Statistical Science
2017, Vol. 32, No. 4, 600–615
DOI: 10.1214/17-STS627

Contemporary Frequentist Views of the 2×2 Binomial Trial

Enrico Ripamonti, Chris Lloyd and Piero Quatto

Abstract. The 2×2 table is the simplest of data structures yet it is of immense practical importance. It is also just complex enough to provide a theoretical testing ground for general frequentist methods. Yet after 70 years of debate, its correct analysis is still not settled. Rather than recount the entire history, our review is motivated by contemporary developments in likelihood and testing theory as well as computational advances. We will look at both conditional and unconditional tests. Within the conditional framework, we explain the relationship of Fisher's test with variants such as mid-p and Liebermeister's test, as well as modern developments in likelihood theory, such as p^* and approximate conditioning. Within an unconditional framework, we consider four modern methods of correcting approximate tests to properly control size by accounting for the unknown value of the nuisance parameter: maximisation (M), partial maximisation (B), estimation (E) and estimation followed by maximisation (E + M). Under the conditional model, we recommend Fisher's test. For the unconditional model, amongst standard approximate methods, Liebermeister's tests come closest to controlling size. However, our best recommendation is the E procedure applied to the signed root likelihood statistic, as this performs very well in terms of size and power and is easily computed. We support our assertions with a numerical study.

Key words and phrases: Approximate conditioning, binomial trial, conditional test, exact tests, Fisher test, Liebermeister test, mid-p test, parametric bootstrap, unconditional test.

1. INTRODUCTION

Testing for a treatment effect in the 2×2 binomial trial is a seminal topic in Statistics, founded on the original contributions of renowned statisticians, Karl Pearson, Jerzy Neyman and Ronald Fisher. Not only is the design of great practical importance, notably in clinical trials, but the admittedly simple model is still rich enough to expose some of the tensions and limitations of frequentist statistics. Consequently, there are now literally dozens of test procedures that have been proposed. Reviews have appeared at regular intervals; see Gart (1969), Yates (1984), Martin Andres (1991), Agresti (1992, 2001) and Lydersen, Fagerland and Laake (2009). Why is it worth discussing tests for 2×2 tables at all and why again now?

First, the 2×2 table is the basic data structure in clinical trials with binary outcome. Modern adaptive designs allow changes to treatments, sample sizes and even hypotheses, but their analysis relies on combining the evidence from different arms, stages and hypotheses. The p-values from each 2×2 table are fed into a more complex p-value (using combination functions, multiple comparison adjustments and the closed testing principle) whose statistical properties are inherited

Enrico Ripamonti is Postdoc Researcher, Department of Economics, Management, and Statistics, University of Milan-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126, Milan, Italy (e-mail: enrico.ripamonti@unimib.it). Chris Lloyd is Professor, Melbourne Business School, The University of Melbourne, 200 Leicester Street, Carlton Victoria 3053, Australia (e-mail: c.lloyd@mbs.edu). Piero Quatto is Professor, Department of Economics, Management, and Statistics, University of Milan-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126, Milan, Italy (e-mail: piero.quatto@unimib.it).

from the component p-values. So, while the 2×2 table might appear a “toy example”, it is the building block of various modern methods/designs.

Second, over the past ten years, there has been considerable progress in the foundational theory of exact or almost exact frequentist inference as well as methods of implementation. Some of these methods require complex computations. Amongst unconditional methods, there are several ways of correcting an approximate test to be exact; see Lloyd (2008b). For conditional methods, the seminal test of Fisher (1935) is limited by discreteness which can be mitigated by the well-known mid-p correction (Lancaster, 1961). Further developments include so-called approximative conditioning (Pierce and Peters, 1992, 1999) as well as the famous p^* formula (Barndorff-Nielsen, 1983).

The objective of this paper is to present a detailed discussion of tests based on the 2×2 binomial trial with an emphasis on more contemporary theories and proposals. We include unconditional and conditional perspectives without taking a definite position on which is better. We do not study Bayesian methods, even though one of the methods we include (Liebermeister, 1877) was originally motivated from a Bayesian approach. Our overall aim is to place the different methods within a coherent framework, to assess and compare their properties, both theoretically and numerically and to arrive at clear recommendations.

In assessing the tests, we focus on four main criteria. First, we require the test to be based on a p-value which measures, possibly approximately, the probability of some observed event. Second, does the test exaggerate the evidence against the null? This is based on comparing the nominal size with the actual *size profile* but also the quoted p-value with its true *profile* (see Lloyd, 2008a). Third, we look at the extent to which the test under-estimates the evidence against the null, commonly called conservatism. Conservative tests tend to have lower power and we confirm this with a numerical study. Lastly, we impose certain natural monotonicity constraints (Barnard, 1947, Röhmel and Mansmann, 1999, Skipka, Munk and Freitag, 2004) on the test statistics. These constraints also have favourable computational implications.

The plan of the article is as follows. In the next section, we establish the basic model, the notation and the theoretical framework for assessing the different methods. In Section 3, we present the conditional approach and in Section 4 the unconditional approach, in both cases emphasising modern perspectives and developments. In Section 5, we assess conditional and unconditional tests within their own frameworks and in Section 6 we report the results of a numerical study on the size and power of 28 different unconditional tests, where some clear conclusions do emerge. Our final recommendations are articulated in Section 7.

2. THEORY RELEVANT TO EXACT TESTS

2.1 Notation

We suppose that n_0 patients are given a comparison treatment, y_0 of whom respond positively with probability p_0, and n_1 are given a new treatment, y_1 of whom respond positively with probability p_1. We henceforth call a positive response a success. Provided patients respond independently, we have the standard binomial model

$$(2.1)\qquad \begin{aligned} &Y_0 \sim \mathrm{Bi}(n_0, p_0) \quad \text{and} \\ &Y_1 \sim \mathrm{Bi}(n_1, p_1) \quad \text{with } Y_0 \perp\!\!\!\perp Y_1. \end{aligned}$$

We mainly focus on one-sided hypotheses

$$(2.2)\qquad H_0 : p_1 \le p_0 \quad \text{vs.} \quad H_1 : p_1 > p_0$$

though the theory in this section applies to two-sided tests without modification. We denote the total number of successes by $S = Y_0 + Y_1$ and the proportion of successes under treatment and control as $\hat{p}_1 = y_1/n_1$ and $\hat{p}_0 = y_0/n_0$.

For the sake of giving general definitions and results, we will refer to the data (Y_0, Y_1) as Y, taking values in a sample space $\mathscr{Y}$ and the parameter as $\omega = (\theta, \varphi)$, where θ is the interest parameter and φ a nuisance parameter vector. We wish to test the null hypothesis that $\theta \in \Theta_0$ without specifying the value of the nuisance parameter φ. For the binomial trial, θ can be taken as any contrast of p_1 and p_0 (such as the difference, or the log-odds ratio), the nuisance parameter is $\varphi = p_0$ and for the hypotheses in (2.2) the null parameter space is $\Theta_0 = \{\theta : \theta \le 0\}$.

2.2 Size and Power

All tests can be expressed in the form *reject the null if $P(Y)$ is less than or equal to α*, where α is the nominal size of the test and $P(Y)$ is called a p-value. The probability of rejecting the null hypothesis is

$$(2.3)\qquad \beta(\theta, \varphi) := \Pr\bigl[P(Y) \le \alpha | \theta, \varphi\bigr].$$

The size of the test is typically defined as $a(\varphi) = \sup_{\theta \in \Theta_0} \beta(\theta, \varphi)$ and the test is *valid* if $a(\varphi)$ is less than α for all φ (Lehmann, 1959). The power of the test is the probability of rejecting the null when $\theta \notin \Theta_0$ and is desired to be as large as possible, subject to validity.

2.3 What Is an Exact Test?

Ideally, we would want the size to equal α but for discrete models $a(\varphi)$ is a polynomial and can never equal a constant. If we further maximise with respect to φ, then a bound can be given but again, because of discreteness, this bound almost never equals α exactly. In summary, if we define an exact test to have "exact size α" then such tests almost never exist for discrete models.

For this reason, Lloyd (2008a) instead looks at the p-value, specifically at the so-called *profile* of a p-value which is defined as

$$\pi(y,\varphi) = \sup_{\theta\in\Theta_0} \Pr\bigl[P(Y) \le P(y);\theta,\varphi\bigr]. \tag{2.4}$$

The putative property of a p-value is that an observed value of say 0.042 means that something unusual has happened and the probability of it happening under the null is 0.042. Thus we would like $\pi(y,\varphi)$ to equal $P(y)$ for all φ. Again, because of discreteness this is impossible, so again it appears as if no exact p-value can exist. However, if $\sup_\varphi \pi(y,\varphi) \le P(y)$ for all y in $\mathscr{Y}$ we call the p-value *guaranteed*. This is identical to $P(Y)$ being stochastically no smaller than uniform for all $\theta \in \Theta_0$. It is the analog of a test being valid and a guaranteed p-value does imply a valid test. However, the advantage of basing the theory on p-values rather than test size is that there always exists a p-value for which

$$\sup_{\varphi} \pi(y;\varphi) = P(y) \quad \forall y \in \mathscr{Y} \tag{2.5}$$

as proven by Röhmel and Mansmann (1999). Such a p-value is called *exact*. It is further shown that amongst p-values that impose the same ordering on the sample space there always exists a smallest p-value and that this p-value is exact. The construction of this p-value is simple and will be given in Section 4.1. The theory is completely general and applies to the conditional or unconditional model, as well as to one-sided or two-sided tests.

2.4 Most Powerful Tests

Lehmann (1959) established the existence of both exact and optimal tests, which is relevant to our purposes. The main class of models where a most powerful test exists is the natural exponential family, where the joint density or probability function of the data Y can be written as

$$f_{\theta,\varphi}(y) = \exp\bigl\{\theta T(y) + S'(y)\varphi + \varsigma(\theta,\varphi)\bigr\}, \tag{2.6}$$

where $T(y)$ is a scalar and $S(y)$ is the sufficient statistic for φ. For model (2.6), uniformly most powerful unbiased (UMPU) tests exist for both one and two-sided alternatives. These procedures are based on tail probabilities of the conditional distribution of T given S but their UMPU properties are also unconditional.

For the binomial trial, the model is of exponential form with $\theta = \operatorname{logit}(p_1) - \operatorname{logit}(p_0)$, the statistic $T(Y) = Y_1$ and its distribution given $S = Y_0 + Y_1 = s$ is

$$\begin{aligned}&\Pr(Y_1 = y_1; s, \theta)\\ &\quad = e^{\theta y_1}\binom{n_1}{y_1}\binom{n_0}{s-y_1}\Big/\kappa(\theta,s),\end{aligned} \tag{2.7}$$

where $\max\{0; s-n_0\} \le y_1 \le \min\{s; n_1\}$ and $\kappa(\theta,s)$ is a normalising constant. When $\theta = 0$, the distribution is hypergeometric. However, this model is discrete.

For discrete exponential models, Lehmann's UMPU tests involve randomisation. This also arises in certain nonexponential continuous models, such as uniform when the support depends on the parameter value. In any case, randomisation is never used in practice. The lack of an optimal test explains the many alternative tests of the 2 × 2 table that have been proposed in the literature. Nevertheless, a key insight that comes from the theory is that optimal tests should be based on the conditional distribution of T given S and that for fixed S there is more evidence against the null hypothesis when T is larger.

2.5 Monotonicity Properties

For some models, there are basic logical properties that we expect any procedure to have. For instance, any statistical procedure for assessing reliability should produce a less favorable assessment if you add any errors to the dataset. Such conditions can be expressed mathematically (see Harris and Soms, 1991 and Kabaila, 2005) and come down to test statistics having certain monotonicity properties. For the 2 × 2 table, the evidence for $p_1 > p_0$ is stronger if Y_1 is larger for fixed Y_0 and if Y_0 is smaller for fixed Y_1. Equivalently, the p-value should be nonincreasing in Y_1 for fixed $S = Y_0 + Y_1$ and nondecreasing in S for fixed Y_1, as noted by Berger and Sidik (2003). While the condition may appear obvious, standard approaches, such as the standard Z-test can violate it. Likelihood ratio tests typically satisfy any required monotonicity conditions.

These monotonicity properties have two important consequences. First, the maximum probability $a(\varphi) =$

$\sup_{\theta \in \Theta_0} \beta(\theta, \varphi)$ of rejecting the null is achieved at the boundary point $\theta = \theta_0$ (Röhmel, 2005). This not only ensures that the test is unbiased, but facilitates computations. Hence, there is no need to search over θ. Second, the tail set $\{P(T, S) \leq P(t, s)\}$ can be simply determined using the fact that $P(T, S)$ is a nonincreasing function of T for fixed S, as noted by Finner and Strassburger (2002).

The above conditions mention nondecreasing rather than strictly increasing. What about ties? For discrete data, it is never advantageous to have ties. It was shown by Röhmel and Mansmann (1999) that if a guaranteed p-value has any ties then breaking the ties appropriately can often make the p-value smaller while still being guaranteed. Similar results for confidence limits were demonstrated in Kabaila and Lloyd (2006).

2.6 Criteria for Comparison

There are several different criteria that can be used to assess the effectiveness of a test. If prior information summarised as a distribution is available on the unknown parameters then an exact Bayesian solution immediately follows. The frequentist properties of the Bayes tests are rarely poor, but neither are they exact. In decision theoretic approaches, various loss functions can be defined and minimised within a specified space of decision functions.

Even within the pure frequentist paradigm that we assume here, there is no nonrandomised test with maximum power and controlled size for discrete models. It seems unsatisfactory that this paradigm does not support an optimal analysis for a simple data structure like the 2 × 2 table. However, based on the four criteria to be listed below, we will find that there is indeed a practically optimal approach.

We now state four criteria that we will use to assess the different tests. The first two relate to their statistical accuracy, that is, to the test size and power. These two descriptors are central to frequentist theory and also to all trial regulation authorities. Tests should firstly be *valid*, or equivalently the p-value should be guaranteed. Gross violations of the size restriction is a serious defect of any test in our review. Ideally, the p-value should also be exact. This means that the test does not under-estimate the evidence against the null and will tend to lead to higher power. Restricting attention to valid tests means that the power achieved by different tests can be compared, without the complicating possibility that any extra power is purchased by size violations.

The other two criteria are more foundational. Tests should be based on a p-value that measures the probability of an observed event. This not only leads to a transparent test decision but provides quantitative information about how unusual the data is under the null hypothesis. Finally, where the model supports it on logical grounds, tests should satisfy certain monotonicity conditions. For the 2 × 2 table, these conditions were listed in the previous section.

3. MODERN PERSPECTIVES ON CONDITIONAL TESTS

It was argued by Fisher (1935) that the number of successes S should be treated as fixed; see Choi, Blume and Dupont (2015) for an overview and historical perspective. We evaluate the merits of this key modelling decision in Section 5.

3.1 Fisher's Exact Test

If we treat $S = s$ as fixed, then the model is given by (2.7). The distribution is stochastically increasing in θ and so we reject $H_0 : \theta \leq 0$ for larger values of y_1, and the p-value is $\Pr[Y_1 \geq y_1 | S = s]$ calculated from (2.7) maximised over $\theta \leq 0$. Because of stochastic monotonicity, the maximum occurs when $\theta = 0$. Fisher's p-value $P_F(y_1; n_1, n_0, s)$ is this tail sum of hypergeometric probabilities:

$$P_F(y_1; n_1, n_0, s) = \sum_{y \geq y_1} \binom{n_1}{y} \binom{n_0}{s-y} \Big/ \binom{n_0 + n_1}{s}. \tag{3.1}$$

The test is exact, in the sense that no approximation or estimation of unknown parameters is involved. Fisher's p-value is also exact in the technical sense of equation (2.5), assuming the model for Y_1 given s. The test generated by this p-value is therefore valid within this same conditional model.

The size of the test can be calculated exactly, since the hypergeometric distribution has no unknown parameters. For given values of n_0, n_1, s and target size α, let c_s be the smallest integer value c such that $P_F(c; n_1, n_0, s) \leq \alpha$. So the test rejects the null exactly when $y_1 \geq c_s$. It follows that the size of Fisher's test is

$$\alpha_s = \sup_{\theta \leq 0} \Pr(Y_1 \geq c_s | s) = P_F(c_s; n_1, n_0, s). \tag{3.2}$$

In words, the true size α_s is equal to the largest observable p-value less than α. So the test is not exact in the sense of having exactly the correct size. The smaller the support of the distribution the less likely it is that

α_s will be close to the chosen α. The support is smaller when the observed value of s is more extreme. In the extreme cases where $s = 0$ or $s = n$, the conditional test never rejects the null and the true size is $\alpha_s = 0$.

3.2 Randomised Version of Fisher's Exact Test

Based on the earlier mentioned theory of Lehmann (1959), there is a randomised version of Fisher's exact test which is UMPU for the one-sided test (see also Tocher, 1950) and for the two-sided alternative. For the one-sided alternative, this comes down to using the randomised p-value

$$\begin{aligned} &P_R(y_1, U; n_0, n_1, s) \\ &\quad = P_F(y_1; n_1, n_0, s) \\ &\qquad - U \Pr(Y_1 = y_1; n_0, n_1, s), \end{aligned} \tag{3.3}$$

where U is a uniformly distributed random number in the interval $(0, 1)$ (e.g., Cox and Hinkley, 1974, page 101). At the expense of introducing the random number U into the inference, we obtain a p-value with exact uniform distribution and a test with exact size α. Apparently, this p-value is always smaller than P_F and so is less conservative. In fact, the test is UMP amongst unbiased tests that are functions of (T, S, U) (Lehmann, 1959) which suggests that the shortcomings of Fisher's test are all due to discreteness. Of course, randomisation is almost never used in practice because we feel that conclusions should not depend on the random number u. If one takes the data to be (T, S, U), then the sufficiency principle states that inference should not depend on U. The conditionality principle would also recommend conditioning on the value of U which, since U is independent of (T, S), again means just using (T, S). There are alternative decision theoretic perspectives where the inference can be a distribution and randomisation is used to generate from this distribution. In this approach, U is not considered part of the data. However, this paper takes a frequentist approach.

3.3 Lancaster's and Liebermeister's p-Value

Lancaster (1961) proposed an alternative solution to the problem of conservatism of any discrete test, which has seen a fair degree of application. Lancaster's mid-p-value only counts half of the observed null probability of the observed sample point in the tail probability. Equivalently, it is obtained by subtracting half the observed probability from the usual tail probability. Referring to (3.3), the mid p-value $P_{\text{mid}}(Y_1; n_0, n_1, s)$ is given explicitly by replacing U by its mean value of 0.5. While not uniformly distributed like the randomised p-value, it has the exact mean (0.5) and variance of a uniform distribution (Agresti, 2002). Stronger theoretical justification for the one-sided mid-p are provided by Hwang and Yang (2001) and recently by Wells (2010).

Another test closely related to Fisher's was proposed by Liebermeister (1877). It is based on a Bayesian argument and turns out to equal Fisher's p-value but with a fictitious success added to the treatment group and a failure to the control group so it can be expressed as $P_F(y_1 + 1; n_0 + 1, n_1 + 1, s + 1)$. It was shown by Seneta and Phipps (2001) that it is always between $P_F(y_1 + 1; n_0, n_1, s)$ and $P_F(y_1; n_0, n_1, s)$, though not necessarily half way between. Like Lancaster's p-value, tests based on Liebermeister's p-value are less conservative than those based on Fisher's p-value.

3.4 Modern Approximations

During the 1980s, new developments in likelihood theory led to the proposal of the p^* formula by Barndorff-Nielsen (1983). The theory is complex but is based on a saddlepoint approximation to the density of the maximum likelihood estimator, conditional on a very generally formulated approximate ancillary statistic. Suppose we want to test a null hypothesis that the parameter $\delta = p_1 - p_0$ is less than or equal to δ_0. Until this point, the null value δ_0 has been zero. A general form for the p^* test statistic is

$$r^*(\delta_0) = r(\delta_0) + r(\delta_0)^{-1} \log\left(\frac{q(\delta_0)}{r(\delta_0)}\right), \tag{3.4}$$

where $r(\delta_0)$ is the signed root likelihood ratio statistic for testing $\delta \leq \delta_0$ and $q(\delta_0)$ is very complex in its general formulation but for the 2×2 table reduces to

$$\begin{aligned} q(\delta_0) = \bigl(\bigl\{&\tilde{w}_0\bigl(\operatorname{logit}(\tilde{p}_1) - \operatorname{logit}(\hat{p}_1)\bigr) \\ &- \tilde{w}_1\bigl(\operatorname{logit}(\tilde{p}_0) - \operatorname{logit}(\hat{p}_0)\bigr)\bigr\}\bigr) \\ &/\bigl(\sqrt{\tilde{w}_1/n_1 + \tilde{w}_0/n_0}\bigr), \end{aligned} \tag{3.5}$$

where $\tilde{w}_j = \tilde{p}_j(1 - \tilde{p}_j)$ and $\tilde{p}_j$ is the ML estimate of p_j under the null, as shown in Lloyd (2010b). The corresponding p-value is denoted $p^*(\delta_0) = 1 - \Phi(r^*(\delta_0))$. An advantage of this approach is that it is available in closed form. The normal approximation is held to be accurate to $O(n^{-1})$ in the medium deviation range (Davison, Fraser and Reid, 2006).

The appeal of p^* is that it depends continuously on the null value δ_0. Amongst other consequences, this means we can invert the test to get a confidence interval for δ. In contrast, Fisher's method only works for testing $\delta_0 = 0$, since no conditional distribution free of unknown nuisance parameters exists as δ_0 moves away from 0. The p^* method gives an answer close to the exact conditional solution when one exists but generalises, albeit approximately, to models and hypotheses where no exact conditional inference is possible. The p-value based on r^* is an approximation to the probability of a well-defined event, unlike the Lancaster or Liebermeister p-values.

The approach does present several problems, however. First, the formula for r^* breaks down when either $r = 0$ or $q = 0$. So to properly investigate its exact frequentist properties, it must be redefined. We define $r^* = r$ whenever the absolute value of r is less than 0.1 or when $q = 0$. These problems are completely ignored in the literature.

Another lesser problem is that, even with these modifications, p^* is not necessarily guaranteed (which is a fundamental criterion in our review) and further numerical work is required to evaluate its degree of liberalism. Second, for some quite natural models such as logistic regression with interest on the intercept, the conditional p-value becomes degenerate, even though p^* does not. In this case, what is the relation between the conditional degenerate p-value and p^*, which is supposed to approximate it? According to Pierce and Peters (1999), in such cases p^* is an "approximately conditional" p-value.

3.5 Approximately Conditional p-Values

One novel proposal to mitigate conservatism is to use a less discrete conditional distribution by conditioning on a range of values for the conditioning statistic rather than the exact value. This leads to a distribution with finer support. On the other hand, the nuisance parameter is no longer eliminated. Consider the p-value $P(t, s) = \Pr[T \geq t | S = s]$ calculated under the null. In the current context, this would equal Fisher's p-value. Define a neighbourhood $N_r(s)$ around the observed value of $S = s$, for example, $\{s - r, \ldots, s + r\}$. Then an approximately conditional p-value is defined as

$$\begin{aligned} &\Pr\bigl(P(T, S) \leq P(t, s) | S \in N_r(s_{\text{obs}})\bigr) \\ &\quad = \sum_{s \in N_r(s_{\text{obs}})} \Pr\bigl[P(T, s) \leq p_{\text{obs}} | S = s\bigr] \\ &\qquad \cdot \Pr\bigl[S = s | s \in N_r(s_{\text{obs}}); \varphi\bigr] \end{aligned} \tag{3.6}$$

When the size r of the neighbourhood equals zero, this gives the conditional p-value $P(t, s)$ since there is only one term in the sum. When $r > 0$, it is approximately conditional. Residual dependence on the nuisance parameter could in principle be handled by any of the methods that we will explain in Section 4.1 below.

The main problem is a lack of recommendation for the size of the neighbourhood $N_r(s)$ as well as its shape when S is higher dimensional. Certainly, a different choice of the neighborhood leads to a different p-value. A second problem is that the p-value still depends on the nuisance parameter φ. A third logical problem is a phenomenon known as spurious deflation; see Lloyd (2010a). It is too early to dismiss approximately conditional p-values though theoretical problems remain. They are at least based on the probability of a well-defined event and can be guaranteed by maximising with respect to the nuisance parameter.

3.6 Unconditional Assessment of Conditional Tests

The tests just described are based on the distribution of $T(Y)$ given $S(Y)$. When S really is fixed by design, it seems pertinent to assess the size and power treating it as fixed. When S is not fixed by design, it is still sometimes argued that conditional assessment is appropriate. Certainly though, in future hypothetical repetitions the value of S will vary and to allow for this we have to use the unconditional model. So both conditional and unconditional assessment have plausible arguments in their favour. But how are the two approaches related?

The conditional probability of rejection we will denote by

$$\beta(\theta | s) := \Pr\bigl[P(T, s) \leq \alpha | \theta, S(Y) = s\bigr], \tag{3.7}$$

where we have used the fact that the distribution of T given $S(Y) = s$ does not depend on φ. With slight abuse of notation, the unconditional probability of rejection is the mean value

$$\beta(\theta, \varphi) = \sum_s \beta(\theta | s) \Pr(S = s; \theta, \varphi) \tag{3.8}$$

of $\beta(\theta | S)$ with respect to the distribution of S which depends again on (θ, φ). For the 2 × 2 table, $S = Y_0 + Y_1$ has the distribution of a sum of two binomials and the summation is from $s = 0, \ldots, n_0 + n_1$.

When $\theta = \theta_0$, $\beta(\theta_0 | s)$ is the conditional size which we earlier denoted α_s. The unconditional size is the mean value of α_S with respect to S. For the Fisher test, α_s is almost always strictly less than α for all values of s and so unconditional size is also less than α. So Fisher's test is unconditionally conservative by design. For the Lancaster or Liebermeister test, their conditional size α_s is typically less than α for some values of

s and larger than α for others. The unconditional size is the mean value which is typically quite close to α because of the averaging, though it can exceed α.

4. MODERN PERSPECTIVES ON UNCONDITIONAL TESTS

For testing the null hypothesis $\theta \leq 0$ against $\theta > 0$, there are several commonly used test statistics based on the unconditional joint binomial model.

From a historical point of view, the best known is the chi-squared statistic (Pearson, 1900). Alternatively, tests can be based on the difference $\hat{p}_1 - \hat{p}_0$ divided by a standard error. When this standard error is estimated under the null hypothesis (i.e., assuming $p_1 = p_0$) it gives rise to the so-called pooled statistic. This can be shown to be a particular case of the Rao's score statistic and is identical to a one-sided version of the chi-square statistic. When the standard error is estimated without any restrictions on p_1 and p_0, it is called the unpooled statistic, which is the Wald statistic based on the interest parameter $\theta = p_1 - p_0$. There are other Wald-type statistics that can also be used, for instance based on the difference between the logarithm or the logit of the estimated success rates $\hat{p}_1$ and $\hat{p}_0$ but they are not in common use. The main alternative to these statistics is the likelihood ratio test, or its one-sided version known as the signed root likelihood ratio (SRLR) test. Formulas for these well-known statistics are in Appendix 1 (see the Supplementary Material, Ripamonti, Lloyd and Quatto, 2017).

The problem is that none of these tests are exact. Suppose we start with an approximate p-value $P(Y)$ based on test statistic $Z(Y)$. We remind the reader of the definition of the profile $\pi(y, \varphi)$ of a p-value $P(Y)$ given in equation (2.4). In words, it is the null probability of the p-value being equal or smaller than its observed value $P(y)$, the *true significance* if you will. Ideally, it should equal $P(y)$. It is worth noting that the tail set $\{P(Y) \leq P(y)\}$ in the definition of $\pi(y, \varphi)$ can equally be expressed as $\{Z(Y) \geq Z(y)\}$ and it is partly a matter of taste how the formulas below are presented.

There are several methods of using the profile function to define either an exact or almost exact version of the original p-value $P(Y)$. These ideas are mostly quite recent and can be implemented with modern computational resources.

4.1 The Maximisation Procedure

The "worst case" p-value is $P^*(y) = \sup_\varphi \pi(y, \varphi)$. While this is a completely general method, with general optimality properties stated below, the seminal paper recommending maximising out nuisance parameters was Barnard (1945). For the 2 × 2 table, we take the nuisance parameter φ to be the common value of $p_1 = p_0$ under the null, and this becomes

$$P^*(y_1, y_0) = \sup_{0 \leq p \leq 1} \Pr\big[Z(Y_1, Y_0) \geq Z(y_1, y_0); \\ p_0 = p_1 = p\big] \tag{4.1}$$

computed by enumerating all pairs (y_1, y_0) in the tail set $\{Z(Y_1, Y_0) \geq Z(y_1, y_0)\}$, summing their null probabilities based on the independent binomial distribution and then maximising with respect to p. We will call this adjustment the M-step. The maximised p-value has the following incredibly strong optimality property: amongst all statistics that are nonincreasing functions of $Z(Y)$, $P^*(Y)$ is the smallest function that is a guaranteed p-value. It is also exact in the sense of (2.5). So, if a test is not expressible as an M p-value based on some test statistic, then it can be improved by the M-step.

The test statistic $Z(Y_1, Y_0)$ may be any of the three mentioned in the previous section. Since $P^*(y_1, y_0)$ only depends on the way $Z(y_1, y_0)$ ranks the sample space, dependence on the choice of Z is modest, so long as it is chosen to be one of the standard test statistics. Moreover, while slightly different answers can be obtained, each p-value is exact in the sense of equation (2.5). The maximization procedure can even be applied to any of the conditional p-values from the previous section, thus converting a conditional test into an exact unconditional test (Boschloo, 1970, McDonald, Davis and Milliken, 1977, Mehrotra, Chan and Berger, 2003).

We remind the reader that when $Z(y_1, y_0)$ does not satisfy the monotonicity properties, the tail probability in (4.1) should in principle be maximised over $\{p_1 \leq p_0\}$, as first pointed out by Röhmel (2005). Fisher's p-value, as well as Lancaster and Liebermeister are monotonic. Among the three standard statistics, only the SRLR statistic is necessarily monotonic.

4.2 The Restricted Maximization Procedure

Maximising the profile function over the entire nuisance parameter space seems extreme when many nuisance parameter values will be very unlikely in light of the data. This might lead to unnecessarily conservative inference. When this occurs, it can be traced to the existence of spikes in the profile, often at values of φ far from its estimated value. Such problems may be avoided by using the procedure proposed by Berger and Boos (1994), which narrows the set of values in the domain of the parameter φ to a confidence set before taking the maximum:

$$P_{\mathrm{BB}}(Y) = \sup_{\varphi \in C_\gamma} \Pr[Z(Y) \geq Z(y); \theta_0, \varphi] + \gamma,$$

where C_γ is a $100(1-\gamma)$ percent confidence interval for φ. In the present case, probability on the right-hand side is calculated under $p_1 = p_0 = p$ and a confidence interval for $\varphi = p$ is the well-known Clopper–Pearson interval. Normally, γ is taken to be very small, for example, 0.001. Again, the tail set could be expressed in terms of the p-value instead of the test statistic. This restricted maximization, which we will call the B-step, produces a guaranteed p-value and typically a smaller p-value than using the M-step.

4.3 The Estimation Procedure

A cruder alternative to accounting for the nuisance parameter by maximization is to replace it with an estimate (Storer and Kim, 1990). In its most general form, this gives what we will call the E p-value

$$P_{\mathrm{E}}(y) = \pi(y, \hat{\varphi}_0), \tag{4.2}$$

where $\hat{\varphi}_0$ is an estimator of φ under the null hypothesis. This is a parametric bootstrap p-value if the bootstrap is viewed as a general recommendation to use the data to estimate the null distribution of the test statistic. For simple models, like the 2 × 2 table, no simulation is required. The value of the E p-value is obtained from equation (4.1) but, rather than maximise with respect to p, it is replaced by the estimate $\hat{p}$. The main problem of this approach is that the resulting p-value is not necessarily guaranteed (Berger and Boos, 1994).

The E-step can be performed more than once by iterating the construction of the significance profile in 2.4. The three methods, M-step, B-step and E-step can also be combined. Lloyd (2008a) proposed applying the M-step to $P_{\mathrm{E}}(y)$, known as the the E + M p-value. More explicit formulas for all these methods are given in Appendix 2.

4.4 Numerical Illustration

To clarify exactly how these three adjustments work, we illustrate their application when $Z(Y_1, Y_0)$ is the pooled z-test. The fictitious data is $y_1 = 13$ responses out of $n_1 = 100$ and $y_0 = 2$ responses out of $n_0 = 50$. The observed value of the test statistic is $Z(y_0, y_1) = 1.732$. Practitioners would typically quote the observed p-value $1 - \Phi(1.732) = 0.0416$. How accurate is this? Figure 1 displays the true significance, as measured by the profile $\pi(y = (2, 13); p)$, with the quoted value 0.0416 as a horizontal line. It deviates from the quoted value, mainly for larger values of p but also for $p = 0.5$. The nuisance parameter φ here is again the assumed common value p of $p_1 = p_0$ under the null.

The maximum of the profile is $P^* = 0.0677$ and occurs at $p = 0.968$. On the basis of this M-step, we quote the p-value 0.0677 instead of the original 0.0416. This value is much larger because of the presence of a spike in the profile but considering that $\hat{p} = 15/150 = 0.1$, one might wonder about taking account

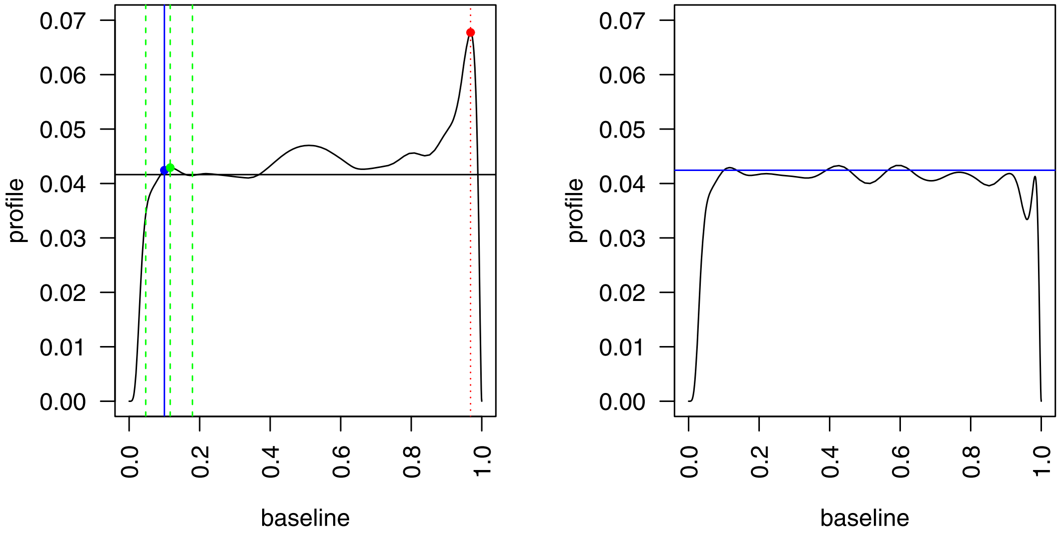

FIG. 1. *Left. Profile for pooled Z p-value for data* ($y_1 = 13, n_1 = 100$; $y_0 = 2, n_0 = 50$), *illustrating the M* (*in red*), *B* (*in green*) *and E-step* (*in blue*) *p-values that this profile generates. Right. Profile of the E p-value for the same data set.*

of the possibility that $p = 0.968$. This motivates the alternative B-step. With $\gamma = 0.01$, an exact 99% confidence interval for p is $(0.0471, 0.1796)$ marked on the plot as dashed green vertical lines. The maximum over this restricted range is 0.0429. After adding the penalty $\gamma = 0.01$, we quote $P_{\mathrm{BB}} = 0.0529$. The E-step involves estimating p by $\hat{p} = 0.1$ and the value of the profile at this point, marked by a vertical blue line, is $P_{\mathrm{E}} = 0.0424$.

It was noted previously that estimated p-values may not be guaranteed. In the right panel of Figure 1, we have calculated the profile of $P_{\mathrm{E}}(Y)$ [which required calculating all possible values of $P_{\mathrm{E}}(y)$] as well as the quoted value as a horizontal line. In such cases, the quoted p-value is extremely close to the true significance profile. This behaviour is typical for E p-values in this context (Lloyd, 2008b). The E + M p-value is the maximum of this profile and equals 0.0427, achieved at $p = 0.42$. The latter M-step removes the practically tiny amount of conservatism or liberality that may be present, and the resulting p-value is exact.

4.5 Two-Sided and Multi-Dimensional Tests

One-sided tests (such as superiority or noninferiority trials) are very common in biomedical contexts, which is why the theory presented to this point is oriented towards one-sided tests. However, the tail set $\{Z(Y) \geq Z(y)\}$ could be based on a two-sided test statistic Z if desired. This is perhaps even clearer when the theory is expressed in terms of the equivalent p-value, where the tail set is $\{P(Y) \leq P(y)\}$.

Also suppressed in the theory is the dimension of the nuisance parameter φ, which is unspecified. In principle then, the theoretical framework is completely general. The M-step, B-step and E-steps are applied in exactly the same way for one or two-sided tests and for any number of nuisance parameters. The M-step retains the same optimality properties stated in Section 4.1 and the B-step always produces a guaranteed p-value. But both these methods become computationally infeasible for many nuisance parameters. Only the E-step maintains the same computational burden as the dimension of φ increases. In the context of 2×2 tables, all three methods are computable for realistic sample sizes.

5. STRUCTURED ASSESSMENT OF COMPETING TESTS

In this section, we review the main arguments for and against the conditional and unconditional model, without taking a position on which is better. We then compare proposed tests within the conditional framework and come to a clear recommendation. Within the unconditional framework, there are literally dozens of plausible tests. We assess these by their theoretical properties as detailed in Section 2.6, as well as their computational burden. Moreover, we support our final recommendation with a numerical study.

5.1 Conditional or Unconditional?

All statistical models involve some conditioning; those things we consider incidental to the data, for instance the sample size, do not have their distributions modeled but, rather, are considered fixed. The dispute between the use of conditional or unconditional tests has a long history and many of the battles have been fought around the 2 × 2 table (see Agresti, 1992, 2001 for a review).

In Fisher's famous tea-tasting experiment, the total number of positive responses s was fixed by design. However, Fisher later argued that it should be considered fixed regardless, arguing that S has much in common with the sample size. Conditioning on the total successes was later proposed not only for comparative trials (e.g., Gail and Gart, 1973, Gart, 1969) but also in matched case-control studies (e.g., Hirji, Mehta and Patel, 1988) and for tables of higher dimension (see Hirji, 2006). The theory extends naturally to generalised linear models with canonical link. For noncanonical link, conditioning on approximate ancillary statistics has led to the p^* formula discussed earlier. The argument for conditioning in 2 × 2 tables cannot be understood without reference to this wider context.

In the narrower context of 2 × 2 tables, the conditional model has a single free variable y_1 and a single parameter θ and the theory is very simple. The unconditional model has two free variables (y_1, y_0) and an additional nuisance parameter. Even though the model is still very simple, it is rich enough to expose all of the difficulties and limitations of frequentist inference. There are plausible arguments for either model, which we now elucidate.

5.2 Arguments for Conditional Inference

The first argument for conditional inference is based on Lehmann's theory. He showed that in full rank exponential families the use of conditioning, with randomisation, leads to an unconditional test. This suggests that the conditional likelihood contains all the relevant information with respect to the parameter of interest. While seldom used in practice, randomisation reveals this basic structure, just as embedding real numbers within the complex number system brings insights into solving polynomials.

A second argument is that the total number of successes S has the same germane properties as the sample size. However, S differs from the sample size in that its distribution depends on the parameters. More refined arguments were based on the idea that S, by itself, is uninformative about the interest parameter θ. Formalising this notion leads to various definitions of approximate ancillarity and sufficiency (Barndorff-Nielsen, 1973, Cox, 1980, Godambe, 1980). However, all these definitions have unsatisfactory implications for some statistical models and it is fair to say that no consensus emerged. There is also debate about the ancillarity argument itself (Berkson, 1978).

A third argument, due to the second author, points to an epistemologically undesirable property of unconditional inference. Consider the most extreme outcome, when there are all successes for treatment and all failures for control; the p-value equals the probability of this single most extreme outcome $y_{\mathrm{E}} = (n_1, 0)$. The unconditional p-value can then be decomposed as

$$\Pr(Y = y_{\mathrm{E}}) = \Pr(Y_1 = n_1 | S = n_1) \times \Pr(S = n_1; p).$$

The first factor is Fisher's p-value; the second factor is a pure probability about S which depends on p but whose maximum value is small. Why should the event $S = n_1$ be counted against the null hypothesis? In words, why should a total of n_1 successes out of $n_0 + n_1$ trials be counted as evidence that the treatment works? Unconditional inference commits us to this inference.

The fourth and last argument for conditioning is simplicity and convenience: conditioning on S eliminates φ from the model and provides an incredibly simple model (2.7) with a single variable y_1 depending on the parameter of interest θ. All good statistical modelling involves treating incidental aspects of the data generating mechanism as fixed so that we can focus on the issue at hand. While eliminating φ from the model is attractive, there are other methods that do not involve conditioning, as detailed in Section 4 (see also Basu, 1977, for an earlier inventory of methods). So conditioning, while one option, is not necessary to account for the nuisance parameter.

5.3 Comparison of Tests Under the Conditional Model

In this section, let us accept the conditional model (2.7) as the model for the number of treatment successes y_1 given the total successes s.

Fisher's p-value is the probability of an observed event. It answers the question: out of $y_1 + y_0 = s$ successes, how often would at least y_1 of them be in the treatment group if the treatment has no effect? It also decreases in y_1 for fixed s and increases in s for fixed y_1, which are the key monotonicity properties in Section 2.5. So the key logical hurdles are passed. On the other hand, for a fixed target nominal size α, the test size a_s given in 3.2 is less than nominal, sometimes much less. This is the source of the common claim that Fisher's test is conservative. The claim is spurious.

Within the conditional framework, some size conservatism is an inevitable consequence of discreteness but an exact p-value still exists, as explained in Section 2.3. Fisher's p-value is exact in this strong technical sense. It is the smallest possible valid p-value that is monotone increasing in y_1. So within the conditional framework, Fisher's test is not unnecessarily conservative. Indeed, any other test that is valid will be even more conservative and any test that is less conservative will be invalid.

The mid-p and Liebermeister proposals are both attractive, but their conditional size can exceed nominal (Hirji, Tan and Elashoff, 1991, Seneta and Phipps, 2001) and the p-values are never guaranteed. Seneta and Phipps (2001) compared the size attained by Fisher's, Liebermeister's and Lancaster's test. These authors showed that Liebermeister's test is the closest to the nominal level (even though it is not valid, exceeding the nominal level) followed by Lancaster's and Fisher's test. So if closeness of attained size, rather than validity, were our key criterion we might be moved towards Liebermeister's test. However, we consider validity of the test and the guaranteed property of a p-value a key criterion.

In addition, both mid-p and Liebermeister suffer from the drawback that they are not the probability of any observed event. While we might consider approximations to a guaranteed test, neither is an approximation to the Fisher p-value. Certainly, neither can be justified within the conditional model. Evaluated unconditionally, their performance may be acceptable and we will present some numerical results in Section 6. However, there are competing tests within the unconditional framework that we will ultimately prefer.

Finally, the randomised version of Fisher's test is UMPU. So Fisher's test may be thought of as the closest valid discrete approximation to the UMPU test. Thus, any conservatism of Fisher p-value is an inevitable artifact of discreteness. In summary, within the conditional framework there appears to be no alternative to Fisher's test. As we shall see later though, the criticisms of Fisher's test are mainly made from an unconditional perspective.

An area for future research is to clarify the properties of p^* and approximately conditional p-values. The former are not necessarily valid and their degree of liberalism should be better assessed. For the latter, it remains unresolved how to determine a general neighborhood for conditioning.

5.4 Arguments for Unconditional Inference

The most persuasive argument for the unconditional model is that in future repetitions of the experiment the value of S will vary. If we want practical assessment of the future performance of the test—which is the key aim of frequentist inference—then we should allow S to vary. At the very least, this suggests augmenting any conditional test with a statement of its unconditional properties.

There are two specific arguments against the conditional model. The first is that conditional inference does not easily generalise to noncanonical parameters. In the context of 2 × 2 tables, we can perform conditional tests of the log-odds ratio but not of the risk difference, as pointed out in Section 3.4. Moreover, even with canonical parameters the relevant conditional distribution can be degenerate, leading to a test with zero size and power. While p^* methods were developed to address these problems, the fact that it gives a nondegenerate answer in this latter case is problematic.

The second argument against the conditional model is conservatism. Basing tests on the unconditional model allows greater unconditional power. This is partly because the distributions involved are much less discrete but also because the conditional size a_s of a test need not be less than α for all s, so long as its mean value is less than α. It is worth noting though that there are cases where conditional tests are more powerful; see Mehrotra, Chan and Berger (2003). Extending the investigation to the case of three binomials (which arises in a three-arm clinical trials), the conditional and unconditional approach seem to achieve similar power (Mehta and Hilton, 1993).

There does not exist a conclusive argument for or against conditioning, either in general or for 2 × 2 tables. Many might argue that this dilemma reveals a fundamental weakness in frequentist inference. For 2 × 2 tables, if the conditioning argument is accepted, then Fisher's exact p-value is exact in the sense of equation (2.5). If the conditioning argument is not accepted, then there is a much wider field of candidate tests which have to be compared. This includes ostensibly conditional tests that are made unconditional by the M, B or E steps.

5.5 Comparison of Tests Under the Unconditional Model

There are many tests in current use: Fisher, mid-p, Liebermeister, pooled-Z, unpooled-Z, the likelihood ratio and various Wald tests. All of these can be assessed within the unconditional model. None of them are exact. All can be adjusted using the M-step, B-step or E-step. A numerical study below will illuminate the properties of the basic and adjusted tests. However, we can say quite a lot about the three adjustments based on theoretical considerations. The example and figure in Section 4.4 serves as an excellent heuristic.

First, all M-step p-values are exact in the sense of (2.5) and subject to the ordering of the sample space induced by the initial test cannot be improved. If there is a spike in the profile then the maximised p-value will tend to be larger and power will be degraded. If there is no spike, then maximisation will just recalibrate the test to remove its conservatism or liberality.

Partially maximised p-values tend to be smaller when there is a spike and pay an insurance premium γ to achieve this, even if there is not a spike. From their definitions, it can be asserted that $P_{\mathrm{B}}(y) < P_{\mathrm{M}}(y) + \gamma$ but when there is a spike $P_{\mathrm{B}}(y)$ will be much smaller than $P_{\mathrm{M}}(y)$. The B-step p-value is guaranteed but is not exact: only M p-values can be exact, and applying the M-step to the B p-value will reduce it slightly (but by no more than γ). For more complex models where φ is a vector, construction of the confidence region C_γ is left unspecified and so partial maximisation is not a well-defined procedure. Indeed, for many models no exact confidence region for φ exists and the method cannot be formally applied.

The estimated p-value is the smallest of the three. It can be easily shown that $P_{\mathrm{E}}(y) < P_{\mathrm{B}}(y) - \gamma < P_{\mathrm{M}}(y)$. The cost is that $P_{\mathrm{E}}(y)$ is not guaranteed and tests based on it can be invalid. However, empirically it is found that the profile of $P_{\mathrm{E}}(y)$ is very flat, much closer than any asymptotic argument might suggest (Lloyd, 2008b). Consequently, $P_{\mathrm{E+M}}(y) \approx P_{\mathrm{E}}(y)$. This supports the use of $P_{\mathrm{E+M}}(y)$ in principle and $P_{\mathrm{E}}(y)$ in practice. Of course, when the original profile is quite flat, all the p-values will be close. However, for all of the standard tests the profile can be far from flat.

A final issue worth mentioning is the choice of initial test to generate the profile. Maximised p-values depend quite a lot on this choice, partially maximised p-values much less and estimated p-values hardly at all. This is a very attractive feature of $P_E(y)$, as it effectively removes any consequences of the user's choice of initial test. All these assertions will be verified in the numerical study below.

We now turn to computational issues. All three adjustments require the set $\{P(Y) \leq P(y)\}$ to be enumerated. Potentially, this requires evaluation of the generating p-value for all possible data sets. So a simpler generating p-value has great computational advantages. Amongst the simplest are Fisher's exact p-value and the maximised version was recommended by Boschloo (1970). The M and B steps require similar computation, while the E-step is faster, since the nuisance parameter φ is estimated rather than maximised. For 2 × 2 tables, all three can be calculated in a few seconds for sample sizes up to 1000.

The theoretically attractive E + M p-value requires computing all possible values of $P_E(y)$. This is currently limited to modest sample sizes of a few 100. Nevertheless, if computation were not an issue we would recommend the E + M p-value, based on the LR test because of its monotonicity properties and consistently high power of the resulting E + M p-value.

For more complex models where φ is a vector, the M and B-steps are not practical to compute. The E-step is not adversely affected by the dimension of φ and can be implemented for generalised linear models using importance sampling; see Lloyd (2012).

6. A NUMERICAL STUDY

To illustrate, verify and compare the unconditional performance of the tests reviewed in this article, we conducted a numerical study. Full details are provided in the online Appendix but it is pertinent to give representative results here. We considered eight test statistics: pooled, unpooled, log Wald, SRLR, p^*, Liebermeister, mid-p and Fisher. Only the last of these is guaranteed and the others can all be liberal for some parameter values. We calculated the unconditional size and power of the tests using five different versions of the basic statistics: raw, M, B (with $\gamma = 0.001$), E, and E + M. We fixed the nominal size $\alpha = 0.05$, the control sample size $n_0 = 40$ and the treatment sample size $n_1 = 60$. This choice is broadly representative of the patterns we have observed across all unbalanced designs.

In Table 1, we report the exact size of the 40 tests using two measures: maximum size with respect to p_0 (upper section) and mean size with respect to p_0 (lower section). In the max part of the table, violations over 0.051 are highlighted in red (over 0.06 is bold). Amongst the raw tests, mid-p is very close to exact but this is not the always case for other sample sizes. M, B and E + M tests are all theoretically valid (which is confirmed numerically in the table), whereas the E test does occasionally violate size by a nontrivial amount, but only when the original statistic is the unpooled or Fisher.

In the lower part of the table, we use colour coding to highlight the largest possible mean size subject to validity. This would imply a flatter profile and would typically lead to higher power. The B procedure is never worse than the M procedure, but is occasionally only slight advantageous. By contrast, the E and E + M procedures are very stable across test statistics, and the advantage is more pronounced.

In Table 2, we show the power results for three selected values of p_0, and corresponding values of p_1

TABLE 1

Maximum size (above) and mean size (below) calculated for 8 test statistics with 5 methods; $n_0 = 40$, $n_1 = 60$, $\alpha = 0.05$

TYPE	pooled	unpooled	log.wald	lr	p*	lieberm.	midp	fisher
raw	**0.066**	**0.088**	0.040	**0.088**	**0.088**	**0.064**	0.051	0.033
M	0.049	0.048	0.047	0.048	0.046	0.050	0.049	0.049
B	0.047	0.048	0.047	0.048	0.050	0.047	0.047	0.049
E	0.050	0.054	0.054	0.050	0.050	0.057	0.057	0.057
E+M	0.050	0.050	0.050	0.050	0.046	0.050	0.050	0.050
raw	0.051	0.052	0.033	0.053	0.053	0.047	0.042	0.028
M	0.037	0.030	0.037	0.029	0.027	0.042	0.041	0.041
B	0.041	0.036	0.039	0.038	0.039	0.042	0.042	0.041
E	0.045	0.045	0.044	0.043	0.043	0.045	0.045	0.046
E+M	0.045	0.045	0.044	0.043	0.043	0.044	0.043	0.043

TABLE 2

Power for three different combinations of p_0 and p_1, for 8 *test statistics and* 5 *methods*; $n_0 = 40$, $n_1 = 60$, $\alpha = 0.05$

p0	p1	TYPE	pooled	unpooled	log.wald	lr	p*	lieberm.	midp	fisher
0.10	0.25	M	0.583	0.583	0.628	0.561	0.508	0.595	0.628	0.613
		B	0.628	0.583	0.628	0.598	0.583	0.628	0.628	0.636
		E	0.628	0.633	0.629	0.633	0.629	0.629	0.629	0.652
		E+M	0.628	0.633	0.629	0.633	0.629	0.628	0.628	0.636
0.50	0.70	M	0.610	0.555	0.610	0.540	0.540	0.634	0.623	0.634
		B	0.623	0.603	0.623	0.623	0.632	0.634	0.634	0.634
		E	0.634	0.634	0.634	0.634	0.634	0.634	0.634	0.634
		E+M	0.634	0.634	0.634	0.634	0.634	0.634	0.634	0.634
0.75	0.90	M	0.609	0.536	0.589	0.586	0.586	0.653	0.608	0.628
		B	0.608	0.608	0.608	0.608	0.628	0.628	0.628	0.628
		E	0.653	0.673	0.672	0.651	0.651	0.653	0.653	0.672
		E+M	0.653	0.653	0.651	0.651	0.651	0.653	0.651	0.651

chosen so that the power is in a practically interesting range. As previously reported in the literature, it emerges that the B procedure leads to more powerful tests than the M procedure. E and E + M tests are always best or the equal best tests; the added value of these procedures is that they seem to work well across all circumstances. The E procedure should be recommended in applications, with the caution that it can occasionally lead to very slight violation of size when sample sizes are unbalanced. The E + M procedure is guaranteed, at the cost of a higher computational burden.

7. CONCLUSIONS

In this paper, we have reviewed both the conditional and unconditional approach to the 2 × 2 table. Both approaches lead to valid frequentist inference within their own different model frameworks.

Fisher originally suggested that the statistical model should depend on the study design. Indeed, when the total sum of successes in a binomial trial is fixed by design, it seems natural to consider the conditional approach and to evaluate power conditionally. When the sum is not naturally fixed by design, as is more often the case, it seems at least pertinent to adopt an unconditional perspective. Treating the sum of successes as fixed when it is not is defensible, but is based on general conditionality arguments whose application to the 2 × 2 table is not completely clear.

Thus, rather than conclusively support one approach against the other, we have reviewed and assessed alternative procedures within the conditional and unconditional models separately.

From the conditional perspective, there does exist an optimal test, namely the randomised version of Fisher's test. This test would be the gold standard for binomial endpoints, but cannot be recommended in practice, because randomisation introduces extra variation into the analysis. It is worth observing that within a decision theory framework randomised tests do not violate the sufficiency principle (e.g., Lehmann and Romano, 2005, page 58) since the randomisation distribution is the same for all data that give the same value of a sufficient statistic. At the point where a random number is drawn to complete the test and reject or accept the null hypothesis, the classical Fisherian sufficiency principle is contradicted. Regardless of these theoretical arguments, however, randomisation is almost never used in practice.

The unrandomised version, which is Fisher's exact test, has long been criticised for its conservatism, both conditionally and unconditionally. However, Fisher's test is valid and satisfies our definition of exactness. In addition, if one accepts the conditional model, then conservatism is an inevitable consequence of the discreteness. Fisher's p-value is exact and is the smallest possible valid p-value that is monotone increasing in y_1. By contrast, both Lancaster's and Liebermeister's tests mitigate the conservatism of Fisher's exact test, but both are necessarily liberal within the conditional framework. Moreover, neither is the probability of any observed event. Hence our recommendation is clear: within the conditional framework, Fisher's is really the only test to recommend.

The newest methods that are motivated from a conditional approach are p^* p-values and approximate conditional p-values, which follow recent developments in likelihood theory. These approaches have the virtue of extending conditional methods to models where exact conditioning is not possible, and provide close approximation to conditional procedures when it is possible. However, for the 2 × 2 table an optimal exact approach is available and there is no need to approximate it. More generally, the validity of such p-values is not guaranteed and they are not easy to write down explicitly or compute.

Amongst unconditional methods, we have explicated four methods for constructing p-values with good size control. From the least to the most computationally intensive these methods are E, M, B and $\mathrm{E}+\mathrm{M}$. The last three are guaranteed to be valid, while the first is very close to valid in practice.

M p-values account for the worst possible scenario and represent the most classical approach to handling the nuisance parameter, which is the key difficulty in the unconditional approach. Maximisation is a straightforward procedure and relatively easy to compute, at least for models with a single nuisance parameter; it guarantees exactness and validity, but allowing for the worst case often leads to unnecessarily low power.

B p-values are a simple method for overcoming the power loss of accounting for an unlikely worst case, especially for highly unbalanced designs. Though not exact, they are necessarily valid, while typically being smaller than M p-values. The difficulty of extending to more general and higher dimensional models and the lack of specification for the level of confidence is a theoretical weakness. However, for 2×2 tables, there is no practical impediment to their use and we would recommend B p-values over M p-values (with $\gamma = 0.001$). B p-values are today available in some statistical softwares.

The easiest p-value to compute (even for sample sizes of several 1000) is the E p-value, which leads to a flat profile. It is very close to exact and could be recommended in practice, provided that tiny violations of the size constraint are acceptable. The method extends to general models in a straightforward manner.

$\mathrm{E}+\mathrm{M}$ tests combine the use of a flatter significance profile (E step) with guaranteeing validity (M step). They are typically slightly more powerful than B p-values and they do not require user choice of a level of confidence. Another attractive feature is that $\mathrm{E}+\mathrm{M}$ (and E) methods lead to almost the same final inference, regardless of the choice of the test statistic. The only weakness of the $\mathrm{E}+\mathrm{M}$ method is its computational burden. R-code for all these methods is available from the authors.

This paper has reviewed contemporary approaches to the 2×2 table from a frequentist point of view. One reason of this choice is practical, as the clinical trials regulators mainly employ frequentist protocols. However, for the sake of comparison, we briefly mention Bayesian methods.

The Bayesian paradigm introduces *a priori* information in the inferential framework, and the plausibility of a hypothesis based on the data is assessed in terms of *a posteriori* probabilities. The key idea is treating parameters as random variables, so that, unlike the frequentist approach, one can integrate out any nuisance parameters. By imposing a continuous distribution on the parameters, the problem of discreteness is naturally solved, and, in case of the 2×2 table, computations are very simple.

Since conclusions are dependent on the prior specification, the frequentist properties of Bayesian methods cannot be stated in general. Nevertheless, there are links between Bayesian and frequentist inference for 2×2 tables. For instance, Liebermeister's test can be generated from a Bayesian argument and has quite good unconditional properties. The one-sided p-value from the pooled z-test can also be derived as the posterior probability of the null hypothesis based on independent Jeffries priors for (p_0, p_1); see Howard (1998). Confidence intervals, which is not a topic we have touched on, can be replaced by highest posterior density intervals; see Brown, Cai and DasGupta (2001) and Brown, Cai and DasGupta (2002).

SUPPLEMENTARY MATERIAL

Supplement to "Contemporary Frequentist Views of the 2×2 Binomial Trial" (DOI: 10.1214/17-STS627SUPP; .pdf). We provide formulas for standard approximate statistics and adjusted p-values. We illustrate in detail the numerical study.

REFERENCES

AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* **7** 131–177. With comments and a rejoinder by the author. MR1173420

AGRESTI, A. (2001). Exact inference for categorica data: Recent advances and continuing controversies. *Stat. Med.* **20** 2709–2722.

AGRESTI, A. (2002). *Categorical Data Analysis*. Wiley, Hoboken, NJ. MR3087436

BARNARD, G. A. (1945). A new test for 2 × 2 tables. *Nature* **156** 177. MR0013274

BARNARD, G. A. (1947). Significance tests for 2 × 2 tables. *Biometrika* **34** 123–138. MR0019285

BARNDORFF-NIELSEN, O. (1973). On M-ancillarity. *Biometrika* **60** 447–455. MR0345255

BARNDORFF-NIELSEN, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365. MR0712023

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72** 355–366. MR0451477

Berger, R. L. and Boos, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. MR1294746

Berger, R. L. and Sidik, K. (2003). Exact unconditional tests for a 2 × 2 matched-pairs design. *Stat. Methods Med. Res.* **12** 91–108. MR1963335

Berkson, J. (1978). In dispraise of the exact test: Do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportions? *J. Statist. Plann. Inference* **2** 27–42.

Boschloo, R. D. (1970). Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. *Stat. Neerl.* **24** 1–35. MR0264827

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci.* **16** 101–133. MR1892660

Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30** 160–201. MR1892660

Choi, L., Blume, J. D. and Dupont, W. D. (2015). Elucidating the foundations of statistical inference with 2 × 2 tables. *PLoS ONE* **10** e0121263.

Cox, D. R. (1980). Local ancillarity. *Biometrika* **67** 279–286. MR0581725

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

Davison, A. C., Fraser, D. A. S. and Reid, N. (2006). Improved likelihood inference for discrete data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 495–508. MR2278337

Finner, H. and Strassburger, K. (2002). Structural properties of UMPU-tests for 2 × 2 tables and some applications. *J. Statist. Plann. Inference* **104** 103–120. MR1900521

Fisher, R. A. (1935). *The Design of Experiments*, 1st ed. Oliver and Boyd, London.

Gail, M. H. and Gart, J. J. (1973). The determination of sample sizes for use with the exact conditional test in 2 × 2 comparative trials. *Biometrics* **29** 441–448.

Gart, J. J. (1969). An exact test for comparing matched proportions in crossover designs. *Biometrika* **56** 75–80.

Godambe, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika* **67** 155–162. MR0570517

Harris, B. and Soms, A. P. (1991). Theory and counterexamples for confidence limits on system reliability. *Statist. Probab. Lett.* **11** 411–417. MR1114531

Hirji, K. F. (2006). *Exact Analysis of Discrete Data*. Chapman & Hall/CRC, Boca Raton, FL. MR2193238

Hirji, K. F., Mehta, C. R. and Patel, N. R. (1988). Exact inference for matched case-control studies. *Biometrics* **44** 803–814. MR0963915

Hirji, K. F., Tan, S. J. and Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Stat. Med.* **10** 1137–1153.

Howard, J. V. (1998). The 2 × 2 table: A discussion from a Bayesian viewpoint. *Statist. Sci.* **13** 351–367. MR1705267

Hwang, J. T. G. and Yang, M.-C. (2001). An optimality theory for mid p-values in 2 × 2 contingency tables. *Statist. Sinica* **11** 807–826. MR1863164

Kabaila, P. (2005). Computation of exact confidence limits from discrete data. *Comput. Statist.* **20** 401–414. MR2236616

Kabaila, P. and Lloyd, C. J. (2006). Improved Buehler limits based on refined designated statistics. *J. Statist. Plann. Inference* **136** 3145–3155. MR2256221

Lancaster, H. O. (1961). Signifiance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56** 223–234. MR0124107

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York. MR0107933

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

Liebermeister, C. (1877). *Über Wahrscheinlichkeitsrechnung in Anwendung Auf Therapeutische Statistik*. Breitkiof and Härtel.

Lloyd, C. J. (2008a). A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* **64** 716–723. MR2526621

Lloyd, C. J. (2008b). Exact P-values for discrete models obtained by estimation and maximization. *Aust. N. Z. J. Stat.* **50** 329–345. MR2474195

Lloyd, C. J. (2010a). P-values based on approximate conditioning and p^*. *J. Statist. Plann. Inference* **140** 1073–1081. MR2574669

Lloyd, C. J. (2010b). Bootstrap and second-order tests of risk difference. *Biometrics* **66** 975–982. MR2758234

Lloyd, C. J. (2012). Computing highly accurate or exact P-values using importance sampling. *Comput. Statist. Data Anal.* **56** 1784–1794. MR2892377

Lydersen, S., Fagerland, M. W. and Laake, P. (2009). Recommended tests for association in 2 × 2 tables. *Stat. Med.* **28** 1159–1175. MR2662203

Martin Andres, A. (1991). A review of classic non-asymptotic methods for comparing two proportions by means of independent samples. *Comm. Statist. Simulation Comput.* **20** 551–583.

McDonald, L. L., Davis, B. M. and Milliken, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2 × 2 contingency tables. *Technometrics* **19** 145–158.

Mehrotra, D. V., Chan, I. S. F. and Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* **59** 441–450. MR1982586

Mehta, C. R. and Hilton, J. F. (1993). Exact power of conditional and unconditional tests: Going beyond the 2 × 2 contingency table. *Amer. Statist.* **47** 91–98.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* **5** 157–175.

Pierce, D. A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **54** 701–737. MR1185218

Pierce, D. A. and Peters, D. (1999). Improving on exact tests by approximate conditioning. *Biometrika* **86** 265–277. MR1705363

Ripamonti, E., Lloyd, C. and Quatto, P. (2017). Supplement to "Contemporary frequentist views of the 2 × 2 binomial trial." DOI:10.1214/17-STS627SUPP.

Röhmel, J. (2005). 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. *Biom. J.* **47** 37–47. MR2135888

RÖHMEL, J. and MANSMANN, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biom. J.* **41** 149–170. MR1693980

SENETA, E. and PHIPPS, M. C. (2001). On the comparison of two observed frequencies. *Biom. J.* **43** 23–43. MR1820037

SKIPKA, G., MUNK, A. and FREITAG, G. (2004). Unconditional exact tests for the difference of binomial probabilities—contrasted and compared. *Comput. Statist. Data Anal.* **47** 757–773. MR2101550

STORER, B. E. and KIM, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *J. Amer. Statist. Assoc.* **85** 146–155.

TOCHER, K. D. (1950). Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* **37** 130–144. MR0036972

WELLS, M. T. (2010). Optimality results for mid p-values. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat.* (*IMS*) *Collect.* **6** 184–198. IMS, Beachwood, OH. MR2798519

YATES, F. (1984). Tests of significance for 2 × 2 contingency tables. *J. Roy. Statist. Soc. Ser. A* **147** 426–463. MR0769998