

EXACT PROPERTIES OF A NEW TEST AND OTHER TESTS FOR DIFFERENCES BETWEEN SEVERAL BINOMIAL PROPORTIONS

*K. Krishnamoorthy**

Department of Mathematics
University of Louisiana at Lafayette
Lafayette, LA 70504

Jie Peng†

Department of Finance, Economics and Decision Science
St. Ambrose University
Davenport, Iowa 52803

Abstract

The problem of testing equality of several binomial proportions is considered. An approximate unconditional test (AU-test) is proposed by extending a method for the two-sample case. Exact binomial distributions are used to evaluate Type I error rates of the usual chi-square test, an exact conditional test, a conditional test based on mid p-value and the AU-test numerically. The AU-test and the conditional test based on mid p-values control the Type I error rates very satisfactorily even for small samples whereas the exact conditional test is too conservative. The powers of the chi-square test, conditional test based on mid-p value, and the AU-test are evaluated and compared. Power comparison shows that all three tests exhibit similar power properties when their sizes are within the nominal level. The AU-test practically behaves like an exact test even for small samples, and can be safely used for applications. The results are illustrated using an example where small proportions are to be compared.

Key Words and Phrases: Conditional exact test; Mid-p value; Parametric bootstrap; Powers; Size

1. Introduction

The problem of testing equality of two proportions is well addressed in the literature. Numerous articles have been written on this topic proposing several approximate and exact tests. Among the available methods, Fisher's conditional test (conditional on the total number of successes) has been a popular choice because of its simplicity and guarantee that the Type I error rates never exceed the nominal level. This test, however, is overly conservative when the sample sizes are small and/or the proportions are at the boundaries. For these reasons, exact unconditional procedures gained popularity among researchers as they are usually less discrete and more powerful than the conditional test. Barnard (1945, 1947) developed an unconditional test by maximizing the p-value with respect to the common unknown parameter under the null hypothesis of equality of proportions.

*E-mail address: krishna@louisiana.edu; Tel. 337 482 5283; Fax 337 482 5346 (Corresponding Author)

†E-mail address: PengJie@sau.edu

This unconditional test is also subject to criticism because it is computationally very intensive, and is also conservative. Several articles addressed the two-sample problem and extensive comparison studies were carried out by many authors. For example, see Upton (1982), Haber (1986), Storer and Kim (1990), Berger (1996), Martin and Silva (1994) and Martin et al. (1998, 2002) and Chan and Zhang (1999) and the references therein.

The problem of comparing several proportions arises when we have samples from m populations (lots of items, opinions of several groups of people on a public issue, products from different suppliers, etc.), and want to test whether there are significant differences in the proportions for these populations. This problem is also well-known, and has been discussed in many text books (e.g., Scheaffer and McClave 1994 and Zar 1999). However, unlike the two-sample case, only very limited results are available. Williams (1988) pointed out a practical situation where small proportions are to be compared, and proposed conditional tests (conditionally given the total number of successes from all samples) based on several statistics appropriate for different alternative hypotheses such as $p_1 \geq p_i$, $i = 2, \dots, m$ or $p_1 \leq p_2 \leq \dots \leq p_j \geq p_{j+1} \geq \dots \geq p_m \geq p_1$ for some j , $1 \leq j \leq m$. Williams also studied conditional Type I error rates of the proposed tests.

To describe the problem formally, consider m independent binomial random variables X_1, \dots, X_m with $X_i \sim \text{binomial}(n_i, p_i)$, $0 < p_i < 1$, $i = 1, \dots, m$. Let k_i be an observed value of X_i , $i = 1, \dots, m$. The hypotheses of interest are

$$H_0 : p_1 = \dots = p_m \text{ vs. } H_a : p_i \neq p_j \text{ for some } i \neq j. \quad (1)$$

Kulkarni and Shah (1995) and Krishnamoorthy et al. (2004) considered the above problem when the common proportion under H_0 is specified. Specifically, Krishnamoorthy et al. proposed a simple exact method for hypothesis testing. If the common proportion is unspecified, then the χ^2 approximate test is commonly used. The χ^2 -test is based on the statistic

$$Q_x = \sum_{i=1}^m \frac{n_i(\hat{p}_{xi} - \hat{p}_x)^2}{\hat{p}_x(1 - \hat{p}_x)}, \quad (2)$$

where $\hat{p}_{xi} = \frac{X_i}{n_i}$, $i = 1, \dots, m$ and $\hat{p}_x = \frac{\sum_{i=1}^m X_i}{\sum_{i=1}^m n_i}$. Under H_0 , Q_x follows a χ_{m-1}^2 distribution approximately. Notice that, for the case of $m = 2$, the statistic Q_x is the squared Z-test statistic

$$\frac{\hat{p}_{x1} - \hat{p}_{x2}}{\sqrt{\hat{p}_x(1 - \hat{p}_x) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

for comparing two proportions. For this case, Storer and Kim (1990) showed that the χ^2 test is often liberal especially when sample sizes are quite different.

In this article, we study unconditional Type I error rates of the usual chi-square test, the conditional test, the conditional test based on mid-p value, and a new test. Specifically, the Type I error rates and powers are evaluated using the exact binomial distributions not normal or chi-square approximations. The new test is obtained by extending an approximate unconditional test (AU-test) first considered in Storer and Kim (1990), later by Krishnamoorthy and Thomson (2002) for the finite population case and Krishnamoorthy and Thomson (2004) for the Poisson case. Even though there are several alternative tests are available for the two-sample case, not all of them can be easily extended to the present problem of comparing several proportions, except the AU-test. Storer and Kim (1990) and Krishnamoorthy and Thomson (2002, 2004) found that the AU-test performed very satisfactorily even for small samples. As the AU-test is actually based on an estimated p-value, Krishnamoorthy and Thomson (2002, 2004) referred to it as the E-test.

In view of the above motivation and discussion, this article is organized as follows. In the following section, we outline the χ^2 -test, the conditional tests and the AU-test. The likelihood ratio test (LRT) for comparing proportions in a multinomial setup (Cai and Krishnamoorthy 2006) exhibited poor size properties and usually inferior to the χ^2 -test. Our preliminary numerical studies in the present setup showed the LRT is in general inferior to the χ^2 -test, and so we will not include the LRT here for comparisons. An exact method of calculating size and power of a test is given in Section 3. Using this method, the sizes of the tests are evaluated for $m = 3$ and 4. Powers are also evaluated for a few parameter and sample size configurations. The size and power studies show that the AU-test followed by the conditional test based on mid-p value are very satisfactory even for small samples. The Type I error rates of the AU-test seldom exceed the nominal level by a negligible amount, and perform almost like an exact test. The results are illustrated using an example in Section 4. Some concluding remarks are given in Section 5.

2. The Tests

Let (k_1, \dots, k_m) be an observed value of (X_1, \dots, X_m) , and Q_k be an observed value of Q_x in (2). That is,

$$Q_k = \sum_{i=1}^m \frac{n_i (\hat{p}_{ki} - \hat{p}_k)^2}{\hat{p}_k (1 - \hat{p}_k)}, \quad (3)$$

where $\hat{p}_{ki} = \frac{k_i}{n_i}$, $i = 1, \dots, m$ and $\hat{p}_k = \frac{\sum_{i=1}^m k_i}{\sum_{i=1}^m n_i}$ is the pooled estimate of the common unknown proportion.

2.1. The χ^2 -test

For a given level α , the χ^2 -test rejects the null hypothesis in (1) if Q_k is greater than or equal to the critical value $\chi_{m-1, \alpha}^2$, where $\chi_{a, \alpha}^2$ denotes the upper α quantile of the χ_a^2 distribution. Equivalently, it rejects the null hypothesis if $P(\chi_{m-1}^2 \geq Q_k | H_0) \leq \alpha$.

2.2. Conditional Tests

An exact conditional test can be developed using the conditional joint distribution of X_1, \dots, X_m given that $\sum_{i=1}^m X_i = x$ and $p_1 = \dots = p_m$. This conditional distribution is multivariate hypergeometric with the probability mass function given by

$$P\left(X_1 = x_1, \dots, X_m = x_m \mid \sum_{i=1}^m X_i = x\right) = \frac{\binom{n_1}{x_1} \binom{n_2}{x_2} \dots \binom{n_m}{x_m}}{\binom{n}{x}}, \quad (4)$$

where $n = n_1 + \dots + n_m$ and $x = x_1 + \dots + x_m$.

The C-test

If we use the Q_x as a test statistic, then the exact conditional p-value can be computed using the expression

$$P\left(Q_x \geq Q_k \mid \sum_{i=1}^m X_i = x\right) = \sum_{x_1=l_1}^{u_1} \dots \sum_{x_{m-1}=l_{m-1}}^{u_{m-1}} \frac{\binom{n_1}{x_1} \binom{n_2}{x_2} \dots \binom{n_m}{x_m}}{\binom{n}{x}} I(Q_x \geq Q_k), \quad (5)$$

where Q_k is an observed value of Q_x , $I(\cdot)$ is the indicator function and, for $i = 1, \dots, m-1$,

$$l_i = \max \left\{ 0, x - \sum_{j=1}^{i-1} x_j - \sum_{j=i+1}^k n_j \right\}, \quad u_i = \min \left\{ n_i, x - \sum_{j=1}^{i-1} x_j \right\} \quad (6)$$

and

$$u_m = \min \left\{ n_m, x - \sum_{j=1}^{m-1} x_j \right\}.$$

Notice that we have used $m-1$ sums in (5), because $x_1 + \dots + x_m = x$. One could use any other test statistic in place of Q_x depending on the alternative hypothesis. As pointed out by Williams (1988), the statistic Q_x given in (2) is appropriate for the general alternative hypothesis in (1). We refer to this test as the C-test.

CM-test

Some authors suggested using mid-p values for hypothesis tests involving discrete distributions. A mid-p value for the above conditional test can be computed as

$$\frac{1}{2} \left[P \left(Q_x \geq Q_k \mid \sum_{i=1}^m X_i = x \right) + P \left(Q_x > Q_k \mid \sum_{i=1}^m X_i = x \right) \right].$$

The null hypothesis in (1) will be rejected whenever this mid-p value is less than or equal to the nominal level α . As the conditional tests are usually very conservative, the use of mid-p value approach will reduce the conservatism. Williams (1988) preferred the mid-p value approach on the basis of arguments favoring the use of mid-p values given in Lancaster (1961), Anscombe (1981) and Frank (1986). Furthermore, for the two-sample case, Martin et al. (1998) compared several tests, and concluded that Fisher's conditional test based on mid-p value is preferable to others for simplicity and power.

2.3. The AU-test

If the common proportion under H_0 is specified as p_0 , then the p-value of the statistic Q_x can be computed using the expression

$$\sum_{x_1=0}^{n_1} \dots \sum_{x_m=0}^{n_m} \left(\prod_{i=1}^m f(x_i; n_i, p_0) \right) I(Q_x \geq Q_k), \quad (7)$$

where

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n, \quad 0 < p < 1,$$

is the binomial probability mass function. The test based on the p-value in (7) is exact in the sense that the Type I error rates are always within the nominal level. The common proportion is usually unknown, and in this case the unconditional p-value can be obtained as

$$\sup_{0 < p < 1} \sum_{x_1=0}^{n_1} \dots \sum_{x_m=0}^{n_m} \left(\prod_{i=1}^m f(x_i; n_i, p) \right) I(Q_x \geq Q_k). \quad (8)$$

The test based on the above p-value is an exact unconditional test which is an extension of Barnard's unconditional test for the two-sample case. As pointed out earlier, this test is computationally intensive even for $m = 2$. As an alternative, Storer and Kim (1990) proposed an approximate unconditional test (AU-test) on the basis of the p-value evaluated at the maximum likelihood estimator of p . That is, p-value is computed as

$$\sum_{x_1=0}^{n_1} \dots \sum_{x_m=0}^{n_m} \left(\prod_{i=1}^m f(x_i; n_i, \hat{p}_k) \right) I(Q_x \geq Q_k), \quad (9)$$

where \hat{p}_k is as defined in (3). As the p-value is evaluated only at \hat{p}_k , it is easier to compute than the p-value in (8). As the test is based on an estimated p-value, Krishnamoorthy and Thomson (2002,2004) referred to this test as the E-test. This AU-test should be used with the convention of not rejecting the null hypothesis in extreme cases where $(k_1, \dots, k_m) = (0, \dots, 0)$ or (n_1, \dots, n_m) .

Even though the p-value is calculated only at \hat{p}_k , calculation is quite time consuming if m and/or sample sizes are large. An alternative simple approach of computing the p-value in (9) is on the basis of an observation by Krishnamoorthy and Thomson (2002). These authors noted that the AU-test (in the context of comparing two proportions in finite populations) is essentially equivalent to the one based on the parametric bootstrap (PB) approach (see Krishnamoorthy and Thomson 2002, Remark 2). In other words, the AU-test described above can be regarded as the PB method using exact binomial distribution rather than simulation. Thus, the p-value in (9) can be computed using the PB approach given in the following algorithm.

Algorithm 1

For a given (n_1, \dots, n_m) and (k_1, \dots, k_m) :

compute the sample proportion $\hat{p}_{ki} = \frac{k_i}{n_i}$, $i = 1, \dots, m$ and

the pooled sample proportion $\hat{p}_k = \frac{\sum_{i=1}^m k_i}{\sum_{i=1}^m n_i}$

compute the observed statistic $Q_k = \sum_{i=1}^m \frac{n_i(\hat{p}_{ki} - \hat{p}_k)^2}{\hat{p}_k(1 - \hat{p}_k)}$

set $T = 0$

For $i = 1$ to N

generate $x_j \sim \text{binomial}(n_j, \hat{p}_k)$, $j = 1, \dots, m$

compute $\hat{p}_{xj} = \frac{x_j}{n_j}$, $j = 1, \dots, m$ and $\hat{p}_x = \frac{\sum_{j=1}^m x_j}{\sum_{j=1}^m n_j}$

compute $Q_x = \sum_{j=1}^m \frac{n_j(\hat{p}_{xj} - \hat{p}_x)^2}{\hat{p}_x(1 - \hat{p}_x)}$

if $Q_x \geq Q_k$, set $T = T + 1$

(end do loop)

$\frac{T}{N}$ is the PB p-value that should be close to the one in (9) if N is large enough.

2.4. Comparison of P-values

We computed the p-values of all four tests when $m = 4$ and presented them in Table 1. The values of (n_1, \dots, n_4) and (k_1, \dots, k_4) were chosen arbitrarily. To judge the accuracies of the PB test based on Algorithm 1, we computed its p-values and presented them in Table 1 along with the exact p-values of the AU-test based on (9). We see from Table 1 that these p-values are practically the same for all the cases considered. Thus, for large m and/or moderate to large sample sizes one can use Algorithm 1 to compute the p-value of the AU-test. As expected, the p-values of the C-test are always greater than or equal to the corresponding p-values of the CM-test. The p-values of the χ^2 -test are the smallest in many cases, especially when the sample sizes are small or the sample sizes are much different.

Table 1. Comparison of p-values of (a) χ^2 -test, (b) C-test, (c) CM-test, (d) AU-test (9) and (e) PB Algorithm 1 with $N = 100,000$ runs

Sample Sizes (n_1, \dots, n_4)	No. of Successes (k_1, \dots, k_4)	Observed Statistic Q_k	p-values				
			a	b	c	d	e
(4, 4, 4, 4)	(4, 4, 1, 1)	9.600	.022	.045	.033	.019	.019
(5, 9, 3, 12)	(4, 3, 2, 9)	4.719	.194	.207	.203	.207	.205
(32, 5, 10, 12)	(16, 4, 4, 4)	3.390	.335	.374	.372	.352	.351
(10, 10, 10, 10)	(9, 5, 5, 9)	7.619	.055	.053	.050	.048	.048
(25, 5, 5, 5)	(20, 4, 3, 1)	7.619	.055	.056	.049	.052	.052
(13, 12, 11, 4)	(4, 4, 4, 4)	6.744	.081	.085	.085	.079	.078
(23, 4, 4, 4)	(16, 3, 2, 1)	3.435	.329	.377	.329	.329	.330
(40, 40, 40, 40)	(12, 9, 6, 3)	7.385	.061	.065	.061	.061	.062
(3, 3, 3, 3)	(3, 3, 1, 1)	6.000	.112	.182	.127	.143	.143
(5, 5, 5, 5)	(5, 2, 1, 1)	8.687	.034	.041	.033	.028	.027
(12, 23, 45, 60)	(4, 9, 23, 36)	4.732	.192	.193	.193	.196	.196
(5, 4, 8, 9)	(5, 3, 3, 7)	6.375	.095	.097	.092	.091	.091
(32, 4, 4, 4)	(16, 4, 3, 3)	4.701	.195	.247	.243	.236	.238

3. Size and Power Studies of the Tests

The exact size or power of the χ^2 -test can be computed using the expression

$$\sum_{k_1=0}^{n_1} \dots \sum_{k_m=0}^{n_m} \left(\prod_{i=1}^m f(k_i; n_i, p_i) \right) I(P(\chi_{m-1}^2 \geq Q_k) \leq \alpha). \quad (10)$$

Notice that the above expression gives Type I error rate when $p_1 = \dots = p_m$ and power otherwise. The exact size and power of the other tests can be computed using the expression

$$\sum_{k_1=0}^{n_1} \dots \sum_{k_m=0}^{n_m} \left(\prod_{i=1}^m f(k_i; n_i, p_i) \right) I(\text{p-value} \leq \alpha). \quad (11)$$

For example, to compute the size or power of the AU-test, use the p-value given in (9).

3.1. Size Properties

To compare the conditional test (C-test) and the one based on the mid-p value (CM-test), we plotted the sizes of these tests as a function of the common p under H_0 . The sizes are plotted in Figure 1. It is clear from these plots that the conditional test is very conservative when the sample sizes are small and/or they are very different. The Type I error rates of the CM-test occasionally exceed the nominal level but not more than 0.06 when the nominal level is 0.05. As the conservatism of the C-test is worse than the liberalism of the CM-test, we prefer to use the latter for further comparison with the other tests.

The sizes of the χ^2 -test, AU-test and the CM-test are computed as a function of the common p under H_0 and plotted in Figure 2 for $m = 3$ and in Figure 3 for $m = 4$. We used the expression in (10) for computing the size of the χ^2 -test, and the one in (11) for computing the size of the CM-test

and the AU-test. As we are interested in small-sample properties of the tests, we did not include large samples. We observe the following from Figures 2 and 3.

1. The sizes of the χ^2 -test exceed the nominal level quite often. In particular, when $n_1 = 25, n_2 = 5, n = 5$, the size of the χ^2 -test exceeds 0.3 when the nominal level is 0.05. This implies that there are situations where the χ^2 -test could be too liberal. Even though the χ^2 -test is still liberal for the case of $m = 4$ (Figure 3), it exhibits slightly better performance than it did for the case of $m = 3$.
2. The CM-test offers improvement over the χ^2 -test in controlling sizes. This test is slightly liberal when some sample sizes are very different.
3. It is clear from Figures 2 and 3 that the AU-test controls the Type I error rates very well regardless of sample sizes except in two cases (Figure 2, $n_1 = n_2 = n_3 = 12$ and Figure 3, $n = 7$) where its Type I error rates barely exceed the nominal level.

The χ^2 -test, in general, performs satisfactorily if sample sizes are not very small and/or drastically different. It appears that the liberalism of the test diminishes with increasing m . The CM-test offers improvement over the χ^2 -test, but it is computationally intensive for $m \geq 4$ and/or the observed number of successes are large. Over all, the AU-test performs very satisfactorily and behaves almost like an exact test; it can be safely used for applications regardless of sample sizes.

3.2. Power Properties

In general, it is fair to compare two tests with respect to powers only if they are level α tests. Therefore, for power comparison studies we chose the situations where all three tests control the Type I error rates within the nominal level. The powers are plotted (alternative hypotheses are specified below the plots) in Figure 4. It is clear from these plots that only little differences exist between the powers. No test dominates the others uniformly except in one case where $n_1 = 5, n_2 = 5, n_3 = 5, n_4 = 10$; in this exceptional case, we observe from Figure 4 that the sizes of the AU-test and the CM-test are greater than or equal to (but within the nominal level) those of the χ^2 -test, and as a result, they provide slightly more powers than the χ^2 -test.

4. An Example

This example is taken from Williams (1988) and the results are outcomes of a chromosome aberration assay study to determine whether or not the compound is clastogenic. The purpose of this study is to assess the toxicity of compounds such as drugs, food additives and pesticides. The data for two assays A and B from Table 1 of Williams (1988) are reproduced here in Table 2a.

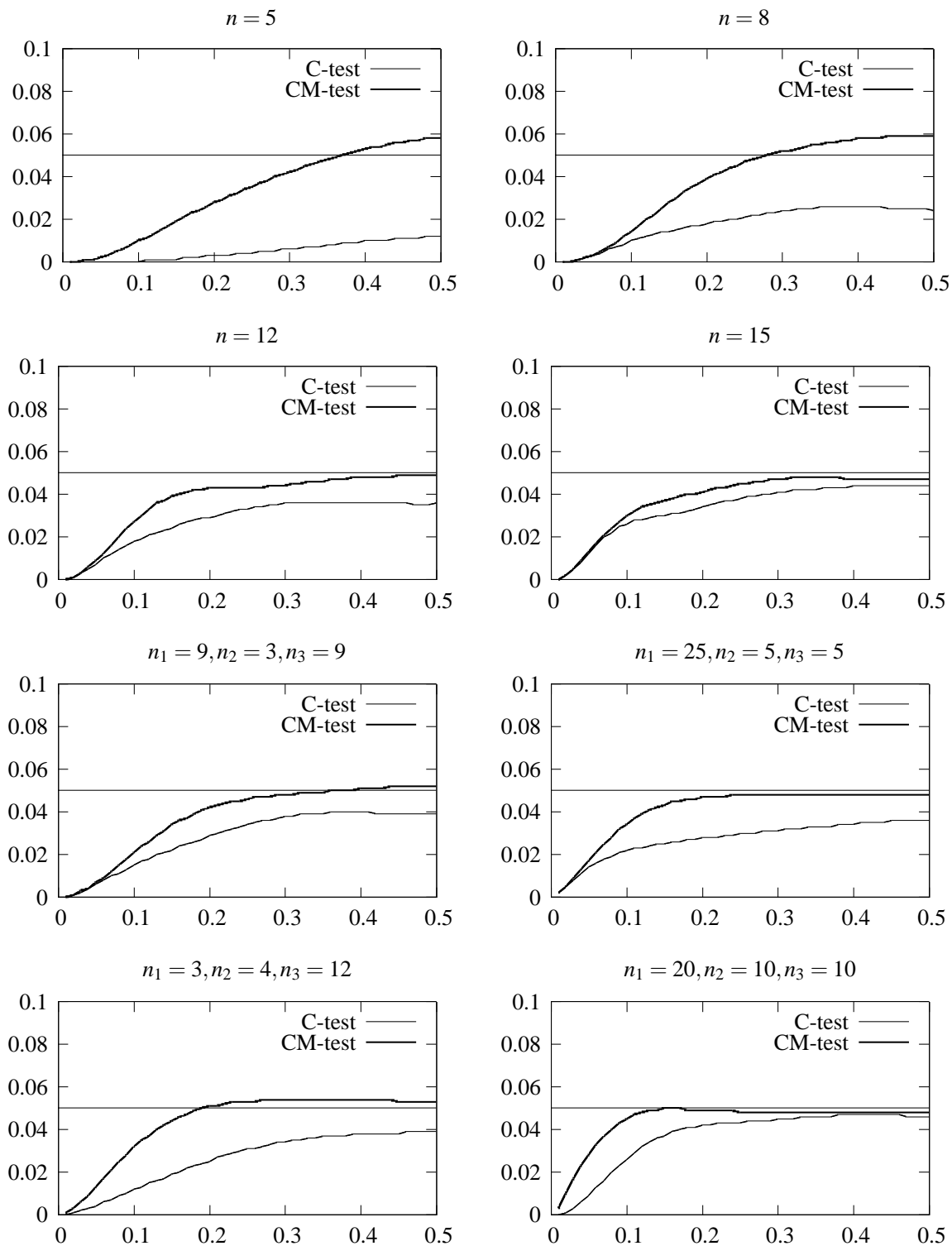


Figure 1. The sizes of the conditional test (C-test) and the conditional test based on mid p-value (CM-test) as functions of p ; $\alpha = 0.05, m = 3$ and $n = n_1 = n_2 = n_3$

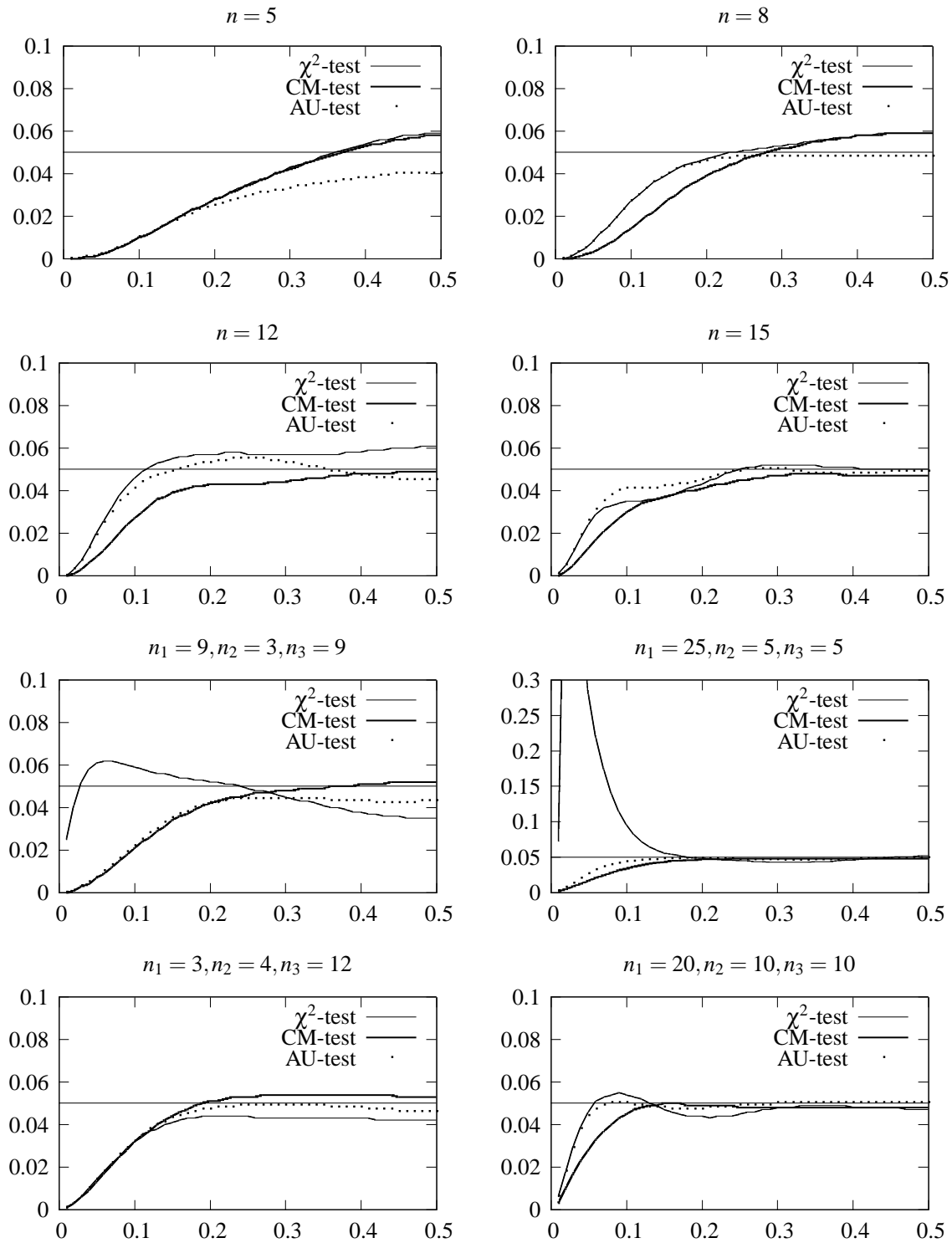


Figure 2. The sizes of the tests as functions of p ; $\alpha = 0.05, m = 3$ and $n = n_1 = n_2 = n_3$.

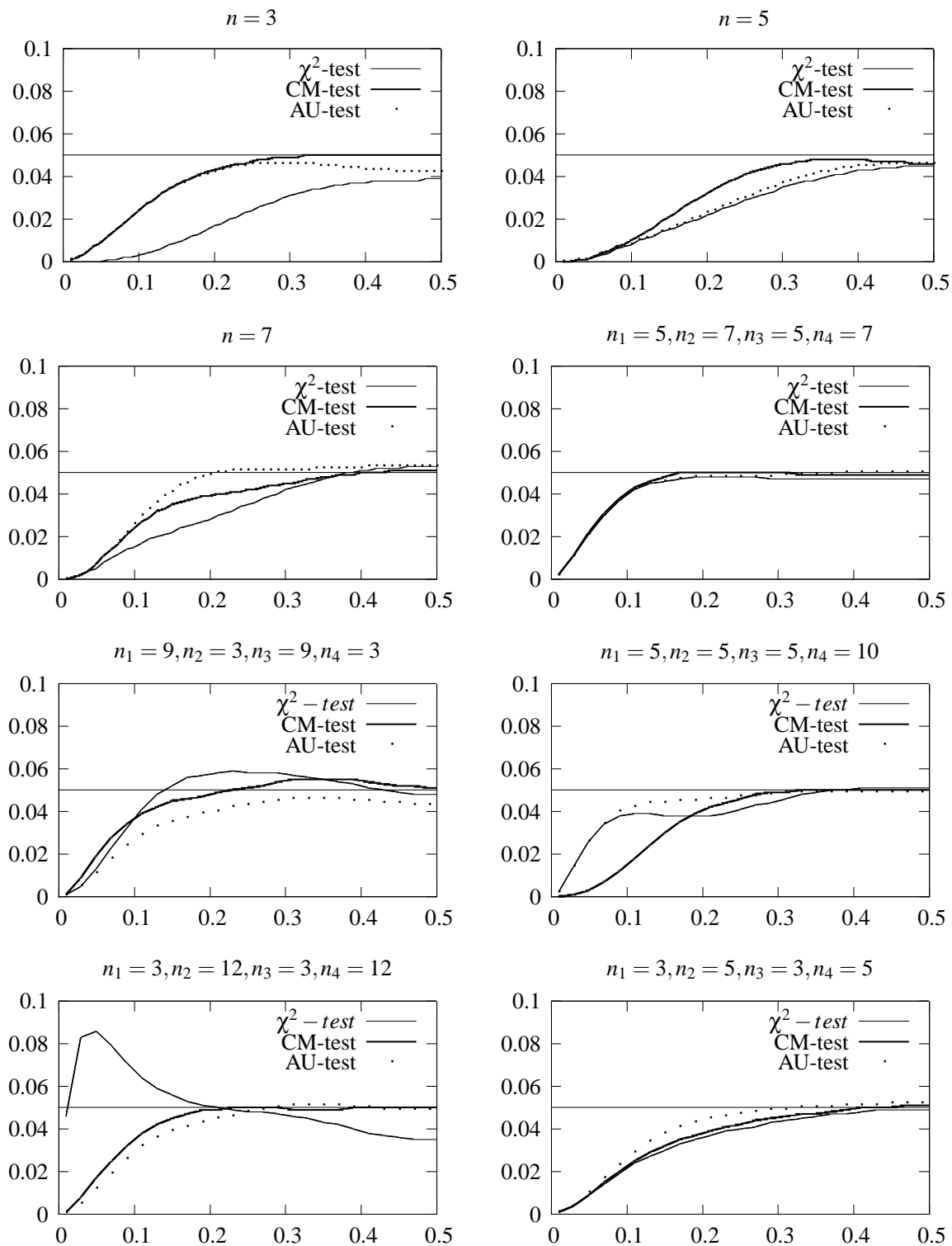


Figure 3. The sizes of the tests as a function of p ; $\alpha = 0.05, m = 4$ and $n = n_1 = \dots = n_4$.

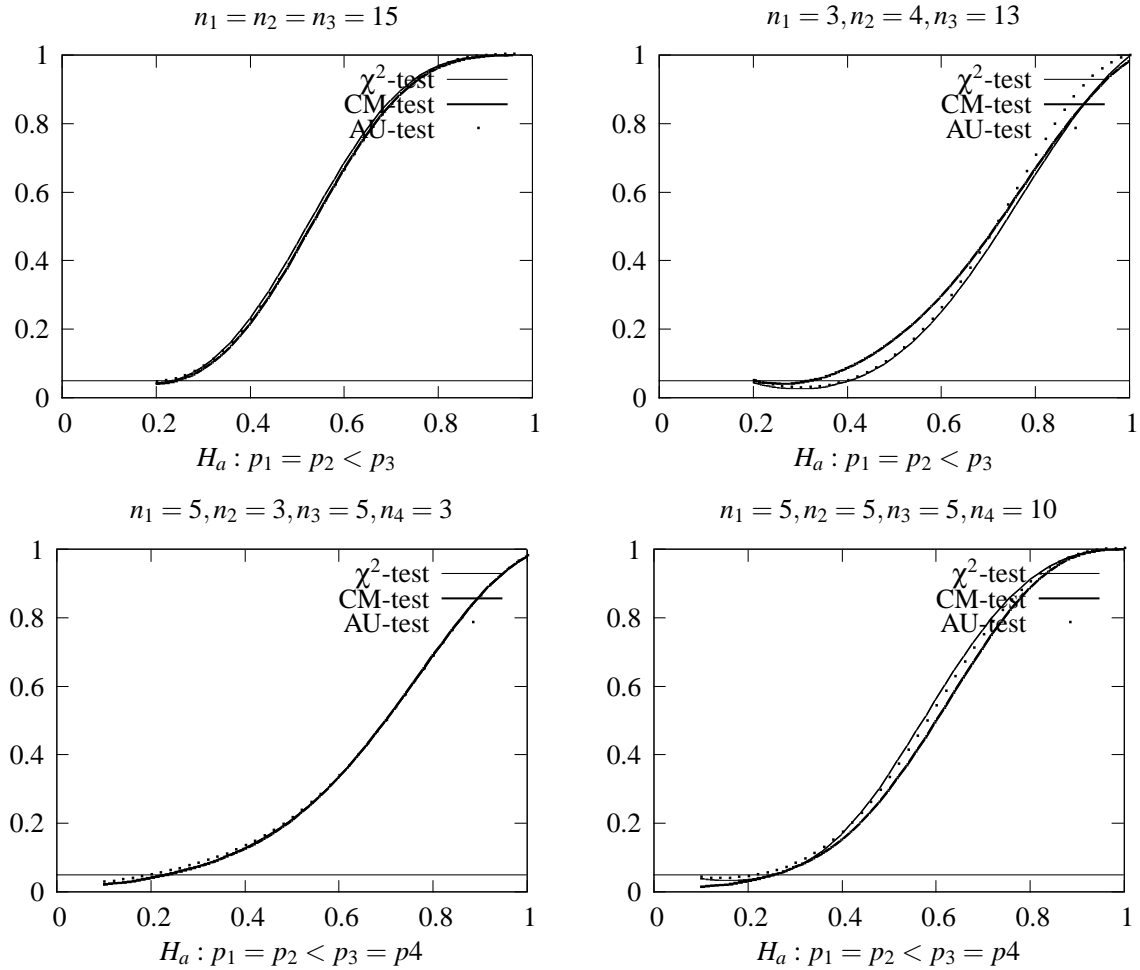


Figure 4. Powers of the tests as a function of p_3 ; $\alpha = 0.05$ and $n = n_1 = \dots = n_4$.

Table 2a. Aberrant cells recorded in two in vitro chromosome aberration assays

Assay	Treatment number, i	Dose levels (mg/ml)	No. of cells sampled, n_i	No. of cells aberrant, k_i	Proportion of aberrant cells, \hat{p}_i
A	1	0	400	3	0.0075
	2	20	200	5	0.025
	3	100	200	14	0.070
	4	200	200	4	0.020
B	1	0	400	5	0.0125
	2	62.5	200	2	0.010
	3	125	200	2	0.010
	4	250	200	4	0.020
	5	500	200	7	0.035

An objective of the statistical analysis is to provide a measure of the evidence for treatment differences. As in Williams' paper, we here use the p-value of a test as a measure of the evidence for the differences between the treatments. The p-values of all tests considered in the preceding sections

are given in Table 2b for assays A and B. To compute the p-values of the PB test, we used simulation consisting of 1,000,000 runs. The p-values of the C-test reported in Williams are 0.00014 for assay A, and 0.239 for assay B, which are in agreement with those in Table 2b. For assay A, the p-values of the tests are close to each others. For assay B, the p-values of the C-test, CM-test and PB-test are not appreciably different while the p-value of the χ^2 -test is 0.061 which is quite different from others.

Table 2b. P-values for the chromosome assay data in Table 2a

	Assay A	Assay B
Method	p-value	p-value
χ^2 -test	0.00011	0.061
C-test	0.00014	0.236
CM-test	0.00013	0.227
PB-test	0.00016	0.220

5. Conclusion

We showed that the AU-test, that is known to produce very satisfactory results even for small samples for comparing two proportions, performs well for comparing more than two proportions also. Even though this test is not so simple as the χ^2 -test, its p-values can be easily computed when the number of proportions to be compared is small and sample size is not too large. In particular, the AU-test is useful in situations where large samples are difficult to obtain. We also demonstrated in Table 1 that the results based on the PB approach and the expression (9) are the same. So one could use the PB approach to compute the p-value of the AU-test regardless of the values of m and sample sizes.

We showed that, unlike the two-sample case, the χ^2 -test controls the size satisfactorily provided sample sizes are moderate, and it offers improved performance when $m = 4$. Computational difficulties of Type I error rates even for comparing two proportions have been pointed out by Storer and Kim (1990). As we here compare several proportions, we found computation of exact sizes of the tests for $m \geq 4$ was laborious and very time consuming, and so we did not study the properties of the tests when $m \geq 5$. However, we have no reason for the tests to behave differently for $m \geq 5$.

References

- Anscombe, F. J. (1981). *Computing in Statistical Science through APL*, Springer: New York.
- Barnard, G. A. (1945). A new test for 2×2 tables. *Nature*. **156**, 177, 783–784.
- (1947). Significance tests for 2×2 tables. *Biometrika*. **34**, 123–138.
- Berger, R. L. (1996). More powerful tests from confidence interval p values. *The American Statistician*. **50**, 314–318.
- Cai, Y. and Krishnamoorthy, K. (2006). Exact size and power properties of five tests for multinomial proportions. *Communications in Statistics - Simulation and Computation*. **35**, 449–460.
- Chan, I. S. F. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics*. **55**, 1202–1209.
- Frank, W. E. (1986). P-values for discrete distributions. *Biometrical Journal*. **20**, 482–491.
- Haber, M. (1986). An exact unconditional test for the 2×2 comparative trial. *Psychological Bulletin*. **99**, 129–132.

- Krishnamoorthy, K. and Thomson, J. (2002). Hypothesis testing about proportions in two finite populations. *The American Statistician*. **56**, 215–222.
- (2004). A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*. **119**, 23–35.
- Krishnamoorthy, K., Thomson J. and Cai Y (2004). An exact method for testing equality of several binomial proportions to a specified standard. *Computational Statistics and Data Analysis*. **45**, 697–707.
- Kulkarni, P. M. and Shah, A. K. (1995). Testing the equality of several binomial proportions to a prespecified standard. *Statistics & Probability Letters*. **25**, 213–219.
- Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*. **56**, 223–234.
- Martin, A. A. and Silva, M. A. (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational Statistics and Data Analysis*. **17**, 555–574.
- Martin, A. A. and Sanchez, Q. M. J. and Silva, M. A. (1998). Fisher's mid-P-value arrangement in 2×2 comparative trials. *Computational statistics and data analysis*. **29**, 107–113.
- (2002). Asymptotical tests in 2×2 comparative trials (unconditional approach). *Computational Statistics and Data Analysis*. **40**:339–354.
- Storer, B. E. and Kim, C. (1990). Exact properties of some exact test statistics comparing two binomial proportions. *Journal of the American Statistical Association*. **85**, 146–155.
- Scheaffer, R. L. and McClave, J. T. (1994). *Probability and Statistics for Engineers*. Duxbury Press: Pacific Grove, CA.
- Upton, G. J. G. (1982). A Comparison of alternative tests for the 2×2 comparative trial. *Journal of the Royal Statistical Society, Ser. A*. **145**, 86–105.
- Williams, D. A. (1988). Test for differences between several small proportions. *Applied Statistics*. **37**, 421–434.
- Zar, J. H. (1999). *Biostatistical Analysis*. Prentice Hall: New York.