

STATISTIQUE ET ANALYSE DES DONNÉES

JEAN-MARIE DUFOUR

MARC HALLIN

An exponential bound for the permutational distribution of a first-order autocorrelation coefficient

Statistique et analyse des données, tome 15, n° 1 (1990), p. 45-56

http://www.numdam.org/item?id=SAD_1990__15_1_45_0

© Association pour la statistique et ses utilisations, 1990, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

AN EXPONENTIAL BOUND FOR THE PERMUTATIONAL DISTRIBUTION OF A FIRST-ORDER AUTOCORRELATION COEFFICIENT¹

Jean-Marie DUFOUR

Marc HALLIN

CRDE
Université de Montréal
C.P. 6128 - succursale A
Montréal - Québec - H3C 3J7
CANADA

CEME and Institut de Statistique
Université Libre de Bruxelles
Campus de la Plaine, C.P. 210
B- 1050 Bruxelles
BELGIQUE

Abstract

Denote by $X = (X_1, \dots, X_n)$ a time series of length n , and consider the null hypothesis H_0 under which the X_i 's are independently (possibly nonidentically) distributed, with unspecified distributions symmetric about a common known median. If H_0 is to be tested against first-order serial dependence, classical similarity and Neyman structure arguments lead one to consider conditional tests, viz. tests based on the distribution of the first-order autocorrelation coefficient r_1 conditional on the series of absolute values. In practice, however, exact critical values for such tests are difficult to derive. We propose an exponential bound for the tail areas of the conditional null-distribution of r_1 . This bound is related to a bound of Edelman (1986) for the tails of permutation t-tests, and allows one to obtain conservative critical values. Finally, a conjecture of Edelman (1986) on a least conservative bound for this type of situation is discussed and disproved by means of three counterexamples.

Keywords : permutation test, similar test, serial correlation, heteroskedasticity, exponential bounds, Neyman structure.

Classification AMS : 62 M 10, 62 G 10

¹ This work has benefited from the financial support of the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Government of Québec (Fonds FCAR and Ministère des Relations internationales) and the Communauté française de Belgique (Commissariat général aux Relations internationales).

Résumé

Soit $X = (X_1, \dots, X_n)$ une série de longueur n . On considère l'hypothèse nulle H_0 sous laquelle les observations X_i sont indépendantes et admettent des distributions symétriques par rapport à une médiane commune, supposée connue; ces distributions ne sont cependant pas forcément identiques entre elles, et demeurent non spécifiées. Si H_0 doit être testée par rapport à une contre-hypothèse de dépendance sérielle du premier ordre, des arguments classiques de similitude et de structure de Neyman conduisent à la considération de tests conditionnels. Ceux-ci sont fondés sur la distribution conditionnelle, sachant la série des valeurs absolues, du coefficient d'autocorrélation usuel r_1 . En pratique, les valeurs critiques de tels tests sont cependant difficiles à calculer de façon explicite. C'est pourquoi nous proposons une borne, de type exponentiel, pour les probabilités de queue des distributions conditionnelles de r_1 sous H_0 . Cette borne est apparentée à une borne proposée par Edelman (1986) pour les probabilités de queue dans les test t permutatinnels. Elle permet la construction de valeurs critiques conservatives. Enfin, une conjecture d'Edelman (1986) à propos d'une borne optimale pour ce genre de situation est réfutée au moyen de trois contre-exemples.

Mots-Clefs : test de permutation, test semblable, corrélation sérielle, hétéroscédasticité, bornes exponentielles, structure de Neyman.

1. INTRODUCTION

Denote by $\mathbf{X} = (X_1, \dots, X_n)$ a series of length n . In many situations of practical interest, one may wish to test the null hypothesis that the X_t 's are independently distributed, against alternatives of first-order serial dependence. If the observations under the null hypothesis were assumed to be identically distributed with known median m_0 , a classical solution would consist in rejecting the hypothesis for large (or small) values of the first-order autocorrelation coefficient

$$r_1 = \frac{\sum_{t=1}^{n-1} (X_t - m_0) (X_{t+1} - m_0)}{\sum_{t=1}^n (X_t - m_0)^2}. \quad (1.1)$$

Under fairly general regularity conditions the distribution of r_1 is known to be approximately normal, with mean zero and standard error $n^{-1/2}$; see e.g. Anderson (1971) or Brockwell and Davis (1987).

Now, if the X_t 's under the null hypothesis cannot be assumed to be identically distributed, this procedure is clearly inappropriate. In this paper, we deal with the much more general null hypothesis H_0 under which the observations X_t are independent but possibly nonidentically distributed, with distributions symmetric about a common specified median m_0 . Further, the distributions of the observations may be discrete. Without loss of generality, the median m_0 can be assumed to be zero, and we henceforth set $m_0 = 0$. A sufficient complete statistic for H_0 is then the series $|\mathbf{X}| = (|X_1|, \dots, |X_n|)$ of absolute values. Further, classical arguments of similarity and Neyman structure lead one to consider tests which are conditional with respect to the complete sufficient statistic $|\mathbf{X}|$; see Lehmann (1986, chapter 4).

Denoting the sign of x by

$$\begin{aligned} \text{sgn}(x) &= -1, & \text{if } x < 0, \\ &= 0, & \text{if } x = 0, \\ &= 1, & \text{if } x > 0, \end{aligned} \quad (1.2)$$

the first-order autocorrelation coefficient r_1 can be written as

$$r_1 = \frac{\sum_{t=1}^{n-1} X_t X_{t+1}}{\sum_{t=1}^n X_t^2} = \sum_{t=1}^{n-1} \operatorname{sgn}(X_t X_{t+1}) c_t, \quad (1.3)$$

where

$$c_t = \frac{|X_t X_{t+1}|}{\sum_{s=1}^n X_s^2}. \quad (1.4)$$

Accordingly, conditional on $|X|$, r_1 is a weighted sum of the variables $\operatorname{sgn}(X_t X_{t+1})$. As shown in Section 2 below, such sums are conditionally distribution-free and thus theoretically allow for the derivation of exact α -level critical values.

In practice, however, the conditional distribution of r_1 depends on each specific sample and may be difficult to compute (e.g., by enumeration), a problem also met in the more traditional case of permutation t-tests. A possible remedy to this consists in replacing tight critical points with conservative ones obtained from bounds on the tail areas of permutational distributions. A simple exponential bound of this type was proposed recently for the t-test by Edelman (1986); for related results, see also Hoeffding (1963), Efron (1969) and Eaton (1970). In Section 2 of this paper, we establish a bound similar to Edelman's for the distribution of the autocorrelation coefficient r_1 ; for a more general case, with other bounds, see Dufour and Hallin (1990). Our main result (Proposition 1) provides a simple exponential bound on the conditional tail areas of r_1 given $|X|$. Further, critical values yielding conservative unconditional tests based on r_1 are derived from the latter.

Along with his exponential bound, Edelman (1986) also stated (in a somewhat ambiguous form) a conjecture on a least conservative bound that would yield a considerable improvement over the exponential one. The most natural and straightforward interpretation of this conjecture is discussed in Section 3, and disproved by means of three counterexamples.

2. EXPONENTIAL BOUND

In the following proposition, we derive a simple exponential bound for the conditional tail areas of the autocorrelation coefficient r_1 when the observations X_1, \dots, X_n are independent with distributions symmetric about zero.

Proposition 1: Let X_1, \dots, X_n be independent random variables with distributions symmetric about zero, $|X| = (|X_1|, \dots, |X_n|)$, and

$$r_1 = \frac{\sum_{t=1}^{n-1} X_t X_{t+1}}{\sum_{t=1}^n X_t^2},$$

with $r_1 \equiv 0$ for $X_1 = X_2 = \dots = X_n = 0$. Then, for all $y > 0$,

$$P \left[r_1 \geq y \mid |X| \right] \leq \exp \left[- \frac{y^2}{2 D(|X|)^2} \right], \tag{2.1}$$

where

$$D(|X|) = \left[\sum_{t=1}^{n-1} X_t^2 X_{t+1}^2 \right]^{1/2} / \sum_{t=1}^n X_t^2 \tag{2.2}$$

and $D(|X|) \equiv 0$ when $X_1 = X_2 = \dots = X_n = 0$.

Proof : When $X_1 = X_2 = \dots = X_n = 0$, we have $r_1 = 0$ and $P[r_1 \geq y \mid |X|] = 0$ for $y > 0$, so that (2.1) holds trivially. Suppose now that $X_t \neq 0$ for at least one t and set $0/0 = 0$. Then we can write

$$\begin{aligned} r_1 &= D(|X|) \frac{\sum_{t=1}^{n-1} X_t X_{t+1}}{\left[\sum_{s=1}^{n-1} X_s^2 X_{s+1}^2 \right]^{1/2}} \\ &= D(|X|) \sum_{t=1}^{n-1} w_t \operatorname{sgn}(X_t X_{t+1}), \end{aligned} \tag{2.3}$$

where $w_t = |X_t X_{t+1}| / \left[\sum_{s=1}^{n-1} X_s^2 X_{s+1}^2 \right]^{1/2}$. Let $A(|X|) = \{t : |X_t| \neq 0, 1 \leq t \leq n\}$.

Since each X_t is symmetrically distributed with respect to zero, though possibly with a probability mass at zero, we have, for each $t \in A(|X|)$,

$$P \left[\operatorname{sgn}(X_t) = -1 \mid |X_t| \right] = P \left[\operatorname{sgn}(X_t) = 1 \mid |X_t| \right] = \frac{1}{2}.$$

Further, since X_1, \dots, X_n are mutually independent, the variables in the set $\{\operatorname{sgn}(X_t) : t \in A(|X|)\}$ are mutually independent conditional on $|X|$, with

$$P \left[\operatorname{sgn}(X_t) = -1 \mid |X| \right] = P \left[\operatorname{sgn}(X_t) = 1 \mid |X| \right] = \frac{1}{2}. \quad (2.4)$$

Let $Z_t = X_t X_{t+1}$, $t = 1, \dots, n-1$, and $B(|X|) = \{t : |Z_t| \neq 0, 1 \leq t \leq n-1\}$. If the set $B(|X|)$ is empty, we have $r_1 = 0$ and (2.1) again holds trivially. We can thus assume that $B(|X|)$ is not empty. Clearly, $t \in B(|X|)$ if and only if $t \in A(|X|)$ and $t+1 \in A(|X|)$. Conditional on $|X|$, the variables in the set $\{\operatorname{sgn}(X_t) : t \in A(|X|) \text{ and } t+1 \in A(|X|)\}$ are mutually independent and satisfy (2.4). Thus the variables in the set $\{\operatorname{sgn}(Z_t) : t \in B(|X|)\}$ are also mutually independent conditional on $|X|$, with

$$P \left[\operatorname{sgn}(Z_t) = -1 \mid |X| \right] = P \left[\operatorname{sgn}(Z_t) = 1 \mid |X| \right] = \frac{1}{2}; \quad (2.5)$$

for a similar argument, see Dufour (1981, proof of Theorem 1). Let $S_t = \operatorname{sgn}(Z_t)$, $t = 1, \dots, n-1$, and

$$R_1 = \sum_{t=1}^{n-1} w_t S_t = \sum_{t \in B(|X|)} w_t S_t, \quad (2.6)$$

where $\sum_{t=1}^{n-1} w_t^2 = 1$. The moment generating function of R_1 given $|X|$ is

$$\begin{aligned} E \left\{ \exp[\tau R_1] \mid |\mathbf{X}| \right\} &= \prod_{t \in B(|\mathbf{X}|)} E \left\{ \exp[\tau w_t S_t] \mid |\mathbf{X}| \right\} \\ &= \prod_{t \in B(|\mathbf{X}|)} \cosh(w_t \tau) = \prod_{t=1}^{n-1} \cosh(w_t \tau), \end{aligned}$$

for all real τ , where the last identity follows by observing that $\cos(w_t \tau) = \cosh(0) = 1$ for $t \notin B(|\mathbf{X}|)$. By Markov's inequality,

$$\begin{aligned} P \left[R_1 \geq y \mid |\mathbf{X}| \right] &\leq P \left[\exp(\tau R_1) \geq \exp(\tau y) \mid |\mathbf{X}| \right] \tag{2.7} \\ &\leq E \left[\exp(\tau R_1) \right] / \exp(\tau y) \\ &\leq \prod_{t=1}^{n-1} \cosh(w_t \tau) / \exp(\tau y) \end{aligned}$$

for all $\tau \geq 0$ and $y > 0$. Taking $\tau = y$, we then obtain, for $y > 0$,

$$\begin{aligned} P \left[R_1 \geq y \mid |\mathbf{X}| \right] &\leq \exp(-y^2) \prod_{t=1}^{n-1} \cosh(w_t y) \tag{2.8} \\ &\leq \exp(-y^2) \prod_{t=1}^{n-1} \exp(w_t^2 y^2 / 2) = \exp(-y^2 / 2), \end{aligned}$$

where the last inequality follows from $\cosh(x) \leq \exp(x^2/2)$; on this latter inequality, see Edelman (1986). Finally, since $r_1 = D(|\mathbf{X}|) R_1$ and $D(|\mathbf{X}|) > 0$, we get

$$\begin{aligned} P \left[r_1 \geq y \mid |\mathbf{X}| \right] &= P \left[R_1 \geq \frac{y}{D(|\mathbf{X}|)} \mid |\mathbf{X}| \right] \\ &\leq \exp \left[-y^2 / 2 D(|\mathbf{X}|)^2 \right] \end{aligned}$$

for any $y > 0$. Q.E.D.

From (2.1), we see that $\exp\{-r_1^2 / 2 D(|\mathbf{X}|)^2\}$ provides an upper bound on p-values for (right-one-sided) sign-permutation tests of independence based on r_1 , i.e. for tests based on the conditional distribution of r_1 given $|\mathbf{X}|$. Note also that the latter distribution is symmetric about zero (this is easy to see from (2.3) or (2.6) above), so that bounds on p-values for left-one-sided and two-sided tests may be computed from inequality (2.1) in a similar way.

Since the bound in (2.1) is valid for any vector of absolute values $|\mathbf{X}|$, it is straightforward to derive unconditional bounds from it. The following corollary restates inequality (2.1) in a possibly more attractive form and gives the corresponding unconditional bound.

Corollary 1: Under the assumptions of Proposition 1,

$$P\left[r_1 > y D(|\mathbf{X}|) \mid |\mathbf{X}|\right] \leq e^{-y^2/2} \quad (2.9)$$

for any $|\mathbf{X}|$ and $y > 0$, and hence

$$P\left[r_1 > y D(|\mathbf{X}|)\right] \leq e^{-y^2/2} \quad (2.10)$$

for all $y > 0$.

In (2.9) and (2.10), \geq is replaced by $>$ because $D(|\mathbf{X}|)$ can be zero. Using (2.9) and the fact that the unconditional distribution of r_1 is also symmetric about zero, it is easy to select $y = c(\alpha)$ so that $c(\alpha) D(|\mathbf{X}|)$ is a conservative critical value for a level- α (one-sided or two-sided) conditional test of the null hypothesis. Clearly, the same critical value $c(\alpha) D(|\mathbf{X}|)$ also yields an unconditional unbiased test, the size of which does not exceed α .

3. ON A CONJECTURE OF EDELMAN

The bound given by Edelman (1986) as well as the one we provide in Proposition 1 follow from bounding the distribution of an arbitrary linear combination of independent Bernoulli variables [see (2.6) above]. Considering this latter problem, Edelman (1986) stated a conjecture on a least conservative bound that would yield a considerable improvement over the exponential one. In this section, we disprove this conjecture by three counterexamples.

Let U_1, \dots, U_n be independent Bernoulli random variables such that $P(U_i = 1) = P(U_i = -1) = 1/2, i = 1, \dots, n$, and consider n constants a_1, \dots, a_n , not all zero. Edelman's conjecture states that "distributions of sums of the form

$Y = \sum_{i=1}^n a_i U_i / \left(\sum_{j=1}^n a_j^2 \right)^{1/2}$ with the largest tail probabilities are the binomial distributions averaged at all points of discontinuity". As stated, this implies the inequality

$$P \left[\left(\sum_{i=1}^n a_i U_i \right) / \left(\sum_{j=1}^n a_j^2 \right)^{1/2} > x \right] \leq P \left[n^{-1/2} \sum_{i=1}^n U_i > x \right] \tag{3.1}$$

for any n -tuple (a_1, \dots, a_n) such that $\sum_{i=1}^n a_i^2 > 0$, at all positive continuity points x of the distribution function of $n^{-1/2} \sum_{i=1}^n U_i$. For discontinuity points, Edelman proposes an averaging device which, however, does not affect the validity of the conjecture.

At first sight, this conjecture looks highly plausible. Among all sums of the form considered, $Z_{(n)} = n^{-1/2} (U_1 + \dots + U_n)$ indeed yields the largest and smallest possible values ($n^{1/2}$ and $-n^{1/2}$, respectively), the largest moments of all orders (Efron, 1969) and, more generally, the largest expectations for a fairly general class of positive functions (Eaton, 1970). Nevertheless, as shown below, (3.1) is not generally true : Examples 1 and 2 are simple counterexamples based on small values of n , while Example 3 shows that the conjecture need not be true for arbitrary values of n .

Example 1. Let $n = 2$, and consider $a_1 = 1, a_2 = 2$, i.e. $Y = (U_1 + 2U_2) / \sqrt{5}$. The probability functions of $Z_{(2)} = (U_1 + U_2) / \sqrt{2}$ and Y are

| | | | |
|------------------|-------------|------|------------|
| x | $-\sqrt{2}$ | 0 | $\sqrt{2}$ |
| $P[Z_{(2)} = x]$ | 0.25 | 0.50 | 0.25 |

and

| | | | | |
|------------|---------------|---------------|--------------|--------------|
| x | $-3/\sqrt{5}$ | $-1/\sqrt{5}$ | $1/\sqrt{5}$ | $3/\sqrt{5}$ |
| $P[Y = x]$ | 0.25 | 0.25 | 0.25 | 0.25 |

respectively. Though 2 is the smallest non-trivial value for n , (3.1) clearly does not hold : for $0 < x < 1/\sqrt{5} \approx 0.4472$,

$$P[Y > x] = 0.50 > 0.25 = P[Z_{(2)} > x].$$

Counterexample 1 is not relevant for applications. The following case is perhaps more convincing, since it shows that Edelman's conjectured conservative critical value can be too liberal at the 10 % level (one-sided).

Example 2. Let $n = 4$, $a_1 = a_2 = a_3 = 1$ and $a_4 = 5$, so that $Y = (U_1 + U_2 + U_3 + 5U_4) / \sqrt{28}$. The probability functions of $Z_{(4)} = (U_1 + U_2 + U_3 + U_4) / 2$ and Y are (for $x \geq 0$)

| | | | |
|------------------|------|------|------|
| x | 0 | 1 | 2 |
| $P[Z_{(4)} = x]$ | 6/16 | 4/16 | 1/16 |

and

| | | | | | |
|------------|-----|---------------|---------------|---------------|---------------|
| x | 0 | $2/\sqrt{28}$ | $4/\sqrt{28}$ | $6/\sqrt{28}$ | $8/\sqrt{28}$ |
| $P[Y = x]$ | 0 | 1/16 | 3/16 | 3/16 | 1/16 |

respectively. Clearly, for $1 < x < 6 / \sqrt{28} \approx 1.3389$, $P[Y > x] = 0.2500 > P[Z_{(4)} > x] = 0.0625$. The upper 10 % cutoff point for $Z_{(4)}$ is 1, which is severely non-conservative with respect to the actual distribution of Y , since $P[Y > 1] = 0.25$.

Example 3 (arbitrary n). Let $n \geq 2$, $a_1 = \dots = a_{n-2} = n^{-1/2}$, $a_{n-1} = n^{-1/2} + \epsilon_1$, $a_n = n^{-1/2} - \epsilon_2$, where ϵ_1 and ϵ_2 are positive, such that $\epsilon_1 + \epsilon_2 < n^{-1/2}$ and $\sum_{i=1}^n a_i^2 = 1$.

It is easy to check that this is possible, for example by taking $0 < \epsilon_1 < n^{-1/2} (\sqrt{3} - 1)/2$ and $\epsilon_2 = n^{-1/2} - (n^{-1} - 2\epsilon_1 n^{-1/2} - \epsilon_2^2)^{1/2}$. In such a case, we also have $-n^{-1/2} < \epsilon_1 - \epsilon_2 < 0$. Then, for $n^{-1/2} (n - 2) < x < n^{-1/2} (n - 2) + \epsilon_1 + \epsilon_2$,

$$\begin{aligned} P[Y > x] &= P [\{Y = n^{1/2} + (\epsilon_1 - \epsilon_2)\} \cup \{Y = n^{-1/2} (n - 2) + \epsilon_1 + \epsilon_2\}] \\ &\geq P [U_1 = \dots = U_{n-2} = 1] P [\{U_{n-1} = U_n = 1\} \cup \{U_{n-1} = 1, U_n = -1\}] \\ &= (1/2)^{n-2} [(1/2)^2 + (1/2)^2] = (1/2)^{n-1}, \end{aligned}$$

whereas $P [Z_{(n)} > x] = P [Z_{(n)} = n^{1/2}] = (1/2)^n$. Hence

$$P[Y > x] \geq (1/2)^{n-1} > (1/2)^n = P [Z_{(n)} > x]$$

for $n^{-1/2} (n - 2) < x < n^{-1/2} (n - 2) + \epsilon_1 + \epsilon_2$. Accordingly (3.1) may not hold for arbitrary n and values of x close to the largest possible value $n^{1/2}$ of $Z_{(n)}$.

REFERENCES

Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. New York : Wiley.

Brockwell, P.J. and Davis, R.A. (1987). *Time Series : Theory and Methods*. New York : Wiley.

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

Dufour, J.-M. (1981). Rank tests for serial dependence. *Journal of Time Series Analysis* 2, 117-128.

Dufour, J.-M. and Hallin, M. (1990). *Bounds for nonparametric tests against serial dependence*. Unpublished manuscript.

Eaton, M.L. (1970). A note on symmetric Bernoulli random variables. *Annals of Statistics* 41, 1223-1226.

Edelman, D. (1986). Bounds for a nonparametric t table. *Biometrika* 73, 242-243.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* 63, 1278-1302.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58, 13-20.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, second Edition. New York : Wiley.