

Comparing Exact, Mid-p, and Wilson Score Tests for a Single Binomial Proportion

Hasan M Rhoma^a 

^aDepartment of Statistics, Faculty of Science, Tripoli University, Tripoli, Libya

*Corresponding email address: h.rhoma@uot.edu.ly

Received 03 Jul 2026 | Accepted 22 Sep 2026 | Available online 23 Sep 2026 | DOI: 10.26629/uzjns.2026.08

ABSTRACT

Recent methodological studies have continued to refine and scrutinize exact binomial inference, providing comparisons between exact tests, approximate methods, and mid- p corrections. This ongoing work reflects that even after many decades, the simple binomial problem continues to yield valuable theoretical insights. A consistent finding across studies is the inadequacy of methods based on the normal approximation. In particular, the exact binomial test, while exact in principle, is known to be highly conservative, with its actual significance level often substantially lower than the nominal level (actual significance level $\leq \alpha$), implying a loss of statistical power and thus a reduced ability to detect a true effect when it exists. This study compares four methods for inference on a single binomial proportion: the standard exact binomial test (using the nominal significance level $\alpha = 0.05$), the exact binomial test (using the actual significance level), the mid- p test, and the widely recommended Wilson score test. Using Monte Carlo simulation with 20000 replications, we compare these tests in terms of empirical Type I error rate and statistical power across a comprehensive grid of sample sizes ($n = 10$ to 500) and nine null success probabilities ($p_0 = 0.02, 0.04, 0.06, 0.08, 0.10, 0.20, 0.30, 0.40, 0.50$). The evidence strongly indicates that the mid- p test outperforms the standard exact binomial test and the Wilson score test for the analysis of a single binomial proportion, particularly in small-to-moderate sample sizes. It achieves a Type I error rate close to the nominal level while consistently providing higher power. The use of an actual significance threshold, although a valuable theoretical exercise, offers no practical advantage over the simpler and more effective mid- p procedure. We therefore strongly recommend the adoption of the mid- p test in practice.

Keywords: Exact Binomial Test, Nominal Level, Actual Level, Mid- p Test, Wilson Score Test, Monte Carlo Simulation.

مقارنة بين الاختبارات Exact Test، و Mid-p Test، و Wilson Score Test لنسبة ذات حدين

حسن محمد ارحومة

قسم الإحصاء، كلية العلوم، جامعة طرابلس، طرابلس، ليبيا

الملخص

لاتزال الدراسات المنهجية الحديثة تسعى إلى تحسين وتدقيق الاستدلال حول نسبة ذاتي الحدين، من خلال مقارنة الاختبارات الدقيقة والطرق التقريبية، وتصحيحات mid- p . وتعكس هذه الجهود المستمر أنه حتى بعد عقود عديدة، لا تزال مسألة التوزيع الثنائي البسيطة تقدم رؤى نظرية قيمة. ومن النتائج الثابتة عبر الدراسات عدم كفاية الأساليب القائمة على التقريب الطبيعي. الاختبار الثنائي الدقيق، رغم كونه دقيقاً من حيث المبدأ، إلا أنه شديد التحفظ، إذ غالباً ما يكون مستوى الدلالة الفعلي أقل بكثير من المستوى الاسمي (مستوى الدلالة الفعلي $\geq \alpha$)، مما يعني فقدان القوة الإحصائية وبالتالي انخفاض القدرة على كشف تأثير حقيقي عند وجوده. تقارن هذه الدراسة أربع طرق للاستدلال على نسبة ذات حدين المفردة: الاختبار الثنائي الدقيق القياسي (باستخدام مستوى الدلالة الاسمي $\alpha = 0.05$)، والاختبار الثنائي الدقيق (باستخدام مستوى الدلالة الفعلي)، واختبار mid- p ، واختبار ويلسون Wilson Score الموصى به على نطاق واسع. أجرينا عملية المقارنة باستخدام محاكاة مونت كارلو مع 20000 تكرار، قارنا خلالها هذه الاختبارات من حيث معدل الخطأ من النوع الأول والقوة الإحصائية عبر شبكة من أحجام العينات ($n = 10$ إلى 500)، واحتمالات النجاح في الفرضية الصفرية ($P_0 = 0.02, 0.04, 0.06, 0.08, 0.10, 0.20, 0.30, 0.40, 0.50$). تشير الأدلة بقوة إلى أن اختبار mid- p يتفوق على الاختبار الدقيق القياسي واختبار ويلسون في تحليل نسبة ذاتي الحدين، لا سيما في أحجام العينات الصغيرة إلى المتوسطة. فهو يحقق معدل خطأ من النوع الأول قريباً من المستوى الاسمي مع تقديمه باستمرار قوة إحصائية أعلى. رغم أن استخدام عتبة دلالة فعلية يُعد ممارسة نظرية قيمة، إلا أنه لا يقدم أي ميزة عملية مقارنة باختبار mid- p ، ولذلك نوصي بشدة باعتماد اختبار mid- p في الممارسة العملية.

الكلمات المفتاحية: اختبار ذاتي الحدين الدقيق، المستوى الاسمي، المستوى الفعلي، اختبار Mid- p ، اختبار درجة ويلسون، محاكاة مونت كارلو.

1. Introduction

Exact tests represent a cornerstone of frequency statistical inference, particularly when dealing with discrete data distributions such as the binomial, Poisson, or hypergeometric. Unlike asymptotic tests, which rely on large-sample approximations (e.g., normal or chi-square distributions), exact tests derive their sampling distributions directly from the underlying probability model without approximation. The modern formulation of the exact binomial test is most directly associated with Clopper and Pearson [1]; they introduced a procedure for constructing confidence intervals for a binomial proportion that guarantees the coverage probability is at least the nominal level. This work provided:

- Exact confidence intervals, known today as Clopper–Pearson intervals, which ensure that the actual confidence level does not fall below the stated level.
- Formal justification for the exact test of a binomial proportion, based directly on the binomial distribution rather than large-sample approximations.

Reference [1] recognized that exact intervals are necessarily conservative (i.e., the actual coverage probability often exceeds the nominal level) due to the discrete nature of the binomial distribution. Nonetheless, this conservatism is precisely what makes the test exact in the sense of controlling the Type I error rate: it never exceeds the nominal significance level α . The term exact refers to the use of precisely defined distributions rather than the normal or chi-square distribution to obtain p -values and confidence intervals. Nevertheless, the inferences are not exact in the sense that the error probabilities exactly equal the nominal values. Rather, the nominal values serve as upper bounds on the true error probabilities, a point well established in the statistical significance testing literature [2]. Despite its conservative nature, the exact test is widely recommended when sample sizes are small or when asymptotic approximations are questionable. It provides a gold standard for validity, even if it is sometimes criticized for being overly conservative. Modern computational advances have made exact tests feasible even for relatively large samples, through algorithms that enumerate or sample from the reference set. Consequently, exact inference remains an essential tool in biostatistics, clinical trials, and categorical data analysis [3]. However, the conservatism of the exact test has motivated the

development of alternative approaches. The mid- p test, proposed by Lancaster [4], offers a compromise that reduces conservatism while maintaining near-nominal Type I error rates. More recently, the Wilson score test, proposed by Wilson [5], has gained popularity as a simple and robust approximate method for inference on proportions, often outperforming the normal approximation. This study aims to provide a comprehensive comparison of these three approaches alongside the standard exact test, to offer clear practical guidance for researchers.

1.1. The Exact Binomial Test

The exact binomial test is based on the binomial distribution. For a test of the null hypothesis $H_0: p = p_0$, the test statistic is the number of successes X , which under the null hypothesis follows a Binomial (n, p_0) distribution. For a two-sided test, the p -value is calculated as the sum of the probabilities of all outcomes that are as likely as, or less likely than, the observed outcome under the null hypothesis.

A key characteristic of the exact test is its conservativeness. Because the binomial distribution is discrete, the test cannot achieve the exact nominal level for all sample sizes; instead, it attains a level strictly less than or equal to α . While this ensures finite-sample validity, it may result in reduced power compared to approximate methods. This property is demonstrated in Figure 2, which shows the actual significance level of the two-sided exact binomial test as a function of sample size n for various null proportions p_0 , with a nominal significance level of $\alpha = 0.05$. The figure illustrates how the actual significance level fluctuates and is almost always below the nominal significance level.

1.2. The Mid- p Test

When analyzing binomial data, statisticians face a persistent dilemma: exact tests are conservative, with their actual significance level often considerably lower than the nominal significance level, leading to a loss of statistical power. The mid- p test offers a compromise, balancing the conservativeness of exact methods with the higher power of approximate ones. The concept of the mid- p -value is generally credited to authors of reference [4], who proposed subtracting half the probability of the observed value to mitigate excessive conservatism. The mid- p test does not provide the theoretical guarantee that the actual Type I error rate will not exceed the nominal significance level. However, extensive simulations, for example [6], has shown that violations are extremely rare and small in

magnitude. For a nominal level of 0.05, no violation was observed across 9595 scenarios. A rigorous theoretical foundation was later provided by [7] that developed a theory of expected p -values based on the Neyman-Pearson lemma and decision-theoretic principles. They showed that the expected p -value coincides exactly with the mid- p -value.

1.3. The Wilson Score Test

The Wilson score test is an asymptotic method for inference on a binomial proportion that is known to perform much better than the standard normal approximation (Wald test). It is based on inverting the score test and results in a test statistic that follows a chi-square distribution asymptotically. The test is often recommended for constructing confidence intervals for proportions, as the resulting Wilson interval has better coverage properties than the Wald interval, especially for small samples or proportions near 0 or 1, as shown by [8]. The Wilson test can be implemented easily using standard statistical software. For a two-sided test of $H_0: p = p_0$, the test statistic is compared to a chi-square distribution with one degree of freedom. Despite its asymptotic nature, the Wilson test is generally considered a robust and simple alternative.

1.4. Nominal vs. Actual significance level

In exact tests in hypothesis testing, the nominal significance level α is the pre-specified threshold for the probability of a Type I error, chosen by the researcher before data collection. Typical values include 0.10, 0.05, or 0.01. In contrast, the actual significance level is the exact probability of rejecting the null hypothesis when it is true, given the specific test procedure and the discrete nature of the test statistic. The exact binomial test relies on the binomial distribution, which is inherently discrete. As a result, it is generally impossible to achieve a test whose actual significance level exactly matches a pre-specified nominal level (e.g., $\alpha = 0.05$). The discrete set of possible probabilities under the null hypothesis means that only certain significance levels can be realised. Consequently, the actual significance level is typically slightly lower. We show this for the specific case (see Figure 1). Suppose we test this hypothesis $H_0: P = 0.5$ vs $H_1: P < 0.5$. The test statistic: $X \sim \text{Bin}(16, 0.5)$. The actual significance level is the sum of probabilities of all outcomes in the rejection region under H_0 . Then

$$\alpha_{\text{actual}} = P_{H_0}(\text{Reject } H_0)$$

$$\alpha_{\text{actual}} = P_{H_0}(x \leq k) \leq \alpha_{\text{nominal}},$$

$$\alpha_{\text{actual}} = P_{H_0}(x \leq 4) = 0.0384 \leq 0.05.$$

In exact hypothesis testing for a binomial proportion, the actual significance level (also known as the size of

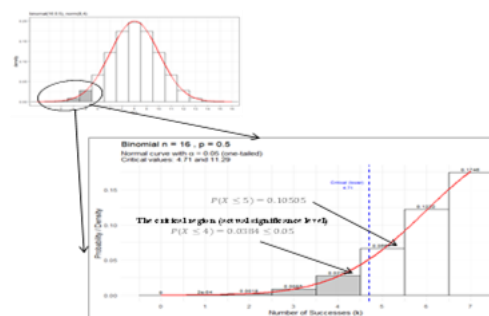


Figure 1. illustrates the critical region, showing how the discrete nature of the binomial distribution forces the actual significance level below the nominal level.

the test) is defined as the maximum probability of

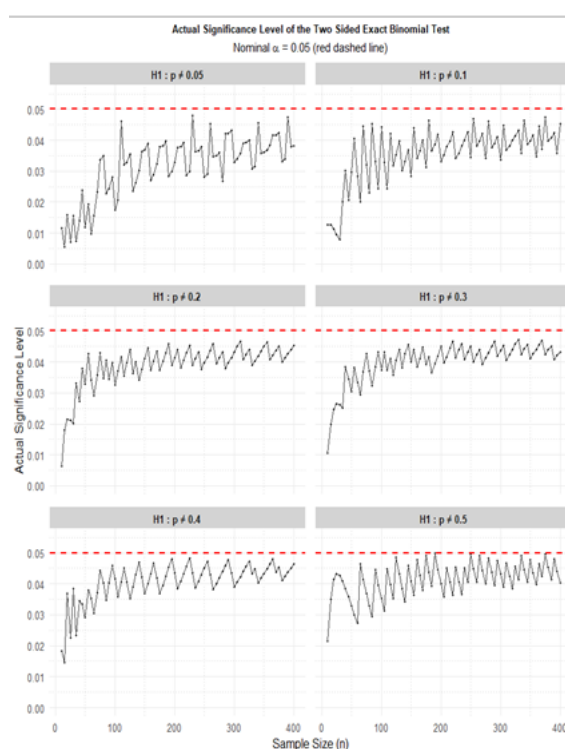


Figure 2. The actual significance level as a function of n and p

rejecting the null hypothesis under H_0 given the discrete nature of the test statistic. Unlike continuous tests, where the nominal level α can be achieved exactly, the discrete binomial distribution forces the actual level to be less than or equal to α . This is the main reason why there is a weakness in the test. The following Figure 2 illustrates how the actual significance level of a two-sided exact binomial test varies across sample sizes ($n = 10, 15, 20, 25, 30, 35, \dots, 400$) and null proportions ($P_0 = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$). The nominal level is fixed at $\alpha = 0.05$.

2. Method

2.1. Evaluation Criteria

The two most fundamental criteria for evaluating hypothesis tests are the Type I error rate (size) and the statistical power. They constitute the primary metrics considered in this work. An ideal hypothesis test should satisfy two key properties:

1. The actual Type I error rate must not exceed the nominal significance level α (i.e., the test should be valid or at most conservative).
2. The power should be as high as possible; a common heuristic benchmark for acceptable power is 0.8, though this is context-dependent and not a strict mathematical requirement.

When comparing tests, the preferred procedure should achieve optimal (or near-optimal) power while avoiding excessive conservatism; that is, the actual Type I error rate should be as close as possible to α without exceeding it.

2.2. Test Procedures

Four test procedures were evaluated in this simulation study:

1. **Exact Test (Nominal Level):** The standard two-sided exact binomial test using the nominal significance level $\alpha = 0.05$. The null hypothesis is rejected if the exact p -value is less than or equal to $\alpha = 0.05$.
2. **Exact Test (Actual Level):** The two-sided exact binomial test using the largest attainable significance level (α_{actual}) that does not exceed the nominal significance level $\alpha = 0.05$. For a given n and p_0 , the actual significance level is determined by enumerating all possible outcomes of X under the null hypothesis. For each outcome, the two-sided p -value is calculated. The set of all possible p -values is then examined, and the actual significance level is defined as the maximum p -value that is less than or equal to α . The null

hypothesis is rejected if the p -value of the observed outcome is less than or equal to (α_{actual}).

3. **Mid-p Test:** The test based on the *mid-p-value*, defined as $\text{mid-}p\text{-value} = \text{exact-}p\text{-value} - 0.5 \times P(X = x)$, where *exact- p -value* is the two-sided exact p -value and $P(X = x)$ is the probability of the observed outcome. The null hypothesis is rejected if $\text{mid-}p\text{-value} \leq \alpha$.
4. **Wilson Score Test:** a statistical method used to assess whether a sample proportion significantly differs from a hypothesized population proportion p_0 . Unlike the Wald test, it uses the standard error under the null hypothesis, which yields superior performance for small samples and skewed probabilities, and it is the basis for the widely used Wilson confidence interval. The test statistic is calculated as the standardized difference between the observed proportion ($\hat{p} = x/n$) and the null proportion p_0 , divided by the null standard error:

$$Z = \frac{p_0 - \hat{p}}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (1)$$

This Z-score asymptotically follows a standard normal distribution. The square of this statistic, Z^2 asymptotically follows a chi-square distribution with 1 degree of freedom, which is often used for two-tailed testing.

3. Simulation Design

A comprehensive Monte Carlo simulation was conducted to compare the performance of the four tests. All simulations were performed using R with $n_{sims} = 20000$ replications per scenario to ensure stable estimates. The simulation covered the following parameter space,

Sample sizes: ($n = 10, 20, 30, \dots, 500$).

Table 1. Alternative proportion values (p_1) used for power computation

p_0	p_1 values								
0.02	0.04	0.06	0.08	0.10	0.15	0.20	0.30	0.40	
0.04	0.06	0.08	0.10	0.15	0.20	0.30	0.40	0.50	
0.06	0.08	0.10	0.15	0.20	0.30	0.40	0.50	0.60	
0.08	0.10	0.10	0.20	0.30	0.40	0.50	0.60	0.65	
0.10	0.20	0.30	0.40	0.50	0.60	0.65	0.70		
0.20	0.30	0.40	0.50	0.60	0.65	0.70	0.75		
0.30	0.40	0.50	0.60	0.65	0.70	0.75	0.80		
0.40	0.50	0.60	0.65	0.70	0.75	0.80	0.85		
0.50	0.60	0.65	0.70	0.75	0.80	0.85	0.90		

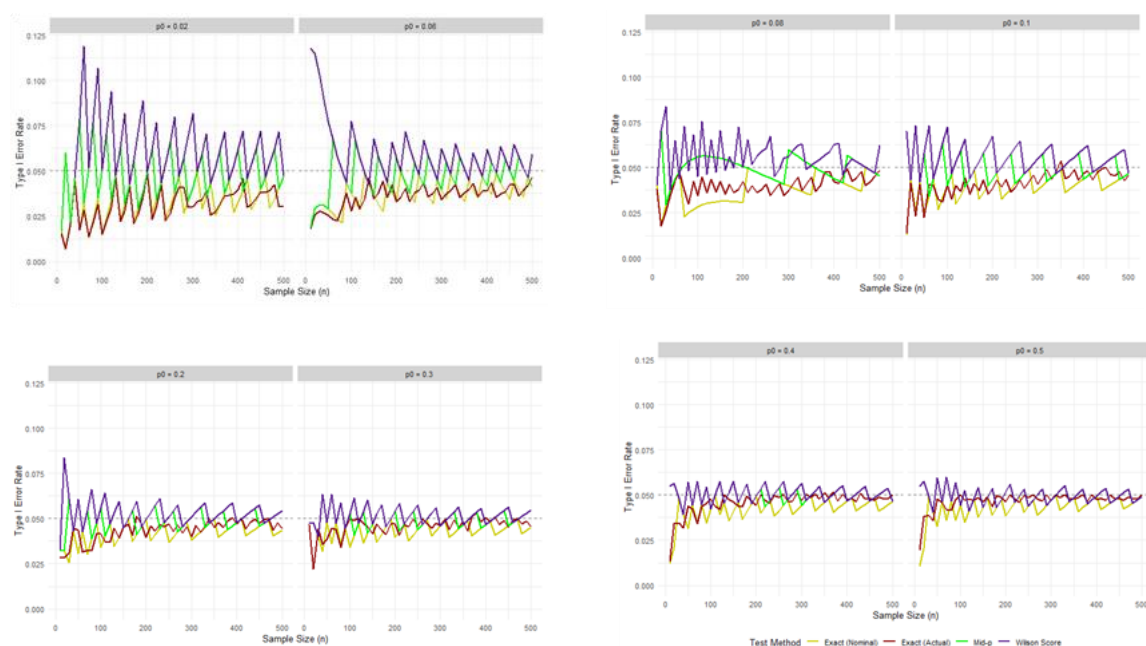


Figure 3. Type I error rate comparison across methods

Null proportions: ($p_0 = 0.02, 0.04, 0.06, 0.08, 0.10, 0.20, 0.30, 0.40$, and 0.50). Alternative proportions: For each p_0 , a set of p_1 values was used to compute power. These sets are shown in Table 1.

For Type I error rate estimation, data were generated under the null hypothesis p_0 . For power estimation, data were generated under the alternative hypothesis p_1 . For each scenario, the rejection rate of each test was recorded. The Average Power for a given n and p_0 was calculated as the mean of the power estimates across all corresponding p_1 values. Standard errors (SE) for each estimate were calculated using the formula $SE = \sqrt{p(1-p)/n_{sims}}$, where p is the estimated rejection rate.

4. Results

This section presents the simulation results for Type I error rates and statistical power for the four test procedures. Now covering all nine null proportions.

The simulations were performed using 20000 replications per scenario to ensure stable estimates. Tables 2 and 3 summarize the key findings across sample sizes and null proportions.

4.1. Type I Error Rate

Figure 3 displays the average Type I error rates for the four tests across all nine simulated p_0 values, with a nominal level of 0.05 indicated by a dashed horizontal line. Table 2 presents selected values of the Type I error rates, averaged across all p_0 values, along with their standard errors.

As shown in Table 2, the standard exact test using the nominal level is consistently conservative. Its empirical Type I error rate averages between 0.022 and 0.046 across all scenarios, remaining well below the nominal 0.05 level, particularly for small samples and extreme null proportions. The exact test using the actual attainable level improves upon the nominal

Table 2. Average Type I error rate (across all p_0 values) and standard errors

Sample Sizes	Exact Nominal SE	Exact Actual SE	Mid-p SE	Wilson SE
$n = 10$	0.0219 (0.0145)	0.0221 (0.0132)	0.0437 (0.0183)	0.0732 (0.0478)
$n = 30$	0.0336 (0.099)	0.0304 (0.0077)	0.0428 (0.0167)	0.0776 (0.0306)
$n = 50$	0.0340 (0.0106)	0.0367 (0.0116)	0.0537 (0.0137)	0.0591 (0.0122)
$n = 80$	0.0334 (0.0098)	0.0338 (0.0069)	0.0516 (0.0114)	0.0546 (0.0113)
$n = 100$	0.0359 (0.0097)	0.0393 (0.0118)	0.0475 (0.0068)	0.0555 (0.0121)
$n = 150$	0.0398 (0.0056)	0.0388 (0.0072)	0.0443 (0.0082)	0.0578 (0.0152)
$n = 200$	0.0404 (0.0078)	0.0419 (0.0051)	0.0517 (0.0105)	0.0543 (0.0061)
$n = 300$	0.0408 (0.0041)	0.0398 (0.0061)	0.0491 (0.0074)	0.0575 (0.0107)
$n = 400$	0.0429 (0.0059)	0.0441 (0.0047)	0.0485 (0.0062)	0.0546 (0.007)
$n = 500$	0.0458 (0.002)	0.0441 (0.0067)	0.0481 (0.004)	0.0519 (0.0058)

version slightly but remains conservative in most cases. The mid- p test achieves Type I error rates closest to 0.05, with averages ranging from 0.0437 to 0.0537 and no systematic exceeding of the nominal level. The Wilson score test also performs reasonably well, with error rates generally close to 0.05 (averaging 0.0519 to 0.0776), though it tends to be slightly less conservative than the mid- p test, especially for very small samples (e.g., $n = 10$, average 0.0732). Examining individual scenarios (not shown in the Table 2.), the largest deviations occur when p_0 is extreme (e.g., $p_0 = 0.02$) and n is small, where the exact nominal test exhibits error rates as low as 0.016.

4.2. Statistical Power

Figure 4 displays the average power, calculated as the mean rejection rate over all alternative proportions p_1 for each p_0 , then averaged across all p_0 values. The associated standard errors are also reported. Table 3 presents selected values of the Average Power, averaged across all p_0 values, along with their standard errors.

The mid- p test consistently outperforms the other three tests in terms of average power across all sample sizes. Its superiority is most pronounced for small to moderate samples: for $n = 10$, the mid- p test attains an average power of 0.609 compared to 0.523 for the exact nominal test, 0.503 for Wilson, and 0.468 for the exact actual test. As sample size increases, the power of all four tests converges, but the mid- p test retains a noticeable advantage even at $n = 500$, where it ties with the exact nominal test at 0.978 and exceeds Wilson and exact actual (both 0.972). When examining power as a function of the alternative proportion p_1 (for example, $n = 100$, $p_0 = 0.02$), the mid- p test consistently yields a higher power curve across all p_1 values, with the benefit being most evident for alternatives close to the null.

5. Discussion

The results of this simulation study provide clear and consistent evidence regarding the performance of the four competing tests for a single binomial proportion. The key findings revolve around two main criteria: control of the Type I error rate and statistical power.

Conservativeness of the Exact Nominal Test

The standard exact binomial test using the nominal significance level $\alpha = 0.05$ is markedly conservative, particularly for small sample sizes and extreme null proportions. This conservativeness arises directly from the discrete nature of the binomial distribution: it is generally impossible to achieve a test whose actual size exactly equals 0.05, and the exact test defaults to the largest attainable level that does not exceed 0.05. In our simulations, the average Type I error rate for the exact nominal test was as low as 0.0219 at $n = 10$, far below the nominal level. This implies that the test rejects the null hypothesis less often than it should when the null is true, leading to a loss of power. This loss is substantial in practice; for instance, at $n = 10$, the exact nominal test achieved an average power of only 0.523, meaning it failed to detect a true effect in nearly 48% of cases where an effect existed.

5.1. Superiority of the Mid- p Test

The mid- p test effectively corrects this over-conservatism by subtracting half the probability of the observed outcome from the exact p -value. This simple adjustment smooths the discrete distribution and brings the actual Type I error rate much closer to the nominal 0.05 level. Our simulations show that the mid- p test maintains an average Type I error rate consistently near 0.05, ranging from 0.0437 to 0.0537, without systematically exceeding it. More importantly, this improved error control translates directly into higher statistical power. Across almost all scenarios, the mid- p test yielded the highest average power, with the advantage being most pronounced

Table 3. Average power (across all p_1 values) and standard errors

Sample Sizes	Exact Nominal SE	Exact Actual SE	Mid- p SE	Wilson SE
$n = 10$	0.523 (0.1457)	0.468 (0.173)	0.609 (0.112)	0.503 (0.135)
$n = 30$	0.767 (0.091)	0.746 (0.086)	0.775 (0.103)	0.747 (0.086)
$n = 50$	0.827 (0.081)	0.811 (0.080)	0.851 (0.079)	0.800 (0.087)
$n = 80$	0.874 (0.077)	0.854 (0.079)	0.892 (0.066)	0.855 (0.078)
$n = 100$	0.894 (0.069)	0.866 (0.083)	0.906 (0.064)	0.875 (0.074)
$n = 150$	0.927 (0.0606)	0.909 (0.067)	0.930 (0.059)	0.911 (0.067)
$n = 200$	0.943 (0.056)	0.934 (0.058)	0.951 (0.047)	0.932 (0.059)
$n = 300$	0.962 (0.044)	0.951 (0.053)	0.965 (0.040)	0.953 (0.050)
$n = 400$	0.971 (0.036)	0.963 (0.045)	0.973 (0.034)	0.966 (0.042)
$n = 500$	0.978 (0.029)	0.972 (0.036)	0.978 (0.029)	0.972 (0.035)

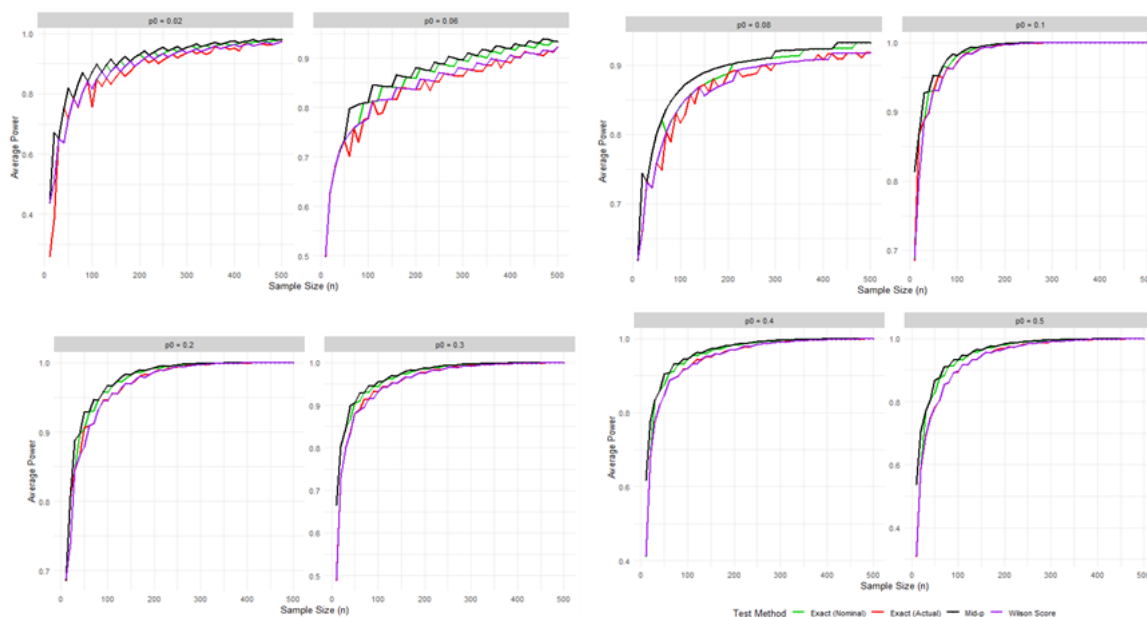


Figure 4. Average power comparison across methods

when the exact nominal test was most conservative (e.g., $p_0 = 0.02$, $n = 20$; average power: exact nominal = 0.506, mid- $p = 0.671$). This makes the mid- p test particularly valuable in small-to-moderate sample settings, where the power loss from the exact test can be scientifically significant.

5.2. Comparison with the Exact Actual Test

The exact test using the actual attainable significance level provides a modest improvement over the nominal version in terms of Type I error control, but it does not consistently match the mid- p test in power. In many scenarios, the actual level is even smaller than the nominal threshold, leading to a test that is still conservative and often less powerful than the standard exact test (e.g., at $n = 50$, average power for exact actual = 0.811 vs. exact nominal = 0.827). Moreover, this approach is impractical for routine use because it requires computing a data-dependent threshold for each combination of n and p_0 , whereas the mid- p test is straightforward to implement and readily available in statistical software.

5.3. Comparison with the Wilson Score Test

The Wilson score test, despite being an asymptotic method, performed reasonably well in terms of Type I error control, with averages close to 0.05. However, it was consistently less powerful than the mid- p test across all sample sizes. For example, at $n = 10$, the Wilson test had an average power of 0.503, compared to 0.609 for mid- p . Even at $n = 500$, the Wilson test (0.972) lagged behind mid- p (0.978). This may be due to the continuity correction or the inherent approximation error in small samples or extreme

proportions, which reduces its ability to detect true differences.

5.4. Convergence for Large Samples

As sample size increases, the binomial distribution approaches continuity, and the performance of all tests converges. However, even at $n = 500$, the mid- p test maintained a slight edge in power and superior Type I error control. This indicates that the benefits of the mid- p test are not limited to small samples, though they are most critical there.

5.5. Practical Implications

The practical implication of these findings is clear: researchers who rely on the standard exact binomial test are unnecessarily sacrificing power. The mid- p test provides a simple, effective, and well-justified alternative that achieves near-nominal Type I error rates while maximizing power. Given its ease of computation and availability, we recommend that the mid- p test become the default choice for inference on a single binomial proportion in most applied settings. The use of the exact actual level, while theoretically interesting, offers no practical advantage over the mid- p procedure and should be avoided in routine analysis.

6. Conclusion

This simulation study provides compelling evidence that, for inference on a single binomial proportion, the mid- p test is superior to the standard exact binomial test (using a nominal level), the exact test using an actual attainable level, and the widely recommended Wilson score test. The mid- p test achieves a Type I error rate that is consistently close to the nominal 0.05

level without systematically exceeding it, while simultaneously providing higher statistical power across nearly all sample sizes and null proportions examined. The standard exact binomial test, despite its "exact" label, is unnecessarily conservative due to the discreteness of the binomial distribution, leading to a substantial loss of power that can have serious practical consequences, particularly in small-sample studies. The exact test using an actual level does not remedy this adequately and is impractical. The Wilson score test, though robust, is consistently less powerful than the mid- p test. We therefore strongly recommend that researchers adopt the mid- p test as the standard procedure for testing a single binomial proportion. This recommendation is especially critical when sample sizes are small to moderate or when the null proportion is extreme, as these are precisely the conditions where the power loss from other tests is most severe. The mid- p test is straightforward to implement, computationally inexpensive, and available in most statistical packages, making it an accessible and practical solution. We also encourage textbook authors and software developers to promote and facilitate the use of the mid- p test, helping to improve the quality and reliability of statistical inference in practice.

Finally, it is worth reiterating that the term "exact" in the context of the binomial test is a misnomer. The test is "exact" only in the sense that it uses the exact binomial distribution rather than an asymptotic approximation; however, the actual Type I error rate is generally less than the nominal level, and the test does not achieve exact error probabilities. The mid- p test,

by providing a closer approximation to the nominal level, offers a more honest and effective approach to inference.

References

- [1] C. J. Clopper, & E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial", *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.
- [2] A. Agresti & A. Gottard, "Independence in multi-way contingency tables: S. N. Roy's breakthroughs and later developments", *Journal of Statistical Planning and Inference*, vol. 137, pp. 3216–3226, 2007.
- [3] C. R. Mehta & N. R. Patel, "A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables", *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 427–434, 1983.
- [4] H. O. Lancaster, "Significance tests in discrete distributions," *Journal of the American Statistical Association*, vol. 56, no. 294, pp. 223–234, 1961.
- [5] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.
- [6] M. W. Fagerland, S. Lydersen, & P. Laake, "Recommended tests for association in 2×2 tables," *Statistics in Medicine*, vol. 34, no. 14, pp. 2159–2175, 2015.
- [7] J. T. Hwang & M. C. Yang, "An optimality theory for mid p -values in 2×2 contingency tables," *Statistica Sinica*, vol. 11, pp. 807–826, 2001.
- [8] A. Agresti & B. A. Coull, "Approximate is better than 'exact' for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998.