Journal of Applied Mathematics, Statistics and Informatics (JAMSI), **7** (2011), No. 1

SOME TESTS FOR EVALUATION OF CONTINGENCY TABLES (FOR BIOMEDICAL APPLICATIONS)

JAN KALINA

Abstract

This paper presents some methods for evaluating contingency tables, especially suitable for biomedical applications. The test of linear trend is recalled, which is derived in an alternative way in the model of weighted regression. The maximal χ^2 test statistic for dichotomized continuous data is studied, which was proposed by Betensky and Rabinowitz [1999] for the evaluation of epidemiological studies; here we correct a mistake of the original paper. Finally Cochran's test and Mantel-Haenszel test are compared; these are tests for the hypothesis of conditional homogeneity in $2 \times 2 \times K$ tables, popular mainly in survival analysis. We propose an exact unconditional version of Cochran's test of conditional homogeneity.

Mathematics Subject Classification 2000: 62H17, 62J12, 62F03
General Terms: Contingency tables, Hypothesis testing
Additional Key Words and Phrases: generalized linear models, exact unconditional test, linear trend, maximal χ^2 statistic, Mantel-Haenszel test

1. INTRODUCTION

Analysis of categorical data is a classical field of mathematical statistics, which still acquires an intensive attention. Standard monographs on analysis of contingency tables [Agresti 1990, Simonoff 2003, Zvárová and Malý 2003] offer a wide variety of classical methods, while many other modern methods have been proposed recently. Contingency tables represent a crucial method for data analysis in various branches of applied research, where their proper analysis allows significant and trustworthy results.

Recently there have appeared both theoretical and applied papers on analyzing categorical data. We point out that many of the following references are published in biostatistical journals. Recent theoretical work includes the study of conditional marginal homogeneity by Tahata et al. [2008], hypothesis tests for tables with structural zero by Tang and Jiang [2011] or robust approaches to the generalized linear models for categorical data by Čížek [2006]. LaMotte [2005] supports the idea that exact inference for generalized linear models is still a big issue in current mathematical statistics. Geenens and Simar [2010] study non-parametric methods for contingency tables and claim that a two-way contingency table is a complicated

This research is fully supported by Center of Biomedical Informatics, project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.

high-dimensional statistical model. In general iterative computations for parameter estimation are required both by the parametric (generalized linear models) and non-parametric methods. Kalina [2010] gives a comparison of unconditional, exact conditional and exact unconditional tests of hypotheses in the context of contingency tables.

Practical papers on contingency tables from recent time include Svozil et al. [2008] analyzing tables of counts, which correspond to different dinucleotides classified to different three-dimensional conformations of DNA molecules. Holinka and Smutek [2010] study two-way contingency tables obtained by different classification methods comparing different approaches to the diagnosis of autoimmune thyroiditis in endocrinologic patients.

This paper has the following structure. Section 2 recalls the Cochran-Armitage test of linear trend, which is later used in Section 3. Nevertheless we derive the test in an alternative way in a weighted linear regression model. Section 3 studies the maximal χ^2 test statistic for continuous data, which are dichotomized (separated into two parts based on comparing the measurements with a threshold). We correct a mistake of Betensky and Rabinowitz [1999], who proposed this test for epidemiological applications. Finally Section 4 compares the null hypotheses of conditional homogeneity and conditional independence in $2 \times 2 \times K$ tables. These are important tests for clinical or epidemiological studies [Zvárová and Malý 2003]. There the Cochran's test and Mantel-Haenszel test are compared and an exact unconditional version of Cochran's test of conditional homogeneity is proposed.

From the methodological point of view, some of the asymptotic tests described in this paper belong to the context of generalized linear models. The paper itself is rather focused on alternative ways to handle their analysis. The test of linear trend (Section 2) is classically derived as the Wald test [Rao 1973] in the logistic regression, but we present an alternative approach based on weighted linear regression. The χ^2 test statistic is classically derived as the Pearson's χ^2 goodness-of-fit test, but can be derived as the score test [Rao 1973] based on the likelihood of the contingency table [Day and Byar 1973]. The asymptotic tests of Section 4 can be also derived as asymptotic tests based on the likelihood function; however we show that the asymptotic tests can be derived directly from the likelihood function of the contingency table, without assuming generalized linear models. In Section 5 we focus on the exact alternative to the classical asymptotic approach to testing conditional homogeneity.

2. TEST OF LINEAR TREND

We describe the classical Cochran-Armitage test of linear trend [Cochran 1954, Armitage 1955], which will be used in Section 3. We present an alternative way for deriving the test; classically the test is derived in the context of logistic regression [Agresti 1990] or as the Pearson's χ^2 goodness-of-fit test [Zvárová and Malý 2003].

Let us assume the random samples to be measured in the total number of K populations (groups) for $K \geq 3$. A binary variable is observed in each randomly selected unit. The columns of the table of observed counts

	Group 1	Group 2	$\cdots$	Group K	$\sum$
Success	n_{11}	n_{12}	$\cdots$	n_{1K}	$n_{1\cdot}$
Failure	n_{21}	n_{22}	$\cdots$	n_{2K}	$n_{2\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K}$	n

(1)

correspond to the binomial model. The table

	Group 1	Group 2	$\cdots$	Group K
Success	π_1	π_2	$\cdots$	π_K
Failure	$1-\pi_1$	$1-\pi_2$	$\cdots$	$1-\pi_K$
$\sum$	1	1	$\cdots$	1

(2)

introduces the notation of the corresponding probabilities.

We assume the columns of table (1) to be ordered according to values of a certain ordinal variable. The test of linear trend is a test of the null hypothesis $H_0:\ \pi_1 = \pi_2 = \cdots = \pi_K$ against the alternative hypothesis

$$H_1:\ \pi_k = \pi + \beta v_k, \quad \text{for } k = 1,\ldots,K,\ \beta \neq 0, \tag{3}$$

with unknown constants π and β and with known regressors (scores) $v_1,\ldots,v_K$. This is actually a test of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$, where β is the slope in the linear trend. The first step is to assign the scores $v_1,\ldots,v_K$ to columns of table (1). It is popular to assign scores which are equidistant or which correspond to a prior expectation of the trend. The test holds the type I error, but its power depends on the subjective choice of the scores.

The test statistic of the Cochran-Armitage test

$$\chi_T^2 = \frac{n\left[n\sum_{k=1}^K v_k n_{1k} - n_{1\cdot}\sum_{k=1}^K v_k n_{\cdot k}\right]^2}{n_{1\cdot}n_{2\cdot}\left[n\sum_{k=1}^K v_k^2 n_{\cdot k} - \left(\sum_{k=1}^K v_k n_{\cdot k}\right)^2\right]} \tag{4}$$

has asymptotically χ_1^2 distribution under H_0. Let us use the notation $\chi_1^2(\alpha)$ for the critical value of χ_1^2 distribution. The null hypothesis is rejected if $\chi_T^2 \geq \chi_1^2(\alpha)$.

We derive the test statistic χ_T^2 and its asymptotic null distribution in a way alternative to Agresti [1990]. Let us denote the estimated probabilities of success in particular groups

$$\hat{\pi}_1 = \frac{n_{11}}{n_{\cdot 1}},\ldots,\hat{\pi}_K = \frac{n_{1K}}{n_{\cdot K}}. \tag{5}$$

Let π denote the true probability of success across groups under H_0. Let us consider the regression model

$$\begin{pmatrix}\hat{\pi}_1\\ \vdots\\ \hat{\pi}_K\end{pmatrix} = \begin{pmatrix}1 & v_1\\ \vdots & \vdots\\ 1 & v_K\end{pmatrix}\begin{pmatrix}\pi\\ \beta\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_K\end{pmatrix}, \tag{6}$$

which means

$$\hat{\pi}_k = \pi + v_k\beta + \varepsilon_k \quad \text{for } k \in \{1,2,\ldots,K\}. \tag{7}$$

Theorem 2.1. *The Cochran-Armitage statistic χ^2_T is the Wald test statistic for the slope in the weighted regression model (6).*

Proof. The random variable $\hat{\pi}_k$ has the variance

$$\mathsf{var}\,\hat{\pi}_k = \mathsf{var}\,\frac{n_{1k}}{n_{\cdot k}} = \frac{1}{n^2_{\cdot k}}\,\mathsf{var}\,n_{1k} = \frac{\pi(1-\pi)n_{\cdot k}}{n^2_{\cdot k}} = \frac{\pi(1-\pi)}{n_{\cdot k}} \tag{8}$$

for $k = 1, \ldots, K$. Assuming H_0 and large sample sizes, the vector of responses $(\hat{\pi}_1, \ldots, \hat{\pi}_K)^T$ is asymptotically K-variate normally distributed with expectation $(\pi, \pi, \ldots, \pi)^T$ and diagonal variance matrix

$$\mathbf{W}^{-1} = \hat{\pi}(1-\hat{\pi}) \cdot \mathsf{diag}\left\{\frac{1}{n_{\cdot 1}}, \ldots, \frac{1}{n_{\cdot K}}\right\}. \tag{9}$$

The diagonal variance matrix corresponds to the independence of $\varepsilon_1, \ldots, \varepsilon_K$ in the model (6).

This is a model of weighted regression. In general assuming a normally distributed response $\mathbf{Y} \sim \mathsf{N}_K(\mathbf{X}\boldsymbol{\gamma}, \mathbf{W}^{-1})$ the least squares estimator of the vector parameter $\boldsymbol{\gamma}$ is equal to

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \tag{10}$$

and its variance matrix is $\mathsf{var}\,\hat{\boldsymbol{\gamma}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$. In our case the two-dimensional parameter $(\pi, \beta)^T$ plays the role of $\boldsymbol{\gamma}$. The estimator of β turns out to be

$$\hat{\beta} = \frac{n\sum_{k=1}^K v_k n_{1k} - n_{1\cdot}\sum_{k=1}^K v_k n_{\cdot k}}{n\sum_{k=1}^K v_k^2 n_{\cdot k} - \left(\sum_{k=1}^K v_k n_{\cdot k}\right)^2} \tag{11}$$

with the variance

$$\mathsf{var}\,\hat{\beta} = \frac{n_{1\cdot}n_{2\cdot}}{n\left[n\sum_{k=1}^K v_k^2 n_{\cdot k} - \left(\sum_{k=1}^K v_k n_{\cdot k}\right)^2\right]}. \tag{12}$$

The Wald test [Rao 1973] of $H_0 : \beta = 0$ against the two-sided alternative hypothesis is based on the statistic

$$\frac{\hat{\beta}^2}{\mathsf{var}\,\hat{\beta}} = \frac{n\left[n\sum_{k=1}^K v_k n_{1k} - n_{1\cdot}\sum_{k=1}^K v_k n_{\cdot k}\right]^2}{n_{1\cdot}n_{2\cdot}\left[n\sum_{k=1}^K v_k^2 n_{\cdot k} - \left(\sum_{k=1}^K v_k n_{\cdot k}\right)^2\right]} = \chi^2_T, \tag{13}$$

which is asymptotically χ^2_1 distributed under H_0. This completes the proof. □

In practice it may be useful to test the null hypothesis of linear trend (3) against the alternative hypothesis of the saturated model. Such test is easily performed with the likelihood ratio test [Rao 1973, Agresti 1990], which is equivalent to the residual deviance in the context of generalized linear models [Hardin and Hilbe 2007]. Let us introduce the notation M_0, M_1 and M_2 for the models of homogeneity, linear trend and saturated model, respectively. It holds

$$M_0 : \text{homogeneity} \subset M_1 : \text{linear trend} \subset M_2 : \text{saturated model}, \tag{14}$$

where $\subset$ denotes a submodel. Let us introduce a general notation G^2 for the test statistic of the likelihood ratio test; for example $G^2(M_1|M_2)$ will stand for the likelihood ratio statistic of the test of the null hypothesis of model M_1 against the alternative hypothesis of model M_2. It holds that the test statistic

$$G^2(M_1|M_2) = G^2(M_0|M_2) - G^2(M_0|M_1) \tag{15}$$

is asymptotically χ^2 distributed assuming M_1, where the number of degrees of freedom corresponds to the difference between the numbers of parameters in M_2 and M_1 [Agresti 1990]. This is the test of linear trend against the saturated model. From the computational point of view the test can be computed easily, because $G^2(M_0|M_2)$ is a classical formula [Agresti 1990] and $G^2(M_0|M_1)$ can be obtained using logistic regression in the model (3).

3. MAXIMAL χ^2 TEST STATISTIC

This section studies the test of homogeneity (or linear trend) using such test statistic, which is maximal over all possible values of the threshold [Betensky and Rabinowitz 1999]. It happens often in epidemiological studies that values of a continuous variable are divided to two groups based on comparing the measured values with a threshold [Miller and Siegmund 1982]. Betensky and Rabinowitz [1999] illustrate the methodology on data from an AIDS clinical trial. The comparison of measurements with a threshold simplifies the structure of the data, but obviously leads to a loss of information. If the measurement is performed on several samples or under several conditions, a contingency table is created. An example is the body temperature of patients; the measurements can be divided to the group of healthy patients and the group of patients with a high temperature.

Let us assume the following measurement study. A continuous variable is measured in the total number of K groups with fixed sample sizes. We order the values from all samples jointly and denote them by $x_1 \leq x_2 \leq \cdots \leq x_n$. The measurements are compared with a given threshold. We assume that none of the values is equal to the threshold. We use the notation D for the set $D = \mathbb{R} \backslash \{x_1, x_2, \ldots, x_n\}$. If a particular value exceeds the threshold, the situation is denoted as success; otherwise the situation is denoted as failure. Each of the columns of the table of observed counts (1) corresponds to the multinomial model.

We consider the test of homogeneity of the binomial distributions. Clearly the χ^2 test statistics depends on the value of the threshold. For a fixed threshold x, let us denote the χ^2 test statistic as $\chi^2(x)$. The value of the test statistic

$$\max_{x \in \mathbb{R}} \chi^2(x) \tag{16}$$

is considered, which cannot be however compared with the standard p-value of the χ^2 test, because such decision rule would increase the type I error. Further we study the asymptotic null distribution of (16).

Test of homogeneity

We describe the test of homogeneity based on the maximal χ^2 test statistic (16). Our computation reveals a mistake of Betensky and Rabinowitz [1999]. We assume large sample sizes $n_{\cdot 1}, n_{\cdot 2}, \ldots, n_{\cdot K}$ and the null hypothesis of homogeneity to be fulfilled. Nevertheless for practical purposes it is not reasonable to maximize the threshold over $\mathbb{R}$, because extreme values of the threshold are not desirable [Halpern 1982].

Let us use the notation $F(x)$ for the function defined in the following way. The argument $x \in D$ of the function corresponds to a given threshold. A table of counts (1) is considered corresponding to the threshold x. The value of $F(x)$ is defined as $1-\hat{\pi} = n_{2\cdot}/n$. This is an estimate of the probability of failure under H_0, which is the probability that (any) observation is smaller than the threshold.

Let

$$\{B_1(t);\ t \in [0,1]\}, \{B_2(t);\ t \in [0,1]\}, \ldots, \{B_{K-1}(t);\ t \in [0,1]\} \tag{17}$$

be independent Brownian bridges [Billingsley 1999]. Let us consider (any) such fixed values x_L and x_R that $x_L > x_1$ and $x_R < x_n$. Betensky and Rabinowitz [1999] prove that

$$\max_{x \in [x_L, x_R]} \chi^2(x) \tag{18}$$

has the same distribution as

$$\max_{t \in [F(x_L), F(x_R)]} \frac{\sum_{k=1}^{K-1} B_k^2(t)}{t(1-t)} \tag{19}$$

for large sample sizes under H_0.,

Let us describe an approximation to this distribution. We denote

$$r = \frac{F(x_R) \cdot [1-F(x_L)]}{F(x_L) \cdot [1-F(x_R)]}. \tag{20}$$

For $b \to \infty$ there holds an approximative relationship of [James et al. 1987]

$$\begin{aligned} &P_{H_0}\left(\max_{t \in [F(x_L), F(x_R)]} \sqrt{\frac{\sum_{k=1}^{K-1} B_k^2(t)}{t(1-t)}} \geq b \right) = \\ &= \frac{b^{K-1} \exp\{-\frac{1}{2}b^2\}}{2^{\frac{K-3}{2}} \Gamma\left(\frac{K-1}{2}\right)} \cdot \left[\frac{1}{2}\left(1 - \frac{K-1}{b^2}\right) \log r + \frac{2}{b^2} + o\left(\frac{1}{b^2}\right) \right], \end{aligned} \tag{21}$$

where Γ denotes the gamma function. The special case $K=2$ was derived by Miller and Siegmund [1982]. For a positive b it holds

$$\begin{aligned} &P_{H_0}\left(\max_{t \in [F(x_L), F(x_R)]} \sqrt{\frac{\sum_{k=1}^{K-1} B_k^2(t)}{t(1-t)}} \geq b \right) = \\ &= P_{H_0}\left(\sqrt{\max_{t \in [F(x_L), F(x_R)]} \frac{\sum_{k=1}^{K-1} B_k^2(t)}{t(1-t)}} \geq b \right) = \\ &= P_{H_0}\left(\max_{t \in [F(x_L), F(x_R)]} \frac{\sum_{k=1}^{K-1} B_k^2(t)}{t(1-t)} \geq b^2 \right). \end{aligned} \tag{22}$$

Therefore taking $w = b^2$ to (21) leads us to

$$P_{H_0}\left(\max_{t\in[F(x_L),F(x_R)]}\frac{\sum_{k=1}^{K-1}B_k^2(t)}{t(1-t)}\geq w\right)=$$
$$=\frac{w^{\frac{K-1}{2}}\exp\{-\frac{1}{2}w\}}{2^{\frac{K-3}{2}}\Gamma\left(\frac{K-1}{2}\right)}\cdot\left[\frac{1}{2}\left(1-\frac{K-1}{w}\right)\log r+\frac{2}{w}+o\left(\frac{1}{w}\right)\right] \tag{23}$$

for $w \to \infty$.

For the practical computation Miller and Siegmund [1982] recommend to select $\varepsilon \in (0, \frac{1}{2})$. Let us use the notation D_ε for the set of possible thresholds $x \in D$ fulfilling

$$\varepsilon \leq \hat{\pi} = \frac{n_{1.}}{n} \leq 1 - \varepsilon. \tag{24}$$

We keep only the middle $\varepsilon \cdot 100$ % observations out of $x_1, \ldots, x_n$ so that the set D_ε does not contain the extreme values. Let us find the maximal $\chi^2(x)$ statistic over D_ε. The approximative p-value of the set is computed from (23). There w is replaced by

$$\max_{x\in D_\varepsilon}\chi^2(x), \tag{25}$$

the term $o(\frac{1}{w})$ is ignored and the maximization is computed over $t \in [\varepsilon, 1-\varepsilon]$. Therefore we obtain an approximation for the p-value of the test in the form

$$p \doteq \frac{w^{\frac{K-1}{2}}\exp\{-\frac{1}{2}w\}}{2^{\frac{K-3}{2}}\Gamma\left(\frac{K-1}{2}\right)}\cdot\left[\frac{1}{2}\left(1-\frac{K-1}{w}\right)\log\left(\frac{1-\varepsilon}{\varepsilon}\right)^2+\frac{2}{w}\right]. \tag{26}$$

The test rejects the null hypothesis of homogeneity if $p \leq \alpha$. The choice of ε should be based on the experience of the researcher; a too small value of ε leads to a poor approximation in (26).

Test of linear trend

In a similar manner the test of linear trend (Section 2) can be used for continuous measurements, which are divided into two parts based on comparison with a threshold. We assume the data to be measured in K samples to form the the table (1). Each column corresponds to the binomial distribution and (2) is the corresponding table of probabilities. The aim is to test the null hypothesis $H_0 : \pi_1 = \pi_2 = \cdots = \pi_K$ against the alternative hypothesis (3) with selected fixed scores $v_1, \ldots, v_K$. Similarly as above we consider the situation popular in epidemiological studies [Betensky and Rabinowitz 1999] that the threshold is found to maximize the Cochran's test statistic χ^2_T for the corresponding table. We repeat the detailed steps of the approximation to obtain an approximative p-value of the test.

The observed table of counts depends on the choice of the threshold. The test can be based on the maximal possible $\chi^2_T(x)$ statistic of the Cochran-Armitage test. The threshold x is optimized over the interval $x \in [x_L, x_R]$, trimming away extreme values using $x_L > x_1$ and $x_R < x_n$. The statistic

$$\max_{x\in[x_L,x_R]}\chi^2_T(x) \tag{27}$$

has approximately the same distribution as

$$\max_{t \in [F(x_L), F(x_R)]} \frac{B^2(t)}{t(1-t)} \tag{28}$$

for large sample sizes under H_0 and this distribution can be approximated by (23) using $K = 2$ [Betensky and Rabinowitz 1999].

Again let us choose $\varepsilon \in (0, \frac{1}{2})$ and find the maximal $\chi^2_T(x)$ for thresholds $x \in D_\varepsilon$ fulfilling (24). Let us denote

$$w = \max_{x \in D_\varepsilon} \chi^2_T(x). \tag{29}$$

Particularly we evaluate the approximation to the p-value of the test by

$$\begin{aligned} p &\doteq \frac{\sqrt{2w}\exp\{-\frac{1}{2}w\}}{\sqrt{\pi}} \left[\frac{1}{2}\left(1 - \frac{1}{w}\right)\log\left(\frac{1-\varepsilon}{\varepsilon}\right)^2 + \frac{2}{w}\right] = \\ &= \frac{\exp\left\{-\frac{1}{2}w\right\}}{\sqrt{2\pi w}}\left[(w-1)\log\left(\frac{1-\varepsilon}{\varepsilon}\right)^2 + 4\right]. \end{aligned} \tag{30}$$

The test rejects the null hypothesis of homogeneity in favor of the alternative hypothesis of the linear trend if $p \leq \alpha$.

4. CONDITIONAL HOMOGENEITY AND CONDITIONAL INDEPENDENCE

In clinical and epidemiological studies it is often important to deal with confounding, which is the effect of an additional variable correlated with both the response and regressors. One of possible solutions is a sophisticated design of the experiment, which controls the level of the confounding variable [Zvárová and Malý 2003]. Different approaches to the design often lead to the models of conditional homogeneity or conditional independence. We describe the difference between the two models. The Cochran's test and Mantel-Haenszel test are both used interchangeably for both hypotheses. Here we point out that the Cochran's test is suitable for the null hypothesis of conditional homogeneity, while the Mantel-Haenszel test arises only from the model with conditional independence. Finally Section 5 proposes an exact unconditional test of conditional homogeneity.

We assume a binary variable to be measured in two independent random samples with fixed sample sizes over K strata ($K \geq 2$), corresponding to different populations or different conditions. The resulting table of counts

	Group 1	Group 2	$\sum$	$\cdots$	Group 1	Group 2	$\sum$
Success	n_{111}	n_{121}	$n_{1\cdot 1}$	$\cdots$	n_{11K}	n_{12K}	$n_{1\cdot K}$
Failure	n_{211}	n_{221}	$n_{2\cdot 1}$	$\cdots$	n_{21K}	n_{22K}	$n_{2\cdot K}$
$\sum$	$n_{\cdot 11}$	$n_{\cdot 21}$	$n_{\cdot\cdot 1}$	$\cdots$	$n_{\cdot 1K}$	$n_{\cdot 2K}$	$n_{\cdot\cdot K}$

(31)

corresponds to the model with the table of probabilities

	Group 1	Group 2	$\cdots$	Group 1	Group 2
Success	π_{11}	π_{21}	$\ldots$	π_{1K}	π_{2K}
Failure	$1 - \pi_{11}$	$1 - \pi_{21}$	$\ldots$	$1 - \pi_{1K}$	$1 - \pi_{2K}$
$\sum$	1	1	$\ldots$	1	1

. (32)

The probability of success in the j-th group ($j = 1, 2$) in the k-th stratum ($k = 1, \dots, K$) is denoted by π_{jk}. At the end of this section we describe a numerical example to illustrate the model.

In general we consider the test of the null hypothesis

$$H_0 : \ \pi_{11} = \pi_{21} \quad \& \quad \pi_{12} = \pi_{22} \quad \& \quad \dots \quad \& \quad \pi_{1K} = \pi_{2K} \tag{33}$$

against the alternative hypothesis that the null hypothesis is not true. Here H_0 is the hypothesis of homogeneity of binomial distribution over strata, in other words the hypothesis of conditional homogeneity, in other words homogeneity of two binomial populations for each fixed value of a certain level of an additional factor variable. Particularly in the example with medicines A and B the null hypothesis is tested that in each hospital there are both medicines equivalently effective. At the same time we allow a difference of the effect of the medicines among hospitals. This allows a different level of the medical care in each hospital to be involved in the model.

Cochran [1954] proposed an asymptotic test of conditional homogeneity based on the statistic

$$T_C = \sum_{k=1}^{K} \frac{n_{11k}n_{\cdot 2k} - n_{12k}n_{\cdot 1k}}{n_{\cdot\cdot k}} \Bigg/ \sqrt{\sum_{k=1}^{K} \frac{n_{1\cdot k}n_{2\cdot k}n_{\cdot 1k}n_{\cdot 2k}}{n^3_{\cdot\cdot k}}} \ . \tag{34}$$

The test rejects the null hypothesis, if $|T_C| \geq \Phi^{-1}(1-\alpha/2)$, where Φ is the quantile function of $\mathsf{N}(0,1)$ distribution. This is an asymptotic unconditional test.

We compare the Cochran's test and Mantel-Haenszel test and stress the difference between them, which will be illustrated in details in the proofs of the asymptotic null distributions of both test statistics. The Cochran's test is derived as a test of conditional homogeneity, while the Mantel-Haenszel test is derived as a test of conditional independence.

Theorem 4.1. *The Cochran's test statistic T_C of conditional homogeneity is asymptotically $\mathsf{N}(0,1)$ distributed under H_0.*

Proof. Assuming H_0, the table (31) is a set of K two-way tables. The probability of success in them under H_0 will be denoted by $\pi_1, \dots, \pi_K$. Random variables n_{11k} and n_{12k} for each $k = 1, \dots, K$ follow binomial distribution

$$n_{11k} \sim \mathsf{Bi}(n_{\cdot 1k}, \pi_k) \quad \text{and} \quad n_{12k} \sim \mathsf{Bi}(n_{\cdot 2k}, \pi_k). \tag{35}$$

The values $\pi_1, \dots, \pi_K$ are however unknown. The following approach is approximative for large sample sizes. The unknown value π_k will be replaced by its maximum likelihood estimator in the corresponding binomial model. That replaces (35) by the binomial distributions

$$n_{11k} \sim \mathsf{Bi}\left(n_{\cdot 1k}, \frac{n_{1\cdot k}}{n_{\cdot\cdot k}}\right) \quad \text{and} \quad n_{12k} \sim \mathsf{Bi}\left(n_{\cdot 2k}, \frac{n_{1\cdot k}}{n_{\cdot\cdot k}}\right). \tag{36}$$

Let us define for each $k = 1, \dots, K$ the expressions

$$d_k = \frac{n_{11k}}{n_{\cdot 1k}} - \frac{n_{12k}}{n_{\cdot 2k}} = \frac{n_{11k}n_{\cdot 2k} - n_{12k}n_{\cdot 1k}}{n_{\cdot 1k}n_{\cdot 2k}} \tag{37}$$

and

$$w_k = \frac{n_{\cdot 1k} n_{\cdot 2k}}{n_{\cdot 1k} + n_{\cdot 2k}} = \frac{n_{\cdot 1k} n_{\cdot 2k}}{n_{\cdot\cdot k}}. \tag{38}$$

Let us denote $w = \sum_{k=1}^{K} w_k$ and

$$\bar{d} = \sum_{k=1}^{K} \frac{w_k d_k}{w} = \frac{1}{w} \sum_{k=1}^{K} \frac{n_{11k} n_{\cdot 2k} - n_{12k} n_{\cdot 1k}}{n_{\cdot\cdot k}}. \tag{39}$$

For each k it holds

$$\mathsf{E} d_k = \frac{1}{n_{\cdot 1k}} \mathsf{E} n_{11k} - \frac{1}{n_{\cdot 2k}} \mathsf{E} n_{12k} = 0 \tag{40}$$

and therefore $\mathsf{E}\bar{d}$ is equal to

$$\mathsf{E}\bar{d} = \frac{1}{w} \sum_{k=1}^{K} w_k \mathsf{E} d_k = 0. \tag{41}$$

Because d_k is a linear combination of independent random variables n_{11k} and n_{12k} with binomial distributions, we easily obtain

$$\begin{aligned} \mathsf{var}\, d_k &= \frac{1}{n^2_{\cdot 1k} n^2_{\cdot 2k}} \left(n^2_{\cdot 2k} n_{\cdot 1k} \frac{n_{1\cdot k} n_{2\cdot k}}{n^2_{\cdot\cdot k}} + n^2_{\cdot 1k} n_{\cdot 2k} \frac{n_{1\cdot k} n_{2\cdot k}}{n^2_{\cdot\cdot k}} \right) = \\ &= \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot 1k} n_{\cdot 2k}}{n^2_{\cdot 1k} n^2_{\cdot 2k} n^2_{\cdot\cdot k}} \left(n_{\cdot 2k} + n_{\cdot 1k} \right) = \frac{n_{1\cdot k} n_{2\cdot k}}{n_{\cdot 1k} n_{\cdot 2k} n_{\cdot\cdot k}}. \end{aligned} \tag{42}$$

From the independence of $d_1, d_2, \ldots, d_K$ we obtain

$$\begin{aligned} \mathsf{var}\, \bar{d} &= \frac{1}{w^2} \sum_{k=1}^{K} w_k^2 \mathsf{var}\ d_k = \\ &= \frac{1}{w^2} \sum_{k=1}^{K} \frac{n^2_{\cdot 1k} n^2_{\cdot 2k}}{n^2_{\cdot\cdot k}} \cdot \frac{n_{1\cdot k} n_{2\cdot k}}{n_{\cdot 1k} n_{\cdot 2k} n_{\cdot\cdot k}} = \frac{1}{w^2} \sum_{k=1}^{K} \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot 1k} n_{\cdot 2k}}{n^3_{\cdot\cdot k}}. \end{aligned} \tag{43}$$

For large values of constants

$$n_{\cdot 11}, n_{\cdot 21}, n_{\cdot 12}, n_{\cdot 22}, \ldots, n_{\cdot 1K}, n_{\cdot 2K} \tag{44}$$

the statistic $\bar{d}$ is asymptotically normally distributed. We obtain that the statistic

$$\frac{\bar{d} - \mathsf{E}\bar{d}}{\sqrt{\mathsf{var}\, \bar{d}}} = \sum_{k=1}^{K} \frac{n_{11k} n_{\cdot 2k} - n_{12k} n_{\cdot 1k}}{n_{\cdot\cdot k}} \Bigg/ \sqrt{\sum_{k=1}^{K} \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot 1k} n_{\cdot 2k}}{n^3_{\cdot\cdot k}}} = T_C \tag{45}$$

is asymptotically $\mathsf{N}(0, 1)$ distributed under H_0, which completes the proof. □

Let us note that for the special case $K = 1$ this test is equivalent to the asymptotic Pearson χ^2 test of homogeneity.

The most commonly used test of conditional homogeneity in $2 \times 2 \times K$ tables is however the asymptotic Mantel-Haenszel test, as claimed by Freidlin and Gastwirth [1999]. The test is based on the test statistic

$$\chi^2_{MH} = \left[\sum_{k=1}^{K} \frac{n_{11k} n_{\cdot 2k} - n_{12k} n_{\cdot 1k}}{n_{\cdot\cdot k}} \right]^2 \Bigg/ \sum_{k=1}^{K} \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot k} n_{\cdot 2k}}{(n_{\cdot\cdot k} - 1) n^2_{\cdot\cdot k}} \tag{46}$$

and rejects the null hypothesis if $\chi^2_{MH} \geq \chi^2_1(\alpha)$ [Agresti 1990]. It is also popular in survival analysis applications [Zvárová and Malý 2003]. [Day and Byar 1979] derive the Mantel-Haenszel as the score test [Rao 1973] based on the joint likelihood function of the table.

Nevertheless the Mantel-Haenszel test is derived for a different situation, namely for a table with fixed marginal counts corresponding to rows and columns

$$n_{1\cdot 1}, n_{2\cdot 1}, n_{1\cdot 2}, n_{2\cdot 2}, \ldots, n_{1\cdot K}, n_{2\cdot K}, \quad n_{\cdot 11}, n_{\cdot 21}, n_{\cdot 12}, n_{\cdot 22}, \ldots, n_{\cdot 1K}, n_{\cdot 2K}. \tag{47}$$

Therefore it is a test of conditional independence [Agresti 1990], assuming (31) to be a set of K two-way tables with testing independence (rather than homogeneity) in each of them. In the following theorem we stress that the Mantel-Haenszel test is inherently connected to the conditional independence (rather than to conditional homogeneity).

Theorem 4.2. *The Mantel-Haenszel test statistic χ^2_{MH} is asymptotically χ^2_1 distributed under the null hypothesis of conditional independence.*

Proof. Assuming the hypothesis of conditional independence, we assume fixed marginal counts corresponding to rows and columns to be fixed. Random variables n_{11k} for each $k = 1, \ldots, K$ follow the same hypergeometric distribution as the number of successes in the situation with selecting $n_{\cdot 1k}$ objects out of $n_{\cdot 1k}$ objects. The probability of success is estimated by $n_{1\cdot k}/n_{\cdot\cdot k}$. Using the formulas for the expectation and variance

$$\mathsf{E} n_{11k} = \frac{n_{1\cdot k} n_{\cdot 1k}}{n_{\cdot\cdot k}}, \quad \mathsf{var}\ n_{11k} = \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot 1k} n_{\cdot 2k}}{n_{\cdot\cdot k}(n_{\cdot\cdot k} - 1)}, \tag{48}$$

we can compute

$$n_{11\cdot} - \mathsf{E} n_{11\cdot} = \sum_{k=1}^{K} \left(n_{11k} - \frac{n_{1\cdot k} n_{\cdot 1k}}{n_{\cdot\cdot k}} \right) = \sum_{k=1}^{K} \frac{n_{11k} n_{\cdot 2k} - n_{12k} n_{\cdot 1k}}{n_{\cdot\cdot k}}. \tag{49}$$

Now it is obtained simply

$$\frac{(n_{11\cdot} - \mathsf{E} n_{11\cdot})^2}{\mathsf{var}\, n_{11\cdot}} = \left[\sum_{k=1}^{K} \frac{n_{11k} n_{\cdot 2k} - n_{12k} n_{\cdot 1k}}{n_{\cdot\cdot k}} \right]^2 \Bigg/ \sum_{k=1}^{K} \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot 1k} n_{\cdot 2k}}{(n_{\cdot\cdot k} - 1) n^2_{\cdot\cdot k}} = \chi^2_{MH}. \tag{50}$$

Here $n_{11\cdot}$ is a sum of independent random variables with hypergeometric distributions (not necessarily equal). For large sample sizes the distribution of $n_{11\cdot}$ can be approximated by normal distribution. The statistic χ^2_{MH} is a square of the normalized random variable $n_{11\cdot}$ and the χ^2_1 distribution under H_0 follows immediately. □

5. EXACT UNCONDITIONAL TEST OF CONDITIONAL HOMOGENEITY

Now we describe an exact unconditional test of conditional homogeneity. Exact unconditional approach to hypothesis testing was studied by Suissa and Shuster [1985], Berger and Boos [1994] and Kalina [2010]. The idea is to replace the nuisance parameter by the least favorable value over a confidence interval for this nuisance parameter. This is such value of the parameter, for which the power function is the maximal. The exact unconditional test exploits the test statistic of an asymptotic (unconditional) test and is based on an enumeration algorithm yielding the exact p-value. We propose the test of conditional homogeneity based on the asymptotic Cochran's T_C test statistic. The computation is computationally feasible also for larger samples. Here we illustrate the method on a numerical data set from a clinical study comparing the effect of two medicines.

The whole table (31) is understood as a set of K two-way tables 2×2 and the probabilities of success under H_0 will be denoted by $\pi_1, \dots, \pi_K$. The joint probability of the outcome is equal to

$$P = \prod_{k=1}^{K} \binom{n_{\cdot 1k}}{n_{11k}} \binom{n_{\cdot 2k}}{n_{12k}} \pi_k^{n_{1\cdot k}} (1-\pi_k)^{n_{2\cdot k}} . \tag{51}$$

Let $\boldsymbol{\pi}$ denote the vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$. In each of the K two-way tables 2×2 let us find the confidence interval for the probability of success on the level $1-\beta/K$. It can be recommended to choose β much smaller than the level of the test α, for example $\alpha = 0.05$ and $\beta = 0.001$ [Berger and Boos 1994]. The confidence intervals for parameters $\pi_1, \dots, \pi_K$ will be denoted by $C_\beta^1, \dots, C_\beta^K$. Then the K-dimensional interval

$$C_\beta = C_\beta^1 \times C_\beta^2 \times \dots \times C_\beta^K \tag{52}$$

covers the true value of $\boldsymbol{\pi}$ with probability at least $1-\beta$.

We use the notation $T_C(a_1, b_1, \dots, a_K, b_K)$ for Cochran's statistic T_C computed for the table

$$\begin{array}{cc||c||cc} a_1 & b_1 & \cdots & a_K & b_K \\ n_{\cdot 11} - a_1 & n_{\cdot 21} - b_1 & \cdots & n_{\cdot 1K} - a_K & n_{\cdot 2K} - b_K \\ \hline n_{\cdot 11} & n_{\cdot 21} & \cdots & n_{\cdot 1K} & n_{\cdot 2K} \end{array} . \tag{53}$$

Let $\mathbb{N}_0$ denote the set of all integers including zero. Let us define the set S as

$$\begin{aligned} S = \big\{ &(a_1, b_1, \dots, a_K, b_K)^T ; a_1 \in \mathbb{N}_0, b_1 \in \mathbb{N}_0, \dots\ ,\ a_K \in \mathbb{N}_0, b_K \in \mathbb{N}_0, \\ &a_1 \le n_{\cdot 11}, b_1 \le n_{\cdot 21}, \dots, a_K \le n_{\cdot 1K}, b_K \le n_{\cdot 2K}, \\ &T_C^2(a_1, b_1, \dots, a_K, b_K) \ge T_C^2(n_{111}, n_{121}, n_{112}, n_{122}, \dots, n_{11K}, n_{12K}) \big\} . \end{aligned} \tag{54}$$

Finally the p-value of the test equals

$$p_\beta = \max_{\boldsymbol{\pi} \in C_\beta} \left\{ \sum \prod_{k=1}^{K} \binom{n_{\cdot 1k}}{a_k} \binom{n_{\cdot 2k}}{b_k} \pi_k^{a_k + b_k} (1-\pi_k)^{n_{\cdot\cdot k} - a_k - b_k} \right\} + \beta, \tag{55}$$

where the sum is computed over all elements $(a_1, b_1, \dots, a_K, b_K)^T \in S$.

Example

As an example let us consider the following study comparing the effect of two medicaments, which leads to the test of conditional homogeneity. There is the total number of K hospitals selected. In the k-th hospital there is a fixed number $n_{\cdot\cdot k}$ of patients selected, who suffer from a particular disease. There is the total number of $n_{\cdot 1k}$ patients selected out of them. These utilize medicine A, while the remaining number of $n_{\cdot 2k}$ patients utilize medicine B. Individual cells of table (31) contain observed counts of healed and non-healed patients.

Table I. Results of the example with the clinical trial. Test of conditional homogeneity in table $2 \times 2 \times 2$.

	Cochran	Mantel-Haenszel
Asymptotic test	$T_C = -2.02$ $p = 0.0433$	$\chi^2_{MH} = 3.27$ $p = 0.0707$
Exact unconditional test over $[0,1] \times [0,1]$	$p = 0.0346$ $(\pi_1, \pi_2)^T = (0; 0.400)^T$	$p = 0.0346$ $(\pi_1, \pi_2)^T = (0; 0.400)^T$
Exact unconditional test over C_β	$p = 0.0344$ $(\pi_1, \pi_2)^T = (0.993; 0.600)^T$	$p = 0.0353$ $(\pi_1, \pi_2)^T = (0.993; 0.600)^T$

We illustrate the test of conditional homogeneity in tables $2 \times 2 \times K$ by a numerical example. We consider a clinical study comparing the effect of two medicines on healing the same disease. The observed data are presented as the $2 \times 2 \times 2$ contingency table

	Hospital 1		Hospital 2	
Patients	Medicine A	Medicine B	Medicine A	Medicine B
Healed	0	3	1	2
Non-healed	1	1	2	0
$\sum$	1	4	3	2

(56)

and the results of the hypotheses tests are summarized in Table I. It contains test statistics and p-values of asymptotic tests and p-values of exact unconditional tests based on statistics T_C and χ^2_{MH}. The vector $(\pi_1, \pi_2)^T$ contains the values of the probability of success in both two-way 2×2 tables, for which the maximal p-value is obtained. Here C_β is a two-dimensional interval $[0.038; 0.993] \times [0.038; 0.993]$. We choose $\alpha = 0.05$ and $\beta = 0.001$. Exact unconditional tests agree that the p-value equals approximately 3.5 %. The results of asymptotic tests cannot be trusted because of small sample sizes, while we trust the results of the exact unconditional test.

REFERENCES

Agresti A. (1990): *Categorical data analysis.* Wiley, New York.

Armitage P. (1955): Tests for linear trend in proportions and frequencies. *Biometrics* **11**, 375 – 386.

Berger R.L, Boos D.D. (1994): P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**, 1012 – 1016.

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

Billingsley P. (1999): *Convergence of probability measures.* Second edition. Wiley, New York.

Čížek P. (2006): Trimmed likelihood-based estimation in binary regression models. *Austrian Journal of Statistics* **35**, No. 2 & 3, 223 – 232.

Cochran W. G. (1954): Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417 – 451.

Day N.E., Byar D.P. (1979): Testing hypotheses in case-control studies. Equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* **35**, 623 – 630.

Freidlin B., Gastwirth J. L. (1999): Unconditional versions of several tests commonly used in the analysis of contingency tables. *Biometrics* **55**, 264 – 267.

Geenens G., Simar L. (2010): Nonparametric tests for conditional independence in two-way contingency tables. *Journal of Multivariate Analysis* **101**, No. 4, 765 – 788.

Halpern J. (1982): Maximally selected chi-square statistics for small samples. *Biometrics* **38**, 1017 – 1023.

Hardin J.W., Hilbe J.M. (2007): *Generalized linear models and extensions.* Stata Press, College Station.

Holinka Š., Smutek D. (2010): Comparison of texture classifier and human observer in diagnosis of autoimmune thyroiditis, observer variability evaluation. *European Journal for Biomedical Informatics* **6**, No. 1.

James B., James K.L., Siegmund D. (1987): Tests for a change-point. *Biometrika* **74**, 71 – 83.

Kalina J. (2010): Logistic and Poisson regression. Analysis of contingency tables. In Kupka K. (Ed.): *Data analysis 2010/II. Statistical methods for technology and research.* Trilobyte Statistical Software, Pardubice, 45 – 56. (In Czech.)

LaMotte L.R. (2005): A note on separated data and exact likelihood-ratio p-values in logistic regression. *Journal of Statistical Computation and Simulation* **75**, No. 8, 667 – 672.

Miller R., Siegmund D. (1982): Maximally selected chi-square statistics. *Biometrics* **38**, 1011 – 1016.

Rao C.R. (1973): *Linear statistical inference and its applications.* Wiley, New York.

Simonoff J.S. (2003): *Analyzing categorical data.* Springer, New York.

Suissa S., Shuster J. (1985): Exact unconditional sample sizes for the 2×2 binomial trials. *Journal of the Royal Statistical Society, Series A* **148**, 317 – 327.

Svozil D., Kalina J., Omelka M., Schneider B. (2008): DNA conformations and their sequence preferences. *Nucleic Acids Research* **36**, No. 11, 3690 – 3706.

Tahata K., Iwashita, T., Tomizawa, S. (2008): Measure of departure from conditional marginal homogeneity for square contingency tables with ordered categories. *Statistics* **42**, No. 5, 453 – 466.

Tang N.-S., Jiang S.-P. (2011): Testing equality of risk ratios in multiple 2×2 tables with structural zero. *Computational Statistics and Data Analysis* **55**, No. 3, 1273 – 1284.

Zvárová J., Malý M. (2003): *Statistical methods in epidemiology.* Karolinum, Prague. (In Czech.)

Jan Kalina
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2
CZ-182 07 Praha 8
Czech Republic
e-mail: kalina@euromise.cz

Received February 2011