Statistical Papers 46, 397-409 (2005)

Statistical Papers
© Springer-Verlag 2005

Divergence-based estimation and testing with misclassified data[1]

E. Landaburu[1], D. Morales[2], and L. Pardo[1]

[1] Department of Statistics & O. R., Complutense University of Madrid, Madrid
[2] Operations Research Center, Miguel Hernández University of Elche, Elche

Received: January 20, 2003; revised version: August 6, 2003

Abstract

The well-known chi-squared goodness-of-fit test for a multinomial distribution is generally biased when the observations are subject to misclassification. In Pardo and Zografos (2000) the problem was considered using a double sampling scheme and ϕ-divergence test statistics. A new problem appears if the null hypothesis is not simple because it is necessary to give estimators for the unknown parameters. In this paper the minimum ϕ-divergence estimators are considered and some of their properties are established. The proposed ϕ-divergence test statistics are obtained by calculating ϕ-divergences between probability density functions and by replacing parameters by their minimum ϕ-divergence estimators in the derived expressions. Asymptotic distributions of the new test statistics are also obtained. The testing procedure is illustrated with an example.

Key words and phrases: Misclassification; Double sampling; Divergence estimators; Goodness–of–fit tests; Divergence statistics.

AMS Classification: 62F05, 62B10.

1. Introduction

Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables taking on values in $\mathcal{X} = \{1, \ldots, M\}$. Let $p = (p_1, \ldots, p_M)^T$ be a probability vector, i.e. $p \in \Delta_M$ with

$$\Delta_M = \left\{ (p_1, \ldots, p_M) \in R^M : \sum_{j=1}^{M} p_j = 1,\ p_j \geq 0,\ j = 1, \ldots, M \right\}, \tag{1.1}$$

such that $p_j = \mathsf{P}_p(Y_i = j)$, $j = 1, \ldots, M$, $i = 1, \ldots, n$. Define $N_j = \sum_{i=1}^{n} 1_{\{j\}}(Y_i)$ and $\widehat{p}_j = \frac{N_j}{n}$, $j = 1, \ldots, M$, so that $(N_1, \ldots, N_M)$ is a sufficient statistics for $p \in \Delta_M$ and multinomially distributed, i.e.

$$\mathsf{P}_p(N_1 = n_1, \ldots, N_M = n_M) = \frac{n!}{n_1! \cdots n_M!}\, p_1^{n_1} \cdots p_M^{n_M},$$

for every integers $n_1 \geq 0, \ldots, n_M \geq 0$ such that $n_1 + \ldots + n_M = n$.

[1] Supported by the grant BMF2000-0800.

The statistician is often interested in testing $H_0 : p \in \mathcal{P}$ versus $H_1 : p \in \Delta_M - \mathcal{P}$, with

$$\mathcal{P} = \left\{ p(\theta) = (p_1(\theta), \ldots, p_M(\theta))^T \in \Delta_M : \ \theta \in \Theta \right\}, \tag{1.2}$$

$\Theta \subset R^k$ open and $k < M - 1$.

For simplicity we restrict ourselves to unknown true parameters θ_0 satisfying the classical regularity conditions given by Birch (1964):

1. θ_0 is an interior point of Θ.

2. $p_i(\theta_0) > 0$ for $i = 1, \ldots, M$. Thus $p(\theta_0) = (p_1(\theta_0), \ldots, p_M(\theta_0))^T$ is an interior point of the set Δ_M.

3. The mapping $p : \Theta \mapsto \Delta_M$ is totally differentiable at θ_0 so that the partial derivatives of p_i with respect to each θ_j exist at θ_0 and $p_i(\theta)$ has a linear approximation at θ_0 given by

$$p_i(\theta) = p_i(\theta_0) + \sum_{j=1}^{k} (\theta_j - \theta_{0j}) \frac{\partial p_i(\theta_0)}{\partial \theta_j} + o(\|\theta - \theta_0\|),$$

   where $o(\|\theta - \theta_0\|)$ denotes a function verifying

$$\lim_{\theta \to \theta_0} \frac{o(\|\theta - \theta_0\|)}{\|\theta - \theta_0\|} = 0.$$

4. The Jacobian matrix

$$J(\theta_0) = \left(\frac{\partial p(\theta)}{\partial \theta} \right)_{\theta = \theta_0} = \left(\frac{\partial p_i(\theta_0)}{\partial \theta_j} \right)_{\substack{i=1,\ldots,M \\ j=1,\ldots,k}}$$

   is of full rank (i.e. of rank k).

5. The inverse mapping $p^{-1} : \mathcal{P} \mapsto \Theta$ is continuous at $p(\theta_0)$.

6. The mapping $p : \Theta \mapsto \Delta_M$ is continuous at every point $\theta \in \Theta$.

Under the assumption that H_0 is true, there exists an unknown parameter θ_0 such that $p = p(\theta_0)$ and the problem of point estimation appears in a natural way. Maximum likelihood estimator (MLE), $\widehat{\theta}$, is obtained by maximizing

$$\ln \mathrm{P}_{p(\theta)}(N_1 = n_1, \ldots, N_M = n_M),$$

with respect to θ with $\theta \in \Theta$ or, equivalently, by minimizing the Kullback-Leibler divergence

$$D(\widehat{p}, p(\theta)) = \sum_{j=1}^{M} \widehat{p}_j \ln \frac{\widehat{p}_j}{p_j(\theta)},$$

with respect to θ with $\theta \in \Theta$, where $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_M)^T$ and $p(\theta) = (p_1(\theta), \ldots, p_M(\theta))^T$. The χ^2 test statistic

$$n\chi^2(\widehat{p}, p(\theta)) = n\sum_{j=1}^{M} \frac{(\widehat{p}_j - p_j(\widehat{\theta}))^2}{p_j(\widehat{\theta})}, \tag{1.3}$$

which is under 1–6 asymptotically chi-square distributed with $M-k-1$ degrees of freedom, is a well known statistic to test H_0. Morales et al. (1995) extended (1.3) in two directions:

1. They considered minimum ϕ-divergence estimators

$$\widehat{\theta}_\phi = \operatorname{argmin}_{\theta \in \Theta} D_\phi(\widehat{p}, p(\theta)), \tag{1.4}$$

where

$$D_\phi(\widehat{p}, p(\theta)) = \sum_{j=1}^{M} p_j(\theta)\phi\left(\frac{\widehat{p}_j}{p_j(\theta)}\right) \tag{1.5}$$

was introduced by Ciszár (1963) and Ali and Silvey (1966) for every ϕ in the set Φ of real convex functions defined on $[0, \infty)$ and satisfying $\phi(1) = \phi'(1) = 0$. In formula (1.5) if either $p_j(\theta)$ or $p_j(\theta)$ and $\widehat{p}_j$ are zero, expressions $0\phi(u/0)$ and $0\phi(0/0)$ are defined as $\lim_{u\to\infty} \frac{\phi(u)}{u}$ and 0 respectively.

2. They introduced and studied the ϕ-divergence statistics

$$T_{\phi_1,\phi_2} = \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right), \quad \phi_1, \phi_2 \in \Phi, \tag{1.6}$$

for testing $H_0 : p \in \mathcal{P}$ versus $H_1 : p \in \Delta_M - \mathcal{P}$.

>From a statistical point of view the most important family of ϕ-divergences is, perhaps, the power-divergence family introduced by Cressie and Read (1984) and Read and Cressie (1988), which is obtained from (1.5) by taking

$$\begin{aligned} \phi(x) \equiv \psi_\lambda(x) &= \frac{1}{\lambda(\lambda+1)}\left(x^{\lambda+1} - x - \lambda(x-1)\right); \quad \lambda \neq 0, \lambda \neq -1, \\ \psi_0(x) &= \lim_{\lambda\to 0} \psi_\lambda(x) = x\log x - x + 1 \\ \psi_{-1}(x) &= \lim_{\lambda\to -1} \psi_\lambda(x) = -\log x + x - 1. \end{aligned} \tag{1.7}$$

Let us assume that $Y_1, \ldots, Y_n$ give the result of classifying n individuals into M classes, so that $Y_i = j$ if individual i is classified into category j. The above mentioned results rely on the implicit hypothesis that no false classifications are allowed. However this assumption is not very realistic in practice and misclassifications often occurs. In this paper Tenenbein's (1970, 1971, 1972) double sampling schemes are used to provide a basis for the study of problems of estimation and testing composite null hypothesis when misclassification is allowed.

2. Double sampling scheme

Let Y be the random variable giving the true classification and let X be the random variable giving a classification obtained through a possibly fallible device. Let us denote the marginal probabilities of Y and X by

$$p_i = \mathsf{P}(Y = i) \quad \text{and} \quad q_j = \mathsf{P}(X = j), \quad i, j = 1, \ldots, M. \tag{2.1}$$

To describe misclassifications we introduce conditional probabilities

$$q_{j/i} = \mathsf{P}(X = j \,/\, Y = i), \tag{2.2}$$

so that $\sum_{j=1}^{M} q_{j/i} = 1$ and $q_j = \sum_{i=1}^{M} p_i q_{j/i}$. In this situation a double sampling scheme can be described as follows:

i) A sample of n units is drawn at random and the true and fallible classifications, denoted by $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_n$, respectively, are obtained for each unit. For $i, j = 1, \ldots, M$, let n_{ij} be the number of units in the sample whose true category is i and whose fallible category is j, and define $n_{i\cdot} = \sum_{j=1}^{M} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{M} n_{ij}$.

ii) A further sample of $N - n$ units is drawn and the fallible classifications $X_{n+1}, \ldots, X_N$ are obtained for each unit. We denote by

$$m_j = \sum_{\ell=n+1}^{N} 1_{\{X_\ell = j\}}, \quad j = 1, \ldots, M,$$

the number of units whose fallible category is j and by $(m_1, \ldots, m_M)$ the vector of absolute frequencies associated to the random sample $X_{n+1}, \ldots, X_N$.

The joint likelihood function associated to the observed data, $(Y_1, X_1), \ldots, (Y_n, X_n)$, $X_{n+1}, \ldots, X_N$, is

$$L(p, Q) = \left\{ \prod_{i=1}^{M} \prod_{j=1}^{M} (p_i q_{j/i})^{n_{ij}} (q_j - p_i q_{j/i})^{n_{\cdot j} - n_{ij}} \right\} \left\{ \prod_{\ell=1}^{M} q_\ell^{m_\ell} \right\},$$

with parameters $p = (p_1, \ldots, p_M)^T$ and $Q = (q_{j/i})_{M \times M}$, and the MLE are

$$\widehat{p}_i = \sum_{j=1}^{M} \frac{(m_j + n_{\cdot j}) n_{ij}}{N n_{\cdot j}} \quad \text{and} \quad \widehat{q}_{j/i} = \frac{(m_j + n_{\cdot j}) n_{ij}}{N n_{\cdot j} \widehat{p}_i}, \quad i, j = 1, \ldots, M. \tag{2.3}$$

Tenenbein (1972) and Cheng, Hsueh and Chien (1998) proved that

$$\sqrt{N} \left(\widehat{p}_1 - p_1, \ldots, \widehat{p}_M - p_M \right) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \Sigma) \tag{2.4}$$

as $N \to \infty$ and $n/N \to f > 0$, where $\mathbf{0} = (0, \dots, 0)^T$, $\Sigma = (\sigma_{ij})_{M \times M}$,

$$\sigma_{ij} = \begin{cases} \dfrac{p_i(1-p_i)}{f}[1-(1-f)K_i] & \text{if } i = j \\ \left(1 - \dfrac{1}{f}\right) \displaystyle\sum_{\ell=1}^{M} q_\ell \lambda_{i\ell} \lambda_{j\ell} - p_i(1-p_j) & \text{if } i \neq j, \end{cases} \tag{2.5}$$

with $K_i = \frac{p_i}{1-p_i}\left\{\sum_{\ell=1}^{M} \frac{q_{\ell/i}^2}{q_\ell} - 1\right\}$ and $\lambda_{ij} = \frac{p_i q_{j/i}}{q_j}$, $i, j = 1, \dots, M$. Throughout this paper $\xrightarrow{L}$ is used to denote convergence in law.

In Pardo and Zografos (2000) the problem of testing a simple null hypothesis, $H_0 : p = p_0$, was considered on the basis of the test statistic

$$\frac{2n}{\phi''(1)} D_\phi\left(\widehat{p}, p_0\right), \quad \phi \in \Phi,$$

where $\widehat{p}$ is given in (2.3). In the following Section this problem is extended to the case of a composite null hypothesis. In addition, the family of ϕ-divergence estimators (1.4) is adapted to the case of misclassified data and some of their asymptotic properties are obtained.

3. Test statistics based on ϕ-divergences

In this section we assume that the probabilities p_i and q_j, introduced in (2.1), depends on an unknown parameter $\theta \in \Theta \subset R^k$ with Θ open and $k < M - 1$, i.e.

$$p_i(\theta) = \mathsf{P}_\theta(Y = i), \; q_j(\theta) = \mathsf{P}_\theta(X = j), \quad i, j = 1, \dots, M,$$

with $q_j(\theta) = \sum_{j=1}^{M} p_i(\theta) q_{j/i}$, $j = 1, \dots, M$. We also assume that true parameter θ_0 and mapping $p : \Theta \mapsto \Delta_M$ satisfy conditions 1–6 of Birch (1964). We consider the M-vector $p(\theta) = (p_1(\theta), \dots, p_M(\theta))^T$, the $M \times k$ Jacobian matrix $J(\theta) = (J_{j\ell}(\theta))_{j=1,\dots,M,\ell=1,\dots,k}$ with $J_{j\ell}(\theta) = \frac{\partial}{\partial\theta_\ell} p_j(\theta)$, the $M \times k$ matrix $A(\theta) = \operatorname{diag}\left(p(\theta)^{-1/2}\right) J(\theta)$ and the $k \times k$ Fisher information matrix

$$I(\theta) = \left(\sum_{j=1}^{M} \frac{1}{p_j(\theta)} \frac{\partial p_j(\theta)}{\partial \theta_r} \frac{\partial p_j(\theta)}{\partial \theta_s}\right)_{r,s=1,\dots,k} = A(\theta)^T A(\theta),$$

where $\operatorname{diag}\left(p(\theta)^{-1/2}\right) = \operatorname{diag}\left(\frac{1}{\sqrt{p_1(\theta)}}, \dots, \frac{1}{\sqrt{p_M(\theta)}}\right)$.

The above defined matrices are considered at the points $\theta \in \Theta$ where the derivatives exist and all the coordinates $p_j(\theta)$ are positive.

For the observed data $(Y_1, X_1), \dots, (Y_n, X_n), X_{n+1}, \dots, X_N$, the minimum ϕ-divergence estimator is

$$\widehat{\theta}_\phi = \arg\min_{\theta \in \Theta} D_\phi(\widehat{p}, p(\theta)), \quad \phi \in \Phi, \tag{3.1}$$

where $\widehat{p} = (\widehat{p}_1, \dots, \widehat{p}_M)^T$ is given in (2.3). In the following theorem the asymptotic distribution of $\widehat{\theta}_\phi$ is given.

Theorem 3.1. Let $\phi \in \Phi$, let $p : \Theta \mapsto \Delta_M$ be twice continuously differentiable in a neighborhood of θ_0 and assume that conditions 1–6 of Section 1 hold. If $N \rightarrow \infty$ and $n/N \rightarrow f > 0$, then the minimum ϕ-divergence estimator $\widehat{\theta}_\phi$ satisfies the relation

$$\sqrt{N}(\widehat{\theta}_\phi - \theta_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, S(\theta_0)),$$

where $S(\theta_0) = C(\theta_0)\Sigma(\theta_0)C^T(\theta_0)$, $C(\theta_0) = I(\theta_0)^{-1}A(\theta_0)^T \text{diag}(p(\theta_0)^{-1/2})$ and $\Sigma(\theta_0)$ is the $M \times M$ matrix whose elements $\sigma_{ij}(\theta_0)$ are obtained from (2.5) by substituting p_i and q_j by $p_i(\theta_0)$ and $q_j(\theta_0)$ respectively.

Proof. In the same way as in Morales et al (1995), it can be established that

$$\widehat{\theta}_\phi = \theta_0 + I(\theta_0)^{-1}A(\theta_0)^T \text{diag}\left(p(\theta_0)^{-1/2}\right)(\widehat{p} - p(\theta_0))^T + o_p(||\widehat{p} - p(\theta_0||), \qquad (3.2)$$

where $\widehat{p}$ is given in (2.3). Now by (2.4) we have

$$\sqrt{N}(\widehat{\theta}_\phi - \theta_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, S(\theta_0)).$$

Theorem 3.2 is needed to derive the asymptotic distribution of the proposed test statistic for misclassified data.

Theorem 3.2. Under the assumptions of Theorem 3.1, we have

$$\sqrt{N}(\widehat{p} - p(\widehat{\theta}_\phi)) \xrightarrow{L} \mathcal{N}(\mathbf{0}, B(\theta_0)),$$

where

$$\begin{aligned} B(\theta_0) &= \Sigma(\theta_0) - \Sigma(\theta_0)L(\theta_0)^T - L(\theta_0)\Sigma(\theta_0) + L(\theta_0)\Sigma(\theta_0)L(\theta_0)^T \\ L(\theta_0) &= J(\theta_0)I(\theta_0)^{-1}A(\theta_0)^T \text{diag}(p(\theta_0)^{-1/2}). \end{aligned}$$

Proof. A first order Taylor expansion gives

$$p(\widehat{\theta}_\phi) = p(\theta_0) + J(\theta_0)(\widehat{\theta}_\phi - \theta_0)^T + o(||\widehat{\theta}_\phi - \theta_0||). \qquad (3.3)$$

>From (3.2) and (3.3) we obtain

$$p(\widehat{\theta}_\phi) = p(\theta_0) + J(\theta_0)I(\theta_0)^{-1}A(\theta_0)^T \text{diag}(p(\theta_0)^{-1/2})(\widehat{p} - p(\theta_0))^T + o(||\widehat{p} - p(\theta_0)||).$$

Therefore the random vectors

$$\begin{pmatrix} \widehat{p} - p(\theta_0) \\ p(\widehat{\theta}_\phi) - p(\theta_0) \end{pmatrix}_{2M \times 1} \quad \text{and} \quad \begin{pmatrix} I \\ L(\theta_0) \end{pmatrix}_{2M \times M} (\widehat{p} - p(\theta_0))_{M \times 1},$$

where I is the $M \times M$ unity matrix, have the same asymptotic distribution. Furthermore, it is clear that

$$\sqrt{N}(\widehat{p} - p(\theta_0)) \xrightarrow{L} \mathcal{N}(0, \Sigma(\theta_0)).$$

implies

$$\sqrt{N}\begin{pmatrix} \widehat{p}-p(\theta_0) \\ p(\widehat{\theta}_\phi)-p(\theta_0)\end{pmatrix} \xrightarrow{L} \mathcal{N}\left(0, \begin{pmatrix} I \\ L(\theta_0)\end{pmatrix}\Sigma(\theta_0)\left(I, L(\theta_0)^T\right)\right). \tag{3.4}$$

Finally, from (3.4) we get

$$\sqrt{N}(\widehat{p}-p(\widehat{\theta}_\phi)) \xrightarrow{L} \mathcal{N}(\mathbf{0}, B(\theta_0)).$$

Now we consider the problem of testing the hypothesis $H_0 : p \in \mathcal{P}$ given in (1.2). Our proposal is based on the divergence test statistics

$$S_{\phi_1,\phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right), \quad \phi_1, \phi_2 \in \Phi, \tag{3.5}$$

where $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_M)^T$ and $\widehat{\theta}_{\phi_2}$ have been introduced in (2.3) and (3.1) respectively. In the following (see Section 4) $S_{\psi_{\lambda_1},\psi_{\lambda_2}}$ is used to denote ϕ-divergence statistics (3.5) with $\phi_1 \equiv \psi_{\lambda_1}$ (power divergence of order λ_1 for the test statistic) and $\phi_2 \equiv \psi_{\lambda_2}$ (minimum power divergence estimator of order λ_2 for the unknown parameter).

Theorem 3.3. Under the assumptions of Theorem 3.1, we have

$$S_{\phi_1,\phi_2} \xrightarrow{L} \sum_{j=1}^{r} \beta_j(\theta_0) Z_j^2, \quad \phi_1, \phi_2 \in \Phi,$$

where $r = \text{rank}\,(B(\theta_0)\text{diag}(p(\theta_0)^{-1}))B(\theta_0))$, $\beta_1(\theta_0), \ldots, \beta_r(\theta_0)$ are the non null eigenvalues of the matrix $\text{diag}\,(p(\theta_0)^{-1})\, B(\theta_0)$ and $Z_1, \ldots, Z_r$ are i.i.d. $\mathcal{N}(0,1)$.

Proof. A second order Taylor expansion gives

$$\frac{2N}{\phi''(1)} D_{\phi_1}(\widehat{p}, p(\widehat{\theta}_{\phi_2})) = \sqrt{N}(\widehat{p}-p(\widehat{\theta}_{\phi_2}))^T diag(p(\theta_0)^{-1})\sqrt{N}(\widehat{p}-p(\widehat{\theta}_{\phi_2})) + o_p(1).$$

Taking into account Theorem 3.2, and applying Corollary 2.1. of Dik and Gunst (1985), the result follows.

Based on the asymptotic result presented in Theorem 3.3 for large N and n, we may reject the null hypothesis $H_0 : p \in \mathcal{P}$, at the asymptotic test size $\alpha \in (0,1)$, if

$$S_{\phi_1,\phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right) > t_\alpha, \quad \phi_1, \phi_2 \in \Phi, \tag{3.6}$$

where $t_\alpha > 0$ is the lowest positive number satisfying the condition

$$\sup_{\theta\in\Theta} \mathsf{P}\left(\sum_{j=1}^{r} \beta_j(\theta) Z_j^2 > t_\alpha\right) \leq \alpha.$$

with $Z_1, \ldots, Z_r$ i.i.d. $\mathcal{N}(0,1)$ and $\beta_1(\theta), \ldots, \beta_r(\theta)$ being the non null eigenvalues of the matrix $\text{diag}\,(p(\theta)^{-1})\, B(\theta)$ defined above.

3.1. Approximations to the linear combination of chi-square distributions

One has to take into account that asymptotic distribution of S_{ϕ_1,ϕ_2} under H_0 depends on $\theta \in \Theta$, so from a practical point of view the worst situation may be considered and $w(\theta) = \mathsf{P}\left(\sum_{j=1}^{r} \beta_j(\theta) Z_j^2 > s\right)$ can be calculated for the observed value s of S_{ϕ_1,ϕ_2} and for each $\theta \in \Theta$. If $\sup_{\theta\in\Theta} w(\theta) < \alpha$, then we have evidence to reject the null hypothesis. To apply this procedure, it is important to have approximations to the distribution of a linear combination of independent and chi-square distributed random variables with one degree of freedom. Here we describe two direct and easy to calculate approximations and we give references to somewhat better, and at the same time more complicated, techniques.

First approximation is taken from Rao and Scott (1981). Let us define $\beta_{max}(\theta) = \max\{\beta_1(\theta), \dots, \beta_r(\theta)\}$ and $\beta_{max} = \sup_{\theta\in\Theta} \beta_{max}(\theta)$. Then

$$w(\theta) = \mathsf{P}\left(\sum_{j=1}^{r} \beta_j(\theta) Z_j^2 > s\right) \leq \mathsf{P}\left(\chi_r^2 > s\beta_{max}(\theta)^{-1}\right) = w_1(\theta).$$

Therefore, if $\sup_{\theta\in\Theta} w_1(\theta) < \alpha$ we should reject H_0. In this case we get an asymptotically conservative decision rule.

Second approximation is based on the relations

$$\begin{aligned}
\mathsf{E}\left[\sum_{j=1}^{r} \beta_j(\theta) Z_j^2\right] &= r\overline{\beta}(\theta) = \mathsf{E}\left[\overline{\beta}(\theta)\chi_r^2\right], \quad \text{with } \overline{\beta} = \frac{1}{r}\sum_{j=1}^{r} \beta_j(\theta), \\
\mathsf{Var}\left[\sum_{j=1}^{r} \beta_j(\theta) Z_j^2\right] &= 2\sum_{j=1}^{r} \beta_j(\theta)^2 = 2\sum_{j=1}^{r} (\beta_j(\theta) - \overline{\beta}(\theta))^2 + 2r\overline{\beta}(\theta)^2 \geq \mathsf{Var}\left[\overline{\beta}(\theta)\chi_r^2\right], \\
w(\theta) &= \mathsf{P}\left(\sum_{j=1}^{r} \beta_j(\theta) Z_j^2 > s\right) \approx \mathsf{P}\left(\chi_r^2 > s\overline{\beta}(\theta)^{-1}\right) = w_2(\theta).
\end{aligned}$$

If $\sup_{\theta\in\Theta} w_2(\theta) < \alpha$, H_0 should be rejected. Note that

$$\mathsf{E}\left[\sum_{i=1}^{r} \beta_i(\theta) Z_i^2\right] = \sum_{i=1}^{r} \beta_i(\theta) = \operatorname{trace}\left(\operatorname{diag}\left(p(\theta)^{-1}\right) B(\theta)\right) = \sum_{i=1}^{M} \frac{g_{ii}(\theta)}{p_i(\theta)},$$

where $g_{ii}(\theta)$ are the diagonal elements of the matrix $B(\theta)$. Therefore

$$\overline{\beta}(\theta) = \frac{1}{r}\sum_{i=1}^{M} \frac{g_{ii}(\theta)}{p_i(\theta)}.$$

Satterthwaite (1946) presented an approximation to the asymptotic distribution of the statistic

$$R_{\phi_1,\phi_2} = c\, S_{\phi_1,\phi_2} + d$$

where c and d are chosen in such a way that expectation and variance of R_{ϕ_1,ϕ_2} coincide with the expectation and variance of a χ_r^2 random variable. Jensen and Solomon (1972) presented a normal approximation and employed a Wilson-Hilferty type scheme to accelerate the rate of convergence to normality. Imhof (1961) considered a nonstatistical approximation based directly in the numerical inversion of the characteristic function. Apart from the above approximations tables of the cumulative distribution of $\sum_{j=1}^{r} a_j Z_j^2$, with $Z_1, \ldots, Z_r$ i.i.d standard normal, are available in the case of small r (see Solomon (1960), Johnson and Kotz (1968), Eckler (1969), and Gupta (1963)).

3.2. Asymptotic power function

To check the consistency of (3.6) and to obtain an approximation to its power function, we consider the case $p \notin \mathcal{P}$. First we introduce the following regularity assumptions.

(A1) There exists $\theta_a = \arg\inf_{\theta\in\Theta} D_\phi\left(p, p(\theta)\right)$ such that

$$p(\widehat{\theta}_\phi) \xrightarrow{a.s.} p(\theta_a) \quad \text{as} \quad N \to \infty,\ n/N \to f > 0.$$

(A2) There exists $\theta_a \in \Theta$, $\Sigma^* = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, $\Sigma_{11} = \Sigma = \left(\sigma_{ij}\right)_{i,j=1,\ldots,M}$ with σ_{ij} given in (2.5) and $\Sigma_{12} = \Sigma_{21}$, such that

$$\sqrt{N}\begin{pmatrix} \widehat{p} - p \\ p(\widehat{\theta}_\phi) - p(\theta_a) \end{pmatrix} \underset{n\longrightarrow\infty}{\overset{L}{\longrightarrow}} \mathcal{N}_2\left(0, \Sigma^*\right) \quad \text{as} \quad N \to \infty,\ n/N \to f > 0.$$

Theorem 3.4. If (A1) holds, then test (3.6) is asymptotically consistent as $N \to \infty$, $n/N \to f > 0$.

Proof. Let $p \notin \mathcal{P}$, then

$$D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right) \xrightarrow{P} D_{\phi_1}\left(p, p(\theta_a)\right) > 0, \quad \text{as} \quad N \to \infty,\ n/N \to f > 0.$$

Therefore

$$P\left(\frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right) > t_\alpha\right) = P\left(D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right) > \frac{\phi_1''(1)t_\alpha}{2N}\right) \longrightarrow 1.$$

Theorem 3.5. Let $N \to \infty$ and $n/N \to f > 0$. If (A1)-(A2) hold, then

$$\sqrt{N}\left[D_{\phi_1}\left(\widehat{p}, p(\widehat{\theta}_{\phi_2})\right) - D_{\phi_1}\left(p, p(\theta_a)\right)\right] \xrightarrow{L} \mathcal{N}\left(0, \sigma^2\right),$$

where

$$\sigma^2 = T\Sigma_{11}T^T + T\Sigma_{12}S^T + S\Sigma_{21}T^T + S\Sigma_{22}S^T, \tag{3.7}$$

$$T = \left(t_1, \ldots, t_M\right)^T, \quad \text{with} \quad t_i = \left(\frac{\partial}{\partial p_i^1} D_{\phi_1}\left(p^1, p^2\right)\right)_{p^1 = p,\, p^2 = p(\theta_a)}, \quad i = 1, \ldots, M,$$

and

$$S=(s_1,\dots,s_M)^T,\quad \text{with}\quad s_i=\left(\frac{\partial}{\partial p_i^2}D_{\phi_1}\left(p^1,p^2\right)\right)_{p^1=p,p^2=p(\theta_a)},\quad i=1,\dots,M.$$

Proof. A first order Taylor expansion gives

$$\begin{aligned} D_{\phi_1}\left(\widehat{p},p(\widehat{\theta}_{\phi_2})\right) &= D_{\phi_1}\left(p,p(\theta_a)\right)+T\left(\widehat{p}-p\right)+S\left(p(\widehat{\theta}_{\phi_2})-p(\theta_a)\right)^T \\ &+\ o\left(\left\|\widehat{p}-p\right\|+\left\|p(\widehat{\theta}_{\phi_2})-p(\theta_a)\right\|\right). \end{aligned}$$

The result follows from the assumed hypotheses.

Theorem 3.5 can be used to obtain the following approximation to the power of test (3.6). Approximated power function is

$$\beta(p)=\mathsf{P}_p\left(S_{\phi_1,\phi_2}>t_\alpha\right)\approx 1-F_N\left(\frac{\phi_1''(1)t_\alpha-2ND_{\phi_1}\left(p,p(\theta_a)\right)}{2N^{1/2}\sigma}\right) \tag{3.8}$$

where σ is given in (3.7) and $F_N(x)$ is a sequence of distributions functions tending uniformly to the standard normal distribution $F(x)$. Note that if $p\notin\mathcal{P}$, then for any fixed test size α the probability of rejecting $H_0: p\in\mathcal{P}$ with the rejection rule $S_{\phi_1,\phi_2}>t_\alpha$ tends to one as $N\to\infty$ and $n/N\to f>0$.

Obtaining the approximate sample size N, guaranteeing a power β for a given alternative p, is an interesting application of formula (3.8). If N^* is the positive root of the equation

$$\beta=1-F\left(\frac{\sqrt{N}}{\sigma}\left[\frac{\phi_1''(1)t_\alpha}{2N}-D_{\phi_1}\left(p,p\left(\theta_a\right)\right)\right]\right),$$

where F stands for the standard normal cumulative distribution function, then

$$N^*=\frac{A+B\sqrt{A\left(A+2B\right)}}{2D_{\phi_1}\left(q,p\left(\theta_a\right)\right)},$$

with

$$A=\sigma^2\left(F^{-1}\left(1-\beta\right)\right)^2\quad\text{and}\quad B=\phi''(1)t_\alpha D_{\phi_1}\left(p,p\left(\theta_a\right)\right),$$

and the required sample size is $N=[N^*]+1$, where $[.]$ denotes "integer part of".

4. Numerical example

In this section we present an example to illustrate the application of the family of tests introduced in Section 3. We consider the power divergence family (ϕ-divergences with ϕ from (1.7)). For this family, test statistics given in (3.5) have the expression

$$S_{\psi_\lambda,\psi_{\lambda_2}}=2N\,D_{\psi_\lambda}\left(\widehat{p},p(\widehat{\theta}_{\psi_{\lambda_2}})\right)=\sum_{i=1}^{M}\widehat{p}_i\left[\left(\frac{\widehat{p}_i}{p_i(\widehat{\theta}_{\psi_{\lambda_2}})}\right)^\lambda-1\right], \tag{4.1}$$

if $-\infty < \lambda < \infty$, $\lambda \neq -1, 0$, and

$$S_{\psi_0,\psi_{\lambda_2}} = \sum_{i=1}^{M} \widehat{p}_i \log \frac{\widehat{p}_i}{p_i(\widehat{\theta}_{\psi_{\lambda_2}})}, \quad S_{\psi_{-1},\psi_{\lambda_2}} = \sum_{i=1}^{M} p_i(\widehat{\theta}_{\psi_{\lambda_2}}) \log \frac{p_i(\widehat{\theta}_{\psi_{\lambda_2}})}{\widehat{p}_i}.$$

For $\lambda_2 = 0$, $\widehat{\theta}_{\psi_0}$, is the MLE and S_{ψ_λ,ψ_0} is the family of power divergence test statistics when the parameter is estimated by the MLE.

We consider the data

		Fallible device				
		1	2	3		
True device	1	37	2	2	41	First sample
	2	1	23	1	25	
	3	1	3	30	34	
		39	28	33	100	

160	57	183	400	Second sample

with sample sizes $n = 100$ and $N - n = 400$. We are interested in testing $H_0 : p \in \mathcal{P}$, where

$$\mathcal{P} = \left\{ (\theta^2, (1-\theta)^2, 2\theta(1-\theta) \in \Delta_3 : \ \theta \in (0,1) \right\},$$

and Δ_3 is defined in (1.1) for $M = 3$. By applying (2.3) to the data, we get

$$\widehat{p}_1 = 0.4159, \ \widehat{p}_1 = 0.1629, \ \widehat{p}_1 = 0.4212, \ \widehat{q}_{1/1} = 0.9078, \ \widehat{q}_{2/1} = 0.0292, \ \widehat{q}_{3/1} = 0.0629$$
$$\widehat{q}_{1/2} = 0.0626, \ \widehat{q}_{2/2} = 0.8570, \ \widehat{q}_{3/2} = 0.0803, \ \widehat{q}_{1/3} = 0.0242, \ \widehat{q}_{2/3} = 0.0433, \ \widehat{q}_{3/3} = 0.9325.$$

Maximum likelihood estimator of θ is minimum ϕ-divergence estimator with $\phi(x) = \psi_0(x) = x \log x - x + 1$. For this function ψ_0 we have,

$$D_{\psi_0}(\widehat{p}, p(\theta)) = \widehat{p}_1 \log \frac{\widehat{p}_1}{p_1(\theta)} + \widehat{p}_2 \log \frac{\widehat{p}_2}{p_2(\theta)} + \widehat{p}_3 \log \frac{\widehat{p}_3}{p_3(\theta)}, \tag{4.2}$$

where $\widehat{p} = (\widehat{p}_1, \widehat{p}_2, \widehat{p}_3)$ and $p(\theta) = (p_1(\theta), p_2(\theta), p_3(\theta))^T = \left(\theta^2, (1-\theta)^2, 2\theta(1-\theta)\right)^T$. In (4.2) minimum is obtained at

$$\widehat{\theta}_{\psi_0} = \frac{2\widehat{p}_1 + \widehat{p}_3}{2} = 0.6265,$$

so that $p_1(\widehat{\theta}_{\psi_0}) = 0.3925$, $p_2(\widehat{\theta}_{\psi_0}) = 0.1629$ and $p_3(\widehat{\theta}_{\psi_0}) = 0.4680$. By plugging these probabilities in (4.1), numerical values t_λ of test statistics $S_{\psi_\lambda,\psi_{\lambda_0}}$ are calculated and presented in Table 4.1.

λ	-2	-1	-0.5	0	2/3	1	2
t_λ	9.8621	9.9971	10.086	10.193	10.362	10.459	10.802

Table 4.1. Observed values of test statistics.

Let θ be the true value of the parameter. Then test statistics $S_{\psi_\lambda,\psi_{\lambda_0}}$ are asymptotically distributed as $\sum_{j=1}^{3} \beta_j(\theta) Z_j^2$, where Z_1, Z_2, Z_3 are i.i.d. $\mathcal{N}(0,1)$ and $\beta_1(\theta), \beta_2(\theta), \beta_3(\theta)$ are the eigenvalues of the matrix $M(\theta) = \text{diag}\,(p(\theta)^{-1})\, B(\theta)$. After some algebraic calculations one can check that $M(\theta)$ has only two nonnull eigenvalues.

For $\theta \in \tilde{\Theta} = \{0.01, 0.02, \ldots, 0.99\}$, nonnull eigenvalues $\beta_1(\theta)$ and $\beta_2(\theta)$ of $M(\theta)$ are obtained and probabilities $P_{\lambda,\theta} = P(\beta_1(\theta)Z_1^2 + \beta_2(\theta)Z_2^2 > t_\lambda)$ are calculated by simulating 100.000 standard normal random numbers, z_1 and z_2, and by counting the number of times that inequality $\beta_1(\theta)z_1^2 + \beta_2(\theta)z_2^2 > t_\lambda$ holds. Alternatively, if no device for random number generation is available, $P_{\lambda,\theta}$ can be approximated by any of the methods presented in Section 3.1.

For each considered λ, p–values $P_\lambda = \sup_{\theta\in\tilde{\Theta}} P_{\lambda,\theta}$ are given in 4.1.

λ	-2	-1	-0.5	0	2/3	1	2
P_λ	0.0273	0.0263	0.0256	0.0248	0.0230	0.0237	0.0209

Table 4.2. Observed p–values of test statistics.

As $P_\lambda < 0.05$ for every considered λ, we may reject H_0.

Acknowledgments. We thank the referees for a careful examination of the paper and suggestions leading to its improvements.

References

[1] Ali, S.M. and Silvey, S.D. (1966). A general class of coefficient of divergence of one distribution from another. *J. of Royal Statistical Society, Series B*, **286**, 131–142.

[2] Birch, M. H. (1964). A new proof of the Pearson-Fisher theorem. *Annals of Mathematical Statistics*, **35**, 817-824.

[3] Cheng, K.F., Hsueh, H.M. and Chien, T.H. (1998). Goodness of fit tests with misclassified data. *Communications in Statistics - Theory and Methods*, **27**, 1379–1393.

[4] Csiszár, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, **8**, Ser. *A*, 85–108.

[5] Cressie, N. and Read, T.R.C. (1984). Multinomial Goodness-of-fit Tests. *J. of the Royal Statistical Society, Series B*, **46**, 440–464.

[6] Dik, J. J. and Gunst, M. C. M: (1985): The distribution of general quadratic forms in normal variables. *Statistica Neerlandica*, **39**, 14-26.

[7] Eckler, A. R. (1969): A survey of coverage problems associated with point and area targets. *Tecnometrics*, **11**, 561-589.

[8] Gupta, S. S. (1963): Bibliography on the multivariate normal integrals and related topics. *Annals of Mathematical Statistics*, **34**, 829-838.

[9] Imhof, J. P. (1961): Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419-426.

[10] Jensen, D. R.and Solomon, H. (1972): A Gaussian appproximation to the distribution of a definite quadratic form. *J. of the American Statistical Association*, **67**, 898-902.

[11] Johnson, N. L. and Kotz, S. (1968): Tables of distributions of positive definite quadratic forms in central normal variables. *Sankhya*, **30**, 303-314.

[12] Morales, D., Pardo, L. and Vajda, I. (1995): Asymptotic divergence of estimates of discrete distributions. *J. of Statistical Planning and Inference*, **48**, 347–369.

[13] Pardo, L. and Zografos K. (2000): Goodness of fit tests with misclassified data based on ϕ–divergences. *Biometrical Journal*, **42**, 223–237.

[14] Rao, J. N. K. and Scott, A. J. (1981). The Analysis of categorical data from complex sample surveys: Chi-squared tests for goodness-of-fit and independence in two-way tables. *J. of the American Statistical Association*, **76**, 221-230.

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

[16] Satterhwaite, F. E. (1946): An approximation distribution of estimates of variance components. *Biometrics*, **2**, 110-114.

[17] Solomon, H. (1960): Distribution of quadratic forms-tables and applications, Technical Report 45, Applied Mathematics and statistics laboratories, Stanford University, Stanford, CA.

[18] Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassification. *J. of the American Statistical Association*, **65**, 1350–1361.

[19] Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassification: Sample size and determination. *Biometrics*, **27**, 935–944.

[20] Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics*, **14**, 187–202.