RT-MAE 9602

PERFORMANCE OF WALD'S TEST FOR THE HARDY-WEINBERG EQUILIBRIUM WITH FIXED SAMPLE SIZES

by

Julio M. Singer, Clovis A. Peres
and
Carlos E. Harle

Palavras Chaves: **Hardy-Weinberg equilibrium, MN blood group, Wald's test**
(Key words) **statistics.**

Classificação AMS: 62F04 e 62P10.
(AMS Classification)

- Janeiro de 1996 -

Performance of Wald's test for the Hardy-Weinberg equilibrium with fixed sample sizes

Julio M. Singer, Clóvis A. Peres and Carlos E. Harle

Instituto de Matemática e Estatística
Universidade de São Paulo
Caixa Postal 66281
São Paulo, SP 05389-970
Brazil

Abstract

We show via an example that Wald's test for the Hardy-Weinberg equilibrium has an aberrant behaviour for fixed sample sizes although it is asymptotically equivalent to the Likelihood Ratio and Pearson's chi-squared tests.

Key words: Hardy-Weinberg equilibrium, MN blood group, Wald's test statistics.

1 Introduction

Singer et al. (1991) have proposed Wald's tests for the Hardy-Weinberg (HW) equilibrium in generalized ABO genetic systems. Such tests are based on explicit expressions involving the observed phenotype relative frequencies and are asymptotically equivalent to the usual Pearson's chi-squared or Wilks' Likelihood Ratio tests, both of which depend on iterative computing procedures in the general case. Since, in practice, we must work with fixed sample sizes, the behaviour of the test statistics under such circumstances is of major concern. Given the results of Væth (1985), which indicate that for one parameter exponential families, Wald's test statistics may decrease to zero as the parameter estimate moves away from the null value, we conjecture that a similar behaviour is true for the multiparameter case as well.

In this note we show that such an aberrant behaviour is followed by the proposed Wald's test for the HW equilibrium in the special case of the MN blood group classification system. In Section 2 we introduce the model and the proposed Wald's statistic and in Section 3 we show that for fixed sample sizes this test statistic may approach zero in situations where the parameters' estimates move towards directions where the HW model is not tenable.

2 Tests for the HW model in the MN blood system

The MN blood group classification system is a genetic system with two codominant alleles, M and N, occurring with probabilities q_M and $q_N = 1 - q_M$, respectively in a given population. We say that such population satisfies the HW (equilibrium) model if the following relations hold:

$$p_M = q_M^2, \quad p_N = q_N^2 \text{ and } p_{MN} = 2q_M q_N \tag{2.1}$$

where p_M, p_N and p_{MN}, $(p_M + p_N + p_{MN} = 1)$ denote the probabilities with which the phenotypes M, N and MN, respectively occur in the population.

A problem of general concern to geneticists is to test whether a given population satisfies the HW model based on the evidence provided by a sample of n units for which the observed phenotype frequencies are n_M, n_N and n_{MN}, $(n_M + n_N + n_{MN} = n)$.

The Likelihood Ratio test statistic for the hypothesis (2.1) is

$$Q_L = 2[n_M \log\{\frac{n_M}{n\hat{q}_M^2}\} + n_N \log\{\frac{n_N}{n(1-\hat{q}_M)^2}\} + n_{MN} \log\{\frac{n_{MN}}{2n\hat{q}_M(1-\hat{q}_M)}\}] \tag{2.2}$$

where $\hat{q}_M = (2n_M + n_{MN})/2n$. Pearson's chi-squared test statistic for the same hypothesis is

$$Q_P = \frac{\{n_M - n\hat{q}_M^2\}^2}{n\hat{q}_M^2} + \frac{\{n_N - n(1-\hat{q}_M)^2\}^2}{n(1-\hat{q}_M)^2} + \frac{\{n_{MN} - 2n\hat{q}_M(1-\hat{q}_M)\}^2}{2n\hat{q}_M(1-\hat{q}_M)}. \tag{2.3}$$

Alternatively, observing that (2.1) is equivalent to

$$\theta = \frac{p_{MN}}{2\sqrt{p_M p_N}} = 1 \tag{2.4}$$

we may follow in Singer et al. (1991), and consider a test for the HW equilibrium using Wald's statistic

$$Q_W = Q_W(\hat{p}_M, \hat{p}_N) = n\left(1 - \frac{1}{\hat{\theta}}\right)^2 \left(\frac{1}{4\hat{p}_M} + \frac{1}{4\hat{p}_N} + \frac{1}{\hat{p}_{MN}}\right)^{-1} \tag{2.5}$$

where $\hat{p}_M = n_M/n$, $\hat{p}_N = n_N/n$, $\hat{p}_{MN} = n_{MN}/n$ and $\hat{\theta} = \hat{p}_{MN}/2\sqrt{\hat{p}_M\hat{p}_N}$. Under the null hypothesis, as n increases, the distribution of Q_L, Q_P and Q_W may be approximated by a chi-squared distribution with one degree of freedom. In practice, however, we must work with fixed sample sizes (of moderate magnitude, in many instances) and the behaviour of these statistics under such circumstances must be evaluated. This is considered in the next section.

3 Behaviour of Wald's statistic for fixed sample sizes

First note that the parameter θ may be used as a measure of departure from the HW equilibrium; it was considered by Pereira and Rogatko (1984) in a

Bayesian analysis of the same problem. Now, write for simplicity, $p_N = x$ and $p_M = y$ and note that the domain of the function $\theta = \theta(x,y)$ corresponds to the two - dimensional simplex defined by $x \in \mathbb{R}$, $y \in \mathbb{R}$, $x \geq 0$, $y \geq 0$ and $x + y \leq 1$. Furthermore observe that, given $a > 0$, the set of points for which $\theta(x,y) = a$ corresponds to the curve with equation

$$y = (\sqrt{1 + (a^2 - 1)x} - a\sqrt{x})^2;$$

in particular, for the null value $a = 1$ we have

$$y = (1 - \sqrt{x})^2.$$

Note also that $\theta(x,y) \to 0$ as $(x,y) \to (p, 1-p)$, where $0 < p < 1$ and that $\theta(x,y) \to \infty$ as $(x,y) \to (0,p)$ or $(x,y) \to (p,0)$.

Let us now analyse the behaviour of the function

$$\begin{aligned} Q_W(x,y) &= n\left(1 - \frac{2\sqrt{xy}}{1-x-y}\right)^2 \left(\frac{1}{4x} + \frac{1}{4y} + \frac{1}{1-x-y}\right)^{-1} \\ &= n\left(\frac{1-x-y}{2\sqrt{xy}} - 1\right)^2 \left\{\frac{(1-x-y)^2}{16x^2y} + \frac{(1-x-y)^2}{16xy^2} + \frac{1-x-y}{4xy}\right\}^{-1} \end{aligned} \tag{3.1}$$

as $\theta(x,y)$ moves away from the null value, i.e. as x and y approach the boundaries of the domain of $\theta(x,y)$.

From the second expression in (3.1) it is clear that $Q_W(x,y) \to \infty$ as $(x,y) \to (p, 1-p)$ for any $0 < p < 1$, i.e. as $\theta(x,y) \to 0$. However, from the first expression in (3.1) we may conclude that $Q_W(x,y) \to 0$ as $(x,y) \to (0,p)$, $(x,y) \to (p,0)$ or $(x,y) \to (0,0)$ for any $0 < p < 1$ which correspond to $\theta(x,y) = \infty$. It remains to consider the behaviour of $Q_W(x,y)$ in the neighbourhood of the points $(0,1)$ and $(1,0)$. In this direction, let $y = f(x)$ where $f : [0,1]$ is a nonnegative function with derivatives $f'(x) = df(x)/dx$ and $f''(x) = d^2f(x)/dx^2$ and such that $f(1) = 0$, $-1 \leq f'(1) \leq 0$ and let us examine the behaviour of $\theta(x,y)$ and $Q_W(x,y)$ as $(x,y) \to (1,0)$ along the curve $y = f(x)$. Applying L'Hôpital's rule we get

$$\lim_{(x,y)\to(1,0)} \theta(x,y) = \lim_{x\to 1} \frac{1-x-f(x)}{2\sqrt{xf(x)}} = \begin{cases} 0 & \text{if } f'(1) \neq 0 \\ \dfrac{1}{\sqrt{2f''(1)}} & \text{if } f'(1) = 0,\ f''(1) \neq 0 \\ \infty & \text{if } f'(1) = f''(1) = 0. \end{cases} \tag{3.2}$$

Writing $A(x) = 1 - x - f(x)$ and $B(x) = A(x)\{x + f(x)\} + 4xf(x)$ we get

$$Q_W(x,y) = 16nx^2 \left\{\frac{1-x-y}{2\sqrt{xy}-1}\right\}^2 \left\{\frac{f^2(x)}{A(x)B(x)}\right\} \tag{3.3}$$

Applying L'Hôpital's rule to $f^2(x)/\{A(x)B(x)\}$ we obtain

$$\lim_{x\to 1} \frac{f^2(x)}{A(x)B(x)} = \frac{-f'(1)}{\{3f'(1)-1\}\{f'(1)+1\}} \tag{3.4}$$

and then, from (3.1)-(3.4) we may conclude that

$$\lim_{(x,y)\to(1,0)} Q_W(x,y) = \begin{cases} 0 & \text{if } f'(1) = 0,\ f''(1) \neq 0 \\ \dfrac{16nf''(1)}{\{3f'(1)-1\}\{f'(1)+1\}} & \text{if } f'(1) \neq 0 \\ \infty & \text{if } f'(1) = f''(1) = 0. \end{cases}$$

From the above discussion and (2.5) we may conclude that for fixed sample sizes, $Q_W = Q_W(\hat{p}_M, \hat{p}_N)$ may converge to zero in situations which clearly violate the HW model, indicating that such statistic should be used with caution, specially when the observed M and N phenotype proportions are close to 0 or 1. If we rewrite the expressions for Q_L and Q_P in terms of the observed phenotype proportions by substituting $n\hat{p}_M$ for n_M, $n\hat{p}_N$ for n_N, $n\hat{p}_{MN}$ for n_{MN}, and $\hat{p}_M + \hat{p}_{MN}/2$ for $\hat{q}_M$ in (2.2) and (2.3) and proceed with a similar analysis as in the case of Wald's statistic, we may conclude that even for fixed sample sizes, both the Likelihood Ratio and Pearson's statistics tend to reject the HW equilibrium hypothesis in all situations where the observed phenotype frequencies clearly violate the null hypothesis.

Since (2.4) is not the only alternative to characterize the HW equilibrium in this context, it seems worthwhile to consider the behaviour of other functions of the phenotype frequencies for such purposes. In particular, a natural substitute for (2.4) is

$$\log \theta = 0. \tag{3.5}$$

The corresponding Wald statistic for this parametrization is

$$Q_{WL} = Q_{WL}(\hat{p}_M, \hat{p}_N) = n \log^2 \hat{\theta} \left(\frac{1}{4\hat{p}_M} + \frac{1}{4\hat{p}_N} + \frac{1}{\hat{p}_{MN}} \right)^{-1}. \tag{3.6}$$

Following the same arguments employed in the case of Q_W, we may show that Q_{WL} has a similar aberrant behaviour near the boundary of the parameter space. However, by noting that

$$\lim_{(x,y)\to(p,0)} Q_{WL}(x,y)/Q_W(x,y) = \infty$$

we may conclude that $Q_{WL}(x,y)$ converges to zero at a slower rate than $Q_W(x,y)$ when $(x,y) \to (p,0)$. Similar results hold when $(x,y) \to (0,p)$, suggesting that (3.6) might be a better alternative than (2.5) for the purposes of testing the HW equilibrium.

Further evidence of these facts are suggested in Figures 3.1 - 3.4, where the values of the four test statistics discussed above are plotted as functions of n_M for $n = 40$ and $n_N = 1, 2, 3, 4, 5$ and 15. First, the non-monotone nature of either Q_W or Q_{WL} as functions of $\hat{\theta}$ clearly contrasts with the corresponding (and desirable) monotone behaviour of Q_L or Q_P (note that the value of $\hat{\theta}$ increases as n_M approches 0). Also, the more stable behaviour of Q_{WL} with respect to that of Q_W is depicted in Figures 3.1 and 3.2.

Finally, in Table 3.1 we present numerical examples of situations where the aberrant behaviour of the Wald's statistic mentioned above is evident. For $n = 30$, Q_W does not lead to the rejection of the HW equilibrium even with $\hat{\theta} = 14.00$ while the other three alternative statistics do so. A similar picture holds for $n = 40$ and seems less evident for $n = 50$, as expected. Although for larger values of n, (e.g. 100 or more) erroneous conclusions will only occur for pathological cases (e.g. n_M or n_N equal 0 or 1), this note may serve as a warning against a routine application of the type of statistics investigated here. A similar analysis for the general case discussed in Singer et al. (1991) seems mathematically intractable; however, we believe that the corresponding Wald's statistics have the same type of aberrant behaviour in that case, too.

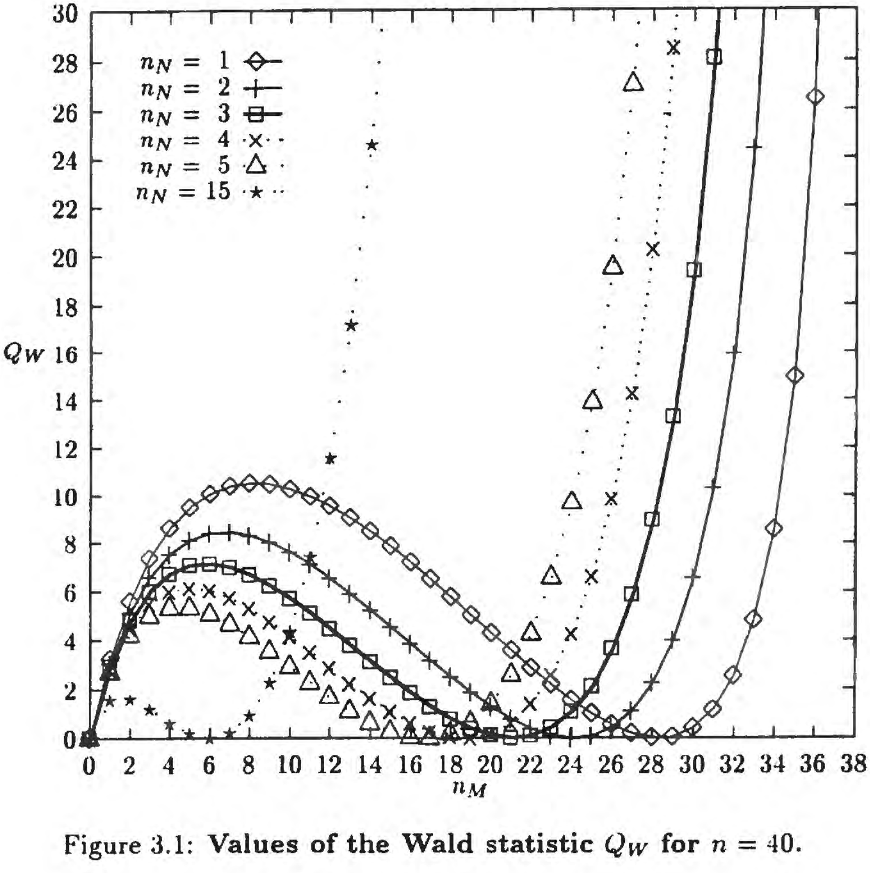

Figure 3.1: **Values of the Wald statistic Q_W for $n = 40$.**

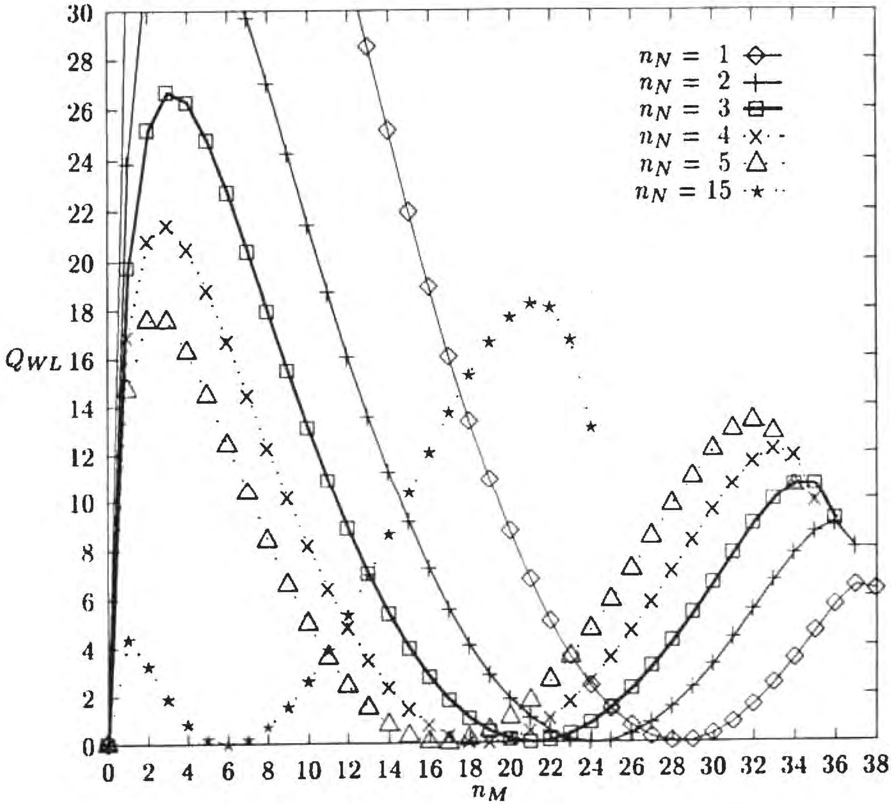

Figure 3.2: **Values of the repametrized Wald statistic Q_{WL} for $n = 40$.**

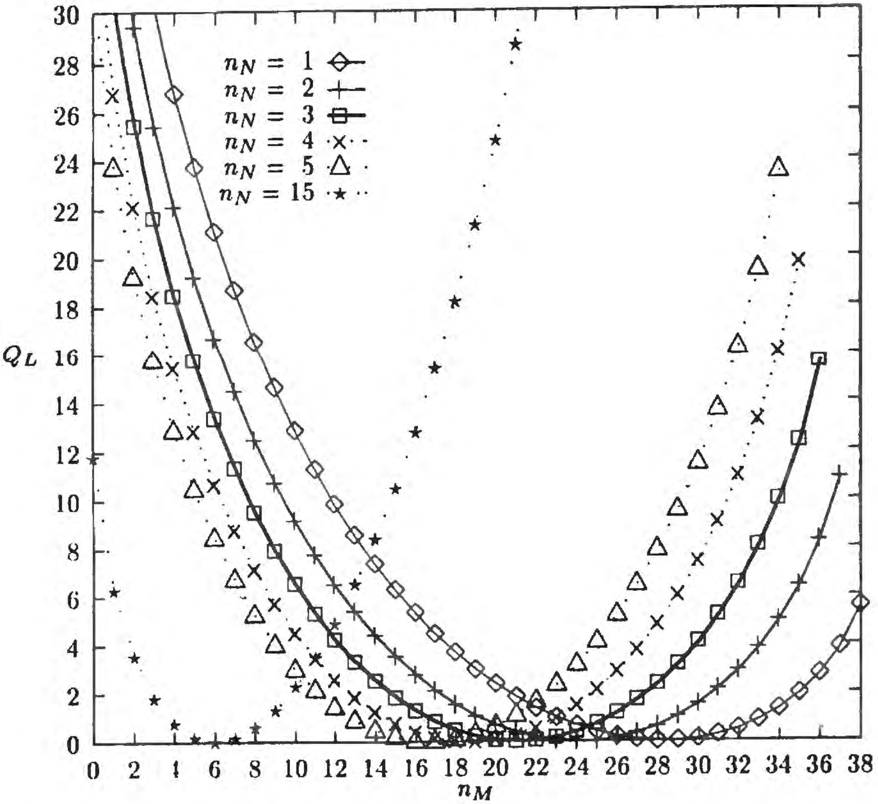

Figure 3.3: **Values of the Likelihood Ratio statistic Q_L for $n = 40$.**

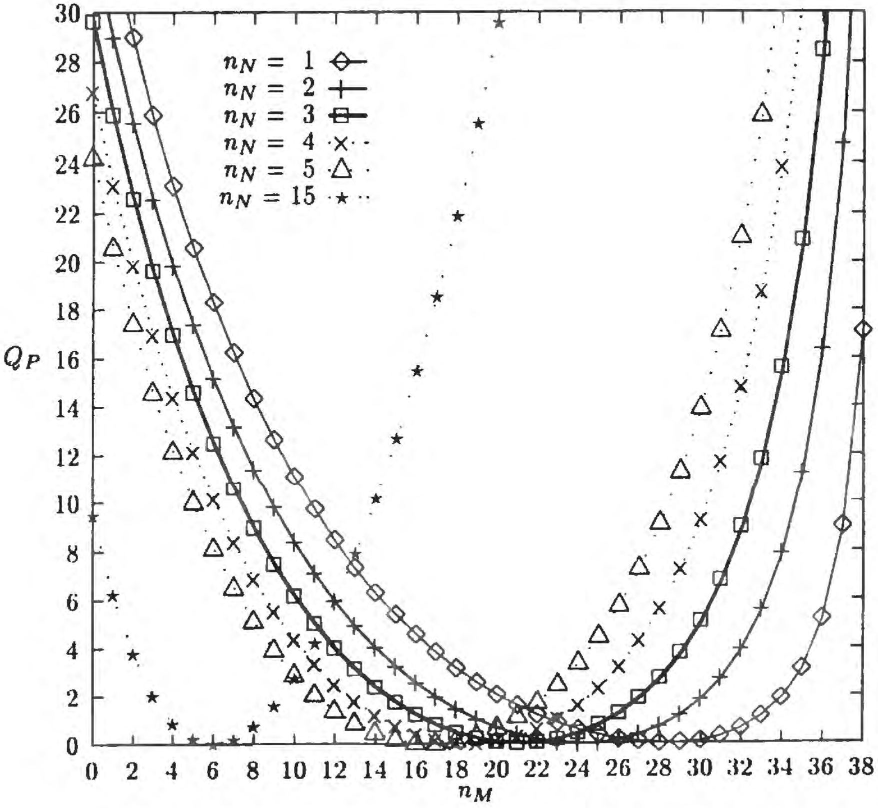

Figure 3.4: **Values for Pearson's statistic Q_P for $n = 40$.**

Table 3.1: **Numerical examples.**

Phenotype frequencies			Total	HW eq. param.	Test statistics			
n_M	n_N	n_{MN}	n	$\hat{\theta}$	Q_W	Q_{WL}	Q_L	Q_P
1	1	28	30	14.00	1.61	13.00	26.89	22.53
2	2	26	30	6.50	2.48	12.15	18.03	16.13
2	3	25	30	5.10	2.60	10.70	14.69	13.41
3	3	24	30	4.00	2.70	9.22	11.56	10.80
3	4	23	30	3.32	2.58	7.61	9.07	8.59
4	4	22	30	2.75	2.38	6.00	6.79	6.53
5	5	20	30	2.00	1.67	3.20	3.40	3.33
1	1	38	40	19.00	1.71	16.47	39.57	32.40
2	2	36	40	9.00	2.84	17.38	29.45	25.60
2	3	35	40	7.14	3.12	16.32	25.46	22.57
3	3	34	40	5.67	3.46	15.35	21.64	19.60
3	4	33	40	4.76	3.54	13.83	18.45	16.95
4	4	32	40	4.00	3.60	12.30	15.42	14.40
5	5	30	40	3.00	3.33	9.05	10.46	10.00
1	1	48	50	24.00	1.76	19.39	52.52	42.32
2	2	46	50	11.50	3.07	21.95	41.44	35.28
2	3	45	50	9.19	3.44	21.33	36.97	32.06
3	3	44	50	7.33	3.94	20.96	32.62	28.88
3	4	43	50	6.21	4.16	19.71	28.92	25.97
4	4	42	50	5.25	4.40	18.48	25.35	23.12
5	5	40	50	4.00	4.50	15.37	19.27	18.00

Acknowledgements

Work of the first author was partially supported by a grant from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil. The authors are grateful to the associate editor and to the referees for their enlightening comments.

References

Pereira, C.A.B. and Rogatko, A. (1984). The Hardy-Weinberg equilibrium under a Bayesian perspective. *Revista Brasileira de Genética*, **7**, 689-707.

Singer, J.M., Peres, C.A. and Harle, C.E. (1991). A note on the Hardy-Weinberg model in generalized ABO systems. *Statistics and Probability Letters*, **11**, 173-175.

Væth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review*, **53**, 199-214.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

9501 - FAVA, L. On Interpenetrating Order Statistics. São Paulo, IME-USP. 1995. 32p. (RT-MAE-9501)

9502 - FERRARI,P.A.,KESTEN,H., MARTÍNEZ,S. R-Positivity, Quasi-Stationary Distributions and Ratio Limit Theorems for a Class of Probabilistic Automata. São Paulo, IME-USP. 1995. 39p. (RT-MAE-9502)

9503 - FERRARI, S.L.P., CORDEIRO, G.M., URIBE-OPAZO, M.A., CRIBARI-NETO, F. Improved Score Tests for One-Parameter Exponential Family Models. São Paulo, IME-USP. 1995. 21p. (RT-MAE-9503)

9504 - ARIZONO, H., BUENO, V.C. Bounds for System Reliability of Used Components under MIFR/$\Im_t$ Assumptions. São Paulo, IME-USP. 1995. 10p. (RT-MAE-9504)

9505 - SINGER, J.M., ANDRADE, D.F. Regression Models for the Analysis of Pretest/Posttest Data. São Paulo, IME-USP. 1995. 8p. (RT-MAE-9505)

9506 - SILVA, A.J., ZUAZOLA, P.L.I., WECHSLER, S. Partial exchangeability and extendibility. São Paulo, IME-USP. 1995. 21p. (RT-MAE-9506)

9507 - MADRUGA, M.G., OCHI-LOHMANN, T.H., OKAZAKI, K., PEREIRA, C.A.B., RABELLO-GAY, M.Z. Bayesian Dosimetry II: Credibility Intervals for Radiation Dose. São Paulo, IME-USP. 1995. 12p. (RT-MAE-9507)

9508 - FERRARI, S.L.P.; URIBE-OPAZO, M.A.; CRIBARI-NETO, F. Second order asymptotics for score tests in exponential family nonlinear models. São Paulo, IME-USP. 1995. 07p. (RT-MAE-9508)

9509 - WECHSLER, S.; BALLAS, D.; CARDOSO, R.H.A. Bayesian Optimization of Quality-of-life. São Paulo, IME-USP. 1995. 06p. (RT-MAE-9509)

9510 - BOTTER, D.A; CORDEIRO, G.M. Bartlett Corrections for Generalized Linear Models with Dispersion Covariates. São Paulo, IME-USP. 199g5. 31p. (RT-MAE-9510)

9511 - PAULA, G.A.; ROJAS, O.V. On Restricted Hypotheses in Extreme Value Regression Models. São Paulo, IME-USP. 1995. 14p. (RT-MAE-9511)

9512 - ARIZONO, H.; BUENO, V.C. Bounds for System Reliability of Used Components Under $NBU|\mathfrak{I}_t$ Assumptions. São Paulo, IME-USP. 1995. 10p. (RT-MAE-9512)

9513 - FERRARI, S.L.P.; BOTTER, D.A.; CORDEIRO, G.M.; CRIBARI-NETO, F. Second and Third Order Bias Reduction for One-Parameter Family Models. São Paulo, IME-USP. 1995. 07p. (RT-MAE-9513)

9514 - BELITSKY, V.; KOHAYAKAWA, Y. Inference on the Flip Rates from Density Dynamics for a Class of Spin-Flip Systems. São Paulo, IME-USP. 1995. 21p. (RT-MAE-9514)

9515 - BELITSKY, V. The change of sign of derivatives of the density function of a spin flip system with the order of derivation. São Paulo, IME-USP. 1995. 27p. (RT-MAE-9515)

9516 - PAULA, G.A. On Approximation of the Level Probabilities for Testing Ordered Parallel Regression Lines. São Paulo, IME-USP. 1995. 15p. (RT-MAE-9516)

9517 - PAULA, G.A. The Concept of Leverage in Restricted Regression. São Paulo, IME-USP. 1995. 09p. (RT-MAE-9517)

9518 - YOSHIDA, O.S.; LEITE, J.G.; BOLFARINE, H. Bayes Estimation of the Number of Component Processes of a Superimposed Process. São Paulo, IME-USP. 1995. 19p. (RT-MAE-9518)

9519 - SIMONIS, A. Metastability of the d-Dimensional Contact Process. São Paulo, IME-USP. 1995. 17p. (RT-MAE-9519)

9520 - FONTES, L.R.G.; ISOPI, M.; SIDORAVICIUS, V. Analyticity of the Density and Exponential Decay of Correlations in 2-d Bootstrap Percolation. São Paulo, IME-USP. 1995. 17p. (RT-MAE-9520)

9521 - FERRARI,S.L.P.; BOTTER, D.A.; CRIBARI-NETO, F. Local power of three classic criteria in generalised linear medels with unknown dispersion. São Paulo, IME-USP. 1995. 5p.(RT-MAE-9521)

9522 - CORDEIRO, G.M.; FERRARI, S.L.P.; A general method for approximating to the distribution of some statistics. São Paulo, IME-USP. 1995. 18p. (RT-MAE-9522)

9523 - BOLFARINE, H.; LIMA, C.; SANDOVAL, M.C. Linear Calibration in Functional Regression Models. São Paulo, IME-USP. 1995.24p. (RT-MAE-9523)

9524 - WECHSLER, S.; BRANCO, M.D. The exercise of coherent statistical inference. São Paulo, IME-USP. 1995. 8p. (RT-MAE-9524)

9525 - ROJAS, O.V.; BOLFARINE, H.; PAULA, G.A. On restricted hypothesis in structural regression models. São Paulo, IME-USP. 1995. 16p. (RT-MAE-9525)

9526 - RODRIGUES, J.; BOLFARINE, H.; CORDEIRO, G.M. NonLinear Quasi-Bayesian Theory and Inverse Linear Regression. São Paulo, IME-USP. 1995. 14p.(RT-MAE-9526)

9527 - INOUE, L.Y.T.; WECHSLER, S. Classes of comparable experiments. São Paulo, IME-USP. 1995. 7p. (RT-MAE-9527)

9528 - CHIANN, C.; MORETTIN, P.A. A Wavelet Analysis for Stationary Processes. São Paulo, IME-USP. 1995. 40p. (RT-MAE9528)

9529 - SÁFADI, T.; MORETTIN, P.A. A Bayesian Analysis of the Threshold Arma Model.São Paulo, IME-USP. 1995. 13p. (RT-MAE9529)

9530 - FONTES, L.R.G.; ISOPI, M. Bouds for gap of the metropolis dynamicss for the Random Energy Model. São Paulo, IME-USP. 1995. 09p. (RT-MAE9530)

9531 - BUENO, V.C. A Note on Simulation of Total Time on Test. São Paulo. IME-USP. 1995. 08p. (RT-MAE9531)

9532 - CRIBARI-NETO, F.; BOTTER, D.A.; CORDEIRO, G.M.; FERRARI, S.L.P. Bias Reduction in One-Parameter Exponential Family Models. São Paulo. IME-USP. 1995. 21p. (RT-MAE9532)

9601 - MENTZ, R.P.; MORETTIN, P.A. and TOLOI, C.M.C. Bias correction for estimators of the residual variance in the ARMA (1,1) Model. São Paulo, IME-USP, 1995. 21p. (RT-MAE-9601)

The complete list of Relatórios do Departamento de Estatística, IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05389-970 - São Paulo, Brasil