mathematics

MDPI

Article

Hypothesis Test to Compare Two Paired Binomial Proportions: Assessment of 24 Methods

José Antonio Roldán-Nofuentes [1,*], Tulsi Sagar Sheth [1,2] and José Fernando Vera-Vera [3]

[1] Department of Statistics and Operations Research, School of Medicine, University of Granada, 18016 Granada, Spain
[2] Department of Applied Sciences and Humanities, Parul Institute of Engineering and Technology, Parul University, Vadodara 391760, Gujarat, India
[3] Department of Statistics and Operations Research, Faculty of Sciences, University of Granada, Fuentenueva s/n, 18071 Granada, Spain; jfvera@ugr.es
[*] Correspondence: jaroldan@ugr.es

Abstract: The comparison of two paired binomial proportions is a topic of interest in statistics, with important applications in medicine. There are different methods in the statistical literature to solve this problem, and the McNemar test is the best known of all of them. The problem has been solved from a conditioned perspective, only considering the discordant pairs, and from an unconditioned perspective, considering all of the observed values. This manuscript reviews the existing methods to solve the hypothesis test of equality for the two paired proportions and proposes new methods. Monte Carlo simulation methods were carried out to study the asymptotic behaviour of the methods studied, giving some general rules of application depending on the sample size. In general terms, the Wald test, the likelihood-ratio test, and two tests based on association measures in 2 × 2 tables can always be applied, whatever the sample size is, and if the sample size is large, then the McNemar test without a continuity correction and the modified Wald test can also be applied. The results have been applied to a real example on the diagnosis of coronary heart disease.

Keywords: hypothesis test; paired binomial proportions; power; sample size; type I error rate

MSC: 62P10; 62F05

check for updates

Citation: Roldán-Nofuentes, J.A.; Sheth, T.S.; Vera-Vera, J.F. Hypothesis Test to Compare Two Paired Binomial Proportions: Assessment of 24 Methods. *Mathematics* **2024**, *12*, 190. https://doi.org/10.3390/math12020190

Academic Editors: Elvira Di Nardo and Jose Luis Vicente Villardon

Received: 7 November 2023
Revised: 29 December 2023
Accepted: 1 January 2024
Published: 6 January 2024

1. Introduction

The comparison of two proportions is a topic of special interest in statistics [1], with important applications in medicine and health sciences in general. Of special interest is the case in which the two proportions are paired, as is the case in which, in a sample of n individuals, a binary variable is observed before and after a certain treatment or when the sensitivities (specificities) of two binary diagnostic tests are compared with respect to the same gold standard [2,3]. This problem also frequently arises in clinical trials [4], such as when assessing the effectiveness of a new drug or treatment. These situations give rise to the analysis of a 2 × 2 table, in which the only value set by the researcher is the sample size n. There are numerous statistical methods in the statistical literature to solve these problems. Classically, the problem has been solved by conditioning in discordant pairs and thus neglecting the frequencies of discordant pairs. This way of solving the problem has given rise to different methods, and the McNemar test [5] is the best known of all of them [6–8]. The problem can be solved with exact tests (conditioned and unconditioned) and with approximate tests (conditioned and unconditioned). All test statistics of the approximate methods are distributed approximately according to a chi-square distribution with one degree of freedom.

In the statistical literature, there are numerous methods to solve the hypothesis test to compare two paired proportions. May and Johnson [9], Park [10], and, more recently, Fagerland et al. [11–13] have compared different methods to solve this problem. However, in these works, only some of the existing methods have been studied. This is one of the main motivations for our study together with the proposal of new methods, comparing a large number of different methods to solve the hypothesis testing to compare two paired binomial proportions.

An alternative method to the hypothesis test, one directly related to it, consists of comparing the two paired proportions using confidence intervals for the difference (or ratio) of the two paired proportions. A review of more common confidence intervals can be seen in Pradhan et al. [4] and Tan et al. [14]. In addition, new intervals are proposed in Pradhan et al. [15], more recently in Fay et al. [16], and in Chan et al. [17]. A review of different methods to solve the hypothesis test as well as confidence intervals for the difference and the ratio of two paired proportions can be seen in Fagerland et al. [13].

Therefore, the purpose of this manuscript is to compare the asymptotic behaviour in terms of type I error rates and powers of different methods to solve the hypothesis test to compare two paired binomial proportions and to provide general rules of application for the methods. The rest of the article is structured as follows. Section 2 describes 24 methods to solve the hypothesis test for comparing two paired binomial proportions. Section 3 describes the criteria used to compare the asymptotic behaviour of the 24 methods. Section 4 carries out extensive Monte Carlo simulation experiments to study the type I error rates and the powers of the methods. Section 5 presents general rules of application for the methods to solve the problem posed. In Section 6, the results are applied to a real example on the diagnosis of coronary heart disease, and Section 7 discusses the conclusions obtained.

2. Notation and Methods

In general terms and focusing on common problems in the field of medicine, let us consider a binary random variable, with the categories of 'success' and 'failure', which is observed in a random sample of n individuals before and after a treatment. This situation gives rise to Table 1, where the only value set from the researcher is the sample size n. This table also shows the theoretical probabilities of each cell. The data observed in this table, $\boldsymbol{n} = (n_{11}, n_{12}, n_{21}, n_{22})^T$, were the product of a multinominal distribution with probability vector $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})^T$, verifying that $\sum p_{ij} = 1$. Variance–covariance matrix of $\boldsymbol{p}$ was as follows:

$$\sum_{\hat{\boldsymbol{p}}} = \frac{diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T}{n},$$

and the estimator of p_{ij} was $\hat{p}_{ij} = n_{ij}/n$.

Table 1. Observed frequencies and theoretical probabilities.

Observed Frequencies				
		After		
		Success	Failure	Total
Before	Success	n_{11}	n_{12}	$n_{1\cdot}$
	Failure	n_{21}	n_{22}	$n_{2\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n
Theoretical probabilities				
		After		
		Success	Failure	Total
Before	Success	p_{11}	p_{12}	$p_{1\cdot}$
	Failure	p_{21}	p_{22}	$p_{2\cdot}$
	Total	$p_{\cdot 1}$	$p_{\cdot 2}$	1

In this situation, the comparison of two paired binomial proportions consisted of solving the hypothesis test:

$$H_0 : p_{1\cdot} = p_{\cdot 1} \; vs \; H_1 : p_{1\cdot} \neq p_{\cdot 1}, \tag{1}$$

which was equivalent to solving the test:

$$H_0 : p_{12} = p_{21} \; vs \; H_1 : p_{12} \neq p_{21}. \tag{2}$$

Estimators of $p_{1\cdot}$ and $p_{\cdot 1}$ were as follows:

$$\hat{p}_{1\cdot} = \frac{n_{11} + n_{12}}{n} = \frac{n_{1\cdot}}{n} \text{ and } \hat{p}_{\cdot 1} = \frac{n_{11} + n_{21}}{n} = \frac{n_{\cdot 1}}{n}.$$

The following describes 24 statistical methods to solve this hypothesis test. Of these 24 methods, two were exact, one was quasi-exact, and 21 were approximate (of which five were new).

1. *Conditional exact test (CET)*

The probabilities p_{11} and p_{22} did not intervene in the hypothesis test (1) so that these probabilities could be ignored, as frequencies n_{11} and n_{22} could, because they did not influence the results of the hypothesis test (1). Conditioning was in the sum of discordant frequencies, i.e., an exact test was obtained using the binomial distribution [13,18]. Conditioning on $n' = n_{12} + n_{21}$, it was verified that $p_{12} + p_{21} = 1$, and therefore, n_{12} was the product of a binomial distribution of parameters n' and p_{12}, i.e., $n_{12} \to Bin(n', p_{12})$. If $H_0 : p_{12} = p_{21}$ was true, then $p_{12} = p_{21} = 1/2$, and the hypothesis test (1) was equivalent to the test:

$$H_0 : p_{12} = 1/2 \; vs \; H_1 : p_{12} \neq 1/2$$

The p-value could be calculated directly from the binomial distribution. If we assumed that $n_{12} \geq n_{21}$, then the following was derived:

$$\textit{two-sided exact p-value} = P(X \geq n_{12} \; or \; X \leq n_{21}) = 2 \times P(X \geq n_{12}),$$

where $X \to Bin(n', 1/2)$. Finally, the two-sided exact p-value for the comparison test of the two paired binomial proportions was as follows:

$$\textit{two-sided exact p-value} = 2 \times \sum_{j=0}^{Min(n_{12}, \mathrm{n}_{21})} \binom{n'}{j} \left(\frac{1}{2}\right)^{n'}. \tag{3}$$

Conditional exact test is a conservative test; that is, when H_0 is true, the p-value is typically less than α% of the time, where α is the nominal error level.

2. *Conditional exact mid-p test* (*MidpT*)

The conditional exact *mid-p* test [19] is a modification of the *CET* that consists of subtracting the probability of the observed outcome n_{12} from (3), as in the following:

$$P(X = n_{12}) = \binom{n'}{n_{12}} \left(\frac{1}{2}\right)^{n'}$$

Then, the mid-p value to compare the two proportions is as follows:

$$\textit{mid-p value} = \textit{two-sided exact p-value} - \binom{n'}{n_{12}} \left(\frac{1}{2}\right)^{n'}.$$

Conditional exact mid-p test is a less conservative method than the *CET*.

3. *McNemar Test* (*MT*)

The McNemar test [4,13,18] is the asymptotic version of the *CET*. Conditioning in $n' = n_{12} + n_{21}$ and applying the central limit theorem, the test statistic for hypothesis test (1) is as follows:

$$z = \frac{\hat{p}_{12} - \hat{p}_{21}}{\sqrt{Var(\hat{p}_{12} - \hat{p}_{21})}},$$

whose distribution is approximately a standard normal distribution and where the following occurs:

$$Var(\hat{p}_{12} - \hat{p}_{21}) = \frac{p_{12} + p_{21} - (p_{12} - p_{21})^2}{n'},$$

Since it is being conditioned in $n' = n_{12} + n_{21}$ (frequencies n_{11} and n_{22} are disregarded), then $\hat{p}_{12} = n_{12}/n'$ and $\hat{p}_{21} = n_{21}/n'$. If $H_0 : p_{12} = p_{21}$ is true, then the following are derived:

$$Var_0(\hat{p}_{12} - \hat{p}_{21}) = \frac{p_{12} + p_{21}}{n'}$$

and

$$\hat{V}ar_0(\hat{p}_{12} - \hat{p}_{21}) = \frac{n_{12} + n_{21}}{(n')^2}.$$

Substituting p_{ij} with $\hat{p}_{ij} = n_{ij}/n'$ and $Var(\hat{p}_{12} - \hat{p}_{21})$ for $\hat{V}ar_0(\hat{p}_{12} - \hat{p}_{21})$ in the expression of the test statistic z, the test statistic of the McNemar test (without continuity correction) is as follows:

$$z_{MT} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}},$$

whose distribution is approximately (it is traditionally required that $n_{12} + n_{21} > 10$) a standard normal. Very often, the test statistic is expressed in terms of the chi-square distribution:

$$\chi^2_{MT} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}},$$

whose distribution is approximately one chi-square with a degree of freedom. *MT* is a method that has good asymptotic behaviour in terms of type I error rate and power.

4. *McNemar test with Yates continuity correction* (*MTYcc*)

The McNemar test approximates the binomial distribution to the normal distribution. In this situation, it is common to apply a continuity correction (*cc*), whose objective is to improve the approximation to the normal distribution. Edwards [20] proposed the following test statistic with Yates *cc* [21]:

$$z_{MTYcc} = \frac{(\hat{p}_{12} - \hat{p}_{21}) - \frac{1}{n'}}{\sqrt{\hat{V}ar_0(\hat{p}_{12} - \hat{p}_{21})}},$$

whose distribution is approximately a standard normal distribution. It is also common to express this test statistic in terms of the chi-square distribution [13,18]:

$$\chi^2_{MTEcc} = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}.$$

5. *McNemar test with continuity correction* (*MTcc1*)

Conditioning in $n' = n_{12} + n_{21}$, the random variable $n_{12} - n_{21}$ jumps from 1 to 1, so a *cc* is 1/2 (half the jump) [22]. Therefore, another test statistic of the McNemar test with *cc* is as follows:

$$z_{MTcc1} = \frac{(\hat{p}_{12} - \hat{p}_{21}) - \frac{1}{2n'}}{\sqrt{\hat{V}ar_0(\hat{p}_{12} - \hat{p}_{21})}} = \frac{(n_{12} - n_{21}) - \frac{1}{2}}{\sqrt{\hat{V}ar_0(\hat{p}_{12} - \hat{p}_{21})}},$$

or what is the same:

$$\chi^2_{MTcc1} = \frac{(|n_{12} - n_{21}| - 0.5)^2}{n_{12} + n_{21}}.$$

This *cc* has been used by Chang et al. [17] to estimate the difference between two paired binomial proportions using confidence intervals. These authors have also proposed other continuity corrections: 0.125 and 0.25. We proposed applying these continuity corrections to the McNemar test statistics, obtaining the following new test statistics (called *MTcc*2 and *MTcc*3, respectively):

$$\chi^2_{MTcc2} = \frac{(|n_{12} - n_{21}| - 0.25)^2}{n_{12} + n_{21}} \text{ and } \chi^2_{MTcc3} = \frac{(|n_{12} - n_{21}| - 0.125)^2}{n_{12} + n_{21}}.$$

6. *Modified McNemar test* (*MMT*)

Bennett and Underwood [23] proposed a modification of the McNemar test statistic by adding 1/2 to the observed frequencies, with the aim of improving the approximation to the chi-square distribution. Thus, the test statistic is as follows:

$$\chi^2_{MMT} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21} + 1}.$$

7. *Wald test* (*WT*)

The hypothesis test (1) can be solved by applying the Wald method [24,25]. Since $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})^T$ is the probability vector of a multinomial distribution, its variance-covariance matrix is as follows:

$$\sum_{\hat{\boldsymbol{p}}} = \frac{diag(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p}}{n}.$$

The hypothesis test (2) is equivalent to checking the following:

$$H_0 : \boldsymbol{\Delta}^T\boldsymbol{p} = 0 \; vs \; H_1 : \boldsymbol{\Delta}^T\boldsymbol{p} \neq 0,$$

where

$$\boldsymbol{\Delta} = (0, 1, -1, 0)^T,$$

It is easy to verify that the estimated variance of $\hat{p}_{12} - \hat{p}_{21}$ is as follows:

$$\hat{V}ar(\hat{p}_{12} - \hat{p}_{21}) = \hat{V}ar\left(\boldsymbol{\Delta}^T\hat{\boldsymbol{p}}\right) = \hat{V}ar(\hat{p}_{12}) + \hat{V}ar(\hat{p}_{21}) - 2\hat{C}ov(\hat{p}_{12}, \hat{p}_{21}) = \frac{n_{12}(n-n_{12})+n_{21}(n-n_{21})+2n_{12}n_{21}}{n^3},$$

Applying the central limit theorem, the following is derived:

$$\frac{\hat{p}_{12} - \hat{p}_{21} - (p_{12} - p_{21})}{\sqrt{Var(\hat{p}_{12} - \hat{p}_{21})}} \to N(0,1).$$

By performing algebraic operations, it was obtained that the Wald test statistic for test (2) was as follows:

$$\chi^2_{WT} = \frac{n(n_{12} - n_{21})^2}{4n_{12}n_{21} + (n_{11} + n_{22})(n_{12} + n_{21})},$$

whose distribution was approximately a chi-square distribution with a degree of freedom.

8. *Modified Wald test* (*MWT*)

May and Johnson [9] proposed modifying the Wald test statistic by adding $1/2$ to n_{12} and to n_{21}. Thus, the modified Wald test statistic is as follows:

$$\chi^2_{MWT} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21} + 1 - \frac{(n_{12}-n_{21})^2}{n}}.$$

This method has good asymptotic behaviour and is recommended as one of the best methods to solve the hypothesis test [9].

9. *Likelihood-ratio test* (*LRT*)

The hypothesis test (1) can be solved by applying the likelihood-ratio test [26]. The likelihood function of the data is as follows:

$$L(\boldsymbol{p}, \boldsymbol{n}) = k p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}},$$

where $k = n!/\prod_{i,j=1}^{2} n_{ij}!$. If $H_0 : p_{12} = p_{21}$ is true, then it is verified that the likelihood function of the data is as follows:

$$L_0(\boldsymbol{p}, \boldsymbol{n}) = k p_{11}^{n_{11}} p_{12}^{n_{12}+n_{21}} p_{22}^{n_{22}}$$

and that the following is derived:

$$\hat{p}_{12} = \hat{p}_{21} = \frac{n_{12} + n_{21}}{2n} = \frac{n'}{2n}.$$

Applying the likelihood-ratio test [25,26], the likelihood-ratio test statistic to compare the two proportions was as follows:

$$\chi^2_{LRT} = -2\log\left(\frac{\left(\frac{n'}{2n}\right)^{n_{12}+n_{21}}}{(n_{12}/n)^{n_{12}}(n_{21}/n)^{n_{21}}}\right) = 2n_{12}\log\left(\frac{2n_{12}}{n_{12}+n_{21}}\right) + 2n_{21}\log\left(\frac{2n_{21}}{n_{12}+n_{21}}\right),$$

whose distribution was approximately one chi-square with a degree of freedom. Therefore, the test statistic of the *LRT* method only contained the frequencies of the discordant pairs.

10. *Unconditional exact test* (*UET*)

The *CET* method condition on $n' = n_{12} + n_{21}$. Suissa and Shuster [27] have proposed, from the McNemar test statistic, an exact test that uses all the observed frequencies and therefore does not condition in $n_{12} + n_{21}$. When the two proportions were compared, the power function of the test was as follows:

$$P(p_{12}, p_{21}) = \sum_{C} \binom{n}{n_{12} \quad n_{21} \quad n-m} p_{12}^{n_{12}} p_{21}^{n_{21}} (1 - p_{12} - p_{21})^{n-m},$$

where $m = n_{12} + n_{21}$ and $C = \{(n_{12}, m) : n_{12} \geq h(m);\ n_{12} = 0, 1, \ldots, m;\ m = 0, 1, \ldots, n\}$, with $h(m) = \left(z_M\sqrt{m} + m\right)/2$ and z_M as the calculated value of the McNemar statistic. If $H_0 : p_{12} = p_{21}$ was true, then the distribution of $(n_{12}, m, n-m)$ was a trinomial distribution with parameters n and probability vector $(\delta/2, \delta/2, 1-\delta)^T$, and the power function was as follows:

$$P(\delta) = \sum_{C} \binom{n}{n_{12} \quad n_{21} \quad n-m} \left(\frac{\delta}{2}\right)^m (1-\delta)^{n-m},$$

where $\delta = p_{12} + p_{21}$ was the nuisance parameter. El nuisance parameter was eliminated by maximizing this function over the range of δ. The function $P(\delta)$ was simplified as follows:

$$P(\delta) = \sum_{j=k}^{n} \binom{n}{j} \delta^{j} (1-\delta)^{n-j} F_j\left(j - i_j - 1\right),$$

where $k = \text{int}\left[z_m^2 + 1\right]$, $i_j = \text{int}[h(j)]$, $\text{int}[.]$ was the integer function and F_j was the cumulative binomial distribution function with parameters j and $1/2$. Finally, the two-sided exact p-value was calculated as follows:

$$\textit{two-sided exact p-value} = 2 \times \sup_{0<\delta<1} \{P(\delta)\}.$$

11. *Unconditional McNemar test* (*UMT*)

Lu [28] has proposed a test statistic for the McNemar test that does not condition on $n' = n_{12} + n_{21}$. Hypothesis test (1) was equivalent to the following hypothesis test:

$$H_0 : \frac{p_{12}}{p_{12} + p_{21}} = \frac{p_{21}}{p_{12} + p_{21}} \; vs \; H_1 : \frac{p_{12}}{p_{12} + p_{21}} \neq \frac{p_{21}}{p_{12} + p_{21}}.$$

If $H_0 : p_{12} = p_{21}$ was true, then n_{12} (or n_{21}) was the product of a binomial distribution with parameters n and $\delta = (p_{12} + p_{21})/2$, that is to say, $n_{12} \to Bin(n, \delta)$. The mean and variance of the estimators of this binomial distribution were as follows:

$$n\hat{\delta} = (n_{12} + n_{21})/2 \text{ and } n\hat{\delta}\left(1 - \hat{\delta}\right) = \frac{(n_{12} + n_{21})(n + n_{11} + n_{22})}{4n},$$

respectively. Approximating the normal distribution and applying the central limit theorem, the unconditional test statistic was as follows:

$$z_{UMT} = \frac{n_{12} - n\hat{\delta}}{\sqrt{n\hat{\delta}\left(1 - \hat{\delta}\right)}} = \frac{n_{12} - n_{21}}{\sqrt{\frac{(n_{12}+n_{21})(n+n_{11}+n_{22})}{n}}},$$

or rather

$$\chi^2_{UMT} = \frac{n(n_{12} - n_{21})^2}{(n_{12} + n_{21})(n + n_{11} + n_{22})},$$

whose distribution was approximately a chi-square distribution with one degree of freedom. In order to apply this method, it was required that $n_{12} + n_{21} \geq 10$, and its asymptotic behaviour was very similar to that of the *CET* [28].

12. *Unconditional likelihood-ratio test (ULRT)*

Lu [29] also proposed a likelihood-ratio test statistic to compare two binomial proportions that contain all frequencies. The likelihood-ratio test statistic is obtained in two phases: (I) the likelihood-ratio test statistic is calculates when the four n_{ij} frequencies are combined in two, n_{12} and $n_{11} + n_{21} + n_{22}$; (II) the likelihood-ratio test statistic is calculated when the four n_{ij} frequencies are combined in another two, n_{21} and $n_{11} + n_{12} + n_{22}$. Corresponding test statistics were as follows:

$$\chi^2_I = 2 \times \left[n_{12} \log\left(\frac{2n_{12}}{n_{12} + n_{21}}\right) + (n_{11} + n_{21} + n_{22}) \log\left(\frac{2(n_{11} + n_{21} + n_{22})}{2n - n_{12} - n_{21}}\right) \right]$$

and

$$\chi^2_{II} = 2 \times \left[n_{21} \log\left(\frac{2n_{21}}{n_{12} + n_{21}}\right) + (n_{11} + n_{12} + n_{22}) \log\left(\frac{2(n_{11} + n_{12} + n_{22})}{2n - n_{12} - n_{21}}\right) \right].$$

Finally, the likelihood-ratio test statistic was calculated as the mean of both likelihood-ratio test statistics:

$$\chi^2_{ULRT} = \frac{\chi^2_I + \chi^2_{II}}{2} = n_{12}\log\left(\frac{2n_{12}}{n_{12}+n_{21}}\right) + n_{21}\log\left(\frac{2n_{21}}{n_{12}+n_{21}}\right) + \\ (n-n_{12})\log\left[\frac{2(n-n_{12})}{2n-n_{12}-n_{21}}\right] + (n-n_{21})\log\left[\frac{2(n-n_{21})}{2n-n_{12}-n_{21}}\right],$$

and its distribution was approximately a chi-square distribution with one degree of freedom. The *ULRT* can be applied in most cases, although the test statistic does not fit well to the chi-square distribution when the difference between n_{12} and n_{21} is large, especially when $n_{11} + n_{22}$ is also large, and in this situation, it was a better method than the *LRT* [29].

13. *New revised version of the McNemar test* (*NMT*)

Lu et al. [30] revised the unconditional McNemar test [28]. Under the hypothesis that is no difference in the number of "success" and "failure" results between "before" and "after", the estimated probability of obtaining a "success" is as follows:

$$\hat{p} = \frac{n_{12} + n_{21} + 2n_{11}}{2n},$$

and the estimated probability of obtaining a "failure" is as follows:

$$\hat{q} = 1 - \hat{p} = \frac{n_{12} + n_{21} + 2n_{22}}{2n}.$$

Frequencies $n_{12} + n_{11}$ and $n_{21} + n_{22}$ correspond to "success" and "failure" in "before" measurements. The estimated mean is as follows:

$$\hat{\mu} = n\hat{p} = \frac{n_{12} + n_{21} + 2n_{11}}{2},$$

and the estimated standard deviation is as follows:

$$\hat{\sigma} = \sqrt{n\hat{p}\hat{q}} = \frac{1}{2}\sqrt{\frac{(n_{12} + n_{21} + 2n_{11})(n_{12} + n_{21} + 2n_{22})}{2}}.$$

Applying the central limit theorem, the statistic test was as follows:

$$z_{NMT} = \frac{n_{12} - n_{21}}{\sqrt{\frac{(n_{12}+n_{21}+2n_{11})(n_{12}+n_{21}+2n_{22})}{n}}},$$

and its distribution was approximately a standard normal distribution when $n_{12} + n_{21} + 2n_{11} \geq 10$ and $n_{12} + n_{21} + 2n_{22} \geq 10$. Alternatively, the following was derived:

$$\chi^2_{NMT} = \frac{n(n_{12} - n_{21})^2}{(n_{12} + n_{21} + 2n_{11})(n_{12} + n_{21} + 2n_{22})}.$$

This method had an asymptotic behaviour that improved that of the UMT [30].

14. *New revised version of the McNemar test with cc* (*NMTcc*)

Lu et al. [30] revised the unconditional McNemar test and proposed the following unconditional test statistic with *cc*:

$$\chi^2_{NMTcc} = \frac{n(|n_{12} - n_{21}| - 1)^2}{(n_{12} + n_{21} + 2n_{11})(n_{12} + n_{21} + 2n_{22})}.$$

15. *Haber test* (*HT*)

Haber [31] has studied the use of continuity correction in hypothesis testing, particularizing the results in 2 × 2 tables. Haber proposed a McNemar test statistic with a *cc* based on the McNemar test statistics:

$$z_{HT} = z_{MT} - \frac{\sqrt{n}}{2m},$$

where $z_{MT} = |n_{12} - n_{21}| / \sqrt{n_{12} + n_{21}}$ is the McNemar test statistic and m is the number of different values z may attain. The number of different achievable values of z_{MT} is very close to $0.9(n+1)^2/4$, and since the range of z_{MT} values is $\left[0, \sqrt{n}\right]$, the *cc* based on the average difference of the successive values gives rise to the test statistic:

$$\chi^2_{HT} = \left[\frac{|n_{12} - n_{21}|}{\sqrt{n_{12} + n_{21}}} - \frac{2\sqrt{n}}{0.9(n+1)^2}\right]^2,$$

and its distribution is approximately a chi-square with one degree of freedom.

16. *Irony et al. test* (*IT*)

Irony et al. [32] have studied the comparison of two binomial proportions from a Bayesian perspective. The Dirichlet distribution is the natural conjugate prior for $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})^T$. Therefore, the distribution for PI is a Dirichlet with parameter $\boldsymbol{a} = (a_{11}, a_{12}, a_{21}, a_{22})^T$, and its posterior distribution is also Dirichlet with parameter $\boldsymbol{A} = (A_{11}, A_{12}, A_{21}, A_{22})^T$, where $A_{ij} = a_{ij} + n_{ij}$. The objective is to solve the hypothesis test:

$$H_o : \delta = 0 \ vs \ H_o : \delta \neq 0,$$

where $\delta = p_{12} - p_{21}$. This hypothesis test is equivalent to the following:

$$H_o : \theta = \frac{1}{2} \ vs \ H_o : \theta \neq \frac{1}{2},$$

where $\theta = p_{12}/(p_{12} + p_{21})$. Therefore, the only parameters of interest are p_{12} and p_{21}, and therefore, only the trinomial data $(n_{12}, n_{21}, n_{11} + n_{22})$ are considered. Likelihood function is written as a product of two factors: one depending only on the parameter of interest θ and the other depending only on the nuisance parameter η. Distribution of θ is as follows:

$$Beta(A_{12}, A_{21}),$$

and distribution of η is as follows:

$$Beta(A_{12} + A_{21}, A_{11} + A_{22}).$$

Parameters θ and η are independent. An interval for δ is constructed by generating a large number of observations from the posterior distribution of $(p_{12}, p_{21}, 1 - \eta)$, that is, a Dirichlet distribution with parameter $(A_{12}, A_{21}, A_{11} + A_{22})$. Irony et al. [32] have shown that posterior mean of δ is as follows:

$$\frac{A_{12} + A_{21}}{A},$$

and posterior variance of δ is as follows:

$$\frac{4A_{12}A_{21} + (A_{12} + A_{21})(A_{11} + A_{22})}{(A+1)A^2}.$$

A confidence interval for δ is as follows:

$$\hat{\delta}^* \pm q\sqrt{\frac{\hat{\eta}^* - \hat{\delta}^{*2}}{n}} \pm \frac{1}{n},$$

where

$$\hat{\delta}^* = \frac{\hat{\delta}}{1+\frac{q^2}{n}},\ \hat{\delta} = \frac{n_{12}-n_{21}}{n},\ \hat{\eta}^* = \frac{\hat{\eta}}{1+\frac{q^2}{n}},\ \hat{\eta} = \frac{n_{12}+n_{21}}{n},$$

and q is the $100(1-\alpha/2)\%$ quantile of the standard normal distribution. From the previous equations, test statistic for the hypothesis test (1) was as follows:

$$\chi^2_{IT} = \begin{cases} \frac{(n_{12}-n_{21}-1)^2}{n_{12}+n_{21}}, & if\, n_{12} > n_{21} \\ \frac{(n_{12}-n_{21}+1)^2}{n_{12}+n_{21}}, & if\, n_{12} < n_{21} \\ 0\ \ , & if\, n_{12} = n_{21} \end{cases},$$

whose distribution was approximately a chi-square distribution with one degree of freedom.

17. *RR test* (*RRT*)

The hypothesis test (1) was equivalent to the hypothesis test:

$$H_0 : RR = 1\ vs\ H_1 : RR \neq 1, \tag{4}$$

where

$$RR = \frac{p_{11}+p_{12}}{p_{11}+p_{21}} = \frac{p_{1\cdot}}{p_{\cdot 1}}.$$

Lui [33] solved this hypothesis test by applying weighted least squares. Estimator of RR is as follows:

$$\widehat{RR} = \frac{n_{1\cdot}}{n_{\cdot 1}},$$

and applying the delta method the estimated variance of $\widehat{RR}$ is as follows:

$$\hat{V}ar\left[\log\left(\widehat{RR}\right)\right] = \left(\frac{\partial RR}{\partial \boldsymbol{p}}\right)_{\boldsymbol{p}=\hat{\boldsymbol{p}}} \hat{\sum}_{\hat{\boldsymbol{p}}} \left(\frac{\partial RR}{\partial \boldsymbol{p}}\right)^T_{\boldsymbol{p}=\hat{\boldsymbol{p}}} = \frac{2(1-\hat{p}^*)}{n\hat{p}^*} - \frac{2\left[\hat{p}_{11}\hat{p}_{22} - (\hat{p}^{**})^2\right]}{n(\hat{p}^*)^2},$$

where

$$\hat{\sum}_{\hat{\boldsymbol{p}}} = \frac{diag(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^T}{n},\ \hat{p}^* = \frac{n_{1\cdot} + n_{\cdot 1}}{2n},\ \text{and}\ \hat{p}^{**} = \frac{n_{12}+n_{21}}{2n}.$$

Applying the central limit theorem, the test statistic for hypothesis test (4) was as follows:

$$z_{RRT} = \frac{\log\left(\widehat{RR}\right)}{\sqrt{\hat{V}ar\left[\log\left(\widehat{RR}\right)\right]}} \underset{n\to\infty}{\to} N(0,1),$$

or equivalently

$$\chi^2_{RRT} = \frac{\left[\log\left(\widehat{RR}\right)\right]^2}{\hat{V}ar\left[\log\left(\widehat{RR}\right)\right]},$$

whose distribution was approximately a chi-square distribution with one degree of freedom.

18. *OD test* (*ODT*)

The hypothesis test (1) was also equivalent to the following:

$$H_0 : OD = 1\ vs\ H_1 : OD \neq 1, \tag{5}$$

where

$$OD = \frac{p_{12}}{p_{21}},$$

Lui [33] solved this hypothesis test by applying the same method as the one used in the *RR* test. Following an analogous procedure, the test statistic for the hypothesis test (5) is as follows:

$$\chi^2_{ODT} = \frac{\left[\log\left(\widehat{OD}\right)\right]^2}{\hat{V}ar\left[\log\left(\widehat{OD}\right)\right]},$$

and where

$$\widehat{OD} = \frac{n_{12}}{n_{21}} \text{ and } \hat{V}ar\left[\log\left(\widehat{OD}\right)\right] = \frac{1}{n\hat{p}^*}.$$

The distribution of the test statistic χ^2_{ODT} is the same as the one in the previous case.

19. *ODM test* (*ODMT*)

The hypothesis test (1) was also the same as the following:

$$H_0 : ODM = 1 \ vs \ H_1 : ODM \neq 1,$$

where

$$ODM = \frac{(p_{11} + p_{12})(p_{21} + p_{22})}{(p_{11} + p_{21})(p_{12} + p_{22})} = \frac{p_{1\cdot}p_{2\cdot}}{p_{\cdot 1}p_{\cdot 2}},$$

Applying the same method as in the two previous cases, Lui [33] proposed the following test statistic:

$$\chi^2_{ODMT} = \frac{\left[\log\left(\widehat{ODM}\right)\right]^2}{\hat{V}ar\left[\log\left(\widehat{ODM}\right)\right]},$$

where

$$\widehat{ODM} = \frac{n_{1\cdot}n_{2\cdot}}{n_{\cdot 1}n_{\cdot 2}} \text{ and } \hat{V}ar\left[\log\left(\widehat{ODM}\right)\right] = \frac{2}{n\hat{p}^*(1-\hat{p}^*)} - \frac{2\left[\hat{p}_{11}\hat{p}_{22} - (\hat{p}^{**})^2\right]}{n(\hat{p}^*)^2(1-\hat{p}^*)^2}.$$

The distribution of test statistic χ^2_{ODMT} was the same as in the previous cases.

20. *RR, OD, and ODM test with cc (RRTcc, ODTcc, and ODMTcc)*

The previous three methods can also be obtained by adding a *cc*. We proposed to add 1/2 to each one of the observed frequencies, i.e., in the following:

$$n'_{ij} = n_{ij} + 1/2.$$

Thus, the expressions of test statistics χ^2_{RRT}, χ^2_{ODT}, and χ^2_{ODMT} were replaced by $\hat{p}_{ij}$, $\hat{p}^*$, and $\hat{p}^{**}$ as follows:

$$\hat{p}'_{ij} = \frac{n'_{ij}}{n'},\ \hat{p}^{*\prime} = \frac{n'_{1\cdot} + n'_{\cdot 1}}{2n'},\ \text{and } \hat{p}^{**\prime} = \frac{n'_{12} + n'_{21}}{2n'},$$

respectively. In this way, new test statistics χ^2_{RRTcc}, χ^2_{ODTcc}, and χ^2_{ODMTcc} were obtained, and their distributions were the same as in previous cases.

3. Criteria for Comparing Methods

The comparison of the asymptotic behaviour of the methods presented in the previous section was made by comparing their type I error rates and their powers, taking as the nominal error level $\alpha = 5\%$. Based on the type I error rates and the powers, the criteria in order to choose the methods with best asymptotic behaviour were as follows:

1. The type I error rate fluctuates around $\alpha = 5\%$ without being much higher than this value, a situation that has been considered when the type I error rate is <7% .
2. The power is higher as long as the type I error rate does not exceed $\alpha = 5\%$. "Step 1" of this method to choose the method with the best asymptotic behaviour establishes that the type I error rate must be lower than 7%. Let $\Delta\alpha = \alpha - \alpha^*$, where $\alpha = 5\%$ and α^* are the type I error rates of the method. Related to a test statistic, if there is a confidence interval (CI), then $\Delta\alpha = \gamma^* - \gamma$, where $\gamma = 1 - \alpha = 0.95$ is the nominal confidence of the CI and γ^* is the coverage probability of the CI calculated. In this method, to choose test statistics, a test statistic is too liberal if $\alpha^* \geq 7\%$ $(\Delta\alpha \leq -2)$, or what amounts to the same if $\gamma^* \leq 93\%$, in which case it is said that the CI fails [34–36]. If a CI fails, then the type I error rate of the corresponding hypothesis test is $\geq 7\%$, and therefore, the hypothesis test is very liberal and leads to too many false significances.

4. Simulation Experiments

Extensive Monte Carlo simulation experiments were carried out in order to study the asymptotic behaviour, measured in terms of type I error rates and powers, of the test statistics presented in Section 2. These experiments, made with the R program [37], consisted of generating $N = 50,000$ random samples of multinomial distribution with probabilities given in Table 1 of $n = \{20, 30, 50, 100, 200\}$ sizes. Following the idea of Fagerland et al. [12], probabilities $(p_{11}, p_{12}, p_{21}, p_{22})$ have been re-parameterized as (p_{12}, p_{21}, θ), where $\theta = p_{11}p_{22}/(p_{11}p_{22})$ is the odds ratio. In order to study type I error rates, it was considered that $p_{12} = p_{21}$, and to study the powers, it was considered that $p_{12} \neq p_{21}$. Values $\{0.1, 0.2, \ldots, 0.8, 0.9\}$ were taken as values for $p_{1\cdot}$ and $p_{\cdot 1}$, and values $\{1, 2, 5, 10\}$ were considered for θ. Therefore, a wide range of values were considered to reveal the asymptotic behaviour of each test statistic. In order to calculate type I error rates and the powers, $\alpha = 5\%$ was set. Initial simulation experiments were carried out, generating $N = \{10,000; 20,000; 50,000; 100,000\}$ random samples for several scenarios, obtaining the outcome that the results for $N = \{50,000; 100,000\}$ were stable so that, finally, $N = 50,000$ was considered as a way to save computing time.

4.1. Type I Error Rates

Tables 2–5 show some of the results obtained for the type I error rates of the test statistics in different scenarios. Each scenario also shows basic descriptive statistics of $n_{12} + n_{21}$ (mean and standard deviation). By analyzing the result, the following conclusions can be drawn:

- Both the exact test (*CET* and *UET*) and the quasi-exact test *MidpT*) are conservative methods, and their type I error rates never exceed the nominal error level $\alpha = 5\%$.
- All of the McNemar test statistics (*MT*, *MTYcc*, *MTcc1*, *MTcc2*, *MTcc3*, and *MMT*) are conservative when, in general terms, $E(n_{12} + n_{21})$ is not high. The value of $E(n_{12} + n_{21})$ decreases as the value of the odds ratio θ increases, so if $\theta = 1$, all four methods are conservative when $E(n_{12} + n_{21}) \leq 21$ (rounding up to the nearest whole value), and when $\theta = 10$, all four methods are conservative when $E(n_{12} + n_{21}) \leq 12$. In each scenario, in general terms, the type I error rates of these methods fluctuate around $\alpha = 5\%$ when $E(n_{12} + n_{21})$ is higher than each one of the previous values. Likewise, continuity corrections do not improve the asymptotic behaviour of the type I error rates, especially when $E(n_{12} + n_{21})$ is high (> 20). When $E(n_{12} + n_{21})$ is small (≤ 10) or moderate $(> 10$ and $\leq 20)$, continuity corrections do not have a clear effect on the type I error rate, as sometimes it improves and sometimes it gets worse.
- *MidpT*, *MT,* and *UET* have practically the same type I error rates when $n \leq 30$.
- Test statistics *ODT* and *ODTcc* are methods that lead to many false significances since they have type I error rates that greatly exceed $\alpha = 5\%$. Therefore, both methods should not be used.
- The other approximate methods (which are unconditioned methods) are conservative when $n \leq 50$, and, in very general terms, their type I error rates fluctuate around $\alpha = 5\%$ (without being too much higher) when $n \geq 100$. Some of these methods (*WT*, *LRT*, *RRT*,

and *ODMT*) have type I error rates that fluctuate around $\alpha = 5\%$ (without being too much higher) when $n = 50$. Regarding the continuity corrections of the *RRT* and ODMT methods, they do not improve the asymptotic behaviour of their type I error rates.

Table 2. Type I error rates (in %) for $\theta = 1$ and different scenarios.

	***n* = 20**			***n* = 30**			***n* = 50**			***n* = 100**			***n* = 200**		
Method	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**
CET	0.02	1.86	0.01	0.29	1.67	2.41	1.45	2.87	3.00	2.89	3.18	3.58	3.34	3.80	3.71
MidpT	0.16	3.64	0.12	0.93	3.44	4.18	3.08	4.61	4.53	4.62	4.48	4.87	4.70	4.98	4.71
MT	0.16	3.64	0.12	0.93	3.44	4.18	3.08	4.82	5.28	5.05	5.08	4.91	4.99	5.04	4.95
UET	0.16	3.62	0.12	0.93	3.44	4.18	3.08	4.61	4.53	4.13	4.27	4.38	4.70	4.82	4.63
MTYcc	0.02	1.86	0.01	0.29	1.65	2.21	1.43	2.46	2.76	2.52	3.17	3.42	3.27	3.64	3.67
MTcc1	0.16	3.64	0.12	0.93	4.60	0.87	3.08	4.52	3.08	4.62	5.00	4.55	4.93	4.95	5.07
MTcc2	0.16	3.64	0.12	0.93	4.67	0.87	3.08	4.80	3.08	4.98	5.00	4.94	4.93	5.05	5.08
MTcc3	0.16	3.64	0.12	0.93	4.67	0.87	3.08	4.80	3.08	4.98	5.00	4.94	4.93	5.05	5.08
MMT	0.16	3.62	0.12	0.93	3.40	3.83	3.04	3.98	4.30	4.13	4.46	4.59	4.59	4.84	4.63
WT	0.72	6.62	0.66	2.72	6.38	6.39	5.92	5.68	5.42	5.39	5.24	5.38	5.27	5.19	5.01
MWT	0.16	5.69	0.13	0.93	3.89	5.17	3.08	4.82	5.28	4.69	5.05	4.91	4.93	5.04	4.95
LRT	0.72	6.49	0.66	2.72	6.38	6.25	5.92	5.68	5.35	5.49	5.23	5.12	5.27	5.15	4.95
UMT	0.00	0.87	0.00	0.01	0.42	1.08	0.26	0.88	1.30	0.70	1.07	1.33	0.70	1.07	1.28
ULRT	0.00	0.87	0.00	0.01	0.42	1.08	0.26	0.88	1.30	0.70	1.07	1.33	0.70	1.07	1.28
NMT	0.00	0.53	0.00	0.01	0.30	0.51	0.19	0.49	0.53	0.54	0.54	0.56	0.52	0.55	0.53
NMTcc	0.00	0.18	0.00	0.00	0.07	0.19	0.05	0.19	0.26	0.20	0.29	0.34	0.32	0.37	0.38
HT	0.16	3.64	0.12	0.93	3.44	4.18	3.08	4.61	4.54	4.98	4.88	4.91	4.99	5.04	4.95
IT	0.02	1.86	0.01	0.29	1.65	2.21	1.43	2.46	2.76	2.52	3.17	3.42	3.27	3.64	3.67
RRT	2.53	4.83	0.12	5.38	6.61	6.13	6.88	6.24	5.49	6.34	5.49	5.37	5.67	5.35	5.01
ODT	9.77	18.43	9.83	15.94	18.50	18.47	18.49	17.72	16.96	17.44	17.43	17.52	17.49	16.78	17.00
ODMT	0.72	5.39	0.66	2.72	6.28	5.83	5.92	5.92	5.37	6.21	5.33	5.27	5.60	5.26	4.99
RRTcc	0.16	3.15	0.01	0.93	3.88	0.29	3.31	4.38	3.04	5.33	4.62	3.98	5.17	4.85	4.73
ODTcc	2.96	13.91	2.94	9.30	15.01	9.38	15.88	15.60	15.92	16.61	16.43	16.64	16.50	16.37	16.85
ODMTcc	0.16	3.16	0.12	0.93	3.89	0.87	3.31	4.38	3.28	5.15	4.62	5.07	5.10	4.85	5.26
$\boldsymbol{E}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	4.20	10.02	4.19	5.72	9.64	12.61	9.08	16.01	21.01	18.02	32.02	42.01	36.02	63.98	83.97
$\boldsymbol{SD}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	1.51	2.22	1.50	1.97	2.53	2.70	2.68	3.30	3.49	3.85	4.67	4.93	5.43	6.59	6.97

Scen. 1: $p_{1\cdot} = p_{\cdot 1} = 10\%$. Scen. 2: $p_{1\cdot} = p_{\cdot 1} = 50\%$. Scen. 3: $p_{1\cdot} = p_{\cdot 1} = 90\%$.

Table 3. Type I error rates (in %) for $\theta = 2$ and different scenarios.

	***n* = 20**			***n* = 30**			***n* = 50**			***n* = 100**			***n* = 200**		
Method	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**
CET	0.01	0.31	0.72	0.21	1.27	1.98	1.22	2.58	2.97	2.83	3.04	3.35	3.20	3.70	3.76
MidpT	0.10	1.02	2.09	0.73	2.86	3.90	2.76	4.44	4.66	4.57	4.53	4.73	4.63	4.91	4.87
MT	0.10	1.02	2.09	0.73	2.86	3.90	2.76	4.51	5.08	4.84	5.27	5.00	5.09	4.91	5.01
UET	0.10	1.02	2.09	0.73	2.86	3.90	2.76	4.44	4.66	3.98	4.44	4.35	4.62	4.75	4.77
MTYcc	0.01	0.31	0.72	0.21	1.27	1.94	1.22	2.29	2.53	2.46	3.03	3.29	3.17	3.65	3.63
MTcc1	0.10	2.88	0.07	0.73	4.24	0.62	2.76	4.50	2.68	4.57	4.91	4.72	4.88	5.01	5.20
MTcc2	0.10	2.88	0.07	0.73	4.25	0.62	2.76	5.08	2.68	4.82	4.91	4.95	4.90	5.01	5.23
MTcc3	0.10	2.88	0.07	0.73	4.25	0.62	2.76	5.08	2.68	4.82	4.91	4.95	4.90	5.01	5.23
MMT	0.10	1.02	2.09	0.73	2.85	3.78	2.73	3.90	4.15	3.98	4.52	4.64	4.58	4.84	4.77
WT	0.53	3.01	4.83	2.19	5.77	6.55	5.59	6.02	5.48	5.38	5.31	5.27	5.25	5.22	5.04
MWT	0.11	1.23	2.78	0.73	3.10	4.61	2.76	4.51	5.08	4.59	5.16	5.00	4.88	4.91	5.01
LRT	0.53	3.01	4.83	2.19	5.77	6.51	5.59	6.02	5.47	5.57	5.31	5.24	5.25	5.08	5.03
UMT	0.01	0.09	0.25	0.01	0.27	0.65	0.18	0.72	1.07	0.65	0.96	1.15	0.71	0.95	1.12

Table 3. *Cont.*

	***n* = 20**			***n* = 30**			***n* = 50**			***n* = 100**			***n* = 200**		
Method	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**
ULRT	0.01	0.09	0.25	0.01	0.27	0.65	0.18	0.72	1.07	0.65	0.96	1.15	0.71	0.95	1.12
NMT	0.00	0.05	0.11	0.01	0.14	0.23	0.10	0.28	0.26	0.37	0.32	0.27	0.38	0.33	0.26
NMTcc	0.00	0.01	0.03	0.00	0.03	0.08	0.02	0.09	0.11	0.11	0.15	0.14	0.19	0.20	0.16
HT	0.10	1.02	2.09	0.73	2.86	3.90	2.76	4.44	4.66	4.82	4.90	4.94	5.09	4.91	5.01
IT	0.01	0.31	0.72	0.21	1.27	1.94	1.22	2.29	2.53	2.46	3.03	3.29	3.17	3.65	3.63
RRT	1.95	4.11	4.82	4.21	5.95	5.81	6.42	6.13	5.49	6.24	5.38	5.27	5.58	5.24	5.03
ODT	8.58	16.63	18.42	14.85	18.63	18.44	18.45	18.13	17.45	17.60	17.24	17.64	17.53	17.01	17.09
ODMT	0.53	2.98	4.37	2.19	5.49	5.40	5.55	5.81	5.41	6.08	5.32	5.25	5.48	5.14	5.03
RRTcc	0.10	2.18	0.01	0.73	3.77	0.19	2.88	4.23	2.66	5.08	4.59	4.09	5.11	4.67	4.88
ODTcc	2.47	12.33	3.00	7.83	15.55	8.06	15.52	16.26	15.48	16.38	16.45	16.52	16.61	15.57	16.95
ODMTcc	0.10	2.19	0.07	0.73	3.88	0.62	2.88	4.23	2.78	4.86	4.59	4.99	5.06	4.67	5.37
$\boldsymbol{E}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	3.99	5.86	7.25	5.36	8.49	10.72	8.41	14.04	17.83	16.63	28.08	35.66	33.26	56.13	71.25
$\boldsymbol{SD}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	1.44	1.89	2.08	1.88	2.42	2.61	2.58	3.17	3.38	3.73	4.49	4.79	5.27	6.35	6.80

Scen. 1: $p_{1\cdot} = p_{\cdot 1} = 10\%$. Scen. 2: $p_{1\cdot} = p_{\cdot 1} = 50\%$. Scen. 3: $p_{1\cdot} = p_{\cdot 1} = 90\%$.

Table 4. Type I error rates (in %) for $\theta = 5$ and different scenarios.

	***n* = 20**			***n* = 30**			***n* = 50**			***n* = 100**			***n* = 200**		
Method	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**
CET	0.01	0.08	0.26	0.09	0.61	1.18	0.74	2.04	2.55	2.58	2.99	3.00	3.05	3.57	3.69
MidpT	0.04	0.43	0.88	0.38	1.72	2.78	2.02	3.93	4.44	4.48	4.46	4.45	4.54	4.94	4.91
MT	0.04	0.43	0.88	0.38	1.72	2.78	2.02	3.93	4.51	4.56	5.28	5.28	5.26	4.96	4.92
UET	0.04	0.43	0.88	0.38	1.72	2.78	2.02	3.93	4.44	3.99	4.29	4.38	4.54	4.82	4.71
MTYcc	0.01	0.08	0.26	0.09	0.61	1.17	0.74	1.98	2.28	2.29	2.81	2.98	3.03	3.40	3.63
MTcc1	0.04	1.40	0.03	0.38	3.27	0.35	2.02	4.58	1.86	4.48	4.89	4.35	4.77	4.95	4.94
MTcc2	0.04	1.40	0.03	0.38	3.27	0.35	2.02	4.72	1.86	4.56	4.94	4.42	4.94	4.95	5.08
MTcc3	0.04	1.40	0.03	0.38	3.27	0.35	2.02	4.72	1.86	4.56	4.94	4.42	4.94	4.95	5.08
MMT	0.04	0.43	0.88	0.38	1.72	2.77	2.01	3.73	3.93	3.99	4.29	4.43	4.52	4.73	4.83
WT	0.31	1.52	2.71	1.40	4.22	5.56	4.62	6.42	6.23	5.32	5.34	5.30	5.30	5.26	5.31
MWT	0.04	0.47	1.06	0.38	1.76	2.96	2.02	3.93	4.51	4.48	4.78	5.10	4.76	4.96	4.92
LRT	0.31	1.52	2.71	1.40	4.22	5.56	4.62	6.42	6.23	6.01	5.34	5.30	5.30	5.16	5.06
UMT	0.01	0.01	0.07	0.00	0.07	0.24	0.09	0.46	0.70	0.54	0.84	0.95	0.63	0.79	0.94
ULRT	0.01	0.01	0.07	0.00	0.07	0.24	0.09	0.46	0.70	0.54	0.84	0.95	0.63	0.79	0.94
NMT	0.00	0.00	0.02	0.00	0.04	0.04	0.03	0.07	0.06	0.15	0.10	0.07	0.15	0.08	0.06
NMTcc	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.03	0.03	0.06	0.05	0.04	0.08	0.06	0.05
HT	0.04	0.43	0.88	0.38	1.72	2.78	2.02	3.93	4.44	4.56	5.03	4.88	5.26	4.96	4.92
IT	0.01	0.08	0.26	0.09	0.61	1.17	0.74	1.98	2.28	2.29	2.81	2.98	3.03	3.40	3.63
RRT	1.02	2.05	2.57	2.58	4.16	4.09	5.21	5.54	5.44	6.12	5.34	5.30	5.36	5.17	5.05
ODT	6.36	13.18	16.36	12.35	17.83	18.61	17.69	18.38	18.23	18.07	17.07	17.23	17.23	17.44	17.11
ODMT	0.31	1.46	2.17	1.39	3.57	3.57	4.49	5.14	5.35	6.00	5.32	5.30	5.33	5.08	5.00
RRTcc	0.04	0.83	0.00	0.38	2.71	0.09	2.04	3.93	1.86	4.72	4.50	3.80	4.96	4.73	4.59
ODTcc	1.51	9.05	1.34	5.39	16.40	5.47	13.45	16.20	13.71	16.14	16.96	15.91	16.82	16.25	17.20
ODMTcc	0.04	0.82	0.03	0.38	3.09	0.35	2.04	3.93	1.90	4.55	4.51	4.41	4.89	4.73	5.05
$\boldsymbol{E}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	3.61	4.86	5.71	4.72	6.83	8.25	7.21	11.12	13.62	14.07	22.19	27.22	28.12	44.36	54.41
$\boldsymbol{SD}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	1.31	1.69	1.86	1.72	2.18	2.38	2.37	2.92	3.14	3.48	4.16	4.45	4.92	5.87	6.31

Scen. 1: $p_{1\cdot} = p_{\cdot 1} = 10\%$. Scen. 2: $p_{1\cdot} = p_{\cdot 1} = 50\%$. Scen. 3: $p_{1\cdot} = p_{\cdot 1} = 90\%$.

Table 5. Type I error rates (in %) for $\theta = 10$ and different scenarios.

	n = 20			*n* = 30			*n* = 50			*n* = 100			*n* = 200		
Method	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**	**Scen. 1**	**Scen. 2**	**Scen. 3**
CET	0.01	0.02	0.07	0.04	0.29	0.52	0.44	1.44	1.98	2.17	2.87	2.99	2.96	3.35	3.55
MidpT	0.01	0.15	0.35	0.19	0.90	1.56	1.26	3.00	3.85	4.05	4.68	4.56	4.50	4.68	4.88
MT	0.01	0.15	0.35	0.19	0.90	1.56	1.26	3.00	3.85	4.07	5.10	5.33	5.31	4.98	4.91
UET	0.01	0.15	0.35	0.19	0.90	1.56	1.26	3.00	3.85	3.76	4.19	4.30	4.50	4.67	4.79
MTYcc	0.01	0.02	0.07	0.04	0.29	0.52	0.44	1.42	1.93	2.04	2.49	2.74	2.86	3.28	3.40
MTcc1	0.01	0.55	0.01	0.19	2.11	0.13	1.26	4.17	1.13	4.05	4.57	3.98	4.61	4.96	4.69
MTcc2	0.01	0.55	0.01	0.19	2.11	0.13	1.26	4.18	1.13	4.07	4.96	3.99	5.00	4.96	5.09
MTcc3	0.01	0.55	0.01	0.19	2.11	0.13	1.26	4.18	1.13	4.07	4.96	3.99	5.00	4.96	5.09
MMT	0.01	0.15	0.35	0.19	0.90	1.56	1.26	2.97	3.68	3.76	4.19	4.30	4.37	4.60	4.69
WT	0.13	0.74	1.37	0.79	2.60	3.96	3.42	5.82	6.45	4.83	5.49	5.41	5.34	5.22	5.20
MWT	0.01	0.16	0.38	0.19	0.91	1.59	1.26	3.00	3.85	4.05	4.74	4.82	4.48	4.89	4.91
LRT	0.13	0.74	1.37	0.79	2.60	3.96	3.42	5.82	6.45	6.21	5.61	5.42	5.34	5.22	5.14
UMT	0.00	0.01	0.01	0.00	0.01	0.06	0.04	0.22	0.41	0.40	0.68	0.79	0.56	0.72	0.81
ULRT	0.00	0.01	0.01	0.00	0.01	0.06	0.04	0.22	0.41	0.40	0.68	0.79	0.56	0.72	0.81
NMT	0.00	0.00	0.02	0.00	0.01	0.01	0.02	0.03	0.02	0.06	0.04	0.03	0.06	0.03	0.01
NMTcc	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.04	0.02	0.00
HT	0.01	0.15	0.35	0.19	0.90	1.56	1.26	3.00	3.85	4.07	5.04	5.11	5.31	4.98	4.91
IT	0.01	0.02	0.07	0.04	0.29	0.52	0.44	1.42	1.93	2.04	2.49	2.74	2.86	3.28	3.40
RRT	0.46	0.99	1.23	1.48	2.39	2.33	3.78	4.11	4.53	5.68	5.42	5.33	5.36	5.14	4.92
ODT	4.52	9.56	12.62	9.57	15.69	17.51	16.66	18.49	18.42	18.24	17.51	17.19	17.04	17.61	17.68
ODMT	0.13	0.68	0.95	0.78	1.97	1.98	3.11	3.74	4.42	5.58	5.33	5.33	5.34	5.04	4.91
RRTcc	0.01	0.30	0.00	0.19	1.52	0.02	1.27	3.86	1.13	4.13	4.38	3.71	4.64	4.59	4.47
ODTcc	0.82	5.75	0.78	3.4	13.98	3.35	10.83	15.99	10.70	16.01	17.28	15.86	16.95	16.53	17.47
ODMTcc	0.01	0.30	0.01	0.19	1.93	0.13	1.27	3.86	1.13	4.04	4.38	3.97	4.59	4.61	4.68
$\boldsymbol{E(n_{12}+n_{21})}$	3.30	4.16	4.73	4.19	5.66	6.62	6.17	8.97	10.74	11.80	17.79	21.41	23.55	35.57	42.79
$\boldsymbol{SD(n_{12}+n_{21})}$	1.18	1.50	1.66	1.55	1.95	2.14	2.15	2.66	2.88	3.21	3.82	4.10	4.65	5.41	5.82

Scen. 1: $p_{1\cdot} = p_{\cdot 1} = 10\%$. Scen. 2: $p_{1\cdot} = p_{\cdot 1} = 50\%$. Scen. 3: $p_{1\cdot} = p_{\cdot 1} = 90\%$.

4.2. Powers

Tables 6–9 show some of the results obtained for the power of the test statistics in different scenarios. These tables do not show the results for the test statistics *ODT* and *ODTcc* since their type I error rates are very clearly higher than $\alpha = 5\%$. For each scenario we can also see the basic descriptive statistic of $n_{12} + n_{21}$. From the analysis of the results, the following conclusions are obtained:

- *UET* and *MidpT* have very similar powers, and both are a little more powerful than *CET*, especially when the sample size is small ($n = \{20, 30, 50\}$).
- The classic McNemar test statistic without cc (*MT*) has the same power as the three McNemar test statistics with cc (*MTcc1*, *MTcc2*, and *MTcc3*), and all of them are more powerful than the McNemar test statistic with Yates *cc* (*MTYcc*).
- Methods *MT*, *MTcc1*, *MTcc2*, and *MTcc3* have the same power as *UET* and as *MidpT* when $n \leq 30$.
- Regarding the approximate tests, in general terms, the *WT*, *LRT*, *RRT*, and *ODMT* methods have more power than the other approximate tests, especially when $n \leq 50$. When $n \geq 100$, if the difference between $p_{1\cdot}$ and $p_{\cdot 1}$ is small (for example, $p_{1\cdot} - p_{\cdot 1} = 10\%$), then the *WT*, *LRT*, *RRT*, and *ODMT* methods have more power than the rest of the approximate methods; if the difference between $p_{1\cdot}$ and $p_{\cdot 1}$ is greater (for example, $p_{1\cdot} - p_{\cdot 1} \geq 40\%$), then all of the approximate methods have very similar powers. The continuity corrections in the *RRT* and *ODMT* methods do not improve their powers.

Table 6. Powers (in %) for $p_{1\cdot} = 0.10$, $p_{\cdot 1} = 0.20$, and $\theta = 1$.

Method	*n* = 20	*n* = 30	*n* = 50	*n* = 100	*n* = 200
CET	0.89	5.49	18.58	42.45	77.04
MidpT	2.63	10.38	25.74	48.94	80.68
MT	2.63	10.38	25.83	52.02	80.68
UET	2.63	10.38	25.74	48.71	80.15
MTYcc	0.89	5.49	17.59	42.27	76.62
MTcc1	2.63	10.38	25.74	49.61	80.68
MTcc2	2.63	10.38	25.83	50.47	80.68
MTcc3	2.63	10.38	25.83	50.47	80.68
MMT	2.63	10.36	24.12	48.81	80.37
WT	6.85	17.79	30.93	52.08	81.42
MWT	3.04	10.86	25.83	51.16	80.68
LRT	6.85	17.78	30.93	52.08	80.92
UMT	0.22	1.40	8.22	26.72	57.77
ULRT	0.22	1.40	8.22	26.72	57.77
NMT	0.19	1.24	6.16	21.50	51.22
NMTcc	0.05	0.36	3.13	14.74	45.96
HT	2.63	10.38	25.74	50.47	80.68
IT	0.89	5.49	17.59	42.27	76.62
RRT	11.43	20.05	31.65	53.80	81.73
ODMT	6.85	17.77	31.38	52.89	81.72
RRTcc	2.63	10.66	26.81	51.66	80.78
ODMTcc	2.63	10.38	26.76	51.18	80.70
$\boldsymbol{E(n_{12}+n_{21})}$	5.57	7.99	13.06	26.01	51.92
$\boldsymbol{SD(n_{12}+n_{21})}$	1.84	2.34	3.08	4.39	6.19

Table 7. Powers (in %) for $p_{1\cdot} = 0.20$, $p_{\cdot 1} = 0.80$, and $\theta = 2$.

Method	*n* = 20	*n* = 30	*n* = 50	*n* = 100	*n* = 200
CET	93.70	99.81	100	100	100
MidpT	96.72	99.92	100	100	100
MT	96.72	99.92	100	100	100
UET	96.66	99.92	100	100	100
MTYcc	93.67	99.74	100	100	100
MTcc1	96.72	99.92	100	100	100
MTcc2	96.72	99.92	100	100	100
MTcc3	96.72	99.92	100	100	100
MMT	96.66	99.89	100	100	100
WT	98.46	99.95	100	100	100
MWT	98.15	99.94	100	100	100
LRT	98.40	99.94	100	100	100

Table 7. *Cont.*

Method	*n* = 20	*n* = 30	*n* = 50	*n* = 100	*n* = 200
UMT	88.88	99.48	100	100	100
ULRT	88.88	99.48	100	100	100
NMT	84.18	98.56	100	100	100
NMTcc	71.75	96.96	100	100	100
HT	96.72	99.92	100	100	100
IT	93.67	99.74	100	100	100
RRT	97.60	99.94	100	100	100
ODMT	97.94	99.94	100	100	100
RRTcc	96.20	99.89	100	100	100
ODMTcc	96.20	99.89	100	100	100
$\boldsymbol{E}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	13.26	19.72	32.64	65.00	129.90
$\boldsymbol{SD}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	2.09	2.57	3.35	4.76	6.74

Table 8. Powers (in %) for $p_{1\cdot} = 0.10$, $p_{\cdot 1} = 0.50$, and $\theta = 5$.

Method	*n* = 20	*n* = 30	*n* = 50	*n* = 100	*n* = 200
CET	53.17	91.78	99.87	100	100
MidpT	68.95	95.66	99.92	100	100
MT	68.95	95.66	99.92	100	100
UET	68.95	95.66	99.92	100	100
MTYcc	53.17	91.73	99.83	100	100
MTcc1	68.95	95.66	99.92	100	100
MTcc2	68.95	95.66	99.92	100	100
MTcc3	68.95	95.66	99.92	100	100
MMT	68.95	95.56	99.91	100	100
WT	82.00	97.80	99.94	100	100
MWT	72.05	96.13	99.92	100	100
LRT	82.00	97.79	99.94	100	100
UMT	36.37	80.43	99.50	100	100
ULRT	36.37	80.43	99.50	100	100
NMT	25.52	67.48	97.21	100	100
NMTcc	14.07	52.93	94.56	100	100
HT	68.95	95.66	99.92	100	100
IT	53.17	91.73	99.83	100	100
RRT	82.31	97.51	99.95	100	100
ODMT	80.95	97.22	99.94	100	100
RRTcc	68.91	95.65	99.92	100	100
ODMTcc	68.69	95.60	99.91	100	100
$\boldsymbol{E}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	9.22	13.54	22.20	43.89	87.49
$\boldsymbol{SD}(\boldsymbol{n}_{12}+\boldsymbol{n}_{21})$	2.15	2.66	3.48	4.94	7.02

Table 9. Powers (in %) for $p_{1\cdot} = 0.30$, $p_{\cdot 1} = 0.70$, and $\theta = 10$.

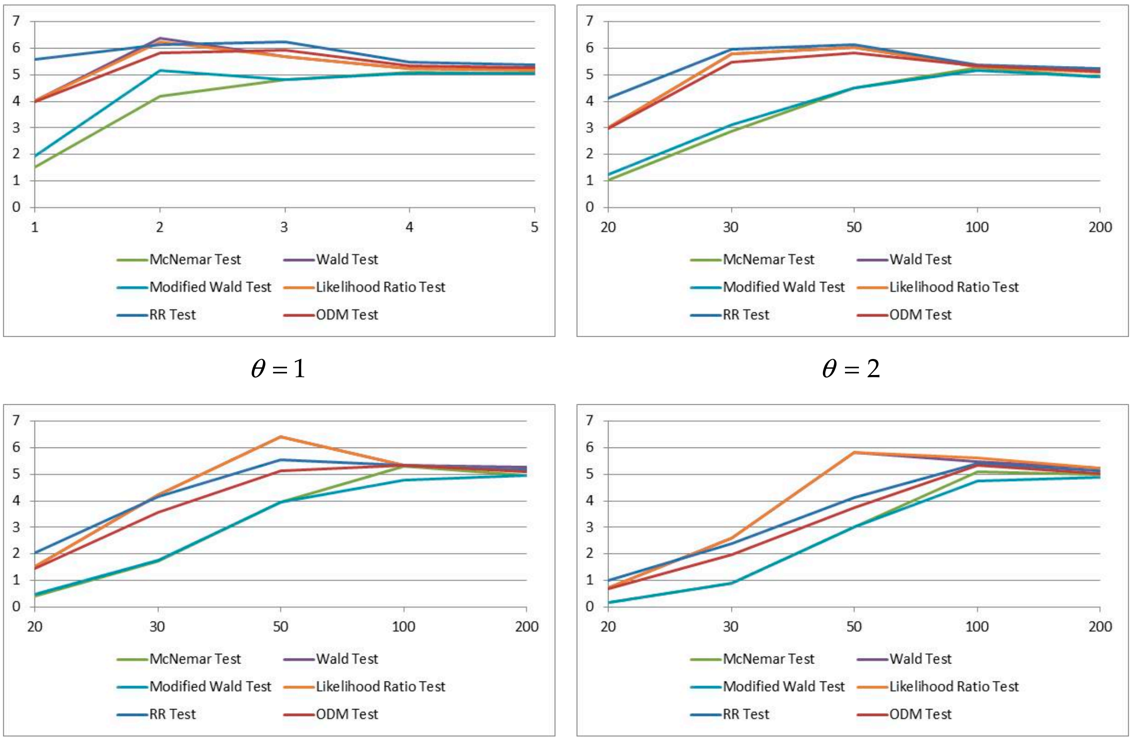
Method	n = 20	n = 30	n = 50	n = 100	n = 200
CET	53.03	91.52	99.87	100	100
MidpT	69.13	95.38	99.93	100	100
MT	69.13	95.38	99.93	100	100
UET	69.13	95.38	99.93	100	100
MTYcc	53.03	91.44	99.83	100	100
MTcc1	69.13	95.38	99.93	100	100
MTcc2	69.13	95.38	99.93	100	100
MTcc3	69.13	95.38	99.93	100	100
MMT	69.13	95.32	99.92	100	100
WT	81.78	97.60	99.96	100	100
MWT	72.10	95.93	99.93	100	100
LRT	81.78	97.60	99.95	100	100
UMT	36.28	80.49	99.39	100	100
ULRT	36.28	80.49	99.39	100	100
NMT	22.60	56.57	94.04	100	100
NMTcc	11.43	42.46	90.09	100	100
HT	69.13	95.38	99.93	100	100
IT	53.03	91.44	99.83	100	100
RRT	71.63	95.93	99.94	100	100
ODMT	72.24	95.93	99.93	100	100
RRTcc	60.79	94.49	99.92	100	100
ODMTcc	60.82	95.07	99.92	100	100
$\boldsymbol{E(n_{12} + n_{21})}$	9.21	13.56	22.23	43.95	87.58
$\boldsymbol{SD(n_{12} + n_{21})}$	2.16	2.67	3.47	4.96	6.96

5. General Rules of Application

From the results obtained in the simulation experiments and only considering sample size n (as it is the only parameter set by the researcher), one can provide the following general rules of application for the test statistics:

- When the sample size is small, use the *WT*, *LRT*, *RRT,* or *ODMT* methods; since they are the least conservative methods, they have the greatest power, and their powers are similar.
- When the sample size is moderate, use the *WT*, *LRT*, *RRT,* or *ODMT* methods; since their type I error rates fluctuate around $\alpha = 5\%$, they have the greatest power, and their powers are similar.
- When the sample size is large, use the *MT*, *WT*, *MWT*, *LRT*, *RRT,* and *ODMT* methods; since their type I error rates fluctuate around $\alpha = 5\%$, they have the greatest power, and the powers of these methods are very similar.

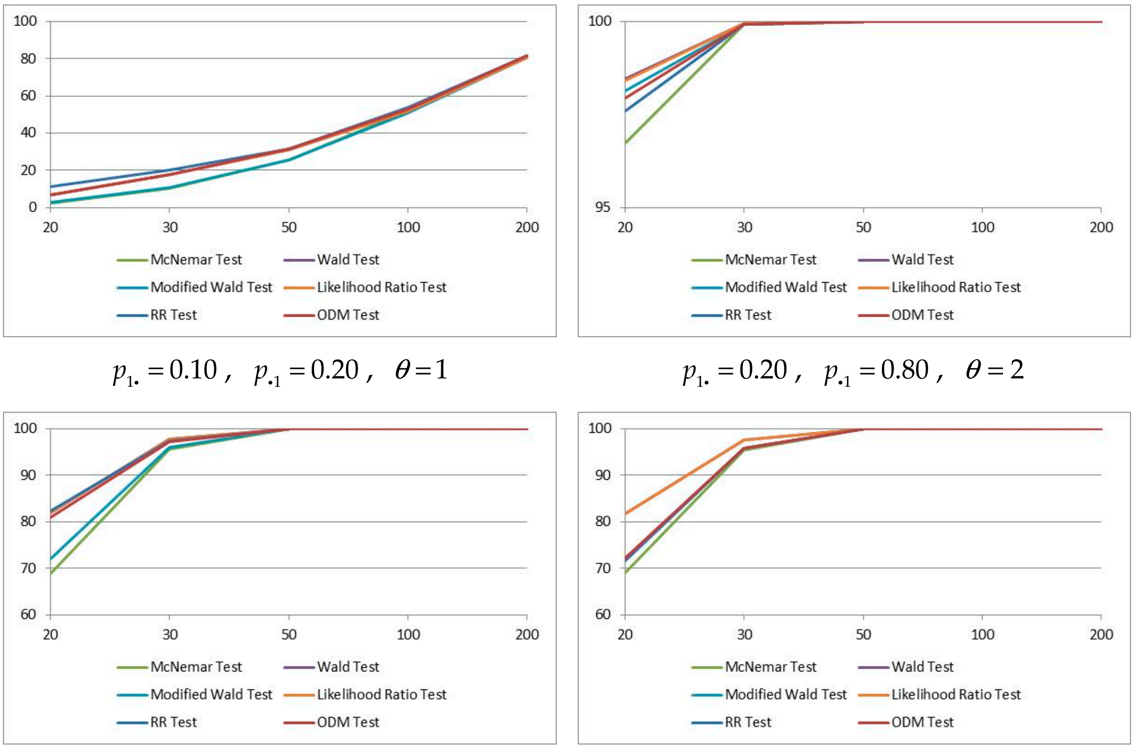
The graphs in Figure 1 show the type I error rates of the selected methods, and the graphs in Figure 2 show the powers of these methods for different scenarios. The graphs in Figure 1 show how the *WT*, *LRT*, *RRT,* and *ODMT* methods have a type I error rate with better behaviour than the *MT* and *MWT* methods when the sample size is small or moderate, with their values being very similar when the sample is large. In the graphs in Figure 2, it can be seen that the power of *MT* is a little lower than that of the other methods when the sample size is small. Likewise, the powers of these methods are very similar when the sample size is moderate or large.



$\theta = 1$ $\theta = 2$

$\theta = 5$ $\theta = 10$

Figure 1. Type I error rates of the methods for $p_{1.} = p_{.1} = 20\%$.



$p_{1.} = 0.10, \quad p_{.1} = 0.20, \quad \theta = 1$ $p_{1.} = 0.20, \quad p_{.1} = 0.80, \quad \theta = 2$

$p_{1.} = 0.10, \quad p_{.1} = 0.50, \quad \theta = 5$ $p_{1.} = 0.30, \quad p_{.1} = 0.70, \quad \theta = 10$

Figure 2. Powers of the methods for different scenarios.

6. Example

The results have been applied to the diagnosis of coronary artery disease using dobutamine echocardiography (*DE*, test 1) and myocardial perfusion scintigraphy (*MPS*, test 2) as diagnostic tests and coronary angiography (*CA*) as the gold standard. The objective of this study is to compare the sensitivities (specificities) of the two diagnostic tests. Table 10 shows the frequencies observed in the study, the estimate of each sensitivity (*Se*) and of each specificity (*Sp*), and the results of each method to resolve the respective comparisons. The comparison of the two sensitivities (specificities) has been carried out using the function "pairedProp", which is a function written in R that allows for comparing two paired binomial proportions using the methods recommended in Section 4. This function is attached as a Supplementary Material to the manuscript. The sentence to compare the two sensitivities is as follows:

$$\text{pairedProp}(152, 17, 7, 36),$$

and the sentence to compare the two specificities is as follows:

$$\text{pairedProp}(25, 10, 11, 290).$$

Table 10. Diagnosis of coronary artery disease: frequencies and results of comparisons of sensitivities and specificities.

Observed Frequencies					
	Positive *DE*		Negative *DE*		
	Positive *MPS*	Negative *MPS*	Positive *MPS*	Negative *MPS*	Total
Positive *CA*	152	17	7	36	212
Negative *CA*	25	10	11	290	336
Total	177	27	18	326	548
Comparison of sensitivities: $H_0 : Se_1 = Se_2$ *vs* $H_1 : Se_1 \neq Se_2$					
MT	*WT*	*MWT*	*LRT*	*RRT*	*ODMT*
$\chi^2 = 4.167$ $p\text{-}value = 0.041$	$\chi^2 = 4.250$ $p\text{-}value = 0.039$	$\chi^2 = 4.077$ $p\text{-}value = 0.0403$	$\chi^2 = 4.296$ $p\text{-}value = 0.038$	$\chi^2 = 4.169$ $p\text{-}value = 0.041$	$\chi^2 = 4.191$ $p\text{-}value = 0.041$
Comparison of specificities: $H_0 : Sp_1 = Sp_2$ *vs* $H_1 : Sp_1 \neq Sp_2$					
MT	*WT*	*MWT*	*LRT*	*RRT*	*ODMT*
$\chi^2 = 0.048$ $p\text{-}value = 0.827$	$\chi^2 = 0.048$ $p\text{-}value = 0.827$	$\chi^2 = 0.045$ $p\text{-}value = 0.831$	$\chi^2 = 0.048$ $p\text{-}value = 0.827$	$\chi^2 = 0.048$ $p\text{-}value = 0.827$	$\chi^2 = 0.048$ $p\text{-}value = 0.827$

In this example, the number of patients with coronary artery disease and the number of patients without coronary artery disease are large, and therefore, all the methods indicated in Section 4 can be applied. The estimates of the sensitivities and specificities of the diagnostic tests are as follows: $\hat{S}e_1 = 0.797$, $\hat{S}e_2 = 0.750$, $\hat{S}p_1 = 1 - 0.104 = 0.896$, and $\hat{S}p_2 = 1 - 0.107 = 0.896$. With fixed $\alpha = 5\%$, the equality of the two sensitivities is rejected, and the equality of the two specificities is not rejected. It is concluded that the sensitivity of the DE test is significantly greater than the sensitivity of the MPS test.

In this example, it can be seen that the *p*-values of all the methods to compare the two sensitivities (specificities) are very similar to each other, and therefore, the conclusions are the same.

7. Discussion

The comparison of two paired binomial proportions is a problem that appears frequently in medical and clinical studies. In the statistical literature, there are diverse methods proposed to solve this hypothesis test, and therefore, it is necessary to determine which methods have the best asymptotic behaviour in terms of the type I error rate and power.

We reviewed 19 existing methods and proposed 5 new ones, and we carried out broad simulation experiments to study their asymptotic behaviour. From the results obtained, we have given some general rules of application for the methods studied.

May and Johnson [9] compared through simulation experiments the asymptotic behaviour of eight methods (*CET*, *MidpT*, *MT*, *MTYcc*, *MMT*, *WT*, *MWT*, and *LRT*) and recommended using the *MidpT*, *MWT*, and *MT* methods when it is verified that $n_{12} + n_{21} \leq 40$. May and Johnson used the criterion that the type I error rate must not be higher than $\alpha = 5\%$.

Park [10] has compared, using the same criteria as May and Johnson, the asymptotic behavior of the *CET*, *MT*, *LRT*, and *WT* methods, concluding that the method with the best behavior is the *MT*.

Fagerland et al. [11–13] also compared through simulation experiments the asymptotic behaviour of five methods: *CET*, *MidpT*, *UET*, *MT*, and *MTYcc*. These authors used the same criterion as May and Johnson and recommended using the *MidpT* and *MT* methods.

The studies of May and Johnson [9] and Fagerland et al. [11–13] used the same criterion to assess the type I error rates, and both studies recommended the *MidpT* and *MT* methods. Park [10] recommends the *MT* method.

Our criterion to assess the type I error rate of each method is more flexible, allowing for a method to be higher than $\alpha = 5\%$ without being too liberal. Regarding the asymptotic behaviour of an approximate test, it is to be expected that its type I error rate will fluctuate around the level of the nominal error when the sample size is large, and therefore, it can be higher than that of the nominal error level. With our criterion, it can be slightly higher than the level of the nominal error. Regarding an exact test, its type I error rate must not be higher than the level of the nominal error, as happens with the results obtained for *CET* and *UET* (Tables 2–5).

The simulation experiments carried out allowed us to establish some general rules of application for the methods. The *WT*, *LRT*, *RRT*, and *ODMT* methods can be used for whatever the sample size is, and if the sample size is large, then the *MT* and *MWT* methods can also be applied. Of these six methods, two are conditioned methods (*MT* and *LRT*), and four are unconditioned (*WT*, *MWT*, *RRT*, *ODMT*); therefore, the problem can be addressed without any problem from both perspectives (conditioned and unconditioned), obtaining results that are very similar. Another important conclusion obtained from the simulation experiments is that continuity corrections do not improve the asymptotic behaviors of the studied methods. Therefore, although in the statistical literature there are different methods that incorporate continuity corrections, their application is not justified.

In this manuscript, we have studied the comparison of two paired proportions using hypothesis tests. An alternative method is to carry out this comparison using confidence intervals instead of hypothesis testing. In this context, there are also numerous intervals (exact and approximate) that can be used [4,13–16]. In Fagerland et al. [12,13], the behaviour of some of the most used is compared, but it may currently be somewhat incomplete. Therefore, given that new confidence intervals have been investigated in recent years [14–16], it is of great interest from a practical point of view to determine which intervals have the best asymptotic behaviour.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math12020190/s1.

Author Contributions: J.A.R.-N., T.S.S., and J.F.V.-V. have collaborated equally in the realization of this work. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Spanish Ministry of Science and Innovation, Grant "PID2021-126095NB-100" funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors thank the anonymous referees and the Academic Editors (Elvira Di Nardo and José Luis Vicente Villardón) for their helpful comments that improved the quality of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fay, M.P.; Hunsberger, S.A. Practical valid inferences for the two-sample binomial problem. *Stat. Surv.* **2021**, *15*, 72–110. [CrossRef]
2. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, 1st ed.; Oxford University Press: New York, NY, USA, 2003.
3. Zhou, X.H.; Obuchowski, N.A.; McClish, D.K. *Statistical Methods in Diagnostic Medicine*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2011.
4. Pradhan, V.; Gangopadhyay, A.K.; Menon, S.M.; Basu, C.; Banerjee, T. *Confidence Intervals for Discrete Data in Clinical Research*, 1st ed.; Chapman & Hall/CRC: New York, NY, USA, 2021.
5. McNemar, Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
6. Davis, C.S. Matched pairs with categorical data. In *Encyclopedia of Biostatistics*; Armitage, P., Colton, T., Eds.; Willey: New York, NY, USA, 1998; Volume 3, pp. 2437–2441.
7. Lachenburch, P.A. McNemar test. In *Encyclopedia of Biostatistics*; Armitage, P., Colton, T., Eds.; Willey: New York, NY, USA, 1998; Volume 3, pp. 2486–2487.
8. Pembury Smith, M.Q.R.; Ruxton, G.D. Effective use of the McNemar test. *Behav. Ecol. Sociobiol.* **2020**, *74*, 133. [CrossRef]
9. May, W.L.; Johnson, W.D. The validity and power of tests for equality of two correlated proportions. *Stat. Med.* **1997**, *16*, 1081–1096. [CrossRef]
10. Park, T. Is the exact test better than the asymptotic test for testing marginal homogeneity in 2 × 2 tables? *Biom. J.* **2002**, *44*, 571–583. [CrossRef]
11. Fagerland, M.W.; Lydersen, S.; Laake, P. The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Med. Res. Methodol.* **2013**, *13*, 91. [CrossRef]
12. Fagerland, M.W.; Lydersen, S.; Laake, P. Recommended tests and confidence intervals for paired binomial proportions. *Stat. Med.* **2014**, *33*, 2850–2875. [CrossRef] [PubMed]
13. Fagerland, M.W.; Lydersen, S.; Laake, P. *Statistical Analysis of Contingency Tables*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2017.
14. Tang, M.L.; Ling, M.H.; Ling, L.; Tian, G. Confidence intervals for a difference between proportions based on paired data. *Stat. Med.* **2010**, *29*, 86–96. [CrossRef]
15. Pradhan, V.; Saha, K.K.; Banerjee, T.; Evans, J.C. Weighted profile likelihood-based confidence interval for the difference between two proportions with paired binomial data. *Stat. Med.* **2014**, *33*, 2984–2997. [CrossRef]
16. Fay, M.P.; Lumbard, K. Confidence intervals for difference in proportions for matched pairs compatible with exact McNemar's or sign tests. *Stat. Med.* **2021**, *40*, 1147–1159. [CrossRef]
17. Chang, P.; Liu, R.; Hou, T.; Yan, X.; Shan, G. Continuity corrected score confidence interval for the difference in proportions in paired data. *J. Appl. Stat.* **2024**, *51*, 139–152. [CrossRef] [PubMed]
18. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: New York, NY, USA, 2013; pp. 416–417.
19. Lancaster, H.O. Significance tests in discrete distribution. *J. Am. Stat. Assoc.* **1961**, *56*, 223–234. [CrossRef]
20. Edwards, A.L. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* **1948**, *13*, 185–187. [CrossRef] [PubMed]
21. Yates, F. Contingency table involving small numbers and the χ^2 test. *J. R. Stat. Soc.* **1934**, *1*, 217–235.
22. Martín-Andrés, A.; de Dios Luna del Castillo, J. *40 ± 10 Horas de Bioestadística*; Norma-Capitel: Madrid, Spain, 2013.
23. Bennett, B.M.; Underwood, R.E. On McNemar's test for the 2×2 table and its power function. *Biometrics* **1970**, *26*, 339–343. [CrossRef]
24. Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **1943**, *5*, 426–482. [CrossRef]
25. Lehmann, E.L.; Romano, J.P. *Testing Statistical Hypotheses*, 4th ed.; Springer: Cham, Switzerland, 2022; Chapter 14.
26. Wilks, S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **1938**, *9*, 60–62. [CrossRef]
27. Suissa, S.; Shuster, J.J. The 2 × 2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics* **1991**, *47*, 361–372. [CrossRef]
28. Lu, Y. A revised version of McNemar's test for paired binary data. *Commun. Stat.-Theory Methods* **2010**, *39*, 3525–3539. [CrossRef]
29. Lu, Y. Considering the concordant observations in likelihood ratio test for paired binary data. *Commun. Stat.-Theory Methods* **2011**, *39*, 4214–4232. [CrossRef]
30. Lu, Y.; Wang, M.; Zhang, G. A new revised version of McNemar's test for paired binary data. *Commun. Stat.-Theory Methods* **2017**, *46*, 10010–10024. [CrossRef]
31. Haber, M. The continuity correction and statistical testing. *Int. Stat. Rev.* **1982**, *50*, 135–144. [CrossRef]

32. Irony, T.Z.; Pereira, C.A.; Tiwari, R.C. Analysis of opinion swing: Comparison of two correlated proportions. *Am. Stat.* **2000**, *54*, 57–62.
33. Lui, K.J. Notes on testing equality in dichotomous data with matched pairs. *Biom. J.* **2001**, *43*, 313–321. [CrossRef]
34. Price, R.M.; Bonett, D.G. An improved confidence interval for a linear function of binomial proportions. *Comput. Stat. Data. Anal.* **2004**, *45*, 449–456. [CrossRef]
35. Martín-Andrés, A.; Álvarez-Hernández, M. Two-tailed asymptotic inferences for a proportion. *J. Appl. Stat.* **2014**, *41*, 1516–1529. [CrossRef]
36. Martín-Andrés, A.; Álvarez-Hernández, M. Two-tailed approximate confidence intervals for the ratio of proportions. *Stat. Comput.* **2014**, *24*, 65–75. [CrossRef]
37. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023. Available online: https://www.R-project.org/ (accessed on 4 December 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.