RESEARCH REPORT SERIES
(Statistics #2013-03)

On a Comparison of Tests of Homogeneity of Binomial Proportions

Martin Klein
Peter Linton[1]

[1] Graduate Student in the Department of Mathematics and Statistics, University of Maryland, Baltimore County

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: April 8, 2013

On a Comparison of Tests of Homogeneity of Binomial Proportions

Martin Klein
Center for Statistical Research and Methodology
U.S. Census Bureau
Washington, DC, USA
martin.klein@census.gov

Peter Linton
Department of Mathematics and Statistics
University of Maryland, Baltimore County
Baltimore, Maryland 21250

Abstract

There are multiple tests of homogeneity of binomial proportions in the statistics literature. However, when working with sparse data, most test procedures may fail to perform well. In this article we review nine classical and recent testing procedures, including the standard Pearson and likelihood ratio tests; exact conditional and unconditional tests; tests based on moment matching chi-squared approximations; a recently proposed test based on a normal approximation in an asymptotic framework for sparse data; and a recent test based on higher order moment corrections using an Edgeworth approximation. For each test we review its theoretical underpinning, and show how to calculate the P-value. Most of the P-values can be readily calculated in a statistical computing software package such as R. We compare type I error probability and power via simulation. As expected, none of the procedures uniformly outperforms the others in terms of type I error probability and power, but we can make some recommendations based on our empirical results. In particular, we indicate scenarios in which certain otherwise reasonable test procedures can perform inadequately.

Key Words and Phrases: Chi-squared approximation, Edgeworth series, exact test, moment-matching approximation, nuisance parameter, power study, simulation.

AMS 2000 Subject Classifications: 62F03.

Disclaimer. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

We consider the problem of testing the homogeneity hypothesis for k binomial populations of possibly unequal sample sizes based on observing one data point on each of the k populations. That is, our data consist of $X_1, \ldots, X_k$ which are distributed such that

$$X_i \sim \text{Binomial}(n_i, \pi_i), \text{ independently, for } i = 1, \ldots, k, \tag{1}$$

where $n_1, \ldots, n_k$ are known, and $0 \leq \pi_1, \ldots, \pi_k \leq 1$ are unknown. The null hypothesis of homogeneity is to be tested against a general alternative; thus we wish to test

$$H_0 : \pi_1 = \cdots = \pi_k \equiv \pi \quad \text{against} \quad H_1 : \pi_i \neq \pi_j \text{ for some } i \neq j, \tag{2}$$

where the common value π under H_0 is not specified. Throughout, we let α denote the nominal level of the test, and we let $n_+ = \sum_{i=1}^k n_i$ and $X_+ = \sum_{i=1}^k X_i$.

There is an extensive literature on this basic testing problem, especially in the case of $k = 2$ (i.e., a 2×2 contingency table) where, for instance, Upton (1982) evaluated 22 different tests. Starting with the likelihood ratio and Pearson's chi-squared tests (Pearson, 1932; Wilks, 1928), various modifications of them have been suggested (Farrington, 1996; McCullagh, 1985; Paul and Deng, 2013). Some new tests have also been developed, especially in the case of sparse data situations, (Nass, 1959; Potthoff and Whittinghill, 1966; Xu, 2011) since the standard tests can perform poorly in sparse data scenarios (Xu, 2011). Sparse data situations can arise when some of the n_i's are small or when some of the π_i's are close to zero or one.

The scope of applications of the above testing problem is equally vast. The literature of statistical meta-analysis (e.g., Hedges and Olkin, 1985; Hartung, Knapp and Sinha, 2008) dwells upon testing homogeneity of the underlying *effect sizes* with proportion as a very important and useful component. The sparse data scenario arises when dealing with rare outcomes such as a rare disease or death in medical experiments. In a different context, statistical agencies may be interested in creating and releasing synthetic microdata for public use in order to provide useful information to the public while protecting confidentiality of respondents. In a synthetic data set, some or all of the original data values are replaced by random draws from an appropriate distribution for the purpose of statistical disclosure control; we refer to Drechsler (2011), Raghunathan, Reiter, and Rubin (2003), Reiter (2003), and Rubin (1993) for details. Generation of synthetic count data in cross-classified contingency tables can be based on an ANOVA type log-linear model for cell probabilities along with a multinomial assumption for the joint distribution of the cell counts. Bhapkar and Koch (1968) and Bishop, Fienberg, and Holland (1977) discuss many aspects of model selections in this context, including choice and interpretation of interaction terms in these models. While a fully saturated log-linear model provides little flexibility, under

the *independence* model, interaction terms are set to zero. This precisely corresponds to the homogeneity of associated cell probabilities across rows or columns in the contingency table, and suggests testing of homogeneity of proportions across rows or columns before using the independence model. The simplest case of homogeneity of binomial proportions arises when testing for absence of interaction in a $k \times 2$ dimensional contingency table.

The primary focus of this article is to provide a comprehensive comparison by simulation among the available tests in terms of type I error probability and power in sparse data settings. The outline of the rest of the article is as follows. We review several test procedures in Section 2. In Section 3 we compare the procedures empirically based on type I error probability and power, and offer some guidance based on these results. We provide some concluding remarks in Section 4. The Appendix contains tables that summarize the simulation results.

2 Test Procedures

In this section we review several procedures for testing the hypotheses (2). We first note that the likelihood function for $(\pi_1, \ldots, \pi_k)$ under the model (1) is

$$L(\pi_1, \ldots, \pi_k; X_1, \ldots, X_k) = \prod_{i=1}^{k} \binom{n_i}{X_i} \pi_i^{X_i} (1-\pi_i)^{n_i - X_i}, \ 0 \le \pi_i \le 1, \tag{3}$$

and the likelihood function for π under the restriction of the parameter space defined by the null hypothesis H_0 is

$$L_0(\pi; X_1, \ldots, X_k) = \left[\prod_{i=1}^{k} \binom{n_i}{X_i} \right] \pi^{X_+} (1-\pi)^{n_+ - X_+}, \ 0 \le \pi \le 1. \tag{4}$$

Thus, under model (1), the maximum likelihood estimator of π_i is $\hat{\pi}_i = X_i/n_i$ for $i = 1, \ldots, k$, and under the null hypothesis H_0, the maximum likelihood estimator of π is $\hat{\pi} = X_+/n_+$.

2.1 Standard chi-squared and likelihood ratio tests

Pearson's chi-squared test (Pearson, 1932) and the usual likelihood ratio test (Wilks, 1928) are two standard tests of the hypotheses (2); the test statistics are defined, respectively, by

$$\begin{aligned} T_P \equiv T_P(X_1, \ldots, X_k) &= \sum_{i=1}^{k} \left\{ \frac{(n_i\hat{\pi}_i - n_i\hat{\pi})^2}{n_i\hat{\pi}} + \frac{[n_i(1-\hat{\pi}_i) - n_i(1-\hat{\pi})]^2}{n_i(1-\hat{\pi})} \right\} \\ &= \sum_{i=1}^{k} \frac{n_i(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}(1-\hat{\pi})}, \end{aligned} \tag{5}$$

$$T_L \equiv T_L(X_1, \ldots, X_k) = -2 \log \left[\frac{L_0(\hat{\pi}; X_1, \ldots, X_k)}{L(\hat{\pi}_1, \ldots, \hat{\pi}_k; X_1, \ldots, X_k)} \right]$$
$$= 2 \sum_{i=1}^{k} X_i \log \left(\frac{\hat{\pi}_i}{\hat{\pi}} \right) + 2 \sum_{i=1}^{k} (n_i - X_i) \log \left(\frac{1 - \hat{\pi}_i}{1 - \hat{\pi}} \right). \qquad (6)$$

Pearson's chi-squared test rejects H_0 if T_P is large and the likelihood ratio test rejects H_0 if T_L is large. Under the null hypothesis, for fixed k and large n_i, the asymptotic distributions of both T_P and T_L are chi-squared on $k-1$ degrees of freedom, and their difference converges in probability to zero (Agresti, 2002). Therefore, letting t_P and t_L denote the observed values of T_P and T_L, respectively, the P-values for Pearson's chi-squared test and the likelihood ratio test are $\Pr\{\chi^2_{k-1} \geq t_P\}$ and $\Pr\{\chi^2_{k-1} \geq t_L\}$, respectively. These tests are justified by an asymptotic theory in which k is fixed and the n_i's are large; therefore they may not perform well in sparse data settings with small cell counts (Haberman, 1988; Mielke, Berry, and Johnston, 2004; Xu, 2011). In fact, when data are sparse, we may not be able to compute T_L as defined in (6) since it is likely that at least one $\hat{\pi}_i$ will equal zero or one.

Below we review several alternative test procedures that have appeared in the literature.

2.2 Exact tests

In the sparse data setting where the asymptotic framework of the standard chi-squared and likelihood ratio tests generally does not hold, exact tests provide a natural alternative. These tests are referred to as *exact* because they use an exact finite sample distribution of the test statistic as opposed to an approximation. We refer to Agresti (1992, 2001, 2002) for an in-depth presentation and discussion of exact methods of inference. Here we will describe a *conditional* and an *unconditional* exact procedure for testing (2) using the Pearson statistic T_P defined in (5). Throughout we let t_P denote the observed value of T_P.

Conditional test. Under the null hypothesis, it readily follows from (4) that X_+ is a sufficient statistic for π and hence the conditional distribution of $X_1, \ldots, X_k$ given X_+ is free of π. The probability mass function of this conditional distribution is given by (Agresti, 2002)

$$p_c(x_1, \ldots, x_k \,|\, X_+) = \frac{(\prod_{i=1}^{k} n_i!)(X_+)!(n_+ - X_+)!}{(n_+)! \prod_{i=1}^{k} [x_i!(n_i - x_i)!]}, \text{ for } (x_1, \ldots, x_k) \in \mathcal{A}_{X_+}, \qquad (7)$$

where $\mathcal{A}_{X_+} = \{(a_1, \ldots, a_k) \in \mathcal{B} : \sum_{j=1}^{k} a_j = X_+\}$, $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_k$, and $\mathcal{B}_i = \{0, 1, \ldots, n_i\}$, $i = 1, \ldots, k$. Then an exact test can be obtained using the P-value defined by $\Pr\{T_P \geq t_P \,|\, X_+\}$ where this probability is computed with respect to the conditional probability distribution (7), i.e.,

$$\Pr\{T_P \geq t_P \,|\, X_+\} = \sum_{\{(x_1, \ldots, x_k) \in \mathcal{A}_{X_+} : T_P(x_1, \ldots, x_k) \geq t_P\}} p_c(x_1, \ldots, x_k \,|\, X_+). \qquad (8)$$

When $k = 2$, this procedure is a two-sided version of *Fisher's exact test*. Unless k and the n_i's are small, the probability (8) is difficult to compute due to the large number of atoms in the sample space $\mathcal{A}_{X_+}$, but we can approximate it via Monte Carlo (Robert and Casella, 2004). For instance, in R (R Development Core Team, 2011) there is a function called *r2dtable* which generates a random sample from the distribution (7) using the algorithm of Patefield (1981). If $(x_1^{(1)}, \ldots, x_k^{(1)}), \ldots, (x_1^{(m)}, \ldots, x_k^{(m)})$ denote m random vectors drawn independently from the distribution (7), then a Monte Carlo estimator of (8) is

$$\widehat{\Pr}\{T_P \geq t_P \,|\, X_+\} = \frac{1}{m} \sum_{j=1}^{m} I[T_P(x_1^{(j)}, \ldots, x_k^{(j)}) \geq t_P] \tag{9}$$

where $I[A]$ is the indicator of the event A. Mehta, Patel, and Senchaudhuri (1988) show that importance sampling can be used to obtain an improved Monte Carlo estimator.

Unconditional test. Under the null hypothesis, the joint distribution of $(X_1, \ldots, X_k)$ is given by

$$p_u(x_1, \ldots, x_k \,|\, \pi) = \left[\prod_{i=1}^{k} \binom{n_i}{x_i} \right] \pi^{\sum_1^k x_i} (1-\pi)^{n_+ - \sum_1^k x_i}, \; 0 \leq \pi \leq 1, \tag{10}$$

which depends on the unknown parameter π. Using T_P as the test statistic, an exact unconditional test can be obtained by defining the P-value as $\sup_{\pi \in [0,1]} \Pr\{T_p \geq t_p\}$ where the probability is computed with respect to the distribution (10) and thus depends on π, i.e.,

$$\sup_{\pi \in [0,1]} \Pr\{T_p \geq t_p\} = \sup_{\pi \in [0,1]} \sum_{\{(x_1, \ldots, x_k) \in \mathcal{B}:\, T_P(x_1, \ldots, x_k) \geq t_P\}} p_u(x_1, \ldots, x_k \,|\, \pi). \tag{11}$$

This test was introduced by Barnard (1945, 1947) in the case of $k = 2$ (later, Barnard [1949] wrote in favor of Fisher's exact test over his unconditional test). When either k or the n_i's are large, the above probability is difficult to compute due to the large number of atoms in $\mathcal{B}$, and hence the P-value is extremely difficult to compute due to the presence of $\sup_{\pi \in [0,1]}$.

Remark. The question of whether one should use the conditional or unconditional test is controversial. There is much debate on this issue in the statistical literature, especially for the case of $k = 2$, and we refer to Little (1989) for a discussion. The debate generally involves issues of statistical philosophy that go beyond power and type I error probability comparisons. But in terms of empirical comparisons, Suissa and Shuster (1985) compared the type I error probability and power of the conditional and unconditional tests when $k = 2$, using the statistic $Z_u = n^{1/2}(\hat{\pi}_2 - \hat{\pi}_1)/[\hat{\pi}_2(1-\hat{\pi}_2) + \hat{\pi}_1(1-\hat{\pi}_1)]^{1/2}$ for the unconditional test, and assumed $n = n_1 = n_2$ and a one-sided alternative hypothesis. Under these conditions, Suissa and Shuster (1985) derived a method for computing the unconditional P-value and found the unconditional test to be more powerful than Fisher's exact test. Mehta and Hilton (1993) considered the case of $k = 3$, and compared the conditional and unconditional tests

based on the Pearson statistic T_P when $n = n_1 = n_2 = n_3$. They concluded that while the unconditional test appeared to hold a power advantage over the conditional test when $k = 2$, when $k = 3$ the power advantage of the unconditional test rapidly diminishes for moderately large values of the common sample size n.

2.3 Test of Nass (1959)

Nass (1959) considered an adjustment to the standard chi-squared test to improve the approximation for sparse data. This approach approximates the conditional distribution of T_P given X_+, under the null hypothesis. Under the null hypothesis, the distribution of $c \times T_P$ is approximated by a chi-squared density on v degrees of freedom, where c and v are chosen so that the conditional mean and variance of $c \times T_P$ match the mean and variance of the approximating chi-squared distribution. In our setting of independent binomial sampling, the approximation is defined by $c \times T_P \,|\, X_+ \sim \chi^2_v$, where c and v are determined such that

$$E(c \times T_P \,|\, X_+) = v, \;\; \mathrm{Var}(c \times T_P \,|\, X_+) = 2v,$$

or equivalently,

$$c = 2E(T_P \,|\, X_+)/\mathrm{Var}(T_P \,|\, X_+), \;\; v = cE(T_P \,|\, X_+). \tag{12}$$

Under the null hypothesis, the mean and variance of T_P, conditional on X_+, were derived by Haldane (1940) and simplified by Dawson (1954) into the following form:

$$E(T_P \,|\, X_+) = \frac{(k-1)n_+}{n_+ - 1},$$
$$\mathrm{Var}(T_P \,|\, X_+) = \frac{2n_+}{n_+ - 3}(\rho - \sigma)(\mu - \tau) + \frac{n_+^2}{n_+ - 1}\sigma\tau,$$

where

$$\rho = \frac{n_+ - 2}{n_+ - 1}, \; \mu = \frac{(k-1)(n_+ - k)}{n_+ - 1}, \; \sigma = \frac{\frac{n_+}{X_+} + \frac{n_+}{n_+ - X_+} - 4}{n_+ - 2}, \; \tau = \frac{n_+ \sum_{i=1}^k n_i^{-1} - k^2}{n_+ - 2}.$$

Thus the P-value of the test is computed as $\Pr\{\chi^2_v \geq c\, t_p\}$ with c and v defined by (12).

2.4 Test of Potthoff and Whittinghill (1966)

Potthoff and Whittinghill (1966) derived a test procedure based on the following argument. First, they suppose that π is known, and the alternative hypothesis specifies the distribution of $X_1, \ldots, X_k$ such that

$$\pi_1, \ldots, \pi_k \sim iid \sim \mathrm{Beta}[\pi a, (1-\pi)a],$$

and conditionally on $(\pi_1, \ldots, \pi_k)$, the variables $X_1, \ldots, X_k$ are distributed as in (1). Then a locally most powerful test (locally in the sense that a is such that the variance of the Beta$[\pi a, (1-\pi)a]$ distribution is small) rejects the null hypothesis for large values of the statistic

$$\mathbb{V}(\pi) = \frac{\sum_{i=1}^n X_i(X_i-1)}{\pi} + \frac{\sum_{i=1}^n Y_i(Y_i-1)}{1-\pi},$$

where $Y_i = n_i - X_i$. The distribution of $\mathbb{V}(\pi)$ under the null hypothesis is approximated as $e(\pi)\mathbb{V}(\pi) + f(\pi) \sim \chi^2_{v(\pi)}$, and the constants $e(\pi)$, $f(\pi)$ and $v(\pi)$ are chosen such that

$$\begin{aligned} v(\pi) &= E\left\{e(\pi)\mathbb{V}(\pi) + f(\pi)\right\}, \\ 2v(\pi) &= \mathrm{Var}\left\{e(\pi)\mathbb{V}(\pi) + f(\pi)\right\}, \\ 8v(\pi) &= E\left\{e(\pi)\mathbb{V}(\pi) + f(\pi) - E[e(\pi)\mathbb{V}(\pi) + f(\pi)]\right\}^3, \end{aligned}$$

i.e., the mean, variance, and third central moment of $e\mathbb{V}(\pi) + f$ match those of the approximating χ^2_v distribution. This results in

$$\begin{aligned} e(\pi) &= \frac{N}{2\gamma(\pi)N + \sum_{i=1}^n n_i(n_i-1)(n_i-2)}, \\ f(\pi) &= e(\pi)(e(\pi)-1)N, \\ N &= \sum_{i=1}^n n_i(n_i-1), \\ \gamma(\pi) &= \frac{1}{4\pi(1-\pi)} - 1, \\ v(\pi) &= e(\pi)^2 N. \end{aligned} \tag{13}$$

To handle the case of unknown π, Potthoff and Whittinghill (1966) obtained a test by setting π equal to the value which minimizes $\mathbb{V}(\pi)$; the resulting values of π and $\mathbb{V}(\pi)$ are

$$\begin{aligned} \pi_{\min} &= \frac{\left[\sum_{i=1}^n X_i(X_i-1)\right]^{1/2}}{\left[\sum_{i=1}^n X_i(X_i-1)\right]^{1/2} + \left[\sum_{i=1}^n Y_i(Y_i-1)\right]^{1/2}}, \\ \mathbb{V}(\pi_{\min}) &= \left\{ \left[\sum_{i=1}^n X_i(X_i-1)\right]^{1/2} + \left[\sum_{i=1}^n Y_i(Y_i-1)\right]^{1/2} \right\}^2. \end{aligned}$$

The P-value of the test is thus computed as $\Pr\{\chi^2_{v(\pi_{\min})} \geq e(\pi_{\min})\mathbb{V}(\pi_{\min}) + f(\pi_{\min})\}$, where of course, $e(\pi_{\min})\mathbb{V}(\pi_{\min}) + f(\pi_{\min})$ and $v(\pi_{\min})$ are fixed at their observed values in the probability computation.

2.5 Adjustment to the Potthoff and Whittinghill (1966) test using the method of Berger and Boos (1994)

In the context of testing a general composite hypothesis of the form $H_0 : \psi = \psi_0$ in the presence of nuisance parameters $\boldsymbol{\theta}$, Berger and Boos (1994) suggested the following approach based on maximization of a suitable P-value. Let $P(\boldsymbol{\theta})$ be a P-value based on some test of H_0 for a specified value of $\boldsymbol{\theta}$, and let C_β be a $(1-\beta)$ level confidence set for $\boldsymbol{\theta}$ under H_0. It is demonstrated in Berger and Boos (1994) that the test which rejects H_0 for small values of

$$P_\beta = \sup_{\boldsymbol{\theta} \in C_\beta} P(\boldsymbol{\theta}) + \beta \tag{14}$$

provides a valid test of H_0. An application of this general procedure to the Potthoff and Whittinghill (1966) test can be formulated as follows. Under the null hypothesis of homogeneity of binomial proportions, the common proportion π is a nuisance parameter, and its $(1-\beta)$ level large sample confidence interval based on the entire data can be computed from the fact that under H_0, $\Pr\{-z_{\beta/2} < \frac{\sqrt{n_+}(\hat{\pi}-\pi)}{\sqrt{\pi(1-\pi)}} < z_{\beta/2}\} \approx 1-\beta$ for large n_+, where $z_{\beta/2} = \Phi^{-1}(1-\beta/2)$ and $\Phi(s) = \int_{-\infty}^{s} e^{-u^2/2} du/\sqrt{2\pi}$. Solving these inequalities for π readily yields the approximate $(1-\beta)$ level confidence interval $C_\beta = \{\pi : \hat{\pi}_L \le \pi \le \hat{\pi}_U\}$ where

$$\hat{\pi}_L = \frac{\hat{\pi} + \frac{z^2_{\beta/2}}{2n_+} - z_{\beta/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_+} + \frac{z^2_{\beta/2}}{4n^2_+}}}{1 + \frac{z^2_{\beta/2}}{n_+}}, \quad \hat{\pi}_U = \frac{\hat{\pi} + \frac{z^2_{\beta/2}}{2n_+} + z_{\beta/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_+} + \frac{z^2_{\beta/2}}{4n^2_+}}}{1 + \frac{z^2_{\beta/2}}{n_+}}.$$

We note that several other confidence intervals are also available for the binomial proportion; C_β as defined above is one of the intervals recommended by Brown, Cai, and DasGupta (2001).

For a given π, the P-value for the Potthoff and Whittinghill (1966) test described in Section 2.4 is

$$P(\pi) = \Pr\left\{\chi^2_{v(\pi)} > e(\pi)\mathbb{V}(\pi) + f(\pi)\right\} \approx \Pr\left\{N(0,1) > \frac{e(\pi)\mathbb{V}(\pi) + f(\pi) - v(\pi)}{[2v(\pi)]^{1/2}}\right\}$$

where the quantities $e(\pi)$, $f(\pi)$ and $v(\pi)$ are defined in (13). The supremum value P_β displayed in (14) is now obtained by computing $P(\pi)$ with the minimum value of $Q(\pi) = \frac{e(\pi)\mathbb{V}(\pi)+f(\pi)-v(\pi)}{[2v(\pi)]^{1/2}}$ with respect to $\pi \in C_\beta$ and adding β. It is easy to verify that $Q(\pi)$ simplifies to $Q(\pi) = \frac{\mathbb{V}(\pi)-N}{[2N]^{1/2}}$, and we compute its minimum subject to $\pi \in C_\beta$. Denoting this minimum value by Q^*_β and following (14), it follows that the test based on the method of Berger and Boos (1994) rejects H_0 for small values of

$$P_\beta = \Pr\left\{N(0,1) > Q^*_\beta\right\} + \beta.$$

Obviously we choose only small values of β in applications.

2.6 Test of Xu (2011)

Xu (2011) proposed an unconditional test which was motivated as follows. First note that

$$E[\hat{\pi}_i - \hat{\pi}]^2 = \frac{n_+ - n_i}{n_i n_+}\pi(1-\pi), \text{ under } H_0,$$

and

$$E\left[\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i - 1}\right] = \frac{\pi_i(1-\pi_i)}{n_i}.$$

Thus defining

$$U_i = (\hat{\pi}_i - \hat{\pi})^2 - \left[\frac{n_+ - n_i}{n_+}\right]\left[\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i - 1}\right], \; i = 1, \ldots, k,$$

it follows that $E(U_i) = 0$ under H_0. Xu (2011) shows that

$$\begin{aligned}\mathrm{Var}(U_i) &= \frac{2\pi^2(1-\pi)^2}{n_i(n_i-1)} + O\left(\frac{1}{n_+}\right), \\ \mathrm{Cov}(U_i, U_j) &= O\left(\frac{1}{n_+}\right) \text{ for } i \neq j,\end{aligned}$$

and thus proposes the test statistic

$$T_D = \frac{\sqrt{k}}{\sqrt{2}}\frac{\overline{V}}{\hat{\pi}(1-\hat{\pi})}$$

where $V_i = \sqrt{n_i(n_i-1)}U_i$ and $\overline{V} = k^{-1}\sum_{i=1}^k V_i$. Letting t_D denote the observed value of T_D, Xu (2011) proposes to compute the P-value of the test as $\Pr\{N(0,1) \geq t_D\}$. The P-value computation is based on a normal approximation under a sparse asymptotic framework in which $k \to \infty$ while the n_i are bounded.

2.7 Test of Paul and Deng (2013)

Extending work of Farrington (1996), Paul and Deng (2013) present a general method for testing goodness of fit of a generalized linear model to sparse data. As discussed by Paul and Deng (2013), their method, which is based on an Edgeworth approximation of the distribution of the modified Pearson χ^2 statistic conditional on $\hat{\pi}$, can be used to test the hypotheses (2). The test statistic is the standardized quantity

$$Z = \frac{X_*^2 - E(X_*^2 \,|\, \hat{\pi})}{[\mathrm{Var}(X_*^2 \,|\, \hat{\pi})]^{1/2}},$$

where X_*^2 is the modified Pearson statistic proposed by Farrington (1996). In our setting, the modified Pearson statistic takes the form:

$$X_*^2 = X^2 - \frac{1-2\hat{\pi}}{\hat{\pi}(1-\hat{\pi})}\sum_{i=1}^{k}(\hat{\pi}_i - \hat{\pi}).$$

The conditional distribution of the test statistic Z under the null hypothesis is approximated by the Edgeworth series:

$$\begin{aligned}\Pr\{Z \geq z \,|\, \hat{\pi}\} \approx 1 - \Phi(z) + \phi(z)\Big\{&(z^2-1)\rho_3(X_*^2 \,|\, \hat{\pi})/6 + (z^3 - 3z)\rho_4(X_*^2 \,|\, \hat{\pi})/24 \\ &+ (z^5 - 10z^3 + 15z)\rho_3^2(X_*^2 \,|\, \hat{\pi})/72\Big\}. \end{aligned} \tag{15}$$

As usual, $\phi(s) = e^{-s^2/2}/\sqrt{2\pi}$ and $\Phi(s) = \int_{-\infty}^{s}\phi(u)du$ denote the standard normal probability density and cumulative distribution functions, respectively. The Edgeworth series involves $\rho_3(X_*^2 \,|\, \hat{\pi})$ and $\rho_4(X_*^2 \,|\, \hat{\pi})$ which are the standardized conditional third and fourth cumulants of X_*^2, respectively, under the null hypothesis. That is, if $\kappa_j(X_*^2 \,|\, \hat{\pi})$ denotes the jth cumulant of X_*^2 conditional on $\hat{\pi}$, then

$$\rho_j(X_*^2 \,|\, \hat{\pi}) = \frac{\kappa_j(X_*^2 \,|\, \hat{\pi})}{[\kappa_2(X_*^2 \,|\, \hat{\pi})]^{j/2}},\ j = 3, 4, \ldots.$$

Paul and Deng (2013) provide the following approximate expressions for the first four conditional cumulants of X_*^2 under the null hypothesis:

$$\begin{aligned}
\kappa_1(X_*^2 \,|\, \hat{\pi}) &= k\left(1 - \frac{1}{k} + \frac{1}{n_+}\right), \\
\kappa_2(X_*^2 \,|\, \hat{\pi}) &= 2(k-1)\left(1 - \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\right), \\
\kappa_3(X_*^2 \,|\, \hat{\pi}) &= 8(k-1)\left[1 - \frac{1}{k}\sum_{i=1}^{k}\frac{5n_i - 4}{n_i^2} + \frac{1}{2k\hat{\pi}(1-\hat{\pi})}\sum_{i=1}^{k}\frac{n_i - 1}{n_i^2}\right], \\
\kappa_4(X_*^2 \,|\, \hat{\pi}) &= 48(k-1)\sum_{i=1}^{k}\left(1 - \frac{1}{n_i}\right)\left[\frac{n_i^2 - 17n_i + 31}{n_i^2} + \frac{3n_i - 7}{n_i^2\hat{\pi}(1-\hat{\pi})} + \frac{1}{6n_i^2\hat{\pi}^2(1-\hat{\pi}^2)}\right] \\
&\quad + \frac{12k^2(1-2\hat{\pi})^2\left(1 - \frac{2}{k}\sum_{i=1}^{k}\frac{1}{n_i}\right)^2}{\hat{\pi}(1-\hat{\pi})n_+}.
\end{aligned}$$

Finally, with z denoting the observed value of the test statistic Z, the P-value is computed by evaluating the right hand side of (15) using the expressions above for the cumulants. This test is designed under a sparse asymptotic framework as mentioned at the end of Section 2.6. The higher order corrections are designed to improve the approximation for moderate values of k.

3 Empirical Comparison of the Tests

In Section 2 we introduced nine different tests for the hypotheses (2). In this section we use simulation to evaluate and compare seven of these tests based on type I error probability and power. We conducted all simulations using the statistical computing software R (R Development Core Team, 2011). Below we list the seven tests included in the comparison, and we give a short name for each test. The short name is used to refer to each test in the following discussion, and in Tables 1 - 12 which summarize the simulation results.

1. Pearson: Pearson's chi-squared test presented in Section 2.1.
2. ExactC: Exact conditional test presented in Section 2.2.
3. Nass: Test of Nass (1959) presented in Section 2.3.
4. PW: Test of Potthoff and Whittinghill (1966) presented in Section 2.4.
5. PWBB: Test of Potthoff and Whittinghill (1966) adjusted using the method of Berger and Boos (1994) as presented in Section 2.5.
6. Xu: Test of Xu (2011) presented in Section 2.6.
7. PD: Test of Paul and Deng (2013) presented in Section 2.7.

We have not included the likelihood ratio test in the comparison because we cannot compute the likelihood ratio test statistic T_L as defined in (6) if at least one $\hat{\pi}_i$ equals zero or one (which is likely in a sparse data situation). We have not included the exact unconditional test due to the computational difficulty in calculating the P-value (11) as discussed in Section 2.2.

In our empirical study we consider twelve different settings where each is of the form:

$$\pi_i = \pi_0, \text{ for } i = 1, \ldots, k-1,$$
$$\pi_k = \pi_0 + \delta.$$

The values of π_0, k, and $n_1, \ldots, n_k$ used in the twelve simulation settings are as follows.

Setting 1. $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = n_3 = 10$

Setting 2. $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = 10$, $n_3 = 60$

Setting 3. $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = 60$, $n_3 = 10$

Setting 4. $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = n_3 = 60$

Setting 5. $\pi_0 = 0.001$, $k = 8$, $n_i = 20 \times 2^{i-1}$

Setting 6. $\pi_0 = 0.001$, $k = 8$, $n_i = 20 \times 2^{8-i}$

Setting 7. $\pi_0 = 0.001$, $k = 8$, $n_1 = \cdots = n_8 = 2560$

Setting 8. $\pi_0 = 0.016$, $k = 8$, $n_i = 20 \times 2^{i-1}$

Setting 9. $\pi_0 = 0.001$, $k = 40$, $n_i = 20 \times 2^{(i-1) \bmod 8}$

Setting 10. $\pi_0 = 0.001$, $k = 40$, $n_i = 20 \times 2^{(40-i) \bmod 8}$

Setting 11. $\pi_0 = 0.001$, $k = 40$, $n_1 = \cdots = n_{40} = 2560$

Setting 12. $\pi_0 = 0.008$, $k = 40$, $n_i = 20 \times 2^{(i-1) \bmod 8}$

For each of the seven tests, we use Monte Carlo simulation to estimate the probability of rejecting the null hypothesis in each of the settings 1 - 12. These probabilities are computed for several values of δ; and we report the probabilities for settings 1 - 12 in Tables 1 - 12, respectively. In all cases, the nominal level of the test is taken as $\alpha = 0.05$, and 10000 iterations are used to obtain the Monte Carlo estimates of type I error probability and power. For the ExactC test, in settings 1 and 2 we compute the P-value directly using (8). In the remaining settings 3 - 12, direct computation of the P-value as defined in (8) is difficult due to larger values of k and/or the n_i's; therefore, in these settings, we obtain the P-value using the Monte Carlo estimator (9) with $m = 10000$ iterations. Notice that in each table, the first row, which corresponds to $\delta = 0$, gives the probability of type I error for a particular value of the common π. As we move down the rows of each table δ increases, and hence we would expect the power to increase because in this way we move further from the null hypothesis of homogeneity. In any particular iteration of the simulation, if the observed value of X_+ equals 0 or n_+, then we do not reject the null hypothesis under any test. Below is a summary of the findings of the simulation study.

1. In setting 1 where we have sparse data and a very small k, we see that all seven tests have type I error probability well below $\alpha = 0.05$. While none of the tests perform ideally in setting 1, we notice here that the probability of type I error and power of the three tests ExactC, PW, and PWBB is similar and well below that of the other four tests. In this setting, the performance of the four tests Pearson, Nass, Xu, and PD is comparable.

2. As expected, we find that the Pearson test exhibits inadequate performance in some sparse data settings where some of the other tests do perform adequately. Specifically, in settings 5, 6, 9, 10, and 12, the Pearson test has probability of type I error in the range of (0.09, 0.20), and hence is well above the nominal level of $\alpha = 0.05$.

3. In all settings, we find that the PW and PWBB tests have probability of type I error below the nominal level $\alpha = 0.05$. In fact, in most of the settings, the probability of type I error is substantially below the nominal level; hence the test seems to be extremely conservative. Furthermore, in settings 2, 3, 4, 6, and 10, PW and PWBB tend to have

extremely low power in comparison to several of the other tests. The test PWBB offers only a slight improvement over PW; generally PW and PWBB perform quite similarly. In spite of these drawbacks, it is interesting that in settings 9 and 12 the tests PW and PWBB have lower type I error probability and tend to have higher power than the other tests.

4. In settings 5 and 6, the PD test exhibits inadequate performance because the type I error probability is 0.661 and 0.6537, respectively, which obviously is substantially above the nominal level of $\alpha = 0.05$. In the settings 2, 3, 9, and 10, we find that PD has type I error probability of 0.1432, 0.0860, 0.0983, 0.09073, and each is well above the nominal level. It is interesting to notice that each setting where we have noticed an increased type I error probability for the PD test is one with unequal n_i's. On the other hand, the PD test performs adequately in settings 7, 8, 11, and 12; settings 8 and 12 also have unequal n_i's.

5. We note that theoretically, the ExactC test is similar to the Nass test. The difference between the two tests, as presented in Sections 2.2 and 2.3, is that the ExactC test computes the P-value directly using the conditional distribution (7), while the Nass test uses (12) to obtain a scaled chi-squared distribution that approximates (7), and then computes the P-value with respect to this approximating distribution. Thus, as one would expect, in each of our simulation settings we find that the ExactC and Nass tests generally yield similar performance and both maintain their level at or below the nominal $\alpha = 0.05$. In settings 1 - 4, we find that the probability of type I error is slightly closer to $\alpha = 0.05$ under the Nass test in comparison with the ExactC test; and also in these settings the power of the Nass test is slightly higher than that of the ExactC test. Notice that in settings 1 - 4, k is quite small, and hence the discrete distribution (7) has a small sample space which makes it difficult for the test to achieve size $\alpha = 0.05$. On the other hand, in setting 5 the ExactC test performs slightly better than the Nass test in terms of both type I error probability and power. In the remaining settings 6 - 12, the performance of the ExactC and Nass tests is nearly identical, indicating that Nass's (1959) chi-squared approximation of (7) works well in these settings. Generally the performance of the ExactC and Nass tests is adequate in the settings we considered.

6. We find that the Xu test performs adequately in each of the simulation settings in terms of both type I error probability and power, though the test tends to be conservative in some settings (e.g., settings 1 - 6). Even in these settings, the power of the Xu test still tends to compare favorably (though it is not always the best) with several of the other test procedures.

7. Since the ExactC, Nass, and Xu tests each tend to perform adequately in the chosen simulation settings, a comparison of the three tests seems appropriate. We have already noted that the ExactC and Nass tests are similar to each other. The choice between the Xu test versus either ExactC or Nass is not so clear, as none dominates the other in power or type I error probability. For instance, in settings 2, 5, 8, 9, and 12, the Xu test has a power advantage over both the ExactC and Nass tests; in settings 3, 6, and 10, the ExactC and

Nass tests have a power advantage over the Xu test; and in settings 4, 7, and 11, the ExactC, Nass, and Xu tests all perform similarly. To complement these findings, it is interesting to observe that in settings 2, 5, 8, 9, and 12, one of the populations with the largest n_i differs from the rest; in settings 3, 6, and 10, one of the populations with the smallest n_i differs from the rest; and in settings 4, 7, and 11, all n_i's are equal. These observations may help to provide some guidance as to when the Xu test is preferable to the ExactC and Nass tests or vice versa.

4 Concluding Remarks

In this article we have reviewed nine procedures for testing the hypothesis of homogeneity of k binomial proportions. In Section 2 we presented the justification for each test, and showed how to calculate each P-value. In Section 3 we used simulation to assess and compare seven of these tests on the basis of type I error probability and power. Through the simulation studies, we located sparse data scenarios in which the otherwise reasonable tests Pearson, PW, PWBB, and PD, performed inadequately. We noted that the ExactC, Nass, and Xu tests exhibited adequate performance in all simulation settings that we considered, and we provided some guidance regarding the choice between these three tests in sparse data situations. As expected, we found the ExactC and Nass tests to be generally similar, but some distinctions emerged between these two tests and the Xu test.

Acknowledgments

We thank Professor Bimal Sinha for encouragement and for several enlightening discussions on this topic. We note that Xu (2011) refers to Dihua Xu's doctoral dissertation which was completed under the supervision of Professor Sinha.

References

[1] Agresti, A. (1992). A Survey of Exact Inference for Contingency Tables. *Statistical Science*, **7**, 131-153.

[2] Agresti, A. (2001). Exact inference for categorical data: Recent advances and continuing controversies. *Statistics in Medicine*, **20**, 2709-2722.

[3] Agresti, A. (2002). *Categorical Data Analysis*, Wiley.

[4] Barnard, G.A. (1945). A new test for 2×2 tables. *Nature*, **156**, 177-177.

[5] Barnard, G.A. (1947). Significance tests for 2×2 tables. *Biometrika*, **34**, 123-138.

[6] Barnard, G.A. (1949). Statistical inference. *Journal of the Royal Statistical Society, Series B*, **11**, 115-149.

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

[8] Bhapker, V.P., and Koch G.G. (1968). Hypotheses of 'No Interaction' in Multidimensional Contingency Tables. *Technometrics*, **10**, 107-123.

[9] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1977). *Discrete Multivariate Analysis*, The MIT Press.

[10] Brown, L.D., Cai, T.T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101-117.

[11] Dawson, R.B. (1954). A simplified expression for the variance of the χ^2-function on a contingency table. *Biometrika*, **41**, 280.

[12] Drechsler, J. (2011), *Synthetic Datasets for Statistical Disclosure Control*, Springer.

[13] Farrington, C.P. (1996). On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society. Series B*, **58**, 349-360.

[14] Haberman, S.J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, **83**, 555-560.

[15] Haldane, J.B.S. (1940). The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small, *Biometrika*, **31**, 346-355.

[16] Hartung, J., Knapp, G., and Sinha, B.K. (2008). *Statistical Meta-Analysis with Applications*, Wiley.

[17] Hedges, L.V., and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press.

[18] Klein, M.D., and Creecy, R.H. (2010). Steps toward creating a fully synthetic decennial census microdata file, *Proceedings of the Joint Statistical Meetings*.

[19] Little, R.J.A. (1989). Testing the equality of two independent binomial proportions, *The American Statistician*, **43**, 283-288.

[20] Mielke, P.W., Berry, K.J., and Johnston, J.E. (2004). Asymptotic log-linear analysis: Some Cautions concerning sparse frequency tables, *Psychological Reports*, **94**, 19-32.

[21] Mehta, C.R., Patel, N.R., and Senchaudhuri, P. (1988). Importance sampling for estimating exact probabilities in permutational inference, *Journal of the American Statistical Association*, **83**, 999-1005.

[22] Mehta, C.R., and Hilton, J.F. (1993). Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. *The American Statistician*, **47**, 91-98.

[23] McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models, *International Statistical Review*, **53**, 61-67.

[24] Nass, C.A.G. (1959). The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika*, **46**, 365-385.

[25] Patefield, W. M. (1981) Algorithm AS159. An efficient method of generating r x c tables with given row and column totals. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **30**, 9197.

[26] Paul, S.R., and Deng, D. (In Press). Assessing goodness of fit of generalized linear models to sparse data using higher order moment corrections. *Sankhya B*.

[27] Pearson, K. (1932). Experimental discussion of the (χ^2, P) test of goodness of fit. *Biometrika*, **24**, 351-381.

[28] Potthoff, R.F. & Whittinghill, M. (1966). Testing for homogeneity: I. The binomial and multinomial Distributions. *Biometrika*, **53**, 167-182.

[29] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

[30] Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003), Multiple Imputation for Statistical Disclosure Limitation, *Journal of Official Statistics*, 19, 1-16.

[31] Reiter, J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets, *Survey Methodology*, 29, 181-188.

[32] Robert, C.P., and Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer.

[33] Rubin, D.B. (1993), Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, 9, 461-468.

[34] Suissa, S., and Shuster, J.J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society. Series A*, **148**, 317-327.

[35] Upton G.J.G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *Journal of the Royal Statistical Society. Series A*, **145**, 86-105.

[36] Wilks, S.S. (1928). The likelihood test of independence in contingency tables. *Biometrika*, **20A**, 263-294.

[37] Xu, D. (2011) *Statistical Issues in Meta-Analysis*. Ph.D. Dissertation, University of Maryland, Baltimore County.

Appendix: Tables Summarizing Simulation Results

Table 1: Rejection probabilities in setting 1; $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = n_3 = 10$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.00	0.0145	0.0014	0.0145	0.0014	0.0014	0.0145	0.0145
0.05	0.0349	0.0054	0.0353	0.0053	0.0054	0.0349	0.0353
0.10	0.0888	0.0237	0.0901	0.0225	0.0237	0.0888	0.0901
0.15	0.1761	0.0622	0.1784	0.0601	0.0622	0.1761	0.1786
0.20	0.2651	0.1218	0.2708	0.1170	0.1218	0.2651	0.2711
0.25	0.3702	0.1982	0.3808	0.1888	0.1982	0.3702	0.3818
0.30	0.4912	0.3042	0.5080	0.2897	0.3041	0.4912	0.5084
0.35	0.5947	0.4136	0.6161	0.3935	0.4133	0.5947	0.6175
0.40	0.6888	0.5289	0.7107	0.5084	0.5286	0.6888	0.7125
0.45	0.7751	0.6436	0.7985	0.6184	0.6430	0.7751	0.8003
0.50	0.8433	0.7466	0.8649	0.7266	0.7465	0.8433	0.8670
0.55	0.8963	0.8252	0.9139	0.8071	0.8249	0.8963	0.9153
0.60	0.9381	0.8932	0.9518	0.8776	0.8931	0.9381	0.9541

Table 2: Rejection probabilities in setting 2; $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = 10$, $n_3 = 60$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.00	0.0676	0.0286	0.0452	0.0013	0.0022	0.0183	0.1432
0.05	0.0102	0.0058	0.0097	0.0002	0.0002	0.0093	0.0292
0.10	0.0116	0.0110	0.0119	0.0000	0.0000	0.0671	0.1124
0.15	0.0770	0.0753	0.0806	0.0000	0.0000	0.2164	0.2960
0.20	0.2352	0.2205	0.2549	0.0000	0.0000	0.3959	0.4979
0.25	0.4392	0.3988	0.4820	0.0000	0.0000	0.5877	0.6868
0.30	0.6665	0.6177	0.7039	0.0002	0.0004	0.7643	0.8303
0.35	0.8292	0.8006	0.8472	0.0016	0.0052	0.8770	0.9104
0.40	0.9262	0.9188	0.9372	0.0112	0.0326	0.9495	0.9659
0.45	0.9697	0.9722	0.9755	0.0550	0.1189	0.9797	0.9864

Table 3: Rejection probabilities in setting 3; $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = 60$, $n_3 = 10$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.00	0.0533	0.0356	0.0433	0.0045	0.0141	0.0352	0.0860
0.05	0.1397	0.1075	0.1236	0.0077	0.0179	0.0901	0.1261
0.10	0.2635	0.2152	0.2413	0.0158	0.0298	0.1785	0.2125
0.15	0.4040	0.3551	0.3840	0.0334	0.0592	0.3025	0.3320
0.20	0.5377	0.4907	0.5199	0.0681	0.1060	0.4274	0.4578
0.25	0.6723	0.6323	0.6575	0.1137	0.1679	0.5753	0.6036
0.30	0.7760	0.7433	0.7654	0.1811	0.2501	0.6894	0.7156
0.35	0.8565	0.8308	0.8463	0.2548	0.3370	0.7839	0.8048
0.40	0.9102	0.8961	0.9057	0.3570	0.4472	0.8657	0.8820
0.45	0.9475	0.9391	0.9450	0.4577	0.5555	0.9194	0.9282

Table 4: Rejection probabilities in setting 4; $\pi_0 = 0.05$, $k = 3$, $n_1 = n_2 = n_3 = 60$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.00	0.0414	0.0318	0.0435	0.0110	0.0147	0.0405	0.0611
0.02	0.0651	0.0539	0.0682	0.0243	0.0282	0.0636	0.0893
0.04	0.1284	0.1100	0.1324	0.0566	0.0689	0.1264	0.1618
0.06	0.2306	0.2073	0.2356	0.1282	0.1490	0.2287	0.2735
0.08	0.3659	0.3393	0.3697	0.2306	0.2616	0.3640	0.4111
0.10	0.4946	0.4677	0.4993	0.3470	0.3876	0.4934	0.5466
0.12	0.6357	0.6106	0.6393	0.4814	0.5230	0.6349	0.6798
0.14	0.7400	0.7201	0.7435	0.6117	0.6517	0.7392	0.7800
0.16	0.8306	0.8140	0.8328	0.7196	0.7568	0.8299	0.8609
0.18	0.9001	0.8883	0.9019	0.8171	0.8489	0.8996	0.9197
0.20	0.9451	0.9370	0.9459	0.8871	0.9106	0.9446	0.9578

Table 5: Rejection probabilities in setting 5; $\pi_0 = 0.001$, $k = 8$, $n_i = 20 \times 2^{i-1}$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.000	0.1096	0.0410	0.0359	0.0068	0.0098	0.0211	0.6661
0.001	0.0702	0.0300	0.0286	0.0003	0.0015	0.0236	0.4297
0.002	0.0515	0.0211	0.0194	0.0019	0.0075	0.0949	0.2987
0.003	0.0638	0.0243	0.0221	0.0220	0.0530	0.2434	0.3534
0.004	0.1226	0.0418	0.0328	0.0858	0.1681	0.4499	0.5194
0.005	0.2531	0.0999	0.0760	0.2176	0.3397	0.6532	0.7022
0.006	0.4254	0.2162	0.1720	0.3976	0.5322	0.8017	0.8329
0.007	0.6037	0.3844	0.3207	0.5883	0.7096	0.9004	0.9171
0.008	0.7372	0.5509	0.4886	0.7230	0.8166	0.9496	0.9585
0.009	0.8525	0.7067	0.6526	0.8416	0.9023	0.9773	0.9817

Table 6: Rejection probabilities in setting 6; $\pi_0 = 0.001$, $k = 8$, $n_i = 20 \times 2^{8-i}$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.00	0.1144	0.0482	0.0428	0.0057	0.0068	0.0212	0.6537
0.01	0.2703	0.2143	0.2111	0.0092	0.0112	0.0423	0.6796
0.02	0.4135	0.3680	0.3649	0.0102	0.0127	0.0838	0.7006
0.03	0.5178	0.4768	0.4754	0.0115	0.0171	0.1395	0.7307
0.04	0.6173	0.5837	0.5839	0.0186	0.0259	0.2173	0.7579
0.05	0.6833	0.6564	0.6564	0.0275	0.0372	0.2757	0.7674
0.06	0.7410	0.7185	0.7202	0.0348	0.0476	0.3571	0.7977
0.07	0.7922	0.7750	0.7760	0.0459	0.0662	0.4251	0.8218
0.08	0.8319	0.8163	0.8180	0.0693	0.0907	0.4973	0.8529
0.09	0.8687	0.8574	0.8574	0.0927	0.1241	0.5623	0.8688

Table 7: Rejection probabilities in setting 7; $\pi_0 = 0.001$, $k = 8$, $n_1 = \cdots = n_8 = 2560$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.0000	0.0424	0.0420	0.0462	0.0187	0.0281	0.0532	0.0462
0.0005	0.0645	0.0643	0.0687	0.0324	0.0483	0.0776	0.0687
0.0010	0.1348	0.1347	0.1388	0.0788	0.1048	0.1512	0.1388
0.0015	0.2523	0.2549	0.2598	0.1757	0.2123	0.2788	0.2598
0.0020	0.4030	0.4030	0.4098	0.3097	0.3534	0.4289	0.4099
0.0025	0.5555	0.5580	0.5636	0.4612	0.5104	0.5817	0.5639
0.0030	0.6852	0.6851	0.6917	0.5984	0.6438	0.7080	0.6919
0.0035	0.7943	0.7938	0.7988	0.7264	0.7615	0.8104	0.7989
0.0040	0.8673	0.8669	0.8700	0.8166	0.8441	0.8784	0.8703
0.0045	0.9251	0.9250	0.9271	0.8938	0.9102	0.9320	0.9272

Table 8: Rejection probabilities in setting 8; $\pi_0 = 0.016$, $k = 8$, $n_i = 20 \times 2^{i-1}$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.000	0.0635	0.0489	0.0457	0.0058	0.0101	0.0486	0.0458
0.002	0.0531	0.0428	0.0408	0.0071	0.0129	0.0547	0.0520
0.004	0.0677	0.0537	0.0494	0.0185	0.0312	0.0909	0.0862
0.006	0.1050	0.0837	0.0781	0.0518	0.0765	0.1606	0.1528
0.008	0.1796	0.1503	0.1421	0.1194	0.1657	0.2763	0.2665
0.010	0.2874	0.2490	0.2396	0.2241	0.2910	0.4178	0.4061
0.012	0.4283	0.3849	0.3753	0.3660	0.4435	0.5645	0.5542
0.014	0.5861	0.5389	0.5313	0.5332	0.6142	0.7215	0.7106
0.016	0.7184	0.6784	0.6743	0.6688	0.7421	0.8252	0.8177
0.018	0.8302	0.8011	0.7959	0.7988	0.8540	0.9068	0.9023

Table 9: Rejection probabilities in setting 9; $\pi_0 = 0.001$, $k = 40$, $n_i = 20 \times 2^{(i-1) \bmod 8}$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.000	0.1932	0.0491	0.0531	0.0169	0.0274	0.0411	0.0983
0.001	0.1702	0.0394	0.0413	0.0882	0.1099	0.0658	0.0917
0.002	0.2007	0.0536	0.0532	0.3112	0.3554	0.2026	0.2179
0.003	0.2877	0.0825	0.0826	0.6037	0.6431	0.4403	0.4495
0.004	0.4616	0.1758	0.1730	0.8297	0.8519	0.7001	0.7041
0.005	0.6540	0.3320	0.3270	0.9391	0.9489	0.8651	0.8658
0.006	0.8142	0.5312	0.5251	0.9838	0.9867	0.9500	0.9500
0.007	0.9098	0.7143	0.7107	0.9956	0.9967	0.9844	0.9846
0.008	0.9644	0.8430	0.8397	0.9990	0.9994	0.9959	0.9958
0.009	0.9897	0.9368	0.9357	0.9997	0.9997	0.9988	0.9988

Table 10: Rejection probabilities in setting 10; $\pi_0 = 0.001$, $k = 40$, $n_i = 20 \times 2^{(8-i) \bmod 8}$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.00	0.1826	0.0495	0.0524	0.0214	0.0319	0.0411	0.0973
0.01	0.3273	0.1260	0.1330	0.0229	0.0316	0.0588	0.1115
0.02	0.4566	0.2073	0.2162	0.0174	0.0301	0.1013	0.1479
0.03	0.5534	0.2948	0.3068	0.0219	0.0327	0.1673	0.2075
0.04	0.6262	0.3673	0.3764	0.0219	0.0353	0.2326	0.2723
0.05	0.6938	0.4410	0.4526	0.0254	0.0391	0.2992	0.3308
0.06	0.7552	0.5126	0.5204	0.0286	0.0418	0.3717	0.3998
0.07	0.7948	0.5737	0.5842	0.0376	0.0534	0.4412	0.4673
0.08	0.8393	0.6349	0.6430	0.0459	0.0649	0.5100	0.5327
0.09	0.8665	0.6865	0.6948	0.0544	0.0755	0.5672	0.5881
0.10	0.8939	0.7276	0.7346	0.0615	0.0860	0.6241	0.6398
0.11	0.9160	0.7719	0.7791	0.0691	0.0947	0.6823	0.6965
0.12	0.9305	0.8053	0.8112	0.0917	0.1190	0.7230	0.7358
0.13	0.9458	0.8404	0.8464	0.1083	0.1436	0.7681	0.7796
0.14	0.9592	0.8648	0.8671	0.1262	0.1619	0.8032	0.8132
0.15	0.9639	0.8817	0.8858	0.1471	0.1856	0.8328	0.8403

Table 11: Rejection probabilities in setting 11; $\pi_0 = 0.001$, $k = 40$, $n_1 = \cdots = n_{40} = 2560$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.000	0.0479	0.0456	0.0487	0.0329	0.0420	0.0569	0.0472
0.001	0.0904	0.0865	0.0910	0.0672	0.0801	0.1038	0.0900
0.002	0.2769	0.2712	0.2780	0.2369	0.2604	0.2968	0.2756
0.003	0.5503	0.5440	0.5519	0.5111	0.5335	0.5752	0.5484
0.004	0.7704	0.7664	0.7713	0.7391	0.7573	0.7846	0.7695
0.005	0.9059	0.9043	0.9064	0.8913	0.9002	0.9146	0.9053
0.006	0.9717	0.9715	0.9720	0.9660	0.9701	0.9750	0.9716

Table 12: Rejection probabilities in setting 12; $\pi_0 = 0.008$, $k = 40$, $n_i = 20 \times 2^{(i-1) \bmod 8}$

δ	Pearson	ExactC	Nass	PW	PWBB	Xu	PD
0.000	0.0911	0.0495	0.0529	0.0272	0.0395	0.0567	0.0470
0.001	0.0868	0.0464	0.0488	0.0382	0.0512	0.0595	0.0490
0.002	0.0931	0.0481	0.0516	0.0604	0.0783	0.0738	0.0591
0.003	0.1085	0.0617	0.0658	0.1285	0.1562	0.0978	0.0805
0.004	0.1339	0.0734	0.0778	0.2175	0.2541	0.1432	0.1230
0.005	0.1711	0.1047	0.1093	0.3375	0.3824	0.2055	0.1787
0.006	0.2405	0.1544	0.1616	0.4880	0.5304	0.2983	0.2669
0.007	0.3186	0.2142	0.2230	0.6313	0.6707	0.4055	0.3728
0.008	0.4179	0.2899	0.3009	0.7516	0.7824	0.5265	0.4887
0.009	0.5230	0.3929	0.4017	0.8440	0.8685	0.6340	0.6051
0.010	0.6338	0.5137	0.5245	0.9115	0.9259	0.7408	0.7185
0.011	0.7371	0.6272	0.6361	0.9545	0.9646	0.8287	0.8076
0.012	0.8047	0.7089	0.7159	0.9762	0.9820	0.8874	0.8714
0.013	0.8801	0.8114	0.8188	0.9901	0.9927	0.9353	0.9240
0.014	0.9290	0.8766	0.8823	0.9954	0.9965	0.9654	0.9593
0.015	0.9609	0.9260	0.9299	0.9984	0.9986	0.9836	0.9802