


A Data-Based Adjustment for Fisher Exact Test

Guolong Zhao*, Huiyu Yang, Junxia Yang, Liufeng Zhang, Xiaolang Yang

*Henan Institute of Medical Sciences, Henan Academy of Medical and Pharmaceutical Sciences,
40 University Road, Zhengzhou, Henan, 450052, China*

*Correspondence: zhaogzu@hotmail.com

ABSTRACT. Fisher exact test is one of most popularly used methods in modern data analyses. However, it is conservative because of discreteness. The mid-p method may reduce the conservativeness but it is defined by the factor $\frac{1}{2}$, an extra term beyond data. This paper considers an adjustment defined by a data-based factor. The adjusted test is compared with other ten tests. Special attention is given to the comparison between the data-based factor and the factor $\frac{1}{2}$. The standardized version of the adjusted test is asymptotically standard normal. The adjustment reduces the conservativeness, as evidenced by increasing test size and power and decreasing p-values. The adjusted test holds such properties as the significance level under control of nominal α , the same modification in the left- and right-sided p-values, and the proportional reduction from Fisher test, which the mid-p method lacks. The mid-p method is more powerful than the adjusted test but the increment of power comes from the factor $\frac{1}{2}$ and is not controlled by α . The unconditional tests are also more powerful but the power comes partly from the unobserved samples. The proper choice of an adjustment is based largely upon a consideration of both the power of test and the origin of power so that the adjusted test is an option in data analyses. It is easy to implement for 2×2 and $r \times c$ contingency tables. Two real examples are given for analyzing 2×2 tables and another example for $r \times c$ tables.

1. INTRODUCTION

Comparison of two independent binomial proportions occurs most frequently in statistical analysis. Fisher exact test (Fisher, 1922; Fisher, 1970; Agresti, 1992) is often the basic requirement. It finds a number of different applications: Although in practice it is employed in the analysis of 2×2 contingency tables when sample sizes are small, it is valid for all sample sizes. Fisher test must be used if the p-value obtained by the chi-squared test is around the significance level, say, 0.05.

Received: 21 Sep 2021.

Key words and phrases. adjustment; Barnard exact test; conservativeness; contingency table; Fisher exact test; mid-p method.

Exact methods guarantee that the size of a hypothesis test is no higher than the nominal level. Subsection 4.2 will give details about the size of test. Conditioning on both sets of marginal totals, it provides a simple way to eliminate nuisance parameters in a variety of problems. Yates (1984) mentioned that tests for independence in a 2×2 table must be conditioned on both margins.

Fisher test requires extensive computations, which once hindered its use in practice. That difficulty does not exist any more however. Nowadays, computers can often implement Fisher test in a few seconds. Now it has often been employed not only in 2×2 tables but also in $r \times c$ tables, even in $2 \times 2 \times k$ and $r \times c \times k$ tables. We will return to this topic in Section 6.

It seems that applied statisticians have still favored Fisher test. However, it is conservative in the sense of its actual test size being lower than the nominal level (see, for example, Berkson, 1978; Haviland, 1990; Crans & Shuster, 2008). Efforts have been taken to reduce the conservativeness. It is of great concern, however, that the properties of Fisher test still hold while reducing the conservativeness. This is not the case in the mid- p method. For details, see Section 5.

This paper considers a data-based adjustment along the following line of thinking. We begin with the conservativeness of Fisher test that is known to be attributable to the discreteness. An intuitive display for the discreteness is shown as the non-exclusivity of the left- and right-sided p -values. The adjustment is just intended to offset the non-exclusivity. An equation is given as a fraction of the table probability plus the more extreme probabilities. It is solved for the fraction under the assumption that the left- and right-sided p -values reduce proportionally to their sum equal to 1. The solution is a data-based factor, which is further converted to the adjustment. Next comes a presentation of properties of the adjusted test such as asymptotic normality, p -values, actual and observed test size, and actual and observed power. The data-based adjustment is also built into $r \times c$ tables. Two real examples are given for analyzing 2×2 tables and another example for $r \times c$ tables. The adjusted test is compared with other ten tests. Section 2 will give a brief overview for these tests.

2. A BRIEF REVIEW OF LITERATURE

Before the development of statistical softwares, statistical inference for contingency tables has relied on large-sample approximations. The most long-standing is Pearson chi-squared test (Fleiss, 1981). It is constructed from a sum of squared errors, or through the sample variance. An often used form is its square root, the two-proportion z test. It is the most powerful test among the ten tests, which will be explained in Subsection 4.3. However it may underestimate true p -values because of discreteness. Often it becomes necessary to use Yates continuity correction (Yates, 1934), which adjusts the formula for Pearson chi-squared test by subtracting 0.5 from the difference between each observed value and its expected value in a 2×2 table based on Euler-Maclaurin theorem. The correction is widely adopted but it may tend to overcorrect. Other corrections are available, for example, Kendal-Stuart correction (Conover, 1974; Haber, 1980). It is the arithmetic average of a

chi-squared statistic with its next smaller possible value. Although it has so far received relatively little attention, we will still include it in our calculations and simulations in Subsection 4.2 and 4.3. It is widely recognized that large-sample approximations can be very poor when the contingency table contains both small and large expected frequencies (Agresti, 1992). Fisher (1925) gave a vivid description that "... the traditional machinery of statistical processes is wholly unsuited to the needs of practical research."

Alternative exact tests had been developed, for example, unconditional tests (Barnard, 1945; 1947; Haber, 1986; Suissa & Shuster, 1985; Berger & Boos, 1994; Routledge, 1992; Lin & Yang, 2009). Barnard exact test considers all possible values of the nuisance parameter(s) and chooses the value(s) that maximizes the p-value. By contrast, Fisher test avoids estimating the nuisance parameter(s) by conditioning on the margins. Barnard test relaxes this constraint on one set of the marginal totals. Berger and Boos (1994) took the supremum for the p-value over a confidence interval of values for the nuisance parameter rather than over all possible values. Berger-Boos test will also be included in all calculations and simulations.

It remains unclear, however, which test statistic is preferred to define the critical region when implementing Barnard test (Wikipedia, the free encyclopedia). The difficulty lies in the fact that the choice of test statistics influences a decision. For example, the binomial model has the critical region defined by the two-proportion z test (Routledge, 1992) and the modified Fisher p-value (Lin & Yang, 2009) by Fisher test. There are some other arguments on conservativeness. An example is what the conservativeness is ascribed to the common practice of fixing the nominal level, say, at 0.05 (Upton, 1992). It is not possible to correct Fisher test without also increasing the true α -level (Berger, 2000). This is implemented in Crans-Shuster test (Crans & Shuster, 2008), which defines an increment of significance level based on unconditional approach with the critical region defined by Fisher test. More information and details about the increment will be given in Subsection 4.2.

Unconditional tests are more powerful than Fisher test (Lydersen, Fagerland & Laake, 2009), but they are not at all commonly used up to date. This is because considerable controversy surrounds their use in statistical literature (see, for example, Agresti, 2001; Agresti, 2002, p95; Cheng, Liou, Aston & Tsai, 2008). Fisher criticized the unconditional approach, arguing that possible samples with quite different numbers of successes than observed were not relevant. In plain words, the unconditional tests require not only the observed sample at hand but also the unobserved samples in statistical inferences. Some other statisticians have argued that the unconditional approach is artificial because it averages what happened in the observed sample with hypothetical response distributions, some of which are much different than observed (Agresti, 2001). Obviously, it has the same meaning as the Fisher argument. A concern is that the power of unconditional tests comes partly from the unobserved samples. It is noted that some authors including Barnard himself refuted Barnard test in favor of Fisher test (Agresti, 2002, p95). As for Crans-Shuster test,

it has so far not gained wide use since fixed significance levels are the standard in real-world applications (Crans & Shuster, 2008).

To this end one may find a way to compensate the discreteness. A randomization test was proposed based on the p -value of Fisher test (Tocher, 1950). Nevertheless, this post hoc test has theoretical interests only and has not been accepted widely in a practical setting (Hirji, Tan & Elashoff, 1991; Liddle, 1976; Mantel & Greenhouse, 1968).

For highly discrete data when large-sample methods are questionable but exact methods may be conservative, one could alternatively use adjustments of exact methods based on the mid- p method (see, for example, Lancaster, 1961; Hwang & Yang, 2001). The mid- p -value is defined as half the conditional probability of the observed statistic plus the conditional probability of more extreme values, given the marginal totals. Thus the mid- p -value is less than the ordinary p -value by half the probability of the observed result. In one view, it has nice properties in terms of Type I error and power and so is recommended by leading statisticians (see, for example, Hirji, Tan & Elashoff, 1991; Routledge, 1992; Agresti, 2001; Agresti, 2002, p21; Lydersen, Fagerland & Laake, 2009). In the other view, a relevant concern is that it is a non-randomized version of Fisher test (Hirji, Tan & Elashoff, 1991). For example, SISA (Simple Interactive Statistical Analysis <http://www.quantitativeskills.com/sisa/>) does not recommend the use of mid- p values. In addition, it is defined by the factor $\frac{1}{2}$, an extra term beyond data, which raises some more concerns. Further details will be given in Section 5. The mid- p method is more powerful than Fisher test but the increment of power comes from the factor $\frac{1}{2}$.

Controversy continues about the appropriateness of some exact methods, however, there is still no consensus (Agresti, 2001). Thus the work continues on the development of the adjustments.

3. DERIVING AN ADJUSTMENT

3.1 Fisher Exact Test and the Conservativeness

Consider the situation in which $Y_j, j = 1, 2$, represents two independent binomial observations with parameters (n_j, π_j) . It follows that the total sample size $n = n_1 + n_2$ with the sample fraction $k_j = n_j/n$, the difference of proportions $\mu = \pi_1 - \pi_2$, and the average $\bar{\pi} = k_1\pi_1 + k_2\pi_2$. The total frequency is $M = Y_1 + Y_2$ with the observed data $M = m = y_1 + y_2$.

Given the parameters n_1, n_2 , and m , the hypergeometric probability density function (pdf) is

$$f(y_1 = t) = \binom{n_1}{t} \binom{n_2}{m-t} / \binom{n}{m},$$

where $\binom{a}{b} = a!/b!(a-b)!$, $\xi_- \leq t \leq \xi_+$, $\xi_- = \max(0, m - n_2)$, and $\xi_+ = \min(n_1, m)$. Then we have $\sum_{\xi_- \leq t \leq \xi_+} f(t) = 1$. The left- and right-sided p -values of Fisher test are given by

$$F_F(y_1) = \sum_{\xi_- \leq t \leq y_1} f(t) \text{ and } S_F(y_1) = \sum_{y_1 \leq t \leq \xi_+} f(t). \quad (3.1)$$

To conduct a two-sided test, a popular approach (Agresti, 1992) is

$$P_F(1) = \sum_{\xi_- \leq t \leq \xi_+} f(t) | (f(t) \leq f(y_1)), \quad (3.2)$$

where $f(y_1)$ is the table probability. It is of interest to see the left and right components of a two-sided p-value. Thus the range of t is divided into the left $[\xi_-, t_{max}]$ and right half $[t_{max}, \xi_+]$, where $t_{max} = t | \max(f(t), t \in [\xi_-, \xi_+])$. Then we define the position of y_1 as

$$y_1 = t | \max(f(t) | (f(t) \leq f(y_1)), t \in [\xi_-, t_{max}])$$

when $y_1 \in [\xi_-, t_{max}]$. Likewise, the position of its opposite point is defined as

$$y_1^* = t | \max(f(t) | (f(t) \leq f(y_1)), t \in [t_{max}, \xi_+]).$$

With both y_1 and y_1^* , (3.2) can be rewritten in the form

$$P_F(1) = \sum_{\xi_- \leq t \leq y_1} f(t) + \sum_{y_1^* \leq t \leq \xi_+} f(t),$$

which shows the left and right components.

It is worth noting that mistakes may occur if ignoring the asymmetrical two-sided p-value when $n_1 \neq n_2$. In this case, the right component can differ from the left substantially. Refer to the example in Pearson (1947) (see Subsection 4.1). The sample sizes are $\{n_1, n_2\} = \{12, 8\}$ and the total frequency is $m = 7$. Given $y_1 = 2$, we have the two-sided p-value $P_F(1) = 0.062$ with the left component 0.052 and the right 0.01. When $n_1 = n_2$, y_1^* is the mirror image of y_1 and so the two components are equal. Suppose the sample sizes are $\{n_1, n_2\} = \{10, 10\}$, the two-sided p-value becomes $P_F(1) = 0.35$ with the left component 0.175 and the right 0.175. Recalling the forms seen in Agresti (2002, p93), they are fit only for symmetrical two-sided p-values when $n_1 = n_2$.

In this way, another possibility of two-sided test (Agresti, 2002, p93) is expressed as

$$P_F(2) = \sum_{\xi_- \leq t \leq E[t]} f(t) | (t - E[t] \leq y_1 - E[t]) + \sum_{E[t] \leq t \leq \xi_+} f(t) | (t - E[t] \geq y_1^* - E[t]),$$

where $E[t] = mn_1/n$. This approach takes $E[t]$ as the boundary of two halves instead of t_{max} . Differences may occur when $E[t] \neq t_{max}$, which will be seen in the fish experiment (Routledge, 1992) in Subsection 7.1. Similarly, the fourth approach (Agresti, 2002, p93) is given by

$$P_F(3) = \min \left(\sum_{\xi_- \leq t \leq \xi_+} f(t) | (t \leq y_1), \sum_{\xi_- \leq t \leq \xi_+} f(t) | (t \geq y_1) \right) \\ + \min \left(\sum_{\xi_- \leq t \leq \xi_+} f(t) | (t \leq y_1^*), \sum_{\xi_- \leq t \leq \xi_+} f(t) | (t \geq y_1^*) \right).$$

When $n_1 = n_2$, it simplifies to

$$P_F(3) = 2 \min \left(\sum_{\xi_- \leq t \leq \xi_+} f(t) | (t \leq y_1), \sum_{\xi_- \leq t \leq \xi_+} f(t) | (t \geq y_1) \right),$$

which happens to be the third approach (Agresti, 2002, p93). Moreover, Dupont (1986) investigated the advantages of doubling the one-sided p-value in conducting a two-sided test:

$$P_F(4) = 2F_F(y_1).$$

To see the conservativeness intuitively, we will use the following procedure: The formula (3.1) specifies that $F_F(y_1)$ and $S_F(y_1)$ are the sums of elements in the sets

$$A_F = \{f(t) | \xi_- \leq t \leq y_1\} \text{ and } A_S = \{f(t) | y_1 \leq t \leq \xi_+\}.$$

The two sets have the union and the intersection

$$A_{F \cup S} = \{f(t) | \xi_- \leq t \leq \xi_+\} \text{ and } A_{F \cap S} = \{f(t) | t = y_1\}.$$

The elements in $A_{F \cup S}$ sum to 1 and in $A_{F \cap S}$ to $f(y_1)$. The sets are related by $A_{F \cup S} = A_F + A_S - A_{F \cap S}$ and the sums by $1 = F_F(y_1) + S_F(y_1) - f(y_1)$ or

$$F_F(y_1) + S_F(y_1) = 1 + f(y_1).$$

This signifies the non-exclusivity when the sum $1 + f(y_1) > 1$, which is an intuitive display of the discreteness. It indicates that the small sample effect may overestimate p-values, meaning the conservativeness.

In a continuous distribution, inclusion or exclusion of the observed point is immaterial so that the left- and right-sided p-values sum to 1. In a discrete distribution, inclusions lead to $F_F(y_1) + S_F(y_1) \geq 1$ (Hirji, Tan, and Elashoff, 1991). Observe what happens to $1 + f(y_1)$ as n increases. It returns the maximum of 2 when $\{n_1, n_2\} = \{1, 0\}$ and $\{y_1, y_2\} = \{1, 0\}$. In the fish experiment, where $n = 6$, we have $1 + f(y_1) = 1.05$. When the sample size magnifies three times, i.e., $n = 18$, we have $1 + f(y_1) = 1.000021$. As n increases indefinitely, $1 + f(y_1)$ approaches nearer and nearer to the minimum of 1. In words, the sum is the measure in the continuous-discrete classification: 2, 1.05, 1.000021, ... refer to discreteness and 1 refers to continuousness. The problem at hand is how to offset the non-exclusivity or how to make the left- and right-sided p-values summing to 1 when sample sizes are limited.

3.2 A Data-based Adjustment

In doing so, an equation is given as a fraction of the table probability plus the more extreme probabilities:

$$F(y_1) = W f(y_1) + P(t < y_1) \text{ and } S(y_1) = (1 - W)f(y_1) + P(t > y_1), \quad (3.3)$$

where $W \in [0, 1]$ is the fraction, $P(t < y_1) = \sum_{\xi_- \leq t < y_1} f(t)$, and $P(t > y_1) = \sum_{y_1 < t \leq \xi_+} f(t)$. Now the equation (3.3) indicates

$$F(y_1) + S(y_1) = 1 \text{ and } S(y_1) = 1 - F(y_1).$$

Stevens (1950) proposed $W = U$, where U is a uniform (0,1) random number and then (3.3) represents the randomized p-value. While the randomized p-value has theoretical interests only, one may turn to the expected value of U (Agresti 2002, p27), i.e., $W = \frac{1}{2}$. Then (3.3) becomes the mid-p method. In another context, the fraction W is regarded as the weight in the weighted average of the two probabilities obtained by inclusion and exclusion of the observed point:

$$W[f(y_1) + P(t < y_1)] + (1 - W)[0 + P(t < y_1)] = W f(y_1) + P(t < y_1).$$

The weight is also known as $W = \frac{1}{2}$ in the mid-p method (Hirji, Tan & Elashoff, 1991).

No matter which mechanism defines it, the factor $\frac{1}{2}$ is an extra term beyond data. But one golden aphorism is clear: Estimation, hypothesis testing, and inference, in general, are based on the data at hand (Insightful Corporation, 2007, p1). This raises two questions: (1) Is $W = \frac{1}{2}$ justified? (2) What does the fraction W equal given a data set? Question (1) will be considered later in Section 5. Question (2) is solved now.

In view of (3.1), the whole table probability is added to both the left- and right-sided p-values of Fisher test:

$$F_F(y_1) = f(y_1) + P(t < y_1) \text{ and } S_F(y_1) = f(y_1) + P(t > y_1) \quad (3.4)$$

so that their sum may be greater than 1. This predicts a reducing process from Fisher test to an adjusted test. It seems only reasonable that the left- and right-sided p-values reduce proportionally to their sum equal to 1:

$$\frac{F_F(y_1)}{S_F(y_1)} = \frac{F(y_1)}{S(y_1)} = \frac{F(y_1)}{1 - F(y_1)}, \quad (3.5)$$

where the right-hand side shows the numerator and denominator summing to 1. Evidently, there is no reason to expect a disproportional reduction. Note that the equality of two ratios in (3.5) does not imply $F_F(y_1) = F(y_1)$ and $S_F(y_1) = S(y_1)$. Converting (3.3) into

$$P(t < y_1) = F(y_1) - W f(y_1) \text{ and } P(t > y_1) = S(y_1) - (1 - W)f(y_1)$$

and putting it into (3.4) produces

$$F_F(y_1) = f(y_1) + F(y_1) - W f(y_1) \text{ and } S_F(y_1) = f(y_1) + S(y_1) - (1 - W)f(y_1)$$

or

$$F_F(y_1) = (1 - W)f(y_1) + F(y_1) \text{ and } S_F(y_1) = W f(y_1) + S(y_1). \quad (3.6)$$

Substituting (3.6) into (3.5) yields

$$\frac{(1 - W)f(y_1) + F(y_1)}{W f(y_1) + S(y_1)} = \frac{F(y_1)}{1 - F(y_1)}$$

or

$$\frac{f(y_1) - W f(y_1) + F(y_1)}{W f(y_1) + 1 - F(y_1)} = \frac{F(y_1)}{1 - F(y_1)}.$$

Taking a crossing multiplication

$$(f(y_1) - W f(y_1) + F(y_1))(1 - F(y_1)) = (W f(y_1) + 1 - F(y_1))F(y_1)$$

gives

$$f(y_1) - W f(y_1) + F(y_1) - f(y_1)F(y_1) + W f(y_1)F(y_1) - F^2(y_1) = W f(y_1)F(y_1) + F(y_1) - F^2(y_1).$$

All terms on the right-hand side are also found on the left and so cancel out:

$$f(y_1) - W f(y_1) - f(y_1)F(y_1) = 0.$$

Finally, we gain such a solution

$$W = 1 - F(y_1) \text{ and } W = S(y_1).$$

So now this is the answer to Question (2). The solution is just the data-based factor.

Now put the data-based factor back into (3.3); then we have

$$F(y_1) = (1 - F(y_1))f(y_1) + P(t < y_1) \text{ and } S(y_1) = (1 - S(y_1))f(y_1) + P(t > y_1). \quad (3.7)$$

This is just the adjusted test, which is comparable with the mid-p method in formula expressions. The calculation requires iteration, however. With initial values of $F(y_1)$ and $S(y_1)$, say, 0.5, adequate convergence usually takes three or four iterations. In addition, putting the data-based factor into (3.6) results in

$$F_F(y_1) = F(y_1) f(y_1) + F(y_1) \text{ and } S_F(y_1) = S(y_1) f(y_1) + S(y_1).$$

With a little arrangement

$$F_F(y_1) = F(y_1)(1 + f(y_1)) \text{ and } S_F(y_1) = S(y_1)(1 + f(y_1)),$$

we obtain

$$F(y_1) = F_F(y_1)/(1 + f(y_1)) \text{ and } S(y_1) = S_F(y_1)/(1 + f(y_1)).$$

Substituting (3.1) into it produces

$$F(y_1) = \sum_{\xi_- \leq t \leq y_1} \frac{f(t)}{1 + f(y_1)} \text{ and } S(y_1) = \sum_{y_1 \leq t \leq \xi_+} \frac{f(t)}{1 + f(y_1)}. \quad (3.8)$$

This is the applied form of the adjusted test. It does not require iteration but its result is identical to that of (3.7). The denominator $(1 + f(y_1))^{-1}$ is just the data-based adjustment.

A comparison of (3.8) and (3.1) provides an insight into the mechanism of the adjustment: using $f(t)/(1 + f(y_1))$ in place of $f(t)$. With this mechanism, the adjustment is used easily in two-sided test. Let $P(1)$, $P(2)$, $P(3)$, and $P(4)$ denote the adjusted two-sided p-values corresponding to

$P_F(1)$, $P_F(2)$, $P_F(3)$, and $P_F(4)$, respectively. Applying the adjustment for the popular approach (3.2) produces

$$P(1) = \sum_{\xi_- \leq t \leq \xi_+} f(t)/(1 + f(y_1)) | (f(t) \leq f(y_1)) = P_F(1)/(1 + f(y_1)). \quad (3.9)$$

The adjusted version of another possibility is given by

$$P(2) = \frac{\sum_{\xi_- \leq t \leq E[t]} f(t) | (t - E[t] \leq y_1 - E[t]) + \sum_{E[t] \leq t \leq \xi_+} f(t) | (t - E[t] \geq y_1^* - E[t])}{1 + f(y_1)}$$

or

$$P(2) = P_F(2)/(1 + f(y_1)).$$

The fourth approach and Dupont approach have the adjusted versions

$$P(3) = P_F(3)/(1 + f(y_1))$$

and

$$P(4) = P_F(4)/(1 + f(y_1)) = 2F(y_1),$$

respectively.

4. PROPERTIES OF THE ADJUSTED TEST

Before seeing properties of the adjusted test, it is good to remember that:

Remark 4.1. *The adjustment $(1 + f(y_1))^{-1}$ is the reciprocal of the non-exclusivity $1 + f(y_1)$ and so it offsets the non-exclusivity. Consequently, the two one-sided p -values become mutually exclusive so that we have $S(y_1) = 1 - F(y_1)$, a property of continuous distributions.*

Remark 4.2. *The adjustment attains its minimum of $(1 + f(y_1))^{-1} = 0.5$ when $\{n_1, n_2\} = \{1, 0\}$ and $\{y_1, y_2\} = \{1, 0\}$. In this situation, Fisher test produces $F_F(y_1) = 1$, $S_F(y_1) = 1$, and $P_F(1) = 2$. It is inexplicable to see a two-sided p -value of 2. The adjusted test gives a favourable turn with the interpretable results $f(y_1) = 1$, $F(y_1) = 0.5$, $S(y_1) = 0.5$, and $P(1) = 1$. It seems as if Fisher test must be accompanied by the adjustment for it to be perfect. The adjustment reaches its maximum of $(1 + f(y_1))^{-1} = 1$ as $n \rightarrow \infty$. This signifies elimination of the non-exclusivity and so the adjustment vanishes into void.*

4.1 Assessing Asymptotic Normality

A standardized version of Fisher test is asymptotically standard normal under H_0 (see, for example, Pearson, 1947 among others) and the same is true for the adjusted test. Given a data set, the adjustment $(1 + f(y_1))^{-1}$ is a constant, which defines a linear transformation from $F_F(y_1)$ to $F(y_1)$ as indicated by (3.8). Linear transformations of normal random variables are themselves normal.

From the pdf of hypergeometric distribution in Subsection 3.1, the expected value of Y_1 is defined to be $E[Y_1] = \sum_{\xi_- \leq y_1 \leq \xi_+} y_1 f(y_1)$ and the variance to be $V[Y_1] = \sum_{\xi_- \leq y_1 \leq \xi_+} \{y_1 - E[Y_1]\}^2 f(y_1)$. They are often written in the corresponding analytic forms $E[Y_1] = mn_1/n$ and

$$V[Y_1] = \frac{m(n-m)n_1n_2}{n^2(n-1)}.$$

Looking at $V[Y_1]$, it is also the variance in Mantel-Haenszel statistic when there is only one stratum, where the term $n - 1$ is the finite population correction factor. With the substitution of n for $n - 1$, the standardized value $(Y_1 - E[Y_1])/\sqrt{V[Y_1]}$ equals the statistic of two-proportion z test. This is known to be asymptotically standard normal $N(0, 1)$ by the central limit theorem. Letting $E[Y_1]$ and $V[Y_1]$ be denoted by μ and σ^2 , the corresponding pdf of normal distribution is expressed as

$$f_{Y_1}(y_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{y_1 - \mu}{\sigma} \right)^2 \right\}.$$

A numerical calculation is provided for comparing the pdf's of hypergeometric and normal distributions. The example in Pearson (1947) is used again with a program available in R language (Venables, Smith, et al., 2019). The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$ and the total frequency $m = 7$. The results are presented in the upper part of Table 4.1.

Table 4.1. *Comparing the hypergeometric and normal probabilities*

| Distributions | Frequency in treatment group Y_1 | | | | | | | |
|--|------------------------------------|--------|--------|--------|--------|--------|--------|--------|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Values of probability density function | | | | | | | | |
| Hypergeometric | 0.0001 | 0.0043 | 0.0477 | 0.1987 | 0.3576 | 0.2861 | 0.0954 | 0.0102 |
| Normal | 0.0002 | 0.0043 | 0.0453 | 0.1989 | 0.3657 | 0.2817 | 0.0909 | 0.0123 |
| Left-sided p-values of Fisher exact test | | | | | | | | |
| Hypergeometric | 0.0001 | 0.0044 | 0.0521 | 0.2508 | 0.6084 | 0.8944 | 0.9898 | 1 |
| Normal | 0.0002 | 0.0045 | 0.0498 | 0.2487 | 0.6144 | 0.8961 | 0.9870 | 0.9993 |
| Left-sided p-values of the adjusted test | | | | | | | | |
| Hypergeometric | 0.0001 | 0.0044 | 0.0497 | 0.2092 | 0.4481 | 0.6955 | 0.9036 | 0.9899 |
| Normal | 0.0002 | 0.0045 | 0.0477 | 0.2075 | 0.4499 | 0.6991 | 0.9047 | 0.9872 |

The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$ for treatment and control group, respectively. The total frequency is $m = 7$.

The left-sided p-values of Fisher test are computed by accumulating the pdf in left to right order, shown in the middle part. In this way, we have the left-sided p-values of the adjusted test by using the adjustment $(1 + f(y_1))^{-1}$ to the pdf's and the results are listed in the lower part. A comparison of the adjusted test and Fisher test is displayed in Figure 1.

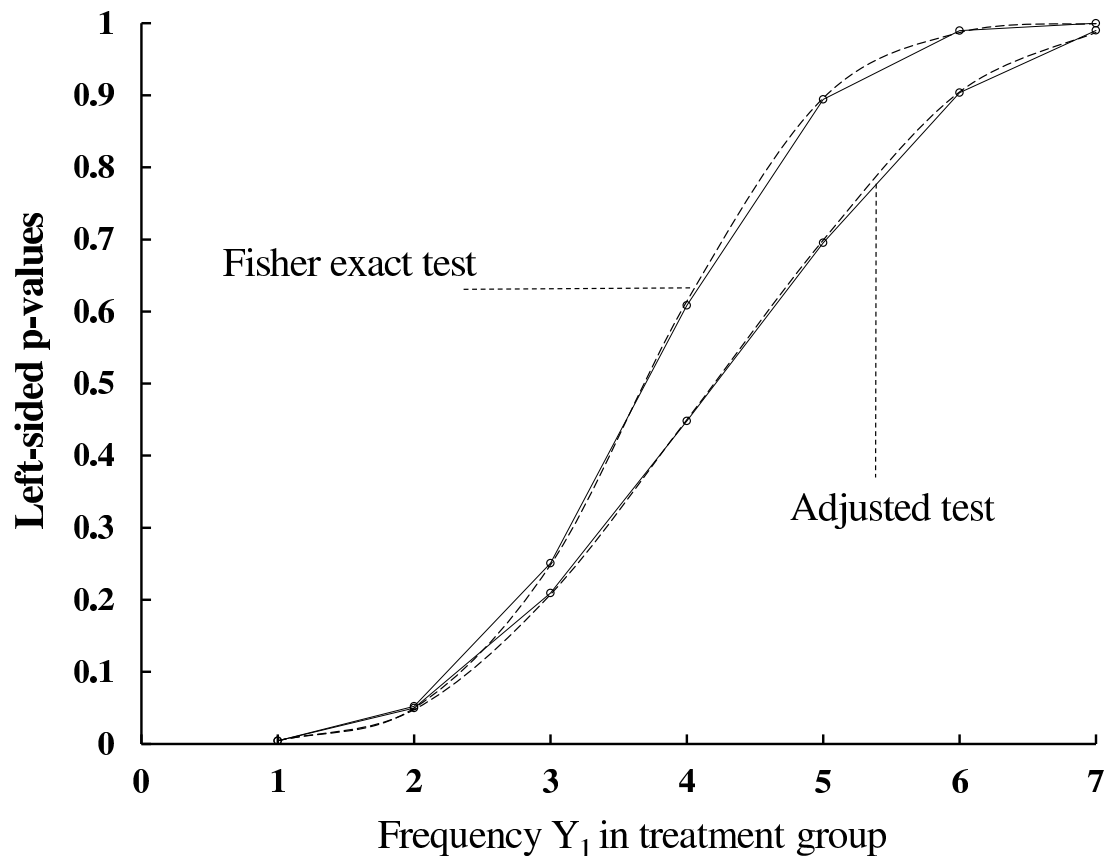


Figure 1. Left-sided p-values of the adjusted test and Fisher exact test. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$ for treatment and control group respectively. The total frequency is $m = 7$. Solid line represents the left-sided p-values calculated from the hypergeometric distributions and dash line from normal distributions.

Both the table and the figure show that the left-sided p-values of the adjusted test are less than those of Fisher test. Moreover, we have seen a satisfactory convergence of hypergeometric distribution to normality, as evidenced by the close agreement between hypergeometric and normal distributions.

Remark 4.3. *The standardized version of the adjusted test is asymptotically standard normal, which inherits from Fisher test. The asymptotic normality is a premise for building the size and power of the adjusted test.*

4.2 Actual and Observed Test Size

The size of test is often the first consideration when conducting a test. To calculate actual test size, we want to use the exact conditional method but this is the power function of Fisher test (Haseman, 1978; Casagrande et al., 1978). Therefore, we take up an extension for the control of

actual test size. A further extension provides a scope to cover other tests. The method is set up on the sum $m = y_1 + y_2$ and the odds ratio $\psi = \pi_1(1 - \pi_2)/\pi_2/(1 - \pi_1)$. Here $\pi_1 = \pi_2 = \pi$ and $\psi = 1$ under H_0 and $\pi_1 \geq \pi_2$ and $\psi \geq 1$ under $H_0 \cup H_1$ over the range of $\mu \in [0, 1)$. The conditional density function of y_1 is given by

$$f(y_1|m, \psi) = \frac{\binom{n_1}{y_1} \binom{n_2}{m-y_1} \psi^{y_1}}{\sum_{\xi_- \leq t \leq \xi_+} \binom{n_1}{t} \binom{n_2}{m-t} \psi^t}$$

in terms of the non-central hypergeometric distribution. The expression must give $\sum_{\xi_- \leq y_1 \leq \xi_+} f(y_1|m, \psi) = 1$. It reduces to the central hypergeometric distribution $f(y_1|m, 1) = \binom{n_1}{y_1} \binom{n_2}{m-y_1} / \binom{n}{m}$ when the null hypothesis is true with $\psi = 1$.

Let y_{1c} be the critical value of y_1 at α -level under H_0 . It can be found by

$$\sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1) \leq \alpha \quad (4.1)$$

and $\sum_{y_{1c}-1 \leq y_1 \leq \xi_+} f(y_1|m, 1) > \alpha$. The left-hand side of (4.1) is just the the right-sided p-value of Fisher test in this context.

With the critical value, it is easy to see the conditional test size $g(1|m) = \sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1)$. The actual test size can then be determined from $g(1|m)$ by taking the supremum over π :

$$g(1, \pi) = \sup_{0 \leq \pi \leq 1} \sum_{1 \leq m \leq n-1} g(1|m) P(m, \pi), \quad (4.2)$$

where $P(m, \pi)$ is the joint distribution of m under H_0 ,

$$P(m, \pi) = \sum_{\xi_- \leq t \leq \xi_+} \binom{n_1}{t} \binom{n_2}{m-t} \pi^m (1 - \pi)^{n-m}.$$

Refer to (3.6) with $W = \frac{1}{2}$ to get

$$F_L(y_1) = F_F(y_1) - \frac{1}{2} f(y_1) \text{ and } S_L(y_1) = S_F(y_1) - \frac{1}{2} f(y_1), \quad (4.3)$$

which is the mid-p method. It states that the mid-p value is Fisher p-value minus half the table probability. Following (4.1), the right-sided p-value of (4.3) is rewritten $\sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1) - \frac{1}{2} f(y_{1c}|m, 1) \leq \alpha$ in this context. Moving the term $\frac{1}{2} f(y_{1c}|m, 1)$ to the right-hand side gives

$$\sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1) \leq \alpha + \frac{1}{2} f(y_{1c}|m, 1). \quad (4.4)$$

It is a heuristic that the decrement of p-values is equivalent to the increment of significance level. Calculating (4.4) requires iteration. Giving an initial value of $f(y_{1c}|m, 1)$, say, 0.5 produces a value of y_{1c} and further a value of $f(y_{1c}|m, 1)$. Repeating the process yields a stationary value of y_{1c} with fast convergence.

Crans-Shuster test uses the significance level plus an increment so that the value of y_{1c} is given by

$$\sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1) \leq \alpha + \varepsilon, \quad (4.5)$$

where ε is known to be the increment. It is calculated with

$$\varepsilon = \inf\{\varepsilon^s : \sup[(g(1, \pi)|\alpha + \varepsilon^s) \leq \alpha]\}, \quad (4.6)$$

where ε^s represents the addend for the gradually increasing significance level.

Here the right-sided p-value of the adjusted test (3.8) is written $\sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1)/(1 + f(y_{1c}|m, 1)) \leq \alpha$ or, in a form more convenient for our present purpose,

$$\sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, 1) \leq \alpha + \alpha f(y_{1c}|m, 1). \quad (4.7)$$

The term $\alpha f(y_{1c}|m, 1)$ is the increment. The calculation of (4.7) requires iteration likewise for (4.4). Now the method may cover the mid-p method, Crans-Shuster test, and the adjusted test as long as using (4.4), (4.5), and (4.7) in place of (4.1).

We show a numerical analysis to look into actual test size by taking the supremum over $0 \leq \pi \leq 1$. A sample of size $n = 20$ is taken with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. That is unequal allocation, which is known to be best. The nominal level of significance is $\alpha = 0, 0.0125, \dots, 0.1$ (one-sided). The increment for Crans-Shuster test is $\varepsilon = 0, 0.0124, \dots, 0.0570$ calculated using (4.6). Shown in the upper part of Table 4.2 are the findings from the exact conditional method with the extensions.

Table 4.2. *Actual test size of the adjusted test and the other tests*

| Tests | Nominal α (one-sided) | | | | | | | | |
|--|------------------------------|--------|-------|--------|-------|--------|-------|--------|-------|
| | 0 | 0.0125 | 0.025 | 0.0375 | 0.05 | 0.0625 | 0.075 | 0.0875 | 0.1 |
| Exact conditional method with the extensions | | | | | | | | | |
| Adjusted test | 0 | 0.004 | 0.012 | 0.012 | 0.025 | 0.033 | 0.033 | 0.042 | 0.057 |
| Fisher exact test | 0 | 0.004 | 0.012 | 0.012 | 0.018 | 0.033 | 0.033 | 0.042 | 0.042 |
| Mid-p method | 0 | 0.009 | 0.018 | 0.033 | 0.042 | 0.057 | 0.061 | 0.096 | 0.096 |
| Crans-Shuster test | 0 | 0.012 | 0.025 | 0.033 | 0.042 | 0.061 | 0.074 | 0.075 | 0.096 |
| Algorithm of Crans and Shuster with the extensions | | | | | | | | | |
| Adjusted test | 0 | 0.004 | 0.012 | 0.012 | 0.025 | 0.033 | 0.033 | 0.042 | 0.057 |
| Fisher exact test | 0 | 0.004 | 0.012 | 0.012 | 0.018 | 0.033 | 0.033 | 0.042 | 0.042 |
| Mid-p method | 0 | 0.009 | 0.018 | 0.033 | 0.042 | 0.057 | 0.061 | 0.096 | 0.096 |
| Crans-Shuster test | 0 | 0.012 | 0.025 | 0.033 | 0.042 | 0.061 | 0.074 | 0.075 | 0.096 |
| Binomial model | 0 | 0.011 | 0.017 | 0.033 | 0.042 | 0.061 | 0.061 | 0.082 | 0.095 |
| Barnard exact test | 0 | 0.011 | 0.017 | 0.033 | 0.042 | 0.061 | 0.061 | 0.082 | 0.095 |
| Berger-Boos test | 0 | 0.011 | 0.018 | 0.033 | 0.042 | 0.061 | 0.061 | 0.082 | 0.095 |
| Modified Fisher p-value | 0 | 0.011 | 0.025 | 0.033 | 0.042 | 0.061 | 0.074 | 0.074 | 0.095 |
| Two-proportion z test | 0 | 0.017 | 0.033 | 0.057 | 0.061 | 0.096 | 0.096 | 0.11 | 0.12 |
| Yates corrected z test | 0 | 0.003 | 0.009 | 0.012 | 0.018 | 0.033 | 0.033 | 0.042 | 0.042 |
| Kendal-Stuart correction | 0 | 0.004 | 0.011 | 0.017 | 0.033 | 0.033 | 0.042 | 0.042 | 0.061 |

The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. The actual test size is calculated at $\alpha = 0, 0.0125, \dots, 0.1$ (one-sided) by taking the supremum over $0 \leq \pi \leq 1$. The increment of significance level is $\varepsilon = 0, 0.0124, \dots, 0.0570$ for Crans-Shuster test.

The calculations are repeated by another method, the algorithm described in Crans and Shuster (2008). We extend it along the same line to cover the eleven tests including three conditional, five unconditional, and three approximate tests. The unconditional tests require not only the observed sample $m = y_1 + y_2$ but also the unobserved samples $[1, m) \cup (m, n - 1]$. The binomial model and Barnard test have the critical region defined by the z test and Berger-Boos test uses the confidence coefficient of 0.999. Details regarding the extensions are available from the author upon request. The actual test size of these tests appears in the lower part of the Table 4.2.

The observed test size is computed by Monte Carlo simulation with the same parameter values as those in calculating the actual size. The observations y_j were sampled from binomial distribution. In each instance, a total of 1000 sets of samples were generated. The nominal level of significance was $\alpha = 0.01, 0.025, \text{ and } 0.05$. The common binomial parameter was prescribed as $\pi = 0, 0.1, \dots, 1$. Computations are performed for the right-sided p-value of the adjusted test

$S(y_1)$. The fraction of times H_0 is rejected for the p-value is calculated. The experiment results in the observed test size

$$\hat{\alpha} = \sum_{i=1}^{1000} I\{\text{p-value} \leq \alpha | H_0\} / 1000.$$

The experiment covers the other tests as well. In the extreme case $\pi = 0$ or 1 , the observed test size is zero for any test at any level. Over the range of π , the observed size increases first and decreases later with the peak at $\pi = 0.6$ so that only the results for $\pi = 0, 0.3, 0.5$, and 0.6 are listed in Table 4.3.

Table 4.3. *Observed test size of the adjusted test and the other tests*

| Tests | $\pi = 0$ | | | $\pi = 0.3$ | | | $\pi = 0.5$ | | | $\pi = 0.6$ | | |
|--------------------------|-----------|-------|------|-------------|-------|-------|-------------|-------|-------|-------------|-------|-------|
| | 0.01 | 0.025 | 0.05 | 0.01 | 0.025 | 0.05 | 0.01 | 0.025 | 0.05 | 0.01 | 0.025 | 0.05 |
| Conditional tests | | | | | | | | | | | | |
| Adjusted test | 0 | 0 | 0 | 0.002 | 0.013 | 0.023 | 0.004 | 0.014 | 0.025 | 0.004 | 0.013 | 0.03 |
| Fisher exact test | 0 | 0 | 0 | 0.002 | 0.013 | 0.013 | 0.004 | 0.014 | 0.021 | 0.004 | 0.013 | 0.023 |
| Mid-p method | 0 | 0 | 0 | 0.002 | 0.013 | 0.031 | 0.008 | 0.021 | 0.042 | 0.011 | 0.023 | 0.045 |
| Unconditional tests | | | | | | | | | | | | |
| Crans-Shuster test | 0 | 0 | 0 | 0.002 | 0.023 | 0.045 | 0.008 | 0.025 | 0.042 | 0.011 | 0.03 | 0.045 |
| Binomial model | 0 | 0 | 0 | 0.008 | 0.013 | 0.045 | 0.009 | 0.021 | 0.042 | 0.011 | 0.023 | 0.045 |
| Barnard exact test | 0 | 0 | 0 | 0.008 | 0.013 | 0.045 | 0.009 | 0.021 | 0.042 | 0.011 | 0.023 | 0.045 |
| Berger-Boos test | 0 | 0 | 0 | 0.008 | 0.023 | 0.045 | 0.008 | 0.021 | 0.042 | 0.007 | 0.023 | 0.045 |
| Modified Fisher p-value | 0 | 0 | 0 | 0.008 | 0.023 | 0.045 | 0.009 | 0.025 | 0.042 | 0.011 | 0.03 | 0.045 |
| Approximate tests | | | | | | | | | | | | |
| Two-proportion z test | 0 | 0 | 0 | 0.013 | 0.028 | 0.063 | 0.014 | 0.033 | 0.066 | 0.013 | 0.033 | 0.069 |
| Yates corrected z test | 0 | 0 | 0 | 0.001 | 0.002 | 0.013 | 0.001 | 0.008 | 0.021 | 0.002 | 0.011 | 0.023 |
| Kendal-Stuart correction | 0 | 0 | 0 | 0.002 | 0.007 | 0.028 | 0.004 | 0.013 | 0.033 | 0.004 | 0.013 | 0.033 |

The observed test size is given by the fraction of p-values less than or equal to α under H_0 in 1000 sets of samples, where $\alpha = 0.01, 0.025, 0.05$ (one-sided). The increment of significance level is $\varepsilon = 0.0081, 0.0272, 0.0522$ for Crans-Shuster test. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$.

Appearing in Figure 2 is a comparison of the adjusted test and the other tests for the observed test size.

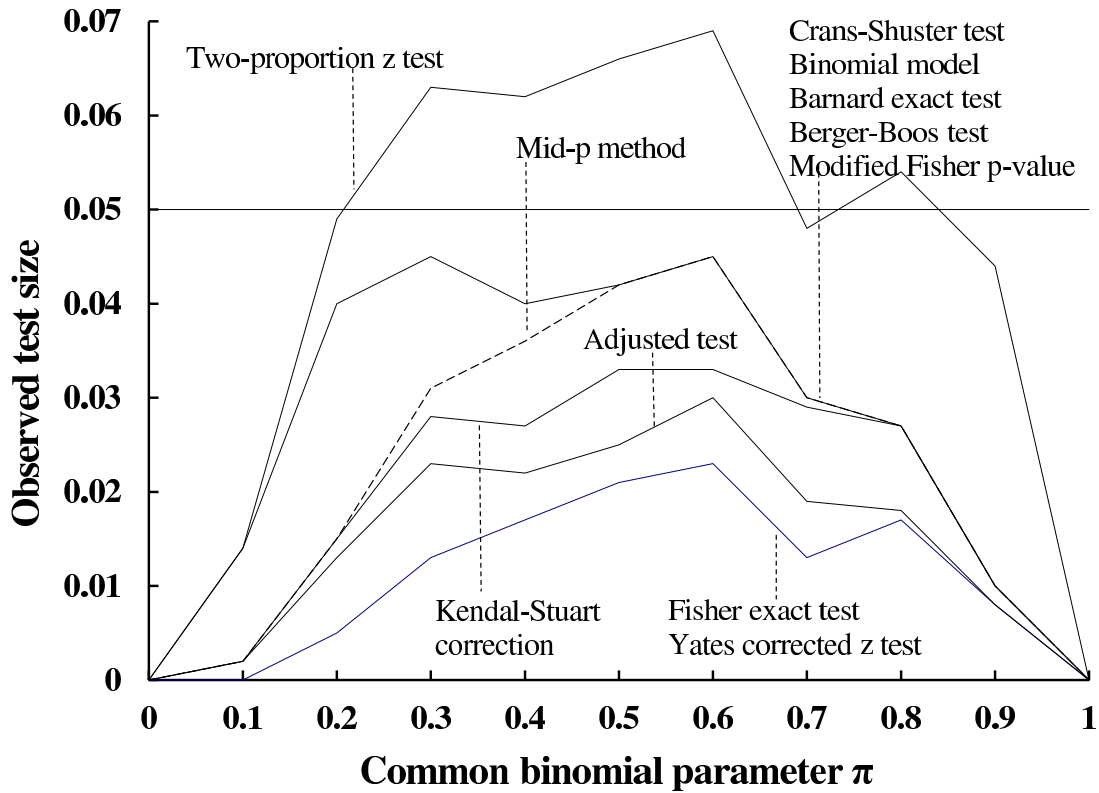


Figure 2. A comparison of the adjusted test and the other ten tests for the observed test size. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. The horizontal straight line refers to the nominal level $\alpha = 0.05$ (one-sided). Dash line represents the mid-p method and solid line the other tests.

4.3 Actual and Observed Power

Actual power is calculated from the exact conditional method with the extensions under $H_0 \cup H_1$ with $\pi_1 \geq \pi_2$ and $\psi \geq 1$. First, we get the conditional power of the level- α test $g(\psi|m) = \sum_{y_{1c} \leq y_1 \leq \xi_+} f(y_1|m, \psi)$. Then the actual power is given by

$$g(\psi, \pi_1, \pi_2) = \sum_{1 \leq m \leq n-1} g(\psi|m)P(m, \pi_1, \pi_2), \tag{4.8}$$

where $P(m, \pi_1, \pi_2)$ is the joint distribution of m under $H_0 \cup H_1$,

$$P(m, \pi_1, \pi_2) = \sum_{\xi_- \leq t \leq \xi_+} \binom{n_1}{t} \pi_1^t (1 - \pi_1)^{n_1-t} \binom{n_2}{m-t} \pi_2^{m-t} (1 - \pi_2)^{n_2-m+t}.$$

Similarly, we give a numerical analysis to look into the actual power. The sample size is the same as that for the actual test size. The significance level is taken to be 0.01, 0.025, 0.05, and 0.1 (one-sided). The other parameters are specified as $\pi_2 = 0.2$ and $\mu = 0, 0.08, \dots, 0.72$. In the upper part of Table 4.4 are given the results at $\alpha = 0.05$ from the exact conditional method with the extensions.

Table 4.4. *Actual power of the adjusted test and the other tests ($\beta \in (0, 1)$)*

| Tests | Difference μ | | | | | | | | | |
|--|------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 0 | 0.08 | 0.16 | 0.24 | 0.32 | 0.4 | 0.48 | 0.56 | 0.64 | 0.72 |
| Exact conditional method with the extensions | | | | | | | | | | |
| Adjusted test | 0.012 | 0.04 | 0.087 | 0.155 | 0.252 | 0.384 | 0.548 | 0.72 | 0.864 | 0.95 |
| Fisher exact test | 0.003 | 0.017 | 0.052 | 0.117 | 0.219 | 0.357 | 0.52 | 0.683 | 0.82 | 0.923 |
| Mid-p method | 0.014 | 0.049 | 0.116 | 0.222 | 0.363 | 0.523 | 0.68 | 0.811 | 0.908 | 0.971 |
| Crans-Shuster test | 0.036 | 0.085 | 0.156 | 0.252 | 0.38 | 0.53 | 0.682 | 0.812 | 0.908 | 0.971 |
| Algorithm of Crans and Shuster with the extensions | | | | | | | | | | |
| Adjusted test | 0.012 | 0.04 | 0.087 | 0.155 | 0.252 | 0.384 | 0.548 | 0.72 | 0.864 | 0.95 |
| Fisher exact test | 0.003 | 0.017 | 0.052 | 0.117 | 0.219 | 0.357 | 0.52 | 0.683 | 0.82 | 0.923 |
| Mid-p method | 0.014 | 0.049 | 0.116 | 0.222 | 0.363 | 0.523 | 0.68 | 0.811 | 0.908 | 0.971 |
| Crans-Shuster test | 0.036 | 0.085 | 0.156 | 0.252 | 0.38 | 0.53 | 0.682 | 0.812 | 0.908 | 0.971 |
| Binomial model | 0.036 | 0.085 | 0.156 | 0.252 | 0.38 | 0.53 | 0.682 | 0.812 | 0.908 | 0.971 |
| Barnard exact test | 0.036 | 0.085 | 0.156 | 0.252 | 0.38 | 0.53 | 0.682 | 0.812 | 0.908 | 0.971 |
| Berger-Boos test | 0.036 | 0.085 | 0.156 | 0.252 | 0.38 | 0.53 | 0.682 | 0.812 | 0.908 | 0.971 |
| Modified Fisher p-value | 0.036 | 0.085 | 0.156 | 0.252 | 0.38 | 0.53 | 0.682 | 0.812 | 0.908 | 0.971 |
| Two-proportion z test | 0.041 | 0.106 | 0.203 | 0.325 | 0.465 | 0.61 | 0.748 | 0.861 | 0.938 | 0.98 |
| Yates corrected z test | 0.003 | 0.017 | 0.052 | 0.117 | 0.219 | 0.357 | 0.52 | 0.683 | 0.82 | 0.923 |
| Kendal-Stuart correction | 0.014 | 0.047 | 0.11 | 0.202 | 0.322 | 0.461 | 0.61 | 0.758 | 0.885 | 0.968 |

The actual power is calculated at $\alpha = 0.05$ (one-sided). The increment of significance level is $\varepsilon = 0.0522$ for Crans-Shuster test. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. The difference is μ , the probability for control group is $\pi_2 = 0.2$, and that for treatment group is $\pi_1 = \pi_2 + \mu$.

Also the calculations are repeated by the algorithm of Crans and Shuster with the extensions. The results are presented in the lower part of Table 4.4.

To see the observed power, we appeal again to Monte Carlo methods. Also we take the same parameter values as those in calculating the actual power. Crossing combination of these quantities

returns 10 patterns. The observed power is given by

$$1 - \hat{\beta} = \sum_1^{1000} I\{\text{p-value} \leq \alpha | H_1\} / 1000.$$

Table 4.5 portrays the results of the experiment sets at $\alpha = 0.05$.

Table 4.5. *Observed power of the adjusted test and the other tests ($\beta \in (0, 1)$)*

| Tests | Difference μ | | | | | | | | | |
|--------------------------|------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 0 | 0.08 | 0.16 | 0.24 | 0.32 | 0.4 | 0.48 | 0.56 | 0.64 | 0.72 |
| Conditional tests | | | | | | | | | | |
| Adjusted test | 0.013 | 0.049 | 0.101 | 0.166 | 0.262 | 0.387 | 0.566 | 0.739 | 0.878 | 0.95 |
| Fisher exact test | 0.005 | 0.019 | 0.063 | 0.129 | 0.222 | 0.362 | 0.538 | 0.698 | 0.843 | 0.933 |
| Mid-p method | 0.015 | 0.058 | 0.127 | 0.229 | 0.377 | 0.532 | 0.696 | 0.832 | 0.918 | 0.975 |
| Unconditional tests | | | | | | | | | | |
| Crans-Shuster test | 0.04 | 0.089 | 0.159 | 0.257 | 0.388 | 0.537 | 0.696 | 0.832 | 0.918 | 0.975 |
| Binomial model | 0.04 | 0.089 | 0.159 | 0.257 | 0.388 | 0.537 | 0.696 | 0.832 | 0.918 | 0.975 |
| Barnard exact test | 0.04 | 0.089 | 0.159 | 0.257 | 0.388 | 0.537 | 0.696 | 0.832 | 0.918 | 0.975 |
| Berger-Boos test | 0.04 | 0.089 | 0.159 | 0.257 | 0.388 | 0.537 | 0.696 | 0.832 | 0.918 | 0.975 |
| Modified Fisher p-value | 0.04 | 0.089 | 0.159 | 0.257 | 0.388 | 0.537 | 0.696 | 0.832 | 0.918 | 0.975 |
| Approximate tests | | | | | | | | | | |
| Two-proportion z test | 0.049 | 0.111 | 0.209 | 0.339 | 0.483 | 0.626 | 0.755 | 0.872 | 0.937 | 0.979 |
| Yates corrected z test | 0.005 | 0.019 | 0.063 | 0.129 | 0.222 | 0.362 | 0.538 | 0.698 | 0.843 | 0.933 |
| Kendal-Stuart correction | 0.015 | 0.057 | 0.119 | 0.203 | 0.331 | 0.472 | 0.63 | 0.777 | 0.895 | 0.971 |

The observed power is given by the fraction of p-values less than or equal to α under H_1 in 1000 sets of samples, where $\alpha = 0.05$ (one-sided). The increment of significance level is $\varepsilon = 0.0522$ for Crans-Shuster test. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. The difference is μ , the probability for control group is $\pi_2 = 0.2$, and that for treatment group is

$$\pi_1 = \pi_2 + \mu.$$

The observed average proportion in the j th group is close to the pre-specified value of π_j . The observed power of these tests can readily be grasped from Figure 3.

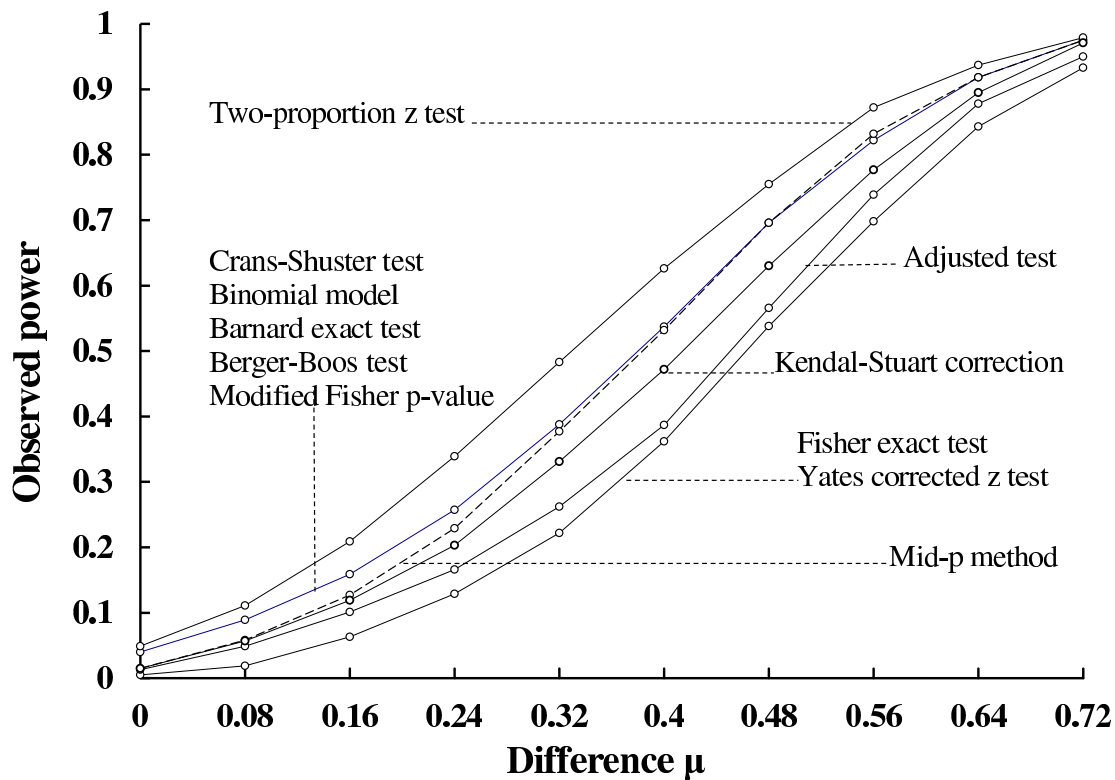


Figure 3. A comparison of the adjusted test and the other ten tests for the observed power. The observed power is given by the fraction of p-values less than or equal to α under H_1 in 1000 sets of samples, where $\alpha = 0.05$ (one-sided) but $0.05 + 0.0522$ for Crans-Shuster test. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. The success probability of control group is $\pi_2 = 0.2$ and that of treatment group is $\pi_1 = \pi_2 + \mu$. Dash line represents the mid-p method and solid line the other tests.

Figure 4 shows a comparison between the observed and actual power of the adjusted test at a variety of significance levels.

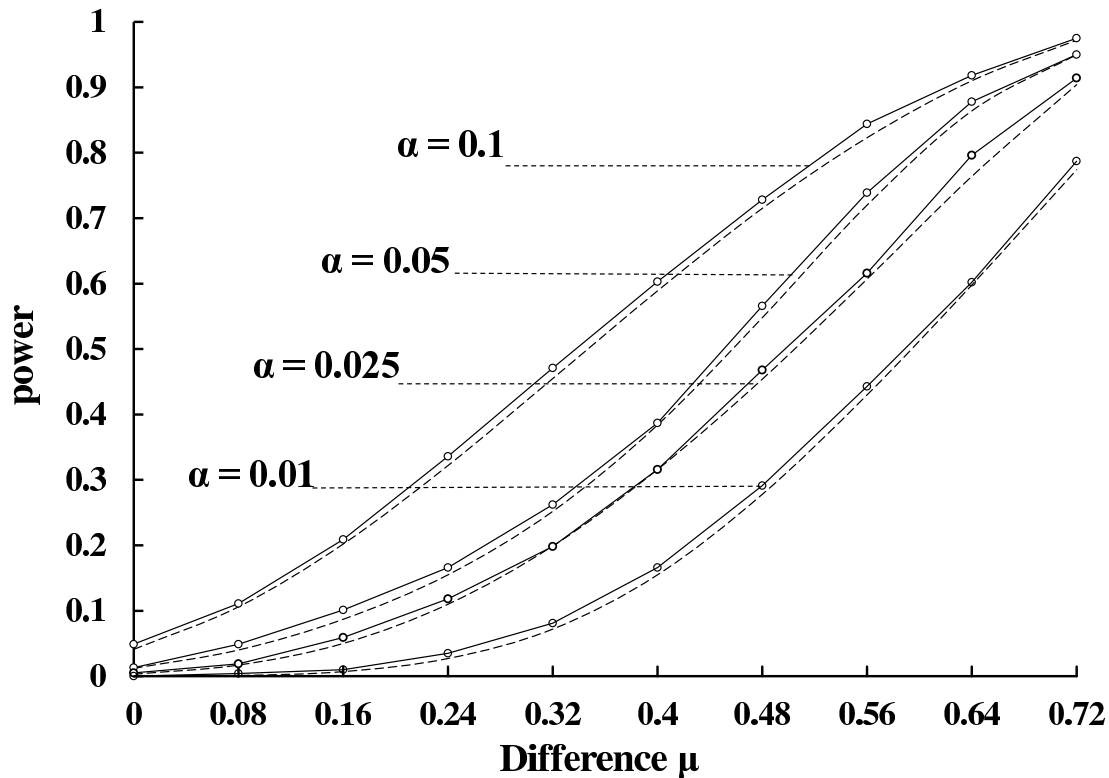


Figure 4. A comparison of the observed and actual power of the adjusted test. Solid lines indicate the observed power from 1000 times of simulations. Dash lines represent the actual power from the exact conditional method (Haseman, 1978, Casagrande et al., 1978) with extensions or the algorithm of Crans and Shuster (2008) with extensions. The sample size is $n = 20$ with sample fractions $\{k_1, k_2\} = \{0.6, 0.4\}$. The success probability of control group is $\pi_2 = 0.2$ and that of treatment group is $\pi_1 = \pi_2 + \mu$.

The above calculations are limited to fixed sample sizes. Now consider a more comprehensive simulation for samples of various sizes. The exact conditional method can return sample sizes but computations are not straightforward and require iteration. Here the sample size is computed by (4.7) and (4.8) for the adjusted test with the parameter values $\alpha = 0.05$ (one-sided), $\beta = 0.1$, $\{k_1, k_2\} = (0.6, 0.4)$, $\pi_2 = 0.2$, $\mu = 0, 0.08, \dots, 0.72$, and $\pi_1 = \pi_2 + \mu$. Needless to say, this requires extensive computations. The actual power is calculated with the same method as that for the fixed sample and the observed power is still computed by Monte Carlo methods. For simplicity in presentation, only the findings for $\mu = 0.32, 0.4, \dots, 0.72$ are displayed in Table 4.6.

Table 4.6. Power of the adjusted test and the other tests for samples of various sizes ($\beta = 0.1$)

| Difference μ | Actual power | | | | | | Observed power | | | | | |
|--------------------------|--------------|-------|-------|-------|-------|-------|----------------|-------|-------|-------|-------|-------|
| | 0.32 | 0.4 | 0.48 | 0.56 | 0.64 | 0.72 | 0.32 | 0.4 | 0.48 | 0.56 | 0.64 | 0.72 |
| Sample size n | 87 | 57 | 41 | 31 | 23 | 17 | 87 | 57 | 41 | 31 | 23 | 17 |
| Conditional tests | | | | | | | | | | | | |
| Adjusted test | 0.9 | 0.9 | 0.895 | 0.903 | 0.907 | 0.901 | 0.908 | 0.9 | 0.894 | 0.903 | 0.907 | 0.908 |
| Fisher exact test | 0.9 | 0.887 | 0.88 | 0.901 | 0.888 | 0.848 | 0.908 | 0.886 | 0.879 | 0.901 | 0.885 | 0.853 |
| Mid-p method | 0.928 | 0.924 | 0.926 | 0.944 | 0.945 | 0.946 | 0.93 | 0.926 | 0.919 | 0.948 | 0.935 | 0.946 |
| Unconditional tests | | | | | | | | | | | | |
| Binomial model | 0.923 | 0.919 | 0.909 | 0.923 | 0.924 | 0.946 | 0.926 | 0.922 | 0.905 | 0.925 | 0.92 | 0.946 |
| Barnard exact test | 0.923 | 0.919 | 0.909 | 0.923 | 0.924 | 0.946 | 0.926 | 0.922 | 0.905 | 0.925 | 0.92 | 0.946 |
| Berger-Boos test | 0.928 | 0.925 | 0.935 | 0.943 | 0.945 | 0.946 | 0.93 | 0.927 | 0.929 | 0.948 | 0.935 | 0.946 |
| Modified Fisher p-value | 0.928 | 0.932 | 0.935 | 0.944 | 0.945 | 0.946 | 0.93 | 0.938 | 0.929 | 0.948 | 0.935 | 0.946 |
| Approximate tests | | | | | | | | | | | | |
| Two-proportion z test | 0.935 | 0.936 | 0.945 | 0.948 | 0.963 | 0.958 | 0.937 | 0.942 | 0.943 | 0.952 | 0.958 | 0.951 |
| Yates corrected z test | 0.893 | 0.887 | 0.88 | 0.896 | 0.888 | 0.848 | 0.901 | 0.886 | 0.879 | 0.9 | 0.885 | 0.869 |
| Kendal-Stuart correction | 0.9 | 0.889 | 0.895 | 0.903 | 0.91 | 0.944 | 0.908 | 0.889 | 0.894 | 0.903 | 0.909 | 0.946 |

The probability for control group is $\pi_2 = 0.2$ and that for treatment group is $\pi_1 = \pi_2 + \mu$. The sample size is computed by (4.7) and (4.8) with iterations, where the sample fractions are $\{k_1, k_2\} = \{0.6, 0.4\}$. The actual power is calculated at $\alpha = 0.05$ (one-sided) and the observed power is given by the fraction of p-values less than or equal to α under H_1 in 1000 sets of samples.

We have seen that the adjusted test has its power around 0.9 but the other tests have a wide variety of power.

Power calculations and simulations were conducted at $\alpha = 0.01, 0.025, 0.05$, and 0.1 and $\beta \in (0, 1)$. In Table 4.4, 4.5, 4.6, however, only the results at the 0.05-level are listed since the results are generally the same for other levels.

Before leaving this section, there are some notes of interest:

Remark 4.4. *On comparing the upper and lower parts of Table 4.2 and 4.4, we see that two distinct algorithms provide identical results of actual test size and power. One is the exact conditional method with the extensions. The other is an unconditional method, the algorithm of Crans and Shuster with the extensions.*

Remark 4.5. *Two tests produce identical values of power (see Table 4.4 and 4.5). One is the modified Fisher p-value, which gives a decreased p-value. The other is Crans-Shuster test, which uses an increased significance level. This is no surprise: Both of them are derived from*

unconditional approach with the critical region defined by Fisher test so that the decrement of p-values is equivalent to the increment of significance level. The equivalence also holds for the size of test except for a few differences literal (see Table 4.2 and 4.3). Crans-Shuster test is not so easy to compute as the modified Fisher p-value. In fact, the increment of significance level ε (4.6) is too time consuming to compute for larger sample sizes. Hence we could not include Crans-Shuster test in the more comprehensive simulation.

Remark 4.6. *The binomial model has the critical region defined by the z test (Routledge, 1992). It remains unclear which test statistic is preferred when implementing Barnard test (Wikipedia, the free encyclopedia). In this context, we also took the z test to define the critical region of Barnard test such that it yields the result identical to that of the binomial model. Different results may be encountered, however, in other contexts. For example, Barnard test may have the critical region defined by Fisher test and then it becomes the modified Fisher p-value (Lin and Yang, 2009).*

5. COMPARING THE DATA-BASED FACTOR AND THE FACTOR $\frac{1}{2}$

The equation (3.3) is the adjusted test when $W = 1 - F(y_1)$ and the mid-p method when $W = \frac{1}{2}$. Both of them address the conservativeness of Fisher test. In comparing the data-based factor and the factor $\frac{1}{2}$, our main interest is to see whether or not the properties of Fisher test hold. Thus Fisher test is taken as the starting point. On the one hand, Fisher test holds the property that the significance level is under control of nominal α , which in turn dominates over the test size and power. On the other, Fisher test gives p-values depending on observed data (Agresti, 2002, p95), especially the ratio of the left- to right-sided p-values $R_F = F_F(y_1)/S_F(y_1)$ may shed light on the nature of data.

5.1 Controlled and Uncontrolled Significance Levels

As we mention in Subsection 4.2, a test may be conducted at the nominal α plus an increment. The increment is ε , $\alpha f(y_{1c}|m, 1)$, and $\frac{1}{2}f(y_{1c}|m, 1)$ for Crans-Shuster test, the adjusted test, and the mid-p method as shown in (4.5), (4.7), and (4.4), respectively. It is of great concern, however, that the increment is still under control of nominal α . This is true for Crans-Shuster test and the adjusted test, arguing that the increment ε in (4.6) and the increment $\alpha f(y_{1c}|m, 1)$ in (4.7) are the functions of α . It is not true for the mid-p method, however, because the increment $\frac{1}{2}f(y_{1c}|m, 1)$ in (4.4) is independent of α . For clarity, the right-hand sides of (4.1), (4.7), and (4.4) are arranged successively as an array

$$\alpha, \quad \alpha + \alpha f(y_{1c}|m, 1), \quad \text{and} \quad \alpha + \frac{1}{2}f(y_{1c}|m, 1), \quad (5.1)$$

which are the significance levels for Fisher test, the adjusted test, and the mid-p method, respectively.

Each increment of significance level has a corresponding increment of test size. For example, we return now to Table 4.2 and see the column "Nominal level $\alpha = 0.05$ ", where the values of

actual test size for the three tests are

$$0.018, \quad 0.018 + 0.007 = 0.025, \quad \text{and} \quad 0.018 + 0.024 = 0.042,$$

respectively. There is a one-to-one correspondence between the increment of significance level and the increment of actual test size. Regarding the adjusted test, the increment of significance level $\alpha f(y_{1c}|m, 1)$ corresponds to the increment of actual test size 0.007. The mid-p method shows the increment of significance level $\frac{1}{2}f(y_{1c}|m, 1)$ and the corresponding increment of actual test size 0.024. As expected, Fisher test has the value of actual size 0.018 much lower than the nominal level 0.05. Using the increment of significance level, the adjusted test gains higher value 0.025 and the mid-p method even higher value 0.042 but both of them are still lower than the nominal level.

Such a correspondence is also seen in the relevant values of power. Looking at Table 4.4, the column "Difference $\mu = 0.64$ " gives the values of actual power for the three tests:

$$0.82, \quad 0.82 + 0.044 = 0.864, \quad \text{and} \quad 0.82 + 0.088 = 0.908.$$

The corresponding increment of actual power is 0.044 for the adjusted test and 0.088 for the mid-p method.

We have seen that all the increments in the mid-p method are larger than the corresponding increments in the adjusted test. However, it is important to note that:

Remark 5.1. *With respect to the adjusted test, the increment of significance level $\alpha f(y_{1c}|m, 1)$ is under control of α and so are the corresponding increments of test size and power.*

Remark 5.2. *The mid-p method uses a higher significance level with the increment $\frac{1}{2}f(y_{1c}|m, 1)$, which is in general larger than $\alpha f(y_{1c}|m, 1)$. This predicts a higher test size and power. However, the increment of significance level $\frac{1}{2}f(y_{1c}|m, 1)$ is not controlled by α and as a consequence, the corresponding increments of test size and power suffer from the same problem.*

5.2 Proportional and Disproportional Reduction

Any modification should be the same for the left- and right-sided p-values. This is the case for the adjusted test. Looking at (3.8), the adjustment $(1 + f(y_1))^{-1}$ is the same for the two one-sided p-values.

By contrast, this is not the case for the mid-p method. For clarity, (4.3) is put into the form

$$F_L(y_1) = \sum_{\xi_- \leq t \leq y_1} \left(f(t) - \frac{\frac{1}{2}f(y_1)}{y_1 - \xi_- + 1} \right) \quad \text{and} \quad S_L(y_1) = \sum_{y_1 \leq t \leq \xi_+} \left(f(t) - \frac{\frac{1}{2}f(y_1)}{\xi_+ - y_1 + 1} \right). \quad (5.2)$$

Since $y_1 - \xi_- + 1 \neq \xi_+ - y_1 + 1$ holds in general, it follows that

$$\frac{\frac{1}{2}f(y_1)}{y_1 - \xi_- + 1} \neq \frac{\frac{1}{2}f(y_1)}{\xi_+ - y_1 + 1}. \quad (5.3)$$

It states that the subtrahend on the left-hand side differs from that on the right. In the fish experiment, for example, the calculation of (5.3) results in $0.025 \neq 0.00625$. Evidently, the mid-p method uses different modifications for the two one-sided p-values.

Further insight is provided by the ratio of the left- to right-sided p-values. The adjusted test (3.8) indicates that the ratio R is identical to R_F :

$$R = \frac{F(y_1)}{S(y_1)} = \frac{F_F(y_1)/(1 + f(y_1))}{S_F(y_1)/(1 + f(y_1))} \equiv R_F,$$

which gives a confirmation of the proportional reduction as mentioned in Subsection 3.2.

Conversely, the mid-p method indicates a disproportional reduction from Fisher test. This can easily be seen from (4.3). When $F_F(y_1) \neq S_F(y_1)$, it gives the ratio R_L unequal to R_F :

$$R_L = \frac{F_L(y_1)}{S_L(y_1)} = \frac{F_F(y_1) - \frac{1}{2}f(y_1)}{S_F(y_1) - \frac{1}{2}f(y_1)} \neq R_F.$$

A consequence of this is that the smaller one in the left- and right-sided p-values is depressed, which predicts an increase in Type I error rate (very significance level and pseudo power are attainable). Again in the fish experiment, the adjusted test gives the ratio $R = 0.048/0.952 = 0.05$, which is identical to $R_F = 0.05/1 = 0.05$, whereas the mid-p method gives the much less ratio $R_L = 0.025/0.975 = 0.026$. In a careful examination, the mid-p method has the left-sided p-value depressed from 0.048 to 0.025 with the decrement 0.023 and the right elevated from 0.952 to 0.975 with the increment 0.023. The increment is just equal to the decrement. That is equivalent to Type I error rate being elevated by 0.023.

Here the thing being counted is the factor $\frac{1}{2}$ that raises some more concerns: the significance level not being entirely controlled by α , different modifications in the two one-sided p-values, and disproportional reduction from Fisher test. This explains that the mid-p method fails to hold the properties of Fisher test. We now return to Question (1): The solution is obvious enough.

6. EXTENSIONS FOR $r \times c$ CONTINGENCY TABLES

As a preparation for extending to $r \times c$ contingency tables, the cell counts in 2×2 tables are denoted by $\{Y_{ij}\}$ for $i = 1, 2$ and $j = 1, 2$. Given a data set $\{Y_{ij}\} = \{y_{ij}\}$, we have the row margins $\{n_1, n_2\}$, the column margins $\{m_1, m_2\}$, and the total number of observations n . The pdf of hypergeometric distribution in Subsection 3.1 is rewritten in a more general form

$$f(t) = \frac{n_1!n_2!m_1!m_2!}{n!y_{11}!y_{12}!y_{21}!y_{22}!},$$

where $f(t)$ is the probability of table t , $\xi_- \leq t \leq \xi_+$, $\xi_- = \max(0, m_1 - n_2)$, and $\xi_+ = \min(n_1, m_1)$. This form provides a room for extending to $r \times c$ tables.

With a little straightforward, $i = 1, 2$ and $j = 1, 2$ are extended to $i = 1, 2, \dots, R$ and $j = 1, 2, \dots, C$, followed by the row margins $\{n_1, n_2, \dots, n_R\}$ and the column margins $\{m_1, m_2, \dots, m_C\}$. One may use the network algorithm (Mehta & Patel, 1983) to generate all possible tables with given margins. Letting $t = 1, 2, \dots, T$ denote all the tables, the pdf is extended to

$$f(t) = \frac{\prod_i (n_i!) \prod_j (m_j!)}{n! \prod_i \prod_j (y_{ij}!)} \quad (6.1)$$

with a multiple hypergeometric distribution. Since (6.1) holds, it must be true that $\sum_{t=1}^T f(t) = 1$.

Assume that $f(y)$ is the probability of the observed table y , so that the popular approach for a two-sided test (3.2)(Agresti, 1992) is extended to

$$P_F = \sum_{t=1}^T f(t) | (f(t) \leq f(y)). \quad (6.2)$$

This is just the statistic of exact conditional test. Note that there are no one-sided p-values in $r \times c$ tables. When $R = 2$ and $C = 2$, (6.2) reduces to (3.2) for 2×2 tables.

Applying the data-based adjustment, (3.9) is extended to

$$P_Z = \sum_{t=1}^T f(t) / (1 + f(y)) | (f(t) \leq f(y)) = P_F / (1 + f(y)). \quad (6.3)$$

This is the statistic of adjusted test for $r \times c$ tables. The reducibility also holds for (6.3): When $R = 2$ and $C = 2$, (6.3) reduces to (3.9) for 2×2 tables.

Special algorithms and software are widely available for computing exact conditional tests for $r \times c$ tables (Agresti, 2002, p98). For larger tables, one can use Monte Carlo method to sample randomly under the multiple hypergeometric distribution from the set of tables with the given margins. The estimated p-value is then the sample proportion of tables having test statistic value at least as large as the value observed.

Using the same principle, the data-based adjustment is easy to be applied in $2 \times 2 \times k$ and $r \times c \times k$ tables. As R and / or C increase, however, the conservativeness issue for conditional tests becomes less problematic and then adjustments become less important.

7. EXAMPLES

7.1 The Fish Experiment

To illustrate the use of the adjusted test, we take the fish experiment from Routledge (1992) for ease of comparison and repetition. It was designed to assess the ability of ozone to control bacteria in a fish tank. Six tanks were randomly allocated into two groups of three each. A fixed amount of bacteria was added to each of them. Tanks in one group were also treated with ozone. This group will be labeled the treatment group; the other, the control group. After some time, all three tanks in the control group contained dead fish, whereas none of the other tanks did. Table 7.1 shows the results of the experiment.

Table 7.1. *The results of the fish experiment*

| Treatment | Response category | | Total |
|-----------|-------------------|--------------|-------|
| | Some dead fish | No dead fish | |
| Treated | 0 | 3 | 3 |
| Untreated | 3 | 0 | 3 |
| Total | 3 | 3 | 6 |

Source: Routledge (1992).

Let π_1 be the probability for the treatment group and let π_2 be that for the control group. State the decision rule for testing $H_0 : \pi_1 = \pi_2$ versus $H_1 : \pi_1 < \pi_2$ at $\alpha = 0.05$. Crans-Shuster test uses the significance level plus an increment ε calculated by a call to (4.6).

The analysis gives the table probability of $f(y_1) = 0.05$. This is just the left-sided p-value of Fisher test in this case. It exactly matches $\alpha = 0.05$. The right-sided p-value is 1 so that we have the sum of 1.05. The two-sided p-values (the left, right components) are $P_F(1) = 0.1$ (0.05, 0.05), $P_F(2) = 0.05$ (0.05, 0), $P_F(3) = 0.1$ (0.05, 0.05), and $P_F(4) = 0.1$. Note that $P_F(2)$ alone gives different values because $E[t] = 1.5$ is different from $t_{max} = 1$ in this case. This reminds us that another possibility of two-sided test must be used with caution, which has already been predicted in Subsection 3.1. The left- and right-sided p-values from the mid-p method are $F_L(y_1) = 0.025$ and $S_L(y_1) = 0.975$ with the sum of 1 and the two-sided p-value is $P_L = 0.05$. Regarding the adjusted test, the left-sided p-value comes to $F(y_1) = 0.048$ that is significant at the 0.05 level. It has a complementary right-sided p-value $S(y_1) = 0.952$. This is in favor of such an interpretation that using ozone is better than not using. The two-sided p-values are $P(1) = 0.095$ (0.048, 0.048), $P(2) = 0.048$ (0.048, 0), $P(3) = 0.095$ (0.048, 0.048), and $P(4) = 0.095$.

It is of interest to see possible results from different frequencies. Thus we write $t = Y_1$ and $m - t = Y_2$, where $t = 0, 1, 2, 3$. The left- and two-sided p-values are computed for all the eleven tests as shown in Table 7.2.

Table 7.2. *Left- and two-sided p-values of the eleven tests in the fish experiment*

| Tests | The left-sided ($Y_1 = t$) | | | | The two-sided ($Y_1 = t$) | | | |
|--------------------------|------------------------------|-------|-------|-------|-----------------------------|-------|-------|-------|
| | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| Conditional tests | | | | | | | | |
| Adjusted test | 0.048 | 0.345 | 0.655 | 0.952 | 0.095 | 0.69 | 0.69 | 0.095 |
| Fisher exact test | 0.05 | 0.5 | 0.95 | 1 | 0.1 | 1 | 1 | 0.1 |
| Mid-p method | 0.025 | 0.275 | 0.725 | 0.975 | 0.05 | 0.55 | 0.55 | 0.5 |
| Unconditional tests | | | | | | | | |
| Crans-Shuster test | 0.05 | 0.5 | 0.95 | 1 | 0.1 | 1 | 1 | 0.1 |
| Binomial model | 0.016 | 0.344 | 0.656 | 0.984 | 0.031 | 0.687 | 0.687 | 0.031 |
| Barnard exact test | 0.016 | 0.344 | 0.656 | 0.984 | 0.031 | 0.687 | 0.687 | 0.031 |
| Berger-Boos test | 0.017 | 0.345 | 0.655 | 0.983 | 0.032 | 0.688 | 0.688 | 0.032 |
| Modified Fisher p-value | 0.016 | 0.25 | 0.766 | 0.781 | 0.031 | 0.5 | 0.5 | 0.031 |
| Approximate tests | | | | | | | | |
| Two-proportion z test | 0.007 | 0.207 | 0.793 | 0.993 | 0.014 | 0.414 | 0.414 | 0.014 |
| Yates corrected z test | 0.051 | 0.207 | 0.793 | 0.949 | 0.102 | 0.414 | 0.414 | 0.102 |
| Kendal-Stuart correction | 0.034 | 0.207 | 0.793 | 0.966 | 0.068 | 0.414 | 0.527 | 0.068 |

Of these, the two-sided p-values of Fisher test and the adjusted test are calculated by the popular approach (3.2) and (3.9), respectively.

The column " $t = 0$ " refers to the results of the observed data. The left-sided p-value of the adjusted test is smaller than that of Fisher test and greater than that of the mid-p method. We have already seen in Section 5, however, that the adjusted test holds the properties of Fisher test and has the increment of power under control of α , which the mid-p method lacks (see Remark 5.1 and 5.2). The binomial model and Barnard test reach the same results. The p-values of Berger-Boos test and the modified Fisher p-value are less than the p-value of the mid-p method. The unconditional tests have smaller p-values than the adjusted test. As mentioned in Subsection 4.2, however, when using the unconditional tests, the inference depends partly on the unobserved samples so that considerable controversy surrounds their use in statistical literature (recall the golden aphorism). The minimum p-value comes from the z test and the maximum from Yates z test. The left-sided p-value of Yates z test resembles that of Fisher test. The p-value of Kendal-Stuart correction is less than that of Yates z test. Kendal-Stuart correction has smaller p-value than the adjusted test. Note, however, that approximate tests are inappropriate for such a small sample since the approximations can be very poor in such cases (see Section 2). They are mentioned here for comparison only. The relative change of the two-sided p-values among these tests is the same as that of the left-sided p-values.

Table 7.2. also lists the results in the range of $t = 1, 2, 3$, which show that the manner of variability among these tests is the same as that for $t = 0$. The left-sided p-values of the adjusted test are greater than those of the mid-p method in the region $t = 0, 1$ but smaller in the region $t = 2, 3$ with the differences $\{0.023, 0.07, -0.07, -0.023 | t = 0, 1, 2, 3\}$. Obviously, the differences in the two regions just cancel each other out.

7.2 The Lady Tasting Tea Experiment

Another example is the lady tasting tea experiment, which is quoted from Agresti (2002, p92). A lady tasted eight cups of tea, four of which had milk added first and four of which had tea added first. She knew there were four cups of each type and had to predict which four had the milk added first. The order of presenting the cups to her was randomized. The results of the experiment are presented in Table 7.3.

Table 7.3. *The results of lady tasting tea experiment*

| Poured First | Guess poured first | | Total |
|--------------|--------------------|-----|-------|
| | Milk | Tea | |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

Source: Agresti (2002, p92).

Thus the hypotheses to be tested are $H_0 : \pi_1 = \pi_2$ vs $H_1 : \pi_1 > \pi_2$, where π_1 is the probability for milk added first and π_2 that for tea added first. The experimental design fixed both marginal distributions, since the researcher had to predict which four cups had milk added first. Thus the hypergeometric applies naturally for the null distribution of Y_1 .

With the table probability $f(y_1) = 0.229$, the right-sided p-value of Fisher test is $S_F(y_1) = 0.243$. It fails to reach significance at $\alpha = 0.05$ so that this result does not establish an association between the actual order of pouring and her predictions. It is not helpful using any other tests. For example, the mid-p method gives $S_L(y_1) = 0.129$ and the adjusted test $S(y_1) = 0.198$.

In an attempt to assess the association, we repeat the calculations with one to sixfold increase in the sample size. Table 7.4 presents the right-sided p-values.

Table 7.4. *The right-sided p-values calculated from samples of various sizes*

| Tests | Multiples of sample sizes | | | | | |
|--------------------------|---------------------------|-------|-------|-------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Conditional tests | | | | | | |
| Adjusted test | 0.198 | 0.062 | 0.019 | 0.006 | 0.002 | 0.001 |
| Fisher exact test | 0.243 | 0.066 | 0.02 | 0.006 | 0.002 | 0.001 |
| Mid-p method | 0.129 | 0.036 | 0.011 | 0.003 | 0.001 | 0 |
| Unconditional tests | | | | | | |
| Crans-Shuster test | 0.243 | 0.066 | 0.02 | 0.006 | 0.002 | 0.001 |
| Binomial model | 0.145 | 0.038 | 0.011 | 0.004 | 0.001 | 0 |
| Barnard exact test | 0.145 | 0.038 | 0.011 | 0.004 | 0.001 | 0 |
| Berger-Boos test | 0.146 | 0.039 | 0.012 | 0.005 | 0.002 | 0.001 |
| Modified Fisher p-value | 0.145 | 0.038 | 0.011 | 0.004 | 0.001 | 0 |
| Approximate tests | | | | | | |
| Two-proportion z test | 0.079 | 0.023 | 0.007 | 0.002 | 0 | 0 |
| Yates corrected z test | 0.24 | 0.067 | 0.021 | 0.007 | 0.002 | 0 |
| Kendal-Stuart correction | 0.159 | 0.057 | 0.019 | 0.006 | 0.002 | 0 |

It is no surprise that as n increases the p-values decrease. Multiple 1 refers to the original sample sizes. All of the right-sided p-values are greater than 0.05. Certainly H_0 is not rejected at $\alpha = 0.05$ by any of the tests. With double sample sizes, however, H_0 is rejected at $\alpha = 0.05$ by the mid-p method, the unconditional tests, or the z test. This implies that there be an association between the actual order of pouring and her predictions. When sample sizes increase by threefold or more, any of the tests may give p-values below 0.05. In addition, the numerical distances between any two p-values decrease as n increases.

7.3 The Illustrative Example for $r \times c$ Tables

For comparing easily, we take the illustrative example from Howell and Gordon (1976), as shown in Table 7.5.

Table 7.5. *An illustrative example for exact conditional test*

| | | | |
|--------------|--------------|--------------|-----------|
| $Y_{11} = 4$ | $Y_{12} = 4$ | $Y_{13} = 0$ | $n_1 = 8$ |
| $Y_{21} = 0$ | $Y_{22} = 4$ | $Y_{23} = 3$ | $n_2 = 7$ |
| $m_1 = 4$ | $m_2 = 8$ | $m_3 = 3$ | $n = 15$ |

Source: Howell and Gordon, 1976.

Let t_1 be the values of Y_{11} and let t_2 be the values of Y_{12} . The range of t_1 is defined as $t_1 \in [\xi_{1-}, \xi_{1+}]$, where $\xi_{1-} = \max(0, m_1 - n_2) = 0$ and $\xi_{1+} = \min(n_1, m_1) = 4$. The range of t_2 is defined as $t_2 \in [\xi_{2-}, \xi_{2+}]$, where $\xi_{2-} = \max(0, m_2 - (n_2 - m_1 + t_1))$ and $\xi_{2+} = \min(n_1 - t_1, m_2)$. Given margins, t_1 and t_2 determine the other cell counts based on $(2 - 1)(3 - 1) = 2$ degrees of freedom. There are a total of 20 possible tables as tabulated in Table 7.6.

Table 7.6. All possible tables and the corresponding probabilities

| t_1 | t_2 | | | | $f(t)$ | | | |
|-------|-------|---|---|---|---------------------|--------|---------------------|---------------------|
| 0 | 5 | 6 | 7 | 8 | 0.0087 ^e | 0.0131 | 0.0037 ^e | 0.0002 ^e |
| 1 | 4 | 5 | 6 | 7 | 0.0435 | 0.1044 | 0.0522 | 0.0050 ^e |
| 2 | 3 | 4 | 5 | 6 | 0.0522 | 0.1958 | 0.1566 | 0.0261 |
| 3 | 2 | 3 | 4 | 5 | 0.0174 | 0.1044 | 0.1305 | 0.0348 |
| 4 | 1 | 2 | 3 | 4 | 0.0012 ^e | 0.0131 | 0.0261 | 0.0109 ^y |

Source: Howell and Gordon 1976. The column t_1 gives the values of frequency Y_{11} , t_2 the values of frequency Y_{12} , and $f(t)$ the probability of table t with $t_1 = Y_{11}$ and $t_2 = Y_{12}$. The symbol ^y indicates the table probability and ^e the extreme probabilities.

The corresponding probabilities are easily calculated by a call to (6.1), shown in the column $f(t)$. We have seen that the 20 probabilities sum to 1.

One may find the table probability from the term with the symbol ^y and the five extreme probabilities from the terms with symbol ^e. Using (6.2) yields $P_F = 0.02968$ the p-value of exact conditional test, which is the same as that in Howell and Gordon, 1976. The calculation of (6.3) gives $P_Z = 0.02936$ the p-value of the adjusted test. The null hypothesis is therefore rejected at $\alpha = 0.05$.

Large-sample tests may be inappropriate for such a small sample. For comparison, however, the usual chi-squared test is also mentioned here. Using the statistic gives 6.96429 with the probability 0.03074, which is the same as that in Howell and Gordon, 1976. Again for comparison, the data are analyzed by the Monte Carlo method. Sampling was repeated 2000 times resulted in 0.02977 for exact conditional test and 0.02945 for the adjusted test, which are similar to those from (6.2) and (6.3). For this example, a simple computer program in R language is available from the author upon request.

8. DISCUSSION

The principle behind the data-based adjustment is quite simple: The conservativeness of Fisher test is known to be due to the discreteness, which is displayed intuitively as the non-exclusivity as noted in Subsection 3.1. The adjustment just offsets the non-exclusivity as shown in Remark

4.1. Accordingly, the two one-sided p-values become mutually exclusive, which is a property of continuous distributions.

The adjustment is derived from an intuitive method: Take account of both the left- and right-sided p-values and treat them equally in the derivation. This is practiced in each of the following three steps: (1) Set up an equation that incorporates a fraction of the table probability and the more extreme probabilities (see (3.3)). (2) Solve the equation for the fraction, which results in the data-based factor (see (3.7)). (3) Convert the factor to the data-based adjustment (see (3.8)). It is important to note that the properties of Fisher test do hold in the process (recall Section 5).

Interesting points of the adjustment are: (1) It gives the results always interpretable in the whole range of the sample sizes, $1 \leq n < \infty$, in which the adjustment has its minimum of 0.5 and maximum of 1. (2) The adjustment vanishes into void as $n \rightarrow \infty$ (recall Remark 4.2). (3) The standardized version of the adjusted test is asymptotically standard normal (see Remark 4.3).

The data-based adjustment reduces the conservativeness, as evidenced by increasing test size and power and decreasing p-values. A check is provided by the fact that two totally different algorithms produce the identical results when calculating the actual test size and power (see Remark 4.4).

The adjustment makes the size of test increased. When data set is small, the size of the adjusted test is greater than that of Fisher test, which can be seen in Table 4.2 and 4.3 as well as Figure 2. As expected, the adjusted test has less size than the mid-p method. The size of the mid-p method resembles the size of the unconditional tests such as Crans-Shuster test, the binomial model, Barnard test, Berger-Boos test, and the modified Fisher p-value. Yates z test behaves like Fisher test not only for the size of test but also for power and p-values. Kendal-Stuart correction has greater size than Yates z test. In most situations, Kendal-Stuart correction has greater size than the adjusted test as shown in Table 4.2 and 4.3 as well as in Figure 2. The z test has the size exceeding nominal level, whereas none of the other tests do.

The adjusted test has a power advantage over Fisher test, as noted in Table 4.4, 4.5, and 4.6 as well as Figure 3. The five unconditional tests are as powerful as the mid-p method. The power is the same for Crans-Shuster test and the modified Fisher p-value as shown in Table 4.4 and 4.5. It is no surprise that the z test is the most powerful. The power of Kendal-Stuart correction always lies between the power of the z test and Yates z test. In general, Kendal-Stuart correction has higher power than the adjusted test in the range of $\mu = 0, 0.08, \dots, 0.72$ as shown in Table 4.4 and 4.5 as well as in Figure 3. The numerical distances of power between any two tests decrease as n increases. For example, Table 4.6 shows that the adjusted test is more powerful than Fisher test for samples of sizes 17 to 57 but the two tests give the same power for the sample of size 87. Again, Table 4.6 says that the adjusted test is less powerful than Kendal-Stuart correction for the sample of size 17 but the two tests show the same power for the sample of size 87. Figure 4 explains

that the data-based adjustment behaves well, as evidenced by the close agreement between the observed and actual power at different significance levels. Concerning the power of test, it is said that the validity is questionable in an unconditional evaluation of a conditional test (Hirji, Tan, and Elashoff, 1991) arguing that a conditional test is naturally less powerful than an unconditional test.

The adjustment decreases the p-values in small samples. This is clearly displayed in Table 4.1 and Figure 1 as well as in the examples (Section 7). In both the fish experiment and the lady tasting tea experiment, we have seen that the p-value of the adjusted test is less than that of Fisher test. The adjusted test has greater p-value than the mid-p method at the observed point. It is the opposite, however, when the frequency is greater than its expected value. In addition, Kendall-Stuart correction has smaller p-values than the adjusted test (see Section 7).

The adjusted test has been compared with the other ten tests but special attention is given to the comparison with the mid-p method. Note that the mid-p method is but a particular form of the adjusted test. In rare cases, when $F(y_1) = S(y_1)$, we have $1 - F(y_1) = \frac{1}{2}$ and $1 - S(y_1) = \frac{1}{2}$ and then the adjusted test (3.7) equals the mid-p method. In addition, they have a common point: The adjusted test has the left- and right-sided p-values summing to 1 and so does the mid-p method.

The adjusted test is easy to implement. With 2×2 tables, one may use (3.8) and (3.9) to calculate the one- and two-sided p-values. The same results are obtained from (3.1) and (3.2) divided by one plus the table probability. As for $r \times c$ tables, one may refer to (6.3) to calculate the p-values of the adjusted test. Also, one may use (6.2) to obtain the p-value of exact conditional test and then divided by one plus the probability of observed table.

The mid-p method is more powerful than the adjusted test but the increment of power comes from the factor $\frac{1}{2}$. The adjusted test holds such properties as the significance level under control of nominal α , the same modification in the left- and right-sided p-values, and the proportional reduction from Fisher test, which the mid-p method lacks as noted in Section 5. Concerning the unconditional tests, they are more powerful as well but the power comes partly from the unobserved samples so that they raise some controversies (recall the golden aphorism). As for approximate tests, they are inappropriate for small samples since the approximations can be very poor in such cases (see Section 2). One pursues high power of test, but must ensure that the power comes from the data at hand and is under control of nominal α . Thus the proper choice of an adjustment is based largely upon a consideration of both the power of test and the origin of power so that the adjusted test is an option in data analyses.

The principle of the data-based adjustment can be employed to deal with other discrete problems as well, which will be reported separately.

ACKNOWLEDGEMENTS

The authors would like to thank Shijun Du, School of Information Engineering, Zhengzhou University, who helped with some software and compiled computer programs for Monte Carlo simulations in this paper.

REFERENCES

- [1] A. Agresti, A survey of exact inference for contingency tables, *Statist. Sci.* 7 (1992). <https://doi.org/10.1214/ss/1177011454>.
- [2] A. Agresti, Exact inference for categorical data: recent advances and continuing controversies, *Statist. Med.* 20 (2001) 2709–2722. <https://doi.org/10.1002/sim.738>.
- [3] A. Agresti, *Categorical Data Analysis*, Second Edition. Hoboken, New Jersey: John Wiley and Sons, Inc., (2002).
- [4] G.A. Barnard, A new test for 2×2 tables, *Nature*, 156 (1945) 177.
- [5] G.A. Barnard, Significance tests for 2×2 tables, *Biometrika*, 34 (1947) 123–138.
- [6] R.L. Berger, D.D. Boos, P-values maximized over a confidence set for the nuisance parameter, *J. Amer. Stat. Assoc.* 89 (1994) 1012–1016.
- [7] V.W. Berger, Pros and cons of permutation tests in clinical trials, *Stat. Med.* 19 (2000), 1319–1328. [https://doi.org/10.1002/\(SICI\)1097-0258\(20000530\)19:10%3C1319::AID-SIM490%3E3.0.CO;2-0](https://doi.org/10.1002/(SICI)1097-0258(20000530)19:10%3C1319::AID-SIM490%3E3.0.CO;2-0).
- [8] J. Berkson, In dispraise of the exact test, *J. Stat. Plan. Inference*, 2 (1978) 27–42.
- [9] J.T. Casagrande, M.C. Pike, P.G. Smith, The power function of the “exact” test for comparing two binomial distributions, *Appl. Stat.* 27 (1978) 176. <https://doi.org/10.2307/2346945>.
- [10] P.E. Cheng, M. Liou, J.A.D. Aston, A.C. Tsai, Information identities and testing hypotheses: power analysis for contingency tables, *Stat. Sinica*, 18 (2008) 535–558.
- [11] W.J. Conover, Some reasons for not using the Yates continuity correction on 2×2 contingency tables, *J. Amer. Stat. Assoc.* 69 (1974) 374. <https://doi.org/10.2307/2285661>.
- [12] G.G. Crans, J.J. Shuster, How conservative is Fisher’s exact test? A quantitative evaluation of the two-sample comparative binomial trial, *Stat. Med.* 27 (2008) 3598–3611. <https://doi.org/10.1002/sim.3221>.
- [13] W.D. Dupont, Sensitivity of Fisher’s exact test to minor perturbations in 2×2 contingency tables, *Stat. Med.* 5 (1986) 629–635.
- [14] R.A. Fisher, On the interpretation of 2×2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* 85 (1922) 87–94.
- [15] R.A. Fisher, *Statistical methods for research workers*, 14th edition. New York: Hafner. (1970).
- [16] R.A. Fisher, *Statistical methods for research workers*. Edinburgh: Oliver and Boyd. (1925).
- [17] J.E. Fleiss, *Statistical methods for rates and proportions*, New York: John Wiley and Sons. (1981).
- [18] M. Haber, A comparison of some continuity corrections for the chi-squared test on 2×2 tables, *J. Amer. Stat. Assoc.* 75 (1980) 510–515. <https://doi.org/10.1080/01621459.1980.10477503>.
- [19] M. Haber, An exact unconditional test for the 2×2 comparative trial. *Psychol. Bull.* 99 (1986) 129–132.
- [20] J.K. Haseman, Exact sample sizes for the use with the Fisher-Irwin test for 2×2 tables. *Biometrics* 34 (1978) 106–109.
- [21] M.G. Haviland, Yates’s correction for continuity and the analysis of 2×2 contingency tables, *Stat. Med.* 9 (1990) 363–367. <https://doi.org/10.1002/sim.4780090403>.
- [22] D.C. Howell, L.R. Gordon, Computing the exact probability of an r by c contingency table with fixed marginal totals. *Behav. Res. Meth. Instrument.* 8 (1976) 317.
- [23] K.F. Hirji, S.J. Tan, R.M. Elashoff, A quasi-exact test for comparing two binomial proportions. *Stat. Med.* 10 (1991) 1137–1153.
- [24] J.T.G. Hwang, M.C. Yang, An optimality theory for mid p-values in 2×2 contingency tables. *Stat. Sinica*, 11 (2001) 807–826.
- [25] Insightful Corporation. *S-PLUS 8 Guide to Statistics*, Volume 1. Seattle, Washington: Insightful Corporation, p1. (2007).
- [26] H.O. Lancaster, Significance tests in discrete distributions, *J. Amer. Stat. Assoc.* 56 (1961) 223–234. <https://doi.org/10.1080/01621459.1961.10482105>.
- [27] D. Liddle, Practical tests of 2×2 contingency tables. *The Statistician.* 25 (1976) 295–305.
- [28] C.-Y. Lin, M.-C. Yang, Improved p-value tests for comparing two independent binomial proportions, *Commun. Stat.-Simul. Comput.* 38 (2008) 78–91. <https://doi.org/10.1080/03610910802417812>.

- [29] S. Lydersen, M.W. Fagerland, P. Laake, Recommended tests for association in 2×2 tables, *Stat. Med.* 28 (2009) 1159–1175. <https://doi.org/10.1002/sim.3531>.
- [30] N. Mantel, S.W. Greenhouse, What is the continuity correction? *Amer. Stat.* 22 (1968) 27–30.
- [31] C.R. Mehta, N.R. Patel, A network algorithm for performing fisher's exact test in $r \times c$ contingency tables, *J. Amer. Stat. Assoc.* 78 (1983) 427–434. <https://doi.org/10.1080/01621459.1983.10477989>.
- [32] E.S. Pearson, The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table, *Biometrika.* 34 (1947) 139. <https://doi.org/10.2307/2332518>.
- [33] R.D. Routledge, Resolving the conflict over fisher's exact test, *Can. J. Stat.* 20 (1992) 201–209. <https://doi.org/10.2307/3315468>.
- [34] W.L. Stevens, Fiducial limits of the parameter of a discontinuous distribution, *Biometrika.* 37 (1950) 117. <https://doi.org/10.2307/2332154>.
- [35] S. Suissa, J.J. Shuster, Exact unconditional sample sizes for the 2 by 2 binomial trial. *J. R. Stat. Soc. Ser. A, Gen.* 148 (1985) 317–327.
- [36] K.D. Tocher, Extension of the Neyman-Pearson theory of tests to discontinuous variates, *Biometrika.* 37 (1950) 130–144. <https://doi.org/10.1093/biomet/37.1-2.130>.
- [37] G. Upton, Fisher's exact test. *J. R. Stat. Soc. Ser. A.* 155 (1992) 395–402.
- [38] W.N. Venables, D.M. Smith, the R Core Team. (2019). *An Introduction to R Notes on R: A Programming Environment for Data Analysis and Graphics.* Version 3.6.1 (2019-07-05).
- [39] F. Yates, Contingency tables involving small numbers and the χ^2 test. *J. R. Stat. Soc. Suppl.* 1 (1934) 217–235.
- [40] F. Yates, Test of significance for 2×2 contingency tables, *J. R. Stat. Soc. Ser. A (Gen.)* 147 (1984) 426. <https://doi.org/10.2307/2981577>.