Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Research Article
Volume 12 Issue 2 - September 2025
DOI: 10.19080/BBOAJ.2025.12.555835

Biostat Biom Open Access J

A Perturbation-Based Method to Boost Exact Tests of Proportions in the Randomized Clinical Trial Setting

Alan D Hutson*

Roswell Park Comprehensive Cancer Center, Department of Biostatistics and Bioinformatics, Elm and Carlton Streets, Buffalo, NY 14623

Submission: September 4, 2024; **Published**: September 30,2025

***Corresponding author:** Alan D Hutson, Roswell Park Comprehensive Cancer Center, Department of Biostatistics and Bioinformatics, Elm and Carlton Streets, Buffalo, NY 14623.

Sample Summary

Small-sample comparisons of binomial proportions often rely on Fisher's exact test, which can be overly conservative and underpowered. We introduce a perturbation testing framework that generates exact tests by perturbing outcomes, with two variants: with replacement (WR) and without replacement (WOR). The WR test maintains nominal Type I error while achieving higher power than Fisher's, Barnard's, and Boschloo's tests, and it extends naturally to multiple groups, outperforming the Freeman-Halton test. Implementable in just a few lines of R code, the WR perturbation test offers a fast, practical, and more powerful alternative to traditional exact methods.

Abstract

Small sample comparisons of binomial proportions are common in early phase clinical trials, laboratory studies, and rare event investigations. The standard test, Fisher's exact, often lacks power due to conservative Type I error control. Alternatives like Barnard's and Boschloo's tests offer less conservative error control and greater power. We present a unified perturbation testing framework for generating exact tests by perturbing the discrete outcome variable, yielding two variants: with replacement (WR) and without replacement (WOR). For the null hypothesis H0: π1=π2, WR controls Type I error at the nominal level while providing uniformly higher power than existing methods. The approach extends naturally to g independent groups, where WR outperforms the Freeman-Halton exact test. Easily implemented in a few lines of R code and executing in seconds, the WR perturbation test offers a practical, more powerful alternative to traditional exact tests.

KeyWords: Perturbation Test; Exact Inference; Small Sample Power; Randomization Test; Monte-Carlo Resampling

Introduction

Consider a randomized trial where subjects are assigned to one of two treatments, T_1 or T_2, using a standard parallel design. The primary outcome is binary, coded as 0 or 1. Suppose there are j = 1, 2, . . , n subjects, where subject j would have response x_{1j} under treatment T_1 and response x_{2j} under treatment T_2. The individual treatment effect for subject j is defined as x_{1j}-x_{2j}. However, in practice, only one of these two potential outcomes is observed for each subject due to random assignment.

Following the principles of randomization inference as outlined by Rosenbaum [1], and focusing on the binary outcome setting, assume that out of $n=n_1+n_2$ total subjects, n_1 are randomly assigned to receive treatment T_1 and n_2 to receive treatment T_2.

There is $m=\binom{n}{n_1}$ possible treatment assignments, each equally likely with probability $\frac{1}{m}$.

Let I_j=1 if subject j is assigned to treatment T_1 and I_j=0 if assigned to T_2. In randomization theory, the only source of randomness is this treatment assignment vector $I= (I_1, I_2, \ldots, I_n)$, where typically $P(I_j = 1) = \theta$ for $j = 1, 2, \ldots, n$. The observed outcome for subject j is then given by the random variable $Y_j = I_j x_{1j} + (1 - I_j) x_{2j}$. This formulation enables the construction of the randomization distribution by enumerating all m possible assignments.

In the context of comparing binary outcomes between the two groups, we are interested in testing whether

$P(Y_j = 1 \mid I_j = 1) = \pi_1$ differs from $P(Y_j = 1 \mid I_j = 0) = \pi_2, j = 1, 2, \ldots, n.$

The corresponding estimators are:

$$\hat{\pi}_1 = \frac{\sum_{j=1}^{n} I_j Y_j}{n_1}, \hat{\pi}_2 = \frac{\sum_{j=1}^{n} (1 - I_j) y_j}{n_2}$$

In a one-sided testing scenario, we may test H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$ (or similarly H_1: $\pi_1<\pi_2$). Define the observed test statistic as $S = \hat{\pi}_1 - \hat{\pi}_2$. For each of the m permutations, compute the corresponding test statistic $S_k = \hat{\pi}_{1k} - \hat{\pi}_{2k}$, for k = 1, 2, . . ., m. The exact p-value is then given by:

$$p = \frac{\sum_{k=1}^{m} I_{(s_k \geq s)}}{m} \text{ for } H_1: \pi_1 > \pi_2,$$

$$p = \frac{\sum_{k=1}^{m} I_{(s_k \leq s)}}{m} \text{ for } H_1: \pi_1 < \pi_2$$

where $I_{(\cdot)}$ is the indicator function.

For a two-sided alternative H_1: $\pi_1 \neq \pi_2$, a common choice for the test statistic is:

$$S^2 = \sum_{l=1}^{2} (\hat{\pi}_l - \hat{\pi})^2,$$

Where $\hat{\pi} = \sum_{j=1}^{n} Y_j / n$ The p-value is then:

$$p = \frac{\sum_{k=1}^{m} I_{(S_k^2 \geq S^2)}}{m},$$

Where $S_k^2 = \sum_{l=1}^{2} (\hat{\pi}_{lk} - \hat{\pi})^2$ for k=1, 2,,,m.

Since enumerating all possible treatment assignments is often computationally infeasible for large sample sizes, Monte Carlo approximations offer a practical alternative. Specifically, by randomly selecting B treatment assignments, the p-value can be estimated as:

$$p = \frac{1}{B} \sum_{k=1}^{B} I_{(S_k \geq S)},$$

for the alternative hypothesis H_1: $\pi_1>\pi_2$, with analogous expressions for other alternatives. This method enables feasible inference under the randomization framework.

The randomization test corresponds to sampling without replacement from the set of all possible assignments. In addition, we will explore a bootstrap-like approach that samples with replacement from the observed outcomes Y. The Monte Carlo approximation to the p-value in this case follows the same logic as that used in the randomization test.

Practically, when comparing two treatments with a binary outcome, the observed data can be summarized in a 2×2 contingency table. In this setting, the p-values obtained through a randomization test is equivalent to those derived from Fisher's exact test. Fisher's exact test, introduced by R. A. Fisher in the 1930s [2], was specifically designed to analyze small 2×2 contingency tables. The test is rooted in conditional probability, famously illustrated by the" Lady Tasting Tea" experiment. The test is based on calculating the exact probability of observing a table as extreme as the one obtained, assuming the null hypothesis and conditioning on fixed row and column margins. This approach leads to the hypergeometric distribution as the basis for inference. Fisher's ex- act test is widely used in biomedical and clinical research, especially where small sample sizes make large-sample approximations unreliable. There are several test statistics that are one-to-one that can be utilized using this approach that generate identical p-values [3].

However, Fisher's test has some important limitations. It assumes fixed marginal totals, a condition that may not reflect many real-world randomization scenarios, such as those described previously where only the treatment assignment is randomized. Furthermore, while computationally feasible for small sample sizes, the test becomes increasingly intensive as the table size grows. Nevertheless, Monte Carlo approximations to Fisher's test can be easily implemented to overcome these limitations in practice.

Fisher's exact test is also known to be the uniformly most powerful (UMP) test among all exact tests in the 2×2 setting. However, this optimality holds only under specific conditions, particularly when the Type I error rate corresponds to attainable probabilities under the hypergeometric distribution, which can be restrictive in practice.

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

As an alternative, Barnard's test, proposed by George A. Barnard in 1945 [4], offers a more flexible framework. Unlike Fisher's test, Barnard's test does not condition on the marginal totals, making it an unconditional test. It evaluates all possible 2×2 tables and typically uses a test statistic such as the difference in sample proportions to assess significance. By allowing the margins to vary and optimizing the rejection region, Barnard's test often achieves greater power than Fisher's test, particularly because it can more accurately control the Type I error rate at common levels such as 0.05 or 0.10.

Historically, the computational burden associated with Barnard's test limited its practical use, leading to a preference for Fisher's exact test despite its conservatism. However, with advances in modern computing, Barnard's test and related methods like Boschloo's test [5], which improves on Barnard's test by modifying the rejection criteria, have gained renewed interest. These methods are now recognized as powerful alternatives for small-sample inference in 2×2 tables, especially when the fixed-margin assumption of Fisher's test is not appropriate.

Recent work by Korn and Freidlin [6] underscores the advantages of unconditional exact tests, particularly Boschloo's test, which preserve the desired Type I error rate while generally offering greater power and, consequently, requiring smaller sample sizes. Although some statisticians have argued that conditional analyses, such as Fisher's exact test, are more appropriate in randomized trial settings, Korn and Freidlin find these arguments either irrelevant or unconvincing. Their conclusions support a broader adoption of unconditional methods in clinical trial design, especially given the practical value of minimizing trial size. Moreover, they propose that incorporating prespecified null and alternative response rates into the test framework could further improve the power of unconditional approaches. This perspective aligns with earlier work by Mehrotra et al. [7], who conducted a comprehensive comparison of Fisher's exact test and various un conditional exact procedures. Their study concluded that "Boschloo's test, in which the *p*-value from Fisher's test is used as the test statistic in an exact unconditional test, is uniformly more powerful than Fisher's test, and is also recommended." See also Lin and Yang [8] and Andres and Mato [9].

Testing binary endpoints across g treatment groups

Now suppose we have g treatment groups, each randomly assigned to subjects, with a binary outcome of interest. We aim to test the hypothesis

H_0: $\pi_1=\pi_2=\cdots=\pi_g$ versus H_1: not all π_i are equal, $i=1, 2, \ldots, g$.

Let $j=1, 2, \ldots, n$ index the subjects, where subject j would have a binary response x_{ij} under treatment T_i, for $i= 1, 2, \ldots, g$. Assume there are a total of $n = n_1+n_2+\cdots+n_g$ subjects, with n_i subjects randomly assigned to treatment group T_i. The number of possible treatment assignments is given by the multinomial coefficient

$$m = \frac{n!}{n_1!n_2!\ldots n_g!},$$

and each assignment is equally likely with probability $\frac{1}{m}$.

The observed outcome for subject j is then represented by the random variable

$$Y_j = I_{1j}x_{1j} + I_{2j} + x_{2j} + \ldots + (1-\sum_{k=1}^{g-1} I_{jk})x_{gj} \quad (1.1)$$

where I_{jk} is the random indicator variable that subject j received treatment T_k. To test the null hypothesis, we may use the test statistic

$$S^2 = \sum_{l=1}^{g} (\hat{\pi}_l - \hat{\pi})^2 \quad (1.2)$$

Where $\hat{\pi}_i = \sum_{j=1}^{n} I_{ij}Y_j / n_i$ is the treatment level sample proportion, $i=1, 2,..,g$, $I_g = (1-\sum_{k=1}^{g-1} I_k)$ and $\hat{\pi} = \sum_{j=1}^{n} Y_j / n$ is the overall sample proportion.

The exact permutation *p*-value is then computed as:

$$p = \frac{\sum_{k=1}^{m} I_{(S_k^2 \geq S^2)}}{m} \quad (1.3)$$

Where $S_k^2 = \sum_{l=1}^{g} (\hat{\pi}_{lk} - \hat{\pi})^2$ is the value of the test statistic under the k-th permutation of treatment assignments.

As before, since enumeration of all possible treatment assignments may be computationally prohibitive for large n, we employ a Monte Carlo approximation. By randomly selecting B treatment assignments, the *p*-value can be estimated as

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

$$p = \frac{1}{B}\sum_{k=1}^{B} I_{(S_k \geq S)} \qquad (1.4)$$

for the alternative hypothesis H_1: $\pi_1 > \pi_2$, with analogous expressions for other alternatives.

In parallel, we also consider a bootstrap-like approach, sampling with replacement from the observed outcomes Y. The Monte Carlo approximation to the *p*-value in this setting follows the same structure as that of the randomization test.

Similar to the two-group setting, the null hypothesis of equal treatment assignment probabilities across *g* groups, H_0: $\pi_1=\pi_2= \cdots =\pi_g$, can be represented using a *2×g* contingency table. This hypothesis may be tested using the exact chi-square test of independence proposed by Freeman and Halton [10], which extends Fisher's exact test to higher dimensional tables. Importantly, the exact p-value obtained from the Freeman-Halton test is identical to that of the corresponding randomization test [11], mirroring the equivalence observed in the 2×2 case.

The data described in this subsection may also be represented in a *K×g* contingency table. Under this representation, the hypothesis stated above can be equivalently tested using the Freeman-Halton extension of Fisher's exact test, which provides an exact chi-square test of independence for multirow by multicolumn tables [12].

In this note we extend these ideas by developing a new approach to resampling based on a Monte Carlo resampling approach. Section 2 introduces our perturbation randomization framework, detailing both the with replacement (WR) and without replacement (WOR) variants and supplying straightforward R code for immediate use. A toy example contrasts the new procedures with Fisher's exact, Barnard's, and Boschloo's tests. Section 3 reports an extensive simulation study that examines the statistical power of each test across a broad spectrum of sample sizes and proportion configurations. Section 4 then analyses two applied data sets, demonstrating how the perturbation tests can alter scientific conclusions relative to traditional exact methods. Finally, Section 5 summarizes practical recommendations.

Perturbation-Based Randomization Testing

The general strategy of our new test involves perturbing the observed data. In the case of Bernoulli outcomes, this means perturbing the binary response (0 or 1). We first describe the method for Bernoulli outcomes in a two-group comparison setting, which naturally generalizes to *g*-group comparisons for both response types. The perturbations are used to generate a less coarse randomization distribution, while utilizing the original observed test statistic $S = \hat{\pi}_1 - \hat{\pi}_2$ for testing H_0: $\pi_1=\pi_2$ versus the one-sided alternative H_1: $\pi_1>\pi_2$. For the two-sided alternative H_1: $\pi_1\neq\pi_2$, we instead use the squared for each subject *j*, we define the observed outcome as the random variable where $I_j=1$ if subject *j* is assigned to treatment T_1, and $I_j=0$ if assigned to treatment T_2. Under this setup, x_{1j} and x_{2j} represent the potential outcomes for subject *j* under treatments T_1 and T_2, respectively.

To introduce variability for resampling, we generate a perturbed response vector with elements

$$Y_j^{'} = I_j x_{1j} + (1 - I_j)x_{2j} + e_j \text{ , } j=1, 2,\ldots n,$$

where the perturbation term $e_j \sim N(0, h^2)$, and *h* is a user specified bandwidth. Unlike traditional smoothing approaches such as kernel density estimation, we fix the bandwidth at $h=1/10000$, independent of the sample size.

Two resampling strategies are considered: sampling with replacement and sampling without replacement from the perturbed vector *Y'*. For each resampled dataset, we compute the one-sided test statistic using the estimators

$$\hat{\pi}_1^{'} = \frac{\sum_{j=1}^{n} I_j y_j^{'*}}{n_1} \text{ , } \hat{\pi}_2^{'} = \frac{\sum_{j=1}^{n} (1-I_j) y_j^{'*}}{n_2} \qquad (2.1)$$

where denotes the resampled value for subject *j* from the perturbed response vector *Y'*. For the sampling with replacement approach, we must also use the resampled group mean $\hat{\pi}^{'} = \frac{\sum_{j=1}^{n} y_j^{'*}}{n}$ when applying the two-sided alternative in two group comparisons, as well as the pooled mean for comparisons involving *g* groups with binary or multinomial outcomes.

It follows immediately that

$$E(\hat{\pi}_1^{'}) = E(\hat{\pi}_1),\ E(\hat{\pi}_2^{'}) = E(\hat{\pi}_2)$$

and

$$Var(\hat{\pi}_1^{'}) = Var(\hat{\pi}_1) + h^2,\ Var(\hat{\pi}_2^{'}) = Var(\hat{\pi}_2) + h^2$$

The probability density function (PDF) of $\hat{\pi}^{'}_1$ is the following normal mixture:

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

$$f_{\hat{\pi}_1'}(y)=\sum_{k=0}^{n_1}\binom{n_1}{k}\pi_1^k(1-\pi_1)^{n_1-k}.\frac{1}{h/\sqrt{n_1}}\phi\left(\frac{y-k/n_1}{h/\sqrt{n_1}}\right)$$

where ϕ is the standard normal PDF. The corresponding cumulative distribution function (CDF) is

$$f_{\hat{\pi}_1'}(y)=\sum_{k=0}^{n_1}\binom{n_1}{k}\pi_1^k(1-\pi_1)^{n_1-k}.\Phi\left(\frac{y-k/n_1}{h/\sqrt{n_1}}\right),$$

where Φ is the standard normal CDF.

Similarly, the PDF of $\hat{\pi}_2'$ is

$$f_{\hat{\pi}_2'}(y)=\sum_{k=0}^{n_2}\binom{n_2}{k}\pi_2^k(1-\pi_2)^{n_2-k}.\frac{1}{h/\sqrt{n_2}}\phi\left(\frac{y-k/n_2}{h/\sqrt{n_2}}\right)$$

and the corresponding CDF is

$$f_{\hat{\pi}_2'}(y)=\sum_{k=0}^{n_2}\binom{n_2}{k}\pi_2^k(1-\pi_2)^{n_2-k}.\Phi\left(\frac{y-k/n_2}{h/\sqrt{n_2}}\right).$$

Thus, $\hat{\pi}_1'$ and $\hat{\pi}_2'$ can be interpreted as kernel-smoothed versions of the observed sample proportions $\hat{\pi}_1$ and $\hat{\pi}_2$.

Each perturbation and resampling iteration yield a value used to construct a Monte Carlo approximation to the null hypothesis distribution. The perturbation-based *p*-value is calculated in accordance with the exact *p*-value framework described in Section 1, by comparing the observed test statistic to the distribution of values generated under the null hypothesis. This framework applies equally whether resampling is performed with or without replacement. From the perspective of Monte Carlo inference, this approach offers a practical and conceptually intuitive approximation to the null distribution.

The estimators of the test statistics based on the perturbed responses retain the same expectation as the observed test statistics defined in (1.1) under the null hypothesis H_0: $\pi_1=\pi_2$. The primary motivation for this perturbation-based strategy, particularly in the context of randomized clinical trials, is to achieve Type I error control that more accurately reflects the nominal significance level α. This approach often yields greater statistical power compared to traditional methods such as Fisher's exact test or the Freeman-Halton extension, especially in scenarios involving small to moderate sample sizes or sparse contingency tables. However, as the sample size increases or the number of treatment arms grows, the advantages of the perturbation-based method tend to diminish, and classical methods generally attain Type I error rates that align with the desired α-level. The extension to the *g*-group setting follows the same strategy as the two-group case, utilizing the test statistic defined in (1.2).

Toy Examples

To illustrate the differences in p-values obtained using Fisher's exact test, Barnard's test, Boschloo's test, and the proposed perturbation method (under both sampling with and without replacement), we consider a simple toy example for testing the hypothesis H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$. The observed and perturbed binary data for one Monte Carlo realization, based on a total sample size of n=10, are presented in (Table 1).

Table 1: Sample values for y^1, y^2 for T^1 and T^2 and their perturbed versions y_1', y_2' Perturbations were generated by adding i.i.d. N (0, h) noise with h=1/10000.

y^1	y^2	y_1'	y_2'
1	0	1.000096	0.0000527
0	1	-0.000017	0.9999023
0	0	-0.000163	0.0000801
0	1	-6.258×10^{-6}	1.0001902
0	0	7.9194×10^{-6}	0.0000652
0	1	-0.000043	0.9997861
1	1	1.000032	1.0001201
0	0	-0:000181	0.0000953
0	0	0.0000504	0.0000715
0	1	0.0000869	1.0001335
$\hat{\pi}_1$ =0.2	$\hat{\pi}_2$ =0.5	$\hat{\pi}_1'$ =0.1999862	$\hat{\pi}_2'$ =0.5000497

The perturbed values y_1' and y_2' corresponding to treatments T_1 and T_2, respectively, were obtained by adding independent noise to the original binary responses:

$$Y_j'=I_jx_{1j}+(1-I_j)x_{2j}+e_j\ j=1,2,..n,$$

where $e_j\sim N(0, h)$ and the bandwidth is fixed at h =1/10000.

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

Here, I_j=1 if subject j is assigned to treatment T_1, and I_j=0 if assigned to treatment T_2.

As demonstrated, the resulting perturbed estimators $\hat{\pi}'_1$ and $\hat{\pi}'_2$ closely match their original (unperturbed) counterparts. This confirms that the perturbation does not meaningfully alter point estimation. Both sets of estimators share the same expectation, and the variance introduced by the added noise is negligible.

To estimate the *p*-value under both sampling with and without replacement, we performed 10,000,000 Monte Carlo resamples. This large number of replicates effectively approximates the full permutation distribution of the test statistic and is computationally efficient in R. Below is the R code used to generate the *p*-values for testing H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$:

```
h <- 1/10000
y0 <- c (1, 0, 0, 0, 0, 0, 1, 0, 0, 0)
y1 <- c (0, 1, 0, 1, 0, 1, 1, 0, 0, 1)
# Observed difference in sample proportions
Obs_diff <- mean(y1)-mean(y0)
# Combine groups
combined <- c (y0, y1)
n <- length(y0) # assumes equal group sizes
# Without replacement: perturb and resample
perm_diffs_wor <- replicate (1e7, {fuzzed <- combined+rnorm(length(combined), mean=0, sd=1) *h shuffled <- sample(fuzzed) mean(shuffled[(n+1) :(2 *n)])-mean (shuffled [1: n])})
pvalue_wor <- mean (perm_diffs_wor >=obs_diff)
# With replacement: perturb and resample with replacement
perm_diffs_wr <- replicate (1e7, {fuzzed <- combined+rnorm(length(combined), mean=0, sd=1) *h shuffled <- sample (fuzzed, replace=TRUE) mean(shuffled[(n+1) :(2 *n)])-mean (shuffled [1: n])})
pvalue_wr <- mean (perm_diffs_wr >=obs_diff)
```

To compare the behavior of different exact and approximate methods for testing H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$, we computed *p*-values using Barnard's test, Boschloo's test, Fisher's exact test, and our proposed perturbation method under both with-replacement (WR) and without-replacement (WOR) resampling schemes. The observed *p*-values are summarized below: (Table 2.1)

These results demonstrate that the proposed perturbation approach, particularly with sampling with replacement, can yield *p*-values aligning more closely with Barnard's and Boschloo's methods. Notably, the WR method produced the smallest *p*-value, suggesting increased sensitivity to detect differences under small-sample conditions. In the following section, we further evaluate the perturbation method through a comprehensive simulation study, examining its power and Type I error performance in the two-group, *g*-group, and multinomial testing settings.

To evaluate the reproducibility of the estimated *p*-values across multiple executions of the same program, we ran the procedure ten times using the same data. The results for the two perturbation strategies, without replacement (WOR) and with replacement (WR), are shown in the Table 2. As can be observed, the estimated *p*-values are highly consistent across all runs. Importantly, no decision regarding the null hypothesis H_0: $\pi_1=\pi_2$ versus the alternative H_1: $\pi_1>\pi_2$ would change under either α=0.05 or α=0.10, both commonly used significance levels in clinical trials (Table 2).

Table 2.1

Method	Barnard	Boschloo	Fisher	Perturbation (WOR)	Perturbation (WR)
p-value	0.132	0.099	0.175	0.102	0.084

Table 2: Estimated *p*-values from 10 repeated runs under two perturbation strategies.

Perturbation (WOR)	Perturbation (WR)
0.1017159	0.0843962
0.1019481	0.0844183
0.1015814	0.0844848
0.1018683	0.0842959
0.1016467	0.0843208
0.1018438	0.0843789
0.1016919	0.0843907
0.1018123	0.0843267
0.1018963	0.0844471
0.1019533	0.0841887

Simulation Study

In this section, we compare the new WOR and WR approaches in three settings: the two-group binary outcome setting and *g*-group binary outcome setting.

Testing binomial endpoints across two treatment groups

In this simulation study, we evaluated Type I error control and power for testing H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$ using Barnard's test, Boschloo's test, Fisher's exact test, and the newly proposed WOR

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

and WR tests. Values for π_1 ranged from 0.05 to 0.95 in increments of 0.10. For each value of π_1, corresponding values of π_2 ranged from π_1 to 0.95, also in increments of 0.10, for sample sizes n=10, 20, 30 per group. The bandwidth parameter h was set to 1/10000. The desired Type I error was set to α=0.05.

For the WOR and WR perturbation tests, 1,000 Monte Carlo resamples were used per scenario to estimate the p-value, and each scenario was simulated 1,000 times. Results for specific combinations of π_1 and π_2 are reported in (Tables 3-5). Full power curves for each test are shown in (Figures 1-3).

Table 3: Power values for each test method across different (π_0, π_1) combinations, n=10.

π_1	π_2	Barnard	Boschloo	Fisher	WOR	WR
0.05	0.05	0.01	0	0	0.005	0.01
0.05	0.45	0.71	0.596	0.516	0.633	0.718
0.05	0.55	0.826	0.8	0.707	0.805	0.846
0.15	0.15	0.042	0.02	0.016	0.03	0.043
0.15	0.65	0.745	0.767	0.632	0.754	0.811
0.15	0.75	0.849	0.882	0.806	0.871	0.906
0.25	0.25	0.04	0.033	0.016	0.035	0.048
0.25	0.75	0.698	0.724	0.619	0.717	0.782
0.25	0.85	0.862	0.879	0.813	0.873	0.91
0.35	0.35	0.048	0.053	0.021	0.052	0.059
0.35	0.85	0.719	0.758	0.628	0.735	0.793
0.35	0.95	0.914	0.915	0.847	0.912	0.936
0.45	0.45	0.043	0.041	0.018	0.046	0.059
0.45	0.95	0.821	0.798	0.713	0.803	0.841
0.55	0.55	0.047	0.04	0.018	0.039	0.058
0.55	0.95	0.701	0.616	0.531	0.634	0.715
0.65	0.65	0.042	0.043	0.024	0.042	0.049
0.65	0.95	0.516	0.369	0.305	0.416	0.518
0.75	0.75	0.046	0.037	0.023	0.039	0.049
0.75	0.95	0.313	0.158	0.139	0.204	0.316
0.85	0.85	0.034	0.015	0.013	0.023	0.034
0.85	0.95	0.116	0.041	0.04	0.066	0.118
0.95	0.95	0.008	0.001	0.001	0.004	0.008

Table 4: Power values for each test method across different (π_0, π_1) combinations, n=20.

π_1	π_2	Barnard	Boschloo	Fisher	WOR	WR
0.05	0.05	0.020	0.005	0.001	0.006	0.020
0.05	0.35	0.770	0.763	0.689	0.782	0.821
0.05	0.45	0.929	0.926	0.886	0.926	0.938
0.15	0.15	0.044	0.035	0.023	0.043	0.066
0.15	0.55	0.847	0.847	0.803	0.859	0.876
0.15	0.65	0.957	0.957	0.928	0.956	0.967
0.25	0.25	0.050	0.050	0.032	0.046	0.053
0.25	0.65	0.806	0.806	0.715	0.807	0.823
0.35	0.35	0.035	0.035	0.024	0.040	0.046
0.35	0.75	0.812	0.812	0.731	0.817	0.828
0.35	0.85	0.948	0.948	0.930	0.950	0.955
0.45	0.45	0.045	0.045	0.025	0.050	0.051
0.45	0.85	0.855	0.855	0.805	0.858	0.877
0.55	0.55	0.042	0.042	0.022	0.044	0.046
0.55	0.95	0.919	0.915	0.878	0.919	0.938
0.65	0.65	0.042	0.042	0.029	0.045	0.052
0.65	0.95	0.768	0.760	0.676	0.790	0.822
0.75	0.75	0.037	0.037	0.025	0.040	0.048
0.75	0.95	0.548	0.489	0.407	0.520	0.603
0.85	0.85	0.034	0.026	0.014	0.036	0.054
0.85	0.95	0.239	0.153	0.083	0.189	0.273
0.95	0.95	0.027	0.007	0.000	0.011	0.025

Table 5: Power values for each test method across different (π_0, π_1) combinations, n=30.

π_1	π_2	Barnard	Boschloo	Fisher	WOR	WR
0.05	0.05	0.037	0.017	0.003	0.021	0.039
0.05	0.25	0.689	0.683	0.569	0.689	0.717
0.05	0.35	0.928	0.928	0.863	0.931	0.936
0.15	0.15	0.039	0.038	0.017	0.041	0.051
0.15	0.45	0.823	0.823	0.739	0.809	0.828
0.15	0.55	0.943	0.943	0.918	0.938	0.950
0.25	0.25	0.041	0.041	0.022	0.039	0.044
0.25	0.55	0.780	0.780	0.699	0.778	0.781
0.25	0.65	0.936	0.936	0.900	0.935	0.936
0.35	0.35	0.054	0.054	0.031	0.050	0.057
0.35	0.65	0.753	0.753	0.647	0.752	0.758
0.35	0.75	0.936	0.936	0.898	0.935	0.939
0.45	0.45	0.038	0.038	0.021	0.039	0.038
0.45	0.75	0.753	0.753	0.674	0.750	0.757
0.45	0.85	0.958	0.958	0.924	0.956	0.958
0.55	0.55	0.046	0.046	0.028	0.043	0.050
0.55	0.85	0.828	0.828	0.746	0.818	0.835
0.55	0.95	0.992	0.992	0.986	0.991	0.994
0.65	0.65	0.042	0.042	0.025	0.041	0.045
0.65	0.85	0.535	0.535	0.432	0.521	0.539
0.65	0.95	0.911	0.911	0.864	0.915	0.919
0.75	0.75	0.043	0.043	0.028	0.041	0.048
0.75	0.95	0.684	0.679	0.554	0.690	0.733
0.85	0.85	0.033	0.033	0.016	0.038	0.044
0.85	0.95	0.310	0.272	0.159	0.291	0.350
0.95	0.95	0.045	0.020	0.002	0.022	0.048

As evident from the results, the WR perturbation test consistently achieves the highest power across all scenarios, with its relative advantage decreasing as sample size increases. The efficiency gains of the WR perturbation test range from approximately 1% to 10%, depending on the parameter configuration. Notably, Barnard's test tends to outperform Boschloo's test at lower values of π_1, while Boschloo's test shows superior performance at higher values of π_1.

Although the performance gain of the WR perturbation test over Barnard's and Boschloo's tests may be modest in some

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

cases, it is consistently superior, and in certain settings, the gain is substantial. This improvement can be especially important in clinical trials, such as cancer immunotherapy studies, where even a small reduction in required sample size can lead to significant cost savings. Moreover, the implementation of the WOR and WR perturbation methods is straightforward, as demonstrated by the R code provided in the previous section.

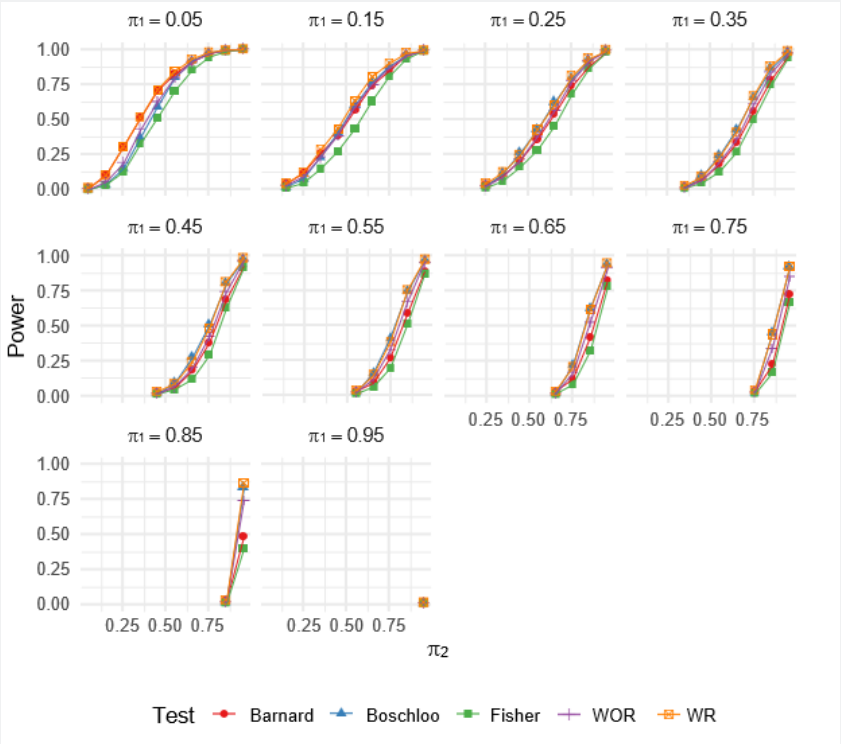

Figure 1: Power curves for various test methods across different (π0, π1) values, n=10.

Testing binomial endpoints across g treatment groups

To evaluate the performance of the proposed methods, we conducted a simulation study using 1,000 Monte Carlo replications per scenario and 1,000 resamples to estimate *p*-values via both the WR and WOR perturbation tests and an approximation to the Pearson exact chi-square test. Type I error and power was assessed under the global null hypothesis for g=3 groups, H_0: $\pi_1=\pi_2=\pi_3$ versus H_1: not all π_i are equal.

Power was evaluated under various alternatives involving increasing divergence in category probabilities. Across all sample sizes (n=10, 20, 30), both the WOR and WR perturbation tests maintained appropriate control of the Type I error near the nominal level of α=0.05, as shown in (Tables 6-8). The WR method tended to be slightly conservative for n=10. In contrast, the Pearson exact chi-squared test exhibited highly conservative behavior at this smallest sample size, with rejection rates under the null as low as 0.002, although performance improved as sample size increased.

Table 6: Simulation results for n=10 for three group comparison.

π_1	π_2	π_3	WOR	WR	X^2
0.1	0.1	0.1	0.021	0.028	0.002
0.1	0.1	0.5	0.527	0.580	0.456
0.1	0.1	0.9	0.991	0.991	0.991
0.1	0.3	0.3	0.167	0.177	0.105
0.1	0.3	0.7	0.733	0.751	0.701
0.1	0.5	0.5	0.461	0.495	0.401
0.1	0.5	0.9	0.966	0.970	0.964
0.1	0.7	0.7	0.858	0.891	0.846
0.1	0.9	0.9	0.992	0.994	0.989
0.3	0.3	0.3	0.05	0.059	0.037
0.3	0.3	0.7	0.384	0.417	0.354
0.3	0.5	0.5	0.110	0.138	0.098
0.3	0.5	0.9	0.714	0.746	0.693
0.3	0.7	0.7	0.432	0.470	0.409
0.3	0.9	0.9	0.880	0.900	0.846
0.5	0.5	0.5	0.045	0.053	0.040
0.5	0.5	0.9	0.469	0.511	0.409
0.5	0.7	0.7	0.129	0.154	0.104
0.5	0.9	0.9	0.537	0.573	0.473
0.7	0.7	0.7	0.054	0.054	0.031
0.7	0.9	0.9	0.192	0.211	0.123
0.9	0.9	0.9	0.020	0.034	0.002

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

Table 7: Simulation results for *n*=20 for three group comparison.

π_1	π_2	π_3	WOR	WR	X^2
0.1	0.1	0.1	0.044	0.040	0.030
0.1	0.1	0.5	0.877	0.877	0.844
0.1	0.1	0.9	1.000	1.000	1.000
0.1	0.3	0.3	0.331	0.330	0.288
0.1	0.3	0.7	0.977	0.976	0.970
0.1	0.5	0.5	0.835	0.846	0.806
0.1	0.5	0.9	1.000	1.000	1.000
0.1	0.7	0.7	0.994	0.994	0.990
0.1	0.9	0.9	1.000	1.000	1.000
0.3	0.3	0.3	0.056	0.059	0.039
0.3	0.3	0.7	0.763	0.776	0.744
0.3	0.5	0.5	0.245	0.258	0.224
0.3	0.5	0.9	0.977	0.976	0.969
0.3	0.7	0.7	0.783	0.798	0.758
0.3	0.9	0.9	0.999	0.999	0.995
0.5	0.5	0.5	0.039	0.049	0.035
0.5	0.5	0.9	0.848	0.852	0.821
0.5	0.7	0.7	0.231	0.242	0.211
0.5	0.9	0.9	0.874	0.880	0.839
0.7	0.7	0.7	0.043	0.042	0.034
0.7	0.9	0.9	0.383	0.371	0.332
0.9	0.9	0.9	0.036	0.037	0.022

Table 8: Simulation results for *n*=30 for three group comparison.

π_1	π_2	π_3	WOR	WR	X^2
0.1	0.1	0.1	0.051	0.045	0.036
0.1	0.1	0.5	0.979	0.977	0.970
0.1	0.1	0.9	1.000	1.000	1.000
0.1	0.3	0.3	0.485	0.486	0.444
0.1	0.3	0.7	1.000	1.000	1.000
0.1	0.5	0.5	0.964	0.963	0.955
0.1	0.5	0.9	1.000	1.000	1.000
0.1	0.7	0.7	1.000	1.000	1.000
0.1	0.9	0.9	1.000	1.000	1.000
0.3	0.3	0.3	0.045	0.047	0.040
0.3	0.3	0.7	0.913	0.919	0.910
0.3	0.5	0.5	0.338	0.345	0.318
0.3	0.5	0.9	1.000	1.000	1.000
0.3	0.7	0.7	0.896	0.907	0.890
0.3	0.9	0.9	1.000	1.000	1.000
0.5	0.5	0.5	0.048	0.051	0.043
0.5	0.5	0.9	0.971	0.976	0.968
0.5	0.7	0.7	0.364	0.375	0.337
0.5	0.9	0.9	0.967	0.967	0.958
0.7	0.7	0.7	0.047	0.049	0.043
0.7	0.9	0.9	0.561	0.560	0.527
0.9	0.9	0.9	0.040	0.036	0.030

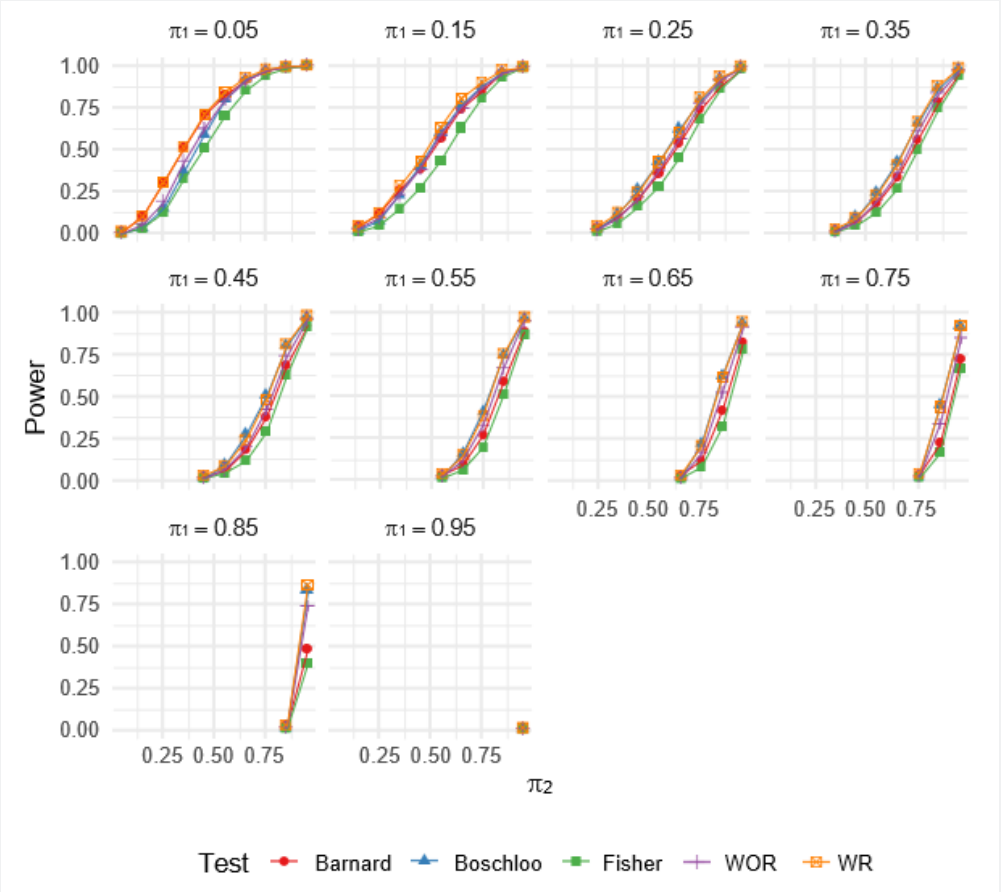

Figure 2: Power curves for various test methods across different (π0, π1) values, n=20.

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

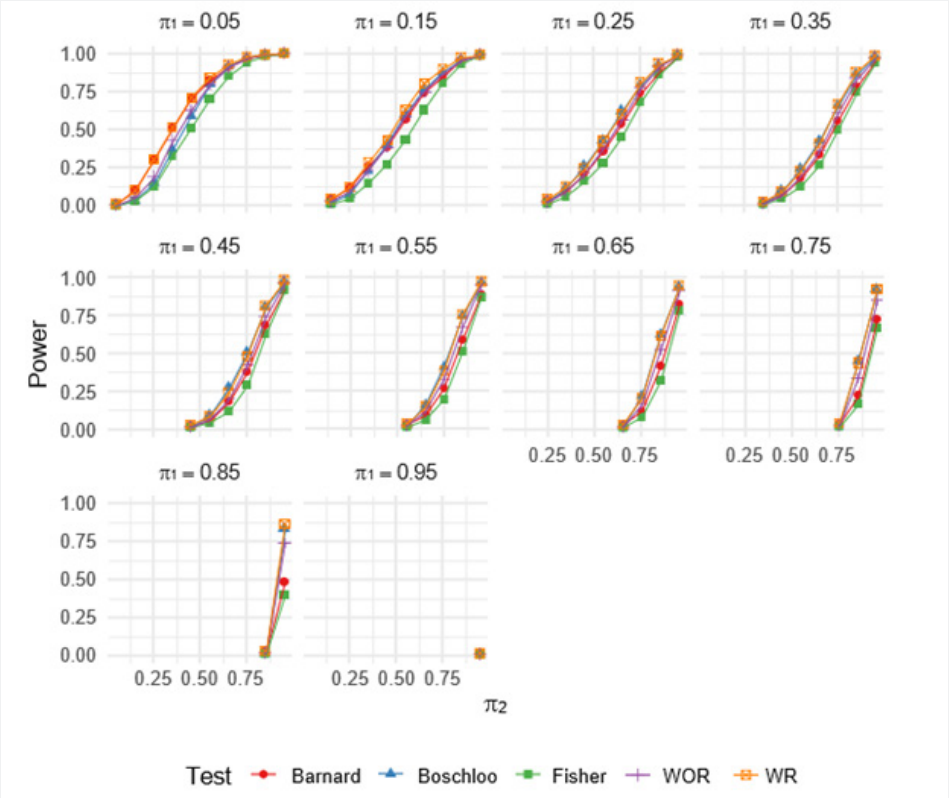

Figure 3: Power curves for various test methods across different (π0, π1) values, n=30.

In terms of power, both the WOR and WR perturbation tests consistently outperformed the Pearson exact chi-squared test, particularly at small to moderate sample sizes and under moderate deviations from the null hypothesis. For example, when $\pi_1=\pi_2=0.1$ and $\pi_3=0.5$, the WR method achieved powers of 0.580, 0.877, and 0.979 for n=10, 20, 30, respectively, compared to 0.456, 0.844, and 0.970 for the Pearson exact chi-squared test (see corresponding rows in Tables 6-8). Similar patterns were observed in other configurations. When the effect size was large (e.g., $\pi_1=\pi_2=0.1$, $\pi_3=0.9$), all three methods approached maximum power even at the smallest sample size.

Overall, these results demonstrate that the WOR and WR perturbation tests offer superior performance in terms of both Type I error control and statistical power, especially in small-sample settings where the assumptions of the chi-squared test may be violated. Among the two resampling-based methods, the WR perturbation test provided the best overall performance, consistently balancing Type I error control and power across all scenarios.

Design Examples

To guide study planning for the one-sided hypothesis H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$, we first computed the per group sample size n required by Boschloo's test to attain 80% power at the α=0.05 level. The resulting n values for several (π_1, π_2) configurations are listed in (Table 9).

Table 9: Example power comparisons across tests.

n	π_1	π_2	Barnard	Boschloo	Fisher	WOR	WR
10	0.05	0.55	0.831	0.799	0.701	0.811	0.851
11	0.05	0.50	0.790	0.809	0.718	0.80	0.849
17	0.15	0.55	0.815	0.807	0.723	0.806	0.830
12	0.35	0.85	0.773	0.807	0.738	0.768	0.823

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

Three findings stand out.

- Fisher's exact test is markedly under powered relative to the other four procedures.
- Contrary to common belief, Boschloo's test does not always dominate Barnard's test; in some settings Barnard offers comparable or even higher power.
- The WR perturbation test uniformly outperforms all competitors.

Because real-world conclusions frequently hinge on *p*-values that lie near the significance threshold, the choice of test can materially affect inference; this will become evident in the applied examples of the next section. Finally, we assessed computational burden for the setting π_1=0.35, π_2=0.55 with *n*=75-85. Power curves for the WOR and WR perturbation methods were produced in minutes on a standard desktop, whereas Boschloo's evaluations stalled for several days without finishing, underscoring its limited practicality for large sample analyses.

Data Examples

Testing Binomial Endpoints Across Two Treatment Groups

The following example, which compares several statistical methods for testing binomial endpoints between two treatment groups, is adapted from the study by Ajani et al. [13]. In this trial, trimodality eligible patients were randomized to receive either no induction chemotherapy (IC; Arm A) or IC consisting of oxaliplatin and fluorouracil (Arm B), followed by concurrent chemoradiation with oxaliplatin/fluorouracil and radiation therapy. One of the primary endpoints was the pathological complete response (pathCR) rate. A total of 55 patients in Arm A and 54 patients in Arm B underwent surgery.

We utilized 100,000 resamples to calculate the WR and WOR perturbation test *p*-values for testing the null hypothesis H_0: $\pi_1=\pi_2$ versus the alternative H_1: $\pi_1\neq\pi_2$. The observed pathCR rates were 13% (7 of 55) in Arm A and 26% (14 of 54) in Arm B. Results from several statistical tests are summarized below:

- Fisher's exact test (two-sided): p=0.094
- Barnard's test: p=0.082
- Boschloo's test: p=0.084
- WOR test: p=0.073
- WR test: p=0.081

The primary hypothesis was evaluated at a significance level of α=0.05. As shown, none of the tests reached conventional statistical significance, although the WOR and WR perturbation tests yielded relatively smaller *p*-values compared to traditional exact tests.

If, hypothetically, the observed pathCR rates were instead 13% (6 of 54) in Arm A and 26% (14 of 54) in Arm B, the test results would be:

- Fisher's exact test (two-sided): p=0.081
- Barnard's test: p=0.065
- Boschloo's test: p=0.058
- WOR test: p=0.053
- WR test: p=0.048

In this scenario, the WR perturbation test demonstrated in our simulation study to maintain appropriate Type I error control while offering consistently greater power, would lead to a different conclusion, suggesting statistical significance, unlike the more conservative traditional tests.

Testing Binomial Endpoints Across g Treatment Groups

For this example, comparing Pearson's exact chi-square test with the WR and WOR perturbation tests, we utilized data from a multicenter, randomized controlled trial conducted across 20 Japanese medical institutions [14]. The study compared three biologics, namely, Infliximab (IFX), Vedolizumab (VED), and Ustekinumab (UST) as treatment arms. The primary endpoint was the clinical remission (CR) rate at week 12, with secondary endpoints including the treatment continuation rate at week 26 and adverse events (AEs). The observed CR rates at week 12 were: 36% (12 of 33) for IFX, 32% (11 of 34) for VED, and 43% (13 of 30) for UST. We used 100,000 resamples to approximate *p*-values for the exact Pearson chi-square test, the WR perturbation test, and the WOR perturbation test when testing the null hypothesis: H_0: $\pi_1=\pi_2=\pi_3$ versus H_1: not all π_i are equal.

- Pearson exact chi-square test: p=0.675
- WOR test: p=0.663
- WR test: p=0.649

For the secondary endpoint of rectal bleeding score of 0 at week 1, the rates were: 39% (13 of 33) for IFX, 50% (17 of 34) for VED, and 70% (21 of 30) for UST. The corresponding *p*-values were:

- Pearson exact chi-square test: p=0.052
- WOR test: p=0.043
- WR test: p=0.044

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

In this case, different conclusions would be drawn from the WR and WOR tests as compared to the Pearson exact chi-square test at a significance level of α=0.05.

Conclusion

In this note we introduced two new perturbation tests for the hypothesis H_0: $\pi_1=\pi_2$ versus H_1: $\pi_1>\pi_2$ (or $\pi_1<\pi_2$),

together with the two-sided alternative H_1: $\pi_1\neq\pi_2$. The with-replacement (WR) and without-replacement (WOR) perturbation tests are both simple to implement, only a handful of Monte-Carlo resampling lines in R suffice, yet their operating characteristics differ in a way that is practically important. Overall, the WR perturbation test was superior.

Some Key Features of the WR Test

- **Consistently higher power**: Across an extensive grid of sample size configurations, we benchmarked WR against Fisher's exact, Barnard's, and Boschloo's tests. In every scenario the WR perturbation test delivered the greatest power, with the advantage most pronounced in the small sample setting that dominates Phase I/II clinical trials and rare-event studies. Even a seemingly modest uptick, for example improving power from 0.80 to 0.85, can flip a borderline *p*-value across the prespecified α threshold, changing the scientific conclusion.

- **Exact type-I error control**: Like Fisher's and Boschloo's procedures, WR maintains the nominal level without the conservatism that plagues Fisher's exact test. Type-I error protection is therefore not sacrificed for power.

- **Generalizes seamlessly**: We extended the method to the *g*-group hypothesis H_0: $\pi_1=\pi_2=\cdots=\pi_g$ versus H_1: not all π_i are equal, and demonstrated that WR outperforms the Freeman-Halton exact test while maintaining the desired Type I error rate. The same resampling blueprint naturally accommodates multinomial outcomes, a direction of future work.

- **Computational simplicity and transparency**: Because the test statistic is distribution-free under H_0, Monte-Carlo *p*-values are obtained in seconds on a lap top, obviating large enumeration tables or specialized software. This lowers the barrier to adoption for practicing analysts.

Conclusion

When a more powerful test requires no additional modelling assumptions, is trivially programmed, and retains exact size, there is little rationale for defaulting to less efficient competitors. The evidence presented here positions the WR perturbation test as the new small sample gold standard for binary proportion comparisons.

Funding

This work was supported by the following NCI grants to Hutson: NRG Oncology Statistical and Data Management Center grant (grant no. U10CA180822); Acquired Resistance to Therapy network (ARTNet) grant (grant no. U24CA274159).

Data Availability

The data used in this study is contained within the manuscript in Section 4.

References

1. Rosenbaum PR (2002) Covariance Adjustment in Randomized Experiments and Observational Studies. Statistical Science 17: 286-327.
2. Fisher RA (1934, 1970) Statistical Methods for Medical Researchers. Edinburgh, Oliver and Boyd.
3. Davis LJ (1986) Exact Tests for 2×2 Contingency Tables. The American Statistician 40: 139-141.
4. Barnard G (1945) A new Test for 2×2 Tables. Nature 177.
5. Boschloo RD (1970) Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities. Statistica neerlandica 24: 1-35.
6. Korn EL, Freidlin B (2024) Design of randomized clinical trials with a binary endpoint: Conditional versus unconditional analyses of a two-by-two table. Statistics in Medicine 43: 3109-3123.
7. Mehrotra DV, Chan IS, Berger RL (2003) A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. Biometrics 59: 441-450.
8. Lin CY, Yang MC (2008) Improved p-Value Tests for Comparing Two Independent Binomial Proportions. Communications in Statistics-Simulation and Computation 38: 78-91.
9. Andr´es AM, Mato AS (1994) Choosing the optimal unconditioned test for comparing two independent proportions. Computational Statistics & Data Analysis 17: 555-574.
10. Freeman GH, Halton JH (1951) Note on an Exact Treatment of Contingency, Goodness of Fit and Other Problems of Significance. Biometrika 38: 141-149.
11. Agresti A, Wackerly D, Boyett JM (1979) Exact Conditional Test for Cross Classifications: Approximation of Attained Significance Levels. Psychometrika 44: 75-83.
12. Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's. exact test in rxc contingency tables. Journal of the American Statistical Association 78: 427-434.
13. Ajani JA, Xiao L, Roth JA, Hofstetter WL, Walsh G, et al. (2013) A phase II randomized trial of induction chemotherapy versus no induction chemotherapy followed by preoperative chemoradiation in patients with esophageal cancer. Annals of Oncology 24: 2844-2849.
14. Naganuma M, Shiga H, Shimoda M, Matsuura M, Takenaka K, et al. (2025) Firstline biologics as a treatment for ulcerative colitis: a multicenter randomized control study. Journal of Gastroenterology 60: 430-441.

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835

DOI: 10.19080/BBOAJ.2025.12.555835

How to cite this article: Alan D H. A Perturbation-Based Method to Boost Exact Tests of Proportions in . Biostat Biom Open Access J. 2025; 12(2): 555835. DOI:10.19080/BBOAJ.2025.12.555835