

María Del Carmen Pardo

An empirical investigation of Cressie and Read tests for the hypothesis of independence in three-way contingency tables

Kybernetika, Vol. 32 (1996), No. 2, 175--183

Persistent URL: <http://dml.cz/dmlcz/124180>

Terms of use:

© Institute of Information Theory and Automation AS CR, 1996

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these

Terms of use.



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library*
<http://project.dml.cz>

AN EMPIRICAL INVESTIGATION OF CRESSIE AND READ TESTS FOR THE HYPOTHESIS OF INDEPENDENCE IN THREE-WAY CONTINGENCY TABLES

MARÍA DEL CARMEN PARDO

This paper is concerned with comparison of Cressie and Read tests for the hypothesis of independence in three-way contingency tables. A Monte Carlo study is done to empirically compare the power of some members of the family of these tests under various spike alternatives.

1. INTRODUCTION

Three-way contingency tables are a straightforward generalization of two-way contingency tables. They can, of course, be imagined as rectangular parallelepipeds built out of cubical boxes each containing a number. In writing out an $I \times J \times K$ table we will usually write out in order each component $I \times J$ table with marginal totals at the bottom and at the right, and finally the sum of all the $I \times J$ tables.

Goldstein et al. [2] proposed a test statistic based on Matusita's distance [3] for the hypothesis of independence in three-way contingency tables. But this statistic as well as the statistic χ^2 of Pearson [5] and the loglikelihood ratio, G^2 , are particular cases of a more general family of, namely the statistic of Cressie and Read [1]. These authors studied this family in goodness-of-fit. Now, our intention is to show how one may employ the Cressie and Read statistic to test the hypothesis of independence in three-way contingency tables as well as to perform a simulation study comparing the power of the different family members of Cressie and Read statistic for 2×3 and $2 \times 2 \times 2$ contingency tables, various spike alternatives and sample sizes.

Suppose X , Y and Z are discrete random variables whose joint probability distribution is given by

$$P_{XYZ} = (p_{ijk})_{\substack{i=1,\dots,I \\ j=1,\dots,J \\ k=1,\dots,K}} = (P(X = x_i, Y = y_j, Z = z_k))_{\substack{i=1,\dots,I \\ j=1,\dots,J \\ k=1,\dots,K}}$$

Let

$$P_X = (p_{i..})_{i=1,\dots,I}, \quad P_Y = (p_{.j.})_{j=1,\dots,J}, \quad \text{and} \quad P_Z = (p_{..k})_{k=1,\dots,K}$$

be the corresponding marginal probability distributions, i.e.,

$$p_{i..} = \sum_{j=1}^J \sum_{k=1}^K p_{ijk}, \quad p_{.j.} = \sum_{i=1}^I \sum_{k=1}^K p_{ijk}, \quad p_{..k} = \sum_{i=1}^I \sum_{j=1}^J p_{ijk}.$$

The Cressie and Read divergence between P_{XYZ} and $P_X \times P_Y \times P_Z$ is given by

$$D_\lambda(P_{XYZ}, P_X \times P_Y \times P_Z) = \frac{1}{\lambda(\lambda + 1)} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K p_{ijk} \left(\left(\frac{p_{ijk}}{p_{i..} p_{.j.} p_{..k}} \right)^\lambda - 1 \right),$$

for $\lambda \neq 0$, $\lambda \neq -1$, and by the corresponding limits for $\lambda = 0$ and $\lambda = -1$.

Consider an $I \times J \times K$ contingency table with entries given by n_{ijk} , $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$. Let the marginal totals $n_{i..}$, $n_{.j.}$ and $n_{..k}$ be defined by $n_{i..} = \sum_{j=1}^J \sum_{k=1}^K n_{ijk}$, $n_{.j.} = \sum_{i=1}^I \sum_{k=1}^K n_{ijk}$, $n_{..k} = \sum_{i=1}^I \sum_{j=1}^J n_{ijk}$ and let us consider the maximum likelihood estimates of p_{ijk} , $p_{i..}$, $p_{.j.}$ and $p_{..k}$ respectively, i.e.,

$$\hat{p}_{ijk} = \frac{n_{ijk}}{n}, \quad \hat{p}_{i..} = \frac{n_{i..}}{n}, \quad \hat{p}_{.j.} = \frac{n_{.j.}}{n}, \quad \hat{p}_{..k} = \frac{n_{..k}}{n}, \quad \text{where } n = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk}.$$

To estimate $D_\lambda(P_{XYZ}, P_X \times P_Y \times P_Z)$, we replace p_{ijk} , $p_{i..}$, $p_{.j.}$ and $p_{..k}$ by their respective maximum likelihood estimates, so we define

$$D_\lambda(\hat{P}_{XYZ}, \hat{P}_X \times \hat{P}_Y \times \hat{P}_Z) = \frac{1}{\lambda(\lambda + 1)} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \hat{p}_{ijk} \left(\left(\frac{\hat{p}_{ijk}}{\hat{p}_{i..} \hat{p}_{.j.} \hat{p}_{..k}} \right)^\lambda - 1 \right).$$

The asymptotic distribution of this statistic was obtained by Pardo [4] as a particular case of the $(\underline{h}, \underline{\phi})$ -divergence measure's. If $P_{XYZ} \neq P_X \times P_Y \times P_Z$, then

$$n^{1/2} \left(D_\lambda(\hat{P}_{XYZ}, \hat{P}_X \times \hat{P}_Y \times \hat{P}_Z) - D_\lambda(P_{XYZ}, P_X \times P_Y \times P_Z) \right) \xrightarrow[n \rightarrow \infty]{L} N(0, \sigma^2),$$

where

$$\sigma^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K p_{ijk} c_{ijk}^2 - \left(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K p_{ijk} c_{ijk} \right)^2$$

and c_{ijk} are given by

$$\begin{aligned} c_{ijk} = & -\frac{1}{(\lambda + 1)} \left(\sum_{r=1}^I \sum_{s=1}^J \left(\frac{p_{rsk}}{p_{..k}} \left(\frac{p_{rsk}}{p_{r..} p_{.s.} p_{..k}} \right)^\lambda \right) \right. \\ & + \sum_{r=1}^I \sum_{t=1}^K \left(\frac{p_{rjt}}{p_{.j.}} \left(\frac{p_{rjt}}{p_{r..} p_{.j.} p_{..t}} \right)^\lambda \right) + \sum_{s=1}^J \sum_{t=1}^K \left(\frac{p_{ist}}{p_{i..}} \left(\frac{p_{ist}}{p_{i..} p_{.s.} p_{..t}} \right)^\lambda \right) \left. \right) \\ & + \frac{1}{\lambda} \left(\frac{p_{ijk}}{p_{i..} p_{.j.} p_{..k}} \right)^\lambda. \end{aligned}$$

Now if $P_{XYZ} = P_X \times P_Y \times P_Z$ then $\sigma^2 = 0$ and the statistic

$$2n D_\lambda = (\hat{P}_{XYZ}, \hat{P}_X \times \hat{P}_Y \times \hat{P}_Z) = \frac{2n}{\lambda(\lambda + 1)} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \hat{p}_{ijk} \left(\left(\frac{\hat{p}_{ijk}}{\hat{p}_{i..} \hat{p}_{.j.} \hat{p}_{..k}} \right)^\lambda - 1 \right)$$

is asymptotically distributed as the chi-square with $IJK - I - J - K + 2$ degrees of freedom.

From this result we obtained that the asymptotically α -size test of independence is given by

$$2n D_\lambda (\hat{P}_{XYZ}, \hat{P}_X \times \hat{P}_Y \times \hat{P}_Z) > \chi_{IJK-I-J-K+2, \alpha}^2.$$

Several important tests can be obtained as particular cases of the Cressie and Read statistic. We present some of them in the following table.

$2n D_\lambda(\hat{P}_{XYZ}, \hat{P}_X \times \hat{P}_Y \times \hat{P}_Z)$		
λ	Symbol	Statistic Name
1	X^2	Pearson [5]
0	G^2	Loglikelihood ratio
-1/2	T^2	Freeman-Tukey
-1	GM^2	Modified loglikelihood ratio
-2	NM^2	Neyman modified X^2

2. NUMERICAL INVESTIGATIONS

In this section the powers of some members of Cressie and Read family are simulated for testing independence in three-way contingency tables to carry out a comparative study of them.

Let us fix a sample size n , a test size α and the null hypothesis $H_0 : p_{ijk} = p_{i..} p_{.j.} p_{..k}$. We consider the spike alternative hypotheses (Goldstein et al. [2])

$$H_i : p_{ijk} = \begin{cases} \frac{1-\delta}{IJK-1} & \text{if } i = 1, \dots, I-1; j = 1, \dots, J-1; k = 1, \dots, K-1 \\ \delta & \text{if } i = I; j = J; k = K \end{cases}$$

i.e., a δ probability is assigned to the (I, J, K) cell, while the rest probabilities are adjusted so that they still sum to one.

The power function is

$$\beta_\lambda(p) = P \left(2n D_\lambda(\hat{P}_{XYZ}, \hat{P}_X \times \hat{P}_Y \times \hat{P}_Z) > \chi_{IJK-I-J-K+2, \alpha}^2 \mid p \right)$$

for $p = (p_{111}, \dots, p_{IJK}) \in H_1$. If $p \in H_0$ then $\beta_\lambda(p)$ is the test size.

The statistic value is calculated for each of the 10 000 trials of the multinomial random variable with parameters $(IJK; p_{111}, \dots, p_{IJK})$. The relative frequency of the case where the statistic value belongs to the critical region is the simulated power.

The powers were simulated for 2×3 and $2 \times 2 \times 2$ contingency tables, significance level $\alpha = 0.05$, and sample sizes 50 and 100, using programs written in FORTRAN.

Table 1 shows that the exact power for 2×3 contingency tables and sample size 50 increases with $|\lambda|$. But it would be false to conclude that the best Cressie and Read statistic is corresponding to the largest $|\lambda|$. The reason is that the values of the first column, which describe the exact size, depart from the desired value $\alpha = 0.05$ with increasing $|\lambda|$ too. Then, although the statistics power increases with $|\lambda|$ their significiorate.

There are many different ways of interpreting the results of Table 1, just as there are many ways of defining what is a "good" test as opposed to a "bad" test. We define ad hoc a class of "acceptable tests" and choose the one with maximal power. A test is said to be acceptable if its size is at most 0.06. In Table 1, the acceptable tests correspond to $0 \leq \lambda \leq 3$ and among these the power is maximized at $\lambda = 3$.

We notice that the statistic with size closest to the designed one is the Pearson's statistic.

The powers of Table 1 are drawn in Figure 1 for different values of λ .

Table 1. Significance level ($\delta = 0.1666$) and power ($\delta \neq 0.1666$) of the statistics D_λ for 2×3 contingency tables, $n = 50$ and $\alpha = .05$.

λ	δ				
	0.1666	0.2	0.3	0.4	0.5
-5	.1688	.1883	.3845	.6112	.7570
-2	.0908	.0997	.2594	.4807	.6446
-1	.0727	.0795	.2297	.4482	.6328
-0.5	.0654	.0736	.2170	.4375	.6234
-0.3	.0627	.0701	.2110	.4315	.6205
0	.0593	.0660	.2064	.4262	.6200
0.3	.0573	.0630	.2014	.4237	.6205
0.5	.0556	.0617	.1990	.4227	.6228
2/3	.0548	.0602	.1972	.4213	.6253
0.7	.0544	.0601	.1967	.4212	.6256
1	.0539	.0595	.1956	.4234	.6333
1.5	.0541	.0590	.1971	.4308	.6472
2	.0547	.0595	.2008	.4434	.6698
2.5	.0567	.0619	.2077	.4594	.6891
3	.0597	.0666	.2181	.4781	.7136
3.5	.0634	.0726	.2301	.4970	.7381
4	.0682	.0780	.2475	.5181	.7622
4.5	.0757	.0864	.2631	.5413	.7837
5	.0847	.0953	.2825	.5669	.8054

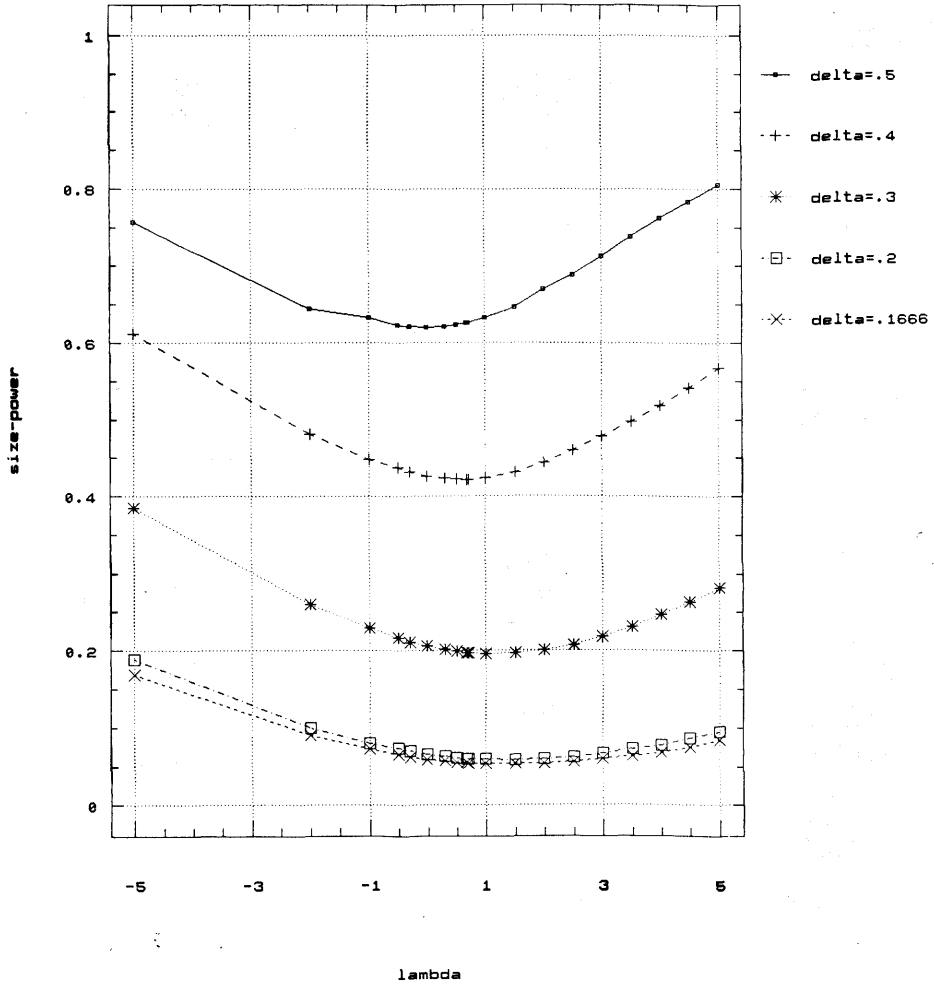


Fig. 1. Significance level ($\delta = 0.1666$) and power ($\delta \neq 0.1666$) of the statistics D_λ for 2×3 contingency tables, $n = 50$ and $\alpha = .05$.

The values of the powers are omitted for 2×3 contingency tables and sample size 100 but they are drawn in Figure 2. Their behaviour is the same as for sample size 50 although the powers are greater and the growth of power with $|\lambda|$ increasing is smaller than before.

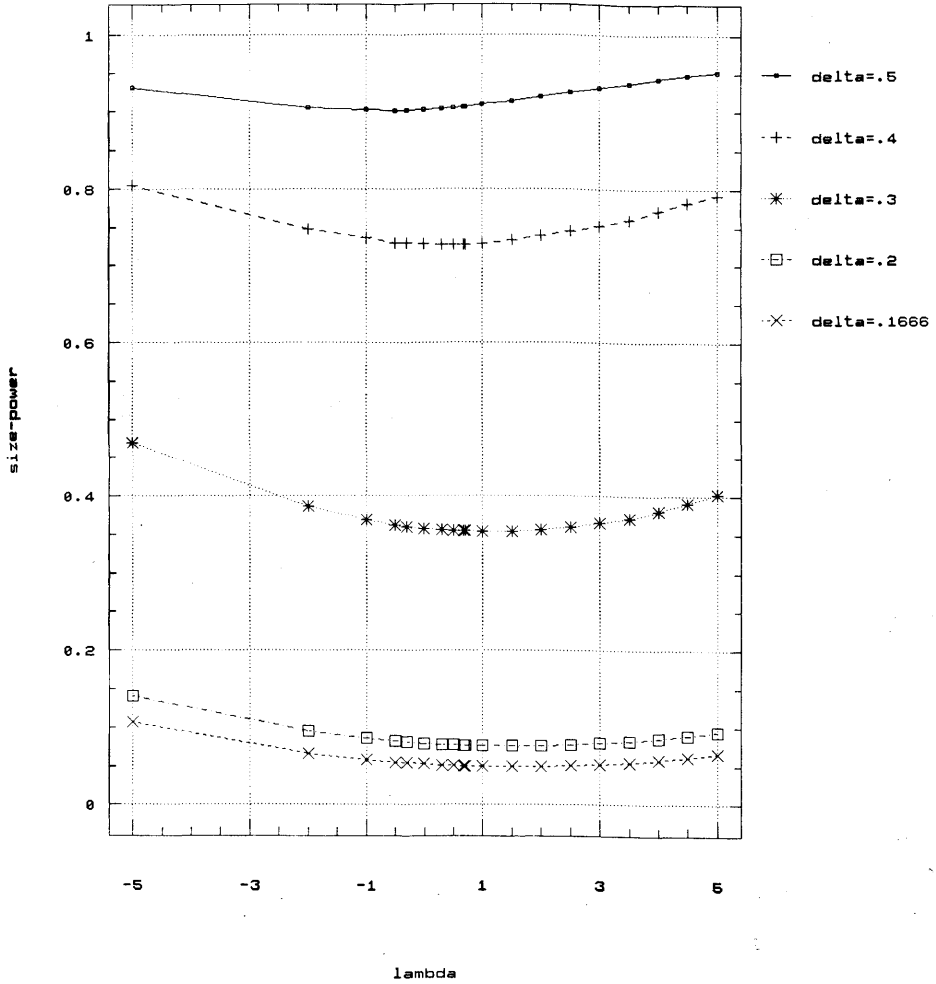


Fig. 2. Significance level ($\delta = 0.1666$) and power ($\delta \neq 0.1666$) of the statistics D_λ for 2×3 contingency tables, $n = 100$ and $\alpha = .05$.

Figures 3 and 4 represent the powers for $2 \times 2 \times 2$ contingency tables and sample sizes 50 and 100 respectively. In these tables the null hypotheses (and, consequently, the test size) corresponds to $\delta = 0.125$. The remaining values represent the alternatives (i.e. the test powers).

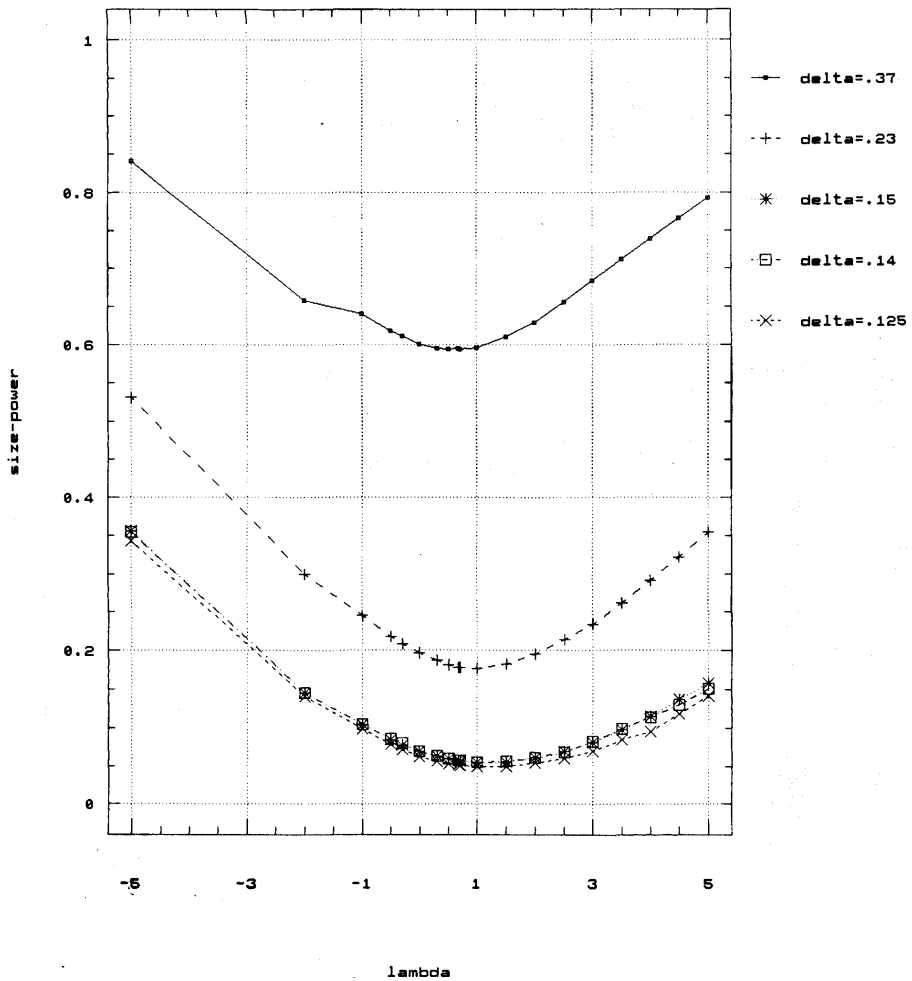


Fig. 3. Significance level ($\delta = 0.125$) and power ($\delta \neq 0.125$) of the statistics D_λ for $2 \times 2 \times 2$ contingency tables, $n = 50$ and $\alpha = .05$.

The behaviour of these powers is similar to the behaviour for 2×3 contingency tables studied before, i. e., the power increases with $|\lambda|$.

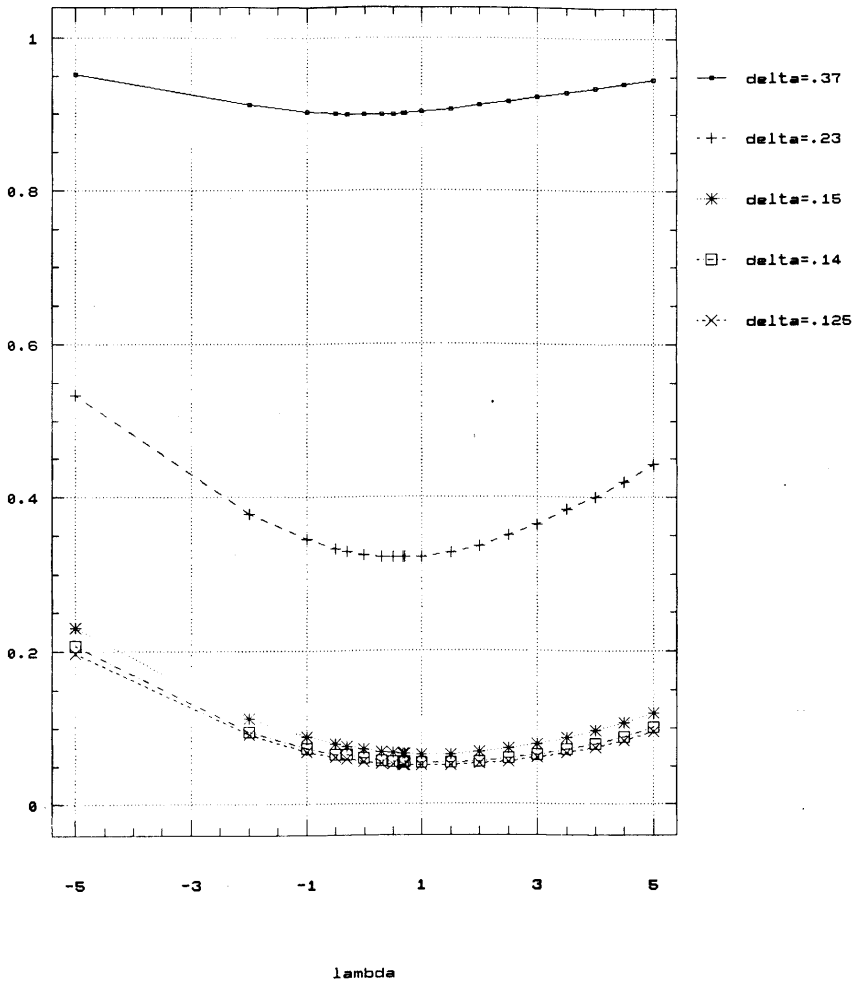


Fig. 4. Significance level ($\delta = 0.125$) and power ($\delta \neq 0.125$) of the statistics D_λ for $2 \times 2 \times 2$ contingency tables, $n = 100$ and $\alpha = .05$.

The information summarized in this section suggests that it is possible to improve the power to test independence in three-way contingency tables under the previous spike alternative hypothesis if instead of the traditional Pearson χ^2 test one chooses another appropriate member of the family of Cressie and Read statistic. The result of our simulation study suggests to prefer λ larger than 1 corresponding to Pearson's

χ^2 , but the exact location of optimal λ depends on the table dimensionality and also on the sample size. However vague, our recommendations are different from those of Cressie and Read [1] where in a goodness-of-fit testing model $\lambda = 2/3 < 1$ is recommended as a suitable alternative to the Pearson χ^2 , and also from those of Menéndez et al. [6].

ACKNOWLEDGEMENT

The author is grateful to the referees for valuable comments and suggestions.

(Received December 22, 1994.)

REFERENCES

- [1] N. Cressie and T. R. C. Read: Multinomial goodness of fit tests. *J. Roy. Statist. Soc. Ser. B* 46 (1984), 440–464.
- [2] M. Goldstein, E. Wolf and W. Dillon: On a test of independence for contingency tables. *Communications Statist. Theory Methods* 2 (1976), 159–169.
- [3] K. Matusita: Decision rules, based on the distance for problems of fit, two samples and estimation. *Ann. Math. Statist.* 26 (1964), 631–640.
- [4] M. C. Pardo: Tests of independence for multidimensional contingency tables based on (h, ϕ) -divergence measures. *Utilitas Math.* 48 (1994), 75–91.
- [5] 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. *Philosophy Magazine Series (5)* 50 (1900), 157–172.
- [6] M. Menéndez, D. Morales, L. Pardo and I. Vajda: Divergence-based estimation and testing of statistical models of classification. *J. Multivariate Anal.* 54 (1995), 2, 329–354.

Prof. María del Carmen Pardo, Departamento de Estadística e. I. O., Escuela Universitaria de Estadística, Universidad Complutense de Madrid, 28040 Madrid. Spain.