Boulesteix:

Maximally selected chi-square statistics for at least ordinal scaled variables

Sonderforschungsbereich 386, Paper 407 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Maximally selected chi-square statistics for at least ordinal scaled variables

Anne-Laure Boulesteix

anne-laure.boulesteix@stat.uni-muenchen.de

Department of Statistics, University of Munich,

Akademiestrasse 1, D-80799 Munich, Germany.

February 10, 2005

Abstract

The association between a binary variable Y and a variable X with an at least ordinal measurement scale might be examined by selecting a cutpoint in the range of X and then performing an association test for the obtained 2×2 contingency table using the χ^2 statistic. The distribution of the maximally selected χ^2 statistic (i.e. the maximal χ^2 statistic over all possible cutpoints) under the null-hypothesis of no association between X and Y is different from the known χ^2 distribution. In the last decades, this topic has been extensively studied for continuous X variables, but not for non-continuous variables with an at least ordinal measurement scale (which include e.g. classical ordinal or discretized continuous variables). In this paper, we suggest an exact method to determine the distribution of maximally selected χ^2 statistics in this context. This novel approach can be seen as a method to measure the association between a binary variable and variables with an at least ordinal scale of different types (ordinal, discretized continuous, etc). As an illustration, this method is applied to a new data set describing pregnancy and birth for 811 babies.

Key words: Association test, contingency table, exact distribution, variable selection, selection bias.

1 Introduction

The following situation is not uncommon in medical data analysis. An at least ordinal scaled variable X is suspected by the investigator to be associated with a binary variable Y. Let $(x_i, y_i)_{i=1,\ldots,N}$ denote N independently and identically distributed realizations of the variables X and Y. N_1 and N_2 denote the numbers of observations with $y_i = 1$ and $y_i = 2$, respectively. In this paper, we derive the exact distribution of the maximally selected χ^2 statistic for such X variables. This distribution can be used to measure the association between at least ordinal scaled X variables and the binary variable Y and allows the comparison of several X variables with different numbers of possible values.

If X were nominal scaled, association tests such as the asymptotic χ^2 test or Fisher's exact test for small samples (for a binary X) could be employed to examine the association between X and Y using the sample $(x_i, y_i)_{i=1,\ldots,N}$. If X were continuous, tests based on the normality assumption such as the two-samples t-test or rank tests such as Wilcoxon's rank sum test for two samples may be applied. The case of an at least ordinal scaled but not continuous variable is much more difficult to handle. Without loss of generality, such a variable X can be assumed to take K distincts levels $a_1, \ldots, a_K \in \mathbb{R}$ in the sample $(x_i, y_i)_{i=1,\ldots,n}$, where $2 \leq K \leq N$ and $a_1 < \cdots < a_K$. An option to measure the association between X and Y is to transform X into binary variables $X^{(k)}$ for $k = 1, \ldots, K-1$ as follows

$$\begin{aligned} X^{(k)} &= 0 \quad \text{if } X \leq a_k \\ X^{(k)} &= 1 \quad \text{otherwise.} \end{aligned}$$

Fisher's exact test or the asymptotic χ^2 test may then be applied to each variable $X^{(k)}$ successively. However, one must be careful when interpreting the p-values output by these tests. Selecting the $X^{(k)}$ yielding the smallest p-value and claiming that a_k is a relevant cutpoint of X because the p-value is low would be an inappropriate approach. This issue has been extensively studied in the case of continuous variables. Miller and Siegmund (1982) prove that the maximally selected χ^2 statistic converges to a normalized Brownian bridge under the null-hypothesis of no association between X and Y, whereas Halpern (1982) studies the case of small samples in a simulation study.

Koziol (1991) derives the exact distribution of maximally selected χ^2 statistics given N_1 and N_2 using Durbin's combinatorial approach (Durbin, 1971). Maximally selected χ^2 statistics in $k \times 2$ contingency tables are investigated in Betensky and Rabinowitz (1999). The distributions of other maximally selected statistics such as the statistic used in Fisher's exact test (Halpern, 1999) or McNemar's statistic (Rabinowitz and Betensky, 2000) have also been studied in the last few years.

In this paper, we are interested in the exact distribution of the χ^2 statistic for all types of at least ordinal scaled variables. Via simulations, we show in Section 3 that Koziol's approach is inappropriate to measure association between a binary variable Y and a non-continuous variable X with equal realizations in the sample $(x_i, y_i)_{i=1,\ldots,n}$ ($K < N$). More specifically, under the null-hypothesis of no association between X and Y, Koziol's approach tends to detect more association if K is large. In the present paper, we are concerned with at least ordinal scaled non-continuous variables, which include e.g. classical ordinal variables (for instance a variable with possible values "very good", "good", "bad", "very bad"), discrete metric variables (for instance the number of children in a family), or essentially continuous variables which are measured in a discretized form in practice. For instance, the height of a newborn baby is often given in centimeters and can thus take only a few values ranging from about 47 to 54 cm. For the types of variables described above, we generally have $K < N$ if N is large enough, whereas continuous variables may be assumed to take N distinct values in the sample $(x_i, y_i)_{i=1,\ldots,N}$ ($K = N$). In the framework of maximally selected statistics, a variable with $K < N$ cannot be handled as a variable with $K = N$. The fact that some values of X are taken several times in the sample must be taken into account when deriving the distribution of the maximally selected χ^2 statistic, since the number of possible cutpoints is $K - 1$. In this paper, we propose a novel method to derive the exact distribution of the maximally selected χ^2 statistic for all types of at least ordinal scaled variables. This distribution depends on parameters N_1, N_2 and $m_1, \ldots, m_K$, which are defined as

$$m_k = \sum_{i=1}^{n} I(x_i = a_k), \text{ for } k = 1, \ldots, K,$$

where I is the indicator function. Our novel approach is an adaptation of the proce-

dure proposed by Koziol (1991) and uses Durbin's combinatorial approach. The exact distribution of the maximally selected χ^2 statistic can be used to compute a measure of association between X and a binary variable Y using a sample $(x_i, y_i)_{i=1,\ldots,n}$. This method has potentially many applications, especially in the field of medicine.

The paper is organized as follows. In Section 2, a method to compute the exact distribution function of the maximally selected χ^2 statistic given $N_1, N_2, m_1, \ldots, m_K$ is proposed. In Section 3, we show via simulations that our method is more appropriate than Koziol's method to compare variables with different K and different numbers of missing values. As an illustration, we use our method to measure the association between various binary and at least ordinal scaled variables from a data set describing pregnancy and delivery for 811 babies born between 1990 and 2004 in Section 4.

2 Distribution of the maximally selected χ^2 statistic

2.1 Framework

In this section, X is assumed to be a variable with an at least ordinal measurement scale. Y is a binary variable with levels $Y = 1, 2$. For a given sample $(x_i, y_i)_{i=1,\ldots,N}$, let $a_1 < \cdots < a_K$ denote the different values taken by X. We consider the following 2×2 contingency table, for $k = 1, \ldots, K-1$:

	$X \leq a_k$	$X > a_k$	Σ
$Y = 1$	$n_{1,\leq a_k}$	$n_{1,>a_k}$	N_1
$Y = 2$	$n_{2,\leq a_k}$	$n_{2,>a_k}$	N_2
	$n_{.,\leq a_k} = \sum_{j=1}^{k} m_j$	$n_{.,>a_k} = \sum_{j=k+1}^{K} m_j$	N

where N_1 and N_2 denote the numbers of realizations with $y_i = 1$ and $y_i = 2$, respectively and m_j denotes the number of realizations with $X = a_j$ in the sample $(x_i, y_i)_{i=1,\ldots,N}$. The corresponding χ^2 statistic can be computed as

$$\chi^2_k = \frac{N(n_{1,\leq a_k} n_{2,>a_k} - n_{1,>a_k} n_{2,\leq a_k})^2}{N_1 N_2 n_{.,\leq a_k} n_{.,>a_k}}. \tag{1}$$

In this context, we define the maximally selected χ^2 statistic as

$$\chi^2_{max} = \arg \max_{k=1,\ldots,K-1} \chi^2_k. \tag{2}$$

The aim of this paper is to derive the distribution of χ^2_{max} given $N_1, N_2, m_1, \ldots, m_K$ under the null-hypothesis of no association between X and Y. For simplicity, we will omit $N_1, N_2, m_1, \ldots, m_K$ in the following: F denotes the distribution function of χ^2_{max} given the parameters $N_1, N_2, m_1, \ldots, m_K$:

$$F(d) = p(\chi^2_{max} \leq d).$$

2.2 Method

According to Miller and Siegmund (1982), the χ^2 statistic obtained for the binary variable $X^{(k)}$ can be formulated as $\chi^2_k = A^2_k$, where

$$A_k = \frac{N}{N_1}\left(\frac{n_{2,\leq a_k}}{N_2} - \frac{n_{.,\leq a_k}}{N}\right) \Big/ \sqrt{\frac{n_{.,\leq a_k}}{N}\left(1 - \frac{n_{.,\leq a_k}}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}, \tag{3}$$

for all $k = 1, \ldots, K-1$. Let d be an arbitrary strictly positive real number. After simple computations, one obtains from Equation 3 that $\chi^2_{max} \leq d$ if and only if all the points with coordinates $(n_{.\leq a_k}, n_{2,\leq a_k})$ for $k = 1, \ldots, K-1$ lie on or above the function

$$\text{lower}_d(x) = \frac{N_2 x}{N} - \frac{N_1 N_2 \sqrt{d}}{N}\sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \tag{4}$$

and on or below the function

$$\text{upper}_d(x) = \frac{N_2 x}{N} + \frac{N_1 N_2 \sqrt{d}}{N}\sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}. \tag{5}$$

These curves may be denoted as boundaries. Let $x_{(1)} \leq \cdots \leq x_{(N)}$ denote the ordered realizations of X. Let $N_2(i)$ denote the number of realizations with $Y = 2$ and $X \leq x_{(i)}$. The functions $\text{lower}_d(x)$ and $\text{upper}_d(x)$ can also be represented on the graph $(i, N_2(i))$. A sufficient and necessary condition for $\chi^2_{max} \leq d$ is that the graph $(i, N_2(i))$ does not pass through any point of integer coordinates (i, j) with

$$i = n_{.,\leq a_k}$$

and

$$\text{upper}_d(i) < j \leq i \text{ or } \max(0, i - N_1) \leq j < \text{lower}_d(i),$$

where $k = 1, \ldots, K-1$. Let us denote these points as $B_1, \ldots, B_q$ and their coordinates as $(i_1, j_1), \ldots, (i_q, j_q)$, where $B_1, \ldots, B_q$ are labeled in order of increasing i and

increasing j within each i. Under the null-hypothesis of no association between X and Y, the probability that the path $(i, N_2(i))$ passes through at least one of the points $B_1, \ldots, B_q$ can be computed using Durbin's combinatorial approach (Durbin, 1971), as described by Koziol (1991). Here, we follow Koziol's formulation. The number b_s of paths that pass through point B_s but do not pass through points $B_1, \ldots, B_{s-1}$ is computed recursively as

$$\begin{aligned} b_s &= \begin{pmatrix} i_s \\ j_s \end{pmatrix} - \sum_{r=1}^{s-1} \begin{pmatrix} i_s - i_r \\ j_s - j_r \end{pmatrix}, \quad s = 2, \ldots, q \\ b_1 &= \begin{pmatrix} i_1 \\ j_1 \end{pmatrix}. \end{aligned}$$

The probability that the path $(i, N_2(i))$ passes through at least one of the points $B_1, \ldots, B_q$ is then obtained as

$$p(\chi^2_{max} > d) = \begin{pmatrix} N \\ N_2 \end{pmatrix}^{-1} \sum_{r=1}^{q} \begin{pmatrix} N - i_r \\ N_2 - j_r \end{pmatrix} b_r. \tag{6}$$

It follows

$$F(d) = 1 - \begin{pmatrix} N \\ N_2 \end{pmatrix}^{-1} \sum_{r=1}^{q} \begin{pmatrix} N - i_r \\ N_2 - j_r \end{pmatrix} b_r. \tag{7}$$

Our method, which is strongly related to the procedure described in Koziol (1991), allows explicitly $K < N$. The two approaches differ in the definition of the points $B_1, \ldots, B_q$. In Koziol (1991), the boundaries are formed by the points $B_1, \ldots, B_q$ of coordinates (i, j) satisfying

$$i = \max\{x \in \mathbb{N} : j > \frac{N_2 x}{N} + \frac{N_1 N_2 d}{N} \sqrt{\frac{x}{N}(1 - \frac{x}{N})(\frac{1}{N_1} + \frac{1}{N_2})}\} \text{ and } 1 \leq j \leq N_2,$$

or

$$i = \min\{x \in \mathbb{N} : j < \frac{N_2 x}{N} - \frac{N_1 N_2 d}{N} \sqrt{\frac{x}{N}(1 - \frac{x}{N})(\frac{1}{N_1} + \frac{1}{N_2})}\} \text{ and } 0 \leq j \leq N_2 - 1.$$

As an example, the boundaries obtained with Koziol's method and our new method are represented in Figure 1 for $N_1 = 30$, $N_2 = 40$, $m_1 = 25$, $m_2 = 10$, $m_3 = 25$, $m_4 = 10$ and $d = 3$. It can seen that the points $B_1, \ldots, B_q$ defined by Koziol form a 'closed corridor'. With Koziol's approach, paths which pass through a boundary point

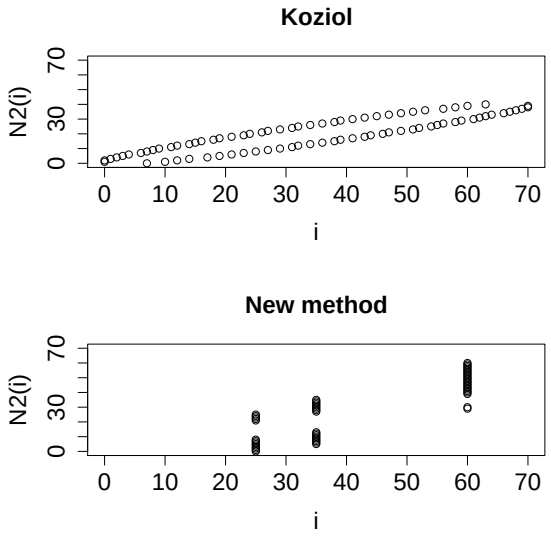

Figure 1: Boundaries obtained with Koziol's approach (top) and our new approach (bottom) for $d = 3$ $N_1 = 30$, $N_2 = 40$, $m_1 = 25$, $m_2 = 10$, $m_3 = 25$ and $m_4 = 10$.

with abscissa i_0 such that $x_{(i_0)} = x_{(i_0+1)}$ are counted for the computation of $F(d)$, although they do not correspond to any concrete possible cutpoint. Thus, this approach is inappropriate for X variables with possibly equal realizations. In our approach, only the paths that would yield $\chi^2_k > d$ for at least one k are counted. These are the paths that are strictly above the upper boundary or strictly below the lower boundary at abscissa $n_{\cdot,\leq a_k}$ ($k = 1, \ldots, K-1$) only. Note that the obtained distribution function F is the same with both methods in the special case $K = N$. In this special case, Koziol's approach is recommended, since computationally faster.

The formula given in Equation 6 can be used to measure the association between a binary variable Y and an at least ordinal scaled variable X using a sample $(x_i, y_i)_{i=1,\ldots,N}$ as follows. For all $k = 1, \ldots, K-1$, the value χ^2_k of the χ^2 statistic for the sample $(x_i, y_i)_{i=1,\ldots,N}$ is computed using Equation 1 and the maximal χ^2 statistic χ^2_{max} is obtained from Equation 2. $F(\chi^2_{max})$ is a measure of association between X and Y. In the following section, we show via simulations that our approach based on the maximally selected χ^2 statistic is more appropriate to measure association between a binary variable Y and a non-continuous predictor variable X than Koziol's approach.

3 Simulations

3.1 Motivation

Koziol's method as well as our method can be used to identify predictor variables which are strongly associated with a binary variable Y. Thus, they can be seen as variable selection methods. This section deals with the variable selection bias of these methods.

In the whole simulation, we make the hypothesis of no association between the binary variable Y and some at least ordinal scaled variables X_1, X_2, X_3. Suppose that we compute a measure of association for each pair (X_1, Y), (X_2, Y) and (X_3, Y) based on a given sample and select the variable X_i with the highest association measure. An effective measure of association is expected to select X_1, X_2 and X_3 with probability $\frac{1}{3}$. In Sections 3.2 and 3.3, we show that Koziol's approach selects variables with large K more often than variables with small K. In contrast, the frequency of selection does not depend on the number of different values in the sample with our method. Two cases are examined:

- Classical ordinal variables for which the set of possible values $\{a_1, \ldots, a_K\}$ does not depend on the specific sample. Such variables are examined in section 3.2.
- Essentially continuous variables which are measured as discrete variables. For such variables, the set $\{a_1, \ldots, a_K\}$ depends on the considered sample. This topic is examined in section 3.3.

For each case, N_{run} data sets are simulated with different values of N ($N = 50$ and $N = 100$) and different a priori probabilities for the classes $Y = 1$ and $Y = 2$:

- *First case (I)*

  The two classes have equal probabilities:

$$\begin{aligned} p_1 = P(Y = 1) &= 0.5 \\ p_2 = P(Y = 2) &= 0.5. \end{aligned}$$

- *Second case (II)*

  The two classes have non-equal probabilities:

$$\begin{aligned} p_1 = P(Y = 1) &= 0.7 \\ p_2 = P(Y = 2) &= 0.3. \end{aligned}$$

For each simulated data set and each variable, χ^2_{max} is determined and $F(\chi^2_{max})$ is computed using successively Koziol's boundaries and the boundaries defined in Section 2. The variable(s) with the highest $F(\chi^2_{max})$ is (are) selected. This is done for several distributions of X_1, X_2, X_3. Sections 3.2 and 3.3 give the description of the variables X_1, X_2, X_3 for each examined case as well as tables containing the obtained frequencies of selection for $N_{run} = 1000$ simulated data sets.

In addition, the problem of predictor variables with different numbers of missing values is addressed in Section 3.4. In Shih (2004), it is shown that with some classical split selection criteria used in the context of classification trees, the selection probability of a given predictor variable depends highly on its number of missing values. In Section 3.4, we show via simulations that our method does not induce such selection bias, whereas Koziol's method does. This makes our method able to measure association between a binary variable and predictor variables with different numbers of missing values, which is a very common situation in practical medical studies.

3.2 Ordinal variables

We simulate data sets containing a binary variable Y and ordinal variables X_1, X_2, X_3 with different numbers of possible values. The set of possible values is $\{1, 2, 3\}$ for X_1, $\{1, 2, 3, 4, 5, 6, 7\}$ for X_2 and $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ for X_3. Let K_i denote the number of possible values of variable X_i. We study two cases successively for the distribution of the variables X_i:

- *First case (A)*

  For each variable X_i, the different levels have equal probability:

$$P(X_i = 1) = \ldots = P(X_i = K_i) = \frac{1}{K_i}$$

	X_1	X_2	X_3
$N = 50, p_1 = 0.5, p_2 = 0.5$	38, *17*	32, *39*	30, *47*
$N = 50, p_1 = 0.7, p_2 = 0.3$	36, *16*	33, *37*	32, *49*
$N = 100, p_1 = 0.5, p_2 = 0.5$	31, *14*	34, *37*	36, *50*
$N = 100, p_1 = 0.7, p_2 = 0.3$	33, *15*	37, *40*	30, *45*
$N = 50, p_1 = 0.5, p_2 = 0.5$	33, *17*	36, *41*	32, *44*
$N = 50, p_1 = 0.7, p_2 = 0.3$	36, *18*	34, *39*	30, *45*
$N = 100, p_1 = 0.5, p_2 = 0.5$	35, *16*	32, *38*	33, *47*
$N = 100, p_1 = 0.7, p_2 = 0.3$	34, *19*	33, *37*	33, *45*

Table 1: **Classical ordinal variables:** Frequency of selection (in %) of X_1, X_2, X_3 for different N, different p_1, p_2, with our method (normal font) and Koziol's method (italic). Top: Case A (equal probabilities), bottom: Case B (non-equal probabilities).

- *Second case (B)*

  For each variable X_i, for $k = 1, \ldots, K$

$$\begin{aligned} P(X_i = k) &= c \cdot 0.1 \quad \text{if } k \text{ is odd,} \\ P(X_i = k) &= c \cdot 0.2 \quad \text{if } k \text{ is even,} \end{aligned}$$

  where c is a normalizing factor such that $\sum_{k=1}^{K_i} p(X_i = k) = 1$.

Thus, four configurations (I/A,I/B,II/A and II/B) are studied. For each configuration, $N_{run} = 1000$ simulation data sets are drawn randomly. The obtained frequencies of selection for each configuration (I/A,I/B,II/A and II/B) and each N can be found in Table 1. As can be seen from Table 1, variable selection using Koziol's criterion is strongly biased. Variables with large K are selected more often. With our new criterion, no bias is observed. Similar results are obtained for the different values of N, p_1 and p_2 and for the different distributions of X_1, X_2, X_3. The bias can be visualized for $N = 100$ and $p_1 = 0.5$ in Figure 2.

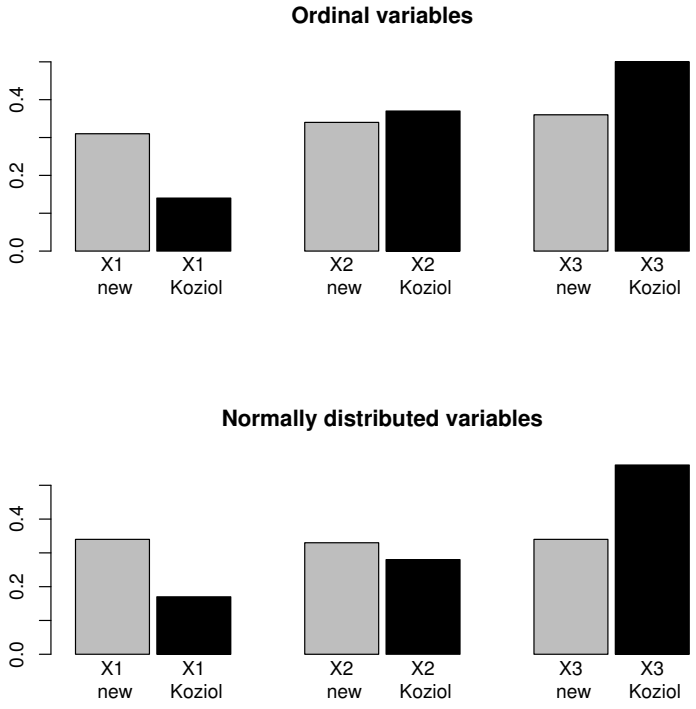

Figure 2: Barplot representing the frequencies of selection of X_1, X_2 and X_3 for $N = 100$ and $p_1 = 0.5$, with our method (gray) and with Koziol's method (black) for classical ordinal variables (Case A: equal probabilities) and discretized normally distributed variables.

3.3 Discretized continuous variables

In this section, continuous variables which are measured as discrete variables are examined. For the binary variable Y, we follow the same scheme as in Section 3.2. Let Z_i, $i = 1, 2, 3$ be identically distributed continuous variables. For $i = 1, 2, 3$, X_i is defined as

$$X_i = round(Z_i/\alpha_i) \cdot \alpha_i,$$

where $\alpha_1 = 1$, $\alpha_2 = 0.5$, $\alpha_3 = 0.1$ and $round(x)$ denotes the integer approximation of x. Thus, X_1, X_2, X_3 correspond to different measurement precisions of identically distributed variables Z_1, Z_2, Z_3. Two cases are examined for the distribution of Z_1, Z_2, Z_3.

- *First case (A)*

  Each variable Z_i is normally distributed:

  $$Z_i \sim \mathcal{N}(0, 1),$$

  for $i = 1, 2, 3$.

- *Second case (B)*

  Each variable Z_i is exponentially distributed:

  $$Z_i \sim exp(1),$$

  for $i = 1, 2, 3$.

As for ordinal variables, N is set successively to $N = 50$ and $N = 100$. For each of the configurations (I/A,IB,II/A and II/B) and each N, $N_{run} = 1000$ data sets are drawn randomly and the frequencies of selection are computed, either with Koziol's approach or with our new method. The results can be found in Table 2. Whereas the frequencies of selection of X_1, X_2, X_3 are approximately equal with our new approach, Koziol's criterion selects more often variables with large K. This difference between the two approaches is observed for the different values of N, p_1 and p_2 and for both normally and exponentially distributed Z_i. The frequencies of selection for normally distributed variables can be visualized for $N = 100$ and $p_1 = 0.5$ in Figure 2.

	X_1	X_2	X_3
$N = 50, p_1 = 0.5, p_2 = 0.5$	39, *22*	30, *27*	31, *55*
$N = 50, p_1 = 0.7, p_2 = 0.3$	36, *19*	33, *31*	31, *52*
$N = 100, p_1 = 0.5, p_2 = 0.5$	34, *17*	33, *28*	34, *56*
$N = 100, p_1 = 0.7, p_2 = 0.3$	32, *15*	34, *30*	34, *56*
$N = 50, p_1 = 0.5, p_2 = 0.5$	36, *19*	32, *28*	32, *54*
$N = 50, p_1 = 0.7, p_2 = 0.3$	35, *18*	33, *31*	31, *53*
$N = 100, p_1 = 0.5, p_2 = 0.5$	38, *21*	30, *29*	32, *53*
$N = 100, p_1 = 0.7, p_2 = 0.3$	32, *18*	34, *30*	34, *55*

Table 2: Discretized continuous variables: Frequency of selection (in %) of X_1, X_2, X_3 for different N, different p_1, p_2, with our method (normal font) and Koziol's method (italic). Top: Case A (normal distribution), bottom: Case B (exponential distribution).

3.4 Selection bias due to missing values

In this subsection, X_1, X_2, X_3 are identically distributed and missing values are introduced at random. The number of missing values differs for the three variables. Four cases are examined for the distribution of X_1, X_2, X_3:

- Ordinal A: The K_i different values have equal probability.
- Ordinal B: The K_i different values do not have equal probability (see Section 3.2 for the description of the distribution).
- Normal: X_1, X_2, X_3 are obtained by rounding normally distributed variables Z_1, Z_2, Z_3.
- Exponential: X_1, X_2, X_3 are obtained by rounding exponentially distributed variables Z_1, Z_2, Z_3.

Using the notations of Sections 3.2 and 3.3, we fix $K_i, i = 1, 2, 3$ at 7 for ordinal variables and $\alpha_i, i = 1, 2, 3$ at 0.5 for continuous variables. Similar results could be obtained with other values of K_i and α_i. N is fixed at 50 and the two levels of the binary variable Y have equal probability 0.5. For all four cases, the number of missing

	X_1 (0 MV)	X_2 (10)	X_3 (20 MV)
Ordinal A	30, *27*	35, *36*	35, *38*
Ordinal B	34, *29*	33, *32*	33, *38*
Normal	32, *30*	32, *34*	36, *37*
Exponential	33, *34*	32, *33*	34, *35*

Table 3: Frequency of selection (in %) of X_1, X_2, X_3 for different distributions and different numbers of missing values. The number of missing values is specified in parentheses. $N = 50$. $p_1 = p_2 = 0.5$. $N_{run} = 1000$.

values is set to 0 for X_1, 10 for X_2 and 20 for X_3. The results obtained with Koziol's method and our new approach for $N_{run} = 1000$ simulated data sets are presented in Table 3.4. It can be seen that that the three variables X_1, X_2 and X_3 are selected with the same frequency by our method, whereas Koziol's method selects variables with many missing values a little more often.

4 Application to pregnancy and birth data

To illustrate our approach, we consider a pregnancy and birth data set which we collected by ourselves directly from internet users recruited on french-speaking pregnancy and birth websites. Table 4 describes the investigated binary variables (top) and the candidate predictor variables (bottom). Each binary variable takes value 1 if the answer is no, 2 if the answer is yes. The candidate predictor variables are discrete metric variables ($PREVIOUS$) and discretized continuous variables (AGE, $HEIGHTMO$, $WEIGHTMO$, $PREVIOUS$, $HEIGHTBB$, $WEIGHTBB$, $HEAD$, $DURATION$, $DIFF$). For each of them, Table 4 gives the number K of different values taken in the sample.

All pairs formed by a binary response variable and a candidate predictor variable from Table 4 are examined successively. The following procedure is applied to each pair.

Variable	K	Description
MEMBRANE	–	Did the membranes rupture before the beginning of labor ?
CESA	–	Did the mother have a cesarean section ?
EPISIO	–	Did the obstetrician perform an episiotomy ?
INDUCED	–	Was the delivery induced medically ?
SEX	–	Sex of the baby (1 - male, 2 - female)
AGE	25	Age of the mother (in year).
HEIGHTMO	32	Height of the mother (in cm).
WEIGHTMO	90	Weight of the mother before pregnancy (in kg).
PREVIOUS	7	Number of previous deliveries.
HEIGHTBB	39	Height of the baby at birth (in cm).
WEIGHTBB	39	Weight of the baby at birth (in g).
HEAD	30	Head circumference of the baby (in cm). About 30% missing values.
DURATION	16	Duration of the pregnancy (in weeks).
DIFF	61	Weight put on by the mother during pregnancy.

Table 4: Binary response variables (top) and candidate predictor variables (bottom)

	MEMBRANE	CESA	EPISIO	INDUCED	SEX
AGE	0.2306	0.9766	0.7870	0.7322	0.7407
HEIGHTMO	0.8850	0.9977	0.5681	0.0932	0.2475
WEIGHTMO	0.9424	0.0536	0.6335	0.9928	0.3844
PREVIOUS	0.9993	0.9883	1	0.5735	0.4437
HEIGHTBB	0.5743	1	0.6455	0.9190	0.9999
WEIGHTBB	0.3869	1	0.2269	0.9771	0.9986
HEAD	0.3671	0.9954	0.8490	0.9986	0.9999
DURATION	0.8881	1	0.5555	0.9995	0.5700
DIFF	0.8922	0.4543	0.7417	0.700	0.2540

Table 5: Measure of association $F(\chi^2_{max})$ between binary variables and predictor variables

1. Denote as $a_1, \ldots, a_K$ the different values taken by the candidate predictor variable X in the sample, with $a_1 < \cdots < a_K$. If X is a classical ordinal variable with values $1, \ldots, K$, we have $a_1 = 1, \ldots, a_K = K$. In the extreme case of X taking different values for all N observations, we have $a_1 = x_{(1)}, \ldots, a_N = x_{(N)}$.

2. For $k = 1, \ldots, K-1$, compute the χ^2 statistic χ^2_k from the 2×2 contingency table

	$X \leq a$	$X > a_k$
$Y = 1$	$n_{1,\leq a_k}$	$n_{1,>a_k}$
$Y = 2$	$n_{2,\leq a_k}$	$n_{2,>a_k}$

3. Determine $\chi^2_{max} = \max_{k=1,\ldots,K-1} \chi^2_k$.

4. Compute $F(\chi^2_{max})$ with the parameters $N_1, N_2, m_1, \ldots, m_K$.

The results can be found in Table 5. Most of the results from Table 5 agree with previous obstetrical knowledge. For instance, the high association between the variables EPISIO and PREVIOUS may be explained by the current french obstetrical policy: episiotomies are still routinely performed for nulliparous women. It is also

well-known that male babies are heavier in average than female babies (Liebermann et al., 1997). The high association between the binary variable INDUCED and the variable WEIGHTBB, HEAD and DURATION can be explained by the fact that post-term pregnancies are one of the most common indications for labor induction. Cesarean sections are known to be more common for big babies (James, 2001), which agrees with the high association found for the binary variable CESA and the variables HEIGHTBB, WEIGHTBB, HEAD and DURATION. A study by Cnattingius et al. (1998) also pointed out that the risks for cesarean increase with maternal age or decrease with maternal height, which is consistent with our results.

5 Discussion

In this paper, we proposed a simple procedure based on Durbin's combinatorial approach (Durbin, 1971) to compute the exact distribution of maximally selected χ^2 statistics in the context of at least ordinal scaled variables. This procedure can be used to identify prognostic factors in clinical studies. In contrast to Koziol's method, the proposed method does not induce selection bias when the candidate predictor variables have different numbers of distinct values or different numbers of missing values in the available sample. For essentially continuous predictor variables, a possible drawback of our method is that it computes the distribution of the maximally selected χ^2 statistic given the observed $m_1, \ldots, m_K$, not given the distribution of X. However, this feature can also be seen as an advantage: the procedure requires no assumptions on the distribution of the predictor variable X. As an exact procedure, it is also appropriate for small sample sizes which are common in clinical studies. It can be seen as a global framework to measure association between a binary variable and all types of at least ordinal scaled variables, for small and large sample sizes. The results obtained on the pregnancy and birth data set agree with previous obstetrical knowledge.

In future work, our method could be interestingly applied to classification trees (Breiman et al., 1984). In recent papers, maximally selected statistics and their associated p-values have been successfully applied to the problem of variable and cutpoint selection in classification and regression trees (Shih, 2004; Lausen et al., 2004). The

association measure developed in this paper might also be used as a selection criterion for choosing the best splitting variable and the best cutpoint. Since it allows the comparison of predictor variables of different types and avoids the selection bias due to missing values, we expect it to perform better than the usual criteria in some cases.

Acknowledgments

I thank Gerhard Tutz and Korbinian Strimmer for critical comments and Stefanie Pildner von Steinburg for helping me to interpret the data.

References

Betensky, R. A., Rabinowitz, D., 1999. Maximally selected χ^2 statistics for $k \times 2$ tables. Biometrics 55, 317–320.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, J. C., 1984. Classification and Regression Trees. Wadsworth, Monterey, CA.

Cnattingius, R., Cnattingius, S., Notzon, F. C., 1998. Obstacles to reducing cesarean rates in a low-cesarean setting: the effect of maternal age, height, and weight. Obstetrics and Gynecology 92, 501–506.

Durbin, J., 1971. Boundary-crossing probabilities for the brownian motion and poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. Journal of Applied Probabilities 8, 431–453.

Halpern, A. L., 1999. Minimally selected p and other tests for a single abrupt changepoint in a binary sequence. Biometrics 55, 1044–1050.

Halpern, J., 1982. Maximally selected chi square statistics. Biometrics 38, 1011–1016.

James, W. H., 2001. Gestational diabetes, birth weigth, sex ratio, and ceasarean section. Diabetes Care 24, 2018–2019.

Koziol, J. A., 1991. On maximally selected chi-square statistics. Biometrics 47, 1557–1561.

Lausen, B., Hothorn, T., Bretz, F., Schumacher, M., 2004. Assessment of optimal selected prognostic factors. Biometrical Journal 46, 364–374.

Liebermann, E., Lang, J. M., Cohen, A. P., Frigoletto, F. D., Acker, D., Rao, R., 1997. The association of fetal sex with the rate of cesarean section. American Journal of Obstetrics and Gynecology 176, 667–671.

Miller, R., Siegmund, D., 1982. Maximally selected chi square statistics. Biometrics 38, 1011–1016.

Rabinowitz, D., Betensky, R. A., 2000. Approximating the distribution of maximally selected mcnemar's statistics. Biometrics 56, 987–902.

Shih, Y. S., 2004. A note on split selection bias in classification trees. Computational Statistics and Data Analysis 45, 457–466.