Published in final edited form as:

J Stat Comput Simul. 2018 ; 88(6): 1200–1210. doi:10.1080/00949655.2018.1425690.

Accurate unconditional p-values for a two-arm study with binary endpoints

Guogen Shan^{1,*}, Le Kang², Min Xiao³, Hua Zhang⁴, and Tao Jiang³

¹Epidemiology and Biostatistics Program, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA

²Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA

³Department of Statistics, Zhejiang Gongshang University, Hangzhou, People's Republic of China

⁴School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, People's Republic of China

Abstract

Unconditional exact tests are increasingly used in practice for categorical data to increase the power of a study and to make the data analysis approach being consistent with the study design. In a two-arm study with a binary endpoint, p-value based on the exact unconditional Barnard test is computed by maximizing the tail probability over a nuisance parameter with a range from 0 to 1. The traditional grid search method is able to find an approximate maximum with a partition of the parameter space, but it is not accurate and this approach becomes computationally intensive for a study beyond two groups. We propose using a polynomial method to rewrite the tail probability as a polynomial. The solutions from the derivative of the polynomial contain the solution for the global maximum of the tail probability. We use an example from a double-blind randomized Phase II cancer clinical trial to illustrate the application of the proposed polynomial method to achieve an accurate p-value. We also compare the performance of the proposed method and the traditional grid search method under various conditions. We would recommend using this new polynomial method in computing accurate exact unconditional p-values.

Keywords

Global maximum; Grid search; Independent proportions; Polynomial; Unconditional tests

1 Introduction

Unconditional exact tests become popular in categorical data analysis to compare two or more independent proportions [1, 5], dependent proportions [6], count data, and so on. Unlike the commonly used exact conditional test (e.g., Fisher's exact test) that assumes both marginal totals of a contingency table fixed, exact unconditional tests always assume only

*Corresponding author. Epidemiology and Biostatistics Program, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. guogen.shan@unlv.edu.

one marginal total or only the total sum fixed. In a study to compare the response rates from two groups, it is often the case that the number of participants from each group is pre-specified by the study design. Therefore, in the data analysis stage, the assumption of one marginal total fixed from the exact unconditional test is aligned with the study design. It should be noted that the final observed sample sizes of the two groups may be slightly different as planned in some cases. Then, the marginal total of sample size for each group is considered as fixed in the data analysis under the unconditional framework. In addition to nature of data generating mechanism, the exact unconditional test has been shown to be more powerful than the exact conditional test in many important statistical and medical research problems [4]. The exact unconditional test is often criticized for its computational intensity. However, with the development of computer hardware and software, the unconditional test becomes feasible when used in practice and it can be implemented on a personal computer.

In a study comparing two independent proportions, the exact unconditional test computes the p-value by maximizing the tail probability over the nuisance parameter which is the common response rate from the two groups. The rejection region is determined by a test statistic, such as the Wald test, according to the pre-specified hypotheses. Traditionally, the grid search method and numeric search methods are utilized in finding the maximum of the tail probability curve. When the curve is erratic, the grid search method has better performance than numerical search methods, since numerical methods depend on multiple factors in finding the accurate global maximum: initial value, search range, search algorithm, and convergence issue. However, the grid search method only computes an approximate maximum, not the accurate global maximum. In addition, the grid search method becomes computationally intensive in a study with multiple groups or multiple stages [10].

The probability density function of a binomial distribution can be rewritten as a polynomial. Then the tail probability is a sum of multiple polynomials, and the maximum possible degree of this polynomial is the total sample size of the study. We can then use functions from the statistical software R, such as *deriv* to obtain the derivative of the polynomial, and *solve* to solve the real solutions of the derivative being equal to zero. The final solution that leads to the global maximum of the tail probability is included in the solution set. It follows that the accurate global maximum is achieved by using this polynomial method. Due to the precision limitation from computer hardware and software, we utilize a piecewise approach to guarantee that the global maximum is correctly achieved.

The remainder of this article is organized as follows. In Section 2, we introduce in detail the new polynomial method for the accurate p-value calculation for a study comparing two independent proportions. We then use an example in Section 3 to illustrate the application of the proposed polynomial method and compare it with the traditional grid search method. Latter, we conduct extensive numerical studies to identify the required number of pieces in the new approach to guarantee accurate p-values. Finally, in Section 4, we provide some remarks on computing exact unconditional p-values.

2 Methods

When a new treatment is compared with the gold standard in a binary two-arm study, the hypotheses can be expressed as

$$H_0: p_E \leq p_C \text{ against } H_a: p_E > p_C,$$

where p_C and p_E are the probabilities of responses from the control group and the treatment group, respectively. Suppose the numbers of participants in the control group and the treatment group are N_C and N_E , respectively. In this study, the number of responses are X_C and X_E , which independently follow binomial distributions with parameters (N_C, p_C) and (N_E, p_E) , respectively. In a two-arm study with a binary outcome, the Wald test is commonly used,

$$T(X_E, X_C | N_E, N_C) = \frac{\frac{X_E}{N_E} - \frac{X_C}{N_C}}{\sqrt{\frac{X_C + X_E}{N_C + N_E} \left(1 - \frac{X_C + X_E}{N_C + N_E}\right) \left(\frac{1}{N_C} + \frac{1}{N_E}\right)}}. \quad (1)$$

In this test statistic, the pooled variance estimate is used.

The null hypothesis is rejected for a large test statistic. Therefore, the tail area for observed data (x_E, x_C) can be expressed as

$$\Omega(x_E, x_C) = \{(X_E, X_C) : T(X_E, X_C) \geq T(x_E, x_C)\},$$

and its associated tail probability is calculated as

$$\sum_{(X_E, X_C) \in \Omega(x_E, x_C)} P(X_E, X_C).$$

The null hypothesis space is a tri-angle space, and the actual type I error rate based on the test statistic T is obtained at the boundary with $p_C = p_E = p$ [8, 9], since the Barnard convexity condition is satisfied. Then, the tail probability

$$\sum_{(X_E, X_C) \in \Omega(x_E, x_C)} P(X_E, X_C | p) = \sum_{(X_E, X_C) \in \Omega(x_E, x_C)} \binom{N_C}{X_C} p^{X_C} (1-p)^{N_C - X_C} \binom{N_E}{X_E} p^{X_E} (1-p)^{N_E - X_E}$$

is a function of the unknown nuisance parameter, p , with a range from 0 to 1. In order to control for this nuisance parameter, several approaches have been proposed. The first

approach was to replace the nuisance parameter with its estimate, such as the maximum likelihood estimate, $\hat{p} = (x_E + x_C)/(N_E + N_C)$. This approach is often referred to be as the estimation approach, also known as the parametric bootstrap approach [2, 11]. Although this approach is computationally easy, in general it does not guarantee the significance level. For this reason, the second approach based on maximization is recommended for use in a study with a small to medium sample size, e.g., an early phase clinical trial. Let

$$H(p : x_E, x_C) = \sum_{(X_E, X_C) \in \Omega(x_E, x_C)} \binom{N_C}{X_C} \binom{N_E}{X_E} p^{X_C + X_E} (1-p)^{N_E + N_C - X_E - X_C}. \quad (2)$$

Then, the tail probability by using the exact unconditional approach is computed as

$$\max_{0 \leq p \leq 1} H(p : x_E, x_C). \quad (3)$$

The naive grid search method is traditionally used to search for the maximization of the tail probability over all possible p values from 0 to 1. It often requires a large number of points to get a good approximate solution. Alternatively, numerical search algorithms can be used in searching for the global maximum of the tail probability. The performance of these numerical algorithms is not satisfactory when a significant number of peaks appear in the tail probability curve. The tail probability curve is often erratic for a study with a binary outcome.

To obtain an accurate tail probability under the exact unconditional framework, we propose first using the polynomial method, then changing the summation index. We first rewrite the tail probability as a polynomial of p , specifically

$$\begin{aligned} p^{X_C + X_E} (1-p)^{N_E + N_C - X_E - X_C} &= p^{X_C + X_E} \sum_{i=0}^{N_E + N_C - X_E - X_C} \binom{N_E + N_C - X_E - X_C}{i} (-p)^{N_E + N_C - X_E - X_C - i} \\ &= \sum_{i=0}^{N_E + N_C - X_E - X_C} \binom{N_E + N_C - X_E - X_C}{i} (-1)^{N_E + N_C - X_E - X_C - i} p^{N_E + N_C - i}. \end{aligned}$$

It follows that

$$\binom{N_C}{X_C} \binom{N_E}{X_E} p^{X_C + X_E} (1-p)^{N_E + N_C - X_E - X_C} = \sum_{i=0}^{N_E + N_C} g(X_E, X_C, i) p^i,$$

where $g(X_E, X_C, i)$ is the coefficient of p^i , $i = 0, 1, \dots, N_E + N_C$, and

$$g(X_E, X_C, i) = \binom{N_C}{X_C} \binom{N_E}{X_E} \binom{N_E + N_C - X_E - X_C}{i - X_E - X_C} (-1)^{i - X_E - X_C} \text{ when } i \geq X_E + X_C, \text{ and } 0$$

otherwise. Let

$$F(p : x_E, x_C) = \sum_{i=0}^{N_E + N_C} \left[\sum_{(X_E, X_C) \in \Omega(x_E, x_C)} g(X_E, X_C, i) \right] p^i. \quad (4)$$

Then, the tail probability under the null hypothesis in Equation (3) can be alternatively expressed as

$$\max_{0 \leq p \leq 1} F(p : x_E, x_C). \quad (5)$$

Since $F(p : x_E, x_C)$ is a polynomial of p with the highest degree of $N_E + N_C$, it is straightforward to compute its derivative. The function *deriv* from the package *polynom* in the statistical software *R* can be used to calculate the derivative of a polynomial. Once the derivative is obtained, its solutions are then solved from the function *solve* in *R*. The derivative is a polynomial with the degree of $N_E + N_C - 1$, thus, there are $N_E + N_C - 1$ complex solutions. We only need to keep the real solutions, which contain all possible local maximum values except the boundary values, 0 and 1. We name this set, including the real solutions and boundary values, as the solution set, $\mathcal{S}$. Therefore, the maximum of $F(p : x_E, x_C)$ over this solution set is the accurate tail probability,

$$\max_{p \in \mathcal{S}} F(p : x_E, x_C). \quad (6)$$

We know that $F(p : x_E, x_C)$ in Equation (4) is a polynomial version of $H(p : x_E, x_C)$ in Equation (2). Therefore, they should be identical for any p . However, we find that they could be different from each other for a large p value due to the precision issue that occurs in $F(p : x_E, x_C)$. The coefficient of p^i is the sum of $g(X_E, X_C, i)$, where $(X_E, X_C) \in \Omega(x_E, x_C)$. In the case that $\left[\sum_{(X_E, X_C) \in \Omega(x_E, x_C)} g(X_E, X_C, i) \right]$ is smaller than the precision level, the computed summation is not the true value but a misleading result caused by the precision problem. We use a balanced randomized study with sample sizes $N_E = N_C = 50$ and $(x_E, x_C) = (16, 6)$, to illustrate this issue by using the polynomial method, in Figure 1. The blue line in the figure is the curve of $H(p : x_E, x_C)$, which is the actual curve. The red line is based on $F(p : x_E, x_C)$. It is easy to see that they are the same when p is small. As the value of p goes up, they no longer overlap. For some cases, $F(p : x_E, x_C)$ could even be negative, or above 1. This is caused by the floating point precision limit in any computer system.

Despite this limitation, the good news is that they are the same when p is small. One reason for the behavior as seen in Figure 1 where $F()$ is different from $H()$ when p is large, is caused by the inaccurate value of the inner sum over g in Equation (4)

$$\left[\sum_{(X_E, X_C) \in \Omega(x_E, x_C)} g^{(X_E, X_C, i)} \right].$$

For example, the actual computed value of this inner sum is 10^{-200} , while the floating point precision limit on a computer system is assumed to be 10^{-10} . Then, the calculated sum could be close to the precision limit of 10^{-10} , instead of the true value of 10^{-200} . Obviously, this is not the correct value it should be. When p is small, p^i is even smaller, and $10^{-10} \times p^i$ is also a very small value, which does not significantly affect the final $F()$ value. When p is large, $10^{-10} \times p^i$ could highly affect the $F()$ value.

For this reason, we consider the piecewise approach by dividing the range of p into K complementary equally-spaced ranges, $((i-1)/K, i/K)$, $i = 1, \dots, K$. For the i -th range of p from $(i-1)/K$ to i/K , it can be shifted to a new range of $(-1/(2K), 1/(2K))$. The original probability density function at the i -th piece

$$\binom{N_C}{X_C} \binom{N_E}{X_E} p^{X_C + X_E} (1-p)^{N_E + N_C - X_E - X_C}, \text{ where } p \in \left(\frac{i-1}{K}, \frac{i}{K}\right)$$

can be rewritten as

$$\binom{N_C}{X_C} \binom{N_E}{X_E} \left(q + \frac{2i-1}{2K}\right)^{X_C + X_E} \left(1 - \frac{2i-1}{2K} - q\right)^{N_E + N_C - X_E - X_C}, \text{ where } q \in \left(\frac{-1}{2K}, \frac{1}{2K}\right).$$

It follows that for any given p from 0 to 1, its probability density function can be shifted to a density function with the parameter range from $-1/(2K)$ to $1/(2K)$.

When K is large enough to obtain an accurate tail probability, its derivative is also accurate and matches the true derivative. The solutions of the derivative, can be obtained within each range. In addition to these solutions, the boundary values of these pieces with a total of $K + 1$ points are the possible values that lead to the maximum of a tail probability. Thus, the global maximum of a tail probability is the maximum of its value over the solution set.

3 Numerical study

We first use an example from a randomized clinical trial to illustrate the application of the proposed polynomial method as compared to the traditional grid search method. This is a double-blind randomized Phase II cancer clinical trial investigating the effectiveness of zoledronic acid for breast cancer patients with high-risk bone metastases [3]. The gold standard treatment is pamidronate, and the primary outcome in this trial is a decrease in serum C-telopeptide (sCTx) after a 12-week treatment when compared to the baseline

measurement. At the end of this study, 61 patients completed the study and had their blood tested to measure sCTx values. Among these patients, $N_C = 29$ patients were from the pamidronate group, and $N_E = 32$ patients were enrolled in the zoledronic acid group. The number of patients that achieved a fall in sCTx were $X_C = 18$ and $X_E = 26$ from the pamidronate group and the zoledronic acid group, respectively. Thus, the estimated response rates are 62.07% in the pamidronate group and 81.25% in the zoledronic acid group. The new treatment, zoledronic acid, has a higher response rate compared to the gold standard.

Figure 2 shows the tail probability curve when comparing the response rate between the zoledronic acid group and the pamidronate group, by using the Barnard's exact unconditional approach as in Equation (3). Multiple local maximum and minimum values exist in this curve, and the accurate p-value is 0.05401857 when using the proposed new approach. The p-value is achieved at $p = 0.06600976$, the red line in the figure. For the grid search method, the search points between 0 and 1 are equally spaced in this numerical study. Figure 3 displays the calculated p-value based on the grid search method with different numbers of grids, starting from 50 to 1000. It can be seen that this traditional p-value eventually approaches the actual value when the number of points is large enough. However, the p-value based on the grid search method could decrease as the number of grids increases in some cases. It is a saw-type shape when the number of points is not very large. The grid search method has the p-value closest to the accurate p-value when the number of grids is over 800. For the proposed method, K can be as small as 3 in this example to guarantee the actual p-value.

We then compare the p-value of each sample point from the sample space, calculated from the proposed polynomial method and the grid search method with sample sizes $(N_E, N_C) = (30, 50), (50, 50), (100, 50),$ and $(100, 100)$. In the grid search method, the number of points of 2,000 are used in finding the maximum of a tail probability. For the proposed polynomial approach, its performance depends on the value of K , the number of pieces in the piecewise step. It is obvious that a large K value is required when many local maximums exist. We study the following values of K , 50, 80, 100, 150, and 200. We use the sample size of $(N_E, N_C) = (100, 50)$ as an example to identify the required K value to guarantee accurate tail probabilities. Tail area is determined by the test statistic, and a total of 1969 unique test statistics are computed from the complete sample space. It should be noted that the sample points with the largest or smallest test statistic are not included in the comparison since the two approaches always have the same p-value. Table 1 shows that the percentage of the proposed polynomial method has at least the same tail probability as the grid search method. It is as expected that this percentage increases as the number of pieces, K , goes up. When K is as large as the total sample size from the study, 150 in this example, the proposed polynomial method can achieve accurate tail probabilities for all sample points in the sample space. The results with $K = 150$ are the same as those with a large K value (e.g., 200). We observe similar results from other sample sizes.

The proposed accurate p-value calculation is always at least as large as the p-value computed by using the traditional grid search. Therefore, the required value of K in the accurate p-value calculation should meet the following two conditions: (i) the proposed p-value calculation is larger than or equal to the p-value based on the grid search; (ii) the proposed p-

value is a constant when K is larger than a threshold. Following these two conditions, we define the minimum required value of K as: (i) the proposed p-value at K is at least as large as that based on the grid search method with sufficiently enough accuracy (e.g., 100, 500 points between 0 and 1); (ii) the proposed p-values at five consecutive $K + a \times \delta$ values are the same, $a = 0, 1, 2, 3, 4$, where δ is a pre-specified value, e.g., 5. Then, this identified K is considered as the minimum required value in the accurate p-value calculation.

We compare the proposed exact p-value calculation with the existing grid search method implemented by the function *barnard.test* from an R package *Barnard*. We use the sample size of $(N_E, N_C) = (50, 30)$ as an example to compare these two methods. The same test statistic as in Equation (1) is used in both methods, and the hypothesis is one-sided. The size of the sample space is $51 \times 31 = 1581$ with 785 samples having positive test statistics. For a sample with a non-positive test statistic, the *barnard.test* function switches the order of the two treatments to make the test statistic non-negative. For these samples, we can not compare the two methods directly since their rejection regions are different from each other. We use these 785 samples to compare the two methods, and the results indicate that the proposed accurate p-value is always at least as large as that by using the *barnard.test*, where 1000 equal-spaced points between 0 and 1 are used in the existing approach. Although the difference is small, the proposed method identifies the p-value precisely.

4 Discussion

In this article, we propose the polynomial method when computing exact unconditional p-values based on the Barnard's approach, in comparing two proportions in a clinical trial with a binary endpoint as the primary outcome. The traditional grid search method computes an approximate p-value, and a significant number of points are generally needed for evaluation. The proposed polynomial method is accurate with the true maximum being calculated. We use the piecewise strategy to overcome the precision limitation from computers. We conduct extensive numerical studies to find out the required number of K for the proposed polynomial method. The proposed accurate p-value calculation is always at least as large as the p-value computed by using the traditional grid search (p_g). Therefore, the lowest required value of K is the one such that the proposed p-value calculation is larger than or equal to p_g . Let K_0 be this lowest required value. The next step is to find the minimum K such that the proposed p-value is a constant when K is beyond that minimum threshold, which can be identified as the minimum of K such that the proposed p-values at five consecutive $K + a \times \delta$ values are identical, where $K > K_0$, $a = 0, 1, 2, 3, 4$, and δ is the incremental value, e.g., 10, 20. Then, this minimum K is used in the accurate p-value calculation.

For a study with multiple stages (Simon's two-stage design [10]) or a study with multiple binary primary outcomes (response rate and toxicity rate), the tail probability is a function of multiple nuisance parameters when the unconditional approach is used in p-value calculation. An alternative to the exact unconditional approach, is the exact conditional approach that considers both marginal totals fixed. In such, the nuisance parameters in the tail probability can be eliminated. Mehta and Hilton [7] found that the exact conditional approach may be slightly more powerful than the exact unconditional approach in some

configurations to test the equality of three independent proportions under a two-sided hypothesis testing problem. These configurations were the ones that the conditional approach had a larger size as compared to the unconditional approach, e.g., sample size of 150 per group and the common response rate of 50%. Multiple factors would affect the performance of the exact approaches with regard to power. The actual size of an exact approach is partially responsible for that. The closer the size to the nominal level, the possibly higher the power could be attained. Another factor is the considered test statistic to order the sample space. In general, a test statistic with the finest possible partition is often preferable.

When multiple nuisance parameters occur in the tail probability, it would be computationally intensive to use the traditional grid method, and computer memory could be another technical issue even for a study with a medium sample size. The proposed polynomial method could be potentially used to reduce the computational intensity and obtain accurate tail probabilities.

References

- [1]. Barnard GA. A new test for 2×2 tables. *Nature*, 156:177, 1945.
- [2]. Basu Debabrata. On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 72(358):355–366, 1977.
- [3]. Jacobs C, Kuchuk I, Bouganim N, Smith S, Mazzarello S, Vandermeer L, Dranitsaris G, Dent S, Gertler S, Verma S, Song X, Simos S, Cella D, and Clemons M. A randomized, double-blind, phase II, exploratory trial evaluating the palliative benefit of either continuing pamidronate or switching to zoledronic acid in patients with high-risk bone metastases from breast cancer. 155(1):77–84, 2016. [PubMed: 26643085]
- [4]. Lloyd Chris J.. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics*, 64(3):716–723, 2008. [PubMed: 18047530]
- [5]. Lloyd Chris J.. Exact p-values for discrete models obtained by estimation and maximization. *Australian and New Zealand Journal of Statistics*, 50(4):329–345, 2008.
- [6]. Lloyd Chris J. and Moldovan Max V.. A more powerful exact test of noninferiority from binary matched-pairs data. *Statistics in Medicine*, 27(18):3540–3549, 8 2008. [PubMed: 18314932]
- [7]. Mehta Cyrus R. and Hilton Joan F.. Exact Power of Conditional and Unconditional Tests: Going beyond the 2×2 Contingency Table. *The American Statistician*, 47(2):91–98, 5 1993.
- [8]. Joachim Röhm. Problems with existing procedures to calculate exact unconditional P-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal*, 47(1):37–47, 2 2005. [PubMed: 16395995]
- [9]. Joachim Röhm and Ulrich Mansmann. Unconditional Non-Asymptotic One-Sided Tests for Independent Binomial Proportions When the Interest Lies in Showing Non-Inferiority and/or Superiority. *Biometrical Journal*, 41(2):149–170, 5 1999.
- [10]. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled clinical trials*, 10(1):1–10, 3 1989. [PubMed: 2702835]
- [11]. Storer Barry E. and Kim Choongrak. Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association*, 85(409):146–155, 1990.

Figure 1:
The actual tail probability curve from a binomial distribution (blue), and the polynomial version of the curve (red) when sample sizes $N_E = N_C = 50$ and $(x_E, x_C) = (16, 6)$.

Figure 2:

Tail probability curve comparing the response rate from the pamidronate group and the zoledronic acid group of breast cancer patients by Jacobs et al. [3]. The red line represents the value of p where the actual p-value is obtained.

Figure 3:
P-value based on the grid search algorithm as a function of the number of grids. The flat red line is the accurate p-value when using the proposed polynomial method.

Table 1:

Proportion of the sample points in the sample space for which the proposed polynomial method has at least the same tail probability as the grid method when the sample size $(N_E, N_C) = (100, 50)$

$K = 50$	$K = 80$	$K = 100$	$K = 150$	$K = 200$
92.9%	98.2%	99.2%	100%	100%