JOURNAL OF MULTIVARIATE ANALYSIS **62**, 137–168 (1997)
ARTICLE NO. MV971680

Some New Statistics for Testing Hypotheses in Parametric Models[1]

D. Morales,* L. Pardo,* and I. Vajda†

**Complutense University of Madrid, Madrid, Spain;*
†Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, Prague, Czech Republic

The paper deals with simple and composite hypotheses in statistical models with i.i.d. observations and with arbitrary families dominated by σ-finite measures and parametrized by vector-valued variables. It introduces ϕ-divergence testing statistics as alternatives to the classical ones: the generalized likelihood ratio and the statistics of Wald and Rao. It is shown that, under the assumptions of standard type about hypotheses and model densities, the results about asymptotic distribution of the classical statistics established so far for the counting and Lebesgue dominating measures (discrete and continuous models) remain true also in the general case. Further, these results are extended to the ϕ-divergence statistics with smooth convex functions ϕ. The choice of ϕ-divergence statistics optimal from the point of view of power is discussed and illustrated by several examples. © 1997 Academic Press

1. INTRODUCTION

The idea of using the functionals of information theory, such as the entropy or divergence, in statistical inference is not new. In fact the so-called statistical information theory has provided several useful methods over the past years. For example, minimum divergence point estimators have been elaborated for models with continuous and discrete data (Beran, 1977; Győrfi, Vajda, and van der Meulen, 1994a, 1994b). Or divergence statistics, obtained by replacing unknown parameters by suitable estimates, have become successful competitors to the classical likelihood ratio-based statistics (Zografos, *et al.*, 1990; Pardo *et al.*, 1993; Morales *et al.*, 1994, 1995).

Received July 24, 1996; revised March 31, 1997.

AMS 1991 classifications: 62F05, 62B10.

Key words and phrases: parametric models, simple and composite hypotheses, asymptotic distributions of test statistics, generalized likelihood ratio, Wald's statistic, Rao's statistic, divergence statistics.

[1] This work was supported by Grants UCM PR219/94-5307, DGICYT PB94-0308, and GA CR201/96/0415.

0047-259X/97 $25.00

Copyright © 1997 by Academic Press
All rights of reproduction in any form reserved.

However, except for isolated cases such as the testing of independence in contingency tables (Pardo *et al.*, 1993; Zografos, 1993; Pardo, 1994), the divergence statistics have been applied so far to simple hypotheses only. The purpose of this work is to introduce the divergence statistics for testing general composite hypotheses about parameters in parametric models with i.i.d. observations. These statistics are compared with the classical statistics for testing general composite hypotheses, such as the generalized likelihood ratio or the Wald and Rao statistics (Serfling, 1980; Sen and Singer, 1993). Preferences between all these statistics based on power considerations are studied as well.

Let $\{P_\theta: \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ with open $\Theta \subset R^d$, $d \geqslant 1$. Measures P_θ are assumed to be described by densities

$$f_\theta = dP_\theta/d\mu$$

w.r.t. a dominating σ-finite measure μ on $\mathcal{X}$. We consider estimation and testing procedures based on a sequence of observations $\mathbf{X}_n = (X_1, ..., X_n)$ with components i.i.d. by a density from the family $\{f_\theta: \theta \in \Theta\}$.

Kupperman (1957) suggested a test of a null hypothesis $H \equiv \{\theta_0\} \subset \Theta$ about the true parameter using the *Kullback statistics*

$$K_n^0 = 2nK(\hat{\theta}_n, \theta_0), \tag{1}$$

obtained by replacing in the *Kullback divergence*

$$K(\theta, \theta_0) = \int_{\mathcal{X}} f_\theta \log \frac{f_\theta}{f_{\theta_0}} d\mu \tag{2}$$

the free parameter $\theta \in \Theta$ by the *maximum likelihood estimator* (MLE) $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ of the true parameter, and by a subsequent normalization by $2n$. He found regularity conditions such that, under H, $K_n^0 \xrightarrow{L} \chi_d^2$ (cf. also Section 5.5 in Kullback, 1959).

Note that, here and in the sequel, all convergences are considered for $n \to \infty$, χ_d^2 denotes the χ^2-distributed random variable with d degrees of freedom, and $\xrightarrow{L}$, $\xrightarrow{P}$ denote the stochastic convergences in law and in probability.

Salicrú *et al.* (1994) extended the validity of Kupperman's result to all *ϕ-statistics*

$$D_{\phi n}^0 = 2nD_\phi(\hat{\theta}_n, \theta_0), \tag{3}$$

where $\phi(t)$ is twice continuously differentiable and convex on $(0, \infty)$ with $\phi(1) = \phi'(1) = 0$, $\phi''(1) = 1$, and

$$D_\phi(\theta, \theta_0) = \int_{\mathscr{X}} f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right) d\mu \tag{4}$$

is the *ϕ-divergence* of densities from the family $\{f_\theta: \theta \in \Theta\}$ introduced by Csiszár (1963) (for a systematic theory of these divergences see Liese and Vajda, 1987).

Obviously, (1) and (2) are obtained from (3) and (4) for $\phi(t) = 1 - t + t \log t$ or, equivalently, for $\phi(t) = t \lg t$. The same authors proposed for testing the special composite null hypotheses $H \equiv \Theta_0 = \Theta_1 \times \{\theta_{20}\}$ and $H^* = \Theta_0^* = \{\theta_{10}\} \times \Theta_2$ in models with $\Theta = \Theta_1 \times \Theta_2$ the ϕ-statistics

$$D_{\phi n} = 2nD_\phi((\hat{\theta}_{1n}, \hat{\theta}_{2n}), (\hat{\theta}_{1n}, \theta_{20})), \qquad D_{\phi n}^* = 2nD_\phi((\hat{\theta}_{1n}, \hat{\theta}_{2n}), (\theta_{10}, \hat{\theta}_{2n})), \tag{5}$$

using the MLE $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n})$.

In the present paper we derive, under appropriate null hypotheses $H \equiv \Theta_0 \subset \Theta$ in suitably regular models, asymptotic distributions of the *general ϕ-statistics*,

$$D_{\phi n} = 2nD_\phi(\hat{\theta}_n, \tilde{\theta}_n), \tag{6}$$

where $\hat{\theta}_n$ is the same MLE as in (1) and $\tilde{\theta}_n$ is a *restricted maximum likelihood estimator* (RMLE), with values limited to the hypothetic subset Θ_0. Obviously, the ϕ-statistics (5) are particular versions of (6) obtained for the special hypotheses H and H^* in the special model.

There is a similarity between our results and the result of Simpson (1989) leading, under somewhat stronger regularity than considered in this paper, to the asymptotic distribution of statistic

$$\Delta_{\phi n} = 2n(D_\phi(\mathbf{X}_n, \tilde{\theta}_n) - D_\phi(\mathbf{X}_n, \hat{\theta}_n))$$

for the particular convex function $\phi(t) = (1 - \sqrt{u})^2$. In this formula

$$D_\phi(\mathbf{X}_n, \theta_0) = \int f_{\theta_0} \phi\left(\frac{\hat{f}_n}{f_{\theta_0}}\right) d\mu$$

denotes the ϕ-divergence between the unknown density f_{θ_0} and its arbitrary estimate $\hat{f}_n$. Simpson considered a nonparametric kernel estimate $\hat{f}_n$. If $\hat{f}_n = f_{\hat{\theta}_n}$ then $\Delta_{\phi n}$ reduces to our statistic (6), for any ϕ considered there. While the simulations of Simpson indicated a loss of power of his tests based on the Hellinger difference statistic $\Delta_{\phi n}$ relative to the likelihood ratio tests, our simulations indicated a gain of power of our tests based on

the Hellinger distance statistics $D_{\phi n}$, defined by (6) for the above considered particular $\phi(t)$. On the other hand, the Simpson tests exhibited a robustness which does not seem to be justified in our case.

2. REGULARITY ASSUMPTIONS

In the rest of paper we suppose that $\mathcal{X}$ is the support of σ-finite measure μ (otherwise we replace $(\mathcal{X}, \mathcal{A})$ by a restricted measurable space). Further, we suppose that the statistical model $\langle (\mathcal{X}, \mathcal{A}), \{f_\theta: \theta \in \Theta\}, \mu \rangle$ satisfies the standard regularity assumptions considered in the parametric asymptotic statistics (see, e.g., Chap. 4 in Serfling, 1980):

(M1) $\Theta \subset R^d$ is open and, for each $\theta \in \Theta$, f_θ is positive and the derivatives

$$\frac{\partial \log f_\theta}{\partial \theta_i}, \qquad \frac{\partial^2 \log f_\theta}{\partial \theta_i \, \partial \theta_j}, \qquad \frac{\partial^3 \log f_\theta}{\partial \theta_i \, \partial \theta_j \, \partial \theta_k}$$

exist everywhere on $\mathcal{X}$ for all $1 \leqslant i, j, k \leqslant d$.

(M2) For each $\theta_0 \in \Theta$, there exist measurable functions $\alpha, \beta, \gamma: \mathcal{X} \to [0, \infty)$, possibly depending on θ_0, such that for all θ in a neighborhood $N(\theta_0)$ and all above considered i, j, k the relations

$$\left|\frac{\partial f_\theta}{\partial \theta_i}\right| \leqslant \alpha, \qquad \left|\frac{\partial^2 f_\theta}{\partial \theta_i \, \partial \theta_j}\right| \leqslant \beta, \qquad \left|\frac{\partial^3 \log f_\theta}{\partial \theta_i \, \partial \theta_j \, \partial \theta_k}\right| \leqslant \gamma$$

hold on $\mathcal{X}$, and

$$\int_{\mathcal{X}} \alpha \, d\mu < \infty, \qquad \int_{\mathcal{X}} \beta \, d\mu < \infty, \qquad \int_{\mathcal{X}} \gamma \, f_\theta \, d\mu < \infty.$$

(M3) For each $\theta \in \Theta$, the *Fisher information matrix*

$$I_\theta = \left[\int_{\mathcal{X}} \frac{\partial \log f_\theta}{\partial \theta_i} \frac{\partial \log f_\theta}{\partial \theta_j} f_\theta \, d\mu \right]_{d \times d}$$

exists and is positive definite, with all elements continuous in the variable θ.

For example, any *natural exponential model* with densities

$$f_\theta = e^{T\theta^t - b(\theta)} \qquad \text{for} \quad T: \mathcal{X} \to R^d, \tag{7}$$

and with an open natural parameter space $\Theta \subset R^d$, which is minimal in the sense of Brown (1986) (i.e., is not overparametrized), satisfies (M1)–(M3). In this model

$$b(\theta) = \log \int_{\mathscr{X}} e^{T\theta^t}\, d\mu$$

is finite and analytic on Θ,

$$\tau(\theta) = \nabla b(\theta)$$

is invertible on Θ, and

$$\hat{\theta}_n = \tau^{-1}(T_n) \qquad \text{for} \quad T_n = T_n(X_1, ..., X_n) \triangleq \frac{1}{n} \sum_{i=1}^{n} T(X_i)$$

is the unique solution of the likelihood equation

$$\nabla \left[\left(\sum_{i=1}^{n} T(X_i) \right) \theta^t - nb(\theta) \right] = 0.$$

The Fisher information matrix is given by

$$I_\theta = \nabla^t \tau(\theta) = \nabla^t \nabla b(\theta).$$

Obviously, the elements of I_θ are continuous in the parameter θ.

Concerning the null hypothesis $H \equiv \Theta_0 \subset \Theta$ we suppose the following.

(H1) Θ_0 is a subset of R^d, and there exists $1 \leqslant d_0 < d$, open $B \subset R^{d-d_0}$ and mappings

$$g\colon \Theta \to R^{d_0}, \qquad h\colon B \to \Theta,$$

such that $\Theta_0 = \{h(\beta)\colon \beta \in B\}$ and $g(\theta) = 0$ for all $\theta \in \Theta_0$;

(H2) The $d_0 \times d$-matrix

$$G_\theta = \nabla^t g(\theta) \qquad \text{for} \quad \nabla = \left(\frac{\partial}{\partial \theta_1}, ..., \frac{\partial}{\partial \theta_d} \right)$$

exists and is of rank d_0 for all $\theta \in \Theta_0$, with all elements continuous on Θ_0;

(H3) The $d \times (d - d_0)$-matrix

$$H_\beta = [\tilde{\nabla}^t h(\beta)]^t \qquad \text{for} \quad \tilde{\nabla} = \left(\frac{\partial}{\partial \beta_1}, ..., \frac{\partial}{\partial \beta_{d-d_0}} \right)$$

exists and if of rank $d - d_0$ for all $\beta \in B$, with all elements continuous on B;

(H4) The statistical submodel $\langle (\mathcal{X}, \mathcal{A}), \{p_\beta = f_{h(\beta)}: \beta \in B\}, \mu \rangle$ satisfies (M1)–(M3).

Since $g(h(\beta)) = 0$ for all $\beta \in B$, the matrices $G_{g(\beta)}$ and H_β are orthogonal for all $\beta \in B$. This implies that the linear space generated by d_0 linearly independent row of vectors of $G_{g(\beta)}$ and $d - d_0$ linearly independent column vectors of H_β are orthogonal. For example, let

$$G = [G_{ij}]_{d_0 \times d}, \qquad H = [H_{ij}]_{d \times (d - d_0)} \tag{8}$$

be two orthogonal matrices of rank d_0 and $d - d_0$, respectively, and let B be an open subset of R^{d-d_0}. Then $\Theta_0 = \{\beta H^t: \beta \in B\}$ represents a null hypothesis H which fulfils (H1)–(H4) for $h(\beta) = \beta H^t$ and $g(\theta) = \theta G^t$ in any statistical model with parameter space $\Theta \subset R^d$ containing Θ_0 and satisfying (M1)–(M3).

The last set of assumptions concerns the function ϕ from the class Φ considered in (6):

(Φ1) The function $\phi: [0, \infty) \to (-\infty, \infty]$ is convex and continuous. Its restriction on $(0, \infty)$ is finite, twice continuously differentiable, with $\phi(1) = \phi'(1) = 0$ and $\phi''(1) = 1$;

(Φ2) Each $\theta_0 \in \Theta$ has an open neighborhood $N(\theta_0)$ such that for all $\theta \in N(\theta_0)$ and $1 \leqslant i, j \leqslant d$ it holds:

$$\frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right) d\mu = \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \left[f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right)\right] d\mu,$$

$$\frac{\partial^2}{\partial \theta_i \, \partial \theta_j} \int_{\mathcal{X}} f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right) d\mu = \int_{\mathcal{X}} \frac{\partial^2}{\partial \theta_i \, \partial \theta_j} \left[f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right)\right] d\mu,$$

and the derivatives of integrals are continuous on $N(\theta_0)$.

Every convex function $\phi: [0, \infty) \to (-\infty, \infty]$ finite and continuous on $(0, \infty)$ has finite derivatives $\phi'_+(t)$ from the right at all points $t \in (0, \infty)$. This function defines the same divergence (4) as the nonnegative convex function $\phi^*(t) = \phi(t) - \phi'_+(1)(t-1)$. The divergence is nonnegative if and only if $\phi(1) = 0$ (which is equivalent to $\phi^*(1) = 0$). In the sequel we often refer to the equivalence between differentiable convex functions $\phi(t)$ and $\phi(t) - \phi'_+(1)(t-1)$. E.g., $\phi(t) = t \log t$ is equivalent to $\phi^*(t) = 1 - t + t \log t$ or $\phi(t) = -\log t$ is equivalent to $\phi^*(t) = -1 + t - \log t$. The function ϕ^* in both cases satisfies (Φ1) while ϕ not because $\phi'(1) = 1$ or -1. If $\phi(t)$ is twice continuously differentiable at $t = 1$ with $\phi''(1) \neq 0$ then

$$\phi^*(t) = \frac{\phi(t) - \phi'(t)(t-1) - \phi(1)}{\phi''(1)}$$

satisfies ($\Phi 1$) and

$$D_{\phi^*}(\theta, \theta_0) = \frac{D_\phi(\theta, \theta_0) - \phi(1)}{\phi''(1)}.$$

This explains that ($\Phi 1$) is not an essential restriction on differentiable convex functions ϕ. Obviously, if $\psi: (0, \infty) \to R$ is convex or concave and twice continuously differentiable on $(0, \infty)$, with $\psi''(1) \neq 0$, then

$$\phi(t) = \frac{\psi(t) - \psi'(1)(t-1) - \psi(1)}{\psi''(1)}$$

satisfies ($\Phi 1$).

Next we present two conditions sufficient for ($\Phi 2$).

LEMMA 1. *Let us consider a model satisfying* (M1) *and a function* ϕ *satisfying* ($\Phi 1$). *If for each* $\theta_0 \in \Theta$ *there exist measure* μ*-integrable functions* $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}: \mathcal{X} \to [0, \infty)$, *possibly depending on* θ_0, *such that for all* θ *in a neighborhood* $N(\theta_0)$ *and* $1 \leqslant i, j \leqslant d$

$$\left| f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right) \right| \leqslant \tilde{\alpha},$$

$$\left| \phi'\left(\frac{f_\theta}{f_{\theta_0}}\right) \frac{\partial f_\theta}{\partial \theta_i} \right| \leqslant \tilde{\beta},$$

$$\left| \frac{1}{f_{\theta_0}} \phi''\left(\frac{f_\theta}{f_{\theta_0}}\right) \frac{\partial f_\theta}{\partial \theta_i} \frac{\partial f_\theta}{\partial \theta_j} + \phi'\left(\frac{f_\theta}{f_{\theta_0}}\right) \frac{\partial^2 f_\theta}{\partial \theta_i \partial \theta_j} \right| \leqslant \tilde{\gamma}$$

on $\mathcal{X}$, *then* ($\Phi 2$) *holds.*

Proof. For the considered model and ϕ,

$$\frac{\partial}{\partial \theta_i}\left[f_{\theta_0} \theta\left(\frac{f_\theta}{f_{\theta_0}}\right)\right] = \phi'\left(\frac{f_\theta}{f_{\theta_0}}\right) \frac{\partial f_\theta}{\partial \theta_i} \tag{9}$$

and

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j}\left[f_{\theta_0} \phi\left(\frac{f_\theta}{f_{\theta_0}}\right)\right] = \frac{1}{f_{\theta_0}} \phi''\left(\frac{f_\theta}{f_{\theta_0}}\right) \frac{\partial f_\theta}{\partial \theta_i} \frac{\partial f_\theta}{\partial \theta_j} + \phi'\left(\frac{f_\theta}{f_{\theta_0}}\right) \frac{\partial^2 f_\theta}{\partial \theta_i \partial \theta_j}. \tag{10}$$

By the theorem about differentiation behind integrals (see, e.g., Theorem 18 in Section 5.11 of Fleming, 1965), our assumptions are sufficient for differentiation behind the integrals considered in ($\Phi 2$) in the neighborhood

$N(\theta_0)$. Moreover, the derivatives behind integrals are assumed to be continuous and dominated. Therefore the continuity of derivatives of integrals follows from the Lebesgue dominated convergence theorem for integrals.

The function $\phi(t) = -\log t$ considered in the next lemma defines the *reversed Kullback divergence*,

$$\tilde{K}(\theta, \theta_0) \triangleq D_{-\log t}(\theta, \theta_0) = K(\theta_0, \theta) \qquad \text{(cf. (2))}. \tag{11}$$

LEMMA 2. *If the model satisfies* (M1)–(M3) *then* ($\Phi 2$) *holds for* $\phi(t) = -\log t$.

Proof. Define

$$\lambda_{\theta_{ij}} = \frac{\partial^2 \log f_\theta}{\partial \theta_i \, \partial \theta_j}, \qquad \psi_{\theta_{ij}} = \frac{\partial \log f_\theta}{\partial \theta_i} \frac{\partial \log f_\theta}{\partial \theta_j}.$$

It follows from (M2), (M3), and the mean value theorem that there exists a subneighborhood $N_{ij}(\theta_0)$ of the neighborhood $N(\theta_0)$ considered in (M2) and a constant $c_{ij} > 0$ such that

$$|\lambda_{\theta_{ij}}| \leqslant c_{ij}\gamma + |\lambda_{\theta_{0ij}}| \qquad \text{for all} \quad \theta \in N_{ij}(\theta_0). \tag{12}$$

By (M2)

$$\int_{\mathscr{X}} \gamma f_{\theta_0} \, d\mu < \infty,$$

and by the definition of $\lambda_{\theta_{0ij}}$ and $\psi_{\theta_{0ij}}$,

$$\int_{\mathscr{X}} f_{\theta_0} |\lambda_{\theta_{0ij}}| \, d\mu = \int_{\mathscr{X}} \left| f_{\theta_0} \psi_{\theta_{0ij}} - \left(\frac{\partial^2 f_\theta}{\partial \theta_i \, \partial \theta_j} \right)_{\theta = \theta_0} \right| d\mu,$$

where by (M3)

$$I_{\theta_0} = \left[\int_{\mathscr{X}} f_{\theta_0} \psi_{\theta_{0ij}} \, d\mu \right]_{d \times d}$$

is the positive definite and finite Fisher information matrix, and by (M2)

$$\int_{\mathscr{X}} \left| \frac{\partial^2 f}{\partial \theta_i \, \partial \theta_j} \right|_{\theta = \theta_0} d\mu \leqslant \int_{\mathscr{X}} \beta \, d\mu < \infty.$$

Therefore,

$$\int_X f_{\theta_0} |\lambda_{\theta_{0ij}}| \, d\mu < \infty;$$

i.e., the function

$$\tilde{\gamma} = \max_{1 \leqslant i, j \leqslant d} c_{ij} \gamma f_{\theta_0} + \max_{1 \leqslant i, j \leqslant d} |\lambda_{\theta_{0ij}}| f_{\theta_0}$$

is μ-integrable. Thus (12) implies that, for all

$$\theta \in \bigcap_{1 \leqslant i, j \leqslant d} N_{ij}(\theta_0)$$

and all $1 \leqslant i, j \leqslant d$, the expression $f_{\theta_0} |\lambda_{\theta_{ij}}|$ is bounded above by the μ-integrable function $\tilde{\gamma}$. On the other hand, for $\phi(t) = -\log t$ the expression (10) equals

$$\frac{1}{f_{\theta_0}} \left(\frac{f_\theta}{f_{\theta_0}} \right)^{-2} \frac{\partial f_\theta}{\partial \theta_i} \frac{\partial f_\theta}{\partial \theta_j} - \left(\frac{f_\theta}{f_{\theta_0}} \right)^{-1} \frac{\partial^2 f_\theta}{\partial \theta_i \partial \theta_j}$$

$$= f_{\theta_0} \frac{\partial \log f_\theta}{\partial \theta_i} \frac{\log f_\theta}{\partial \theta_j} - \frac{f_{\theta_0}}{f_\theta} \frac{\partial^2 f_\theta}{\partial \theta_i \partial \theta_j} = -f_{\theta_0} \lambda_{\theta_{ij}}.$$

This implies the validity of the last condition of Lemma 1. Proofs of validity of the previous two conditions of Lemma 1 are similar but easier. In the verification of the second inequaltiy for $\phi(t) = -\log t$, it follows from the mean value theorem that

$$\phi' \left(\frac{f_\theta}{f_{\theta_0}} \right) \frac{\partial f_\theta}{\partial \theta_i} = -\frac{\partial f_{\theta_0}}{\partial \theta_i} - f_{\theta_0} \lambda_{\theta_{ij}^*}$$

for some θ_{ij}^* from a subneighborhood $N_{ij}^*(\theta_0)$ of $N_{ij}(\theta_0)$. Therefore $|\phi'(f_\theta / f_{\theta_0})(\partial f_\theta / \partial \theta_i)|$ is bounded above by the μ-integrable function $\alpha + \tilde{\gamma}$ for all $\theta \in \bigcap_{1 \leqslant i, j \leqslant d} N_{ij}^*(\theta_0)$. In the verification of the first condition it suffices to use the equality

$$f_{\theta_0} \phi \left(\frac{f_\theta}{f_{\theta_0}} \right) = f_{\theta_0} \log \frac{f_{\theta_0}}{f_\theta}$$

and the inequality

$$-\frac{f_\theta}{e} \leqslant f_{\theta_0} \log \frac{f_{\theta_0}}{f_\theta} \leqslant f_{\theta_0} \left(\frac{f_{\theta_0}}{f_\theta} - 1 \right),$$

and to approximate f_θ by f_{θ_0} uniformly in a sufficiently small neighborhood $N(\theta_0)$, by employing the uniform bound α on the partial derivative $\partial f_\theta / \partial \theta_i$ considered in (M2).

LEMMA 3. *If the model is natural exponential as in* (7) *then* ($\Phi 2$) *holds for* $\phi(t) = -\log t$ *and for* $\phi(t) = t \log t$.

Proof. Clear from (2), (11) and from the relations

$$\log \frac{f_\theta}{f_{\theta_0}} = T(\theta - \theta_0)^t + b(\theta_0) - b(\theta)$$

and

$$\int_{\mathscr{X}} T f_\theta \, d\mu = \nabla b(\theta) \qquad \text{(cf. (H2))},$$

which follow from (7), and from the possibility to differentiate behind the integral in the formula for $b(\theta)$ next to (7) (as to the differentiability behind the integral; see Theorem 2.2 in Brown, 1986).

3. ASYMPTOTIC RESULTS

We start with some concepts and results for models satisfying (M1)–(M3). The first results stated below are not principally new, but will be needed in the present generality for references later. We shall use the random function

$$\lambda_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i),$$

its gradient

$$\delta_n(\theta) = \nabla \lambda_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla \log f_\theta(X_i),$$

and the random matrix

$$M_n(\theta) = \nabla^t \delta_n(\theta).$$

Unless otherwise explicitely stated, we consider a hypothesis $H_0 \equiv \{\theta_0\}$ about the true parameter value. Then it follows from (M1)–(M3) that for $Y_\theta = \nabla \log f_\theta(X_i)$

$$E(Y_{\theta_0}) = 0, \qquad E(Y_{\theta_0} Y_{\theta_0}^t) = I_{\theta_0}, \qquad E(\nabla^t Y_\theta)_{\theta=\theta_0} = -I_{\theta_0},$$

where I_{θ_0} is the Fisher information matrix defined in (M3). By the multivariate Lindeberg–Lévy central limit theorem, the first two equalities imply

$$\sqrt{n}\,\delta_n(\theta_0) \xrightarrow{L} N(0, I_{\theta_0}). \tag{13}$$

By the law of large numbers, the third equality implies

$$M_n(\theta_0) \xrightarrow{P} -I_{\theta_0}. \tag{14}$$

If $\mathcal{X} \subset R^d$ and μ is restriction of the Lebesgue measure on $\mathcal{X}$, then the theorem on page 145 and the arguments on pages 147–149 in Serfling (1980) imply that the MLE $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ exists and

$$\hat{\theta}_n \xrightarrow{P} \theta_0, \tag{15}$$

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, I_{\theta_0}^{-1}). \tag{16}$$

One can verify step by step that all arguments used there remain valid under the present generality.

For hypotheses $H \equiv \Theta_0 \subset \Theta$ satisfying (H1) and (H2) we shall define the *generalized Wald statistic*

$$W_n = ng(\hat{\theta}_n)(G_{\hat{\theta}_n} I_{\hat{\theta}_n}^{-1} G_{\hat{\theta}_n}^t)^{-1} g(\hat{\theta}_n)^t$$

(cf. Wald, 1943, and also p. 157 in Serfling, 1980, in the *cointinuous model*, where $\mathcal{X} \subset R^k$ and μ is the above-mentioned Lebesgue measure). Here $\hat{\theta}_n$ is the MLE, I_θ^{-1} denotes the inverse to the Fisher information matrix figuring in (M3), and $g(\theta)$, G_θ are the mapping and the matrix figuring in (H1) and (H2). If, moreover, the hypothesis satisfies (H3) and (H4), then we consider also the *auxiliary statistic*

$$A_n = n(\hat{\theta}_n - \tilde{\theta}_n)\, I_{\hat{\theta}_n}(\hat{\theta}_n - \tilde{\theta}_n)^t,$$

the *generalized Rao statistics*

$$R_n = n\delta_n(\tilde{\theta}_n)\, I_{\tilde{\theta}_n}^{-1} \delta_n(\tilde{\theta}_n)^t,$$

and the *generalized likelihood ratio test statistic* (GLRT statistic)

$$L_n = 2n[\,\lambda_n(\hat{\theta}_n) - \lambda_n(\tilde{\theta}_n)\,],$$

where

$$\tilde{\theta}_n = h(\hat{\beta}_n)$$

is the RMLE defined by the mapping h figuring in (H1), (H3) and by the MLE $\hat{\beta}_n(\mathbf{X}_n)$ for the restricted model considered in (H4) (cf. Rao, 1947, Section 6e in Rao, 1973, and pp. 158, 159 in Serfling, 1980, for the continuous models).

For arbitrary $\beta \in B$ it follows from the definition of the restricted model that the Fisher information $(d-d_0)\times(d-d_0)$-matrix J_β in this model satisfies the relation

$$J_\beta = H_\beta^t I_{h(\beta)} H_\beta,$$

where H_β is the matrix figuring in (H3) and $I_{h(\beta)}$ is the Fisher information of the unrestricted model for $\theta = h(\beta) \in \Theta_0$. Also it follows from (15) and (16) that if the true parameter satisfies for some $\beta_0 \in B$ the relation $\theta_0 = h(\beta_0)$, then

$$\hat{\beta}_n \xrightarrow{P} \beta_0, \qquad \tilde{\theta}_n \xrightarrow{P} \theta_0 \tag{17}$$

and

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{L} N(0, J_{\beta_0}^{-1}), \qquad n^{1/2}(\tilde{\theta}_n - \theta_0) \xrightarrow{L} N(0, H_{\beta_0} J_{\beta_0}^{-1} H_{\beta_0}^t). \tag{18}$$

In the following theorem we consider reductions A_n^0, R_n^0, and L_n^0 of the statistics A_n, R_n, and L_n to a simple hypothesis $H \equiv \{\theta_0\}$, obtained by replacing the argument $\tilde{\theta}_n$ by θ_0.

Note that in testing a simple hypothesis the statistic W_n is undefined ((H1) and (H2) cannot be satisfied). In fact, Wald (1943) proposed in such a case A_n^0. Our extension A_n of A_n^0 to composite hypotheses seems to be less practical than W_n or R_n, since it requires both the MLE $\hat{\theta}_n$ and RMLE $\tilde{\theta}_n$. (The same objection applies to L_n.) But we shall see that A_n is useful as an auxiliary mathematical tool. Note also that L_n^0 has been first introduced by Neyman and Pearson (1928).

LEMMA 4. *If the model satisfies* (M1)–(M3) *then the reductions* A_n^0, R_n^0, L_n^0 *of the statistics* A_n, R_n, L_n *to a simple hypothesis* $H \equiv \{\theta_0\}$ *tend, under this hypothesis in law, to* χ_d^2.

Proof. For continuous models see the theorem on page 155 of Serfling (1980). All arguments used in its proof are applicable in our dominated models.

LEMMA 5. *If the model satisfies* (M1)–(M3) *then the statistics* W_n, A_n, R_n, *and* L_n *tend, under any hypothesis H satisfying* (H1)–(H4) *in law, to* $\chi_{d_0}^2$, *where the degrees of freedom are specified in* (H1).

Proof. For continuous models the proofs concerning W_n, L_n, and practically also A_n can be found on page 158 of Serfling (1980). For the proof concerning R_n, Serfling refers to Rao (1973) (see also pp. 321–324 in Cox and Hinkley, 1974). The arguments of these proof are applicable in our dominated models.

Now we can formulate our main results.

THEOREM 1. *Let the model and ϕ satisfy* (M1)–(M3) *and* (Φ1), (Φ2). *Then, under any simple hypothesis $H \equiv \{\theta_0\}$, the ϕ-statistic $D^0_{\phi n}$ defined by* (3) *and the reversed ϕ-statistic*

$$\tilde{D}^0_{\phi n} = 2nD^0_{\phi}(\theta_0, \theta_n)$$

converge in law to χ^2_d.

Proof. Under the considered assumptions the Taylor theorem implies for

$$B(\theta) = D_{\phi}(\theta, \theta_0), \qquad M(\theta) = \nabla' \nabla B(\theta),$$

and arbitrary $\theta \in \Theta$ the relation

$$B(\theta) = B(\theta_0) + \nabla B(\theta_0)(\theta - \theta_0)' + \tfrac{1}{2}(\theta - \theta_0)\, M(\theta^*)(\theta - \theta_0)',$$

where $B(\theta_0) = 0$ and θ^* is on the line joining θ_0 and θ. Therefore,

$$D^0_{\phi n} = 2n\, \nabla B(\theta_0)(\hat{\theta}_n - \theta_0) + n(\hat{\theta}_n - \theta_0)\, M(\theta^*_n)(\hat{\theta} - \theta_0)',$$

where θ^*_n is on the line between $\hat{\theta}_n$ and θ_0. By (Φ2),

$$\nabla B(\theta) = \int_{\mathscr{X}} \nabla f_{\theta} \phi' \left(\frac{f_{\theta}}{f_{\theta_0}} \right) d\mu,$$

where the identity $\phi'(1) = 0$ assumed in (Φ1) implies $\nabla B(\theta_0) = 0$. By (Φ2), the elements of the matrix $M(\theta)$ are continuous in a neighborhood of θ_0 and, by (9), (10), and (Φ1), $M(\theta_0)$ is the Fisher information I_{θ_0} defined in (M3). Since under (M1)–(M3) the matrix function $M(\theta) = I_{\theta}$ is continuous in θ, (15) implies that the difference $M(\theta^*_n) - I_{\hat{\theta}_n}$ converges in probability to the zero matrix. Moreover, by (16), $\sqrt{n}\,(\hat{\theta}_n - \theta_0) = O_P(1)$. Therefore

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0)[\, M(\theta^*_n) - I_{\hat{\theta}_n}\,] \sqrt{n}\,(\hat{\theta}_n - \theta_0)' \xrightarrow{P} 0,$$

i.e.,

$$D^0_{\phi n} - A^0_n \xrightarrow{P} 0.$$

This, together with the convergence in law of A_n^0 to χ_d^2, established in Lemma 4 implies $D_{\phi n}^0 \xrightarrow{L} \chi_d^2$. Since A_n^0 is symmetric in the variables $\hat{\theta}_n$ and θ_0, $\tilde{D}_{\phi n}^0 \xrightarrow{L} \chi_d^2$ holds too.

THEOREM 2. *Let the model and ϕ satisfy* (M1)–(M3) *and* (Φ1), (Φ2). *Then under any hypothesis H with the properties* (H1)–(H4), *the generalized ϕ-statistic $D_{\phi n}$ defined by* (6) *and the reversed generalized ϕ-statistic*

$$\tilde{D}_{\phi n} = 2nD_\phi(\tilde{\theta}_n, \hat{\theta}_n)$$

converge in law to $\chi_{d_0}^2$.

Proof. Similarly as in the previous proof, if $\tilde{\theta} \in \Theta$ then the Taylor theorem implies for

$$B(\theta) = D_\phi(\theta, \tilde{\theta}), \qquad M(\theta) = \nabla^t \nabla B(\theta),$$

and arbitrary $\theta \in \Theta$, the relation

$$B(\theta) = B(\tilde{\theta}) + \nabla B(\tilde{\theta})(\theta - \tilde{\theta})^t + \tfrac{1}{2}(\theta - \tilde{\theta})\, M(\theta^*)(\theta - \tilde{\theta})^t,$$

where $B(\tilde{\theta}) = \nabla B(\tilde{\theta}) = 0$ and θ^* is on the line between $\tilde{\theta}$ and θ. Therefore,

$$D_{\phi n} = n(\hat{\theta}_n - \tilde{\theta}_n)\, M(\theta_n^*)(\hat{\theta}_n - \tilde{\theta}_n)^t,$$

where θ_n^* is on the line between $\tilde{\theta}_n$ and $\hat{\theta}_n$. Similarly as in the previous proof, (15) and (17) imply under any $H_0 \equiv \{\theta_0\} \subset \Theta_0$,

$$M(\theta_n^*) \xrightarrow{P} I_{\theta_0}, \qquad I_{\tilde{\theta}_n} \xrightarrow{P} I_{\theta_0}.$$

Therefore, under H we have the asymptotic matrix relation

$$M(\theta_n^*) - I_{\hat{\theta}_n} \xrightarrow{P} 0.$$

By taking into account (16) and (18) one obtains from here and from the definition of the auxiliary statistic:

$$D_{\phi n} - A_n \xrightarrow{P} 0 \qquad \text{under } H.$$

Since A_n is symmetric in the variables $\hat{\theta}_n$ and $\tilde{\theta}_n$, this implies also

$$\tilde{D}_{\phi n} - A_n \xrightarrow{P} 0 \qquad \text{under } H.$$

The desired results follow from here and from the convergence

$$A_n \xrightarrow{L} \chi^2_{d_0} \qquad \text{under } H$$

in Lemma 5.

COROLLARY 1. *If the model satisfies* (M1)–(M3) *then, under any hypothesis H satisfying* (H1)–(H4), *the Kullback statistic*

$$K_n = 2nK(\hat{\theta}_n, \tilde{\theta}_n)$$

and the reversed Kullback statistic

$$\tilde{K}_n = 2nK(\tilde{\theta}_n, \hat{\theta}_n)$$

fulfil the asymptotic relations

$$K_n \xrightarrow{L} \chi^2_{d_0} \qquad \tilde{K}_n \xrightarrow{L} \chi^2_{d_0}.$$

Proof. Clearly follows from Lemma 2 and Theorem 1.

COROLLARY 2. *If the model is natural exponential as in* (7), *then the assertion of Corollary* 1 *remains true.*

Proof. Clearly follows from Lemma 3 and Theorem 1.

Lemma 5, together with Theorem 2 and its corollaries establish an asymptotic equivalence between the classical Wald's statistic W_n, auxiliary statistic A_n, Rao's statistic R_n, the GLRT statistic L_n, and the ϕ-statistics $D_{\phi n}$ and $\tilde{D}_{\phi n}$. The following result restates this in a slightly stronger form.

THEOREM 3. *Let the model satisfy* (M1)–(M3) *and let* Φ *be the class of all functions satisfying* (Φ1) *and* (Φ2). *Then, under a hypothesis H satisfying* (H1)–(H4), *the difference between any two statistics from the class*

$$\mathscr{C} = \{W_n, A_n, R_n, L_n\} \cup \{D_{\phi n} : \phi \in \Phi\} \cup \{\tilde{D}_{\phi n} : \phi \in \Phi\} \tag{19}$$

tends in probability to zero.

Proof. The fact that the difference between any pair of statistics from the first subclass $\{W_n, A_n, R_n, L_n\}$ tends in probability to zero follows from the arguments presented in the proof of Lemma 5 and their obvious modifications. In the proof of Theorem 1 we proved that the differences $D_{\phi n} - A_n$ and $\tilde{D}_{\phi n} - A_n$ tend in probability to zero for all ϕ satisfying (Φ1) and (Φ2). The desired assertion follows from here by the symmetry and transitivity of the relation $\xrightarrow{P}$.

Remark 1. In particular models $\langle (\mathcal{X}, \mathcal{A}), \{f_\theta : \theta \in \Theta\}, \mu \rangle$ one can find more intimate relations between statistics from the class (19) than those specified by Theorems 1–3. Consider, e.g., the natural exponential model (7) and an arbitrary hypothesis $H \equiv \Theta_0 \subset \Theta$. Then, by using the formulas following (7) and the two formulas presented in the proof of Lemma 3, one obtains from (2) and (6) the Kullback statistic,

$$K_n = 2nK(\hat{\theta}_n, \tilde{\theta}_n) = 2n \int_{\mathcal{X}} f_{\hat{\theta}_n} \log \frac{f_{\hat{\theta}_n}}{f_{\tilde{\theta}_n}} \, d\mu$$

$$= 2n \int_{\mathcal{X}} f_{\hat{\theta}_n} [\, T_n(\hat{\theta}_n - \tilde{\theta}_n)^t + b(\tilde{\theta}_n) - b(\hat{\theta}_n)\,] \, d\mu$$

$$= 2n[\, \tau(\hat{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n)^t + b(\tilde{\theta}_n) - b(\hat{\theta}_n)],$$

and from (16) the GLRT statistic,

$$L_n = 2n[\, \lambda_n(\hat{\theta}_n) - \lambda_n(\tilde{\theta})\,] = 2n[\, T_n(\hat{\theta}_n - \tilde{\theta})^t + b(\tilde{\theta}_n) - b(\hat{\theta}_n)],$$

where $\hat{\theta}_n$ is the MLE and $\tilde{\theta}_n$ is the RMLE. Since the MLE is given by the formula $\hat{\theta}_n = \tau^{-1}(T_n)$, we see that $K_n = L_n$. Further, by Proposition 1.5 in Brown (1986), an arbitrary exponential model which is not over-parametrized can be transformed in a one–one manner to the natural form considered in (7). Consequently in such a model the Kullback and GLRT statistics for testing any hypothesis coincide.

Remark 2. In some models it might be more convenient to consider the functions

$$D_\phi^\psi(\theta, \theta_0) = \psi(D_\phi(\theta, \theta_0))$$

of ϕ-divergences for ϕ satisfying (Φ1), where ψ is a convenient function possibly depending on ϕ. E.g., for the *power divergences* $D_{\phi_a}(\theta, \theta_0)$, defined by the continuous extensions of

$$\phi_a(t) = \frac{t^a - a(t-1) - 1}{a(a-1)} \tag{20}$$

to all a from the closure R of $R - \{0, 1\}$, one can consider the continuous extension of the functions

$$\psi_a(D) = \frac{1}{a(a-1)} \ln(1 + a(a-1)D),$$

$$0 \leqslant D < \begin{cases} \dfrac{1}{a(a-1)}, & \text{if} \quad 0 < a < 1, \\ \infty, & \text{otherwise,} \end{cases}$$

to all a from the closure R of $R-\{0, 1\}$. In this manner one obtains the *Rényi divergences*

$$D_a(\theta, \theta_0) = \frac{1}{a(a-1)} \ln \int_{\mathscr{X}} f_\theta^a f_{\theta_0}^{1-a} \, d\mu \qquad \text{for} \quad a \in R-\{0, 1\}$$

and

$$D_a(\theta, \theta_0) = D_{\phi_a}^{\psi_a}(\theta, \theta_0) \qquad \text{for} \quad a \in \{0, 1\}$$

(see Rényi, 1961, and generalization of his proposal in Liese and Vajda, 1987). Since $\phi_1(t) = t \ln t - t + 1$ and $\phi_0(t) = -\ln t + t - 1$, it holds that

$$D_1(\theta, \theta_0) = K(\theta, \theta_0) \qquad \text{(cf. (2))}$$

and

$$D_0(\theta, \theta_0) = D_1(\theta_0, \theta) = \tilde{K}(\theta, \theta_0) \qquad \text{(cf. (11))}.$$

The last two divergences can also be obtained by the continuous extension of $D_a(\theta, \theta_0)$, $a \in (0, 1)$ to all a from the closure $[0, 1]$ (see Proposition 2.9 in Liese and Vajda, 1987).

For the exponential model (7) the Rényi distances are relatively simple. If $a \in R-\{0, 1\}$ then

$$D_a(\theta, \theta_0) = \begin{cases} \dfrac{1}{a(a-1)} [b(a\theta + (1-a)\,\theta_0) - ab(\theta) - (1-a)\,b(\theta_0)], \\ \qquad \text{if} \quad a\theta + (1-a)\,\theta_0 \in \Theta, \\ \infty, \qquad \text{otherwise}, \end{cases} \tag{21}$$

and

$$D_1(\theta, \theta_0) = (\theta - \theta_0)\,\tau(\theta)^t + b(\theta_0) - b(\theta) \qquad \text{for} \quad \tau(\theta) = \nabla b(\theta). \tag{22}$$

Similarly as in (2) and (6), we define for arbitrary models the *Rényi statistics*

$$D_{an}^0 = 2n\, D_a(\hat{\theta}_n, \theta_0), \qquad D_{an} = 2n\, D_a(\hat{\theta}_n, \tilde{\theta}_n), \qquad a \in R. \tag{23}$$

The class D_{an}, $a \in R$ (similarly for D_{an}^0, $a \in R$), contains as particular cases the Kullback statistics

$$D_{1n} = K_n, \qquad D_{0n} = \tilde{K}_n. \tag{24}$$

in the terminology of discrete multivariate statistic (cf. Section 2.2 in Read and Cressie, 1988), D_{1n} is a *generalized loglikelihood ratio statistics* and D_{0n} is a *modified generalized loglikelihood ratio statistic*. Further,

$$D_{2n} = n \ln(1 + X_n^2/n), \qquad D_{-1n} = n \ln(1 + \tilde{X}_n^2/n)$$

with

$$X_n^2 = n \int_{\mathscr{X}} \frac{(f_{\hat{\theta}_n} - f_{\tilde{\theta}_n})^2}{f_{\tilde{\theta}_n}} d\mu, \qquad \tilde{X}_n^2 = \int_{\mathscr{X}} \frac{(f_{\hat{\theta}_n} - f_{\tilde{\theta}_n})^2}{f_{\hat{\theta}_n}} d\mu \tag{25}$$

are respectively a *generalized Pearson's statistic* and a *modified generalized Pearson's statistic* (cf. *ibid.*). Finally,

$$D_{(1/2)n} = -8n \ln(1 - FT_n/8n)$$

with

$$FT_n = 4n \int_{\mathscr{X}} (\sqrt{f_{\hat{\theta}_n}} - \sqrt{f_{\tilde{\theta}_n}})^2 \, d\mu \tag{26}$$

is a *generalized Freeman–Tukey statistic* (cf. *ibid.*). All these statistics, (24)–(26), are well known in the case that $\mathscr{X}$ is finite and μ is the counting measure on $\mathscr{X}$ (so that the integrals become sums over $\mathscr{X}$), and the null hypothesis is simple, $H \equiv \{\theta_0\}$, so that the RMLE $\tilde{\theta}_n$ is reduced to the constant θ_0. In accordance with the notation employed in this paper, we denote these particular versions by K_n^0, $\tilde{K}_n^0$, $X_n^{2,0}$, $\tilde{X}_n^{2,0}$, and FT_n^0.

In general,

$$D_{an} = \frac{2n}{a(a-1)} \ln[1 + a(a-1) D_{\phi_a n}/2n], \qquad a \in R, \tag{27}$$

where $D_{\phi_a n}$ are the ϕ_a-statistics defined by (6) for the functions (20) (and their limits when $a \in \{0, 1\}$). These statistics are called in the sequel *power divergence statistics*. For finite $\mathscr{X}$ the particular versions $D_{\phi_a}^0$ of these statistics, obtained by reducing $\tilde{\theta}_n$ to a constant $\theta_0 \in \Theta$, have been systematically studied in the book of Read and Cressie (1988).

THEOREM 4. *For any model and hypothesis H, the asymptotic distribution of a Rényi statistic D_{an} or D_{an}^0 exists if and only if the asymptotic distribution of the corresponding power divergence statistic $D_{\phi_a n}$ or $D_{\phi_a n}^0$ exists, and these two distributions then coincide.*

Proof. If $a \in \{0, 1\}$ then $D_{an} = D_{\phi_a n}$ and the assertion is trivial. Let $a \in R - \{0, 1\}$. If $D_{\phi_a n} = O_p(1)$ then

$$\log\left[1 + \frac{a(a-1)}{2n} D_{\phi_a n}\right] = \frac{a(a-1)}{2n} D_{\phi_a n} + o_p(n^{-1})$$

so that

$$D_{an} = D_{\phi_a n} + o_p(1).$$

It follows from here that $D_{\phi_a n} \xrightarrow{L} Y$ for some Y implies $D_{an} \xrightarrow{L} Y$. The reversed implication follows from (26) in a similar way. For the reduced versions D^0_{an} and $D^0_{\phi_a n}$ the argument remains the same.

4. PREFERENCES BETWEEN STATISTICS

The most natural optimality of an α-level (or an asymptotically α-level) test $\mathscr{T}^{\alpha}_{*}$ of a hypothesis $H \equiv \Theta_0$ is that for all alternative values $\theta \in \Theta - \Theta_0$ it maximizes the power

$$\beta_n(\mathscr{T}^{\alpha}; \theta) = P_{\theta, n}(T_n \in C_{\alpha, n})$$

in a sufficiently wide class $\mathbf{T}^{\alpha}$ of α-level tests $\mathscr{T}^{\alpha}$ defined by statistics T^{α}_n depending on $\mathbf{X}_n \sim P_{\theta, n}$ and critical regions $C_{\alpha, n}$ (so that $\sup_{\theta \in \Theta_0} P_{\theta, n}(T_n \in C_{\alpha, n}) \leqslant \alpha$ when the test is nonasymptotically α-level). Unfortunately, it is well known (cf., e.g., Lehmann, 1986) that the uniformly most powerful tests exist only for special hypotheses in special statistical models. As indicated by the examples of the following two sections, the class of tests under consideration has in typical situations no uniformly most powerful members.

A more suitable method for choosing a best test in a given class is based on local alternatives (cf. Section 13 in Chap. 7 of Lehmann, 1986). However, in our case this method is not applicable. Indeed, by using similar arguments as presented in the mentioned Section 13, one obtains that all the tests of present paper achieve on local alternatives the same asymptotic power (they are consistent in the sense of Fraser, 1957).

In this section we present an alternative concept of optimality based on "relative inefficiency" in a given class $\mathbf{T}^{\alpha}$ of tests $\mathscr{T}^{\alpha} \equiv \{T_n; C_{\alpha, n}\}$. By the relative inefficiency $\eta_n(T^{\alpha}; \theta)$ of $\mathscr{T}^{\alpha} \in \mathbf{T}^{\alpha}$ at a given alternative $\theta \in \Theta - \Theta_0$ we mean how far the power of $\mathscr{T}^{\alpha}$ is behind the maximum power achieved in the class, i.e.,

$$\eta_n(\mathscr{T}^{\alpha}; \theta) = \sup_{\mathscr{T} \in \mathbf{T}^{\alpha}} \beta_n(\mathscr{T}; \theta) - \beta_n(\mathscr{T}^{\alpha}; \theta).$$

Optimum represents the test $\mathscr{T}^{\alpha}_{*} \in \mathbf{T}^{\alpha}$ with the minimax relative inefficiency, i.e., with the property

$$\sup_{\theta \in \Theta - \Theta_0} \eta_n(T^{\alpha}_{*}; \theta) = \min_{\mathscr{T}^{\alpha} \in \mathbf{T}^{\alpha}} \sup_{\theta \in \Theta - \Theta_0} \eta_n(\mathscr{T}^{\alpha}; \theta).$$

If the power $\beta_n(\mathscr{T}^{\alpha}; \theta)$ figuring in these formulas cannot be calculated analytically, it can be replaced by the estimate

$$\hat{\beta}_{N;n}(\mathscr{T}^{\alpha}; \theta) = \frac{1}{N} \sum_{j=1}^{N} 1_{C_{\alpha,n}}(T^{j}_{n}),$$

where $C_{\alpha,n}$ is the critical region of $\mathscr{T}^{\alpha}$, $T^{j}_{n} = T_n(\mathbf{X}^{(j)}_{n})$ and $\mathbf{X}^{(1)}_{n}, ..., \mathbf{X}^{(N)}_{n}$ is a sufficiently large number of independent replications of the sample $\mathbf{X}_n = (X_1, ..., X_n)$ with components i.i.d. by the model density f_θ. In this case, of course, $\mathbf{T}^{\alpha}$ must be finite and, if the alternative $\Theta_1 = \Theta - \Theta_0$ is infinite, one has to replace $\sup_{\Theta_1}$ by $\max_{\Theta_1^*}$, where $\Theta_1^* = \Theta_1 \cap \Theta_*$ and Θ_* is a finite subset of Θ "sufficiently dense" in Θ. By the law of large numbers, the estimate $\hat{\beta}_{N;n}(\mathscr{T}^{\alpha}; \theta)$ is consistent in the sense $\hat{\beta}_{N;n}(\mathscr{T}^{\alpha}; \theta) \xrightarrow{P} \beta_n(\mathscr{T}^{\alpha}; \theta)$ as $N \to \infty$.

This concept of optimality is based on the methodology for preferences between various tests developed previously for discrete models (see Read and Cressie, 1988; Menéndez *et al.*, 1995; Morales *et al.*, 1996). Next we describe this methodology in more detail, separately for cases where the distribution of statistics under consideration is unknown and known.

(a) *Unknown Distributions of* $D_{\phi n}$

By using the known asymptotic distribition G of all the statistics $D_{\phi n}$ under a hypothesis $H \equiv \Theta_0 \subset \Theta$, one can construct for any $0 < \alpha < 1$ a system of asymptotically α-level tests $\mathscr{T}^{\alpha}_{\phi}$, $\phi \in \Phi$. Obviously, $\mathscr{T}^{\alpha}_{\phi}$ is rejecting H if and only if $D_{\phi n}$ exceeds the $(1-\alpha)$-quantile $q_{1-\alpha}$ of distribution G.

The evaluation of test $\mathscr{T}^{\alpha}_{\phi}$, best in a given finite set $\Phi_0 \subset \Phi$ of candidates in the sense specified above, requires the selection of a finite set $\Theta^* \subset \Theta$ of parameters reasonably dense in Θ. Then, for every $\theta \in \Theta^*$ one simulates a large number N of independent replications $\mathbf{X}^{(j)}_{n}$ of the sample $\mathbf{X}_n = (X_1, ..., X_n)$ with components i.i.d. by f_θ. For each replication one evaluates the ϕ-divergence statistic $D^{(j)}_{\phi n}$ and then calculates the estimates

$$\hat{\beta}_{N;n}(\mathscr{T}^{\alpha}_{\phi}; \theta) = \frac{1}{N} \sum_{j=1}^{N} 1_{(q_{1-\alpha}, \infty)}(D^{(j)}_{\phi n}) \tag{28}$$

of the size of T_ϕ^α at the hypothesis points $\theta \in \Theta_0^* = \Theta_0 \cap \Theta^*$, and of the power of $\mathscr{T}_\phi^\alpha$ at the alternative points $\theta \in \Theta_1^*$, $\Theta_1^* = (\Theta - \Theta_0) \cap \Theta^*$. By minimizing the maximum observed relative inefficiency

$$\eta_{N;n}(\mathscr{T}_\phi^\alpha) = \max_{\theta \in \Theta_1^*} [\max_{\psi \in \Phi_0} \hat{\beta}_{N;n}(\mathscr{T}_\psi^\alpha; \theta) - \hat{\beta}_{N,n}(\mathscr{T}_\phi^\alpha; \theta)] \tag{29}$$

over Φ_0 one obtains the best test in the stated sense. If for some $\psi \in \Phi_0$ there is detected a "nonnegligible" deviation

$$\delta_{N;n}(\mathscr{T}_\psi^\alpha) = \max_{\theta \in \Theta_0^*} |\hat{\beta}_{N;n}(\mathscr{T}_\psi^\alpha; \theta) - \alpha|$$

from the nominal level α then the corresponding ψ is excluded from the maximization in (29) and also from the subsequent minimization.

Consider $0 < \alpha < 1$ and denote by $\mathscr{T}^\alpha$ any asymptotically α-level test. The test parameters $\hat{\beta}_{N;n}(\mathscr{T}^\alpha; \theta)$, $\eta_n(\mathscr{T}^\alpha)$, $\delta_n(\mathscr{T}^\alpha)$ can be evaluated as described above and, consequently, the defined optimization procedure can be extended also to $\mathscr{T}^\alpha$. This applies in particular to the tests based on the classical statistics like those of Wald and Rao, asymptotically distributed by the same G as the ϕ-divergence statistics. In the next section this procedure is illustrated by two examples.

(b) *Known Distributions of $D_{\phi n}$*

For typical hypotheses in exponential models (binomial, Poisson, modular, normal, etc.) the distributions of estimates $\hat{\theta}_n$ and $\tilde{\theta}_n$ figuring in the Rényi statistics (23) are known. Then the relative inefficiencies of these statistics can be explicitly calculated.

Consider for simplicity a simple hypothesis $H \equiv \{\theta_0\}$ tested against the alternative $H \equiv \Theta - \{\theta_0\}$ in the exponential model (7) by using one of the Rényi statistics $D_{an}^0 = 2n\, D_a(\hat{\theta}_n, \theta_0)$ defined by (21)–(23). If the distribution $Q_{\theta_0,n}$ of $\hat{\theta}_n$ on Θ is known then it obviously suffices to specify subdomains $K_a^\alpha(\theta_0) \subset \Theta$ with the property

$$\hat{\theta}_n \in K_a^\alpha(\theta_0) \qquad \text{if and only if} \quad 2n\, D_a(\hat{\theta}_n, \theta_0) \geqslant \kappa_\alpha, \tag{30}$$

where κ_α is the smallest positive constant for which $Q_{\theta_0,n}(K_a^\alpha(\theta_0)) \leqslant \alpha$. The rejection rule of $\mathscr{T}_a^\alpha$ is $D_{an}^0 \geqslant \kappa_\alpha$ or, equivalently, $\hat{\theta}_n \in K_a^\alpha(\theta_0)$.

The test power under a simple alternative $\{\theta\} \subset \Theta - \{\theta_0\}$ is given by the formula

$$\beta_n(\mathscr{T}_a^\alpha; \theta) = Q_{\theta,n}(K_\alpha(\theta_0)). \tag{31}$$

We can consider finite sets $A \subset R$ specifying similar preselected test candidates as above. By minimizing the maximum relative inefficiency

$$\eta_n(\mathscr{T}_a^{\alpha}) = \max_{\theta \in \Theta - \{\theta_0\}} [\max_{b \in A} \beta_n(\mathscr{T}_b^{\alpha}; \theta) - \beta_n(\mathscr{T}_a^{\alpha}; \theta)] \tag{32}$$

over A one obtains the candidate best from the point of view of power.

Since the Fisher information matrix is in the models under consideration explicitly given in a relatively simple form, one can extend this optimization procedure also to the above considered GLRT, Wald, and Rao tests $\mathscr{T}^{\alpha}$. This optimization procedure is illustrated below by a simple example.

5. EXAMPLE 1: SIMPLE HYPOTHESIS

Let us consider the model with distributions exponential in the narrow sense; i.e., let $\Theta = \mathscr{X} = (0, \infty)$ and let for every θ and x from $(0, \infty)$ the density with respect to the restriction μ of Lebesgue measure on $(0, \infty)$ be

$$f_\theta(x) = \theta e^{-\theta x} = e^{-\theta x + \ln \theta} \qquad \text{(cf. (7))}.$$

By (22) the Rényi distance is for $a \in R - \{0, 1\}$ given by

$$D_a(\theta, \theta_0) = \begin{cases} \dfrac{1}{a(a-1)} [a \ln \theta + (1-a) \ln \theta_0 - \ln(a\theta + (1-a)\theta_0)], \\ \qquad \text{if} \quad a(\theta_0 - \theta) > \theta_0, \\ \infty, \qquad \text{otherwise}, \end{cases} \tag{33}$$

and for the remaining a's by

$$D_1(\theta, \theta_0) = \frac{\theta_0}{\theta} - 1 - \ln \frac{\theta_0}{\theta} = D_0(\theta_0, \theta). \tag{34}$$

Let for a fixed $\theta_0 > 0$ the simple null hypothesis $H \equiv \{\theta_0\}$ be tested against the composite alternative $\bar{H} \equiv (0, \theta_0) \cup (\theta_0, \infty)$ by using the Rényi statistics,

$$D_{an}^0 = 2n\, D_a(\hat{\theta}_n, \theta_0) = 2n\, D_a(1/\bar{\mathbf{X}}_n, \theta_0),$$

where

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the sample mean and $\hat{\theta}_n = 1/\bar{\mathbf{X}}_n$ is the MLE. Thus for a different from 0 and 1 one obtains from (33)

$$D^0_{an} = \begin{cases} -2n\left[\frac{1}{a}\ln(\theta_0\bar{\mathbf{X}}_n) + \frac{1}{a(a-1)}\ln[a+\theta_0\bar{\mathbf{X}}_n(1-a)]\right], \\ \qquad \text{if} \quad a(\theta_0\bar{\mathbf{X}}_n - 1) > \theta_0\bar{\mathbf{X}}_n \\ \infty, \qquad \text{otherwise,} \end{cases}$$

and from (34)

$$D^0_{1n} = K^0_n = 2n[\,\theta_0\bar{\mathbf{X}}_n - 1 - \ln(\theta_0\,\bar{\mathbf{X}}_n)\,].$$

As is easy to verify in statistical textbooks,

$$2n\theta_0\bar{\mathbf{X}}_n \equiv \chi^2_{2n}; \qquad \text{i.e.,} \quad \frac{dP_{\theta_0, n}}{dx} = \frac{1}{2^n(n-1)!}\,x^{n-1}e^{-x/2}1_{(0,\,\infty)}(x), \quad (35)$$

where $P_{\theta_0, n}$ denotes the distribution of

$$Y_n = 2n\theta_0\bar{\mathbf{X}}_n. \quad (36)$$

In other words Y_n is χ^2-distributed with $2n$ degrees of freedom, $Y_n \sim \chi^2_{2n}$.

We shall illustrate the choice of the best test or test statistics in case (b) of Section 4 for the hypothesis with $\theta_0 = 1$ in the class of selected tests $\{\mathscr{T}^{\alpha}_{1/2}, \mathscr{T}^{\alpha}_1, \mathscr{T}^{\alpha}_2\}$ with $\alpha = 0.05$. Here $\mathscr{T}^{\alpha}_a$ denotes the exact α-level or test using the Rényi statistics D^0_{an}. Therefore $\mathscr{T}^{\alpha}_1$ represents the classical α-level GLRT. The set of alternative parameters is $\Theta_1 = (0, 1) \cup (1, \infty)$.

Let us first consider the case $a = \frac{1}{2}$. It follows from (33) that $D_{1/2}(\theta, 1) \geqslant \kappa$ for some real κ (cf. (30)) if and only if for some $\gamma \leqslant \infty$

$$\psi(\theta) \leqslant \gamma, \qquad \text{where} \quad \psi(\theta) = \sqrt{\theta} + (1/\sqrt{\theta}). \quad (37)$$

Since the function $\sqrt{\theta} + 1/\sqrt{\theta}$ is decreasing on $(0, 1)$ and increasing on $(1, \infty)$, the set $K^{\alpha}_{1/2}(1)$ defined in accordance with (30) is of the form $(0, a_n) \cup (b_n, \infty)$, where $0 < a_n < 1 < b_n < \infty$. Since $\hat{\theta}_n = 1/\bar{\mathbf{X}}_n$, the constants $c_n = 1/b_n$ and $d_n = 1/a_n$ must satisfy the condition that the probability of the event $\bar{\mathbf{X}}_n \in (c_n, d_n)$ under $H \equiv \{1\}$ is $1 - \alpha$. Hence,

$$P_{1, n}(\,Y_n \in (c^*_n, d^*_n)) = 1 - \alpha, \quad (38)$$

where

$$c^*_n = 2nc_n, \qquad d^*_n = 2nd_n$$

and $P_{1,n}$, Y_n are defined by (35), (36) for $\theta_0 = 1$. Further, (37) implies $\psi(c_n) = \psi(d_n) = \gamma$, i.e.,

$$\sqrt{c_n^*} + \frac{2n}{\sqrt{c_n^*}} = \sqrt{d_n^*} + \frac{2n}{\sqrt{d_n^*}}. \tag{39}$$

Equation (38) is of the form

$$F_{\chi^2_{2n}}(d_n^*) - F_{\chi^2_{2n}}(c_n^*) = 1 - \alpha.$$

Obviously, for every n and $0 < \alpha < 1$ there exist unique solutions $0 < c_n^* < 2n < d_n^* < \infty$ of the last two equations and these solutions can easily be obtained by means of programs for evaluating distribution functions $F_{\chi^2_k}$ of random variables χ^2_k.

The test $\mathscr{T}^{\alpha}_{1/2}$ rejects the hypothesis $H \equiv \{1\}$ if and only if $2n\bar{\mathbf{X}}_n \notin (c_n^*, d_n^*)$. The power $\beta_n(\mathscr{T}^{\alpha}_{1/2}; \theta)$ of this test considered in (31) is given for every $\theta \neq 1$ by

$$\begin{aligned}
\beta_n(\mathscr{T}^{\alpha}_{1/2}; \theta) &= 1 - P_{\theta,n}(2n\bar{\mathbf{X}}_n \in (c_n^*, d_n^*)) \qquad \text{(cf. (35))} \\
&= 1 - P_{\theta,n}(2n\bar{\mathbf{X}}_n \in (\theta c_n^*, \theta d_n^*)) \\
&= 1 - \frac{1}{2^n(n-1)!} \int_{\theta c_n^*}^{\theta d_n^*} x^{n-1} e^{-x/2}\, dx.
\end{aligned} \tag{40}$$

The *last column of* Table II below presents the powers of the test $\mathscr{T}^{0.05}_{1/2}$ when the sample size $n = 10$, i.e., the values of integrals

$$\beta_{10}(\mathscr{T}^{0.05}_{1/2}; \theta) = 1 - \frac{1}{2^{10} \times 9!} \int_{\theta c_{10}^*}^{\theta d_{10}^*} x^9 e^{-x/2}\, dx \tag{41}$$

for the listed values of θ. Equations (38), (39) for c_{10}^* and d_{10}^* take on the form

$$\frac{1}{2^{10} \times 9!} \int_{c_{10}^*}^{d_{10}^*} x^9 e^{-x/2}\, dx = 0.95,$$

$$\sqrt{c_{10}^*} + \frac{20}{\sqrt{c_{10}^*}} = \sqrt{d_{10}^*} + \frac{20}{\sqrt{d_{10}^*}}.$$

Numerical solutions of these equations are

$$c_{10}^* = 10.4854, \qquad d_{10}^* = 38.1483.$$

These solutions, as well as the integrals (41) presented in Table II, were calculated by means of the instruction "NIntegrate" of the software package MATHEMATICA, with the inaccuracy below 10^{-4}.

Similarly one easily obtains that the test $\mathscr{T}_2^{\alpha}$ rejects $H \equiv \{1\}$ if and only if $2n\bar{\mathbf{X}}_n \notin (c_n^*, d_n^*)$, where $0 < c_n^* < d_n^* < \infty$ are solutions of Eq. (38) and

$$c_n^{*2} - \frac{4n}{c_n^*} = d_n^{*2} - \frac{4n}{d_n^*}. \tag{42}$$

The power of $\mathscr{T}_2^{\alpha}$ is given for these new solutions by the former formula (40). For $\alpha = 0.05$ and $n = 10$ we obtained from (38) and (42) the solutions

$$c_{10}^* = 7.9159, \qquad d_{10}^* = 32.0841.$$

The power function $\beta_{10}(\mathscr{T}_2^{0.05}; \theta)$ is given by the right-hand side of (41) for these c_n^* and d_n^*. Its values are presented in the *first column* of Table II.

TABLE I

Sizes, Powers, and Maximum Relative Inefficiencies of $\mathscr{T}_{\alpha}^{0.05}$ for $n = 100$

	$\beta_{100}(\mathscr{T}_a^{0.05}; \theta)$		
θ	$a = 2$	$a = 1$	$a = 1/2$
0.6	0.9991	0.9987	0.9984
0.65	0.9917	0.9888	0.9870
0.7	0.9561	0.9441	0.9369
0.75	0.8529	0.8237	0.8073
0.8	0.6629	0.6172	**0.5930**
0.85	0.4277	**0.3796**	0.3555
0.9	0.2241	0.1887	0.1720
0.95	0.0990	0.0816	0.0741
1	*0.0516*	*0.0502*	*0.0507*
1.05	0.0643	0.0772	0.0849
1.1	0.1274	0.1549	0.1699
1.15	0.2368	0.2783	0.2999
1.2	0.3822	0.4329	0.4583
1.25	**0.5424**	0.5945	0.6195
1.3	0.6922	0.7378	0.7588
1.35	0.8126	0.8471	0.8623
1.4	0.8969	0.9195	0.9292
1.45	0.9486	0.9617	0.9671
1.5	0.9768	0.9835	0.9861
$\eta_{100}(\mathscr{T}_a^{0.05})$	0.0771	0.0481	0.0699

TABLE II

Sizes, Powers, and Maximum Relative Inefficiencies of $\mathscr{T}_a^{0.05}$ for $n = 10$

	$\beta_{10}(\mathscr{T}_a^{0.05}; \theta)$		
θ	$a = 2$	$a = 1$	$a = 1/2$
0.2	0.9983	0.9965	0.9940
0.4	0.8844	0.8259	0.7615
0.6	0.5058	**0.3902**	**0.2958**
0.8	0.1787	0.1131	0.0729
1	*0.0500*	*0.0500*	*0.0500*
1.2	0.0313	0.0847	0.1063
1.4	0.0571	0.1668	0.2056
1.6	0.1089	0.2792	0.3326
1.8	0.1823	0.4076	0.4700
2.0	0.2730	0.5368	0.6011
2.2	**0.3741**	0.6544	0.7145
2.4	0.4780	0.7532	0.8048
2.6	0.5779	0.8305	0.8719
2.8	0.6683	0.8878	0.9191
3.0	0.7464	0.9281	0.9506
3.2	0.8109	0.9552	0.9707
3.4	0.8623	0.9729	0.9831
3.6	0.9019	0.9840	0.9905
3.8	0.9314	0.9907	0.9948
$\eta_{10}(\mathscr{T}_a^{0.05})$	0.3404	0.1156	0.2100

Finally, the GLRT $\mathscr{T}_1^{\alpha}$ rejects $H \equiv \{1\}$ if and only if $2n\bar{\mathbf{X}}_n \notin (c_n^*, d_n^*)$, where $0 < c_n^* < d_n^* < \infty$ are solutions of Eq. (38) and

$$c_n^* - 2n \ln d_n^* = d_n^* - 2n \ln d_n^*. \tag{43}$$

For $\alpha = 0.05$ and $n = 10$ we obtained

$$c_{10}^* = 9.9579, \qquad d_{10}^* = 35.2267.$$

The power function $\beta_{10}(\mathscr{T}_1^{0.05}; \theta)$ is given by the right-hand side of (41) for the present c_{10}^* and d_{10}^*. Its values are tabulated in the *middle column* of Table II.

It follows from Table II that none of the tests under consideration is uniformly most powerful. In each column the bold figures indicate the alternative θ attaining the maximum relative inefficiency. The value of this inefficiency, defined by (32), is presented in the *last row* of the table. We see that the best in the exact nonasymptotic sense of Section 4 is the GLRT $\mathscr{T}_1^{0.05}$, for which the maximum relative inefficiency is 0.1156. Note that

TABLE III

Solutions c^*_{100} and d_{100^*} for $\mathscr{T}_a^{0.05}$

	$a=2$	$a=1$	$a=1/2$
c^*_{100}	161.1715	163.3164	164.3750
d^*_{100}	238.8285	241.8040	243.3460

$\mathscr{T}_1^{0.05}$ is known to be optimum also in a more fundamental sense. Namely, according to Section 2 in Chapter 4 of Lehmann (1986), $\mathscr{T}_1^{0.05}$ is uniformly most powerful in the class of all unbiased tests $\mathscr{T}^{0.05}$. It follows from Table II that $\mathscr{T}_2^{0.05}$ and $\mathscr{T}_{1/2}^{0.05}$ are biased.

This example can also be used to illustrate the approach presented in part (a) of Section 4. The asymptotic distribution G of all statistics D^0_{an} is χ^2_1, so that the 0.95-quantile is $q_{0.95}=3.842$. By the law of large numbers, for large numbers N of replications the estimates $\hat{\beta}_{N;n}(\mathscr{T}_a^{\alpha};\theta)$ will be close to the values $\beta_n(\mathscr{T}_a^{\alpha};\theta)$. These can be explictly evaluated for any n in a similar way as done the above for $n=10$. For $n=100$ the solutions c^*_{100} and d^*_{100} of the above considered equations are in Table III. The sizes and powers defined by (40) can be found in Table I.

We see from Table I that the difference $\delta_n(\mathscr{T}_a^{0.05})=|\beta_{100}(\mathscr{T}_a^{0.05};1)-0.05|$ are small, i.e., that all the test sizes $\beta_{100}(\mathscr{T}_a^{0.05};1)$ are well adapted to the designed level $\alpha=0.05$. We also see that the differences between powers $\beta_{100}(\mathscr{T}_a^{0.05};\theta)$ for various values of a are less sharp than in the case of $\beta_{10}(\mathscr{T}_a^{0.05};\theta)$. Similarly as for $n=10$, none of the tests $\mathscr{T}_1^{0.05}$, $\mathscr{T}_1^{0.05}$, $\mathscr{T}_{1/2}^{0.05}$ is uniformly most powerful. In each column the bold figures indicate the argument θ of the maximum relative inefficiency defined by (32). The maximum inefficiencies $\eta_{100}(\mathscr{T}_a^{0.05})$ can be found in the last row of Table I. Since these inefficiencies are limiting values of $\eta_{N;100}(\mathscr{T}_a^{0.05})$ defined by (29) for $N\to\infty$, one can expect for large N the minimax inefficiency at the GLRT $\mathscr{T}_1^{0.05}$. One can thus expect that for large N the approach of part (a) of Section 4 will lead to the same optimum test $\mathscr{T}_1^{0.05}$ as the approach of part (b).

6. EXAMPLE 2: COMPOSITE HYPOTHESIS

Let us consider the normal model with $\mathscr{X}=R$ and parameters $\theta=(\mu,\sigma)$, where $(\mu,\sigma)\in\Theta=(-\infty,+\infty)\times(0,\infty)$. We shall test the hypothesis

$$H_0\equiv\Theta_0=\{(\mu,\mu/3):\mu>0\}\qquad\text{versus}\quad H_1\equiv\Theta_1\equiv\Theta-\Theta_0.$$

Here we have the well-known MLE,

$$\hat{\theta}_n \equiv \left(\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \hat{\sigma}_n = \left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \right]^{1/2} \right)$$

and the Θ_0-restricted RMLE (cf. Morales *et al.*, 1996)

$$\tilde{\theta}_n = (\tilde{\mu}_n, \tilde{\mu}_n/3) \qquad \text{for} \quad \tilde{\mu}_n = \tfrac{3}{2}(-3\hat{\mu}_n + \sqrt{13\hat{\mu}_n^2 + 4\hat{\sigma}_n^2}).$$

As shown in Morales *et al.*, 1996, we have in this case

$$A_n = \frac{n}{\hat{\sigma}_n^2} [(\hat{\mu}_n - \tilde{\mu}_n)^2 + 2(\hat{\sigma}_n^2 - \tilde{\mu}_n/3)^2] \qquad \text{(Auxiliary)}$$

$$W_n = \frac{2n(\hat{\mu}_n - 3\hat{\sigma}_n)^2}{11\hat{\sigma}_n^2} \qquad \text{(Wald)}$$

$$R_n = \frac{9n(\hat{\mu}_n - \tilde{\mu}_n)^2}{\tilde{\mu}_n^2} + \frac{n}{2}\left[8 + \frac{9}{\tilde{\mu}^2}(\hat{\sigma}_n^2 + \hat{\mu}_n^2 - 2\hat{\mu}_n\tilde{\mu}_n) \right]^2 \qquad \text{(Rao)}$$

$$L_n \equiv D_{1n} = \frac{9n[(\hat{\mu}_n - \tilde{\mu}_n)^2 + \hat{\sigma}_n^2]}{\tilde{\mu}_n^2} - n + n \ln \frac{\tilde{\mu}_n^2}{9\hat{\sigma}_n^2} \qquad \text{(GLRT)}$$

$$D_{0n} = \frac{n[9(\hat{\mu}_n - \tilde{\mu}_n)^2 + \tilde{\mu}_n^2]}{9\hat{\sigma}_n^2} - n + n \ln \frac{9\tilde{\sigma}_n^2}{\tilde{\mu}_n^2} \qquad \begin{pmatrix} \text{reversed} \\ \text{GLRT} \end{pmatrix}$$

$$D_{an} = \begin{cases} \infty \\ \dfrac{n}{a(a-1)} \ln \dfrac{\hat{\sigma}_n^{2(1-a)}(\tilde{\mu}_n/3)^{2a}}{a(\tilde{\mu}_n/3)^2 + (1-a)\,\hat{\sigma}_n^2} + \dfrac{n(\hat{\mu}_n - \tilde{\mu}_n)^2}{a(\tilde{\mu}_n/3)^2 + (1-a)\,\hat{\sigma}_n^2} \end{cases} \qquad \text{(Rényi)},$$

where the last formula holds for $a \neq 0$, $a \neq 1$, and the first possibility in this formula takes place when $a(\tilde{\mu}_n/3)^2 + (1-a)\,\hat{\sigma}_n^2 \leqslant 0$.

Let us consider an infinite parameter subspace $\Theta^* \subset \Theta$,

$$\Theta^* = \{(\mu; c_j\mu): \mu > 0, 1 \leqslant j \leqslant J\},$$

consisting of half-lines specified by auxiliary constants $0 < c_1 < \cdots < c_J < \infty$, where

$$c_{j_0} = \tfrac{1}{3}$$

for some $1 < j_0 < J$. Then $\Theta_0 \subset \Theta^*$, so that the sets considered in part (a) of Section 4 satisfy the relations

$$\Theta_0^* = \Theta_0 \cap \Theta^* = \Theta_0$$

and

$$\Theta_1^* \equiv \Theta_1 \cap \Theta^*.$$

Consider finally $0 < \alpha < 1$ and denoty by $\mathbf{T}^\alpha$ the set of all asymptotically α-level tests defined in accordance with Theorem 3 by the above-considered statistics A_n, ..., D_{an} for rational a of the form $k/10$, where k is an integer.

It is easy to see that under any parameter $\theta_0 = (\mu_0, c_0\mu_0) \in \Theta$, the distributions of the above listed statistics A_n, ..., D_{an} depend on $c_0 > 0$ and not on $\mu_0 > 0$. Hence, for any $\mathscr{T}^\alpha \in \mathbf{T}^\alpha$ and $\mu > 0$ the estimates

$$\hat{\beta}_{N;n}(\mathscr{T}^\alpha; (\mu, c\mu)), \qquad \eta_{N;n}(\mathscr{T}^\alpha), \qquad \text{and} \qquad \delta_{N;n}(\mathscr{T}^\alpha),$$

defined by (28), (29), and the formula next to (29), can be obtained from replications $\mathbf{X}_n^{(1)}$, ..., $\mathbf{X}_n^{(N)}$, generated by the normal model under consideration with the location and scale parameters $\theta_0 = (1, c)$.

In Table IV we present the estimates obtained for $N = 1000$ replications of data of sample size 50 and for $J = 11$ auxiliary constants with $j_0 = 5$. The independent data distributed by $N(1, c)$ for the values of c given in Table III were obtained by calling the "RANDOM" procedure in the IMSL package and a subsequent application of the Box–Muller subroutine. We consider the class $\mathbf{T}^{0.05}$ of tests defined by the statistics $T_{50} \in \{A_{50}, ..., D_{a;50}\}$ and the critical value 3.84 which is the quantile $q_{0.95}$ of χ_1^2.

Table IV presents estimates $\hat{\beta}_{1000;50}(\mathscr{T}^{0.05}; (\mu; c\mu))$ and $\eta_{1000;50}(\mathscr{T}^{0.05})$ obtained from the simulated replications (so that the values $\hat{\alpha} = \hat{\beta}_{1000;50}(\mathscr{T}^{0.05}; (\mu, \mu/3))$ printed in italics estimate the actual test size achieved at the sample size $n = 50$). In 14 columns there are results concerning 14 statistics selected for the table (for statistics $D_{k/10;n}$, which were not selected, the estimates can be obtained by extrapolating in the table). Bold figures indicate the half-lines in the parameter space (auxiliary parameters c) at which is achieved the maximum relative inefficiency $\eta_{1000;50}(\mathscr{T}^{0.05})$, presented in the last line.

We see from the last line that the test using the Hellinger distance statistics $T_n = D_{0.5;50}$ is attaining the minimax relative inefficiency. The relative inefficiency of the test using, e.g., the Rao statistic $T_n = R_{50}$ is more than 100% higher. We also see that the deviations

$$\delta_{1000;50}(\mathscr{T}^{0.05}) = |\hat{\beta}_{1000;50}(\mathscr{T}^{0.05}; (\mu, \mu/3)) - 0.05|$$

for $\mathscr{T}^{0.05} \equiv T_n \in \{W_{50}, A_{50}, \text{and all } D_{a;50} \text{ with } a \leqslant 0\}$ exceed by more than 20% the nominal test size $\alpha = 0.05$ (in the case of generalized modified Pearson statistics $D_{-1;50}$ the deviation is almost 100%). The exclusion of the corresponding tests by the method described in part (a) of Section 4 has, however, no impact on the established optimality of the Hellinger distance test.

TABLE IV

Estimates $\hat{\beta}_{1000;\,50}(\mathscr{T}^{0.05};(\mu, c\mu)) \equiv \hat{\beta}(T_{50};(\mu, c\mu))$ Obtained for the Tests $\mathscr{T}^{0.05} \in \mathbf{T}^{0.05}$ Based on the 14 Listed Statistics T_{50}, for Auxiliary Constants $c = c_j = (17+3j)/(101-j)$, $1 \leqslant j \leqslant 11$, and for All $\mu > 0$

Auxiliary contants c	Statistics T_{50}													
	W_{50}	R_{50}	A_{50}	$D_{-1;\,50}$	$D_{-0.6;\,50}$	$D_{-0.3;\,50}$	$D_{0;\,50}$	$D_{0.3;\,50}$	$D_{0.5;\,50}$	$D_{0.7;\,50}$	$D_{1;\,50}$	$D_{1.3;\,50}$	$D_{1.6;\,50}$	$D_{2;\,50}$
$10/50 = 0.200$	1.00	0.997	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.999	0.997	0.997
$23/99 = 0.232$	0.965	0.853	0.965	0.979	0.975	0.970	0.960	0.951	0.946	0.936	0.927	0.913	0.894	0.880
$13/49 = 0.266$	0.675	**0.431**	0.677	0.742	0.710	0.689	0.664	0.640	0.615	**0.602**	**0.573**	**0.540**	**0.516**	**0.477**
$29/97 = 0.299$	0.252	0.099	0.253	0.312	0.280	0.260	0.238	0.215	0.205	0.191	0.176	0.159	0.147	0.126
$10/30 = 0.333$	*0.068*	*0.044*	*0.068*	*0.095*	*0.078*	*0.074*	*0.067*	*0.059*	*0.056*	*0.054*	*0.055*	*0.052*	*0.049*	*0.050*
$07/18 = 0.389$	0.083	0.194	0.084	0.067	0.069	0.082	0.095	0.109	0.119	0.133	0.147	0.162	0.181	0.214
$20/47 = 0.426$	**0.301**	0.477	**0.303**	**0.222**	**0.264**	**0.294**	**0.326**	**0.349**	**0.363**	0.377	0.404	0.437	0.463	0.496
$41/93 = 0.441$	0.544	0.715	0.545	0.459	0.504	0.536	0.563	0.596	0.620	0.663	0.659	0.677	0.701	0.730
$11/23 = 0.478$	0.811	0.907	0.813	0.754	0.786	0.806	0.825	0.841	0.850	0.861	0.874	0.888	0.897	0.915
$47/91 = 0.516$	0.910	0.962	0.910	0.872	0.895	0.907	0.915	0.931	0.933	0.943	0.950	0.958	0.960	0.966
$50/90 = 0.555$	0.972	0.989	0.972	0.964	0.967	0.969	0.974	0.977	0.979	0.980	0.981	0.986	0.989	0.992
$\eta(T_{50})$	0.195	0.311	0.193	0.274	0.232	0.202	0.170	0.147	0.133	0.140	0.169	0.202	0.208	0.226

Note. The last line presents the relative inefficiencies $\eta_{1000;\,50}(\mathscr{T}^{0.05}) \equiv \eta(T_{50})$ for the tests $\mathscr{T}^{0.05}$ based on the listed statistics T_{50}.

REFERENCES

Beran, R. (1977). Minimum Hellinger distance estimator for parametric models. *Ann. Statist.* **5** 455–463.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families*. Lecture Notes, Vol. 9. Institute Mathematical Statistics, Hayward, CA.

Csiszár, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci. Ser. A* **8** 85–108.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman Hall, London.

Fleming, W. H. (1965). *Functions of Several Variables*. Addison–Wesley, New York.

Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. Wiley, New York.

Győrfi, L., Vajda, I., and van der Meulen, E. (1994a). Minimum Hellinger distance point estimates under weak family regularity. *Math. Methods Statist.* **3** 25–45.

Győrfi, L., Vajda, I., and van der Meulen, E. (1994b). Parameter estimation by projecting on structural parametric models. In *Asymptotic Statistics* (P. Mandl and M. Hušková, Eds.), pp. 261–272. Physica Verlag, Vienna.

Kupperman, M. (1957). *Further Applications of Information Theory to Multivariate Analysis and Statistical Inference*. Ph.D. dissertation. George Washington University.

Lehman, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

Liese, F., and Vajda, I. (1987). *Convex Statistical Distances*, Teubner, Leipzig.

Menéndez, M. L., Morales, D., Pardo, L., and Vajda, I. (1995). Divergence based estimation and testing of statistical models of classification. *J. Multivariate Anal.* **54** 329–354.

Menéndez, M. L., Morales, D., Pardo, L., and Vajda, I. (1996). About divergence-based goodnes-of-fit tests in the Dirichlet-multinomial model. *Commun. Statist. Theory Methods* **25** 119–133.

Morales, D., Pardo, L., Salicrú, M., and Menéndez, M. L. (1994). Asymptotic properties of divergence statistics in a stratified random sampling and its applications to test statistical hypotheses. *J. Statist. Planning Infer.* **38** 201–224.

Morales, D., Pardo, L., and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *J. Statist. Planning Infer.* **48** 347–369.

Morales, D., Pardo, L., and Vajda, I. (1996). *About the Behaviour of Rényi's Divergence Statistics Under Null Composite Hypotheses*. Res. Rep. 1870, Inst. of Inform. Theory and Automation, Czech. Acad. Sci., Prague.

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika A* **20** 175–240; 263–294.

Pardo, L., Morales, D., Salicrú, M., and Menéndez, M. L. (1993). The ϕ-divergence statistics in bivariate multinomial applications including stratification. *Metrika* **40** 223–235.

Pardo, M. C. (1994). On testing independence for multidimensional contingency tables with stratified random sampling. *Inform. Sci.* **78** 101–118.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Phil. Soc.* **44** 50–57.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

Read, A., and Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer, New York.

Rényi, A. (1961). On measures of entropy and information. In *Proc. 4th Berkeley Symp. on Probab. Math. Statist.*, Univ. of California, Berkeley, Vol. 1, pp. 547–561.

Salicrú, M., Menéndez, M. L., Morales, D., and Pardo, L. (1994). On the applications of divergence type measures in testing statistical hypotheses. *J. Multivariate Anal.* **51** 372–391.

Sen, P. K., and Singer, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Simpson, D. G. (1989), Hellinger deviance tests: Efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.* **84**, 107–113.

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

Zografos, K., Ferentinos, K., and Papaioannou, T. (1990). ϕ-divergence statistics: sampling properties, multinomial goodness of fit and divergence tests. *Communications in Statistics, Theory and Methods*, **18**, 1785–1802.

Zografos, K. (1993). *Asymptotic properties of ϕ-divergence statistics and its application in contingency tables*. Res. Rep. 185, Math. Dept., University of Ioannina, Ioannina, Greece.

Printed in Belgium