Received: 10 February 2018 | Revised: 21 November 2018 | Accepted: 27 December 2018

DOI: 10.1002/pst.1929

MAIN PAPER

WILEY

Tests for noninferiority trials with binomial endpoints: A guide to modern and quasi-exact methods for biomedical researchers

Enrico Ripamonti[1] | Chris J. Lloyd[2]

[1]IRCCS Fondazione Don Carlo Gnocchi, Milan, Italy

[2]Melbourne Business School, The University of Melbourne, Melbourne, Australia

Correspondence
Enrico Ripamonti, IRCCS Fondazione Don Carlo Gnocchi, via Alfonso Capecelatro 66, 20148 Milan, Italy.
Email: enrico.ripamonti@unimib.it

Applied statisticians and pharmaceutical researchers are frequently involved in the design and analysis of clinical trials where at least one of the outcomes is binary. Treatments are judged by the probability of a positive binary response. A typical example is the noninferiority trial, where it is tested whether a new experimental treatment is practically not inferior to an active comparator with a prespecified margin δ. Except for the special case of $\delta = 0$, no exact conditional test is available although approximate conditional methods (also called second-order methods) can be applied. However, in some situations, the approximation can be poor and the logical argument for approximate conditioning is not compelling. The alternative is to consider an unconditional approach. Standard methods like the pooled z-test are already unconditional although approximate. In this article, we review and illustrate unconditional methods with a heavy emphasis on modern methods that can deliver exact, or near exact, results. For noninferiority trials based on either rate difference or rate ratio, our recommendation is to use the so-called E-procedure, based on either the score or likelihood ratio statistic. This test is effectively exact, computationally efficient, and respects monotonicity constraints in practice. We support our assertions with a numerical study, and we illustrate the concepts developed in theory with a clinical example in pulmonary oncology; R code to conduct all these analyses is available from the authors.

KEYWORDS

estimation, exact test, maximisation, noninferiority, nuisance parameters

1 | INTRODUCTION

In a two-arm clinical trial, n patients are randomised into two groups: n_1 patients are assigned to a new (experimental) treatment, and the remaining n_0 patients are assigned to a control treatment. In practice, some subjects are lost during follow-up, but we will assume noninformative dropout and n_0 and n_1 will indicate sample sizes at the end of the study. We will also adopt an intention-to-treat approach; namely, trial results will be evaluated on the basis of the planned treatment regimen regardless of patients' compliance. This maintains the benefits of randomisation and is preferred by international agencies and the pharmaceutical industry.

Generally, more than one outcome is collected during the observation period, but for the purposes of the present article, we focus on one outcome only (either primary or secondary). The targeted outcome could be the score attained on a

clinical scale (eg, a pain scale) or survival during follow-up, or another clinically meaningful measure. In this paper, we are only interested in binary outcomes, although much of the theory is more general.

Let p_1 be the probability of success for the treatment group and p_0 for the control group. The larger is p_1 compared with p_0, the more beneficial is the treatment. However, there are many ways to compare p_1 with p_0. Let G be any strictly increasing function and measure the benefit of treatment by $\Delta = G(p_1) - G(p_0)$. We consider the one-sided hypotheses:

$$H_0 : \Delta \leq \delta \text{ vs. } H_1 : \Delta > \delta, \tag{1}$$

where $\delta < 0$ is a small negative quantity set a priori by the researcher on the basis of theoretical and pragmatic considerations. In words, the null hypothesis is that patients do worse on treatment by an amount $|\delta|$ or larger, so the treatment is practically inferior. The alternative is that the treatment is not inferior. Noninferiority trials are generally used when a new experimental treatment is compared with another active control treatment that is already on the market.[1]

When $G(p) = p$ is the identity function, this is a test of whether the risk difference $p_1 - p_0$ exceeds δ. When $G(p) = \log(p)$, it is a test of whether the risk ratio p_1/p_0 exceeds $\exp(\delta)$. These are the two most common risk comparison measures, although we could also consider $G(p) = \log(p/(1-p))$, the log-odds transformation, which leads to a test of odds ratio. Munk et al[2] give some results for a general nondecreasing function G.

Many readers will be familiar with at least some of the standard tests[3-5] of hypotheses (1) and a nonexhaustive field of options will be detailed in Section 2.2. Unfortunately, the standard tests are based on distributional approximations that can be very poor in practical circumstances. Type 1 error can exceed the stated level, even for quite large sample sizes. This is the key reason why researchers need to know of modern developments over the past decade or so.

In this article, we consider unconditional tests, with a strong emphasis on exact or near exact size control. The theory of exact tests is outlined in Section 4. When treatment effect is measured by log-odds, there is a standard conditional method available known as Fisher's exact test based on the generalised hypergeometric distribution (see Johnson et al[6] for a review). While conditionally exact, this test is unconditionally conservative. For the risk difference or ratio, no exact conditional test exists so it is natural to look for an exact unconditional test. Such tests, appropriately defined, do exist. In general, one-sided hypotheses formalise problems where there is prior expectation concerning the direction of the treatment effect. While we focus here on one-sided tests, the general theory on exact tests applies to two-sided tests without modification.

The plan of the paper is as follows. In Section 2, we introduce the notation used throughout the article, the statistical model adopted as reference point, and a field of standard test statistics. This includes the second-order accurate test statistic, often known as r^*. In Section 3, we discuss why it is important to use modern methods for hypothesis testing in clinical trials. Technical issues on exactness, validity, and monotonicity are presented in Section 4. Exact unconditional tests are introduced in Section 5. In Section 6, we present the results of a numerical study comparing different test statistics and P values across several practical scenarios. Finally, in Section 7, the concepts presented in theory are illustrated with a clinical example in pulmonary oncology, and in Section 8, we outline our general conclusions.

2 | NOTATION, MODEL, AND TEST STATISTICS

This section establishes the notation and describes some of the standard test statistics with which many readers will be familiar. We give the theory generally in terms of an unspecified function G, which in practice will be the identity or logarithmic function.

2.1 | Notation and model

We denote by y_1 the number of positive responses out of n_1 patients in the treatment group and by y_0 the number of positive responses out of n_0 in the control group. These are assumed to be realisations of random variables $Y_j (j = 0, 1)$ with independent Binomial(n_j, p_j) distributions. So, ignoring the additive constant, the log-likelihood is

$$\tilde{\ell}(p_1, p_0; y_1, y_0) = y_0 \log p_0 + (n_0 - y_0)\log(1 - p_0) + y_1 \log p_1 + (n_1 - y_1)\log(1 - p_1). \tag{2}$$

For testing hypothesis (1), it will be convenient to parametrise the likelihood in terms of $\Delta = G(p_1) - G(p_0)$ and the baseline control success rate p_0. To this end, we write $p_1(\Delta, p_0) = G^{-1}(G(p_0) + \Delta)$, which equals $p_0 + \Delta$ when G is the identity and $p_0 e^{\Delta}$ when G is the logarithmic function. The reparametrised log-likelihood is $\ell(\Delta, p_0) = \tilde{\ell}(p_1(\Delta, p_0), p_0)$. The baseline probability p_0 is a nuisance parameter that has to be eliminated in order to make an inference about Δ.

2.2 | Standard first-order tests

The unrestricted maximum likelihood estimates (MLE) of p_j is $\hat{p}_j = y_j/n_j$, the proportion of patients in each arm who responded positively. We want to detect if $\Delta = G(p_1) - G(p_0)$ exceeds δ where G is specified nondecreasing function. A natural approach is to compare $G(\hat{p}_1)$ with $G(\hat{p}_0) + \delta$. Let H be another increasing function and consider the deviation statistic

$$H(G(\hat{p}_1)) - H(G(\hat{p}_0) + \delta). \tag{3}$$

Under mild regularity conditions, this converges to zero when $\Delta = \delta$ and to a value greater than zero if $\delta > 0$. Dividing by an estimate of its standard deviation $\sigma(p_1, p_0)$ produces a statistic that will be asymptotically normal, again under mild regularity conditions. The accuracy of this approximation as $n = \min(n_0, n_1)$ diverges is $O(n^{-1/2})$, and such methods are called *first order*. The standard deviation $\sigma(p_1, p_0)$ can be estimated replacing p_j either by $\hat{p}_j$ or by the restricted MLE $\tilde{p}_{j\delta}$ under the boundary null hypothesis that $\Delta = \delta$.

When G is the identity so that we are testing the risk difference, the most common choice for H is the identity, whence the test statistic equals $\hat{p}_1 - \hat{p}_0 - \delta$ divided by its standard deviation $(p_1(1-p_1)/n_1 + p_0(1-p_0)/n_0)^{1/2}$. The most common version of this test replaces p_j by $\tilde{p}_{j\delta}$. Obtaining the latter involves solving a cubic equation.[4,7] It can be shown that the resulting statistic is identical to Rao's score statistic but it is often known as the *pooled Z statistic*. This largely replaced the method of Dunnett and Gent[3] who estimated p_j under the null using a simpler method of moments. When unrestricted estimates $\hat{p}_j$ are used, it is often called the *unpooled Z statistic*.

When G is the log-function so that we are testing the risk ratio, the most common choice is $H = G^{-1}$, whence the test statistic equals $\hat{p}_1 - \hat{p}_0 e^{\delta}$ divided by its standard deviation $(p_1(1-p_1)/n_1 + e^{2\delta}p_1(1-p_0)/n_0)^{1/2}$. Again, the most common version of this test replaces p_j by $\tilde{p}_{j\delta}$, which now involves solving a quadratic equation.[8] This statistic is identical to Rao's score statistic.

Rather than dividing a measure of deviation by its standard error, an alternative approach is to use the likelihood ratio (LR) statistic whose approximate distribution is χ^2_1. For one-sided hypotheses like (1), it is more convenient to use the so-called signed root likelihood ratio (SRLR) statistic[9] that has the form

$$r(\delta) = \text{sign}(\hat{\Delta} - \delta)(2\ell(\hat{\Delta}, \hat{p}_0) - 2\ell(\delta, \tilde{p}_{0\delta}))^{1/2},$$

where $\hat{\Delta} = G(\hat{p}_1) - G(\hat{p}_0)$ is the unrestricted MLE of Δ.

2.3 | Second-order tests

Conditional tests to compare hypotheses such as (1) typically do not exist. However, a deep strand of research has led to a general form called the r^* formula,[10] given by

$$r^*(\delta) = r(\delta) + r(\delta)^{-1} \log\left\{\frac{q(\delta)}{r(\delta)}\right\}, \tag{4}$$

where $r(\delta)$ is the earlier mentioned SRLR statistic. The formula is based on a double saddle-point approximation to the distribution of the MLE conditional on an approximate ancillary statistic. The component $q(\delta)$ has, in general, a very complicated formula, but it can be simplified in exponential models; in particular, for the present case of noninferiority testing, it can be expressed as

$$q(\delta) = \frac{\{\tilde{w}_{0\delta}(\text{logit}(\tilde{p}_{1\delta}) - \text{logit}(\hat{p_1})) - \tilde{w}_{1\delta}(\text{logit}(\tilde{p}_{0\delta}) - \text{logit}(\hat{p_0}))\}}{\sqrt{\tilde{w}_{1\delta}/n_1 + \tilde{w}_{0\delta}/n_0}}, \tag{5}$$

where $\tilde{w}_{j\delta} = \tilde{p}_{j\delta}(1 - \tilde{p}_{j\delta})$ and $\tilde{p}_{j\delta}$ is the aforementioned ML estimate of p_j under the null.

The statistic $r^*(\delta)$ is also approximately standard normal but with relative error $O(n^{-1})$ rather than the $O(n^{-1/2})$ of first-order methods. The formula is also parametrisation invariant, as shown in Davison et al.[9] Note that r^* is undefined when $r(\delta) = 0$ but this is usually handled by defining $r^*(\delta) = r(\delta)$ when $|r(\delta)|$ is smaller than some small tolerance.

All statistics quoted in this section can have numerical problems for boundary data sets, ie, when there are all successes or all failures in either arm. This is handled numerically by adding or subtracting a small number ϵ when a count y_j equals 0 or n_j, respectively. All statistics are available in closed form and can be computed almost instantly. Indeed, this is a practical computational requirement for the unconditional exact tests that we describe later.

3 | WHY AIM FOR EXACT *P* VALUES?

In this section, we provide some arguments in favour of using *P* values for tests of statistical hypotheses in clinical trials. Exact testing is today computationally feasible and keeps type 1 error under strict control.

3.1 | Why *P* values?

All tests are based on *P* values, and biostatisticians seem to agree that a *P* value is a better data summary than just a test decision at a nominal significance level. However, over the last 20 years, some have asserted the primacy of confidence intervals over *P* values in applied medical research.[11] The reason for this change of paradigm is mainly related to the fact that *P* values were frequently misunderstood and misinterpreted by practitioners. Common errors include the so-called principles of logical fallacy[12] and replication fallacy.[13]

There has been a lot of work aimed at clarifying the pitfalls and shortcomings of *P* values (see Wellek[14]), culminating in the American Statistical Association (ASA) statement[15] in 2016. We have no objections to supplementing *P* values with other measures, such as confidence intervals, effect size, or even Bayes factors. However, *P* values remain a fundamental tool of statistical inference, and they are used everyday by practitioners. They measure something important and interpretable, namely, how unusual the results are under the null hypothesis that we are evaluating whether to reject. Moreover, accurate *P* values are central to proper control of type 1 error, which is a key concern of regulators.

3.2 | Why exact?

For clinical trials enrolling thousands of participants, approximate tests often, but not always, control type 1 error well, especially if the trial is balanced. One example is the PROactive study (NCT00174993), a large trial to test the efficacy and safety of pioglitazone, a new second-line therapy for type 2 diabetes. For this trial, 5238 patients were enrolled in the study.[16] However, in most medical fields, trials are typically much smaller, for instance, those associated with the development and approval of new drugs in clinical oncology. There are trials involving patients with cancer in advanced stage of disease, where the total sample size is of few hundreds, or in some cases smaller than 100.

It is simply a fact that the standard approximate tests do not control type 1 error for such sample sizes, or even for much larger sample sizes when the design is unbalanced. At the same time, regulators demand control of type 1 error. For this reason, in this paper, we will put an emphasis on the calculation of *P* values that are as close to exact as practically possible.

3.3 | How much does it matter?

Until fairly recently, exact *P* values were computationally prohibitive, and so the approximate statistics listed in Section 2.2 were used, most commonly the score statistic. Today, computation for sample sizes up to 1000 is no longer an issue; we can calculate an exact *P* value as well as exact attained size and power.

Consider, for instance, a trial where either docetaxel ($n_0 = 166$) or pemetrexed ($n_1 = 304$) was prescribed in patients with recurrent or progressive non-small cell lung cancer.[17] Common statistical training would have practitioners believe that these sample sizes are large enough for the accurate application of the normal approximation. For illustrative purposes, let us take a margin of $\delta = -0.05$. The estimated success rates of $\hat{p}_0 = 11/166 = 6.6\%$ and $\hat{p}_1 = 13/304 = 4.3\%$ do not provide much evidence against the null hypothesis, so let us suppose a more interesting outcome, namely, that $\hat{p}_1 = 22/304 = 7.2\%$. Using risk difference as our comparison measure (ie, with G the identity), the SRLR statistic equals 2.119 and the *P* value based on the standard normal approximation is 0.0170. We would hope and expect that the probability of obtaining a value as large as 2.119 under the null hypothesis (which measures the evidence against inferiority of the treatment) would be 0.0170.

The left hand panel of Figure 1 plots this probability as a function of the unknown probability p_0, under the null assumption that $p_1 - p_0 = -0.05$. We will later define this function formally as the *profile* of the observed *P* value. In simple words, this is the true significance of the observed test statistic. The horizontal line gives the quoted *P* value 0.0170. The maximum value is 0.0315, marked by a blue point, which is almost double the quoted value. We will point out later that, starting with the SRLR statistic, a frequentist has no choice but to quote this maximum value.

The vertical line is the restricted MLE $\tilde{p}_{0,-.05} = 0.109$ of p_0 under the boundary hypothesis that $p_0 = p_1 + 0.05$. It would seem reasonable to use this estimate in place of the unknown p_0, which gives a significance value of 0.0194. How

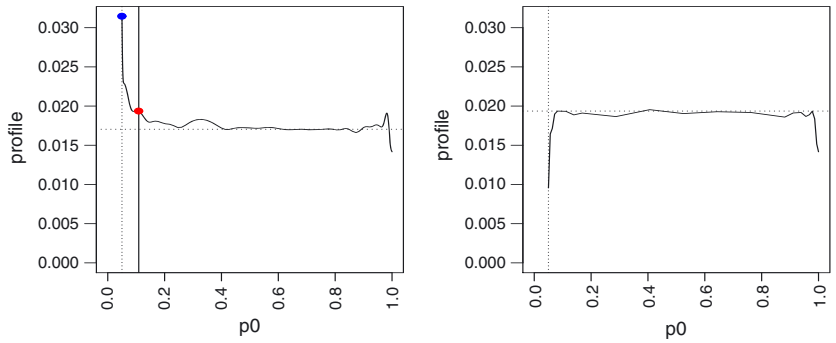

FIGURE 1 Heuristic illustration. *Left.* Exact significance of outcome $11/166 = 6.62\%$ for control and $22/304 = 7.23\%$ for treatment when $\delta = -0.05$ using the signed root likelihood ratio (SRLR) statistic. Horizontal line is quoted significance; blue point is maximum significance; red point is the estimated significance at $\tilde{p}_{0\delta} = 0.109$. *Right.* Exact significance profile of estimated P value, later denoted p_E in Equation 10

can a frequentist justify this though? This "estimated" P value defines a new statistic, more complex than the original score statistic. This statistic has its own profile function, which is displayed in the right panel. Apparently, this profile is extremely close to the quoted value of 0.0194. This is not specific to this example, and much numerical work (and some theoretical work) has shown that estimated P values always perform this well.[18]

4 | THEORY OF EXACT *P* VALUES FOR DISCRETE DATA

We assume that readers are already familiar with basic testing theory. We sketch here the theory of exact tests of a parameter Δ in the presence of a nuisance parameter p_0 from data $Y = (Y_1, Y_0)$. This theory is quite general, and both nuisance parameter and data can be higher dimensional, albeit with computational challenges.

All standard tests can be expressed in terms of a test statistic $T(Y)$, large enough values of which lead to rejection of the null hypothesis. This typically involves using an approximation F to the distribution of $T(Y)$ under the null. We then reject the null if the P value $P(Y) = 1 - F(T(Y))$ is less than a prespecified nominal test size (or significance level) α. Define the probability of rejection of the null hypothesis by

$$\beta(\Delta, p_0) = \Pr(P(Y) \leq \alpha; \Delta, p_0).$$

The null hypothesis specifies that $\Delta \leq \delta$ but does not specify p_0. It will be convenient in what follows to refer to the set of parameter values (Δ, p_0) consistent with the null hypothesis as Ω_0. When G is the identity so that Δ is the risk difference, this is the set $\{\Delta \leq \delta, \max(0, -\delta) \leq p_0 \leq \min(1, 1 - \delta)\}$. When G is the logarithmic function so that Δ is the logged risk ratio, this is the set $\{\Delta \leq \delta, 0 \leq p_0 \leq 1\}$. The size of the test is usually defined to be $\sup\{\beta(\Delta, p_0) : (\Delta, p_0) \in \Omega_0\}$, which is the worst case probability of falsely rejecting the null hypothesis. Subject to the size being controlled at the target level α, we want the test with maximum power.

There are two key observations to make when the data are discrete. First, there will typically not exist any test with size exactly equal to a prespecified α. The best you can do is to ensure that the size is less than α. Second, amongst tests with size less than α there is typically no unrandomised test that is most powerful. These general assertions apply in particular to noninferiority tests based on binomial data. However, it turns out that exact and optimal P values, appropriately defined, do exist.

4.1 | Validity of a *P* value

Ideally, a P value $P(Y)$ should have the uniform distribution under the null hypothesis. In this case, the implied test is exact at any specified nominal size α. We have just asserted that exact tests typically do not exist for discrete models. When Y is discrete, the analogous property is that $\Pr(P(Y) \leq \pi_k) = \pi_k$ for all observable values π_k of $P(Y)$. This is the basis of the interpretation of a P value: If we observe a P value of 0.046, then we have observed an event whose probability is 0.046. However, if we still have free nuisance parameter(s) p_0, then these probabilities are functions of p_0 and cannot equal a constant. So apart from trivial cases where the distribution of $P(Y)$ does not depend on nuisance parameters, uniformity is not achievable.

A P value is called *valid* if

$$\sup_{(\Delta,p_0)\in\Omega_0} \Pr(P(Y) \le \pi_k; \Delta, p_0) \le \pi_k, \tag{6}$$

again for all observable π_k. So the probability of observing a P value as say 0.046 is at most 0.046. Technically, $P(Y)$ is stochastically larger than uniform under the null hypothesis. It is simple to show that a valid P value leads to a valid test, ie, a test of size less than or equal to α, for any prespecified α.

4.2 | The profile of a *P* value

For any fixed value of the nuisance parameter p_0, we can calculate the probability of rejection under the null hypothesis that $\Delta \le \delta$. This gives a function of p_0, which we call the *profile* of the observed P value:

$$\Pi_P(y, p_0) = \sup_{\Delta\le\delta} \Pr(P(Y) \le P(y); \Delta, p_0) \tag{7}$$

defined over the range $p_0 : (\delta, p_0) \in \Omega_0$. We have slightly generalised the notation by replacing π_k by $P(y)$. The profile gives the weight of evidence against the null as a function of the nuisance parameter p_0. So one use of the profile is as a summary of the evidence. Figure 1 already displayed two profiles, one for the SRLR statistic for testing risk difference and one for the bootstrap test to be defined below.

4.3 | Exact *P* values

A P value is called *exact*[18] if

$$\sup_{(\Delta,p_0)\in\Omega_0} \Pr(P(Y) \le \pi_k; \Delta, p_0) = \pi_k \tag{8}$$

for all observable π_k. For a prespecified nominal size that happens to equal an observable value π_k, the test will be exact. However, for all other values of α, the test will only be valid. This seems a reasonable definition of an exact P value.

With this definition, an exact P value can always be constructed from an inexact one $P(y)$ from its profile Π_P. Define

$$P_M(y) = \sup_{p_0} \Pi_P(y, p_0). \tag{9}$$

Then it can be shown[18] that $P_M(Y)$ is exact and that it is the smallest valid P value that is a nondecreasing function of the original P value $P(Y)$. It can also be shown that any exact statistic must be expressible in the form (9) for some initial P value $P(Y)$.

4.4 | Monotonicity properties

The parameter of interest $\Delta = G(p_1) - G(p_0)$ is an increasing function of p_1 and a decreasing function of p_0. It would seem logically essential then that any estimate, confidence limit, or test statistic should be an increasing function of y_1 and decreasing function of y_0. Slightly less transparent conditions are that the evidence for a higher value of Δ should be increasing in y_1 for fixed $s = y_1 + y_0$. Rather unhelpfully, conditions of this type are called convexity[19,20] because tail sets of the form $\{T(Y_1, Y_0) \ge T(y_1, y_0)\}$ are then convex.

Besides the basic logical property, there are several practical consequences of convexity. First, the size of the test that is the supremum of $\beta(\Delta, p_0)$ over $(\Delta, p_0) \in \Omega_0$ is achieved at the boundary $\Delta = \delta$, which saves a full search over the set Ω_0.[21] The same can be said for the supremum over Ω_0 in the profile (7). Convexity also ensures that the set $\{P(Y) \le P(y)\}$ in Π_P can be determined by bisection,[18] without searching the entire sample space.

The general statistic (3) is necessarily convex. However, when it is divided by its standard error the situation is less clear. When the denominator is estimated with unrestricted ML estimates, convexity is almost always violated. When the denominator is estimated with restricted ML estimates, convexity does hold for the cases G=H=id or G=log and $H = G^{-1}$convex,[22] which correspond to the score statistic. The SRLR satisfies convexity numerically[2] although no proof has appeared.

5 | EXACT UNCONDITIONAL *P* VALUES

5.1 | Full maximisation

The key result from the previous section is that an approximate P value can always be adjusted into an exact P value. This requires calculating its significance profile $\Pi_P(y, p_0)$ and then quoting the maximum. This is called the M-step and the P value is termed maximised. The initial choice of approximate P value only affects the profile through the tail set $\{P(Y) \leq P(y)\}$ and so only depends on how $P(Y)$ orders the sample space.

There are two problems with maximised P values. The first is that they can lead to low power because they account for an unlikely worst case possibility. The second is that the value depends quite critically on the initial approximate P value on which it is based. Thus, in a given example not only may the score and LR statistics give practically different results but the maximised versions will also give practically different results, even though both are ostensibly exact.

5.2 | Restricted maximisation

The value of p_0 that corresponds to the maximum can easily be a value that is very unlikely, given the data. Berger and Boos[23] suggested the P value

$$p_B(y) = \sup_{p_0 : (p_0, \Delta) \in \Omega_0, p_0 \in C_\gamma} \Pi_P(y, p_0) + \gamma,$$

where C_γ is an exact $1 - \gamma$ confidence interval for p_0 under the null hypothesis, and proved that this generates a valid test for any nominal α and chosen γ. A common recommendation is $\gamma = 0.001$.

The partially maximised P value may not satisfy Barnard's convexity properties[21] but does mitigate the problems of the fully maximised P value: It loses less power and it is less dependent on the initial approximate P value. It does have two additional problems of its own however. First, it is not exact in the sense of (8) because it is not the maximised value of any profile function. This results in some size conservatism, although this is usually very small. Second, it is not always easy to give an exact confidence interval C_γ for p_0. Our noninferiority hypothesis is a case in point. Third, results depend on choice of the confidence region C_γ, both its coverage and, in models where the nuisance parameter has dimension greater than 1, its shape. Finally, other researchers have found that the efficiency gains are modest.[24] For these reasons, we do not include $p_B(y)$ in our later numerical study.

5.3 | Estimation/bootstrap

Replacing an unknown parameter by its estimate is a very old idea. Indeed, we could not calculate standard errors without this device. This suggests the P value

$$p_E(y) = \Pi_P(y, \tilde{p}_{0\delta}). \tag{10}$$

This has been recommended by several authors[25,26] including for noninferiority trials.[27] It is important that the estimate be calculated under the null hypothesis.[18] One can think of this as a parametric bootstrap, even though no simulation is involved. The fact that the estimate $\tilde{p}_{0\delta}$ is inserted into a P value that is approximately pivotal is analogous to prepivoting, as proposed by Beran[28] for continuous models.

The advantages of p_E is that it hardly depends at all on the choice of the initial P value and that it is much faster to compute than p_M or p_B since no search over p_0 is required. This is particularly pertinent for models where there are more than one nuisance parameter. For discrete models, the error in using p_E is $O(n^{-1})$.[9]

There are two key problems with $p_E(Y)$. First, while its error is $O(n^{-1})$,[9] it is not exact according to (8), despite the title of this section. However, it is much closer to exact than asymptotics suggest.[29] For this reason, it can be referred to as *quasi exact* by which we mean exact for all practical purposes. The tiny levels of inexactness in P_E can be corrected with a further maximisation step,[2,5] resulting in an E+M P value.[18] This typically makes little difference in practice because the profile is already so flat, see the right panel of Figure 1 and Ripamonti et al[30] for an exhaustive numerical assessment.

The second problem is that $p_E(Y)$ is not necessarily convex. Empirically, this is quite rare and occurs for extreme data sets that has no effect on the achieved size of the test. Application of the M-step cannot correct this since the M-step preserves the ordering of the sample space.

6 | NUMERICAL STUDY

In this section, we report the results of a numerical study on the power and size control of the various tests described above. For the risk difference, we will only consider the canonical cases of $G = H = id$ where the pooled statistic is the score statistic as noted in Section 2.2. Likewise, for the risk ratio test, we take $H = G^{-1}$ where again the pooled statistic is the score statistic. For each model then, we have four approximate tests, namely, unpooled, pooled/score, LR, and r^* (the P value based on the r^* statistic is indicated as p^*).

We include three adjustments of these raw statistics: maximisation, estimation, and estimation followed by maximisation. We did not evaluate the restricted maximisation procedure since it is not clear how to construct an honest interval for p_0 under the null when $\delta \neq 0$. We note however that this method has been shown to perform quite well when $\delta = 0$, although not as well as estimation followed by maximisation.[30] So there are 16 different tests (ie, P values) in all. Some of these tests have been considered in previous literature, namely, the maximised version of the score statistic[31] and the E+M version of the SRLR statistic.[2]

For the risk difference, we look at values $\delta = 0, -0.05, -0.10$ and for risk ratio $\delta = 0, -0.15, -0.3$, based on recommendations in Wangge et al.[32] In all cases, we chose $\alpha = 0.05$. The sample size chosen for (n_0, n_1) were (42, 38), (60, 40), (50, 70), (80, 70), and (100, 50), which covers balanced and unbalanced cases with totals ranging from 80 to 150. Note that when $\delta = 0$, the results are invariant to swapping the sample sizes.

To calculate the size of these tests, we computed $\Pr(P(Y) \leq \alpha; \Delta, p_0)$ at the boundary value $\Delta = \delta$, which leaves p_0 free. We used two measures of size. The first maximises over p_0 (see Table 1, upper section) and the second is the simple average respect to p_0 across an even grid of values (lower section). It was mentioned in Section 4.4 that if the tail set generated by P value is not convex, then the maximum probability of rejection is not necessarily obtained at the boundary. So, for the unpooled and r^* statistics, the actual size could possibly be larger than the value quoted.

Full results are in Supporting Information. We present typical size results in Tables 1 and 2 for the risk difference with $(n_0, n_1) = (60, 40)$ and risk ratio with $(n_0, n_1) = (42, 38)$, respectively. The upper sections show maximum size and are colour coded in red if it exceeds 0.05. Clearly, none of the raw tests control size although the score comes closest to respecting the 0.05 limit, consistent with the folklore. The M and E+M tests necessarily control size where as the E test is not guaranteed to do so. However, in this and other cases, the occasional violations are very small in magnitude. It is also noteworthy that the maximum size of the M tests can be well below nominal, indicating some conservatism. This is caused by having to account for the worst case.

Looking at average size in the lower section of the table, for those tests that control maximum size, we colour code larger values green and smaller as red, the idea being that if the maximum size is less than nominal, then average size should be as close to nominal as possible to minimise conservatism. Clearly, the M test imposes much more conservatism than

TABLE 1 Size control of 16 tests of risk difference for $(n_0, n_1) = (60, 40)$. *Top section*: maximum size over p_0; *lower section*: average size with respect to p_0

	n0=60, n1=40, δ=0				n0=60, n1=40, δ=−0.05				n0=60, n1=40, δ=−0.10			
TYPE	Unpool	Score	LR	p*	Unpool	Score	LR	p*	Unpool	Score	LR	p*
raw	0.088	0.066	0.088	0.088	0.192	0.057	0.129	0.129	0.137	0.068	0.080	0.080
M	0.048	0.049	0.048	0.046	0.046	0.049	0.046	0.046	0.031	0.047	0.032	0.035
E	0.051	0.050	0.050	0.050	0.051	0.049	0.051	0.049	0.050	0.050	0.050	0.050
E+M	0.050	0.050	0.050	0.050	0.049	0.049	0.049	0.049	0.050	0.050	0.050	0.050
raw	0.052	0.051	0.053	0.052	0.060	0.049	0.053	0.052	0.057	0.052	0.055	0.055
M	0.030	0.037	0.029	0.027	0.001	0.044	0.023	0.023	0.003	0.042	0.026	0.027
E	0.045	0.044	0.043	0.043	0.045	0.046	0.046	0.046	0.045	0.046	0.046	0.046
E+M	0.044	0.044	0.043	0.043	0.045	0.046	0.046	0.046	0.044	0.045	0.045	0.045

TABLE 2 Size control of 16 tests of risk ratio for $(n_0, n_1) = (42, 38)$. *top section*: maximum size over p_0; *lower section*: average size with respect to p_0

	n0=42, n1=38, δ=0				n0=42, n1=38, δ=−0.05				n0=42, n1=38, δ=−0.10			
TYPE	Unpool	Score	LR	p*	Unpool	Score	LR	p*	Unpool	Score	LR	p*
raw	0.074	0.055	0.091	0.090	0.142	0.055	0.142	0.142	0.112	0.068	0.095	0.095
M	0.047	0.049	0.037	0.037	0.000	0.050	0.038	0.037	0.017	0.049	0.037	0.037
E	0.050	0.050	0.052	0.051	0.050	0.050	0.050	0.050	0.049	0.050	0.049	0.049
E+M	0.050	0.050	0.049	0.048	0.050	0.050	0.050	0.050	0.049	0.050	0.049	0.049
raw	0.051	0.048	0.054	0.050	0.133	0.055	0.132	0.132	0.056	0.051	0.053	0.055
M	0.038	0.041	0.024	0.023	0.000	0.050	0.038	0.037	0.002	0.042	0.029	0.030
E	0.044	0.044	0.045	0.045	0.050	0.050	0.050	0.050	0.044	0.046	0.045	0.045
E+M	0.044	0.044	0.043	0.043	0.050	0.050	0.050	0.050	0.044	0.046	0.045	0.045

TABLE 3 Power control of 16 tests of risk difference for $(n_0, n_1) = (60, 40)$

			n0=60, n1=40, δ=0				n0=60, n1=40, δ=−0.05				n0=60, n1=40, δ=−0.10			
P_0	P_1	TYPE	Unpool	Score	LR	p*	Unpool	Score	LR	p*	Unpool	Score	LR	p*
0.1	0.22	raw	0.603	0.664	0.664	0.664	0.837	0.846	0.840	0.840	0.953	0.965	0.965	0.965
		M	0.510	0.583	0.556	0.556	0.000	0.840	0.758	0.758	0.574	0.947	0.921	0.938
		E	0.627	0.628	0.627	0.627	0.840	0.840	0.840	0.840	0.953	0.953	0.953	0.953
		E+M	0.628	0.628	0.627	0.627	0.840	0.840	0.840	0.840	0.953	0.953	0.953	0.953
0.5	0.65	raw	0.663	0.641	0.651	0.641	0.819	0.811	0.811	0.811	0.919	0.919	0.914	0.914
		M	0.566	0.602	0.538	0.526	0.000	0.795	0.689	0.688	0.529	0.907	0.853	0.853
		E	0.629	0.639	0.629	0.629	0.801	0.801	0.801	0.801	0.909	0.914	0.914	0.914
		E+M	0.629	0.639	0.629	0.629	0.801	0.801	0.801	0.801	0.909	0.914	0.914	0.914
0.7	0.85	raw	0.677	0.657	0.657	0.657	0.864	0.836	0.847	0.836	0.949	0.944	0.947	0.947
		M	0.589	0.593	0.567	0.521	0.000	0.824	0.721	0.718	0.648	0.932	0.899	0.899
		E	0.638	0.644	0.638	0.646	0.827	0.827	0.827	0.827	0.938	0.943	0.943	0.943
		E+M	0.638	0.644	0.638	0.646	0.827	0.827	0.827	0.827	0.938	0.938	0.938	0.938

TABLE 4 Power control of 16 tests of risk ratio for $(n_0, n_1) = (42, 38)$

			n0=42, n1=38, δ=0				n0=42, n1=38, δ=−0.05				n0=42, n1=38, δ=−0.10			
P_0	P_1	TYPE	Unpool	Score	LR	p*	Unpool	Score	LR	p*	Unpool	Score	LR	p*
0.1	0.25	raw	0.574	0.574	0.583	0.550	0.772	0.772	0.790	0.772	0.913	0.913	0.913	0.922
		M	0.519	0.519	0.442	0.412	0.000	0.747	0.629	0.618	0.418	0.900	0.863	0.863
		E	0.550	0.550	0.550	0.550	0.770	0.772	0.772	0.772	0.900	0.912	0.900	0.900
		E+M	0.550	0.550	0.550	0.550	0.770	0.749	0.749	0.749	0.900	0.912	0.900	0.900
0.5	0.7	raw	0.590	0.577	0.577	0.577	0.756	0.745	0.745	0.745	0.869	0.869	0.869	0.869
		M	0.539	0.540	0.440	0.439	0.000	0.714	0.596	0.590	0.354	0.848	0.793	0.793
		E	0.574	0.574	0.574	0.574	0.729	0.729	0.729	0.729	0.858	0.858	0.858	0.858
		E+M	0.574	0.574	0.557	0.557	0.729	0.729	0.729	0.729	0.858	0.858	0.858	0.858
0.7	0.87	raw	0.598	0.597	0.598	0.597	0.793	0.793	0.793	0.793	0.916	0.912	0.916	0.916
		M	0.548	0.548	0.455	0.455	0.000	0.756	0.644	0.636	0.456	0.884	0.850	0.850
		E	0.568	0.568	0.568	0.568	0.764	0.764	0.764	0.764	0.900	0.900	0.900	0.900
		E+M	0.568	0.568	0.568	0.568	0.764	0.764	0.764	0.764	0.900	0.900	0.900	0.900

the other methods. This is because the profiles of the P values have spikes, typically near the boundary. For the unpooled statistic, this leads to a size profile that is extremely conservative. The E and E+M tests have the highest average size that, we are about to see, translates into higher power.

To calculate the power, we selected three values of p_0 (0.10, 0.50, 0.70) and then values of p_1 that would generate powers in a reasonable range for the three values of δ. Results for the two models just considered above are in Tables 3 and 4. Larger/smaller values are coded greener/redder. There are three general patterns to identify, which also hold for the other cases listed in Supporting Information. First, the E and E+M tests have higher power than the M test (while controlling size as shown above). Second, the E and E+M tests have almost identical power. This is because the E and E+M P values are almost identical as a consequence of the former having very flat profiles, one example of which we saw earlier in Figure 1. Thirdly, when the E or E+M test is used, it makes little difference which raw statistic is used to generate it. This is in contrast to traditional asymptotic methods where the choice does make a difference and can potentially be used by unethical analysts to achieve a desired result.

Our general conclusion will be given in Section 8.

7 | ILLUSTRATIVE EXAMPLE

We illustrate the methods explained in this article using data from the clinical trial by Kim et al,[17] already mentioned in Section 3.3. This was a multicentre, randomised phase 3 trial in patients with recurrent progressive non-small cell lung cancer, after failure of initial platinum-based chemotherapy. Pemetrexed (experimental drug) was compared with docetaxel (active comparator), and various outcomes were considered, amongst which the proportion of patients free of serious adverse events. We wish to test noninferiority on the risk difference scale. This is a particularly interesting measure in noninferiority trials, since, provided reasonably comparable efficacy, the new treatment could be preferred in terms of lower toxicity and better safety profile.

It emerges from Table 5 that raw P values are smaller than M P values, but the former are not guaranteed and very unstable across test statistics. Both M and E P values are identical for different test statistics, but E P values are slightly less conservative. Ideally, we would recommend the E+M test based on the LR or score statistics (since these statistics satisfies the monotone property). However, the E procedure is much quicker to compute, gives very similar results here, and is quasi exact according to our reported numerical study.

TABLE 5 *P* values calculated with four methods and 4 test statistics for the illustrative example, with $\delta = -0.05$. There were $y_0 = 11$ successes from $n_0 = 166$ for the active comparator group and $y_1 = 13$ from $n_1 = 304$ for the treatment group

TYPE	Score	Wald	LR	r*
raw	0.139	0.119	0.135	0.133
M	0.165	0.165	0.165	0.165
E	0.143	0.143	0.143	0.143
E+M	0.147	0.147	0.147	0.147

8 | CONCLUSIONS

Clinical trials are the fundamental method of securing evidence on the effectiveness of new medicines. Oft-used non-inferiority trials, where the effect of a new drug is contrasted with that of an active comparator, need to be assessed by contemporary statistical methods. The most relevant trials to the discussion of this article are phase 3 clinical trials, where sample sizes might be under 300 participants, or even under 100, as in case of diseases at very advanced stage. Exact methods are needed in these cases.

Conditioning is one classical approach to the analysis of the 2×2 table; however, this approach is not only unconditionally conservative but is not available when the margin δ is nonzero, as in noninferiority trials. How should we proceed in these cases? One possibility is to use approximate conditioning, which we have included as one of our approximate methods, as well as the three approximated methods of Wald, score, and LR. None of these tests control unconditional type 1 error at an acceptable level.

Maximisation and restricted maximisation both ensure valid tests. Full maximisation is conservative as a consequence of allowing for possibly unlikely worst case values of the nuisance parameter. We did not investigate restricted maximisation in the numerical study since the confidence interval for the nuisance parameter cannot be guaranteed and also because other larger numerical studies[30,33] have shown that it performs somewhere between estimation and maximisation. Estimation/bootstrap produces tests that are not guaranteed but are extremely close to exact in practice. There are large computational savings, and the resulting test is consistently more powerful than maximisation. Moreover, the final *P* value is easy to explain to practitioners, especially if it is based on a familiar test statistic such as the score test.

On the basis of our numerical study as well as theoretical considerations, we recommend beginning with either the SRLR or score statistic and adjusting with estimation/bootstrap. Results using the two initial statistics are barely distinguishable, so we recommend starting with the score statistic, this being the better known of the two. A further M-step can be applied but typically has a negligible effect on the reported *P* value for considerable extra computation. R code for implementing the score test with E step applied is available as Supplementary Information. Code for the E+M *P* value is available from the authors.

ORCID

Enrico Ripamonti https://orcid.org/0000-0002-0584-8401

REFERENCES

1. D'Agostino RB, Massaro J, Sullivan LM. Non-inferiority trials: design, concepts and issues—the encounters of academic consultants in statistics. *Stat Med*. 2003;22:169-186.
2. Munk A, Skipka G, Stratmann B. Testing general hypotheses under binomial sampling: the two sample case—asymptotic theory and exact procedures. *Comput Stat Data An*. 2005;49:723-739.
3. Dunnett CW, Gent M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics*. 1977;33:593-602.
4. Miettinen O, Nurminen M. Comparative analysis of two rates. *Stat Med*. 1985;4:213-226.
5. Martín Andrés A, Silva Mato A. Choosing the optimal unconditioned test for comparing two independent proportions. *Comput Stat Data An*. 1994;17:555-574.
6. Johnson NL, Kotz S, Kemp AW. *Univariate Discrete Distributions*. New York: Wiley; 2013.
7. Mee RW. Confidence bounds for the difference between two probabilities. *Biometrics*. 1984;40:1175-1176.

8. Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics*. 1984;40:513-.517.
9. Davison AC, Fraser DAS, Reid N. Improved likelihood inference for discrete data. *J R Stat Soc B Met*. 2003;68:495-508.
10. Reid N. Asymptotics and the theory of inference. *Ann Stat*. 2003;31:1695-1731.
11. Altman D, Machin D, Bryant T, Gardner M. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. New York: John Wiley & Sons; 2013.
12. Poole C. Low p-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291-294.
13. Gigerenzer G. Mindless statistics. *J Socio-Econ*. 2004;33:587-606.
14. Wellek S. A critical evaluation of the current p-value controversy. *Biometrical J*. 2016;59:854-872.
15. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129-133.
16. Dormandy JA, Charbonnel B, Eckland DJ, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the proactive study (prospective pioglitazone clinical trial in macrovascular events): a randomised controlled trial. *Lancet*. 2005;366:1279-1289.
17. Kim ES, Neubauer M, Cohn A, et al. Docetaxel or pemetrexed with or without cetuximab in recurrent or progressive non-small-cell lung cancer after platinum-based therapy: a phase 3, open-label, randomised trial. *Lancet Oncol*. 2013;14:1326-1336.
18. Lloyd CJ. Exact p-values for discrete models obtained by estimation and maximization. *Aust NZ J Stat*. 2008;50:329-345.
19. Barnard GA. A new test for 2 × 2 tables. *Nature*. 1945;156:177.
20. Almendra-Arao F, Sotre-Ramos D, Zuñiga-Estrada M. Extending the barnard's test to non-inferiority. *Comm Stat A Theory*. 2017;46:6293-6302.
21. Röhmel J, Mansmann U. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical J*. 1999;41:149-170.
22. Röhmel J. 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 J*. 2005;47:37-47.
23. Berger RL, Boos DD. P-values maximized over a confidence set for the nuisance parameter. *J Am Stat Assoc*. 1994;89:1012-1016.
24. Chan ISF, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics*. 1999;55:1202-1209.
25. Liddell.D. Practical test of 2 × 2 tables. *The Statistician*. 1976;25:295-304.
26. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. *J Am Stat Ass*. 1990;85:146-155.
27. Kang SH, Chen JJ. An approximate unconditional test of non-inferiority between two proportions. *Stat Med*. 2000;19:2089-2100.
28. Beran R. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J Am Stat Assoc*. 1988;83:687-697.
29. Lloyd CJ. A numerical investigation of the accuracy of parametric bootstrap for discrete data. *Comput Stat Data An*. 2013;61:1-6.
30. Ripamonti E., Lloyd C, Quatto P. Contemporary frequentist views on the 2 × 2 binomial trial. *Stat Sci*. 2017;32:600-615.
31. Chan ISF. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Stat Med*. 1998;17:1403-1413.
32. Wangge G, Putzeist M, Knol MJ, et al. Regulatory scientific advice on non-inferiority drug trials. *PloS One*. 2013;8:e74818.
33. Lloyd CJ, Ripamonti E. An exhaustive numerical assessment of alternative unconditional tests of a binary treatment effect. *J Stat Comput Simul*. 2018;88:2150-2169.

SUPPORTING INFORMATION

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

How to cite this article: Ripamonti E, Lloyd CJ. Tests for noninferiority trials with binomial endpoints: A guide to modern and quasi-exact methods for biomedical researchers. *Pharmaceutical Statistics*. 2019;1–11. https://doi.org/10.1002/pst.1929