n(1)/mk^d]$$

where $o_n(1)$ approaches zero uniformly in $\theta \varepsilon \Theta \cap \tilde{\Theta}$, C, and k.

PROOF. The proof is immediate from Theorem 2 as expressed in Note 3. The approximating factor is seen to be $(1 + o_n^*(1)/mk^d)/(1 + o_n^*(1)/mk^d)$.

If either P_{θ_0} or $\tilde{P}_{\theta_0}$ (or both) is 1-lattice, then the hypotheses on C must be strengthened as in Theorem 3, while the conclusion is weakened to

$$P_{\theta,n}(C)/\tilde{P}_{\theta,n}(C) = \exp\left(-n[(\psi(\theta) - \psi(\theta_0)) - (\tilde{\psi}(\theta) - \tilde{\psi}(\theta_0))]\right) \cdot [e^{o_n(1)}]$$

as in that theorem.

The most useful case of Theorem 5 for numerical approximation is that where

$\tilde{P}_\theta$ is a normal translation family in θ, for then the denominator $\tilde{P}_{\theta,n}(C)$ can often be evaluated from standard tables. (This technique will be illustrated in Section 6.)

If P_{θ_0} has finite first and second moments, $E_{\theta_0}X \equiv \lambda(\theta_0)$, $\mathrm{Cov}_{\theta_0}(X) \equiv \Sigma(\theta_0)$, (which is always the case for θ_0 interior to Θ) then there exists an *agreeing normal family*

$$\tilde{P}_\theta \sim \mathfrak{N}(\lambda(\theta_0) + \Sigma(\theta_0)[\theta - \theta_0], \Sigma(\theta_0)),$$

(the particular form of the expectation as a function of θ being necessary to achieve the kernal $e^{\theta' x - \tilde{\psi}(\theta)}$). This distribution has the log generating function

$$\tilde{\psi}(\theta) = \lambda_0'[\theta - \theta_0] + \tfrac{1}{2}[\theta - \theta_0]'\Sigma_0[\theta - \theta_0],$$

where we have used the shorter notation $\lambda_0 \equiv \lambda(\theta_0)$ and $\Sigma_0 \equiv \Sigma(\theta_0)$. The exponential term in Theorem 5 becomes

$$\psi(\theta) - \{\psi(\theta_0) + \lambda_0'[\theta - \theta_0] + \tfrac{1}{2}[\theta - \theta_0]'\Sigma_0[\theta - \theta_0]\} \equiv \Delta_{\theta_0}(\theta)$$

in this case, which in just $\psi(\theta)$ minus its second order Taylor expansion around θ_0, a function we have chosen to call $\Delta_{\theta_0}(\theta)$. Thus for the normal agreeing family,[4]

$$P_{\theta,n}(C) = \exp(-n\Delta_{\theta_0}(\theta))\, \tilde{P}_{\theta,n}(C)[1 + o_n(1)/mk^d],$$

where $\tilde{P}_{\theta,n}(C)$ is the probability that a $\mathfrak{N}(n^{\frac{1}{2}}\Sigma_0[\theta - \theta_0], \Sigma_0)$ random variable falls in the set C. For θ interior to Θ, the function $\Delta_{\theta_0}(\theta)$ will behave like $\|\theta - \theta_0\|^3$ for θ near θ_0, and we obtain the familiar result that $P_{\theta,n}$ behaves like a normal translation family for "small deviations", i.e., for $\|\theta_n - \theta_0\| = o(n^{-\frac{1}{3}})$. Beyond this point, $P_{\theta,n}(C)$ deviates exponentially from $\tilde{P}_{\theta,n}(C)$ as indicated, and it is instructive to plot the function $\Delta_{\theta_0}(\theta)$. Values of $\Delta_{\theta_0}(\theta) > 0$ indicate "supernormal" behavior (faster convergence to zero), while $\Delta_{\theta_0}(\theta) < 0$ is "subnormal".

The function $\Delta_{\theta_0}(\theta)$ may also be expressed as

$$\Delta_{\theta_0}(\theta) = I(\theta_0, \theta) - \tfrac{1}{2}J(\theta_0, \theta),$$

where I is the Kullback-Liebler information

$$I(\theta_0, \theta) = E_{\theta_0}(\log(dP_{\theta_0}/dP_\theta)(X))$$

and

$$J(\theta_0, \theta) = \mathrm{Var}_{\theta_0}(\log(dP_{\theta_0}/dP_\theta)(X)).$$

(This is the notation used in [8].)

In addition to its value as a computational device, Theorem 5 offers interesting theoretical insights into the structure of exponential families. We offer two immediate corollaries of Theorem 5:

[4] Going back to Theorem 2, we notice that for the normal agreeing family the approximation given for $\tilde{P}_{\theta,n}(C)$ is exact, that is $\tilde{o}_n(1) \equiv 0$, so that all the error in this formula comes from estimating the numerator $P_{\theta,n}(C)$ in Theorem 5.

COROLLARY 1. *Let K be a bounded convex set, and let $\Lambda(K, \tau)$ be the class of sets defined in Note 1 to Theorem 2. Then if both P_θ and $\tilde{P}_\theta$ are non-lattice and agree at θ_0, the conditional probabilities satisfy*

$$P_{\theta,n}(C \mid K)/\tilde{P}_{\theta,n}(C \mid K) = 1 + o_n(1)$$

uniformly for C in $\Lambda(K, \tau)$ and θ in any bounded subset of $\Theta \cap \tilde{\Theta}$.

COROLLARY 2. *Suppose that P_{θ_0} has finite mean and covariance, and let $\tilde{P}_\theta$ be the agreeing normal family to P_θ at θ_0. Then if we define*

$$\tilde{\theta} = \theta_0 + (\theta - \theta_0)[2I(\theta_0, \theta)/(\theta - \theta_0)'\Sigma_0(\theta - \theta_0)]^{\frac{1}{2}},$$

where $I(\theta_0, \theta) = E_{\theta_0} \log (dP_{\theta_0}/dP_\theta)(X)$,

$$\log P_{\theta,n}(C)/\log \tilde{P}_{\tilde{\theta},n}(C) = 1 + o_n(1)$$

uniformly as in Theorem 5, in both the non-lattice and 1-lattice cases.

(Note that $(\theta - \theta_0)'\Sigma_0(\theta - \theta_0) = 2\tilde{I}(\theta_0, \theta)$.)

4. On the Chernoff bound and a theorem of Hoeffding. We present two examples illustrating the use of the preceding theorems to obtain accurate approximations for the power function of a test of hypothesis. (A third, somewhat different result is given in Section 5.)

First let us consider testing a simple hypothesis versus a simple alternative, say $H_0 : P = P_0$ vs. $H_1 : P = P_1$, where we assume that P_0 and P_1 are absolutely continuous with respect to each other.

In a well-known paper [6] Chernoff has shown that there exists a positive number, which we will call I_2, with the following property: let ξ_0 and $\xi_1 = 1 - \xi_0$ be the *a priori* probabilities for H_0 and H_1, $0 < \xi_0 < 1$. Then if α_n and β_n represent the probabilities of errors of the first and second kinds respectively based on n independent observations, the Bayes risk satisfies

$$\lim_{n \to \infty} [\log (\xi_0\alpha_n + \xi_1\beta_n)]n^{-1} = -I_2.$$

To apply our results to this problem, we embed P_0 and P_1 in an exponential family in the usual way: that is, we consider the real-valued sufficient statistic $X = \log (dP_1/dP_0(Y))$, and define the one-parameter exponential family

$$dP_\theta(x) = e^{\theta x - \psi(\theta)}\, dP_0(x)$$

with parameter space $\Theta = [0, 1]$. Here $\theta = 0$ ($\theta = 1$) corresponds to $P_0(P_1)$, and $\psi(0) = \psi(1) = 0$.

Let θ_2 be the point interior to Θ where the convex function $\psi(\theta)$ achieves its minimum. It is easy to show that $I(\theta_2, 0) = I(\theta_2, 1) \equiv I$ (say), where $I(\cdot, \cdot)$ is the Kullback-Liebler information defined previously. Note that $E_\theta X = d\psi(\theta)/d\theta$ is zero at $\theta = \theta_2$.

The Bayes test versus *a priori* probabilities ξ_0 and ξ_1 on H_0 and H_1 rejects H_0 for $\sum_1^n X_i > \log \xi_0/\xi_1$, or, normalizing under θ_2, for $\sum_1^n X_i/n^{\frac{1}{2}} > t/n^{\frac{1}{2}}$ where $t = \log \xi_0/\xi_1$.

Let us consider the case where X is non-lattice. To find β_n we apply Theorem 2 as in Note 3, with $\theta = 1$, $\theta_0 = \theta_2$ and C which we can take to be $\{(t - 2\log n^{\frac{1}{2}})/n^{\frac{1}{2}} \leqq x \leqq t/n^{\frac{1}{2}}\}$ by Note 2. In this case it is trivial to approximate the integral $\int_0^\infty \nu(C_y)B_n e^{-B_n y}\,dy$ by $[1 + o_n(1)]/(1 - \theta_2)(2\pi n\sigma_2^2)^{\frac{1}{2}}$, $(\sigma_2^2 = \mathrm{Var}_{\theta_2} X = d^2\psi(\theta_2)/d\theta^2)$, and noting that $h(\theta) = t/n^{\frac{1}{2}}$ for $\theta = 1$ we obtain

$$\beta_n = \exp(-nI + (1 - \theta_2)t)\,(1 - \theta_2)^{-1}(2\pi n\sigma_2^2)^{-\frac{1}{2}}(1 + o_n(1))$$

with the corresponding result for α_n being

$$\alpha_n = \exp(-nI - \theta_2 t)\,\theta_2^{-1}(2\pi n\sigma_2^2)^{-\frac{1}{2}}(1 - o_n(1)).$$

Thus $\beta_n/\alpha_n = \theta_2(1 - \theta_2)^{-1}\xi_0\xi_1^{-1}(1 + o_n(1))$, and the Bayes risk is given by

$$\xi_0\alpha_n + \xi_1\beta_n = \xi_0^{\theta_2}\xi_1^{1-\theta_2}\theta_2^{-1}(1 - \theta_2)^{-1}e^{-nI}(2\pi n\sigma_2^2)^{-\frac{1}{2}}(1 + o_n(1))$$

uniformly for ξ_0 bounded away from 0 and 1.[5] We recognize this as a sharpened version of Chernoff's theorem, so that I must equal I_2.

For our second application of Theorems 2 and 3, we return to the case of the exponential family $dP_\theta(x) = e^{\theta' x - \psi(\theta)}\,d\mu(x)$, and consider the problem of testing $H_0 : \theta = \theta_1$ versus the general alternative $H_1 : \theta \varepsilon \Theta - \{\theta_1\}$, where θ_1 is a point interior to Θ. As before we indicate the expectation under θ by $\lambda(\theta) \equiv E_\theta X$, and define the set of all possible expectations as Λ,

$$\Lambda = \{\lambda(\theta) : \theta \varepsilon \Theta\}.$$

Given observations $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, it is simple to show that the maximum likelihood estimate of θ, $\hat{\theta}(x_1, \cdots, x_n)$, is given by $\hat{\theta} : \lambda(\hat{\theta}) = \bar{x} = \sum_1^n x_i/n$, provided that $\bar{x} \varepsilon \Lambda$. (If not, $\hat{\theta}$ is a boundary point of Θ.) More precisely,

$$dP_\theta(x_1, x_2, \cdots, x_n) = e^{-nI(\,\cdot\,,\theta)}\,dP_{\hat{\theta}}(x_1, x_2, \cdots, x_n),$$

under the condition $\bar{x} \varepsilon \Lambda$, where $I(\cdot, \cdot)$ is once again the Kullback-Liebler information. We see that the likelihood ratio tests of H_0 vs. H_1 are of the form "reject H_0 for $\hat{\theta} \varepsilon R_D$, where $R_D = \{\theta : I(\theta, \theta_1) > D\}$".

Hoeffding [12] has shown that in the case of the d-dimensional multinomial distribution ($d + 1$ possible disjoint outcomes),

$$P_{\theta_1,n}(R_D) = n^{(d-2)/2}e^{-nD}[e^{o_n(1)}].$$

(Here we are abusing our notation slightly by writing $P_{\theta_1,n}(R_d)$ for $\mathrm{Prob}_{\theta_1}(\hat{\theta}(X_1, \cdots, X_n) \varepsilon R_D)$.) For a general exponential family we now give the analogue of Hoeffding's theorem:

THEOREM 6. *For a non-lattice exponential family, let θ_1 be a point in the interior of the parameter space Θ, let $R_D = \{\theta : I(\theta, \theta_1) > D\}$, and let $\bar{D} = \sup\{D : \Theta - R_D$ is contained in the interior of $\Theta\}$. Then*

$$P_{\theta_1,n}(R_D) = n^{(d-2)/2}e^{-nD}k_D[1 + o_n(1)]$$

[5] This result does not seem to appear in the literature, but it can be obtained easily from results in both [2] and [15].

uniformly in D for D in any range $\epsilon \leqq D \leqq \bar{D} - \epsilon$, $\epsilon > 0$, where k_D is a constant not depending on n. (In the 1-lattice case, the $1 + o_n(1)$ factor must be replaced by $e^{o_n(1)}$, the uniformity holding as indicated.)

PROOF. We present the proof only in outline form: it is more convenient to work in the Λ space, so let S_D be the mapping of R_D under the transformation $\lambda = \lambda(\theta)$. It is not difficult to show that this mapping is one to one, continuous, and increasing in the sense that for every $\theta_1 \neq \theta_2$ in Θ, $(\theta_1 - \theta_2)'(\lambda(\theta_1) - \lambda(\theta_2)) > 0$. Moreover, for $0 < D < \bar{D}$, the set $A_D = \Lambda - S_D$ is bounded and convex, with an analytic boundary, and if $\lambda(\theta_0)$ is a point on the boundary of A_D, i.e. $I(\theta_0, \theta_1) = D$, then the outward-pointing normal vector to A_D through $\lambda(\theta_0)$ is parallel to $\theta_0 - \theta_1$. [Note: it is incorrectly stated in [5] that A_D is convex for all D.]

Let U be the unit sphere in the Λ space, centered at $\lambda(\theta_1)$, and for a boundary point $\lambda(\theta_0)$ of A_D, let ω_0 be the point on U given by $\omega_0 = (\lambda(\theta_0) - \lambda(\theta_1))/\|\lambda(\theta_0) - \lambda(\theta_1)\|$. Let $V_n(\omega_0)$ be those points on U which are within $n^{-3/4}$ units of ω_0. If $\phi(V)$ represents ordinary Lebesgue measure on the surface U, then $\phi(V_n(\omega_0)) = k_1 n^{-(3/4)(d-1)}[1 + o_n(1)]$ for some constant k_1.

Define $B_n(\omega_0)$ as the cone of vectors originating at $\lambda(\theta_1)$ and passing through $V_n(\omega_0)$, and let $C_n(\omega_0) = B_n(\omega_0) \cap S_D$. Theorem 2 is now used—in the non-lattice case—to show that there exists a constant $k(\omega_0)$ such that

$$P_{\theta_1,n}(C_n(\omega_0)) = k(\omega_0) n^{-(d+1)/4} e^{-nD}[1 + o_n(1)],$$

so that the ratio $r(\omega_0) \equiv [P_{\theta_1,n}(C_n(\omega_0))]/\phi(V_n(\omega_0))$ is given by

$$r(\omega_0) = k(\omega_0) k_1^{-1} n^{(d-2)/2} e^{-nD}[1 + o_n(1)].$$

Finally, we note that the $o_n(1)$ term can be made uniform over the choice of ω_0 on U (a compactness argument yields this easily, upon examination of the proof of Lemma 1). Since $P_{\theta_1,n}(R_D) = \int_U r(\omega_0)\, d\phi(\omega_0)$, the theorem is verified in the non-lattice case, with $k_D = \int_U k(\omega_0)/k_1\, d\phi(\omega_0)$. A similar argument using Theorem 3 gives the result in the 1-lattice case.

5. Asymptotic power and optimality of the likelihood ratio test at a fixed significance level. In this section we consider testing the null hypothesis $H_0 : \theta = \theta_0$ versus the general alternative $H_1 : \theta \in \Theta - \{\theta_0\}$ at a fixed significance level α, $0 < \alpha < 1$, using the likelihood ratio test "reject H_0 for $I(\hat{\theta}, \theta_0) > D_n$" as described in Section 4. Here θ_0 will be taken in the interior of Θ.

In order to achieve a fixed level α, the constants D_n must approach zero at the asymptotic rate $\chi_d^2(\alpha)/2n$, where $\chi_d^2(\alpha)$ is the upper α-point of the χ^2-distribution with d degrees of freedom. This follows from the Taylor expansion $I(\hat{\theta}, \theta_0) = \frac{1}{2}(\bar{x} - \lambda_0)'\Sigma_0^{-1}(\bar{x} - \lambda_0) + O(\|\bar{x} - \lambda_0\|^3)$ holding for $\bar{x}$ in an open neighborhood R of λ_0 in the Λ space (where as previously we have written $\lambda_0 \equiv \lambda(\theta_0)$ and $\Sigma_0 \equiv \Sigma(\theta_0)$). Transforming to the normalized statistic $x = n^{\frac{1}{2}}(\bar{x} - \lambda_0)$, as before, the acceptance regions C_n for the level α likelihood ratio test will converge to a limiting region C which is the ellipsoid

$$C = \{x : x'\Sigma_0^{-1}x \leqq \chi_d^2(\alpha)\}.$$

We shall first give an optimality property of the region C which is the "large deviations" equivalent of a well known result of Wald [19]. There is no real loss of generality in assuming that $\theta_0 = 0$, $\lambda_0 = 0$, and $\Sigma_0 = I$, so that C becomes the sphere $\|x\| \leqq \chi_d(\alpha) \equiv (\chi_d^2(\alpha))^{\frac{1}{2}}$.

Let B be any other fixed, (for all n), bounded, convex, asymptotically level α test of H_0, i.e., $P(X \varepsilon B) = 1 - \alpha$ for $X \sim \mathfrak{N}(0, I)$. By Theorem 4 we have

$$n^{-\frac{1}{2}} \log P_{\theta,n}(B)/P_{\theta,n}(C) = \|\theta\|(h_B(\theta) - \chi_d(\alpha)) + O([\log n]n^{-\frac{1}{2}})$$

uniformly for θ in any bounded subset of Θ, where

$$h_B(\theta) = \sup_{x \varepsilon B} \theta' x/\|\theta\|.$$

If the d-dimensional sphere $S_d = \{\theta : \|\theta\| = r\}$ is contained in Θ, a natural measure of how poorly the test with acceptance region B behaves on S_d as a whole, compared with the behavior of the asymptotic likelihood ratio region C, is given by

$$\int_{S_d} \log (P_{\theta,n}(B)/P_{\theta,n}(C))\, d\phi_d(\theta),$$

$\phi_d(\theta)$ representing the uniform probability distribution on the surface S_d.

THEOREM 7. *Under the conditions given above,*

$$\lim_{n\to\infty} n^{-\frac{1}{2}} \int_{S_d} \log (P_{\theta,n}(B)/P_{\theta,n}(C))\, d\phi_d(\theta) \geqq 0,$$

with equality if and only if $B = C$ a.e.

PROOF. The limit equals $\|\theta\| \int_{S_d} [h_B(\theta) - \chi_d(\alpha)]\, d\phi_d(\theta)$, so it is equivalent to verify the following statement: among all sets with $\mathfrak{N}(0, I)$ probability $1 - \alpha$, the sphere C minimizes the integral $\int_{S_d} h_B(\theta)\, d\phi_d(\theta)$. This statement is obviously true for $d = 1$ (ϕ_1 is the distribution putting probability mass $\frac{1}{2}$ on the points r and $-r$). For $d > 1$, note that

$$\int_{S_d} h_B(\theta)\, d\phi_d(\theta) = \int_{S_d} \left[\int_{S_d(\theta)} h_{B(\theta)}(\theta')\, d\phi_{d-1}(\theta')\right] d\phi_d(\theta),$$

where $S_d(\theta)$ and $B(\theta)$ are the projections of S_d and B respectively into the $d - 1$ dimensional hyperplane orthogonal to θ. The theorem follows by induction on d.

An accurate large sample expression for the power of the level α likelihood ratio test is given in the next theorem (better formulae for calculating the power in moderate samples are discussed in Section 6).

THEOREM 8. *The power of the level* α *likelihood ratio test of* $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, θ_0 *interior to* Θ, *is given by*

$$P_{\theta,n}(C_n) = \exp\{-nI(\theta_0, \theta) + [n\chi_d^2(\alpha)(\theta - \theta_0)'\Sigma_0(\theta - \theta_0)]^{\frac{1}{2}}\}\, n^{-(d+1)/4}\, [e^{o_n(1)}],$$ *where*

the factor $e^{o_n(1)}$ *is bounded away from* 0 *and* ∞ *uniformly for* θ *in any bounded subset of* Θ.

PROOF. Let us again assume that we have transformed the problem so that $\theta_0 = 0$, $\lambda_0 = 0$, and $\Sigma_0 = I$. The displayed expression for $P_{\theta,n}(C_n)$ is easily seen to be correct for $P_{\theta,n}(C)$, (via Theorem 4), where C is the limiting set

$\|x\|^2 \leqq \chi_d^2(\alpha)$. It remains to show that $P_{\theta,n}(C_n)/P_{\theta,n}(C) = e^{o_n(1)}$, which follows immediately from the following lemma:

LEMMA. *There exists a constant* $\tau > 0$ *such that*

$$\{\|x\|^2 \leqq \chi_d^2(\alpha) - \tau n^{-\frac{1}{2}}\} \subset C_n \subset \{\|x\|^2 \leqq \chi_d^2(\alpha) + \tau n^{-\frac{1}{2}}$$

for all sufficiently large n.

PROOF. Using our previous notation for the Taylor expansion of $I(\hat{\theta}, \theta_0)$, we have

$$C_n = \{\|x\|^2 + n^{-\frac{1}{2}}O(\|x\|^3) \leqq l_n\},$$

where in terms of the original constants D_n, $l_n = 2nD_n$. Let $\chi_d^2(\alpha) \equiv l$, and note that there exists a positive constant τ_1, such that

$$\{\|x\|^2 \leqq l' - \tau_1 n^{-\frac{1}{2}}\} \subset \{\|x\|^2 + n^{-\frac{1}{2}}O(\|x\|^3) \leqq l'\} \subset \{\|x\|^2 \leqq l' + \tau_1 n^{-\frac{1}{2}}\}$$

for l' in a bounded interval containing l.

Esseen [10] has shown that there exists a constant $\tau_2 > 0$ such that

$$|P_{\theta_0}(\|\textstyle\sum_1^n X_i/n^{\frac{1}{2}}\|^2 \leqq l') - (1 - \alpha(l'))| < \tau_2 n^{-d/d+1}$$

uniformly for the l' as above, where $\alpha(l')$ is the probability that a χ_d^2 random variable exceeds l'. For any positive number j we have

$$\begin{aligned} P_{\theta_0}(2_n I(\hat{\theta}, \theta_0) &\leqq l + (j\tau_2 + \tau_1)n^{-\frac{1}{2}}) \\ &\geqq P_{\theta_0}(\|\textstyle\sum_1^n X_i/n^{\frac{1}{2}}\|^2 \leqq l + j\tau_2 n^{-\frac{1}{2}}) \\ &\geqq 1 - \alpha(l + j\tau_2 n^{-\frac{1}{2}}) - \tau_2 n^{-d/d+1} \geqq 1 - \alpha \end{aligned}$$

for j sufficiently large. Thus $l_n \leqq l + (j\tau_2 + \tau_1)n^{-\frac{1}{2}}$ and likewise $l_n \geqq l - (j\tau_2 + \tau_1)n^{-\frac{1}{2}}$. Going back to the set inequalities for a second time, we see that the lemma holds with $\tau = j\tau_2 + 2\tau_1$.

6. Numerical approximation of power functions. Suppose we wish to evaluate the power at the point $\theta \neq \theta_0$ of the level α likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on n independent observations from our exponential family P_θ. In the notation of Section 5, we wish to find the "error rate" $P_{\theta,n}(C_n)$. Unless n is enormous, the approximation given in Theorem 8 will not yield useful numerical results since that theorem is based on very large sample considerations.

Much more accurate approximations, which are useful in samples as small as $n = 10$, can be obtained via the application of Theorem 5 and the remarks following it in Section 3. In the notation of that section, we consider only the case where P_{θ_0} has finite mean and covariance, λ_0 and Σ_0, respectively, and let $\tilde{P}_\theta$ be the agreeing normal family $\tilde{P}_\theta \sim \mathfrak{N}(\lambda_0 + \Sigma_0(\theta - \theta_0), \Sigma_0)$. Our approximation formula is then

$$P_{\theta,n}(C_n) \approx e^{-n\Delta_{\theta_0}(\theta)}\tilde{P}_{\theta,n}(C_n), \tag{3}$$

where $\tilde{P}_{\theta,n}(C_n)$ is the probability that a $\mathfrak{N}(n^{\frac{1}{2}}\Sigma_0(\theta - \theta_0), \Sigma_0)$ random variable falls into C_n, and as before,

$$\Delta_{\theta_0}(\theta) = \psi(\theta) - \{\psi(\theta_0) + \lambda_0'(\theta - \theta_0) + \tfrac{1}{2}(\theta - \theta_0)'\Sigma_0(\theta - \theta_0)\}.$$

From the proof of Theorem 5 we know that there is only one source of error in (3), that which arises from replacing the exact distribution of $(\sum_1^n X_i - n\lambda_0)n^{-\frac{1}{2}}$ *under* H_0, $P_{\theta_0,n}$, by its limiting distribution $\nu \sim \mathfrak{N}(0, \Sigma_0)$ in the integral $\int_0^\infty P_{\theta_0,n}(C_{n,y})n^{\frac{1}{2}}e^{-n\frac{1}{2}y}\,dy$.

Two contradictory tendencies are at work in this approximation: when n is small, $P_{\theta,n}$ may deviate considerably from ν. On the other hand the integrating kernel $n^{\frac{1}{2}}e^{-n\frac{1}{2}y}$ is "diffuse" for small n, so that our estimate is averaged over a large set, and we would expect our error ratio to be not much worse than that for the global central limit theorem, i.e., of the same order of magnitude as $P_{\theta_0,n}(C_n)/\nu(C_n)$. As n grows larger, $P_{\theta_0,n}$ rapidly approaches ν, but on the other hand $n^{\frac{1}{2}}e^{-n\frac{1}{2}y}$ approaches a delta function at the origin. As we have seen, the ultimate error ratio depends on the local central limit theorem.

Of course the approximation (3) is of no use to us if we cannot evaluate the normal probability $\tilde{P}_{\theta,n}(C_n)$. In general this is a difficult task, but if we are willing to substitute the limiting set $C = \{x : x'\Sigma_0^{-1}x \leqq \chi_d^2(\alpha)\}$ for C_n we can simply read $\tilde{P}_{\theta,n}(C)$ out of a table of the non-central χ^2 distribution. Let us denote by $E_{d,\alpha}[\delta^2]$ the non-central χ^2 probability

$$E_{d,\alpha}[\delta^2] = \text{Prob}\,(\|Y\|^2 \leqq \chi_d^2(\alpha))$$

for a d-dimensional normal random vector $Y \sim \mathfrak{N}(\mu, I)$, $\|\mu\|^2 \equiv \delta^2$. Formula (3) now becomes

$$P_{\theta,n}(C_n) \approx e^{-n\Delta_{\theta_0}(\theta)}E_{d,\alpha}[n(\theta - \theta_0)'\Sigma_0(\theta - \theta_0)]. \tag{4}$$

In addition to the "probability error" discussed above, (4) involves a "set error" arising from the replacement of C_n by C. In general there does not seem to be a simple correction for set error (though see [8] for the case $d = 1$). On the other hand, since it is customary to use the approximate acceptance region C instead of the more complicated set C_n when actually performing the likelihood ratio test, what we have called set error may very well be quite appropriate to the approximation procedure.

"Probability error," on the other hand, can impair or destroy the accuracy of (4), particularly for large values of $\|\theta - \theta_0\|$. An obvious remedy for probability error is to use an improved estimate of $P_{\theta_0,n}$ in the integral $\int_0^\infty P_{\theta_0,n}(C_{n,y})n^{\frac{1}{2}}e^{-n\frac{1}{2}y}\,dy$. Instead of $\nu \sim \mathfrak{N}(0, \Sigma_0)$, we might, for instance, modify ν by the next term in the Cramér expansion of $P_{\theta_0,n}$. Such a modification [3] is computationally impractical in dimensions higher than 1 since it involves the relation of the entire third order central moment matrix to Σ_0.

The nature of the integral we are trying to estimate suggests a simpler approximation. Since the variable y equals $(\theta - \theta_0)'(x_0 - x)$ in the original deri-

vation of Theorem 2, where $(\theta - \theta_0)'x_0 = \sup_{x \in C_n} (\theta - \theta_0)'x$, the integrating kernel $n^{\frac{1}{2}}e^{-n^{\frac{1}{2}}y}$ varies entirely along the direction $\theta - \theta_0$ in the x space. Therefore we may expect good results if we use Cramér type corrections to ν only along the direction $\theta - \theta_0$. By first transforming to the case $\Sigma_0 = I$, it is simple to carry out this calculation using the familiar expansions [7] for the central limit theorem in one dimension. We obtain the improved approximation formula

$$P_{\theta_0,n}(C_n) \approx e^{-n\Delta_{\theta_0}(\theta)}[E_{d,\alpha}(\delta^2) + \tfrac{1}{6}\gamma n^{-\frac{1}{2}}E^*_{d,\alpha}(\delta^2)] \tag{5}$$

where, as before,

$$\delta^2 = n(\theta - \theta_0)'\Sigma_0(\theta - \theta_0),$$

γ is the skewness coefficient

$$\gamma = E_{\theta_0}[(\theta - \theta_0)'(X - \lambda_0)]^3/[(\theta - \theta_0)'\Sigma_0(\theta - \theta_0)]^{3/2}$$

and, using superscripts to denote components,

$$E^*_{d,\alpha}(\delta^2) = \int_{\|x\|^2 \leqq \chi^2_{d(\alpha)}} \exp\left(-\tfrac{1}{2}[(x^{(1)} - \delta)^2 + \textstyle\sum_2^d (x^{(j)})^2]\right) \cdot ((x^{(1)})^3 - 3x^{(1)})(2\pi)^{-d/2}\, dx^{(1)}\, dx^{(2)} \cdots dx^{(d)}.$$

A table of $E^*_{d,\alpha}(\delta^2)$ for $\alpha = .05$ is given below. The function $E_{d,\alpha}(\delta^2)$ can be obtained from the Pearson and Hartley non-central F tables [17] by setting $\nu_1 = d$, $\nu_2 = \infty$, and $\phi = (\delta^2/(d+1))^{\frac{1}{2}}$. All the other quantities involved in (5) can be calculated by differentiating the function ψ: let

$$\psi_\theta(t) \equiv \psi(\theta_0 + t(\theta - \theta_0)),$$

and let $\dot{\psi}_\theta(t)$, $\ddot{\psi}_\theta(t)$ and $\dddot{\psi}_\theta(t)$ indicate the first three derivatives of $\psi_\theta(t)$ with respect to the real parameter t. Then we have

$$\dot{\psi}_\theta(0) = (\theta - \theta_0)'\lambda_0, \qquad \ddot{\psi}_\theta(0) = (\theta - \theta_0)'\Sigma_0(\theta - \theta_0),$$

and

$$\dddot{\psi}_\theta(0) = E_{\theta_0}[(\theta - \theta_0)'(X - \lambda_0)]^3.$$

The efficacy of (4) and (5) in a univariate situation is illustrated by the following computations for the case $dP_\theta(x) = \theta e^{-\theta x}dx$, $x \geqq 0$, with parameter space $\Theta = (0, \infty)$. Taking $\theta_0 = 1$, alternative $\theta = .6$, and $\alpha = .05$, we obtain:

Sample size n	10	20	30	50
actual value $P_{\theta,n}(C_n)$	.465	.241	.120	.0270
approximation (4)	.518	.264	.130	.0287
approximation (5)	.468	.242	.120	.0270

(Here the set C_n is actually the acceptance region for the level α one-sided test of $\theta \geqq \theta_0$ versus $\theta < \theta_0$. Moreover, the effect of "set error" has been removed by calculating the E and E^* terms for the actual region C_n instead of for C.)

We now discuss, briefly and heuristically, the problem of testing a composite null hypothesis $H_0 : \theta \varepsilon \Theta_0$ versus the general alternative $H_1 : \theta \varepsilon \Theta - \Theta_0$, where Θ_0 is a d_0 dimensional subspace of Θ, $d_0 < d$, or more generally a d_0 dimensional differentiable manifold.

For a point $\theta \varepsilon \Theta - \Theta_0$, we would like to evaluate $P_\theta(\bar{X} \varepsilon R_n)$, where $\bar{X} = \sum_1^n X_i/n$ and R_n is the acceptance set in the $\bar{x}$ space of the level α likelihood ratio test of H_0 versus H_1. A general method or procedure would be to partition the $\bar{x}$ space, or equivalently the space Λ, into small blocks K_j of diameter approximately $1/n^{\frac{1}{2}}$, and for each block choose a point θ_{0j} such that $\lambda(\theta_{0j}) \varepsilon K_j$. The probability $P_\theta(\bar{x} \varepsilon R_n \cap K_j)$ is then approximated by re-normalizing to the random variable $(\sum_1^n X_i - \lambda(\theta_{0j}))/n^{\frac{1}{2}}$ and applying one of the previous theorems, the total probability $P_\theta(\bar{X} \varepsilon R_n)$ being obtained by summation over the blocks K_j (cf. [5]).

Asymptotically the dominant factor in the expression for $P_\theta(\bar{X} \varepsilon R_n \cap K_j)$ will be $e^{-nI(\theta_{0j},\theta)}$. Suppose now that there exists a unique point θ_{0*} in Θ_0 which is "nearest" to θ in the Kullback-Liebler distance,

$$I(\theta_{0*}, \theta) = \inf_{\theta_0 \varepsilon \Theta_0} I(\theta_0, \theta).$$

It is seen that the shape of R_n near $\lambda(\theta_{0*})$ will determine the asymptotic behavior of $P_\theta(\bar{X} \varepsilon R_n)$. As n grows large this shape approaches that of R, the acceptance region for H_{0*} : "$\theta - \theta_{0*}$ exists in a given d_0 dimensional subspace," under the agreeing normal family at θ_{0*}.

This reasoning suggests the following approximation for the power of the level α likelihood ratio test:

$$P_\theta(\bar{X} \varepsilon R_n) \approx e^{-n\Delta_{\theta_0*}(\theta)} E_{d_1,\alpha}(\delta^2), \tag{6}$$

where $d_1 = d - d_0$ and $\delta^2 = n(\theta - \theta_{0*})'\Sigma(\theta_{0*})(\theta - \theta_{0*})$. Taking skewness into account as in the case of the simple null hypothesis yields the estimate

$$P_\theta(\bar{X} \varepsilon R_n) \approx e^{-n\Delta_{\theta_0*}(\theta)}[E_{d_1,\alpha}(\delta^2) + \tfrac{1}{6}\gamma n^{-\frac{1}{2}} E^*_{d_1,\alpha}(\delta^2)], \tag{7}$$

where

$$\gamma = E_{\theta_0*}[(\theta - \theta_{0*})'(X - \lambda_{0*})]^3/[(\theta - \theta_{0*})'\Sigma_{0*}(\theta - \theta_{0*})]^{\frac{3}{2}}.$$

NOTE. We are assuming here that θ_{0*} is in the interior of Θ. Then the gradient relations $d\theta_0' \triangle_{\theta_0} I(\theta_0, \theta) = 0$ holding for a set of infinitesimal vectors $d\theta_0$ locally spanning Θ_0 near θ_0 must be satisfied by the point θ_{0*}. Using $\triangle_{\theta_0} I(\theta_0, \theta) = \Sigma(\theta_0)(\theta_0 - \theta)$ and $d\lambda(\theta) = \Sigma(\theta)\, d\theta$, this becomes $d\lambda_0'(\theta_0 - \theta) = 0$ for a set of infinitesimal vectors locally spanning Λ_0, the image of Θ_0, near $\lambda_0 = \lambda(\theta_0)$. In certain cases these equations have a trivial solution:

LEMMA. *Suppose that H_0 is $\theta_{(1)} = 0$, where $\theta = (\theta_{(1)}, \theta_{(2)})'$, $\theta_{(1)}$ being the first d_1 coordinates of θ. Then if the cross-covariance matrix $\Sigma_{12}(\theta)$ between $\theta_{(1)}$ and $\theta_{(2)}$ is identically zero for $\theta \varepsilon \Theta_0$, the gradient relations will have as their unique solution $\theta_{0*} = (0, \theta_{(2)})'$ for any $\theta \varepsilon \Theta$.*

An example which allows us to easily assess the accuracy of (6) is the familiar

student's t problem. We have n independent observations from a univariate normal distribution with unknown mean and variance, $X_i \sim \mathfrak{N}(\mu, \sigma^2)$, and we wish to test $H_0 : \mu = 0$. Applying the theory to the relevant two parameter exponential family yields the approximate error rate $P_n \approx E_{1,\alpha}((\mu/\sigma)^2)$. That is, instead of obtaining the actual non-central F probability $F_{1,n-1;\alpha}(\delta^2)$, we get the non-central χ^2 probability $F_{1,\infty;\alpha}(\delta^2)$. If we take $\mu/\sigma = \frac{1}{2}$, the numerical approximation is quite good:

n	16	32	48	64
Actual error probability	.53	.22	.079	.024
Approximation (6)	.48	.19	.066	.020

(In this example, $\gamma = 0$ so approximations (6) and (7) are identical.)

It may be shown that if the problem of testing H_0 is invariant under a group of transformations, then approximations (6) and (7) will depend only on the maximal invariant in the parameter space, e.g. $(\mu/\sigma)^2$ in the case above.

TABLE OF $E^*_{d,\alpha}(\delta^2)$, $\alpha = .05$

δ	d									
	1	2	3	4	5	6	7	8	9	10
0.5	−.4035	−.4164	−.2239	−.1243	−.0704	−0.403	−.0233	−.0134	−.0077	−.0044
1.0	−.6207	−.5352	−.1977	−.0447	.0202	.0435	.0475	.0435	.0367	.0295
1.5	−.5956	−.2994	.1267	.2866	.3177	.2930	.2481	.2004	.1570	.1205
2.0	−.4174	.0819	.5539	.7121	.7055	.6261	.5225	.4199	.3290	.2530
2.5	−.2215	.3417	.8115	.9842	.9759	.8778	.7448	.6085	.4842	.3779
3.0	−.0891	.3751	.7783	.9628	.9868	.9171	.8017	.6729	.5486	.4375
3.5	−.0266	.2631	.5474	.7122	.7667	.7445	.6765	.5875	.4939	.4049
4.0	−.0055	.1334	.2934	.4084	.4665	.4771	.4536	.4101	.3572	.3024
4.5	−.0007	.0510	.1218	.1833	.2239	.2427	.2427	.2295	.2081	.1827
5.0		.0150	.0394	.0647	.0850	.0981	.1037	.1030	.0976	.0892
5.5		.0034	.0100	.0180	.0255	.0315	.0353	.0370	.0367	.0350
6.0		.0006	.0020	.0039	.0061	.0080	.0096	.0106	.0111	
6.5			.0003	.0007	.0011	.0016	.0021	.0024	.0027	
7.0					.0002	.0003	.0004	.0004	.0005	

We conclude with an application to classical multivariate analysis: testing for independence between the components of a normally distributed vector. Let $Y_1, Y_2, \cdots, Y_n$ be independent $\mathfrak{N}(0, \theta^{-1})$ vectors in p dimensional space, $n \geqq p$, where θ is a positive definite matrix and thus has $p(p+1)/2 \equiv d$ independent components. The Wishart matrix $S = \sum_{l=1}^{n} Y_l Y_l'$ is a sufficient statistic, and a change of variables to $X_{ij} = -S_{ij}/(1 + \delta_{ij})$ yields an exponential family of our form,

$$dP_\theta(x) = \exp\left(\sum_{i \leqq j} \theta_{ij} x_{ij} + \tfrac{1}{2} n \log |\theta|\right) d\mu(x).$$

(Note that we can think of having either n independent observations $Y_i Y_i'$, or one observation of S. The approximation theory is not affected by such groupings, and in this case we prefer the latter because of the known form of the Wishart distribution for $n \geqq p$.)

Let us divide the components of the Y vectors into two classes, say the first p_1 components in one class and the last $p_2 = p - p_1$ components in the other. We partition the matrix θ in a corresponding way,

$$\theta = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix},$$

where $\theta_{21} = \theta_{12}'$, and we wish to test the independence hypothesis $H_0 : \theta_{12} = 0$.

To compute the approximate power of the level α likelihood ratio test [1], Chapter 9, we first note that the conditions of the lemma are satisfied. Then for any $\theta \, \varepsilon \, H_0$, θ positive definite, the "nearest point" in the null hypothesis is simply

$$\theta_{0*} = \begin{pmatrix} \theta_{11} & 0 \\ 0 & \theta_{22} \end{pmatrix}.$$

(Here it is not possible for the minimum of $I(\theta_0, \theta)$ to occur at the boundary of the Θ space—it is easily seen that $I(\theta_0, \theta)$ approaches infinity as θ_0 goes to the boundary.)

The computations can now be completed in a straightforward manner, either by differentiating

$$\psi_\theta(t) = -\tfrac{1}{2} n \log \left| \begin{pmatrix} \theta_{11} & t\theta_{12} \\ t\theta_{21} & \theta_{22} \end{pmatrix} \right|,$$

or computing the required quantities directly from the properties of the Wishart distribution. Applying (6), the approximate error probability is found to be

$$P_n \approx \exp\left(-\tfrac{1}{2} n[\log |I - D| + \tfrac{1}{2} \operatorname{tr} D]\right) E_{d_1,\alpha}(\delta^2 = n \operatorname{tr} D),$$

where $D = \theta_{12}\theta_{22}^{-1}\theta_{21}\theta_{11}^{-1}$, and $d_1 = p_1 p_2$. We note that this expression depends only on the eigenvalues of D, that is on the squares of the canonical correlations, which are the maximal invariant in this problem.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] BAHADUR, R. R. and RAO, R. R. (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027.

[3] BERGSTROM, H. (1945). On the central limit theorem in the space R_k. *Skand Actura* **28** 106–127.

[4] BIRNBAUM, A. (1955). Characterizations of complete classes of tests of some multiparametric hypotheses with applications to likelihood ratio tests. *Ann. Math. Statist.* **26** 21–36.

[5] BOROVKOV, A. A. and ROGOZIN, B. A. (1965). On the central limit theorem in the multidimensional case (in Russian). *Teor. Verojatnost. i. Primenen* **10** 61–69.

[6] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507.
[7] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.
[8] EFRON, B. (1967). The power of the likelihood ratio test. *Ann. Math. Statist.* **38** 802-806.
[9] EGGLESTON, H. G. (1958). *Convexity*. Cambridge Univ. Press.
[10] ESSEN, G. (1944). Fourier analysis of distribution functions. *Acta Math.* **77** 1–125.
[11] HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–401.
[12] HOEFFDING, W. (1967). On probabilities of large deviations. *Proc. Fifth Berkeley Symp.Math. Statist. Prob.*
[13] LEHMAN, E. L. (1959). *Testing Statistical Hypothoses*. Wiley, New York.
[14] MATTHES, T. K. and TRUAX, D. R. (1967). Tests of composite hypotheses for the multivariate exponential family. *Ann. Math. Statist.* **38** 681-697.
[15] PETROV, V. V. (1965). On the probabilities of large deviations for sums of independent random variables. *Theor. Prob. Appl.* **10** 287–298.
[16] RVAČEVA, E. L. (1954). On domains of attraction of multi-dimensional distributions. *Lvov. Gos. Univ. Uč. Zap.* **29** 5–49.
[17] SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.
[18] STONE, C. (1965). A local limit theorem for non-lattice multi-dimensional distribution functions. *Ann. Math. Statist.* **36** 546–551.
[19] WALD, A. (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.