Pergamon

PII: S0016-0032(96)00063-4

J. Franklin Inst. Vol. 334B, No. 2, pp. 307–318, 1997
Copyright © 1997 The Franklin Institute
Published by Elsevier Science Ltd
Printed in Great Britain
0016-0032/97 $17.00+0.00

The Jensen–Shannon Divergence

by M. L. MENÉNDEZ

Department of Applied Mathematics, Technical University of Madrid, Madrid, Spain

J. A. PARDO, L. PARDO *and* M. C. PARDO

Department of Statistics and O.R., Complutense University of Madrid, Madrid, Spain

(Received 5 March 1996; accepted 14 May 1996)

ABSTRACT: *In this paper we investigate the Jensen–Shannon parametric divergence for testing goodness-of-fit for point estimation. Most of the work presented is an analytical study of the asymptotic differences between different members of the family proposed in goodness of fit, together with an examination of closer approximations to the exact distribution of these statistics than the commonly used chi-squared distribution. Finally the minimum Jensen–Shannon divergence estimates are introduced and compared with other well-known estimators by computer simulation.* Copyright © 1997 Published by Elsevier Science Ltd

I. Introduction

Let a and $1-a$, $0<a<1$, be the weights of the two probability distributions $P=(p_1,\dots,p_M)$ and $Q=(q_1,\dots,q_M)$, respectively. Lin (**1**) defined the Jensen–Shannon divergence in the following way:

$$L(P,Q) = H(aP+(1-a)Q)-aH(P)-(1-a)H(Q) \tag{1}$$

with $H(P) = -\Sigma_{i=1}^{M} p_i \log p_i$ being Shannon's entropy (**2**). This is a particular case of the Jensen difference divergences introduced by Burbea and Rao (**3**). The Jensen–Shannon divergence $L(P,Q)$ is called the increment of the Shannon entropy by Wong and You (**4**) and was used to measure the distance between random graphs. It was introduced as a criterion for the synthesis of random graphs. In fact, it was shown that the divergence given in Eq. (1) provides both the lower and upper bounds to the Bayes probability error. Lin (**1**) said that one of the major features of the Jensen–Shannon divergence is that it is possible to assign different weights to the distributions involved according to their importance. This is particularly useful in the study of decision problems.

If we consider the function $\varphi(x) = ax\log x-(ax+(1-a))\log(ax+1-a)$, the measure of divergence (1) can be seen as a particular case of the φ-divergence introduced by Csiszár (**5**) in the following way:

$$D_\varphi(P,Q) = \sum_{j=1}^{M} q_j\varphi\left(\frac{p_j}{q_j}\right) \tag{2}$$

for any continuous convex function $\varphi:[0,\infty)\to R\cup\{\infty\}$, where $0\varphi(0/0)=0$ and $0\varphi(p/0) = p\lim_{u\to\infty}(\varphi(u)/u)$. Since the Jensen–Shannon divergence is the only common measure of divergence to the families, Jensen difference divergences and φ-divergence measures, it is important to study its behaviour. In this paper we will study it as a function of the parameter a in the goodness-of-fit and estimation problems.

II. Optimality of the Jensen–Shannon Family of Statistics

Multinomial test for the fit of iid observations $X_1,\ldots,X_n$ to a specified distribution F is based on the counts N_i, $i=1,\ldots,M$, of observations falling in M cells $C_1,\ldots,C_M$ that partition the range of the X_j. The earliest such test is based on the Pearson (**6**) chi-squared statistic: $\chi^2 = \Sigma_{i=1}^{M}(N_i-np_i)^2/np_i$, where $p_i = P_F(X_j\in C_i)$, $i=1,\ldots,M$ are the cell probabilities under the null hypothesis. Later many papers solving this problem have appeared in the literature. In this sense, the following papers are important: Read and Cressie (**7**), Zografos *et al.* (**8**), Menéndez *et al.* (**9**) and so on.

If we denote by $\hat{P}=(\hat{p}_1,\ldots,\hat{p}_M)$, $\hat{p}_i = n_i/n$, the statistic

$$D^a(\hat{P},P_0) = \frac{2nL(\hat{P},P_0)}{\varphi''(1)}$$
$$= 2n\left\{\frac{1}{1-a}\sum_{i=1}^{M}\hat{p}_i\log\frac{\hat{p}_i}{a\hat{p}_i+(1-a)p_{i0}} + \frac{1}{a}\sum_{i=1}^{M}p_{i0}\log\frac{p_{i0}}{a\hat{p}_i+(1-a)p_{i0}}\right\} \quad (3)$$

under the null hypothesis $H_0:Q=P_0$ (**10**), is a chi-squared distribution with M–1 degrees of freedom.

For $a\to 0$, we have the loglikelihood ratio statistic, G^2, and for $a\to 1$ the reversed Shannon information statistic. The statistic $D^a(\hat{P},P_0)$, $0\leqslant a\leqslant 1$, will be called, in the following, the Jensen–Shannon family of statistics.

Each statistic of the Jensen–Shannon family has the same asymptotic distribution as Pearson's χ^2 statistic, so that

$$\Pr(D^a(\hat{P},P_0)\geqslant\chi^2_{M-1,\alpha})\to\alpha \quad \text{if} n\to\infty \quad (4)$$

for each $a\in[0,1]$ and each $\alpha\in(0,1)$ where $\chi^2_{M-1,\alpha}$ is such that $\Pr(\chi^2_{M-1}>\chi^2_{M-1,\alpha})=\alpha$.

If $H_1:Q=\pi=(\pi_1,\ldots,\pi_M)$ is the alternative hypothesis then the asymptotic power is given by $\beta_n(\pi_1,\ldots,\pi_M) = \Pr(D^a(\hat{P},\pi)\geqslant\chi^2_{M-1,\alpha})$ and $\lim_{n\to\infty}\beta_n(\pi_1,\ldots,\pi_M)=1$. That is to say, the test is asymptotically consistent in the sense of Fraser (**11**).

To produce a non-trivial asymptotic power, Cochran (**12**) suggested using a set of local alternative hypotheses that converge to the null probability vector as n increases. In particular, consider

$$H_{1,n}:p_i^* = p_{i0}+n^{-1/2}\gamma_i \quad i=1,\ldots,M$$

where the vector $\gamma'=(\gamma_1,\ldots,\gamma_M)$ satisfies $\Sigma_{i=1}^{M}\gamma_i=0$. If we consider the result given in Menéndez *et al.* (**9**) and Zografos *et al.* (**8**), we have

$$\lim_{n\to\infty}\beta_n(p_1^*,\ldots,p_M^*) = \Pr(\chi^2_{M-1}(\delta) \geqslant \chi^2_{M-1,\alpha}) \quad a\in[0,1].$$

Here $\chi^2_{M-1}(\gamma)$ represents a non-central chi-squared random variable with $M-1$ degrees of freedom and non-centrality parameter $\delta = \gamma' A\gamma$, where $A = \text{diag}(p_i^{-1})_{i=1,\ldots,M}$. The Pitman asymptotic relative efficiency of $D^{a_1}(\hat{P}, P_0)$ to $D^{a_2}(\hat{P}, P_0)$ is given by the ratio of their non-centrality parameters and is therefore equal to one for each a_1 and a_2. Consequently, we cannot choose the best a according to the Pitman definition of efficiency.

Another concept of efficiency was introduced by Bahadur (**13**). A generalization of Bahadur example 8.3, p. 31 [see also Refs (**14**), p. 447 and (**8**)] gives the following method to evaluate the exact Bahadur slope of the test statistic $D^a(\hat{P}, P_0)$. If we consider the goodness of fit test, then $C^a(P) = \inf_{v\in B^a} 2D^0(v, Q)$, where $B^a = \{v\in\Delta_M^+ : D^a(v, Q) \geqslant D^a(P, Q)\}$, and

$$\Delta_M^+ = \left\{P = (p_1,\ldots,p_M) : p_i > 0, i = 1,\ldots,M;\ \sum_{i=1}^{M} p_i = 1\right\}.$$

The exact Bahadur efficiency, between two tests obtained from the Jensen–Shannon family of statistics, is equal to the ratio of its exact Bahadur slopes. A straightforward generalization of example 8.3 of Bahadur (**13**) provides that the loglikelihood ratio statistics obtains maximal Bahadur efficiency among all tests based on the Jensen–Shannon family of statistics.

It is also possible to compare arbitrary goodness of fit tests

$$(D^a(\hat{P}, P_0), \chi^2_{M-1,(1-\alpha)}, \quad 0 < \alpha < 1$$

from the point of view of how the first two moments of the test statistic $D^a(\hat{P}, P_0)$ match the first two moments of the limiting chi-square random variable χ^2_{M-1}. The method used here is similar to the method given by Cressie and Read (**14**). We therefore omit justification or motivation of it.

The proximity is interpreted as a coincidence between moments. For the moment $\mu_\beta(D^a(\hat{P}, P_0)) = E[(D^a(\hat{P}, P_0))^\beta]$ of given order β we evaluate the asymptotic expansion

$$\mu_\beta(D^a(\hat{P}, P_0)) = m_{\beta,0}(a) + \frac{m_{\beta,1}(a)}{n} + o(n^{-1}),$$

where the parameters $m_{\beta,i}(a)$, $i = 0, 1$, are given by

$$m_{1,0}(a) = M-1,$$

$$m_{2,0}(a) = M^2-1,$$

$$m_{1,1}(a) = M-1-\frac{(1+a)}{3n}\left(2-3M+\sum_{i=1}^{M} p_{i0}^{-1}\right)$$

$$+\frac{a^2+a+1}{2n}\left(1-2M+\sum_{i=1}^{M} p_{i0}^{-1}\right)+O(n^{-3/2}),$$

and

$$m_{2,1}(a) = M^2-1+\frac{1}{n}\left(2-2M-M^2+\sum_{i=1}^{M} p_{i0}^{-1}\right)$$

$$-\frac{a+1}{3n}\left(10-13M-6M^2+(M+8)\sum_{i=1}^{M} p_{i0}^{-1}\right)+\frac{(a+1)^2}{3n}\left(4-6M-3M^2+5\sum_{i=1}^{M} p_{i0}^{-1}\right)$$

$$+(a^2+a+1)\left(3-5M-2M^2+(M+3)\sum_{i=1}^{M} p_{i0}^{-1}\right)+O(n^{-3/2}).$$

According to this criterion, as optimal values of a are considered, the set R_β of roots of the equations $m_{\beta,1}(a) = 0$, $\beta = 1, 2$. An interesting situation appears when $p_{i0} = 1/M$, $i = 1, \ldots, M$ and $M \to \infty$. In this case we have that the two roots, irrespective of β, are $\lambda_{\beta,1} = 1$, and $\lambda_{\beta,2} = 0$. For $\lambda_{\beta,1} = 1$, we have the loglikelihood ratio statistic and for $\lambda_{\beta,2} = 0$, we found an optimal statistic, the reversed Shannon information statistic, as an alternative to the loglikelihood ratio statistic.

The power function is one of the most important criteria for comparing test statistics for finite samples although often mathematically intractable. However, here it is accessible on the computer for the $D^a(\hat{P}, P_0)$ test statistics for testing the equiprobable null hypothesis against the specified alternative model

$$H_1 : p_i = \begin{cases} \dfrac{M-1-\delta}{M(M-1)} & \text{if } i = 1, \ldots, M-1 \\ \dfrac{1+\delta}{M} & \text{if } i = M \end{cases}$$

where $-1 \leqslant \delta \leqslant M-1$ is fixed. This alternative results from the Mth probability being perturbed by δ/M, while the rest are adjusted so that they still sum to one.

Firstly, it is necessary to choose a test size α and calculate the associated critical region. If we rely on the chi-squared approximation studied already then it is clear that we commit an approximation error in calculating a level α test. Therefore we calculate the exact powers based on exact critical regions of the $D^a(\hat{P}, P_0)$ test statistics. To calculate the critical value of the exact size α test, $C_a(\alpha)$, we follow the steps:

(i) choose n and M and calculate all possible partitions x of n into M classes. At the same time calculate the associated multinomial probability for each partition x and the value of the statistics $D^a(x/n, P_0)$;

(ii) rank the partition according to the statistic values from largest to smallest;

(iii) the critical value, $C_a(\alpha)$, is an achievable value of $D^a(\hat{P}, P)$ such that starting from the smallest rank, the sum of the associated probabilities under the null hypothesis, until just before the inequality $D^a(\hat{P}, P) \leqslant C_a(\alpha)$ is satisfied, is equal $\alpha_{1,a}$ and until just before $D^a(\hat{P}, P) < C_a(\alpha)$, is equal to $\alpha_{2,a}$, where $\alpha_{1,a} < \alpha \leqslant \alpha_{2,a}$.

Due to the discrete nature of the critical regions we use the randomized size α test which rejects H_0 with probability

TABLE I

	$(n = 20, M = 5)$ Alternatives				
a	-0.9	-0.5	0.5	1.0	1.5
0.000	0.4463	0.1213	0.1076	0.3056	0.5103
0.100	0.4841	0.1244	0.1012	0.2765	0.5678
0.200	0.4948	0.1252	0.1001	0.2718	0.5588
0.300	0.5260	0.1267	0.0953	0.2521	0.5279
0.400	0.5336	0.1268	0.0933	0.2424	0.5128
0.500	0.5627	0.1281	0.0886	0.2205	0.4705
0.600	0.5689	0.1272	0.0849	0.1971	0.4222
0.700	0.5670	0.1251	0.0821	0.1759	0.3646
0.800	0.5844	0.1282	0.0800	0.1640	0.3361
0.900	0.5849	0.1270	0.0790	0.1503	0.2792
1.000	0.5803	0.1262	0.0789	0.1474	0.2574

$$\begin{cases} 1 & \text{if } D^a(\hat{P}, P) > C_a(\alpha) \\ \dfrac{\alpha - \alpha_{1,a}}{\alpha_{2,a} - \alpha_{1,a}} & \text{if } D^a(\hat{P}, P) = C_a(\alpha). \\ 0 & \text{if } D^a(\hat{P}, P) < C_a(\alpha) \end{cases}$$

Therefore, the power of the randomized size α test is

$$\beta_a = \beta_{1,a} + \frac{\alpha - \alpha_{1,a}}{\alpha_{2,a} - \alpha_{1,a}}(\beta_{2,a} - \beta_{1,a})$$

where $\beta_{1,a}(\beta_{2,a})$ is obtained following steps (i)–(iii). Starting from the smallest rank, the associated probabilities under H_1 are added until just before the inequality $D^a(\hat{P}, P) \leqslant C_a(\alpha)$ $(D^a(\hat{P}, P) < C_a(\alpha))$. The value of the accumulative probabilities sum equal $\beta_{1,a}(\beta_{2,a})$.

We tabulate this power for five δ values, $\delta = -0.9, -0.5, 0.5, 1, 1.5$, $M = 5$, $n = 20$, $\alpha = 0.05$ and several a values (see Table I).

For alternatives $\delta > 0$ for the power increases with a. For alternatives $\delta < 0$ the reverse occurs. For fixed a we can see that the power increases with increasing $|\delta|$. If we are interested in choosing a test with reasonable power against all the alternatives, Table I indicates that one should choose $a \in [0.3, 0.6]$.

III. Improving the Jensen–Shannon Family of Statistics Asymptotic Distribution for Small Samples

In finite samples, the relation

$$\Pr(D^a(\hat{P}, P_0) \leqslant c) = \Pr(\chi^2_{M-1} \leqslant c) + o(1) \tag{5}$$

can be assumed to hold approximately. In this section we explore the accuracy of applying asymptotic results in cases where the sample size cannot be assumed to be large. We will assume the equiprobable hypothesis, that is to say, $p_{io} = 1/M$, $i = 1, \ldots, M$ since if tests of fit are based on continuous functions, then in general they are biased for testing an arbitrary simple hypothesis for multinomial distribution (**15**). However, if tests of fit are based on convex functions, then they are unbiased for testing the equiprobable hypothesis (**16**). Apart from asymptotic χ^2, we have two closer approximations to the $D^a(\hat{P}, P_0)$ exact distribution.

(*a*) *Corrected χ^2 distribution*

The first is a corrected χ^2 distribution

$$\Pr(\chi^2_{M-1} < (c - \gamma_a)/\delta_a^{1/2}) \tag{6}$$

with γ_a, and δ_a being such that the mean and variance of

$$D_M^a(\hat{P}, P_0) = (D^a(\hat{P}, P_0) - \gamma_a)/\delta_a^{1/2}$$

are $M-1$ and $2(M-1)$, respectively, to $o(n^{-1})$. In our case [see Menéndez *et al.* (**17**)] we have

$$\gamma_a = (M-1)(1-\delta_a^{1/2}) - \frac{1}{3n}(a+1)(2-3M+M^2) + \frac{a^2+a+1}{2}(1-2M+M^2)$$

and

$$\delta_a = 1 - \frac{1}{n} - \frac{2}{(M-1)n}(a+1)(2-3M+M^2) + \frac{(a+1)^2}{3n(M-1)}(2-3M+M^2) + \frac{a^2+a+1}{n(M-1)}(1-2M+M^2).$$

The idea of modifying the statistic as above was given by Read (**18**).

(*b*) *Second-order corrected term*

The second closer approximation is derived by extracting the a-dependent second-order component from the $o(1)$ term. This result was obtained in Menéndez *et al.* (**17**) for the family of the (h, ϕ)-divergence measure and it generalizes the results of Yarnold (**19**) and Read (**18**). This approximation is given by

$$\Pr(D^a(\hat{P}, P_0) < c) \simeq J_1^a + J_2^a \tag{7}$$

with

$$J_1^a = \Pr(\chi^2_{M-1} < c) + \frac{1}{24n}\Pr(\chi^2_{M-1} < c)2(1-M^2) + \frac{1}{24n}\Pr(\chi^2_{M+1} < c)[-6(a^2+a+1)(M-1)^2]$$

$$+\frac{1}{24n}\Pr(\chi^2_{M+1} < c)2(1+a)^2(M^2-3M+2)+6(M^2-M)]$$

$$+\frac{1}{24n}\Pr(\chi^2_{M+3} < c)[-4(a+1)^2(M^2-3M+2)+6(a^2+a+1)(M-1)^2]$$

$$+\frac{1}{24n}\Pr(\chi^2_{M+5} < c)2a^2(M^2-3M+2)$$

and

$$J_2^a = [N^a(c)-n^{(M-1)/2}V^a(c)]e^{-c/2}/(2\pi n)^{(M-1)}(1/M)^M]^{1/2},$$

where $N^a(c)$ is the number of lattice points $(w_1, \ldots, w_{M-1})$ which satisfy $w_i = n^{1/2}(N_i/n - 1/M)$, $N_i = 0, 1, 2, \ldots$ such that $\Sigma_{i=1}^M N_i = n$ and $D^a(\hat{P}, P_0) < c$, and

$$V^a(c) = \frac{(\pi c)^{(M-1)/2}}{\Gamma((M+1)/2)}(1/M)^{M/2}\left\{1+\frac{c}{24(M+1)n}[2(1+a)^2(M^2-3M+2)\right.$$
$$\left.-6(a^2+a+1)(M-1)^2]\right\}+O(n^{-3/2}).$$

Now we compare the approximations (5), (6) and (7) which will be denoted by $D_\chi^a(c) \equiv \Pr(\chi^2_{M-1} < c)$, $D_M^a(c) \equiv \Pr(\chi^2_{M-1} < ((c-\gamma_a)/(\delta^a)^{1/2})$, and $D_D^a \equiv J_1^a + J_2^a$, respectively, for small samples through two different criteria. In both criteria we need the exact distribution of the statistic $D^a(\hat{P}, P_0)$, $D_E^a(c) = \Pr(D^a(\hat{P}, P_0) < c)$. To calculate it, we must follow steps (i)–(iii) of Section II. Starting from the largest rank, the associated probabilities are added until just before the inequality $D^a(\hat{P}, P_0) > c$ is satisfied. The value of the accumulative probabilities sum equals $D_E^a(c)$.

Criterion 1 consists of recording the maximum approximation error incurred by each of the three approximations to D_E^a. We then calculate

$$\max_{\hat{P}} |D_E^a(D^a(\hat{P}, P_0)) - D_i^a(D^a(\hat{P}, P_0))|$$

for $i = \chi$, M and D. The sign associated with the maximum difference is also recorded, so we know if the maximum error is an over-estimate or an under-estimate.

Figure 1 illustrates the maximum approximation error resulting from using the approximations D_χ^a, D_M^a and D_D^a for the true distribution function D_E^a which are labelled Apr1, Apr2 and Apr3, respectively, on the graphs. The results are illustrated for specific values of a in the range [0, 1], classes number $M = 5$ and sample sizes $n = 10$ and 20.

At first sight, it is clear that the D_χ^a approximation is worse than D_M^a and D_D^a is a slight improvement on D_M^a. As n increases from 10 to 20, the error curves flatten over the range of a considered so the size of the maximum errors decreases overall.

Secondly, we assess the accuracy of the approximation in calculating the size of the a test. We use the standard χ^2 approximation D_χ^a to give an approximate size α test, i.e. choose c_α such that $1 - D_\chi^a(c_\alpha) = \alpha$. We then calculate

$$1 - D_i^a(c_\alpha) \quad i = E, M, D.$$

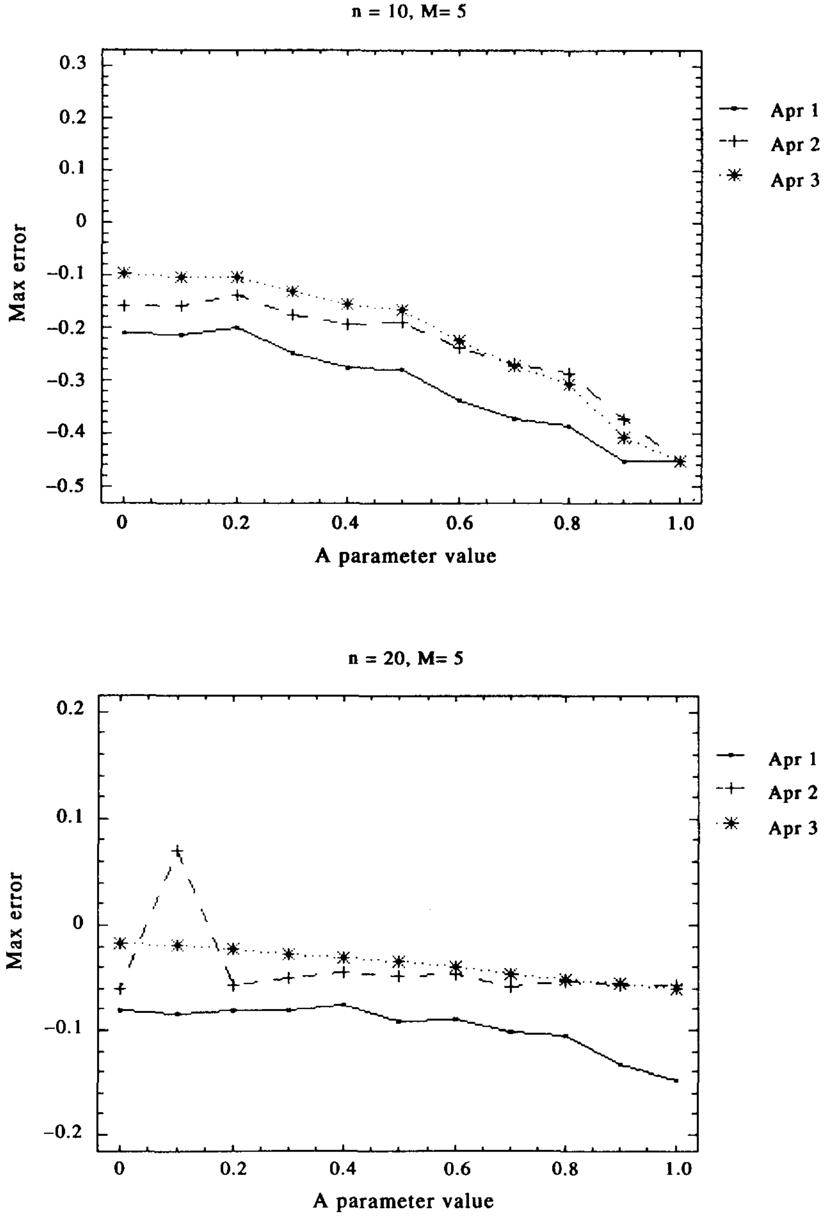

FIG. 1

There are two reasons to take the χ^2 approximation as reference, on the one hand this is the most commonly used for the tests based on X^2 and G^2 and on the other, the critical region obtained from this approximation is independent of a. Figure 2 compares the exact and nominal test levels for the three approximations for the same parameter values as criterion 1 and $\alpha = 0.1$.

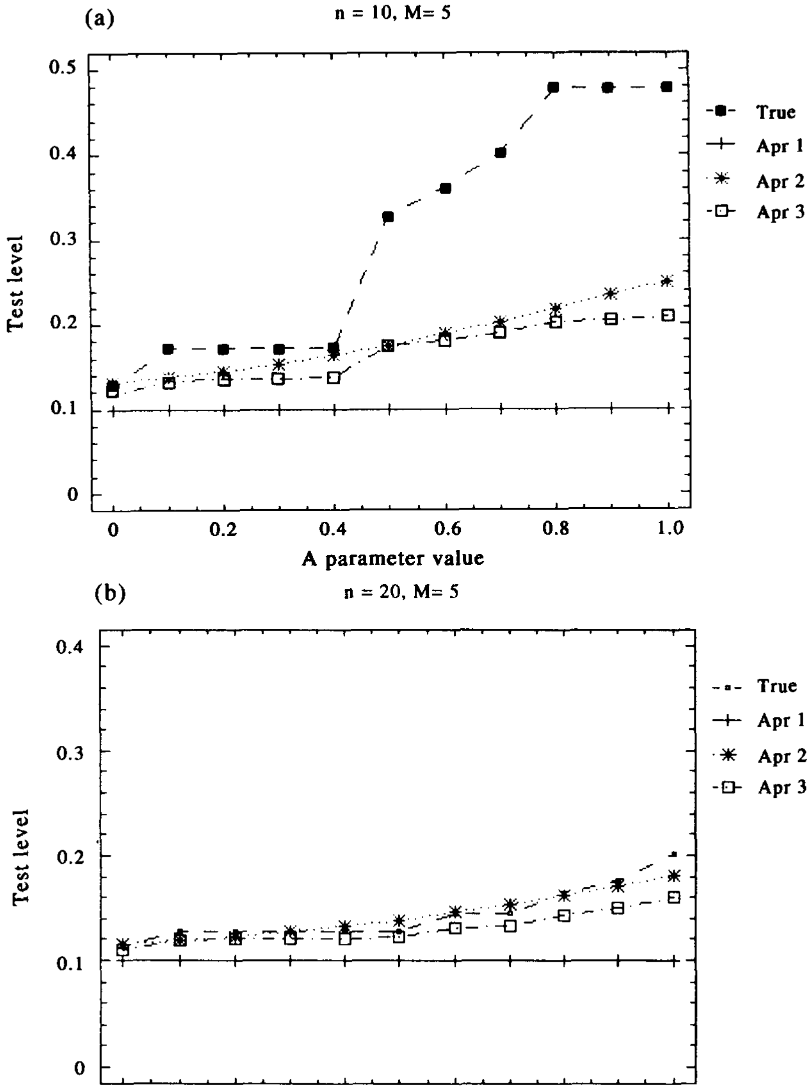

FIG. 2

Although the order of preference among the different approximations is the same as before, the superiority of D_D^a over D_M^a according to criterion 1 is not so clear as criterion 2. The corrected chi-squared has the added bonus that it is much easier to calculate than the second-order chi-squared.

From criteria 1 and 2 we see that if we want to use the standard chi-squared then we should use an a value in the range $[0, 0.4]$.

IV. Minimum Jensen–Shannon Estimates

If the distribution F depends on an unknown parameter $\theta \in \Theta \subset R^{M_0}$, then $q_i(\theta) = Pr_{F_\theta}(X_j \in C_i)$ and we can estimate θ with the maximum likelihood estimator or equivalently minimizing on $\theta \in \Theta \subset R^{M_0}$ the Kulback–Leibler divergence, $D(P, Q)$, since

$$\log P_\theta(N_1 = n_1, \ldots, N_M = n_M) = -nD(\hat{P}, Q(\theta)) + b$$

where b is a constant not depending on $\hat{P}$. Instead of using this divergence it is possible to consider the divergence given in Eq. (1) which leads to the minimum Jensen–Shannon estimator (MJSE) of $\theta \in \Theta$. This estimator is any $\hat{\theta} \in \bar{\Theta}$ (the closure of $\hat{\Theta}$) for which

$$L(\hat{P}, Q(\hat{\theta})) = \inf_{\theta \in \Theta \subset R^{M_0}} L(\hat{P}, Q(\theta)).$$

The asymptotic properties of this estimator can be obtained as a particular case from the paper of Morales *et al.* (**20**) about the minimum φ–divergence estimates. For this reason we only present in this paper an example of its behaviour to estimate the parameters of a normal population. Furthermore, these estimators are compared with those given by other well-known estimators, namely the maximum likelihood estimator for the original data (MLE) and the minimum D_n estimator (MD_nE). The MD_nE is the value $\hat{\theta} \in \Theta$ such that

$$D_n(\hat{\theta}) = \min \{D_n(\theta), \theta \in \Theta\}$$

being

$$D_n(\theta) = \sup_{x \in R} \{|F_n^*(x) - F_\theta(x)|\} = \max \{D_n^+(\theta), D_n^-(\theta)\}$$

$$D_n^+(\theta) = \sup_{x \in R} \{F_n^*(x) - F_\theta(x)\} = \max\left\{0, \max_{i=1,\ldots,n} \left\{\frac{i}{n} - F_\theta(x_{(i)})\right\}\right\}$$

$$D_n^-(\theta) = \sup_{x \in R} \{F_\theta(x) - F_\theta^*(x)\} = \max\left\{F_\theta(x_{(i)}) - \frac{i-1}{n}\right\}$$

where $F_n^*(x)$ is the empiric distribution function of a population sample $x_1, \ldots, x_n$ and $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ are the order statistics.

Table II estimates the parameters of a normal population with mean (μ) equal zero and typical deviation (σ) equal to 1 through the MLE, MD_nE and MJSE. These values have been calculated by computer simulation for 1000 samples, class number = 6 and sample sizes = 10 and 20. The sum of the mean quadratic error also appears in the table.

We choose the members of the Jensen–Shannon family of estimators corresponding to $a = 0$ and 1 because they are known and $a_{\min}$ which is the a value of the Jensen–Shannon divergence that minimizes the mean quadratic error. The $a_{\min}$ values are 0.9464522 and 0.4 for $n = 10$ and 20, respectively.

Looking at Table II we see that the minimum Jensen–Shannon estimators are best for $n = 10$ according to the mean quadratic error criterion, although the MLE is based on the original normal values. The minimum D_n estimator is always worse than $MJSE_{a_{\min}}$.

TABLE II

$N(0,1)$		$n=10$	$n=10$
EMV	$\hat{\mu}$	-0.0352	-0.113
	$\hat{\sigma}$	0.9686	0.9854
	mqe	0.0787	0.0385
ED_n	$\hat{\mu}$	-0.0181	0.0132
	$\hat{\sigma}$	0.9761	0.9878
	mqe	0.0868	0.0430
MJSE_0	$\hat{\mu}$	-0.0191	-0.0103
	$\hat{\sigma}$	0.9973	0.9895
	mqe	0.0496	0.0443
MJSE_1	$\hat{\mu}$	-0.0155	-0.0102
	$\hat{\sigma}$	1.0083	1.0096
	mqe	0.0504	0.0524
$\text{MJSE}_{a_{\min}}$	$\hat{\mu}$	-0.0158	-0.0114
	$\hat{\sigma}$	0.9920	0.9817
	mqe	0.0438	0.0411

Acknowledgement

This work was supported by grants DGICYT PB94-0408 and PB93-0022.

References

(1) J. Lin, "Divergence measures based on the Shannon entropy", *IEEE Trans. Information Theory*, Vol. 37, No. 1, pp. 145–151, 1991.

(2) C. Shannon, "A mathematical theory of Communications", *Bell. System. Tech. J.*, Vol. 27, pp. 379–423, 1948.

(3) J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions", *IEEE Trans. Information Theory*, Vol. 28, pp. 489–495, 1982.

(4) A. K. C. Wong and M. You, "Entropy and distance of random graphs with application to structural pattern recognition", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. PAMI-7, No. 5, pp. 599–609, 1985.

(5) I. Csiszár, "Information type measures of difference of probability distributions and indirect observations", *Studia Sci. Mat. Hung.*, Vol. 2, pp. 299–318, 1967.

(6) K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Phil. Mag.*, Vol. 50, pp. 157–172, 1900.

(7) T. R. C. Read and N. A. C. Cressie, "Goodness-of-fit Statistics for Discrete Multivariate Data". Springer-Verlag, 1988.

(8) K. Zografos, "Asymptotic properties of φ-divergence statistics and its application in contingency tables", *Int. J. Mathematical Stat. Sci.*, Vol. 2, pp. 5–21, 1993.

(9) M. L. Menéndez, D. Morales, L. Pardo and M. Salicrú, "Asymptotic behaviour and statistical applications of divergence measures in multinomial populations: A unified study", *Statistical Papers*, Vol. 36, pp. 1–29, 1995.

(10) D. Morales, L. Pardo, M. Saalicrú and M. L. Menéndez 'Asymptotic properties of divergence statistics in astratified sampling and its applications to test statistical hypotheses", *J. Statistical Planning and Inference*, Vol. 38, pp. 201–222, 1994.

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

(12) W. G. Cochran, "The χ^2 test of goodness of fit", *Annals of Mathematical Statistics*, Vol. 23, pp. 315–345, 1952.

(13) R. R. Bahadur, "Some Limit Theorems in Statistics", Society for Industrial and Applied Mathematics, Philadelphia, 1971.

(14) N. Cressie and T. R. C. Read, "Multinomial goodness-of-fit tests", *J. R. Statistical Soc. Series B*, Vol. 46, pp. 440–464, 1984.

(15) T. Bednarski and T. Ledwina, "A note on biasedness of tests of fit", *Mathematische Operationsforschung und Statistik, Series Statistics*, Vol. 9, pp. 191–193, 1978.

(16) J. E. Cohen and H. B. Sackrowitz, "Unbiasedness of the chi-square, likelihood ratio, and other goodness of fit tests for the equal cell case", *Annals of Statistics*, Vol. 3, pp. 959–964, 1975.

(17) M. L. Menéndez, J. A. Pardo, L. Pardo and M. C. Pardo, "Asymptotic approximations for the distributions of the (h, ϕ)-divergence goodness-of-fit statistics: application to Renyi's statistics". To appear in *Kybernetes*, 1996.

(18) T. R. C. Read, "Small sample comparisons for the power divergence goodness of fit statistics". *J. Am. Stat. Assoc.*, Vol. 79, pp. 929–935, 1984.

(19) J. K. Yarnold, "Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set". *Annals of Mathematical Statistics*, Vol 43, pp. 1566–1580, 1972.

(20) D. Morales, L. Pardo and I. Vajda, "Asymptotic divergence of estimates of discrete distributions". *J. Statistical Planning and Inference*, Vol. 48, pp. 347–369, 1995.