Received: 27 May 2022 | Revised: 25 August 2023 | Accepted: 28 August 2023

DOI: 10.1111/stan.12320

ORIGINAL ARTICLE

WILEY

The Yates, Conover, and Mantel statistics in 2×2 tables revisited (and extended)

Antonio Martín Andrés[1] | **María Álvarez Hernández**[2] | **Francisco Gayá Moreno**[3]

[1]Bioestadística, Facultad de Medicina, Universidad de Granada, Granada, Spain

[2]Centro Universitario de la Defensa, Escuela Naval Militar, Marín (Pontevedra), Spain

[3]Hospital Universitario de La Paz, Unidad de Estadística, Madrid, Spain

Correspondence
Antonio Martín Andrés, Bioestadística. Facultad de Medicina. Universidad de Granada. 18071 Granada. Spain.
Email: amartina@ugr.es

Abstract

Asymptotic inferences about the difference, ratio or odds-ratio of two independent proportions are very common in diverse fields. This article defines for each parameter eight conditional inference methods. These methods depend on: (1) using a chi-squared type statistic or a z type one; (2) using the classic Yates continuity correction or the less well-known Conover one; and (3) whether the p-value of the test is determined by doubling the one-tailed p-value or by the Mantel method (asymmetrical approach). In all cases, the conclusions are: (i) the methods based on the chi-squared statistic should not be used, as they are too liberal; (ii) for those in favor of using the criterion of doubling the p-value, the best method is using the z statistic with Conover continuity correction; and (iii) for those in favor of the asymmetrical approach, the best method is based on the z statistic with Conover continuity correction and the Mantel p-value.

KEYWORDS

confidence intervals, Conover chi-square, difference of proportions, Mantel's method, odds ratio, relative risk, two-tailed test, Yates chi-square, z-statistics

1 | INTRODUCTION

The independence test in 2×2 tables has a long, controversial and complicated history (Martín Andrés, 1998; Martín Andrés, Herranz Tejedor, & Gayá Moreno, 2020), and this applies both to the exact test and the asymptotic test. In both cases, the test to be used depends on whether or not the researcher is in favor of the conditional method (Fisher, 1935) or the unconditional method (Barnard, 1947). From here on, let $\{x_1, y_1, x_2, y_2\}$ be the observed frequencies of a 2×2 table, $x_i+y_i=n_i$ $(i=1, 2)$ the totals of the row, $x_1+x_2=a_1$ and $y_1+y_2=a_2$ the totals of the column, and $n_1+n_2=a_1+a_2=n$ the grand total. From the point of view of the conditional method, which is followed in this article, it is assumed that the values n_i and a_i of the 2×2 table studied are fixed (Model III); therefore, the only random value of the 2×2 table is x_1 (e.g.), which comes from a hypergeometric distribution. In the unconditional method, the only values that are assumed to be fixed are those which were known at the beginning of the experiment, as in the following models. If the data come from two samples from independent binomial distributions of sample sizes n_i (Model II), only the n_i values are fixed and the only random values of the 2×2 table are (x_1, x_2), for example. If the data come from a sample from tetranomial distribution of sample size n (Model I), only the n value is fixed and the only random values of the 2×2 table are (x_1, y_1, x_2), for example; this origin of the data will be not considered in the following.

From an asymptotic point of view, which is used in this article, the three models lead to the classic chi-square statistic $\chi^2_{P(i)}=n(x_1y_2-x_2y_1)^2/\{a_1a_2n_1n_2\}$ (Pearson, 1947); in this expression, subindex (i) refers to the fact that the statistic is used to carry out the independence test. Under the null hypothesis of independence, this statistic follows a chi-square distribution with 1 degree of freedom (df). To distinguish one model from another, it is necessary to provide the previous statistic with a different continuity correction (*cc* from now on) (see, e.g., Martín Andrés, 2008). Subject to Model III, the traditional *cc* is that of the chi-square statistic of Yates (1934) $\chi^2_{Y(i)}=n(\mid x_1y_2-x_2y_1\mid -0.5)^2/\{a_1a_2n_1n_2\}$. Conover (1974) proposed another *cc* which leads to the statistic $\chi^2_{C(i)}=\chi^2_{Y(i)}+n^3/\{4a_1a_2n_1n_2\}$; this statistic is less conservative than the statistic $\chi^2_{Y(i)}$ and performs better when the minimum quantity expected is higher than 2 (Martín Andrés, Herranz Tejedor, & Luna del Castillo, 1992).

An added problem is how to determine the p-value of the two-tailed test. If the objective of the asymptotic test is to provide a p-value which is approximately equal to that of the exact test (Fisher's, in the present case of Model III), then a similar procedure must be followed in the exact test and in the asymptotic test. If the p-value of the two-tailed is defined as double the p-value of the one-tailed test (as in Armitage, 1971), then the p-value of the asymptotic test is directly obtained from the value of the statistics $\chi^2_{Y(i)}$ or $\chi^2_{C(i)}$. However, the p-value of the two-tailed test is often defined as the sum of the 2 one-tailed p-values: that of the observed table, and that of the table with the other tail, which is just as extreme or even more so than the observed table (Baptista & Pike, 1977; Mantel, 1974). In this case, it must be understood that "just as extreme or even more so" refers to "just as improbable or even more so" in the exact test, while in the asymptotic test it refers to the value of the chi-square statistic ("just as large or even more so"). This is why in the asymptotic test it is necessary to proceed as indicated by Mantel: the first one-tailed p-value is obtained from the value of $\chi^2_{Y(i)}$; the second one is obtained from the value of $\chi'^2_{Y(i)}\geq\chi^2_{Y(i)}$, where $\chi'^2_{Y(i)}$ is the minimum value of the Yates chi-square statistic in the tables with the other tail which satisfy this inequality. A similar issue occurs with $\chi^2_{C(i)}$. It should be noted that Mantel (1990) retracted this position, showing himself favorable to the criterion of doubling the p-value.

Up to this point, we have restricted our focus to the case of conditional independence tests, but these same arguments could be applied to more general tests. When the data come from Model II, with success proportions p_1 and p_2, in the field of Medicine, and in many other disciplines, the parameters $d = p_1 - p_2$ (difference in proportions), $R = p_1/p_2$ (relative risk), and $OR = p_1q_2/p_2q_1$ (odds ratio) are usually of interest, with $q_i = 1 - p_i$. Regarding these parameters, it is possible to propose the null hypotheses H_δ: $d = \delta$, H_ρ: $R = \rho$ or H_θ: $OR = \theta$, with $-1 < \delta < +1$, $0 < \rho < \infty$ and $0 < OR < \infty$, versus alternative two-tailed hypotheses K_δ: $d \neq \delta$, K_ρ: $R \neq \rho$ or K_θ: $OR \neq \theta$, respectively. When $\delta = 0$, $\rho = 1$ or $\theta = 1$ we obtain the independence test from the previous paragraphs. The chi-square statistics for the null hypotheses H_δ, H_ρ, or H_θ, based on conditional estimators of the p_i proportions, are well known (Dunnett & Gent, 1977 Farrington & Manning, 1990, and Cornfield, 1956, respectively), including their versions with a Yates type *cc*. Nevertheless, the version with *cc* by Conover has not been defined in them, nor has the performance been assessed for the data from Model II, with or without the precaution of Mantel, and finally their definitions have not been modified in order to obtain coherent values. The only exception to what is stated previously is that in the case of the *OR* parameter, since recently Martín Andrés et al. (2020) proposed the application of the Mantel method to the Yates statistic. Moreover, when the maximum likelihood estimators of the unknown parameters are obtained under the null hypothesis, the z statistic from the score test is equivalent to the chi-square statistic; but this does not happen in another case. This is why it is also possible to apply the aforementioned *cc* and the precaution of Mantel to the z statistic based on conditional estimators. All of this represents the current objective.

Note that the inversion of each test in δ, ρ, or θ allows us to obtain a two-tailed confidence interval (CI) for the parameter involved d, R, or OR (Agresti & Min, 2001), respectively. The CIs obtained based on what was stated in the previous paragraph, are conditional CIs of the asymptotic type. The CIs obtained based on a noncentral hypergeometric distribution, whose only parameter is *OR*, are exact type conditional CIs (in the case of the *OR* parameter) or of a semi-exact type (in the case of the d or R parameters), since this distribution only depends on *OR* and there is no biunivocal relation between *OR* and d or R. As stated previously, here we only focus on the asymptotic case since, as pointed out by Agresti and Coull (1998), "approximate results are sometimes more useful than exact results, because of the inherent conservativeness of exact methods." We also focus on the conditional case as the *cc* of Yates and the *cc* of Conover are of conditional type.

2 | PROPOSED INFERENCE METHODS

2.1 | Statistical tests based on conditional estimators

If p_i parameters are estimated by conditioning and subject to the null hypothesis, then their p_{iC} estimators must satisfy that $n_1p_{1C} + n_2p_{2C} = a_1$, with $p_{1C}\text{-}p_{2C} = \delta$, $p_{1C}/p_{2C} = \rho$ or $p_{1C}q_{2C}/p_{2C}q_{1C} = \theta$ for the H_δ, H_ρ, or H_θ hypotheses, respectively, where $q_{iC} = 1 - p_{iC}$. In the case of the null hypotheses H_δ, H_ρ, and H_θ (d, R, and *OR* parameters, respectively), Dunnett and Gent (1977), Farrington and Manning (1990), and Cornfield (1956) obtained the following p_{iC} estimators (the other p_{i0} estimators are justified below)

$$\text{Parameter } d: \quad p_{1C} = \frac{a_1 + n_2\delta}{n}, p_{2C} = \frac{a_1 - n_1\delta}{n}, p_{i0} = \max\{0; \min(1; p_{iC})\}, \tag{1}$$

$$\text{Parameter } R: p_{1C} = \frac{\rho a_1}{n_2 + n_1\rho}, p_{2C} = \frac{a_1}{n_2 + n_1\rho}, p_{i0} = \min\{1; p_{iC}\}, \tag{2}$$

$$\text{Parameter } OR: \begin{cases} p_{1C} = \begin{cases} \textbf{If } \theta = \mathbf{1}: a_1/n \\ \textbf{If } \theta \neq 1: \frac{(n_1+a_1)\theta+(n_2-a_1)-\sqrt{\{(n_1-a_1)\theta-(n_2-a_1)\}^2+4n_1n_2\theta}}{2n_1(\theta-1)} \end{cases}, \\ p_{2C} = \frac{a_1-n_1p_{1C}}{n_2}, p_{i0} = p_{iC}, \end{cases} \tag{3}$$

respectively. These estimators only match the maximum likelihood estimators under H_δ, H_ρ, or H_θ in the last case (Miettinen & Nurminen, 1985). In general, in all cases we should find that $0 \leq p_{iC} \leq 1$, but this is only guaranteed in the case of *OR*. That is why Farrington and Manning (1990) suggested that the p_{iC} estimators (1) and (2) take the value of $p_{i0} = 0$ (or 1) when $p_{ic} < 0$ (or > 1); hence the p_{i0} estimators of Expressions (1)–(3). Note that the p_{i0} estimators of Expressions (1) and (2) may not satisfy the null hypotheses or the equality $n_1p_{10} + n_2p_{20} = a_1$. As the latter would prevent us from using the equality $(x_1 - n_1p_{10}) = -(x_2 - n_2p_{20})$, which is crucial for the simplification of the chi-square statistic (see next paragraph), the proposal made here is to define p_{i0} only in the part of the statistic that does not contain the terms $(x_i - n_ip_{iC})^2$.

The classic Pearson chi-square statistic takes the form $\chi^2_P = \sum (O_i - E_i)^2/E_i$, where O_i are the observed quantities and E_i are the quantities expected under the null hypothesis. If the 2×2 table comes from Model II, the four values of O_i (or E_i) are x_1, y_1, x_2, and y_2 (or n_1p_1, n_1q_1, n_2p_2, and n_2q_2), respectively, where the values of the parameters p_i and q_i still need to be estimated subject to the null hypothesis. Grouping together terms, we obtain $\chi^2_P = \sum (x_i - n_ip_i)^2/n_ip_iq_i$. Using the p_{iC} estimators in the numerator, the p_{i0} estimators in the denominator and applying equality $(x_1 - n_1p_{1C}) = -(x_2 - n_2p_{2C})$, the following statistic X is obtained

$$\chi^2_X = \frac{(x_1 - n_1p_{1C})^2}{V_X}, \text{where } V_X = \left(\sum \frac{1}{n_ip_{i0}q_{i0}}\right)^{-1}, \tag{4}$$

where p_{i0} are the appropriate values of Expressions (1)–(3) and $q_{i0} = 1 - p_{i0}$. Statistic X leads to the three statistics *d-X*, *R-X* and *OR-X* (χ^2_{d-X}, χ^2_{R-X} and χ^2_{OR-X}) depending on whether the estimators of Expressions (1), (2), or (3) are used, respectively. Subject to the null hypothesis, χ^2_X asymptotically follows a chi-square distribution with $df = 1$. Note that the actual definition of V_X, based on estimators p_{i0}, prevents the X statistic from giving a negative value. The *d-X* statistic is a definition of compromise between the statistics of Dunnett and Gent (1977) and Farrington and Manning (1990); the *R-X* statistic is a modification of the one proposed by Martín Andrés and Herranz Tejedor (2010); finally, the *OR-X* is by Cornfield (1956).

Other classic statistics are those obtained by typifying the appropriate random variable. The square of the typified value leads to the statistics $\chi^2_T = \left(\overline{p}_1 - \overline{p}_2 - \delta\right)^2 / \sum (p_iq_i/n_i)$ for the *d* case (Dunnett & Gent, 1977), $\chi^2_T = \left(\overline{p}_1 - \rho\overline{p}_2\right)^2 / \left[(p_1q_1/n_1) + \rho^2(p_2q_2/n_2)\right]$ for the *R* case (Katz, Baptista, Azen, & Pike, 1978), and $\chi^2_T = (x_1 - n_1p_{1C})^2 / \left\{\sum (n_ip_iq_i)^{-1}\right\}^{-1}$ for the *OR* case (Miettinen & Nurminen, 1985), with $\overline{p}_i = x_i/n_i$. In all three cases, the parameters p_i and q_i still need to be estimated under the null hypothesis. If operations are carried out and p_i is substituted with the appropriate values of p_{i0} in Expressions (1)–(3), the following type Z statistics are obtained.

$$\chi^2_Z = \frac{(x_1 - n_1p_{1C})^2}{V_Z}, \tag{5}$$

where

$$V_Z = \left(\frac{n_1 n_2}{n}\right)^2 \left(\sum \frac{p_{i0} q_{i0}}{n_i}\right), V_Z = \left(\frac{n_1 n_2}{n_2 + \rho n_1}\right)^2 \left(\frac{p_{10} q_{10}}{n_1} + \rho^2 \frac{p_{20} q_{20}}{n_2}\right), \text{and } V_Z = V_X, \quad (6)$$

for the cases *d*, *R*, and *OR*, respectively. This results in three statistics *d-Z*, *R-Z*, and *OR-Z* (χ^2_{d-Z}, χ^2_{R-Z}, and $\chi^2_{OR-Z} = \chi^2_{OR-X}$) depending on the value of (6) that is assigned to V_Z in Expression (5) (Farrington & Manning, 1990; Miettinen & Nurminen, 1985). Note that Expressions (4) and (5) have the same form, but the V value in each of them (V_X or V_Z) is different. This structural equality will be of use in the next two sections.

If in the expressions of χ^2_P and χ^2_T p_i are replaced by their maximum likelihood estimators subject to the null hypothesis, it is known that $\chi^2_P = \chi^2_T$ (Gart & Nam, 1988; Miettinen & Nurminen, 1985; Nam, 1995), where χ^2_T is the classic score statistic; but that equality does not take place when using the p_{iC} or p_{i0} current conditional estimators. The exception is provided by the case of the *OR* parameter, since p_{iC} is also the maximum likelihood estimator under the null hypothesis (Miettinen & Nurminen, 1985); that is why in the previous paragraph it was indicated that $\chi^2_{OR-Z} = \chi^2_{OR-X}$ and, therefore, the *OR-X* and *OR-Z* statistics are the same.

Note finally that in the case of the independence test, in which $\delta = 0$ or $\rho = \theta = 1$, then $p_{1C} = a_1/n$, $p_{2C} = a_2/n$ and the current values of χ^2_X and χ^2_Z are the same, respectively, as the $\chi^2_{P(i)}$ statistic indicated in the Section 1.

2.2 | Test statistics with continuity correction

When a discrete random variable is approximated through a continuous random variable, as in our case, it is well known that it is advisable to perform a *cc* (Cox, 1970; Hamdan, 1974; Schouten, 1976). In general, a *cc* means replacing the observed value with the average between the "truly observed value" and its "less extreme immediate value" or, equivalently, of adding to or subtracting from the observed value half the gap between both values. In the case of the statistics of Expression (4) or (5), the observed value is $\chi^2_1 = (x_1 - n_1 p_{1C})^2/V$, where V is V_X or V_Z depending on the case. Assuming, without loss of generality, that $x_1 > n_1 p_{1C}$ and that we want to estimate the probability of the right tail, the immediately lower value χ^2_1 is $\chi^2_2 = (x_1 - n_1 p_{1C} - 1)^2/V$ or, in general, $\chi^2_2 = (|x_1 - n_1 p_{1C}| - 1)^2/V$. The *cc* statistic depends on what is considered to be the base variable of the problem. If the base variable is χ_X or χ_Z, as Haber (1980) believes in the context of the test for $H_{\delta=0}$, then we obtain $[(\chi_1 + \chi_2)/2]^2 = (|x_1 - n_1 p_{1C}| - 0.5)^2/V$, which is the chi-square statistic with the classic Yates *cc*. For this reason, the statistics χ^2_X and χ^2_Z with the Yates *cc* will be, respectively

$$\chi^2_{XY} = \frac{(|x_1 - n_1 p_{1C}| - 0.5)^2}{V_X}, \text{with } V_X \text{ as in expression (4)}, \quad (7)$$

$$\chi^2_{ZY} = \frac{(|x_1 - n_1 p_{1C}| - 0.5)^2}{V_Z}, \text{with } V_Z \text{ as in expression (6)}, \quad (8)$$

If the base variable is χ^2_X or χ^2_Z, as Conover (1974) believes in the same context as Haber, then we obtain $\left(\chi^2_1 + \chi^2_2\right)/2 = (|x_1 - n_1 p_{1C}| - 0.5)^2/V + (4V)^{-1}$, which is the chi-square with

Conover's *cc*; that is why statistics χ^2_X and χ^2_Z the *cc* of Conover will be, respectively

$$\chi^2_{XC} = \frac{(|x_1 - n_1 p_{1C}| - 0.5)^2 + 0.25}{V_X}, \quad \text{with } V_X \text{ as in expression (4),} \tag{9}$$

$$\chi^2_{ZC} = \frac{(|x_1 - n_1 p_{1C}| - 0.5)^2 + 0.25}{V_Z}, \quad \text{with } V_Z \text{ as in expression (6).} \tag{10}$$

In every case it must be understood that the statistic equals 0 when $|x_i\text{-}n_i p_{iC}| \leq 0.5$.

Each of the two X statistics with *cc* (*XY* and *XC*) leads to three *cc* statistics, one for each *d*, *R*, and *OR* parameter (χ^2_{d-XY}, χ^2_{d-XC}, χ^2_{R-XY}, χ^2_{R-XC}, χ^2_{OR-XY}, and χ^2_{OR-XC}). Each of the two Z statistics with *cc* (*ZY* and *ZC*) leads to three *cc* statistics, one for each *d*, *R*, and *OR* parameter (χ^2_{d-ZY}, χ^2_{d-ZC}, χ^2_{R-ZY}, χ^2_{R-ZC}, $\chi^2_{OR-ZY} = \chi^2_{OR-XY}$, and $\chi^2_{OR-ZC} = \chi^2_{OR-XC}$). That is why there are 10 statistics to be evaluated. For instance, the χ^2_{d-XY} statistic is obtained by replacing the value of V_X from Expression (4) and the estimators from Expression (1) in Expression (7). Statistics χ^2_{d-ZY} and χ^2_{R-ZY} were proposed by Farrington and Manning (1990), statistic $\chi^2_{OR-ZY} = \chi^2_{OR-XY}$ comes from Cornfield (1956), and statistics χ^2_{d-XY} and χ^2_{R-XY} are a modification of those proposed by Dunnett and Gent (1977) and Martín Andrés and Herranz Tejedor (2010), respectively; the other statistics have never been explicitly proposed (as far as we know).

2.3 | Procedures to obtain the *p*-value of the two-tailed test and inference methods that are obtained

The conditional method means that only random variable in the 2×2 table is x_1 (e.g.), since the other three values in the table are deduced from their marginal a_i and n_i: $x_2 = a_1\text{-}x_1$, $y_1 = n_1\text{-}x_1$, and $y_2 = n_2\text{-}x_2$. Thus,

$$r(a_1) = \max\ \{0;\ a_1 - n_2\} \leq x_1 \leq \min\ \{a_1;\ n_1\} = s(a_1), \tag{11}$$

where functions $r(a_1)$ and $s(a_1)$ will be used later and only depend on a_1, since this article assumes that the data comes from Model II, that is to say that the values of n_i are fixed.

The *p*-value of the two-tailed test is traditionally obtained by doubling the *p*-value of the one-tailed test. Therefore the *p*-value of the *XY* (χ^2_{XY}) statistic will be $P_{XY} = 2 \times \Pr\{z \geq \chi_{XY}\}$, where z refers to a typical normal random variable. Something similar occurs with the *XC*, *ZY*, and *ZC* statistics. However, this approach makes the test very conservative (see e.g., Martín Andrés et al., 2020 in the χ^2_{OR-XY} statistic). The inference methods that are obtained by "doubling the *p*-value" will be denominated in a general way by adding letter *D*: methods *XYD*, *XCD*, *ZYD* and *ZCD*. When these methods in particular are applied to parameter *d*, methods *d-XYD*, *d-XCD*, *d-ZYD*, and *d-ZCD* are obtained, respectively. The same happens with parameters *R* (*R-XYD*, *R-XCD*, *R-ZYD*, and *R-ZCD*) and *OR* (*OR-XYD* = *OR-ZYD* and *OR-XCD* = *OR-ZCD*).

As stated previously in the third paragraph of the Section 1, Mantel (1974) explained how to determine the *p*-value of the two-tailed test based on the statistic $\chi^2_{Y(i)}$ and from the perspective of conditional inference. Martín Andrés et al. (2020) extended the idea to the statistic χ^2_{OR-XY}. Here we also generalize it to the case of the other statistics and the other parameters. As the conditional estimators of p_{iC} are independent from x_1, then the value of χ^2_{XY} (e.g.) only depends on x_1 and will be referred to in this section as $\chi^2_{XY}(x_1)$. Mantel's definition that "the *p*-value of the two-tailed

test is the sum of the *p*-value*s* of the 2 one-tailed tests", applied to the case of the statistic $\chi^2_{XY}(x_1)$, leads to the *XYM* method that is described and justified below in four steps:

1. Determining the value $\chi^2_{XY}(x_1)$ of Expression (7) in the observed table and calculating its *p*-value of one tail $P_1(x_1) = \Pr\{z \geq \chi_{XY}(x_1)\}$.
2. Determining the first table of the other tail (x'_1) whose value $\chi^2_{XY}(x'_1)$ is greater than or equal to the observed value $\chi^2_{XY}(x_1)$. Let us assume that $x_1 < n_1 p_{1C}$. As $\chi^2_{XY}(x_1)$ grows as x_1 moves away from the mean $n_1 p_{1C}$, the values x'_1 that satisfy $\chi^2_{XY}(x'_1) \geq \chi^2_{XY}(x_1)$ are those that satisfy $x'_1 - n_1 p_{1C} \geq n_1 p_{1C} - x_1$ or, equivalently, those that satisfy $x'_1 \geq 2n_1 p_{1C} - x_1$. That is why, in general, the value sought will be $x'_1 = \left[2n_1 p_{1C} - x_1\right]$ where $[A]$ refers to the rounding-up of A in the sense of moving away from $n_1 p_{1C}$, that is:

$$x'_1 = \begin{cases} \left[2n_1 p_{1C} - x_1\right]^+ \text{ if } x_1 \leq n_1 p_{1C}, \\ \left[2n_1 p_{1C} - x_1\right]^- \text{ if } x_1 > n_1 p_{1C}, \end{cases} \tag{12}$$

With $[A]^+$ the rounding above A and $[A]^-$ the rounding below A. The rest of the values in table 2×2 will be given by $x'_2 = a_1 - x'_1$, $y'_1 = n_1 - x'_1$, and $y'_2 = n_2 - x'_2$, although they will not be necessary for what follows. Let $\chi^2_{XY}(x'_1)$ be the value of the Expression (7) in this new table.

3. If it is verified that $r(a_1) \leq x'_1 \leq s(a_1)$, then x'_1 is a valid value and the *p*-value of the other tail will be $P_1(x'_1) = Pr\{z \geq \chi_{XY}(x'_1)\}$. Otherwise x'_1 is not a valid value and $P_1(x'_1) = 0$.
4. Finally the *p*-value of the two-tailed test through the Mantel method is the sum of the previous two $P_{XYM} = P_1(x_1) + P_1(x'_1)$, hence the letter *M* added after the letters *XY* of the statistic used.

It will proceed in a similar way with the other three statistics (*XC*, *ZY*, and *ZC*). When the current Mantel method is applied to the four statistics for the *d* parameter, the *d-XYM*, *d-XCM*, *d-ZYM*, and *d-ZCM* methods are obtained, respectively. Correspondingly with the parameters *R* (*R-XYM*, *R-XCM*, *R-ZYM*, and *R-ZCM*) and *OR* (*OR-XYM* = *OR-ZYM* and *OR-XCM* = *OR-ZCM*).

From all the aforementioned reasons, it follows that for each of the parameters *d* and *R* (or *OR*) eight (or four) ways of obtaining the *p*-value of the two-tailed test have been defined, and the current objective is the comparative evaluation when the data come from Model II, which is the most common one. For example, in the case of the *OR* parameter, only the four methods *OR-ZYD*, *OR-ZCD*, *OR-ZYM*, and *OR-ZCM* are evaluated. All these inference methods are included in Table 1 for easy reference.

TABLE 1 Parameter of inference, statistic, continuity correction and the mode of determining the *p*-value used in this paper (first, second, third and fourth letter of each inference method, respectively).

Parameter (1st letter and hyphen)		**Statistic (2nd letter)**		**cc (3rd letter)**		***p*-value (4th letter)**	
d	Difference	*X*	Chi-square	*Y*	Yates	*D*	Doubling the one-tailed test
R	Ratio	*Z*	Typified value	*C*	Conover	*M*	Mantel
OR	Odds-ratio						

Notes: For each parameter, $2 \times 2 \times 2 = 8$ inference methods are defined, but in the case of *OR* there are only four methods, since the statistics *X* and *Z* are the same. For example, *OR-XYD* = *OR-ZYD*.

3 | COMPARATIVE EVALUATION OF THE PROPOSED INFERENCE METHODS

3.1 | Procedure for obtaining and analyzing the results

In order to comparatively evaluate the eight (or four) inference methods proposed for each parameter d and R (or OR), it is necessary to obtain certain parameters that synthesize the quality of each method. The parameters chosen are the four indicated below—{Θ, $S5$, $S6$, $S7$}—, all defined based on the critical region (CR) that determines each inference method. As in the evaluation it is assumed that the data comes from Model II, each CR—and therefore each quality parameter—will be obtained in a combination of values $\alpha = 5\%$ (the objective Type I error of the test), n_1, n_2 and ϕ (the value of the parameter under the null hypothesis), where $\phi = \delta/\rho/\theta$ for the parameters $d/R/OR$, respectively. Each inference method, applied to the point (x_1, x_2) of the sample space, provides a p-value $P(x_1, x_2)$ that allows constructing the critical region $CR = \{(x_1, x_2) | P(x_1, x_2) \leq \alpha\}$. To form the CR, the most advisable action is to obtain the $CR(a_1)$ in each value of a_1, with $0 \leq a_1 \leq n$, and then join all the $CR(a_1)$ thus obtained: $CR = \cup_{a_1} CR(a_1)$. Note that each $CR(a_1)$ refers to the set of values of licit x_1—that is, those that satisfy the Expression (11)—in which $P(x_1, x_2 = a_1\text{-}x_1) \leq \alpha$.

In type D inference methods (such as the d-XYD method, for example), it is easier to obtain the CR directly from the value χ^2_{exp} of the implied statistic ($\chi^2_{d\text{-}XYD}$ in the example), since $CR = \{(x_1, x_2) | \chi^2_{\text{exp}} \geq \chi^2_\alpha\}$, with χ^2_α the $(1-\alpha)$-percentile of the chi-square distribution with $df = 1$. Solving for x_1 from the inequality $\chi^2_{\text{exp}} \geq \chi^2_\alpha$, with χ^2_{exp} given by any of the Expressions (7)–(10), it is obtained that the CR (a_1) is given by the values of x_1 that satisfy:

$$\underline{XYD/ZYD \text{ methods}} : r(a_1) \leq x_1 \leq n_1 p_{1C} - Y_\alpha \text{ or } n_1 p_{1C} + Y_\alpha \leq x_1 \leq s(a_1) \quad \text{with } Y_\alpha = .5 + \chi_\alpha V^{0.5}, \tag{13}$$

$$\underline{XCD/ZCD \text{ methods}} : r(a_1) \leq x_1 \leq n_1 p_{1C} - C_\alpha \text{ or } n_1 p_{1C} + C_\alpha \leq x_1 \leq s(a_1) \quad \text{with } C_\alpha = .5 + \left(\chi^2_\alpha V - .25\right)^{0.5}, \tag{14}$$

where the value of V (V_X or V_Z) is constant for the current value of a_1.

In type M inference methods (such as the d-XYM, for instance), the solution is not so simple. Now the only thing that can be said is when certain points do or do not belong to the CR. By definition, the p-value of type M inference methods is less than or equal to the p-value of type D inference methods. Therefore, all values of x_1 that satisfy Expressions (13) for XYM/ZYM methods, or Expressions (14) for XCM/ZCM methods, are in fact part of the $CR(a_1)$ of the Type M methods. Additionally, the p-value of the two-tailed test of type D inference methods is less than or equal to twice the p-value of one of the one-tail test. Therefore, if a method of type D is not significant to error 2α at a certain value of x_1, then the type M method will not be significant to error α. That is why all values of x_1 that do not satisfy Expressions (13)/(14) for 2α, will NOT be a part of the $CR(a_1)$, that is the values of x_1 that satisfy:

$$\underline{XYM/ZYM \text{ methods}} : \max\{r(a_1);\ n_1 p_{1C} - Y_{2\alpha}\} < x_1 < \min\{s(a_1);\ n_1 p_{1C} + Y_{2\alpha}\},$$

$$\underline{XCM/ZCM \text{ methods}} : \max\{r(a_1);\ n_1 p_{1C} - C_{2\alpha}\} < x_1 < \min\{s(a_1);\ n_1 p_{1C} + C_{2\alpha}\},$$

where $Y_{2\alpha}$ and $C_{2\alpha}$ are obtained as in expressions (13) and (14), respectively, but changing α for 2α. In the rest of the values of x_1, one has to determine its two-tailed p-value in order to make the decision: x_1 will be from the $CR(a_1)$ if and only if $P(x_1, x_2) \leq \alpha$. These values of x_1, which we have experimentally verified are usually few, are the ones that satisfy:

$$\underline{XYM/ZYM \text{ methods}}: \max\{r(a_1);\ n_1p_{1C} - Y_\alpha\} < x_1 \leq \min\{s(a_1);\ n_1p_{1C} - Y_{2\alpha}\} \text{ or}$$
$$\max\{r(a_1);\ n_1p_{1C} + Y_{2\alpha}\} \leq x_1 < \min\{s(a_1);\ n_1p_{1C} + Y_\alpha\}$$

$$\underline{XCM/ZCM \text{ methods}}: \max\{r(a_1);\ n_1p_{1C} - C_\alpha\} < x_1 \leq \min\{s(a_1);\ n_1p_{1C} - C_{2\alpha}\} \text{ or}$$
$$\max\{r(a_1);\ n_1p_{1C} + C_{2\alpha}\} \leq x_1 < \min\{s(a_1);\ n_1p_{1C} + C_\alpha\}.$$

Once the CR has been determined, it is then possible to determine the four parameters of interest. The first parameter of interest is $\Theta = 100 \times (\text{number of points in the set } CR)/[(n_1+1)(n_2+1)]$, where $(n_1+1)(n_2+1)$ is the total number of points in the sample space. The value of Θ is a good indication of the power of the test (Chen, Hung, & Chen, 2007; Martín Andrés & Silva Mato, 1994; Upton, 1982). The Θ parameter allows us to obtain a global comparison of two tests performed at the same error α, which does not happen when traditional power is used, since it depends on the alternative hypothesis that is considered.

To determine the other three parameters of interest ($S5$, $S6$ and $S7$) it is necessary to calculate the actual error of the test for each value of the unknown parameter p_1.The actual error is given by $\alpha(p_1) = \sum_{CR} \Pr(x_1, x_2|p_1, p_2)$, with $p_2 = p_1 - \delta$, $p_2 = p_1/\rho$ or $p_2 = p_1/(p_1 + \theta q_1)$ depending on the case, and $\Pr(x_1, x_2|p_1, p_2) = C(n_1, x_1) \times C(n_2, x_2) \times p_1^{x_1}(1-p_1)^{y_1} p_2^{x_2}(1-p_2)^{y_2}$. The possible values of p_1 depend on the parameter of interest: $\max\{0; \delta\} \leq p_1 \leq \min\{1; 1+\delta\}$ in case d, $0 \leq p_1 \leq \min\{1; \rho\}$ in the case R, and $0 \leq p_1 \leq 1$ in the case OR. When $\alpha(p_1) < 0.05$ (or > 0.05) the test will be conservative (or liberal). In general, it is advisable that $\alpha(p_1) \leq 0.05$, in order not to obtain false significic. If you take $2^{20}-1$ equispaced values of p_1 and define $S5$, $S6$ or $S7$ as the proportion of p_1 values in which $\alpha(p_1) > 5\%$, 6% or 7% respectively, the parameters $S5$, $S6$, and $S7$ will allow us to determine how liberal the test is. This is because the $S5$, $S6$, or $S7$ parameters refer to the proportion of values of the nuisance parameter p_1 in which the actual error α is greater than 5%, 6%, or 7%, respectively. It is true that in an asymptotic CI at a nominal confidence of 95%, it does not matter very much whether or not the actual coverage is 94.9% or 94.8%, although 93% is excessively low for any researcher; but if $S6 > 0$, then there are p_1 values in which the coverage is 93.2% (for example), which is undesirable. The value $S7$, which should equal 0 (Agresti & Caffo, 2000), is set as a control to make sure that excessive liberality does not occur. That is why we use all three values $S5$, $S6$, and $S7$. Of course, some researchers may disagree with these criteria, as they are as debatable as any others, but we understand that these reflect the most common opinion.

Once we know the values of $\{\Theta, S5, S6, S7\}$ for all of the inference methods on a given parameter, the optimal method will be the one that provides maximum values of Θ and minimum values of $S5$, $S6$ and $S7$. As this situation ideal is not usually verified, priority is therefore given to the values of $S5$, $S6$ and $S7$ being small, especially $S6$ and $S7$, even when Θ is a little lower. In fact, it is to be wished that at least $S7 = 0$.

The previous assessment allows us to determine the best method to carry out the test. However, as the CI is obtained through the inversion of the test, the method selected will also be the best method to obtain the CI for the parameter involved. In order to assess a test, we normally use the parameters *real error* (probability of the critical region when the null hypothesis is true) and *power* (probability of the critical region when a determined alternative hypothesis is true). In

order to assess a CI, we normally use the parameters *real coverage* (probability of the CI containing the true value of the parameter) and *average length* (average value of the width of the CI). Both assessments are equivalent since, on the one hand, the *real coverage* and the *real error* add up to 1 (see e.g., Martín Andrés, Álvarez Hernández, & Herranz Tejedor, 2012). On the other hand, the greater Θ is, the *CR* has more points and, in general, the greater the *power* of the test and the lower the *average length* of the *CI* that is obtained through its inversion.

3.2 | Selection of the optimal method

In order to assess the eight inference methods for parameter d, we consider the values $\delta = 0$, 0.1, 0.2, 0.3, 0.5, 0.7, 0.8 and 0.9, and $n_i = 20$, 40, 60 and 100, with $n_1 \leq n_2$. The values $\delta < 0$ and $n_1 > n_2$ are excluded because of the equivalence between the different null hypotheses which are obtained by permuting the order of the samples and/or by permuting the successes with the failures. Thus, since the original null hypotheses $p_1 - p_2 = \delta$ is equivalent to these other three null hypotheses $p_2 - p_1 = -\delta$, $q_2 - q_1 = \delta$ and $q_1 - q_2 = -\delta$, then the results for a setting of those which are not considered will be the same as for those of another setting which is in fact used.

In order to assess the eight inference methods for parameter R, we consider the values $\rho = 0.05$, 0.10, 0.20, 0.50, 0.80, 1, 1.25, 2, 5, 10 and 20, and $n_i = 20$, 40, 60 and 100, with $n_1 \leq n_2$. The values $n_1 > n_2$ are excluded since, as in case d, their results are already considered in some of the combinations which are in fact used. The reason now is the equivalence of the two null hypotheses which are obtained by permuting the samples ($p_1/p_2 = \rho$ and $p_2/p_1 = 1/\rho$).

Finally, in order to assess the four inference methods for the *OR* parameter, we consider the values $\theta = 1$, 1.25, 2, 5, 10 and 20, and $n_i = 20$, 40, 60, and 100, with $n_1 \leq n_2$. The values $\theta < 1$ and $n_1 > n_2$ are excluded for similar reasons to those of case d: the original null hypothesis $p_1q_2/p_2q_1 = \theta$ is equivalent to these three null hypotheses $p_2q_1/p_1q_2 = 1/\theta$, $q_2p_1/q_1p_2 = \theta$ and

TABLE 2 Average values of the percentage of values of the proportion p_1 in which the test provides an error $\alpha(p_1)$ higher than 5% (*S5*), 6% (*S6*) or 7% (*S7*), and of the percentage of points Θ in the critical region, for the eight inference methods indicated when a two-tailed test to an error $\alpha = 5\%$ is performed for H_d: $d = \delta$, H_R: $R = \rho$ or H_{OR}: $OR = \theta$.

Parameter	***d***				***R***				***OR***			
Inference method	***S5***	***S6***	***S7***	***Θ***	***S5***	***S6***	***S7***	***Θ***	***S5***	***S6***	***S7***	***Θ***
XYD	41.8	37.2	34.0	80.4	17.3	14.8	13.1	79.0	—	—	—	68.3
XCD	43.1	38.2	34.8	80.6	18.1	15.4	13.6	79.2	—	—	—	68.7
XYM	46.8	41.4	37.3	80.9	22.0	18.1	15.4	79.9	—	—	—	69.8
XCM	49.5	44.1	39.9	81.2	25.2	19.8	16.3	80.2	0.5	0.1	—	70.2
ZYD	—	—	—	78.6	—	—	—	77.4	—	—	—	68.3
ZCD	—	—	—	78.8	0.1	—	—	77.7	—	—	—	68.7
ZYM	1.2	—	—	79.5	2.8	—	—	78.6	—	—	—	69.8
ZCM	2.8	0.1	—	79.7	5.3	0.8	0.1	78.9	0.5	0.1	—	70.2

Notes: These values were obtained in all of the combinations of values of (n_1, n_2, $\delta/\rho/\theta$). In the case of the *OR* parameter, the four statistics *X*** are the same as their respective statistics *Z***; and this is why their results are the same. The boxes with "—" indicate that the percentage of the same thing is "0.0."

TABLE 3 Percentage of the values of the proportion p_1 in which the test provides an error $\alpha(p_1)$ higher than 5% (*S5*), 6% (*S6*) or 7% (*S7*), and percentage of points Θ in the critical region, for the four inference methods indicated (*ZYD*, *ZCD*, *ZYM* and *ZCM*) when a two-tailed test to an error $\alpha = 5\%$ is performed for H_d: $d = \delta$.

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*			
n_1	n_2	δ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
20	20	0	—	—	—	46.7	—	—	—	48.5	—	—	—	46.7	—	—	—	48.5
		0.1	—	—	—	49.4	—	—	—	49.9	—	—	—	50.3	—	—	—	50.8
		0.2	—	—	—	53.1	—	—	—	55.3	—	—	—	54.4	—	—	—	56.2
		0.3	—	—	—	59.4	—	—	—	59.9	—	—	—	60.8	—	—	—	61.2
		0.5	—	—	—	69.8	—	—	—	70.3	—	—	—	72.1	—	—	—	72.6
		0.7	—	—	—	82.5	—	—	—	83.0	—	—	—	84.8	—	—	—	84.8
		0.8	—	—	—	89.3	—	—	—	89.8	—	—	—	90.3	—	—	—	90.7
		0.9	—	—	—	94.3	—	—	—	94.8	—	—	—	95.2	—	—	—	95.2
	40	0	—	—	—	53.9	—	—	—	54.6	—	—	—	55.5	—	—	—	56.5
		0.1	—	—	—	56.1	—	—	—	56.8	—	—	—	57.8	—	—	—	58.7
		0.2	—	—	—	59.9	—	—	—	60.5	—	—	—	62.1	—	—	—	63.0
		0.3	—	—	—	64.5	—	—	—	65.0	—	—	—	67.5	—	—	—	67.9
		0.5	—	—	—	75.5	—	—	—	75.6	—	—	—	77.1	—	—	—	77.1
		0.7	—	—	—	86.2	—	—	—	86.3	—	—	—	87.3	—	—	—	87.3
		0.8	—	—	—	91.3	—	—	—	91.6	—	—	—	92.3	—	—	—	92.6
		0.9	—	—	—	95.8	—	—	—	96.1	—	—	—	96.5	—	—	—	96.6
20	60	0	—	—	—	57.3	—	—	—	57.5	—	—	—	58.7	—	—	—	59.0
		0.1	—	—	—	58.6	—	—	—	59.6	—	—	—	60.2	6.7	4.0	—	60.7
		0.2	—	—	—	62.5	—	—	—	62.8	—	—	—	63.6	—	—	—	63.9
		0.3	—	—	—	66.7	—	—	—	66.9	0.4	—	—	68.1	0.4	—	—	68.2
		0.5	—	—	—	76.9	—	—	—	77.1	—	—	—	78.5	—	—	—	78.8
		0.7	—	—	—	87.5	—	—	—	87.7	—	—	—	88.6	—	—	—	88.9
		0.8	—	—	—	92.3	—	—	—	92.4	—	—	—	93.2	—	—	—	93.3
		0.9	—	—	—	96.4	—	—	—	96.6	—	—	—	97.0	—	—	—	97.0
	100	0	—	—	—	59.7	—	—	—	60.1	—	—	—	61.5	—	—	—	62.2
		0.1	—	—	—	61.2	—	—	—	61.6	—	—	—	62.9	5.5	2.6	—	63.4
		0.2	—	—	—	64.6	—	—	—	64.9	—	—	—	66.6	7.3	—	—	66.9
		0.3	—	—	—	68.8	—	—	—	69.1	0.3	—	—	70.6	6.5	—	—	70.8
		0.5	—	—	—	78.7	—	—	—	78.8	—	—	—	79.8	—	—	—	79.9
		0.7	—	—	—	88.6	—	—	—	88.7	—	—	—	89.5	—	—	—	89.7
		0.8	—	—	—	93.2	—	—	—	93.3	—	—	—	93.8	—	—	—	93.9
		0.9	—	—	—	97.1	—	—	—	97.2	—	—	—	97.4	—	—	—	97.5

TABLE 3 (Continued)

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*			
n_1	n_2	δ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
40	40	0	—	—	—	62.6	—	—	—	63.1	—	—	—	62.6	—	—	—	63.1
		0.1	—	—	—	64.5	—	—	—	64.5	—	—	—	64.8	4.3	—	—	65.1
		0.2	—	—	—	67.3	—	—	—	67.3	—	—	—	67.8	—	—	—	68.1
		0.3	—	—	—	70.8	—	—	—	71.0	—	—	—	71.5	—	—	—	71.7
		0.5	—	—	—	79.7	—	—	—	79.7	—	—	—	80.6	—	—	—	80.6
		0.7	—	—	—	89.0	—	—	—	89.2	—	—	—	89.7	—	—	—	90.0
		0.8	—	—	—	93.2	—	—	—	93.3	—	—	—	93.8	—	—	—	93.8
40	40	0.9	—	—	—	97.1	—	—	—	97.1	—	—	—	97.3	—	—	—	97.4
	60	0	—	—	—	66.1	—	—	—	66.4	—	—	—	67.2	—	—	—	67.5
		0.1	—	—	—	67.5	—	—	—	67.7	—	—	—	68.7	3.4	—	—	69.1
		0.2	—	—	—	70.3	—	—	—	70.6	—	—	—	71.5	1.4	—	—	71.5
		0.3	—	—	—	73.8	—	—	—	74.0	5.4	—	—	75.0	14.0	—	—	75.4
		0.5	—	—	—	81.9	—	—	—	82.0	2.5	—	—	83.0	12.4	—	—	83.1
		0.7	—	—	—	90.3	—	—	—	90.4	—	—	—	91.2	—	—	—	91.3
		0.8	—	—	—	94.1	—	—	—	94.2	—	—	—	94.8	—	—	—	94.8
		0.9	—	—	—	97.4	—	—	—	97.5	—	—	—	97.8	—	—	—	97.8
	100	0	—	—	—	68.8	—	—	—	69.2	—	—	—	70.1	—	—	—	70.4
		0.1	—	—	—	70.4	—	—	—	70.5	—	—	—	71.6	5.3	—	—	71.8
		0.2	—	—	—	73.1	—	—	—	73.2	—	—	—	74.1	9.1	—	—	74.3
		0.3	—	—	—	76.2	—	—	—	76.3	4.3	—	—	77.3	15.7	—	—	77.6
		0.5	—	—	—	83.7	—	—	—	83.8	7.3	—	—	84.5	10.9	—	—	84.6
		0.7	—	—	—	91.4	—	—	—	91.4	—	—	—	92.0	—	—	—	92.0
		0.8	—	—	—	94.9	—	—	—	94.9	—	—	—	95.3	—	—	—	95.3
		0.9	—	—	—	97.8	—	—	—	97.9	—	—	—	98.1	—	—	—	98.1
60	60	0	—	—	—	69.7	—	—	—	69.7	—	—	—	69.7	—	—	—	69.7
		0.1	—	—	—	71.1	—	—	—	71.2	—	—	—	71.3	—	—	—	71.4
		0.2	—	—	—	73.5	—	—	—	73.6	—	—	—	73.9	10.8	—	—	74.1
		0.3	—	—	—	76.7	—	—	—	76.8	8.0	—	—	77.1	8.0	—	—	77.2
		0.5	—	—	—	83.9	—	—	—	84.0	—	—	—	84.4	—	—	—	84.5
		0.7	—	—	—	91.4	—	—	—	91.4	—	—	—	91.9	—	—	—	92.0
		0.8	—	—	—	94.8	—	—	—	94.8	—	—	—	95.1	—	—	—	95.2
		0.9	—	—	—	97.8	—	—	—	97.8	—	—	—	98.1	—	—	—	98.1
	100	0	—	—	—	72.9	—	—	—	73.2	—	—	—	73.8	—	—	—	74.1
		0.1	—	—	—	74.3	—	—	—	74.3	3.7	—	—	75.3	5.4	—	—	75.4

TABLE 3 (Continued)

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*			
n_1	n_2	δ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
		0.2	—	—	—	76.4	—	—	—	76.7	5.9	—	—	77.4	10.0	—	—	77.6
		0.3	—	—	—	79.3	—	—	—	79.4	12.7	—	—	80.2	15.3	—	—	80.2
		0.5	—	—	—	85.8	—	—	—	85.9	—	—	—	86.5	10.7	—	—	86.5
		0.7	—	—	—	92.5	—	—	—	92.6	15.0	—	—	93.1	22.3	—	—	93.1
		0.8	—	—	—	95.5	—	—	—	95.6	—	—	—	95.9	—	—	—	95.9
		0.9	—	—	—	98.1	—	—	—	98.2	—	—	—	98.3	—	—	—	98.4
100	100	0	—	—	—	76.8	—	—	—	76.9	—	—	—	76.8	—	—	—	76.9
		0.1	—	—	—	77.8	—	—	—	78.0	—	—	—	78.0	4.0	—	—	78.1
		0.2	—	—	—	79.7	—	—	—	79.8	4.5	—	—	79.9	4.5	—	—	80.0
		0.3	—	—	—	82.2	—	—	—	82.2	8.3	—	—	82.4	8.3	—	—	82.4
		0.5	—	—	—	87.9	—	—	—	88.0	—	—	—	88.2	8.1	—	—	88.3
		0.7	—	—	—	93.6	—	—	—	93.6	14.6	—	—	94.0	14.6	—	—	94.0
		0.8	—	—	—	96.2	—	—	—	96.2	—	—	—	96.4	—	—	—	96.4
		0.9	—	—	—	98.4	—	—	—	98.5	—	—	—	98.6	—	—	—	98.7

Notes: These values were obtained in the combinations of values of (n_1, n_2) which are indicated. To understand the name given to each method, see Table 1. The symbol "—" in some boxes indicates that the percentage is "0.0."

$q_1p_2/q_2p_1 = 1/\theta$ which are obtained by permuting the samples and/or permuting the successes with the failures.

Table 2 provides the average values of the parameters $\{\Theta, S5, S6, S7\}$ in all of the combinations of values of $(n_1, n_2, \delta/\rho/\theta)$, for error $\alpha = 5\%$ and for 24 inference methods which have been defined. For parameters d and R, it can be observed that the methods based on statistic X provide excessive values of $S5$ and $S6$, and even for $S7$. Nevertheless, for the OR parameter these values are always null or very moderate, when $S7 = 0$ always. Therefore, the first conclusion is that we should not apply the methods from statistic X (Dunnett & Gent, 1977, obtained something similar in the case of parameter d). On the contrary, all of the methods based on statistic Z have very small values of $S5$, $S6$, and $S7$, and therefore they are all acceptable methods to carry out the inference. Now we will start to select the optimal inference method for each parameter, restricted our selection to the methods based on statistic Z. In order to do so, Tables 3–5 contain the individual results for the values of the parameters $\{\Theta, S5, S6, S7\}$ in each combination of values of $(n_1, n_2, \delta/\rho/\theta)$, for the methods based on statistic Z and for the parameters d, R, and OR, respectively. The results for statistic X can be seen in the Tables S1 and S2 (only for parameters d and R). Again, the analysis and conclusions that follow are based on the criteria described in the previous section; other criteria may provide different conclusions.

For parameter d, analyzing its results from Table 3, we reach the following conclusion. For those who are in favor of doubling the p-value, the optimal method is ZCD, since it is always equal to or more powerful than ZYD (as was to be expected, due to its very definition) and its $S5$, $S6$, and $S7$ values are always very small. In general, if the four methods are compared, all of them are conservative or slightly liberal, and therefore it is sufficient to select the most powerful method

TABLE 4 Percentage of the values of the proportion p_1 in which the test provides an error $\alpha(p_1)$ higher than 5% (*S5*), 6% (*S6*) or 7% (*S7*), and percentage of points Θ in the critical region, for the four inference methods indicated (*ZYD*, *ZCD*, *ZYM* and *ZCM*) when a two-tailed test to an error $\alpha = 5\%$ is performed for H_R: $R = \rho$.

Method →			***ZYD***				***ZCD***				***ZYM***				***ZCM***			
n_1	n_2	ρ	*S5*	*S6*	*S7*	*Θ*	*S5*	*S6*	*S7*	*Θ*	*S5*	*S6*	*S7*	*Θ*	*S5*	*S6*	*S7*	*Θ*
20	20	0.05	—	—	—	85.9	—	—	—	86.6	—	—	—	87.1	14.6	8.4	—	88.0
		0.1	—	—	—	80.5	—	—	—	81.2	—	—	—	82.5	—	—	—	82.8
		0.2	—	—	—	71.7	—	—	—	72.3	—	—	—	74.2	—	—	—	74.4
		0.5	—	—	—	57.6	—	—	—	58.1	—	—	—	60.3	—	—	—	60.8
		0.8	—	—	—	49.7	—	—	—	50.6	—	—	—	52.4	—	—	—	52.6
		1	—	—	—	46.7	—	—	—	48.5	—	—	—	46.7	—	—	—	48.5
		1.25	—	—	—	49.7	—	—	—	50.6	—	—	—	52.4	—	—	—	52.6
		2	—	—	—	57.6	—	—	—	58.1	—	—	—	60.3	—	—	—	60.8
		5	—	—	—	71.7	—	—	—	72.3	—	—	—	74.2	—	—	—	74.4
		10	—	—	—	80.5	—	—	—	81.2	—	—	—	82.5	—	—	—	82.8
		20	—	—	—	85.9	—	—	—	86.6	—	—	—	87.1	14.6	8.4	—	88.0
	40	0.05	—	—	—	86.3	—	—	—	86.9	—	—	—	87.5	17.3	3.3	—	88.3
		0.1	—	—	—	81.2	—	—	—	81.8	—	—	—	82.8	8.1	4.6	—	83.4
		0.2	—	—	—	72.9	—	—	—	73.3	—	—	—	74.8	—	—	—	75.7
		0.5	—	—	—	60.2	—	—	—	60.6	—	—	—	62.1	—	—	—	62.6
		0.8	—	—	—	54.6	—	—	—	55.1	—	—	—	56.7	—	—	—	57.5
		1	—	—	—	53.9	—	—	—	54.6	—	—	—	55.5	—	—	—	56.5
		1.25	—	—	—	58.3	—	—	—	58.7	—	—	—	60.5	—	—	—	61.0
		2	—	—	—	66.2	—	—	—	66.7	—	—	—	67.0	—	—	—	67.5
		5	—	—	—	78.6	—	—	—	79.1	—	—	—	80.4	—	—	—	80.8
		10	—	—	—	85.1	—	—	—	85.1	3.4	—	—	86.5	3.4	—	—	86.8
		20	—	—	—	90.0	—	—	—	90.1	—	—	—	90.8	—	—	—	91.2
	60	0.05	—	—	—	86.4	2.5	—	—	87.0	—	—	—	87.6	24.7	6.9	3.1	88.4
		0.1	—	—	—	81.6	—	—	—	82.1	—	—	—	82.9	5.2	2.4	—	83.5
		0.2	—	—	—	73.2	—	—	—	73.7	—	—	—	75.1	2.7	—	—	75.6
		0.5	—	—	—	61.3	—	—	—	61.7	—	—	—	63.5	2.7	0.7	—	63.7
		0.8	—	—	—	56.7	—	—	—	57.1	—	—	—	58.9	—	—	—	59.3
		1	—	—	—	57.3	—	—	—	57.5	—	—	—	58.7	—	—	—	59.0
		1.25	—	—	—	62.3	—	—	—	62.5	—	—	—	64.3	6.7	—	—	64.5
		2	—	—	—	70.7	—	—	—	71.0	—	—	—	72.4	11.3	—	—	72.6
		5	—	—	—	82.2	—	—	—	82.4	—	—	—	83.5	—	—	—	83.6
		10	—	—	—	87.5	—	—	—	87.7	—	—	—	88.7	—	—	—	88.7
		20	—	—	—	91.4	—	—	—	91.7	—	—	—	92.2	9.2	—	—	92.5

TABLE 4 (Continued)

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*				
n_1	n_2	ρ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
	100	0.05	—	—	—	86.5	4.0	—	—	87.1	5.0	—	—	87.7	32.8	12.4	6.0	88.5
		0.1	—	—	—	81.7	—	—	—	82.2	—	—	—	83.1	6.8	4.1	2.1	83.6
		0.2	—	—	—	73.6	—	—	—	74.1	—	—	—	75.6	1.5	—	—	75.9
		0.5	—	—	—	62.2	—	—	—	62.3	—	—	—	64.1	4.5	—	—	64.6
		0.8	—	—	—	58.1	—	—	—	58.7	—	—	—	60.3	2.9	—	—	61.1
		1	—	—	—	59.7	—	—	—	60.1	—	—	—	61.5	—	—	—	62.2
		1.25	—	—	—	65.0	—	—	—	65.4	—	—	—	67.0	—	—	—	67.3
		2	—	—	—	74.6	—	—	—	74.7	14.7	—	—	76.3	15.0	—	—	76.5
20	100	5	—	—	—	85.5	—	—	—	85.6	7.5	—	—	86.0	9.7	—	—	86.1
		10	—	—	—	90.0	—	—	—	90.1	—	—	—	90.6	—	—	—	90.6
		20	—	—	—	93.2	—	—	—	93.2	—	—	—	93.7	—	—	—	93.8
40	40	0.05	—	—	—	90.4	—	—	—	90.5	—	—	—	91.1	8.1	4.6	—	91.4
		0.1	—	—	—	85.8	—	—	—	86.0	2.2	—	—	86.9	2.2	—	—	87.2
		0.2	—	—	—	80.1	—	—	—	80.2	0.5	—	—	81.1	0.5	—	—	81.5
		0.5	—	—	—	70.1	—	—	—	70.5	13.5	—	—	72.0	13.5	—	—	72.2
		0.8	—	—	—	64.3	—	—	—	64.7	—	—	—	66.6	—	—	—	66.8
		1	—	—	—	62.6	—	—	—	63.1	—	—	—	62.6	—	—	—	63.1
		1.25	—	—	—	64.3	—	—	—	64.7	—	—	—	66.6	—	—	—	66.8
		2	—	—	—	70.1	—	—	—	70.5	13.5	—	—	72.0	13.5	—	—	72.2
		5	—	—	—	80.1	—	—	—	80.2	0.5	—	—	81.1	0.5	—	—	81.5
		10	—	—	—	85.8	—	—	—	86.0	2.2	—	—	86.9	2.2	—	—	87.1
		20	—	—	—	90.4	—	—	—	90.5	—	—	—	91.1	8.1	4.6	—	91.4
	60	0.05	—	—	—	90.4	—	—	—	90.6	—	—	—	91.2	5.2	2.4	—	91.5
		0.1	—	—	—	86.0	—	—	—	86.2	3.3	—	—	87.2	6.1	—	—	87.4
		0.2	—	—	—	80.4	—	—	—	80.6	0.5	—	—	81.7	1.4	—	—	81.9
		0.5	—	—	—	71.3	—	—	—	71.5	5.7	—	—	72.5	5.7	—	—	72.6
		0.8	—	—	—	66.8	—	—	—	67.0	—	—	—	68.4	—	—	—	68.6
		1	—	—	—	66.1	—	—	—	66.4	—	—	—	67.2	—	—	—	67.5
		1.25	—	—	—	68.6	—	—	—	68.9	1.9	—	—	70.1	1.9	—	—	70.2
		2	—	—	—	74.5	—	—	—	74.7	10.1	—	—	75.7	10.1	—	—	75.8
		5	—	—	—	83.4	—	—	—	83.5	14.2	—	—	84.5	14.7	—	—	84.6
		10	—	—	—	88.1	—	—	—	88.2	—	—	—	89.0	5.5	—	—	89.2
		20	—	—	—	91.9	—	—	—	92.0	—	—	—	92.6	6.7	2.3	—	92.7

TABLE 4 (Continued)

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*			
n_1	n_2	ρ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
	100	0.05	—	—	—	90.5	—	—	—	90.8	—	—	—	91.3	5.7	4.1	2.1	91.6
		0.1	—	—	—	86.2	—	—	—	86.4	—	—	—	87.2	6.3	—	—	87.6
		0.2	—	—	—	80.8	—	—	—	81.0	0.5	—	—	82.0	3.4	1.6	—	82.2
		0.5	—	—	—	72.4	—	—	—	72.5	1.8	—	—	73.4	6.0	—	—	73.5
		0.8	—	—	—	68.9	—	—	—	69.0	—	—	—	70.3	—	—	—	70.4
		1	—	—	—	68.8	—	—	—	69.2	—	—	—	70.1	—	—	—	70.4
		1.25	—	—	—	72.2	—	—	—	72.4	—	—	—	73.3	6.6	—	—	73.4
		2	—	—	—	78.6	—	—	—	78.7	3.0	—	—	79.5	13.2	—	—	79.6
		5	—	—	—	86.7	—	—	—	86.7	10.9	—	—	87.4	10.9	—	—	87.4
		10	—	—	—	90.7	—	—	—	90.7	5.9	—	—	91.2	6.8	—	—	91.3
		20	—	—	—	93.4	—	—	—	93.5	—	—	—	93.9	—	—	—	94.0
60	60	0.05	—	—	—	92.0	—	—	—	92.2	—	—	—	92.6	5.6	3.2	—	92.8
		0.1	—	—	—	88.3	—	—	—	88.4	5.5	—	—	89.2	6.1	—	—	89.4
		0.2	—	—	—	83.9	—	—	—	84.0	9.2	—	—	84.6	11.7	—	—	84.8
		0.5	—	—	—	75.8	—	—	—	75.9	11.5	—	—	77.1	11.5	—	—	77.1
		0.8	—	—	—	71.2	—	—	—	71.4	4.6	—	—	72.6	5.8	—	—	72.7
		1	—	—	—	69.7	—	—	—	69.7	—	—	—	69.7	—	—	—	69.7
		1.25	—	—	—	71.2	—	—	—	71.4	4.6	—	—	72.6	5.8	—	—	72.7
		2	—	—	—	75.8	—	—	—	75.9	11.5	—	—	77.1	11.5	—	—	77.1
		5	—	—	—	83.9	—	—	—	84.0	9.2	—	—	84.6	11.7	—	—	84.8
		10	—	—	—	88.3	—	—	—	88.4	5.5	—	—	89.2	6.1	—	—	89.4
		20	—	—	—	92.0	—	—	—	92.2	—	—	—	92.6	5.6	3.2	—	92.8
	100	0.05	—	—	—	92.1	—	—	—	92.2	—	—	—	92.7	2.3	—	—	92.8
60	100	0.1	—	—	—	88.5	—	—	—	88.6	—	—	—	89.2	0.2	—	—	89.4
		0.2	—	—	—	84.2	—	—	—	84.3	7.9	—	—	85.1	7.9	—	—	85.2
		0.5	—	—	—	76.9	—	—	—	77.0	9.5	—	—	78.0	10.5	1.2	0.2	78.2
		0.8	—	—	—	73.5	—	—	—	73.6	3.5	—	—	74.7	3.5	—	—	74.8
		1	—	—	—	72.9	—	—	—	73.2	—	—	—	73.8	—	—	—	74.1
		1.25	—	—	—	75.2	—	—	—	75.3	7.2	—	—	76.3	8.4	—	—	76.5
		2	—	—	—	80.1	—	—	—	80.3	13.3	—	—	81.1	14.5	—	—	81.2
		5	—	—	—	87.1	—	—	—	87.2	9.6	—	—	87.7	10.5	—	—	87.7
		10	—	—	—	90.8	—	—	—	90.9	1.9	0.4	—	91.4	4.6	1.1	—	91.5
		20	—	—	—	93.6	—	—	—	93.6	—	—	—	94.0	2.8	—	—	94.1

TABLE 4 (Continued)

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*			
n_1	n_2	ρ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
100	100	0.05	—	—	—	93.6	—	—	—	93.7	3.0	—	—	94.1	8.0	1.9	—	94.2
		0.1	—	—	—	91.0	—	—	—	91.1	5.8	—	—	91.5	5.8	—	—	91.6
		0.2	—	—	—	87.6	—	—	—	87.7	10.5	—	—	88.1	10.5	—	—	88.2
		0.5	—	—	—	81.4	—	—	—	81.5	10.4	0.2	—	82.1	10.4	0.2	—	82.2
		0.8	—	—	—	77.9	—	—	—	78.0	4.5	—	—	78.7	6.7	—	—	78.8
		1	—	—	—	76.8	—	—	—	76.9	—	—	—	76.8	—	—	—	76.9
		1.25	—	—	—	77.9	—	—	—	78.0	4.5	—	—	78.7	6.7	—	—	78.8
		2	—	—	—	81.4	—	—	—	81.5	10.4	0.2	—	82.1	10.4	0.2	—	82.2
		5	—	—	—	87.6	—	—	—	87.7	10.5	—	—	88.1	10.5	—	—	88.2
		10	—	—	—	91.0	—	—	—	91.1	5.8	—	—	91.5	5.8	—	—	91.6
		20	—	—	—	93.6	—	—	—	93.7	3.0	—	—	94.1	8.0	1.9	—	94.2

Notes: These values were obtained in the combinations of values of (n_1, n_2) which are indicated. To understand the name given to each method, see Table 1. The symbol "—" in some boxes indicates that the percentage is "0.0."

which, by definition, must be an *M* type method. It can be observed that the method selected is *ZCM*, since it almost always provides a power which is somewhat higher than that of the rest of the methods, especially in the smallest sample sizes.

For parameter *R*, analyzing its results from Table 4, we reach the following conclusion. The four methods are always conservative or slightly liberal, except in the case of method *ZYD*, which is always conservative; therefore, all of them are acceptable. The exception is method *ZCM*, which takes values $S7 > 0$ when ρ is very extreme. For those who are in favor of doubling the *p*-value, the optimal method is *ZCD*, since it is always equal to or slightly more powerful than *ZYD* (as was to be expected, due to its very definition) and its *S5*, *S6*, and *S7* values are always zero or very small. In general, if the four methods are compared, the most powerful one is *ZCM*, which almost always provides a power which is somewhat higher than that of the rest of the methods, especially in the smallest sample. Nevertheless, when ρ is very small (large) and $n_1 < n_2$ $(n_1 > n_2)$ *ZYM* is preferable, since in this method $S7 = 0$.

For the *OR* parameter, analyzing its results from Table 5, we reach the following conclusion. In general, it is observed that the four methods are usually conservative, although sometimes they are somewhat liberal when $n_1 \neq n_1$ and θ takes a value which is far from 1; as the value $S7 > 0$ is only achieved on one occasion (the *ZCM* method, with very different values of n_i and θ which are very far from 1), it can be concluded that the four methods are acceptable to perform the inference. For those in favor of "doubling the *p*-value", the *ZCD* method has values for Θ higher than those of the *ZYD* method, as was to be expected; moreover, *ZYD* is always conservative and *ZCD* is slightly liberal on a few occasions on which $n_1 \neq n_2$ and θ is far from 1. The conclusion is that in this case *ZCD* is the best method. In general, if we compare the four methods, the method selected is *ZCM*, since it has a higher value for Θ and is not very liberal at all; the only exception is when n_1 is very different to n_2 and θ is extreme ($\theta \leq 0.05$ ó $\theta \geq 20$), since then *S7* may be higher than 0. For these cases the *ZYM* method is preferable.

TABLE 5 Percentage of the values of the proportion p_1 in which the test provides an error $\alpha(p_1)$ higher than 5% (*S5*), 6% (*S6*) or 7% (*S7*), and percentage of points Θ in the critical region, for the four inference methods indicated (*ZYD*, *ZCD*, *ZYM* and *ZCM*) when a two-tailed test to an error $\alpha = 5\%$ is performed for H_{OR}: $OR = \theta$.

Method →			***ZYD***				***ZCD***				***ZYM***				***ZCM***			
n_1	n_2	θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
20	20	1	—	—	—	46.7	—	—	—	48.5	—	—	—	46.7	—	—	—	48.5
		1.25	—	—	—	47.9	—	—	—	49.0	—	—	—	49.9	—	—	—	50.3
		2	—	—	—	49.4	—	—	—	49.4	—	—	—	51.7	—	—	—	51.7
		5	—	—	—	54.4	—	—	—	54.4	—	—	—	57.8	—	—	—	58.7
		10	—	—	—	61.5	—	—	—	62.4	—	—	—	63.3	—	—	—	64.6
		20	—	—	—	66.9	—	—	—	67.8	—	—	—	69.8	—	—	—	70.8
	40	1	—	—	—	53.9	—	—	—	54.6	—	—	—	55.5	—	—	—	56.5
		1.25	—	—	—	54.9	—	—	—	55.3	—	—	—	57.0	—	—	—	57.6
		2	—	—	—	55.5	—	—	—	56.1	—	—	—	58.1	—	—	—	58.9
		5	—	—	—	60.6	—	—	—	61.1	—	—	—	63.1	—	—	—	63.7
		10	—	—	—	65.9	—	—	—	66.4	—	—	—	67.8	—	—	—	68.6
		20	—	—	—	71.8	0.2	—	—	72.4	—	—	—	73.5	0.6	0.4	—	74.3
	60	1	—	—	—	57.3	—	—	—	57.5	—	—	—	58.7	—	—	—	58.9
		1.25	—	—	—	57.4	—	—	—	57.8	—	—	—	59.6	—	—	—	60.0
		2	—	—	—	58.4	—	—	—	58.7	—	—	—	60.5	1.4	0.5	—	60.9
		5	—	—	—	63.1	—	—	—	63.6	—	—	—	65.0	0.7	—	—	65.7
		10	—	—	—	67.8	—	—	—	68.5	—	—	—	70.3	0.7	0.4	—	70.7
		20	—	—	—	73.5	—	—	—	73.9	—	—	—	75.2	6.8	0.3	—	76.0
	100	1	—	—	—	59.7	—	—	—	60.1	—	—	—	61.5	—	—	—	62.1
		1.25	—	—	—	59.6	—	—	—	60.0	—	—	—	62.1	1.2	—	—	62.7
		2	—	—	—	60.6	—	—	—	61.1	—	—	—	62.8	—	—	—	63.5
		5	—	—	—	65.2	—	—	—	65.5	—	—	—	67.1	0.4	—	—	67.6
		10	—	—	—	69.9	0.1	—	—	70.4	—	—	—	71.7	0.5	0.2	—	72.1
		20	—	—	—	75.0	0.2	—	—	75.7	—	—	—	76.7	0.5	0.4	0.2	77.3
40	40	1	—	—	—	62.6	—	—	—	63.1	—	—	—	62.6	—	—	—	63.1
		1.25	—	—	—	62.8	—	—	—	62.9	—	—	—	65.0	—	—	—	65.0
		2	—	—	—	63.5	—	—	—	63.8	—	—	—	65.6	—	—	—	65.8
		5	—	—	—	67.2	—	—	—	67.7	—	—	—	69.1	—	—	—	69.6
		10	—	—	—	71.5	—	—	—	72.4	—	—	—	73.2	—	—	—	73.7
		20	—	—	—	76.6	—	—	—	76.9	—	—	—	78.1	—	—	—	78.6

TABLE 5 (Continued)

Method →			*ZYD*				*ZCD*				*ZYM*				*ZCM*			
n_1	n_2	θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ	*S5*	*S6*	*S7*	Θ
	60	1	—	—	—	66.1	—	—	—	66.4	—	—	—	67.5	—	—	—	67.9
		1.25	—	—	—	66.2	—	—	—	66.3	—	—	—	67.7	—	—	—	67.9
		2	—	—	—	67.0	—	—	—	67.1	—	—	—	68.4	—	—	—	68.5
		5	—	—	—	70.5	—	—	—	70.7	—	—	—	72.1	0.3	—	—	72.4
		10	—	—	—	74.3	—	—	—	74.6	—	—	—	75.5	3.3	—	—	76.0
		20	—	—	—	78.5	—	—	—	78.9	—	—	—	79.7	7.7	3.7	—	80.3
	100	1	—	—	—	68.8	—	—	—	69.2	—	—	—	70.1	—	—	—	70.4
		1.25	—	—	—	69.1	—	—	—	69.3	—	—	—	70.5	—	—	—	70.6
		2	—	—	—	69.9	—	—	—	70.0	—	—	—	71.0	—	—	—	71.1
		5	—	—	—	72.9	—	—	—	73.2	—	—	—	74.5	—	—	—	74.6
		10	—	—	—	76.6	—	—	—	76.8	—	—	—	77.8	0.2	—	—	78.0
		20	—	—	—	80.5	0.1	—	—	80.7	—	—	—	81.6	0.2	0.1	—	81.9
60	60	1	—	—	—	69.7	—	—	—	69.7	—	—	—	69.7	—	—	—	69.7
		1.25	—	—	—	69.6	—	—	—	69.9	—	—	—	71.2	—	—	—	71.3
		2	—	—	—	70.4	—	—	—	70.6	—	—	—	71.7	—	—	—	71.8
		5	—	—	—	73.4	—	—	—	73.7	—	—	—	74.9	—	—	—	75.0
		10	—	—	—	76.8	—	—	—	77.2	—	—	—	78.1	—	—	—	78.2
		20	—	—	—	80.7	—	—	—	80.9	—	—	—	81.8	—	—	—	82.0
	100	1	—	—	—	72.9	—	—	—	73.2	—	—	—	73.8	—	—	—	74.1
		1.25	—	—	—	73.0	—	—	—	73.2	—	—	—	74.2	—	—	—	74.4
		2	—	—	—	73.6	—	—	—	73.8	—	—	—	74.9	—	—	—	75.0
		5	—	—	—	76.4	—	—	—	76.5	—	—	—	77.4	—	—	—	77.6
		10	—	—	—	79.4	—	—	—	79.6	—	—	—	80.4	2.9	1.1	—	80.6
		20	—	—	—	82.8	—	—	—	82.9	—	—	—	83.7	0.1	—	—	83.9
100	100	1	—	—	—	76.8	—	—	—	76.9	—	—	—	76.8	—	—	—	76.9
		1.25	—	—	—	76.6	—	—	—	76.8	—	—	—	77.7	—	—	—	77.8
		2	—	—	—	77.4	—	—	—	77.4	—	—	—	78.2	—	—	—	78.3
		5	—	—	—	79.7	—	—	—	79.8	—	—	—	80.5	—	—	—	80.6
		10	—	—	—	82.2	—	—	—	82.3	—	—	—	83.1	—	—	—	83.2
		20	—	—	—	85.1	—	—	—	85.2	—	—	—	85.8	—	—	—	86.0

Notes: These values were obtained in the combinations of values of (n_1, n_2) which are indicated. To understand the name given to each method, see Table 1. The current Z^{**} statistics are the same as the X^{**} statistics referred to in text. The symbol "—" in some boxes indicates that the percentage is "0.0."

TABLE 6 Two-sided conditional confidence intervals for the parameters d, R, and OR when $x_1 = 48$, $n_1 = 102$, $x_2 = 11$ and $n_2 = 46$.

			95% confidence interval	
Parameter	**Inference method (conditional)**	**Type**	**Lower**	**Upper**
d	Doubling the p-value			
	ZCD (typified value with the *cc* of Conover)	A	0.0476	0.3777
	Semi-exact (noncentral hypergeometric)	SE	0.0490	0.3818
	Adding the p-values of each tail			
	ZCM (typified value with the *cc* of Conover)	A	0.0579	0.3735
	Semi-exact (noncentral hypergeometric)	SE	0.0583	0.3754
R	Doubling the p-value			
	ZCD (typified value with the *cc* of Conover)	A	1.130	3.729
	Semi-exact (noncentral hypergeometric)	SE	1.134	3.817
	Adding the p-values of each tail			
	ZCM (typified values with the *cc* of Conover)	A	1.161	3.643
	Semi-exact (noncentral hypergeometric)	SE	1.163	3.684
OR	Doubling the p-value			
	ZCD (typified value with the *cc* of Conover)	A	1.221	6.638
	Exact (noncentral hypergeometrical)	E	1.229	6.837
	Adding the p-values of each tail			
	ZCM (typified value with the *cc* of Conover)	A	1.277	6.446
	Exact (noncentral hypergeometric)	E	1.279	6.537

Note: Data from Maxwell (1961).
Abbreviations: A, asymptotic; E, exact; SE, semi-exact (in the sense of Fleiss, Levin, & Paik, 2003).

It can be observed that the general conclusion is the same in all case: the best inference method is *ZCD* for those in favor of doubling the p-value or *ZCM* in general.

4 | EXAMPLE

Section 2 describes the different statistics that allowed us to perform a conditional asymptotic hypothesis test for the parameters d, R, and OR. If we want a CI for the aforementioned parameters, it is sufficient to invert the appropriate test. In the case of the methods based on doubling the p-value (methods which end in the letter D), the solution is obtained directly from the value for the statistic: if $\chi^2_{\text{exp}}(\delta)$ is the value of a statistic for the test H_d: $d = \delta$ versus K_d: $d \neq \delta$, then the two-tailed CI to 95% of confidence for d is given by $\delta_L \leq d \leq \delta_U$, where $\chi^2_{\text{exp}}(\delta_L) = \chi^2_{\text{exp}}(\delta_U) = \chi^2_{5\%}$. This occurs in a similar way for the other two parameters R and OR. However, in the case of the methods based on the criteria of Mantel (methods which end in the letter M), the solution must be obtained through the p-value of the two-tailed test: the CI $\delta_L \leq d \leq \delta_U$ is such that the p-value $p(x_1, \delta)$ of the two-tailed test in the observed value x_1 verifies that $p(x_1, \delta_L) = p(x_1, \delta_U) = 5\%$.

In order to illustrate the two methods selected for the three parameters (*ZCD* and *ZCM*), Table 6 shows the results from the example cited by Maxwell (1961). The data refer to the rate occurrence of virus infection among the group of those inoculated p_1 ($x_1 = 48$ from among $n_1 = 102$ individuals) and the group of those not inoculated p_2 ($x_2 = 11$ from among $n_2 = 46$ individuals). The StatXact7 package (StatXact7, 2005) indicates that $1.229 \leq OR \leq 6.837$ is the conditional exact 95%-CI, which is obtained through the inversion of the two-tailed test and through the method of doubling the *p*-value. The CI determined by the method of the "tables as improbable or more" of Baptista and Pike (1977) is $1.279 \leq OR \leq 6.537$, which was obtained through the web page https://rdrr.io/cran/ORCI/man/BPexact.CI.html. Note that this second method provides a CI which is narrower than the first one. The semi-exact conditional CIs for *d* and *R* are deduced from the exact conditional CI for *OR* through the procedure of Fleiss, Levin & Paik, 2003, although this CI is only for the values of *d* or *R* in the values of the conditional expected frequencies. For example, if $OR = \theta$ is one of the extremes of the CI for *OR*, then the value of $n_1\hat{p}_1 = \hat{x}_1$ is such that $\theta = \hat{x}_1(a_2 - n_1 + \hat{x}_1)/(a_1 - \hat{x}_1)(n_1 - \hat{x}_1)$, from which it is deduced that $\hat{x}_1 = (A - B)/2(\theta - 1)$ when $A = \theta(a_1 + n_1) + (a_2\text{-}n_1)$ and $B = [A^2\text{-}\, 4n_1a_1\theta(\theta\text{-}1)]^{0.5}$; the consequence is that the extreme of the CI for *R* will be $\rho = \hat{x}_1 n_2/n_1(a_1 - \hat{x}_1)$. It can be observed that the asymptotic methods provide results which are quite close to those of the exact or semi-exact conditional methods, especially in their lower extremes, since the upper are somewhat liberal.

5 | CONCLUSIONS

The determination of a CI for the parameters difference (*d*), ratio (*R*) and odds-ratio (*OR*) of two independent proportions is a frequent objective in statistics. The two-tailed CI is obtained through inversion of the two-tailed test (Agresti & Min, 2001), a test which may be conditional or unconditional. Here the conditional point of view has been adopted (as Prescott, 2019). The CI will be exact when it is obtained based on the noncentral hypergeometric distribution, but that is only possible for the *OR* parameter (the only parameter of the previous distribution). If the CI of *OR* is transferred to *R* or to *d*, the CI is said to be semi-exact, since there is no biunivocal relation between *OR* and *d* or *R*. The intensity of the calculation of the CI is simplified notably when using asymptotic distributions, thus obtaining approximate CIs based on chi-square type statistics (*X*) or *z* type (*Z*). In order to be coherent with the point of view adopted, in both statistics we use the conditional estimators of the nuisance parameter. However, a *cc* must be applied. Here two *cc* have been proposed: the classic *Y* by Yates (1934) and the less well-known *C* by Conover (1974). The four statistics that are obtained in this way (*XY*, *XC*, *ZY* and *ZC*) can be used to obtain the *p*-value of the two-tailed test that must be inverted. It is habitual (Armitage, 1971; Prescott, 2019) to define the *p*-value of two-tailed test as double the *p*-value of a one-tailed test, which leads to method *D* and causes a conservative test. The CI which is obtained in that way, is said to have been obtained through the method called *TOST* (two one-sided test). Another more efficient option is to use the *M* method by Mantel (1974), for which the *p*-value of the two-tailed test is the sum of two *p*-values of one tail: the one in the original table and the one in the table with the other tail which is "as extreme or more than the original table." The CI which is obtained in that way, is said to have been obtained by the method called *TTST* (two two-sided tests). Therefore, the four statistics cited above lead to eight inference methods (see Table 1), although in the case of *OR* there are only four different methods, since only in that case we can find that $XY = ZY$ and $XC = ZC$.

This article has assessed the inference methods highlighted in the previous paragraph, when these are used with the *d*, *R*, or *OR* parameters. One initial conclusion is that for parameters *d* and *R* all of the methods based on the chi-square statistic (*X*) must be ruled out, since they lead to very liberal CIs; on the contrary, the methods based on the *z* statistics (*Z*) are all acceptable and

generally conservative. A second conclusion concerns which inference method is preferable to provide narrower CIs. Whether or not one is in favor of the *TOST* method (i.e., of the *p*-value being obtained through the *D* method) or the *TTST* method (of the *p*-value being obtained through the *M* method), the selection is always that the best inference method is that based in the *ZC* statistic.

ACKNOWLEDGMENTS

This research was supported by the Ministry of Science and Innovation (Spain), Grant PID2021-126095NB-I00 funded by MCIN/AEI/10.13039/5 01100011033 and by "ERDF A way of making Europe". Funding for open access charge: Universidad de Granada / CBUA.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Antonio Martín Andrés https://orcid.org/0000-0002-2548-2638

REFERENCES

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *The American Statistician*, *54*(4), 280–288. https://doi.org/10.2307/2685779

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.2307/2685469

Agresti, A., & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*, *57*, 963–971. https://doi.org/10.1111/j.0006-341X.2001.00963.x

Armitage, P. (1971). *Statistical methods in medical research*. Oxford, England: Blackwell Scientific Publications.

Baptista, J., & Pike, M. C. (1977). Exact two-sided confidence limits for the odds ratio in a 2x2 table. *Applied Statistics*, *26*(2), 214–220. https://doi.org/10.2307/2347041

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika*, *34*, 123–138.

Chen, L. A., Hung, H. N., & Chen, C. R. (2007). Maximum average-power (MAP) test. *Communications in Statistics-Theory & Methods*, *36*, 2237–2249.

Conover, W. J. (1974). Some reasons for not using the Yates' continuity corrections on 2×2 contingency tables. *Journal of the American Statistical Association*, *69*, 374–376.

Cornfield, J. (1956). *A statistical problem arising from retrospective studies*. In *Proceedings of third Berkeley symposium on mathematical statistics and probability* (pp. 135–148). Berkeley: University of California Press.

Cox, D. R. (1970). The continuity correction. *Biometrika*, *57*, 217–219.

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics*, *33*, 593–602.

Farrington, C. P., & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, *9*, 1447–1454. https://doi.org/10.1002/sim.4780091208

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society A*, *98*, 39–54.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York, NY: John Wiley & Sons.

Gart, J. J., & Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics*, *44*, 323–338.

Haber, M. (1980). A comparison of some continuity corrections for the chi-squared test on 2×2 tables. *Journal of the American Statistical Association*, *75*, 510–515.

Hamdan, M. A. (1974). On the continuity correction. *Technometrics*, *16*(4), 631–632.

Katz, D., Baptista, J., Azen, S. P., & Pike, M. C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, *34*, 469–474. https://doi.org/10.2307/2530610

Mantel, N. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables: Comment and a suggestion. *Journal of the American Statistical Association*, *69*, 378–380. https://doi.org/10.1080/01621459.1974.10482959

Mantel, N. (1990). Comment. *Statistics in Medicine*, *9*, 369–370.

Martín Andrés, A. (1998). *Fisher's exact and Barnard's tests*. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences: Update* (Vol. *2*, pp. 250–258). New York, NY: Wiley-Interscience.

Martín Andrés, A. (2008). Comments on "chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations". *Statistics in Medicine*, *27*(10), 1791–1795 (Reply in 1796-1796). https://doi.org/10.1002/sim.3169

Martín Andrés, A., Álvarez Hernández, M., & Herranz Tejedor, I. (2012). Asymptotic two-tailed confidence intervals for the difference of proportions. *Journal of Applied Statistics*, *39*(7), 1423–1435. https://doi.org/10.1080/02664763.2011.650686

Martín Andrés, A., & Herranz Tejedor, I. (2010). Asymptotic inferences about a linear combination of two proportions. *JP Journal of Biostatistics*, *4*(3), 253–277.

Martín Andrés, A., Herranz Tejedor, I., & Gayá Moreno, F. (2020). Comments on "Two-tailed significance tests for 2×2 contingency tables: What is the alternative?". *Statistics in Medicine*, *39*(4), 510–513. https://doi.org/10.1002/sim.8432

Martín Andrés, A., Herranz Tejedor, I., & Luna del Castillo, J. D. (1992). Optimal correction for continuity in the chi-squared test in 2×2 tables (conditioned method). *Communications in Statistic – Simulation and Computation*, *21*(4), 1077–1101.

Martín Andrés, A., & Silva Mato, A. (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational Statistics and Data Analysis*, *17*, 555–574. https://doi.org/10.1016/0167-9473(94)90148-1

Maxwell, A. E. (1961). *Analysing qualitative data*. London, England: Methuen.

Miettinen, O., & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, *4*, 213–226. https://doi.org/10.1002/sim.4780040211

Nam, J.-M. (1995). Confidence limits for the ratio of two binomial proportions based on likelihood scores: Non-iterative method. *Biometrical Journal*, *37*(3), 375–379.

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika*, *34*, 139–167.

Prescott, R. J. (2019). Two-tailed significance tests for 2×2 contingency tables: What is the alternative? *Statistics in Medicine*, *38*(22), 4264–4269. https://doi.org/10.1002/sim.8294

Schouten, H. J. (1976). On the continuity correction. *Statistica Neerlandika*, *30*, 93–95.

StatXact7. (2005). *CYTEL software corporation*. Retrieved from www.cytel.com

Upton, G. J. G. (1982). A comparison of alternative test for the 2×2 comparative trial. *Journal of the Royal Statistical Society A*, *145*(1), 86–105. https://doi.org/10.2307/2981423

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, *1*, 217–235.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Martín Andrés, A., Hernández, M. Á., & Gayá Moreno, F. (2024). The Yates, Conover, and Mantel statistics in 2×2 tables revisited (and extended). *Statistica Neerlandica*, *78*(2), 334–356. https://doi.org/10.1111/stan.12320