Boulesteix:

Maximally selected chi-square statistics and binary splits of nominal variables

Sonderforschungsbereich 386, Paper 449 (2005)

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

Projektpartner

Maximally selected chi-square statistics and binary splits of nominal variables

Anne-Laure Boulesteix

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

Department of Statistics, University of Munich,

Akademiestrasse 1, D-80799 Munich, Germany.

October 18, 2005

Abstract

We address the problem of maximally selected chi-square statistics in the case of a binary Y variable and a nominal X variable with several categories. The distribution of the maximally selected chi-square statistic has already been derived when the best cutpoint is chosen from a continuous or an ordinal X, but not when the best split is chosen from a nominal X. In this paper, we derive the exact distribution of the maximally selected chi-square statistic in this case using a combinatorial approach. Applications of the derived distribution to variable selection and hypothesis testing are discussed based on simulations. As an illustration, our method is applied to a pregnancy and birth data set.

keywords: Categorical variables, association test, contingency table, exact distribution, variable selection, selection bias.

1 Introduction

A classical problem in medical research is the study of the association between a binary clinical outcome Y and a nominal (multicategorical) prognostic factor X having the set of unordered categories $S = \{1, \ldots, K\}$. For instance, health scientists often want their statistical consultants to find statistically relevant binary splitting rules to predict the binary outcome Y using the nominal variable X as predictor. A common approach consists of considering successively all the subsets S_j, $j = 1, \ldots, 2^K - 2$ of S and transforming the nominal variable X into a binary variable as follows:

$$X^{(S_j)} = \begin{cases} 1 & \text{if } X \in S_j \\ 0 & \text{otherwise,} \end{cases}$$

thus producing a 2×2 contingency table. One can then evaluate the association between $X^{(S_j)}$ and Y using any association measure. In medical applications, the partition yielding the highest association measure is

often used to construct a binary prediction rule. Classification trees with binary splits, which are especially appreciated in medical research, are an important application of such binary splittings.

The chi-square statistic is commonly used to test the association between two binary variables using a sample of N independent observations (see Kang and Kim (2004) for an extensive study of association tests in 2×2 contingency tables). The p-value of the chi-square test can be used as association measure for the problem described above. However, the minimal p-value over the $2^K - 2$ partitions of S must be considered with caution. Selecting the partition of S that minimizes the p-value of the chi-square test and claiming that X is a good predictor for Y because this p-value is low would be an inappropriate approach. Indeed, the distribution of the maximally selected chi-square statistic is different from the nominal chi-square distribution. Numerous papers published in the last decades address the problem of maximally selected statistics under the hypothesis of no association between two variables X and Y in different situations. Minimally selected (weighted) misclassification rates are examined by Gail and Green (1976) in the case of a continuous X and a binary Y. Miller and Siegmund (1982) show that the maximally selected chi-square statistic converges to a normalized Brownian bridge under the null-hypothesis of no association between a continuous X and a binary Y. The case of small samples is examined by Halpern (1982) in a simulation study, whereas Koziol (1991) derives the exact distribution of maximally selected chi-square statistics using a combinatorial approach. Boulesteix (2005) proposes a generalization of Koziol's approach to derive the exact distribution of the maximally selected chi-square statistics in the case of an at least ordinally scaled X and a binary Y. Maximally selected chi-square statistics for a continuous X and a nominal Y are investigated in Betensky and Rabinowitz (1999). The distributions of other related optimally 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. Hothorn and Lausen (2003) derive a lower bound for the exact distribution of maximally selected rank statistics. The problem of the asymptotic null distribution of maximally selected statistics for binary, ordered, quantitative or censored response variables is discussed in Lausen and Schumacher (1992, 1996) and Lausen, Lerche and Schumacher (2002).

In this paper, we investigate the case of a binary Y and a nominal (multicategorical) X. The chi-square statistic obtained for the binary variables Y and $X^{(S_j)}$ is denoted as $\chi^2_{S_j}$. In this context, the maximally selected chi-square statistic is defined as

$$\chi^2_{max} = \max_{S_j \in \mathcal{S}} \chi^2_{S_j},$$

where $\mathcal{S}$ denotes the set of the non-empty strict subsets of S. The present paper proposes a novel combinatorial method to compute the exact distribution of the maximally selected chi-square statistic χ^2_{max} under the null-hypothesis of no association between X and Y. The term $(x_i, y_i)_{I=1,\ldots,N}$ denotes N independent 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 and m_k, $k = 1, \ldots, K$ the number of observations with $x_i = k$. The distribution of χ^2_{max} is derived given $N_1, N_2, m_1, \ldots, m_K$. The focus of the paper is on the chi-square statistic. However, our approach might be easily generalized to other association measures satisfying a specific convexity property, such as the cross-entropy criterion (also called deviance) used in

machine learning.

The paper is organized as follows. Our novel approach to derive the distribution of χ^2_{max} in the context of a binary Y and a nominal (multicategorical) X is presented in Section 2. Section 3 discusses possible applications of the new approach to variable selection and hypothesis testing based on two simulation studies and compares it to other related approaches. The new method is illustrated through an application to a birth data set.

2 Method

2.1 Framework and notations

For a given partition $\{S_j, \overline{S}_j\}$ of S, let us consider the contingency table

$$\begin{array}{c|cc|c} & X^{(S_j)}=0 & X^{(S_j)}=1 & \\ \hline Y=1 & n_{1,\overline{S}_j} & n_{1,S_j} & N_1 \\ Y=2 & n_{2,\overline{S}_j} & n_{2,S_j} & N_2 \\ \hline & N_{\overline{S}_j} & N_{S_j} & N \end{array} .$$

The chi-square statistic $\chi^2_{S_j}$ may be computed as

$$\chi^2_{S_j} = \frac{N(n_{1,\overline{S}_j} n_{2,S_j} - n_{1,S_j} n_{2,\overline{S}_j})^2}{N_1 N_2 N_{S_j} N_{\overline{S}_j}}$$

and reformulated as $\chi^2_{S_j} = A^2_{S_j}$, where:

$$A_{S_j} = \frac{N}{N_1}\left(\frac{n_{2,S_j}}{N_2} - \frac{N_{S_j}}{N}\right) \Big/ \sqrt{\frac{N_{S_j}}{N}\left(1-\frac{N_{S_j}}{N}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)},$$

The rest of this section presents a novel method to compute $P(\chi^2_{max} \leq d)$ under the null-hypothesis of no association between X and Y and given $N_1, N_2, m_1, \ldots, m_K$. For simplification, we use the notation $F(d) = P_{H_0}(\chi^2_{max} \leq d)$ in the rest of the paper. In our approach, $P_{H_0}(\chi^2_{max} \leq d)$ is developed into an ordinal framework as follows. Let $\hat{p}_k$, $k = 1, \ldots, K$ be defined as the empirical proportion of observations of class $Y = 2$ within category $X = k$:

$$\hat{p}_k = \frac{n_{2,\{k\}}}{m_k}.$$

Let $\mathcal{Z}$ denote the set of the permutations of $\{1, \ldots, K\}$:

$$\mathcal{Z} = \{(\sigma_1, \ldots, \sigma_K) | \forall k = 1, \ldots, K,\ \sigma_k \in \{1, \ldots, K\} \text{ and } \forall k_1 \neq k_2, \sigma_{k_1} \neq \sigma_{k_2}\}$$

and $\mathbf{Z} = (Z_1, \ldots, Z_K)$ the element of $\mathcal{Z}$ satisfying

$$\hat{p}_{Z_1} \geq \cdots \geq \hat{p}_{Z_K},$$

for the sample $(x_i, y_i)_{i=1,\ldots,N}$. Since some of the $\hat{p}_k$ might be equal, one needs a further convention for the random vector $\mathbf{Z}$ to be defined uniquely: we make the convention that if $\hat{p}_{Z_i} = \hat{p}_{Z_{i+1}}$, then $Z_i < Z_{i+1}$.

Then $F(d) = P_{H_0}(\chi^2_{max} \leq d)$ can be decomposed as

$$\begin{aligned} P_{H_0}(\chi^2_{max} \leq d) &= 1 - P_{H_0}(\chi^2_{max} > d) \quad (1) \\ &= 1 - \sum_{(\sigma_1,\ldots,\sigma_K)\in\mathcal{Z}} P_{H_0}(\chi^2_{max} > d \cap \mathbf{Z} = (\sigma_1,\ldots,\sigma_K)). \quad (2) \end{aligned}$$

As shown by Shih (2001), if $\mathbf{Z} = (\sigma_1,\ldots,\sigma_K)$, then χ^2_{max} is obtained as

$$\chi^2_{max} = \max_{k=1,\ldots,K-1} \chi^2_{\{\sigma_1,\ldots,\sigma_k\}}.$$

Since $\hat{p}_{\sigma_1} \geq \cdots \geq \hat{p}_{\sigma_K}$, we have $A_{\{\sigma_1,\ldots,\sigma_k\}} > 0$, for all $k = 1,\ldots,K-1$. Thus,

$$\chi^2_{max} > d \Leftrightarrow (\max_{k=1,\ldots,K-1} A_{\{\sigma_1,\ldots,\sigma_k\}}) > \sqrt{d}.$$

In the rest of this section, we address the problem of computing

$$P_{H_0}\left(\max_{k=1,\ldots,K-1} A_{\{\sigma_1,\ldots,\sigma_k\}} > \sqrt{d} \cap \mathbf{Z} = (\sigma_1,\ldots,\sigma_K)\right) \quad (3)$$

for any $\{\sigma_1,\ldots,\sigma_K\} \in \mathcal{Z}$, given $N_1, N_2, m_1,\ldots,m_K$. Our exact and computationally efficient novel method is based on binomial coefficients. For a fixed $(\sigma_1,\ldots,\sigma_K) \in \mathcal{Z}$, let B_k, $k = 1,\ldots,K-1$ denote the number of ways to choose simultaneously N_2 observations from a sample of N observations such that

$$\mathbf{Z} = (\sigma_1,\ldots,\sigma_K), \quad (4)$$

$$\forall j < k,\ A_{\{\sigma_1,\ldots,\sigma_j\}} \leq d, \quad (5)$$

$$\text{and } A_{\{\sigma_1,\ldots,\sigma_k\}} > d. \quad (6)$$

The probability (3) may then be obtained as

$$P_{H_0}\left(\max_{k=1,\ldots,K-1} A_{\{\sigma_1,\ldots,\sigma_k\}} > \sqrt{d} \cap \mathbf{Z} = (\sigma_1,\ldots,\sigma_K)\right) = \sum_{k=1}^{K-1} B_k / \binom{N}{N_2}. \quad (7)$$

The combinatorial derivation of the B_k is given in the appendix. The computation of $P_{H_0}(\chi^2_{max} \leq d)$ can be summarized as follows:

1. For each $\{\sigma_1,\ldots,\sigma_K\} \in \mathcal{Z}$,
   - the B_k, $k = 1,\ldots,K-1$, are computed as described in the appendix,
   - $P_{H_0}\left(\max_{k=1,\ldots,K-1} A_{\{\sigma_1,\ldots,\sigma_k\}} > \sqrt{d} \cap \mathbf{Z} = (\sigma_1,\ldots,\sigma_K)\right)$ is obtained from formula (7).
2. $P_{H_0}(\chi^2_{max} \leq d)$ is computed as

$$P_{H_0}(\chi^2_{max} \leq d) = 1 - \sum_{\{\sigma_1,\ldots,\sigma_K\}\in\mathcal{Z}} P_{H_0}\left(\max_{k=1,\ldots,K-1} A_{\{\sigma_1,\ldots,\sigma_k\}} > \sqrt{d} \cap \mathbf{Z} = (\sigma_1,\ldots,\sigma_K)\right).$$

3 Applications and simulations

3.1 Introduction

In this section, we discuss several applications of the derived exact distribution and their advantages and inconveniences over two other approaches based on the chi-square statistic. Here, we give a short summary of the considered approaches.

3.1.1 The 'naive' approach based on χ^2_{max}

The naive approach consisting of (i) selecting the partition of X that yields the maximal chi-square statistic, (ii) performing the chi-square association test for the resulting 2×2 contingency table and (iii) rejecting the null-hypothesis of no association based on the obtained p-value is formally incorrect. In this approach, the p-value is computed from the nominal chi-square distribution with one degree of freedom, which is different from the distribution of the considered 'test statistic' χ^2_{max}. This contradiction makes it inappropriate, as discussed in Section 3.3.

3.1.2 The usual chi-square test for $k \times 2$ contingency tables

The chi-square association test might be performed for $k \times 2$ contingency tables, where $k > 2$. In Section 3.2, this test is compared to our new approach.

3.1.3 Our novel approach based on $F(\chi^2_{max})$

The distribution function F of the maximally selected chi-square statistic provides a statistical association criterion. To measure the association between a binary Y and a nominal X, one has to (i) compute χ^2_{max} for the available sample $(x_i, y_i)_{i=1,\ldots,N}$, (ii) compute $F(\chi^2_{max})$ given $N_1, N_2, m_1, \ldots, m_K$. The quantity $1-F(\chi^2_{max})$ may be seen as the p-value of a test testing the null-hypothesis of no association between X and Y. Under the null-hypothesis, the distribution function of the test statistic χ^2_{max} is F. $N_1, N_2, m_1, \ldots, m_K$ are parameters of the distribution F.

3.2 Our novel approach vs. the chi-square statistics for $k \times 2$ contingency tables

Our method to derive the exact distribution of the maximally selected chi-square statistic may be extended to any association criterion for 2×2 contingency tables satisfying specific convexity properties given in Shih (2001). Thus, our combinatorial method solves a general problem, with applications e.g. in machine learning. However, in the special case of the chi-square statistic examined in this paper, $k \times 2$ tables are explicitly allowed. Hence, a natural question from the point of view of statisticians is 'what is the difference between the novel approach and the chi-square test for $k \times 2$ contingency tables ?'. This topic is addressed in this section.

The first obvious difference between our approach and the chi-square test for $k \times 2$ tables is that, as an exact procedure, our approach is also valid for very small sample sizes, which are common in clinical studies or in the branches of classification trees. In return, it becomes computationally expensive for very large sample sizes or large numbers of X categories.

The power of both approaches can be examined via simulations. Intuitively, one expects our novel approach to perform comparatively better if the conditional probabilities $p_k = P(Y = 2|X = k)$, $k = 1, \ldots, K$, form two well-separated clusters. It is expected to perform comparatively worse if the p_k are approximately equidistantly spread over an interval included in $[0, 1]$. In the rest of this section, the power of both tests in these two extreme situations is examined via simulations.

$N_{run} = 1000$ data sets containing $N = 100$ independent observations of a binary variable Y and nominal variable X are simulated as follows. The number of categories of X is set to $K = 3$ and $K = 4$. Two different distributions of Y given X corresponding to the two situations described above are examined successively.

- **Case A:** The $p_k = P(Y = 2|X = k)$, $k = 1, \ldots, K$, form two distinct clusters.

  $$p_1 = 0.3, \quad p_2 = p_3 = 0.6,$$

  for $K = 3$ and

  $$p_1 = p_2 = 0.3, \quad p_3 = p_4 = 0.6,$$

  for $K = 4$.

- **Case B:** The $p_k = P(Y = 2|X = k)$, $k = 1, \ldots, K$, are equidistantly spread over an interval included in $[0, 1]$:

  $$p_1 = 0.3, \quad p_2 = 0.5, \quad p_3 = 0.7,$$

  for $K = 3$ and

  $$p_1 = 0.3, \quad p_2 = 0.4, \quad p_3 = 0.5, \quad p_4 = 0.6,$$

  for $K = 4$.

The empirical distributions of the logarithmized p-values obtained with both tests are displayed in Figure 1 for the four examined situations ($K = 3$ or $K = 4$, Case A or B). Whereas the two distributions merge for Case B, the empirical distribution function of the p-value obtained for our novel approach lies above the one obtained for the usual chi-square statistic in Case A. In Case A, the p-values are significantly lower with our approach than with the usual chi-square test: for both $K = 3$ and $K = 4$, Wilcoxon's (paired and two-sided) rank sum test rejects the hypothesis of equality of the medians at significance level $< 10^{-12}$. In a word, our novel test and the usual chi-square test for $k \times 2$ contingency tables have approximately equal power if the $p_k = P(Y = 2|X = k)$ are equidistantly spread over an interval included in $[0, 1]$, but our test performs better if the p_k form two clusters.

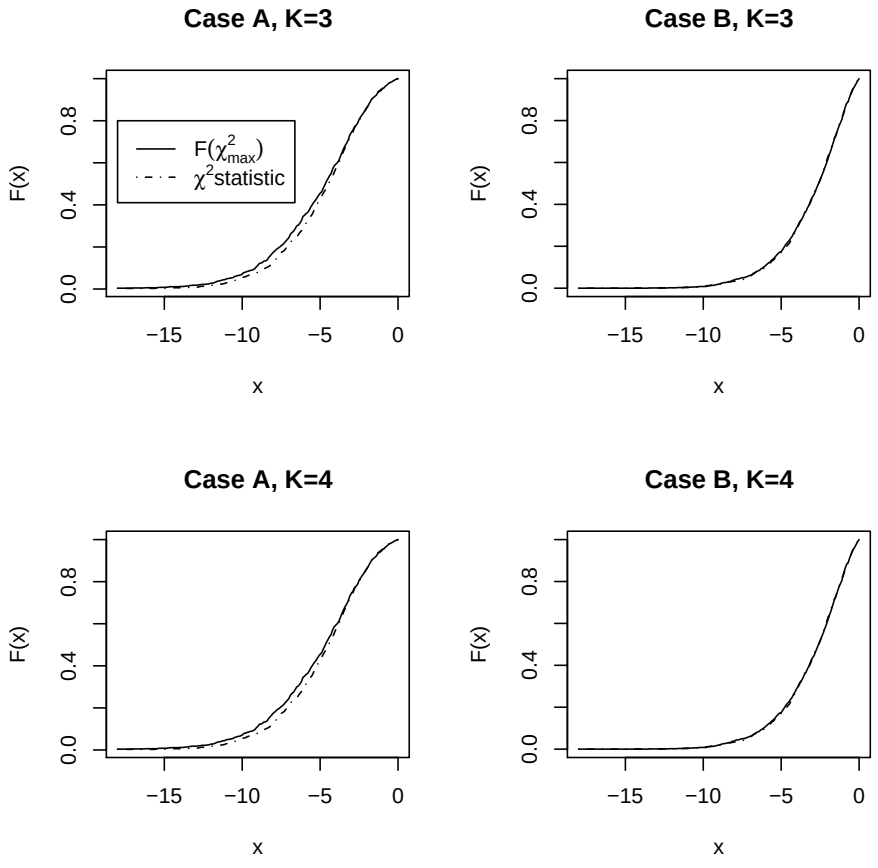

Figure 1: Empirical distribution of the logarithmized p-values obtained using our approach (plain) and the usual chi-square statistic for $k \times 2$ tables (dashed) with $K = 3$ (top) and $K = 4$ (bottom) in Case A (left) and Case B (right).

3.3 Our novel approach vs. the 'naive approach'

3.3.1 Selection bias

A classical approach in recursive partitioning algorithms with binary splittings such as CART (Breiman et al., 1984) is to search through all possible partitions generated by the candidate predictor variables. An association criterion is computed for each of these partitions and the one maximizing the criterion is selected for splitting. However, this approach is biased towards predictor variables with many possible partitions, see e.g. Shih (2004) for a discussion of this problem in the case of continuous predictors with different numbers of missing values. Shih (2004) suggests an alternative p-value criterion based on the distribution of the maximally selected chi-square statistic. In the case of nominal predictors, the selection is biased towards variables with many categories (Kim and Loh, 2001), since the number of possible partitions for a given predictor variable increases exponentially with its number of categories. The distribution of the maximally selected chi-square statistic under the null-hypothesis of no association between X and Y derived in Section 2 provides graphical evidence for this selection bias. As an example, Figure 2 (top) shows the distribution function of the maximally selected chi-square statistic for $N_1 = N_2 = 30$ and $m_1 = m_2 = m_3 = 20$ ($K = 3$, dashed), $m_1 = m_2 = m_3 = m_4 = 15$ ($K = 4$, dotted) or $m_1 = m_2 = m_3 = m_4 = m_5 = 12$ ($K = 5$, plain). It can be seen from Figure 2 that $F(x)$ increases with K for a given x. Thus, variable selection based on the maximally selected chi-square statistic is expected to be biased towards variables with large K. In the next section, a simulation study is performed to quantify this selection bias and to check that variable selection based on $F(\chi^2_{max})$ instead of χ^2_{max} eliminates the selection bias.

3.3.2 Simulations

$N_{run} = 1000$ data sets containing a binary 'response' variable Y and nominal 'predictor variables' X_1, X_2, X_3 are simulated as follows. Each data set contains either $N = 50$ or $N = 100$ independent identically distributed observations of the binary variables Y and of the nominal predictors X_1, X_2, X_3. The number of categories is set to $K = 3$ for X_1, $K = 4$ for X_2 and $K = 5$ for X_3. All four variables are assumed mutually independent and uniformly distributed. Thus, a reliable variable selection criterion is expected to select X_1, X_2 and X_3 with equal probability $\frac{1}{3}$. For each of the N_{run} data sets, variable selection is performed using successively χ^2_{max} and $F(\chi^2_{max})$ as selection criterion. The obtained frequencies of selection are collected in Table 1. Strong selection bias are observed when χ^2_{max} is used as a selection criterion. In contrast, the $F(\chi^2_{max})$ criterion selects all three variables with approximately equal frequency. Another advantage of the $F(\chi^2_{max})$ based method is that, as an exact procedure, it provides valid p-values even for very small samples.

3.3.3 Comparison of the p-values

The performance of the two testing approaches may be visualized via a 'p-value/p-value' plot representing the p-values obtained with our new approach against the p-values obtained with the naive approach which

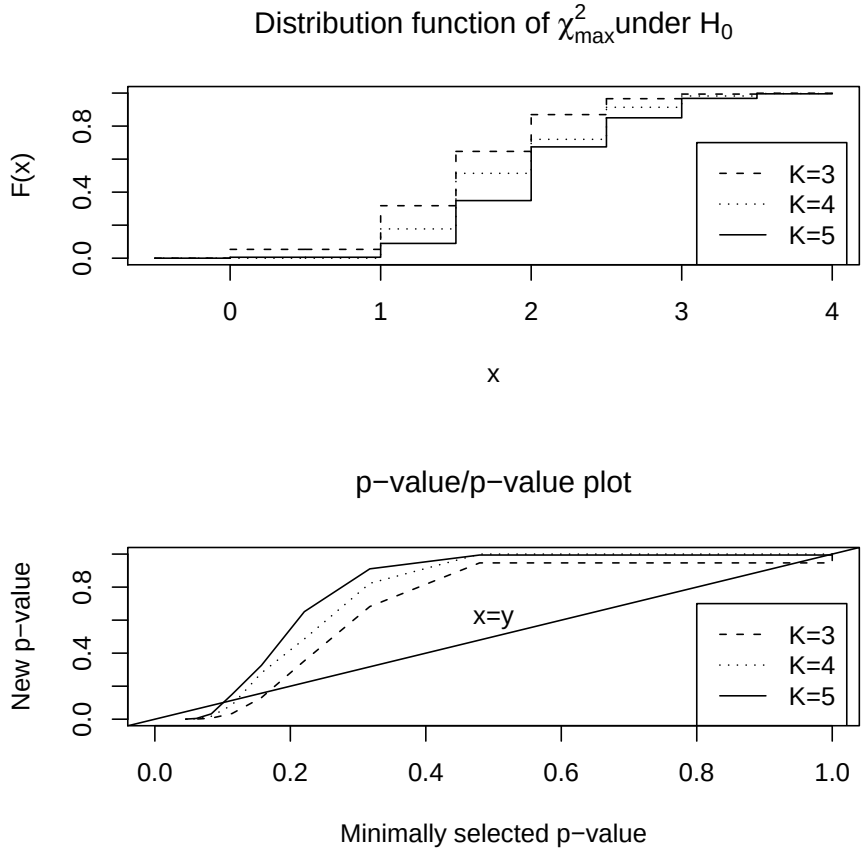

Figure 2: Example: $N_1 = N_2 = 30$ and $m_1 = m_2 = m_3 = 20$ ($K = 3$, dashed), $m_1 = m_2 = m_3 = m_4 = 15$ ($K = 4$, dotted) or $m_1 = m_2 = m_3 = m_4 = m_5 = 12$ ($K = 5$, plain). **Top:** Distribution function of χ^2_{max}. **Bottom:** p-value/p-value plot.

Criterion	N	X_1	X_2	X_3
χ^2_{max}	50	17	35	50
$F(\chi^2_{max})$	50	32	34	34
χ^2_{max}	100	17	32	51
$F(\chi^2_{max})$	100	35	32	34

Table 1: Frequencies of selection (in %) of X_1,X_2 and X_3 for $N = 50$ (top) and $N = 100$ (bottom) with χ^2_{max} and $F(\chi^2_{max})$.

may be denoted as minimally selected p-values. As an example, Figure 2 (bottom) displays the p-value/p-value plot for $N_1 = N_2 = 30$ and $m_1 = m_2 = m_3 = 20$ ($K = 3$, dashed), $m_1 = m_2 = m_3 = m_4 = 15$ ($K = 4$, dotted) or $m_1 = m_2 = m_3 = m_4 = m_5 = 12$ ($K = 5$, plain). Unsurprisingly, for a given χ^2_{max}, the p-value obtained with our method increases with K. All three graphs are above the line of equation $y = x$ when the minimally selected p-values are large enough, which can be interpreted as a multiple comparisons effect. However, when the minimally selected p-value is low (e.g. < 0.1 for $K = 3$), the p-value obtained using our new approach is lower than the p-value yielded by the naive approach. In a word, our new testing approach makes large p-values even larger and low p-values even lower, which is of course a desirable property in terms of power. This feature is illustrated in the next section through a real data example.

3.3.4 A real data example

This section illustrates the novel method through an application to a birth data set introduced by Boulesteix (2005). In developed countries, the proportion of cesarean births has considerably increased in the last decades. The factors influencing the probability to undergo a cesarean section have been the subject of numerous publications, e.g., Cnattingius, Cnattingius and Notzon (1998) and Liebermann et al. (1997). Beside obvious factors such as previous cesarean section(s), fetal-pelvic disproportion or unusual fetal presentation (for instance breech), the way the delivery begins might also influence the probability to need a cesarean section. There are basically three possible scenarios: (i) the woman feels natural contractions, (ii) the membranes rupture or (iii) labor is medically induced, for instance because of fetal distress or overdue pregnancy. If the membranes rupture before the first contractions, most women will go into labor naturally. However, if the amniotic sac has ruptured but labor has not started within 24-48 hours, contractions are often induced medically. Here, we consider 566 singleton births of primiparous women who tried to have a vaginal delivery. The binary variable of interest is whether an emergency cesarean section had to be performed. The considered nominal variable has four categories:

$X = 1:$	Labor started naturally and before the membranes ruptured.
$X = 2:$	Labor started naturally but after the membranes ruptured.
$X = 3:$	Labor was induced medically before the membranes ruptured.
$X = 4:$	Labor was induced medically after the membranes ruptured.

Our new approach, the 'naive' approach and the usual chi-square test for $k \times 2$ tables yield p-values of 10^{-4}, $3 \cdot 10^{-4}$ and $3 \cdot 10^{-5}$, respectively, and $3 \cdot 10^{-4}$, $4 \cdot 10^{-4}$ and $2 \cdot 10^{-4}$ if categories $X = 2$ and $X = 4$ (both corresponding to scenario (ii)) are merged. In both cases, the new approach yields lower p-values than the naive approach, but the difference is greater for $K = 4$ than for $K = 3$, which corroborates Figure 2 as well as the simulation results . The p-values of the usual chi-square test for $k \times 2$ tables are lower than the p-values of our test. It indicates that the probabilities $p_k = P(Y = 2|X = k)$ do not form two clearly separated clusters, which can be confirmed by considering the corresponding empirical probabilities

$$\hat{p}_1 = 0.18,\ \hat{p}_2 = 0.10,\ \hat{p}_3 = 0.33,\ \hat{p}_4 = 0.28.$$

Thus, although X and Y are associated, a binary splitting of X might be somewhat artificial. This topic could be investigated more precisely in future research.

4 Discussion and Perspectives

In this paper, we proposed a combinatorial exact method to compute the distribution of the maximally selected chi-square statistic in the context of a binary Y and a nominal (multicategorical) X. The same approach could be generalized to other association measures satisfying a specific convexity property (Shih, 2001) such as the deviance (also called cross-entropy), the Freeman-Tukey or the Cressie-Read criteria.

Since the distribution of the maximally selected chi-square statistic is different from the nominal chi-square distribution, performing a chi-square test from the 2×2 contingency table corresponding to the best partition of the variable X is an incorrect approach. In particular, it leads to strong selection bias towards X variables with more categories when used for variable selection. In contrast, our approach avoids this selection bias.

The derived distribution function can be used to construct a test of association competing with the usual chi-square test for $k \times 2$ contingency tables. In simulation studies, we found that our new test has a higher power than the usual chi-square test when the conditional probabilities $P(Y = 2|X = k)$, $k = 1, \ldots, K$, form two distinct clusters. A further advantage of our new exact and distribution-free test is that it is also applicable to very small sample sizes. Our method might be employed to prevent biased reporting of artificially low p-values in the context of a "drop-the-losers" design (Sampson and Sill, 2005), where only the most effective treatments are selected for continuation into the subsequent phase of a biopharmaceutical clinical trial, as described by Wittes (2005).

In future work, our approach could be interestingly applied to recursive partitioning algorithms. 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 (Schlittgen, 1999; Shih, 2004; Lausen et al., 2004). The p-value based association measure proposed in the present paper might also be used as a selection criterion for choosing the best nominal predictor variable and the best binary partition. Since it allows the comparison of nominal predictor variables with different numbers of categories and can be used in the case of small sample sizes, we expect it to perform better than the usual criteria in some cases. Together with the method for at least ordinally scaled predictors developed in Boulesteix (2005), it forms an homogeneous class of selection criteria for predictor variables of different types (nominal, ordinal). Alternative exact procedures for ordinal predictors are discussed in Berger (1998). In further research, one could also work on a generalization of our approach based on the minimally selected p-value rather than on the maximally selected statistic itself. Approaches using the minimal p-value as test statistic are developed, e.g., in Berger and Ivanova (2002). Such a modification of our approach would potentially allow the comparison of more complicated splits of various forms, even if the p-values are computed in completely different ways.

Another potential application of our approach is the search for complex binary splits of ordinal variables of the type $\{\{X \leq a \cup X > b\}, \{a \leq X < b\}\}$. In medical applications, such binary splits may be used e.g. when X is a blood value that has to be neither too high nor too low and Y is a binary variable of the type 'healthy/not healthy'.

Acknowledgement

I thank Florian Leitenstorfer and the reviewer for helpful comments.

References

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

Berger, V. (1998). Admissibility of exact conditional tests of stochastic order. *Journal of Statistical Planning and Inference* **66**, 39–50.

Berger, V. and Ivanova, A. (2002). Adaptive tests for ordinal data. *Journal of Modern Applied Statistical Methods* **1**, 269–280.

Boulesteix, A. -L. (2005). Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal* (to appear).

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, J. C. (1984). Classification and regression trees. Wadsworth, Monterey, CA.

Cnattingius, R., Cnattingius, S. and 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.

Gail, M. H. and Green, S. B. (1976). A generalization of the one-sided two-sample Kolmogorov-Smirnov statistic for evaluating diagnostic tests. *Biometrics* **32**, 561–570.

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 for small samples. *Biometrics* **38**, 1017–1023.

Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics and Data Analysis* **43**, 121–137.

Kang, S. -H. and Kim, S. -J. (2004). A comparison of the three conditional exact tests in two-way contingency tables using the unconditional exact power. *Biometrical Journal* **46**, 320–330.

Kim, H. and Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association* **96**, 589–604.

Koziol, J. A. (1991). On maximally selected Chi-square statistics. *Biometrics* **47**, 1557–1561.

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

Lausen, B., Lerche, R. and Schumacher, M. (2002). Maximally selected rank statistics for dose-response problems. *Biometrical Journal* **44**, 131–147.

Lausen, B. and Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics* **48**, 73–85.

Lausen, B. and Schumacher, M. (1996). Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics and Data Analysis* **21**, 307–326.

Liebermann, E. and Lang, J. M. and Cohen, A. P. and Frigoletto, F. D. and Acker, D. and 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. and Siegmund, D. (1982). Maximally selected Chi square statistics. *Biometrics* **48**, 1011-1016.

Rabinowitz, D. and Betensky, R. A. (2000). Approximating the distribution of maximally selected McNemar's statistics. *Biometrics* **56**, 897–902.

Sampson, A. R. and Sill, M. W. (2005). Drop-the-losers design: Normal case. *Biometrical Journal* **47**, 257–268.

Schlittgen, R. (1999). Regression trees for survival data - An approach to select discontinuous split points by rank statistics. *Biometrical Journal* **41**, 943–954.

Shih, Y. S. (2001). Selecting the best splits for classification trees with categorical variables. *Statistics and Probability Letters* **54**, 341–345.

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

Wittes, J. (2005). Discussion of "Drop-the-losers design: Normal case" by A. R. Sampson and M. W. Sill. *Biometrical Journal* **47**, 276–277.

Appendix

To compute the B_k, $k = 1, \ldots, K-1$, we consider the graph representing the points of coordinates $(N_{\{\sigma_1,\ldots,\sigma_j\}}, n_{2,\{\sigma_1,\ldots,\sigma_j\}})$ for $j = 1, \ldots, K-1$. After simple computations, one obtains that for a subset $S \in \mathcal{S}$ of size N_S

$$A_S > \sqrt{d} \Leftrightarrow n_{2,S} > f_\chi(N_S),$$

where f_χ denotes the function

$$f_\chi(x) = \frac{N_2 x}{N} + \frac{N_1 N_2 \sqrt{d}}{N} \sqrt{\frac{x}{N}(1 - \frac{x}{N})(\frac{1}{N_1} + \frac{1}{N_2})}. \tag{8}$$

For conditions (5) and (6) to be satisfied, the graph $(N_{\{\sigma_1,\ldots,\sigma_j\}}, n_{2,\{\sigma_1,\ldots,\sigma_j\}})$ must lie below or on the boundary of equation (8) for $j = 1,\ldots,k-1$ and above it for $j = k$:

$$\begin{aligned}
(5) \quad &\Leftrightarrow \quad \forall j < k,\ n_{2,\{\sigma_1,\ldots,\sigma_j\}} \leq f_\chi(N_{\{\sigma_1,\ldots,\sigma_j\}}) \\
&\Leftrightarrow \quad \forall j < k,\ n_{2,\{\sigma_j\}} \leq f_\chi(N_{\{\sigma_1,\ldots,\sigma_j\}}) - n_{2,\{\sigma_1,\ldots,\sigma_{j-1}\}}, \\
(6) \quad &\Leftrightarrow \quad n_{2,\{\sigma_1,\ldots,\sigma_k\}} > f_\chi(N_{\{\sigma_1,\ldots,\sigma_k\}}) \\
&\Leftrightarrow \quad n_{2,\{\sigma_k\}} > f_\chi(N_{\{\sigma_1,\ldots,\sigma_k\}}) - n_{2,\{\sigma_1,\ldots,\sigma_{k-1}\}}.
\end{aligned}$$

Condition (4) can be interpreted in terms of concavity of the graph $(N_{\{\sigma_1,\ldots,\sigma_j\}}, n_{2,\{\sigma_1,\ldots,\sigma_j\}})$ and rewritten as

$$\begin{aligned}
\forall j = 2,\ldots,K, \quad n_{2,\{\sigma_j\}} \quad &< \tfrac{n_{2,\{\sigma_{j-1}\}}}{m_{\sigma_{j-1}}} m_{\sigma_j} \quad \text{if } \sigma_{j-1} > \sigma_j \\
&\leq \tfrac{n_{2,\{\sigma_{j-1}\}}}{m_{\sigma_{j-1}}} m_{\sigma_j} \quad \text{otherwise.}
\end{aligned}$$

Let $upper(n_{2,\{\sigma_{j-1}\}}, m_{\sigma_{j-1}}, m_{\sigma_j}, \sigma_j, \sigma_{j-1})$ denote the greatest allowed value for $n_{2,\{\sigma_j\}}$ corresponding to condition (4).

To sum up, B_k, $k = 2,\ldots,K-1$, is the number of ways to choose N_2 observations from a sample of N observations such that $n_{2,\{\sigma_j\}} \in I_j^{(B_k)}$, for all $j = 1,\ldots,K-1$, with

$$\begin{aligned}
I_1^{(B_k)} &= [0, \min(f_\chi(m_{\sigma_1}), m_{\sigma_1}, N_2)], \\
I_j^{(B_k)} &= [0, \min(f_\chi(N_{\{\sigma_1,\ldots,\sigma_j\}}) - n_{2,\{\sigma_1,\ldots,\sigma_{j-1}\}}, upper(n_{2,\{\sigma_{j-1}\}}, m_{\sigma_{j-1}}, m_{\sigma_j}, \sigma_j, \sigma_{j-1}))] \\
&\quad \text{for } j = 2,\ldots,k-1, \\
I_k^{(B_k)} &=]f_\chi(N_{\{\sigma_1,\ldots,\sigma_k\}}) - n_{2,\{\sigma_1,\ldots,\sigma_{k-1}\}}, upper(n_{2,\{\sigma_{k-1}\}}, m_{\sigma_{k-1}}, m_{\sigma_k}, \sigma_k, \sigma_{k-1})], \\
I_j^{(B_k)} &= [0, upper(n_{2,\{\sigma_{j-1}\}}, m_{\sigma_{j-1}}, m_{\sigma_j}, \sigma_j, \sigma_{j-1})] \\
&\quad \text{for } j > k.
\end{aligned}$$

For $k = 1$, the intervals $I_j^{(B_1)}$ are simply given as

$$\begin{aligned}
I_1^{(B_1)} &=]f_\chi(m_{\sigma_1}), \min(m_{\sigma_1}, N_2)] \qquad \text{for } j = 1, \\
I_j^{(B_1)} &= [0, upper(n_{2,\{\sigma_{j-1}\}}, m_{\sigma_{j-1}}, m_{\sigma_j}, \sigma_j, \sigma_{j-1})] \quad \text{for } j > 1.
\end{aligned}$$

Using these notations, B_k is then obtained as

$$B_k = \sum_{i_1 \in I_1^{(B_k)}} \binom{m_{\sigma_1}}{i_1} \cdot \left(\ldots \left(\sum_{i_{K-1} \in I_{K-1}^{(B_k)}} \binom{m_{\sigma_{K-1}}}{i_{K-1}} \cdot \binom{m_{\sigma_K}}{N_2 - \sum_{q=1}^{K-1} i_q} \right) \right). \tag{9}$$

The computational efficiency can be considerably improved by noticing that $\hat{p}_{\sigma_1} \geq \cdots \geq \hat{p}_{\sigma_K}$ implies

$$\frac{n_{2,\{\sigma_j\}}}{m_{\sigma_j}} \geq \frac{N_2 - n_{2,\{\sigma_1,\ldots,\sigma_{j-1}\}}}{N - N_{\{\sigma_1,\ldots,\sigma_{j-1}\}}},$$

for $j = 1,\ldots,K$. If $n_{2,\{\sigma_1,\ldots,\sigma_k\}} > f_\chi(N_{\{\sigma_1,\ldots,\sigma_k\}})$, then we even have

$$\frac{n_{2,\{\sigma_j\}}}{m_{\sigma_j}} \geq \frac{f_\chi(N_{\{\sigma_1,\ldots,\sigma_k\}}) - n_{2,\{\sigma_1,\ldots,\sigma_{j-1}\}}}{N_{\{\sigma_1,\ldots,\sigma_k\}} - N_{\{\sigma_1,\ldots,\sigma_{j-1}\}}},$$

for $j = 1,\ldots,k-1$. Computation time can be spared by replacing 0 by the corresponding lower bounds in the intervals $I_j^{(B_k)}$.