Practical Valid Inferences for the Two-Sample Binomial Problem

Michael P. Fay*
National Institute of Allergy and Infectious Diseases
and
Sally A. Hunsberger
National Institute of Allergy and Infectious Diseases

May 16, 2017

Abstract

Consider comparing two independent binomial responses. Our interest is whether the two binomial parameters are different, and if different, which is larger, and if larger, by how much. This apparently simple problem was addressed by Fisher in the 1930's, and has been the subject of many review papers since then. Yet there continues to be new work on this issue and no consensus solution. Previous reviews have focused primarily on testing and power, or primarily on confidence intervals, their coverage, and average length. Here we evaluate both together; we define a frequentist "method" as a parameter estimate, the p-value of a test and its matching confidence interval. For focus, we only examine non-asymptotic inferences, so that most of the methods are valid (i.e., exact) by construction. Within this focus, we review different methods emphasizing many of the properties and interpretational aspects we desire from applied frequentist inference: validity, accuracy, good power, equivariance, unified inferences, coherence, causal interpretation, and parameterization and direction of effect. We show that no one method can meet all the desirable properties and give recommendations based on which properties are given more importance.

Keywords: 2 by 2 table, Barnard's test, Fisher's exact test, Unconditional exact test

*The authors thank Erica Brittain for helpful comments.

1 Introduction

Suppose we observe two independent binomial variates with parameters (n_1, θ_1) and (n_2, θ_2). One question we might have is: are θ_1 and θ_2 equal or not? If we reject the null hypothesis of equality (or even if we do not), we typically want to estimate how much larger one θ parameter is than the other. To answer these two questions, the frequentist typically presents an estimate of the effect, a confidence interval on that effect, and a p-value to test that there is no effect. For such a simple problem, one might think that by now there is a consensus method for testing and creating confidence intervals for this problem. But this is not so. New methods continue to be developed for this problem (see e.g., Lloyd, 2008; Wang, 2010; Wang and Shan, 2015; Fay et al., 2015), and the closely related problems of causal inferences from a two-sample randomized experiment with binary responses (Rigdon and Hudgens, 2015; Ding and Dasgupta, 2016). Many review papers on this problem focus on testing alone (see Lydersen et al., 2009), or confidence intervals alone (see Fagerland et al., 2015; Santner et al., 2007). Here we focus on both together.

We define a method as an estimator of a parameter of interest, a confidence interval, and a p-value function. This approach allows us to compare different methods by examining not just properties of each component (i.e., comparing powers of different p-value functions or expected lengths of different confidence intervals), but also to examine properties of the methods as a whole. For example, within a method we examine inferential agreement between the p-value function and confidence interval procedure. Additionally, we examine what directional inferential statements we can make from the method, such as stating that θ_2 is significantly larger than θ_1.

Although in some different statistical settings (e.g., two-sample normal problem) the standard method will automatically give inferential agreement between p-values and con-

fidence intervals as well as automatically give directional inferential statements, in the two-sample binomial problem those inferential properties are not automatic. Thus, before discussing the binomial problem, we first review the two-sample problem with normally distributed responses with the same variance. We consider the latter problem first, because there is some consensus that one method (the t-test, and its associated p-value and confidence interval) is appropriate for this problem. In the normal case, this t-test method meets some regularity properties that lead to inferences that are intuitive and easy to understand. Because these properties form the basis for a certain statistical intuition about how frequentist inferences ought to be, and because the example uses normal distributional assumptions, we call these properties the "normal intuition". We will show later how the normal intuition breaks down for the two-sample binomial problem, although many of the properties may approximately hold for large samples.

1.1 Background and Notation

Consider a general frequentist problem, where we observe data, $\mathbf{x}$, and denote its random variable as $\mathbf{X}$. Assume some probability model for $\mathbf{X}$ that depends on a parameter vector θ, but we are interested in a function of θ that returns a scalar, $b(\theta) = \beta$. We partition the possible values of θ into two sets, the null hypothesis, Θ_0, and the alternative hypothesis, Θ_1.

In this paper, except for Section 7, we consider only three classes of partitions, where the null and alternative space is defined by β, and separated by a value β_0 on the boundary between the hypotheses. These three classes are two-sided hypotheses,

$$\begin{aligned} H_0 : &\quad \beta = \beta_0 \\ H_1 : &\quad \beta \neq \beta_0 \end{aligned}$$

or one of the one-sided hypotheses,

Alternative is Less	Alternative is Greater
$H_0 : \beta \geq \beta_0$	$H_0 : \beta \leq \beta_0$
$H_1 : \beta < \beta_0$	$H_1 : \beta > \beta_0$.

Let $p(\mathbf{x}, \Theta_0)$ be a p-value associated with the null hypothesis, Θ_0. Typically, we assume a class of hypotheses and write (with a slight abuse of notation) $p(\mathbf{x}, \beta_0)$ as a p-value associated with the null hypothesis indexed by β_0. We reject the null hypothesis at level α if $p(\mathbf{x}, \beta_0) \leq \alpha$. Following Berger and Boos (1994), we define a p-value procedure as valid if

$$P_\theta \left[p(\mathbf{X}, \beta_0) \leq \alpha \right] \quad \leq \quad \alpha,$$

for all $\alpha \in (0, 1)$ and all $\theta \in \Theta_0$. The term *exact* is often used to describe tests that give valid p-values. For example, Fisher's exact test and unconditional exact tests. In this paper, we will follow that convention and make no distinction between the terms *exact* and *valid*. Following Röhmel (2005), we define a p-value procedure as coherent if for every $\mathbf{x}$, $p(\mathbf{x}, \Theta_0^*) \leq p(\mathbf{x}, \Theta_0)$ if $\Theta_0^* \subseteq \Theta_0$.

For the classes of hypotheses above, we can invert the p-value function to get its associated $100(1 - \alpha)\%$ confidence region,

$$C(\mathbf{x}, 1 - \alpha) \quad = \quad \{\beta : p(\mathbf{x}, \beta) > \alpha\}. \tag{1}$$

We define a confidence region as valid if it is guaranteed to have at least nominal coverage for every θ (and hence every $b(\theta) = \beta$); in other words,

$$P_\theta \left[\beta \in C(\mathbf{X}, 1 - \alpha) \right] \quad \geq \quad 1 - \alpha.$$

Often we use asymptotic methods to create p-values and confidence intervals that are not valid for finite samples, but approach validity as the sample size gets large. In this paper, we only consider non-asymptotic methods, and all are valid except the mid-p methods described in Section 9.

1.2 Standard Frequentist Inference: Normal Intuition

Consider the two-sample problem, where the ath group has n_a independent and normally distributed responses, with mean μ_a and variance σ^2. Let $\theta = [\mu_1, \mu_2, \sigma]$, and suppose we are interested in $\beta = b(\theta) = \mu_2 - \mu_1$. The t-test is valid for testing the null that $\beta = \beta_0$ and it is the uniformly most powerful (UMP) unbiased test (Lehmann and Romano, 2005, p. 160) for this problem. UMP unbiasedness means that among the class of unbiased tests for this problem (i.e., tests for which the power for each specific parameter in the alternative space is always greater than the power for every parameter in the null space), the t-test is the most powerful test regardless of which $\theta \in \Theta_1$ we measure power.

We study this case first to define "normal intuition" about frequentist inferences. This normal intuition is a series of properties, that if they are not met, conflict with many statisticians' intuitive feeling of how p-values and confidence regions ought to work. Here are those properties met by the difference in sample means, $\hat{\beta}$; the two-sided p-value from the t-test, p; and the $100(1-\alpha)\%$ confidence interval on β associated with that p-value, (L, U).

Reproducibility: Two statisticians applying the method to the same data always get the same results (as opposed to randomized tests).

Confidence region is an interval: The confidence region created from p through equation 1 is an interval, meaning it can be written as (L, U) with all values within the

interval belonging to the confidence region.

Unified Inferences: $p \leq \alpha$ if and only if the $(1-\alpha)$ confidence interval does not contain β_0. (This idea is similar to the unified report of Hirji, 2006, p. 77).

Accuracy (of coverage): Taken over repeated applications, the probability that the $100(1-\alpha)\%$ confidence interval procedure includes β is equal to $(1-\alpha)$ for all values of β regardless of the nuisance parameters.

Centrality (of CI): The $100(1-\alpha)\%$ CI is a central one, meaning $P[L > \beta] \leq \alpha/2$ and $P[U < \beta] \leq \alpha/2$.

One-sided p-value from Two-sided p-value: Half of the two-sided p-value can be interpreted as a one-sided p-value in the apparent direction of the effect. For example, if $\hat{\beta} > \beta_0$ then we can reject $H_0 : \beta \leq \beta_0$ at level $p/2$.

Directional Coherence (of p-value): The t-test method has "directional coherence", where we have expanded the definition of coherence of one-sided p-values to two-sided p-values with an estimate. Call a two-sided p-value function directionally coherent if the p-values are decreasing as β_0 gets farther from $\hat{\beta}$. In other words, directionally coherent two-sided p-values have $p(\mathbf{x}, \beta_0^*) \leq p(\mathbf{x}, \beta_0)$ when either $\beta_0^* < \beta_0 < \hat{\beta}$ or $\hat{\beta} < \beta_0 < \beta_0^*$. A two-sided p-value with this property can be interpreted as a coherent one-sided p-value in the appropriate direction. For example, if $\hat{\beta} > \beta_0$ then we can reject $H_0 : \beta \leq \beta_0$ at level p. (And for the t-test p-value, we can also reject at a level of $p/2$.)

Monotonicity (of power): As the sample size increases, there is an increase in power under any probability model in the alternative hypothesis.

Nestedness (of CIs): If we had used a larger confidence level, $(1-\alpha^*) > (1-\alpha)$, then the $100(1-\alpha^*)\%$ confidence interval, (L^*, U^*), would completely contain the $100(1-\alpha)\%$ one, (L, U); in other words, $L^* \leq L < U \leq U^*$.

1.3 Two-Sample Binomial: Failure of Normal Intuition

Now we turn to the two-sample binomial problem, where $X_1 \sim Binomial(n_1, \theta_1)$ and independently $X_2 \sim Binomial(n_2, \theta_2)$. Here the parameter of interest is typically one of three functions of $\theta = [\theta_1, \theta_2]$: the difference ($\beta_d = \theta_2 - \theta_1$), the ratio ($\beta_r = \theta_2/\theta_1$), or the odds ratio ($\beta_{or} = \{\theta_2(1-\theta_1)\} / \{\theta_1(1-\theta_2)\}$). In this problem, the inferential methods do not necessarily follow the properties that we would expect from normal intuition. We list several examples using several different valid tests, valid confidence intervals, or methods.

Failure of Reproducibility: The uniformly most powerful unbiased (UMPU) test of $H_0 : \theta_1 \geq \theta_2$ versus $H_1 : \theta_1 < \theta_2$ is a randomized version of a one-sided Fisher's exact test (see e.g., Lehmann and Romano, 2005; Finner and Strassburger, 2001). Testing this hypothesis at the one-sided $\alpha = 0.025$ level for the data $x_1/n_1 = 1/6$ and $x_2/n_2 = 7/9$, the UMPU test rejects 70.3% of the time, and fails to reject 29.7% of the time. So, provided they are not using the same pseudo-random number generator, there is a 41.7% chance that two researchers applying the UMPU test to those data will have different accept/reject decisions.

Associated confidence region not an interval: There are two versions of the two-sided Fisher's exact test and the most common is the Fisher-Irwin test (default in current versions of SAS [version 9.4] and R [version 3.3.2]). The test was designed to test $H_0 : \beta_{or} = 1$, but it can be generalized to test other null hypotheses. Consider the data $x_1/n_1 = 7/262$ and $x_2/n_2 = 30/494$ (see Fay, 2010a, Supplement, Section 3.1).

The two-sided p-value for testing the $H_0 : \beta_{or} = 1$ is $p = 0.04996$, which rejects the null hypothesis at the $\alpha = 0.05$ level. If we slightly change the null and test $H_0 : \beta_{or} = 0.99$, we get $p = 0.05005$, and we fail to reject. But counter-intuitively, if we change the null the other way and test $H_0 : \beta_{or} = 1.01$, we also fail to reject, $p = 0.05006$. So if we create the 95% confidence region by inverting the p-value procedure, this region is not contiguous,

$$C(\mathbf{x}, 0.95) \;=\; \{\beta : \beta \in (0.177, 0.993) \text{ or } \beta \in (1.006, 1.014)\}\,.$$

and includes values of β_{or} both larger and smaller than 1. The cause of this behaviour is the lack of unimodality of the p-value function; see Figure 1.

Non-unified inferences: If the confidence region is not an interval, we can create a valid CI by using the interval that covers the whole confidence region. But this will not give unified inferences. Returning to the Fisher's exact test confidence region example, we can create a 95% confidence interval by "filling in the hole" as $(0.177, 1.014)$ to create the *matching* confidence interval (see Section 3.1 or Blaker, 2000). In this case, the two-sided p-value rejects the null that $\beta_{or} = 1$ at the 0.05 level, but the matching 95% confidence interval includes $\beta_{or} = 1$. This issue is different from the non-unified inferences that often occurs by using different methods to calculate p-values and confidence intervals, which can be quite prevalent in this application. For example, the default for R (`fisher.test` in base R, version 3.3.1) and SAS (exact option in Proc Freq, version 9.4) uses the Fisher-Irwin two-sided p-value, but calculates the two-sided confidence interval on β_{or} by inverting two one-sided Fisher exact p-values (see e.g., Fay, 2010a,b).

Imperfect Accuracy of Coverage: Because of discreteness, the valid confidence interval must have coverage larger than the nominal level for some values of θ, in order

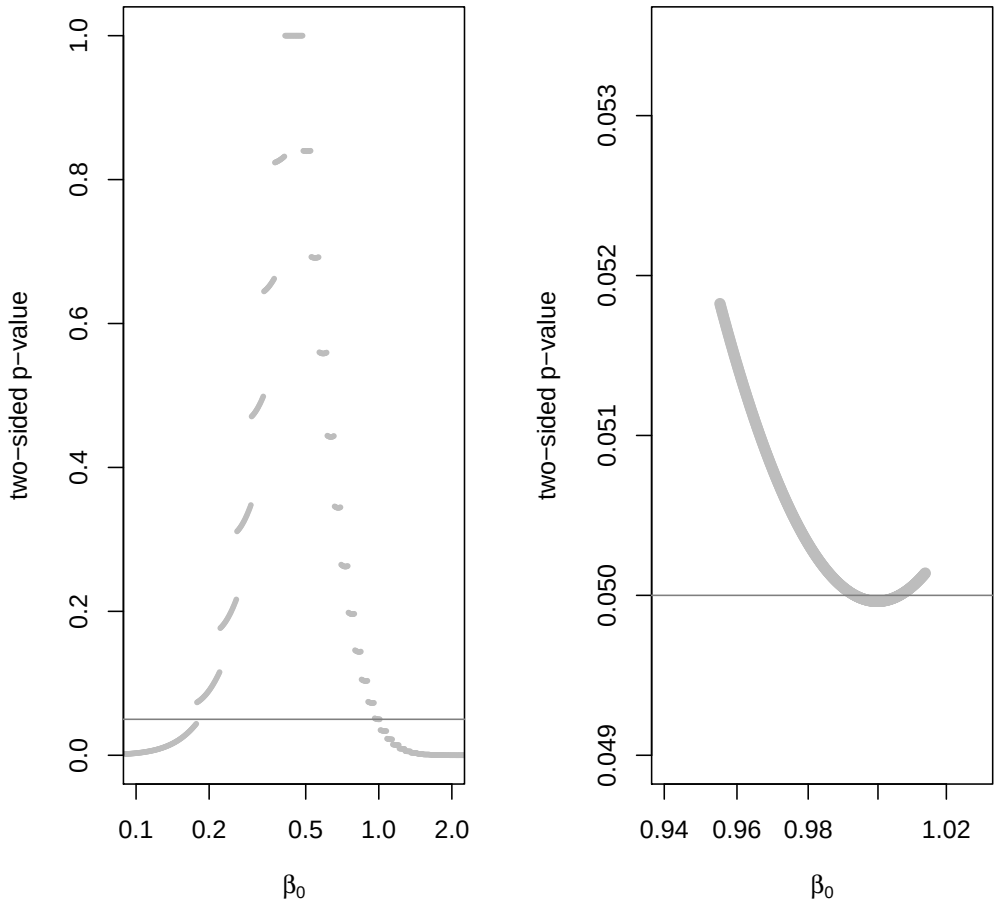

Figure 1: Two-sided Fisher's exact test (Fisher-Irwin version) p-values by β_0 for $x_1/n_1 = 7/262$ and $x_2/n_2 = 30/494$. Right panel is an enlargement of part of the left panel. Reference line is 0.05.

to ensure validity for all values of θ. Unfortunately, the term "exact" is often used to mean valid, so an "exact" confidence interval may have coverage greater than the nominal level and not, as the term might imply, have coverage exactly equal to the nominal level. Section 9 discusses relaxing the requirement of validity in order to have coverage closer to the nominal level "on average", slightly greater than nominal for some parameter values and slightly less for others.

Non-Centrality of Confidence Interval: Although central $(1-\alpha)$ CIs for the binomial problem are important, much has been written on non-central intervals. Agresti and Min (2001) showed that by inverting certain two-sided tests, we get smaller confidence intervals than central ones. For the difference in proportions, this strategy often uses an unconditional exact (i.e., valid) version of a two-sided score test (see Fagerland et al., 2015). For $x_1/n_1 = 5/9$ and $x_2/n_2 = 7/7$ then the difference in proportions is $\hat{\beta}_d = 0.444$ with 95% confidence interval using this method equal to $(0.005, 0.749)$ and the associated two-sided exact p-value for testing $\beta_d = 0$ giving $p = 0.0496$. Because the 95% confidence interval is based on inverting a two-sided test, we cannot use $p/2 = 0.0248$ as a one-sided p-value showing that $\beta_d > 0$ at the 0.025 level. In fact, to ensure validity, we can only use the two-sided p-value as an upper bound on that one-sided p-value.

Non-monotonicity of power: Continuing with the previous example ($x_1/n_1 = 5/9$ and $x_2/n_2 = 7/7$ using the unconditional exact two-sided score test), if we add one more observation to group 2 the two-sided p-value increases regardless of whether the extra observation is a failure (giving $x_2/n_2 = 7/8$ and $p = 0.172$), or success (giving $x_2/n_2 = 8/8$ and $p = 0.0510$) (this example comes from Vos and Hudson, 2008). Thus, it is not surprising that the power to reject at the two-sided 0.05 level when $\theta_1 = .4$ and $\theta_2 = .9$

is higher for $n_1 = 9, n_2 = 7$ (power= 61.9%) than for $n_1 = 9, n_2 = 8$ (power=53.7%). Power non-monotonicity can also exist for common one-sided tests. Using a one-sided Fisher's exact test at the 0.025 level, the power to reject $H_0 : \beta_{or} = 1$ when $\theta_1 = 0.01$ and $\theta_2 = 0.80$ is 71.7% when $n_1 = n_2 = 5$, but 63.2% when $n_1 = n_2 = 6$.

Non-nesting Confidence Intervals: Wang (2010) proposed a method for constructing the smallest one-sided confidence interval for the difference of two proportions. Consider $x_1/n_1 = 2/7$ and $x_2 = 2/5$. The lower one-sided 95% interval on the difference, β_d, is $(-0.467, 1)$, but the 96% interval by the same method is $(-0.442, 1)$. See Figure 2.

Non-Coherence: For testing for non-inferiority on a difference in proportions, Chan and Zhang (1999) recommend the exact unconditional test based on the score test. Röhmel (2005) give the following virtual example: the proportion of failures on control is $x_1/n_1 = 130/248$ and on new treatment is $x_2/n_2 = 76/170$, with the failure rate slightly lower on new treatment, $\hat{\beta}_d = -0.077$. If we want to show that $H_1 : \beta_d < 0.025$ the p-value is $p = 0.0226$, but if we want to show an even less stringent margin, $H_1 : \beta_d < 0.026$ the p-value non-intuitively increases to $p = 0.0239$ (see Figure 3).

For the two-sample binomial problem, many attempts to increase power or get the smallest width CI result in violations of some of these "normal intuition" properties.

1.4 Outline of Paper

We begin in Section 2 by discussing the choice of effect measure. In Section 3 we define matched methods, and discuss properties of methods such as unified inferences, and directionality of inferences. We describe methods for defining valid one-sided decision rules in

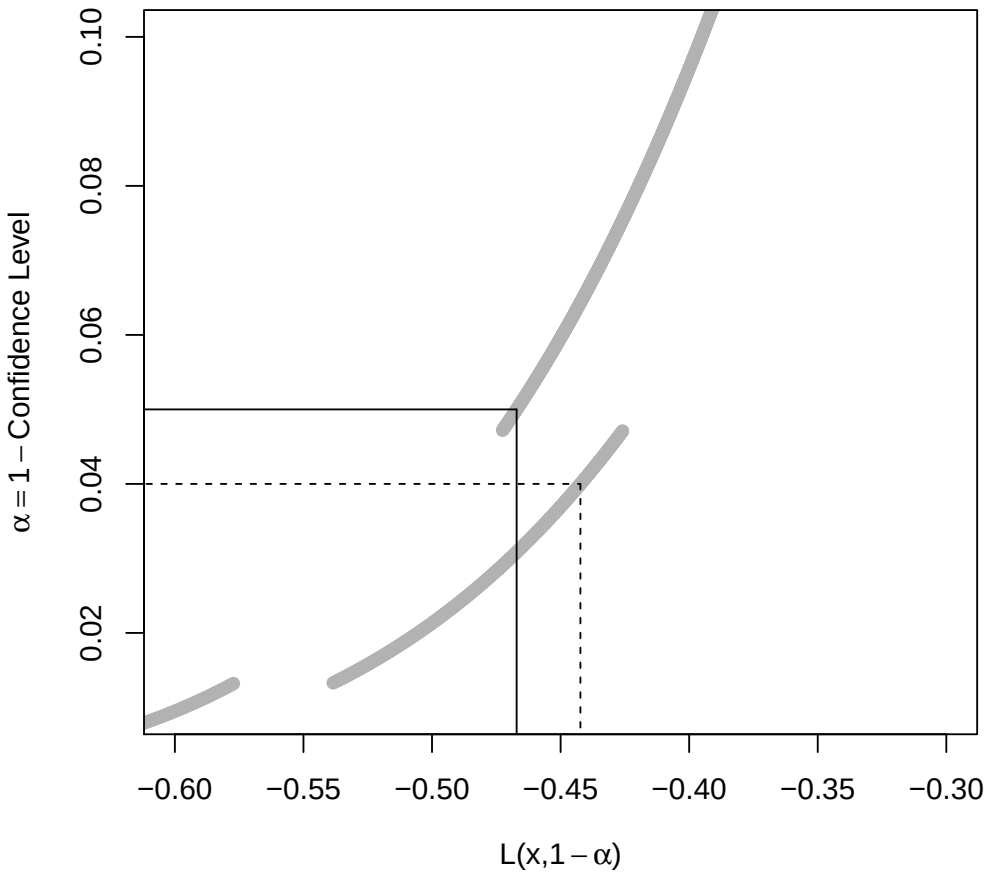

Figure 2: Thick gray lines are lower limits for the smallest one-sided $100(1-\alpha)\%$ confidence limits for β_d from Wang (2010) for $x_1/n_1 = 2/7$ and $x_2/n_2 = 2/5$. Solid black lines show one-sided 95% limit of -0.467, while dotted black lines show one-sided 96% limit of -0.442.

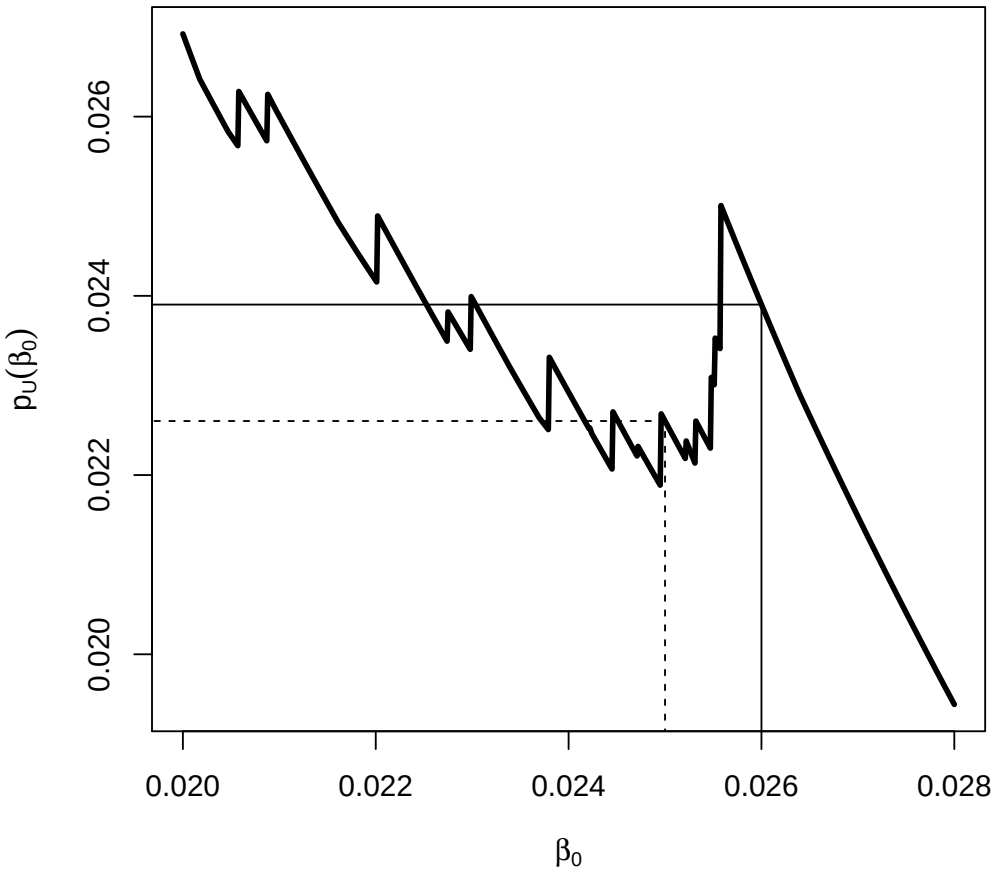

Figure 3: One-sided exact unconditional p-value using the score statistic ordering, $p_U(\mathbf{x}, \beta_0)$. Virtual data example from Röhmel (2005): $x_1/n_1 = 130/248$ and $x_2/n_2 = 76/170$. Thick black line is the p-value function. Thin black lines depict the test of the null $H_0 : \beta \geq 0.026$ and dotted lines depict the test of the null $H_0 : \beta \geq 0.025$.

Sections 4 (unconditional methods) and 5 (conditional methods), including the associated p-values and CIs. In Section 6 we review the melded confidence intervals of Fay et al. (2015), which give associated p-values of the null of $\theta_1 = \theta_2$ that match the one-sided conditional method. In Section 7 we talk about equivalence studies and non-inferiority studies. In Section 8 we discuss non-central confidence intervals and associated tests. In Section 9 we discuss mid-p methods, which are non-asymptotic methods that relax the validity assumption in order to achieve better accuracy. In Section 10 we discuss the computational aspects of the various methods. In Section 11 we review some recent work on causality and the two-sample binomial problem, and relate those results to the rest of this paper. In Section 12 we discuss power and efficiency of methods. In Section 13 we give our final recommendations.

2 Choosing the Effect Measure

Choosing the effect measure is dependent on the application, so we examine a real application to discuss the issues. Coulibaly et al. (2009) studied a parasite called *Mansonella perstans* that infects people in parts of Africa. The usual drugs that kill other similar parasites had not been working on killing *M. perstans*. Coulibaly et al. (2009) realized that in this case there was a symbiotic bacteria, *Wolbachia*, that helped the *M. perstans* live. They suspected that if they gave a common antibiotic, doxycycline, to kill the bacteria, it may in fact help cure the patient of *M. perstans*. To study this, some patients were randomized to the treatment group (received doxycycline) and some to the control group (received no treatment). There are issues of missing data that we will ignore for simplicity. The results are that at 12 months $x_2 = 67$ out of $n_2 = 69$ subjects who received doxycycline had cleared the *M. perstans* from their blood, while only $x_1 = 10$ out of $n_1 = 63$ who got no treatment

cleared the parasite. There are several reasonable choices for how to measure the effect: the difference in clearance rates, the ratio of clearance rates, the ratio of failures, and the odds ratio of clearance rates. Although the choice is often dominated by what is most natural to the intended audience, there are some statistical issues related to this choice.

Without loss of generality, we define the effect measures as measuring how much larger θ_2 is than θ_1. The opposite effect can be measured by switching group labels. But we could also simultaneously switch group labels *and* switch the responses. If the effect remains the same after this double switching, we say that the measure has symmetry equivariance. The measures β_d and β_{or} have symmetry equivariance; however, β_r does not have it, as we demonstrate with the example. Let $\hat{\theta}_2 = 67/69 \approx 0.97$ and $\hat{\theta}_1 = 10/63 \approx 0.16$. An estimate of the rate ratio for success (cleared parasites at 12 months) is $\hat{\theta}_2/\hat{\theta}_1 \approx 6.12$. The rate ratio is often called the relative risk, but in this case the "risk" is the risk of getting cured. A different expression of the same data would be to measure the ratio of the rates of failures (those still having detectable parasites at 12 months). Let $\hat{\theta}_{F2} = 2/69 \approx 0.03$ and $\hat{\theta}_{F1} = 53/63 \approx 0.84$, then an estimate of the relative risk of failure is $\hat{\theta}_{F1}/\hat{\theta}_{F2} \approx 29.0$. In this latter case the control group looks about 29 times worse than the treatment group, while if we look at the rate ratios for success the treatment group looks only about 6 times better than the control group. So how many times better treatment is than control depends on which way we measure risk. This is a violation of symmetry equivariance. Despite this the rate ratio is often used because it is easy to understand (see e.g., Coulibaly et al., 2009), or because it has become the parameter of choice within a field so that its use facilitates comparisons between studies.

The difference has symmetry equivariance. If we measured the difference in rates of disease rather than the difference in rates of cure we get exactly the negative difference as we might expect. Similar to the relative risk, the difference is often used because it is easy

to understand. Additionally, the sample difference in rates is always defined, unlike the ratio which is undefined when $\hat{\theta}_1 = \hat{\theta}_2 = 0$.

Figure 4 gives plots of the three statistics using $\hat{\theta}_2$ and $\hat{\theta}_1$ with $n_1 = n_2 = 8$. The plots go from dark blue ($\hat{\theta}_2$ is larger) to white ($\hat{\theta}_1 = \hat{\theta}_2$) to dark red ($\hat{\theta}_1$ is larger), with black denoting indeterminate. Because of the indeterminate black areas, the ordering of the sample space for the ratio and odds ratio is not straightforward (see Section 4.3). The ordering of the measures on the parameters themselves would give a continuous version of Figure 4, and the black regions would reduce to points at $(\theta_1, \theta_2) = (0, 0)$ or $(1, 1)$. The bottom panels show the lack of symmetry equivariance for the β_r. Comparing the panel for β_{or} with the two different ratio panels, we see that the lower left hand corner of the β_{or} panel is similar to the lower left hand corner of $\hat{\beta}_r = \hat{\theta}_2/\hat{\theta}_1$. For small θ, $\hat{\beta}_{or}$ is a good approximation to $\hat{\beta}_r$. Similarly for both θ values close to 1, $\hat{\beta}_{or}$ is a good approximation of $(1 - \theta_1)/(1 - \theta_2)$ (right bottom panel).

The odds ratio is the more complicated of the three measures, but it has some nice properties. It is very important for the case-control design used to study rare diseases, because the odds ratio of disease given exposure is equal to the odds ratio of exposure given disease (see Breslow, 1996). Also for performing regression on binary observations, logistic regression allows linear predictors to be used to model the log odds, and effects of binary covariates can be expressed as odds ratios. An advantage of the odds ratio for the two-sample binomial case is that by conditioning on the total number of successes in both groups, the probability distribution reduces to a noncentral hypergeometric distribution which is a function of β_{or}. This is discussed more in Section 5.

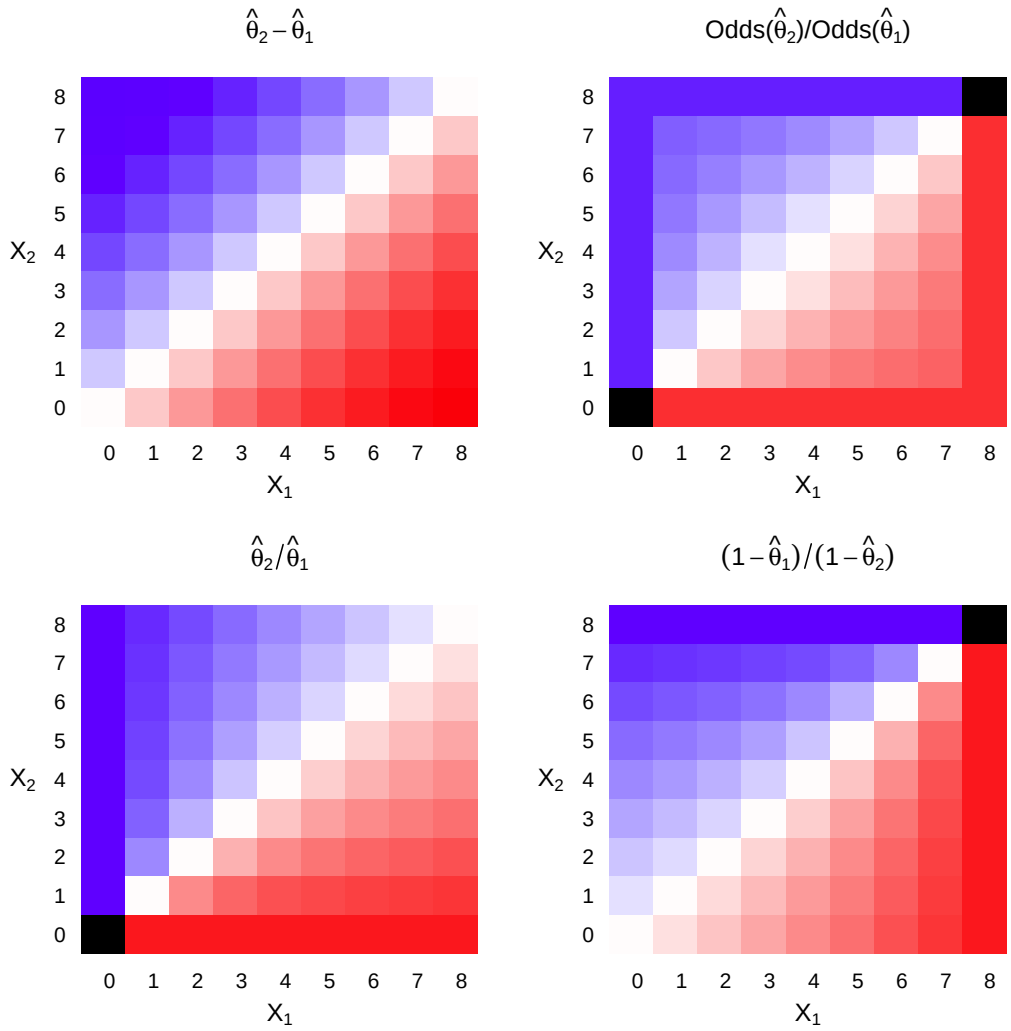

Figure 4: Four simple ordering functions. Dark blue means θ_2 is much larger than θ_1 and dark red is the opposite. White means both treatments appear the same. The functions are based on $n_1 = n_2 = 8$ and using functions of the sample proportions, $\hat{\theta}_1 = x_1/n_1$ and $\hat{\theta}_2 = x_2/n_2$. The sample space is depicted by a 9×9 grid of responses, ranked by the ordering functions: difference in success proportions (upper left), odds ratio (upper right), ratio of success proportions (lower left), and ratio of failure proportions (lower right). Colors rank the functions from the highest values (dark blue) indicating larger θ_2, to middle values (white) indicating $\theta_1 = \theta_2$, to lowest values (dark red), with black indicating no information.

3 Properties of Methods

3.1 Defining a Matched Method

Once we choose an effect measure, we choose an appropriate method (an estimator, confidence interval, and p-value function) for inferences. Deciding the estimator is not the focus of this paper. We will not specify the estimator except to require that it is within the confidence interval. We focus mostly on choosing the CI and p-value function. Except in Section 9, we only consider methods that are valid (i.e., the CI and p-value are both valid) and reproducible. Because we require reproducibility, the method based on the UMP unbiased (and randomized) test is not allowed. Although one could define a method where the p-value function and confidence interval are derived from different procedures, for focus we will not consider those kinds of methods in this paper. We define a matched method as one where its confidence interval is derived from its p-value function or vise versa. A matched method is slightly different from a unified method. For example, as we have shown in Section 1.3, for some p-value functions it is not possible to get a confidence interval to give unified inferences. Loosely speaking, if we start with a valid p-value function, the matching CI is the valid CI that gives unified inferences as much as is possible, and vice versa if we start with a valid CI.

Here is a precise definition of a matched method. If we start with $p(\mathbf{x}, \beta_0)$, an associated confidence region is given by equation 1, and the matching CI is smallest interval that contains that confidence region. In other words, if the confidence region has holes in it, then those holes are "filled in". On the other hand, if we start with $(L, U) = C_I(\mathbf{x}, 1-\alpha)$, then the matching p-value function is the smallest α such that β_0 is outside $C_I(\mathbf{x}, 1-a)$

for all $a \geq \alpha$, or more precisely,

$$p(\mathbf{x}, \beta_0) \quad = \quad \begin{cases} 1 & \text{if } A = \emptyset \\ \min A & \text{otherwise,} \end{cases} \tag{2}$$

where $A \equiv A(\mathbf{x}, \beta_0)$ is the set

$$A(\mathbf{x}, \beta_0) \quad = \quad \{\alpha : \beta_0 \notin C_I(\mathbf{x}, 1-a) \text{for all } 1 > a \geq \alpha\}.$$

3.2 Implications of Unified Inferences

Theorem 3.1 *Consider a valid, reproducible, and matched method. The method has unified inferences*

1. *if and only if the CI is equal to the confidence region associated with the p-value, and*
2. *only if the CI is nested, and*
3. *only if the the p-value function is coherent (for one-sided p-values), or directionally coherent (for two-sided p-values).*

The formal proof of the theorem is in the Appendix. The theorem says we must have nested CIs and coherent p-values in order to have unified inferences. These ideas are best understood graphically. Figure 1 shows lack of directional coherence; for every β_0 there is only one p-value, and the two-sided p-value function is not unimodal. Similarly, Figure 3 shows lack of coherence. Figure 2 shows non-nestedness; for every α there is only one lower limit, and the lower limit is not a monotonic function of the level.

3.3 Directional Inferences

Typically, if a researcher finds a significant difference from the two-sided p-value suggesting that $\beta \neq \beta_0$, they almost always are interested in interpreting the result in terms of whether $\beta > \beta_0$ or $\beta < \beta_0$. In other words, the two-sided hypothesis test is often treated as a three-decision rule: (1) fail to reject $\beta = \beta_0$, (2) reject $\beta = \beta_0$ and conclude $\beta > \beta_0$, or (3) reject $\beta = \beta_0$ and conclude $\beta < \beta_0$. If the two-sided p-value has directional coherence, then if we reject $H_0 : \beta = \beta_0$ at level α, we can additionally reject at level α either $H_0 : \beta \leq \beta_0$ (if $\beta_0 < \hat{\beta}$) or $H_0 : \beta \geq \beta_0$ (if $\beta_0 > \hat{\beta}$).

Consider comparing two unified methods, one with a central CI, and one with a non-central CI. For the non-central method a two-sided hypothesis may be be slightly more powerful, but if the non-central method is applied also to a subsequent one-sided hypothesis (as in the three decision rule), it can be quite a bit less powerful than the central one. To see this, start with a nested central CI, say (L, U), and pair it with its matching two-sided p-value, say p_C. By Theorem 3.1, this means that whenever the $100(1-\alpha)\%$ CI excludes β_0 then $p_C \leq \alpha$, and we can reject $H_0 : \beta = \beta_0$ at level α. After rejecting the two-sided hypothesis at level α, we can reject one of the one-sided hypotheses at level $\alpha/2$; if $\beta_0 < L$ we reject $H_0 : \beta \leq \beta_0$, while if $\beta_0 > U$ we reject $H_0 : \beta \geq \beta_0$. A non-central CI does not allow one-sided rejections at the $\alpha/2$ level. Freedman (2008) discusses this issue in terms of clinical trials, and using these arguments as well as some Bayesian motivation recommends performing two one-sided tests at the $\alpha/2$ level, which is another way of describing the use of central CI methods for three decision rules.

In summary, if we desire directional inferences, and we want to compare the power to detect a one-sided effect in a fair way, then we need to compare a method with a two-sided p-value and its matching $100(1-2\alpha)\%$ non-central CI, with a pair of one-sided p-values and its matching $100(1-\alpha)\%$ central CI. This means that when comparing lengths of CIs,

if directionality of effect is important, we should compare the length of a $100(1-2\alpha)\%$ non-central CI with the length of a $100(1-\alpha)\%$ central CI. Because directionality is usually important, our default recommendation is to use central confidence intervals and perform three-sided inferences as described above.

4 Methods for Creating One-Sided Exact Unconditional Testing Procedures

4.1 Basic Procedure for Defining p-values

Suppose larger θ is better. We want to know if treatment 2 is better than treatment 1 ($\theta_2 > \theta_1$), and if so by how much. Let $T(\mathbf{x})$ be a function of the data, where larger values of $T(\mathbf{x})$ indicate that treatment 2 is better than treatment 1, and $T(\mathbf{X})$ is defined for all possible values of $\mathbf{X}$. For example, a simple $T(\mathbf{x})$ is the difference in observed proportions (see Figure 4 upper left). For this section and the next (Section 4.2), we require that T is a function of $\mathbf{x}$ only. Later in Section 4.5 T may depend on α, and in Section 4.6 T may depend on β_0. Barnard (1947) outlined convexity conditions which ensure that larger values of T suggest treatment 2 is better. Barnard's convexity (BC) conditions are:

$$\begin{aligned} &\text{if } x_2^* > x_2 \quad \text{then} \quad T([x_1, x_2^*]) \geq T([x_1, x_2]) \\ &\qquad\qquad \text{and} \\ &\text{if } x_1^* < x_1 \quad \text{then} \quad T([x_1^*, x_2]) \geq T([x_1, x_2]). \end{aligned} \tag{3}$$

Even within functions that satisfy the BC conditions, there are many choices. In later sections we explore choice of T further, but for now imagine the simple ordering function of $T(\mathbf{x}) = \hat{\theta}_2 - \hat{\theta}_1$ plotted in Figure 4a, which meets the BC conditions.

Once we have decided on the ordering function, T, we can create valid unconditional one-sided p-values: p_U for testing the null H_{U0} (defined as $H_0 : \beta \geq \beta_0$) and p_L for testing H_{L0} ($H_0 : \beta \leq \beta_0$) using

$$\begin{aligned} p_U(\mathbf{x}, \beta_0) &= \sup_{\theta: b(\theta) \geq \beta_0} P_\theta \left[T(\mathbf{X}) \leq T(\mathbf{x}) \right] \\ &\text{and} \\ p_L(\mathbf{x}, \beta_0) &= \sup_{\theta: b(\theta) \leq \beta_0} P_\theta \left[T(\mathbf{X}) \geq T(\mathbf{x}) \right]. \end{aligned} \tag{4}$$

These p-values are valid since

$$\sup_{\theta \in \Theta_0} P_\theta [p(\mathbf{X}, \beta_0) \leq p(\mathbf{x}, \beta_0)] \leq p(\mathbf{x}, \beta_0)$$

where $\Theta_0 = \{\theta : b(\theta) \geq \beta_0\}$ for p_U and $\Theta_0 = \{\theta : b(\theta) \leq \beta_0\}$ for p_L. Further, any other valid p-values that retain the same ordering are inadmissible (that is, they have values that are never less than the valid unconditional p-values and are greater for at least one $\mathbf{x}$) (Lloyd, 2008, p. 333).

These valid one-sided p-values can be inverted to create two $100(1 - \alpha/2)$ one-sided confidence limits using

$$\begin{aligned} U(\mathbf{x}) &= \begin{cases} \sup \{\beta_0 : \quad p_U(\mathbf{x}, \beta_0) > \alpha/2\}, \text{ if } \exists \text{ a } \beta_0 \\ \qquad\qquad \text{with } p_U(\mathbf{x}, \beta_0) > \alpha/2 \\ \beta_{max} \qquad \text{otherwise} \end{cases} \\ &\text{and} \\ L(\mathbf{x}) &= \begin{cases} \inf \{\beta_0 : \quad p_L(\mathbf{x}, \beta_0) > \alpha/2\}, \text{if } \exists \text{ a } \beta_0 \\ \qquad\qquad \text{with } p_L(\mathbf{x}, \beta_0) > \alpha/2 \\ \beta_{min} \qquad \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

where $(\beta_{min}, \beta_{max}) = (-1, 1)$ for β_d and $(0, \infty)$ for β_r or β_{or}. A central $100(1-\alpha)$ confidence interval is the union of the one-sided ones, $(L(\mathbf{x}), U(\mathbf{x}))$, and a central p-value is $p_C(\mathbf{x}, \beta_0) =$

$\min(1, 2p_L, 2p_U)$. These confidence limits are called exact unconditional (see e.g., Mehrotra et al., 2003) or Buehler confidence limits (see Lloyd and Kabaila, 2003). Lloyd and Kabaila (2003) and Wang (2010) show two results about these one-sided intervals. First, the lower and upper one-sided confidence limits retain a logical ordering analogous to Barnard's convexity conditions. Specifically, $(L, U) \in \mathcal{O}_T$, where $\mathcal{O}_T$ is the class of valid central confidence intervals such that if $T(\mathbf{x}_1) < T(\mathbf{x}_2)$ then $L(\mathbf{x}_1) \leq L(\mathbf{x}_2)$ and $U(\mathbf{x}_1) \leq U(\mathbf{x}_2)$. Second, (L, U) calculated in this manner is the smallest confidence interval within $\mathcal{O}_T$. In other words, any other valid central confidence interval (L^*, U^*) in $\mathcal{O}_T$ must have $L^*(\mathbf{x}) \leq L(\mathbf{x})$ and $U(\mathbf{x}) \leq U^*(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

Barnard (1947) proposed a test he called the CSM test based on an ordering function that starts from the most extreme point, and adds points as the ones that have the lowest valid unconditional p-value among those that meet the BC condition and the symmetry equivariance condition. The p-value function used could be the p_C related to testing $\theta_2 = \theta_1$, e.g., $p_C(\mathbf{x}, 0)$ for testing $\beta_d = 0$. Additionally, Barnard (1945) outlined the general exact unconditional test, and those tests are sometimes referred to as "Barnard's test" (see e.g., Sas, 2012; Cytel, 2010), but we avoid that terminology to avoid confusion with Barnard's CSM test. Röhmel and Kieser (2013) discussed one-sided exact unconditional tests using Barnard's CSM p-value ordering, except with breaking more ties to get higher power, an idea discussed in the next section.

4.2 Improving Power by Breaking Ties: Refinement of Ordering Functions

One important way to improve the power of some unconditional exact tests based on a function T is to break any ties that exist in the ordering function. If T is an ordering

function with ties, and T^* is an ordering function that gives the same ordering of T at all the untied values and additionally breaks some ties, then we say T^* is a *refinement* of T. Then the unconditional exact p-values formed with T^* are always less than or equal to those formed with T (see Röhmel and Mansmann, 1999, p. 158). Similarly, one-sided exact unconditional lower confidence limits formed using T^* are always at least as large as the ones formed using T (Kabaila and Lloyd, 2006; Wang, 2010).

We describe one specific refinement or tie breaking algorithm for the difference in proportions next, which as far as we are aware, has not been specifically described in the literature and has not been available in software (although there are some closely related methods). We can order within each set of tied values using Wald statistics for $\hat{\beta}_d$, i.e., ordering by

$$Z(\mathbf{x}) \quad = \quad \frac{\hat{\beta}_d}{\sqrt{\widehat{var}_0(\hat{\beta}_d)}} = \frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{\hat{\theta}(1-\hat{\theta})(1/(n_1+n_2))}}$$

where $\hat{\theta} = (x_1 + x_2)/(n_1 + n_2)$. This leaves the ties for $\hat{\beta}_d = 0$, but otherwise defines points with more precision as more extreme, where extreme is further away from zero. Not all the values with $\hat{\beta}_d \neq 0$ break all the ties. For example, consider the ties at $\hat{\beta}_d = 5/8$ that happen at the $\mathbf{x}$ values $[0, 5]$, $[1, 6]$, $[2, 7]$, and $[3, 8]$, for $n_1 = n_2 = 8$. This method still leaves tied the two pairs of points, $\{[0, 5], [3, 8]\}$ and $\{[1, 6], [2, 7]\}$. These remaining ties we argue should remain tied in order for the ordering to retain symmetry equivariance. Note that this suggested ordering is similar, but not equivalent to just ordering the entire sample space by $Z(\mathbf{x})$ (as was studied in Mehrotra et al., 2003).

If we break the ties in this way, then the BC conditions are still met, because only at the boundaries (where the ties are broken according to the BC conditions) do the ties occur at two points $\mathbf{x}_a$ and $\mathbf{x}_b$ with $x_{a1} = x_{b1}$ or $x_{a2} = x_{b2}$. All of the other ties will not have any $x_{a1} = x_{b1}$ or $x_{a2} = x_{b2}$ so they can be broken in any manner and the overall

ordering function, T^*, will meet the BC conditions. This is important for computation (see Section 10). Further, the proposed T^* (tie-breaking on difference in proportions) does not depend on α or β_0 like some score test based methods (see Section 4.6) so avoids problems with nesting and coherence.

4.3 Ordering Functions for Ratio and Odds Ratio

Performing exact unconditional tests on β_r or β_{or} is not straightforward. We consider β_r first since it is simpler. One problem is that if we observe $\mathbf{x} = [0, 0]$, this could occur with high probability if the true ratio was 100 or if it was 1/100 as long as both θ_1 and θ_2 were very small. So if $T(\mathbf{x})$ is designed so that larger values suggest $\theta_2 > \theta_1$, it is not clear how to define $T([0, 0])$ if our interest is in β_r.

Since $\mathbf{x} = [0, 0]$ gives us no information about β_r, we must deal with that point in a special way; we set the p-value at $\mathbf{x} = [0, 0]$ to 1 for tests of β_r regardless of the null hypothesis. This means that $\mathbf{x} = [0, 0]$ is placed "deepest" within the null. Following equations 4, this implies $T([0, 0])$ can be thought of as the largest value when calculating $p_U(\mathbf{x}, \beta_0)$ and the smallest value when calculating $p_L(\mathbf{x}, \beta_0)$. A similar issue applies to the odds ratio, except in that case in addition to $\mathbf{x} = [0, 0]$, the point $\mathbf{x} = [n_1, n_2]$ also has no information about β_{or}.

For clarity, we rewrite equations 4 applied to all three parameters. Let $\mathcal{X}_I$ denote the set of $\mathbf{X}$ values with information about β. Then if $\mathbf{x} \notin \mathcal{X}_I$ set $p_U(\mathbf{x}, \beta_0)$ and $p_L(\mathbf{x}, \beta_0)$ to 1, otherwise let $p_U(\mathbf{x}, \beta_0)$ be

$$\sup_{\theta: b(\theta) \geq \beta_0} P_\theta \left[T(\mathbf{X}) \leq T(\mathbf{x}) | \mathbf{X} \in \mathcal{X}_I \right] P_\theta \left[\mathbf{X} \in \mathcal{X}_I \right]$$

and analogously, let $p_L(\mathbf{x}, \beta_0)$ be

$$\sup_{\theta: b(\theta) \leq \beta_0} P_\theta \left[T(\mathbf{X}) \geq T(\mathbf{x}) | \mathbf{X} \in \mathcal{X}_I \right] P_\theta \left[\mathbf{X} \in \mathcal{X}_I \right].$$

Since we never reject when $\mathbf{x} \notin \mathcal{X}_I$, these definitions give valid p-values, and additionally when $\mathbf{x} \notin \mathcal{X}_I$ we do not need to define $T(\mathbf{x})$.

The simple ordering function by the estimate of β_r or β_{or} (even when using a tie breaking ordering similar to what was done for β_d) is not very powerful (see Section 12), and is not recommended. Typically, we order using a score function (see Section 4.6) since it gives more reasonable power.

4.4 Other Improvements: E+M and Berger-Boos

Another method to apparently improve the ordering statistic for any efficacy parameter (difference, ratio, or odds ratio) is the estimated and maximized $(E + M)$ p-value (Lloyd, 2008). In this method, we replace an ordering statistic, T, with T^*, where T^* is an estimated p-value when testing H_{L0} (or the negative estimated p-value when testing H_{U0}). We estimate the p-value by plugging in $\hat{\theta}_0$ instead of taking the supremum of θ under the null, where $\hat{\theta}_0$ is the maximum likelihood estimator of $\theta \in \Theta_0$. For example, the approximation for p_L in expression 4 uses $\hat{p}_L(\mathbf{x}, \beta_0) = P_{\hat{\theta}_0} \left[T(\mathbf{X}) \leq T(\mathbf{x}) \right]$. Then we "maximize" using $T^*(\mathbf{x}) = \hat{p}_L(\mathbf{x}, \beta_0)$ instead of T as the ordering function, that is, we calculate the exact conditional p-value using expression 4 by taking the supremum. Lloyd (2008) studied this method and observed that when T^* (the approximate p-value) is used as the ordering statistic, the resulting exact unconditional p-value is generally smaller than the exact unconditional p-value on T. The process can be repeated (replace T^* by its approximate p-value), but the additional reduction appears to be minimal.

Berger and Boos (1994) introduced a popular method that tends to reduce exact unconditional p-values. Instead of taking the supremum over the entire null hypotheses parameter space take the supremum only over C_γ, a $100(1-\gamma)\%$ confidence set of θ restricted to be in the null space, then add γ to ensure validity. This is usually done by reexpressing the parameter space (θ_1, θ_2) as (β, ψ), where ψ is a nuisance parameter, then defining C_γ as the intersection of $\theta \in \Theta_0$ and the set of θ values with ψ in its $100(1-\gamma)\%$ confidence interval. A Berger-Boos version of p_U of expression 4, uses

$$p_{U\gamma}(\mathbf{x}, \beta_0) \quad = \quad \gamma + \sup_{\theta \in C_\gamma} P_\theta \left[T(\mathbf{X}) \geq T(\mathbf{x}) \right].$$

This is not optimal, since we may be able to improve it by using $p_{U\gamma}(\mathbf{x}, \beta_0)$ as an ordering function. Nevertheless, it usually provides some reduction in p-values (see e.g., Lloyd, 2008).

4.5 Ordering Functions That Depend on Significance Level

Kabaila and Lloyd (2003) showed that for one-sided $100(1-\alpha/2)\%$ exact unconditional upper confidence limit, the ordering function, T, that maximizes the asymptotic efficiency is an approximate $100(1-\alpha/2)\%$ one-sided upper confidence limit itself. This means that you would use a different ordering function for the upper and lower limit, and in fact would use a different ordering function for different confidence levels.

Wang (2010) and Wang and Shan (2015) also proposed an ordering function to give the smallest CI, and the calculation of the ordering function itself is iterative and quite involved, similar to the CSM test of Barnard (1947). The precise definition of the ordering is notationally cumbersome, but the idea is roughly as follows. Consider the lower $100(1-\alpha/2)$ one-sided limit. Start from the most extreme point $\mathbf{x} = [0, n_2]$. Then add points one at a time, picking the point, $\mathbf{x}_a$, that gives the largest $L(\mathbf{x}_a, 1-\alpha/2)$ and belongs to the

set of closest neighboring points with the already included points, where closest neighbor is defined in terms of the BC conditions. The algorithm ensures that the lower limit function meets the BC conditions. Because each added $L(\mathbf{x})$ value is as large as possible, this ordering ensures that if the resulting ordering function T gives the finest partition (there are no ties), then any valid $100(1-\alpha/2)\%$ one-sided lower limit that meets the BC conditions and uses T for ordering, say L^*, has $L^*(\mathbf{x}) \leq L(\mathbf{x})$ for all $\mathbf{x}$ (see Wang, 2010; Wang and Shan, 2015).

Although we obtain this optimality property, the price is that the ordering function depends on α. Thus, we can have different ordering functions for different α, which can lead to non-nestedness (see Figure 2).

4.6 Ordering Functions That Depend on Hypothesis Boundaries

Basing the ordering statistic on a score test can increase the power over using simple Wald-type Z statistics (see Chan, 2003). Although this increased power has been shown in several simulation studies, it is not clear whether the increase in power is due to the fewer ties for the score test, or from some other difference between the ordering statistics. A problem with the score statistic is that the induced ordering may change based on the β_0. This can produce non-coherence as was shown in Section 1.3 and Figure 3.

5 One-Sided Conditional Exact Tests

Yates (1984) argues that conditioning on total number of failures is the proper strategy for this problem, and most of the discussants of the paper agreed with this (including Barnard, who first suggested the unconditional approach). One of the main reasons that others had recommended the unconditional approach is an overemphasis on the fixed significance

level and the resulting power, which when used leads to more power for the unconditional tests because the sample space has more values and hence is less discrete. Yates (1984) argues (in his Section 9) that over reliance on the nominal significance level is not a good reason to prefer the unconditional test, and that p-values should be reported instead of accept/reject decisions. Yates (1984) also argues that we should condition on the total number of events (X_1+X_2), because that statistic is approximately ancillary to the effects of interest. Recent reviews (e.g., Lydersen et al., 2009) have emphasized power arguments, and we review the choice of test from that perspective in Section 12. Historically, conditional tests have been important because of their much smaller computational burden compared to unconditional tests. The computational burden for unconditional tests has become less important, although for some applications it may be a non-trivial concern (e.g., big data applications with small sample sizes but very many covariates being tested).

For the unconditional one-sided exact method, to calculate p-values we need to take the supremum of the probability that $T(\mathbf{X})$ is more extreme than the observed $T(\mathbf{x})$ over the parameter space Θ_0 (see e.g., equation 4). This is a difficult calculation (see Section 10). An alternative method is to condition on the sum $s = x_1 + x_2$, and calculate the conditional probability. The resulting conditional distribution is the extended hypergeometric distribution (Johnson et al., 2005) also called Fisher's noncentral hypergeometric distribution (Fog, 2008), which depends only on β_{or}. Additionally, because s is fixed we can write the ordering function in terms of X_2 only. In fact, the only unique ordering function that makes sense and meets the BC conditions is X_2 itself (ordering on $n_1 - X_1$ will be equivalent). So this simplifies the calculations if the effect measure is β_{or}. For example, for testing $H_0 : \beta_{or} \geq \beta_0$ use

$$\begin{aligned} p_{Uc}(\mathbf{x}, \beta_0) &= \sup_{\theta \in \Theta_0} P_\theta \left[T(\mathbf{X}) \geq T(\mathbf{x}) | S \right] = \sup_{\beta_{or}: \beta_{or} \geq \beta_0} P_{\beta_{or}} \left[X_2 \geq x_2 | S \right] \\ &= P_{\beta_0} \left[X_2 \geq x_2 | S \right], \end{aligned} \tag{6}$$

where the last step comes because the conditional distribution is monotone in β_{or} (Mehta et al., 1985). The other conditional one-sided p-value, p_{Lc} is calculated similarly except by reversing the inequality. These conditional p-values for testing $H_0 : \beta_{or} = 1$ (or equivalently $H_0 : \theta_1 = \theta_2$) are Fisher's exact one-sided p-values. We calculate the central confidence intervals on β_{or} using equation 5 except using the conditional exact one-sided intervals instead of the unconditional ones.

Now consider the other measures, β_d and β_r. At the boundary of equality, the one-sided hypotheses are equivalent. For example, the following three null hypotheses give equivalent Θ_0: (odds ratio) $H_{0U} : \beta_{or} \geq 1$, (ratio) $H_{0U} : \beta_r \geq 1$, and (difference) $H_{0U} : \beta_d \geq 0$. Analogously for the other one-sided p-value. But for boundaries not representing equality, Θ_0 changes depending on the effect measure. The simplification of the p-value calculation only works for the odds ratio. For example, for the difference in proportions (i.e., $\beta = \beta_d$) there is not simplification analogous to equation 6. Figure 5 shows that the exact one-sided conditional confidence limit on β_d is not efficient, because the conditional distribution depends on β_{or}. The upper $100(1 - \alpha/2)\%$ limit for β_d, say U_d, based on the upper limit for β_{or}, say U_{or}, is (see Santner and Snell, 1980, Section 2)

$$U_d = \begin{cases} 0 & \text{if } U_{or} \leq 1 \\ \frac{\sqrt{U_{or}}-1}{\sqrt{U_{or}}+1} & \text{if } U_{or} > 1 \end{cases}$$

There are better ways to get confidence intervals on β_d and β_r that provide matching inferences for the one-sided p-values with β_0 representing $\theta_1 = \theta_2$. We show these in the next section.

6 Melded Confidence Intervals

Fay et al. (2015) developed melded confidence intervals, a general method for creating

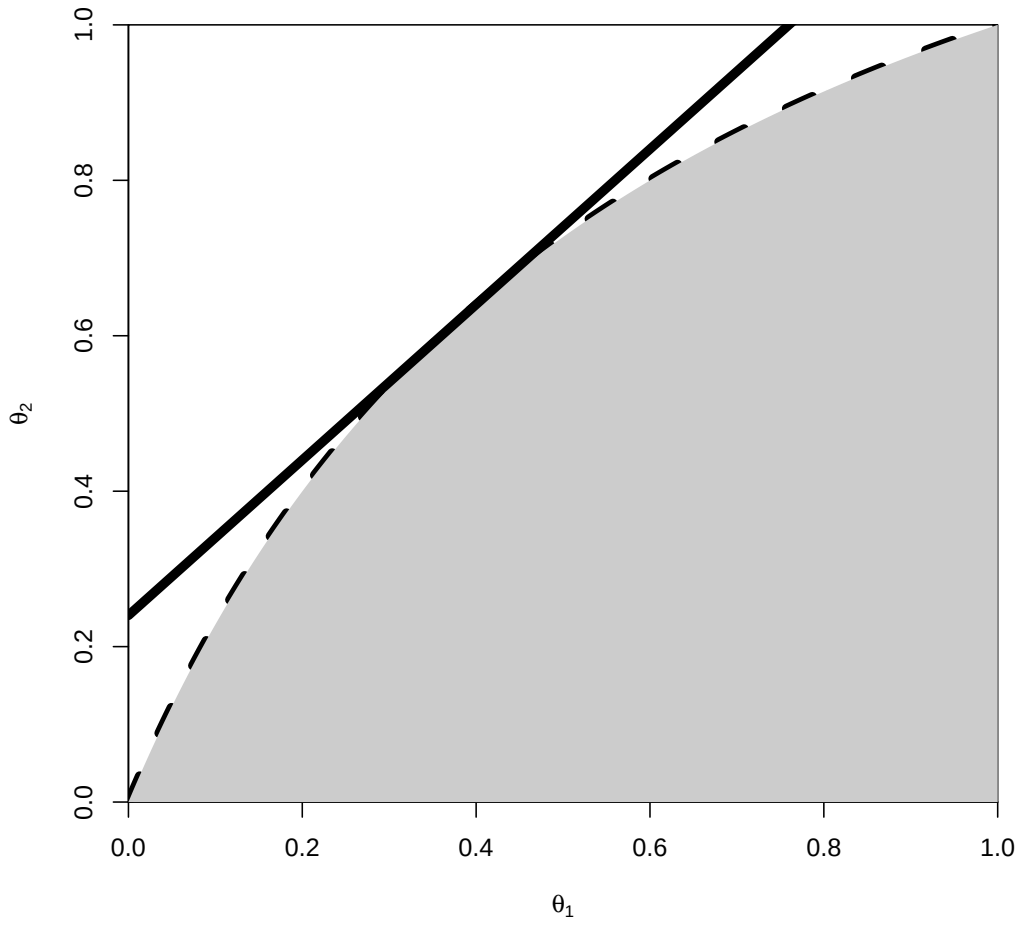

Figure 5: 97.5% Confidence region based on one-sided conditional test of odds ratio (gray shaded area). Data is $x1/n_1 = 4/12$ and $x_2/n_2 = 8/15$. Upper 97.5% exact conditional limit on β_{or} is $U = 2.664$ (dotted line) and on β_d is $U = 0.240$ (solid line). The confidence region based on the upper limit for β_d is the gray region plus the white space between the dotted and solid line. We see that because the conditional probability depends on β_{or} alone, that white space represents the lack of efficiency of basing the confidence region on β_d instead of β_{or}.

confidence intervals for the two-sample case, that is closely related to the confidence distribution (CD) approach (Xie and Singh, 2013). Roughly, the $100(1-\alpha)\%$ melded confidence interval is a central confidence interval that takes the middle $100(1-\alpha)\%$ of a function of random variables, each created from one-sided confidence intervals.

Let $L_{\theta_a}(\mathbf{x}, 1-\alpha/2)$ and $U_{\theta_a}(\mathbf{x}, 1-\alpha/2)$ be exact nested $100(1-\alpha/2)\%$ one-sided confidence limits, for θ_a for $a = 1, 2$. The lower and upper CD random variables for group a are $W_{La} = L_{\theta_a}(\mathbf{x}, A_{a1})$ and $W_{Ua} = U_{\theta_a}(\mathbf{x}, A_{a2})$, where A_{ai} are independent uniform random variables. This gives, $W_{La} \sim Beta(x_a, n_a - x_a + 1)$ with expectation $x_a/(n_a+1)$, and $W_{Ua} \sim Beta(x_a+1, n_a - x_a)$ with expectation $(x_a+1)/(n_a+1)$, and using limits of parameters going to zero we define $Beta(0, n+1)$ as a point mass at 0 and $Beta(n+1, 0)$ as a point mass at 1. If the responses were normally distributed, then the lower and upper CD random variables would be identical, but for the binomial case (and for discrete random variables in general) the lower and upper CD random variables (CD-RVs) are different – the lower CD-RV is stochastically smaller than the upper CD-RV. To get a melded confidence intervals on $b(\theta)$, Fay et al. (2015) require that $b(\theta)$ is a monotonic function of the parameters. Suppose $\beta = b(\theta)$ is increasing in θ_2 and decreasing in θ_1, such as our examples: β_d, β_r and β_{or}. Then the $100(1-\alpha)\%$ (two-sided) melded confidence interval is given by

$$\left(q\left\{b([W_{U1}, W_{L2}]), \alpha/2\right\}, q\left\{b([W_{L1}, W_{U2}]), 1-\alpha/2\right\}\right).$$

where $q(Y, a)$ is the ath quantile of a random variable Y. The interval is designed conservatively by using $[W_{U1}, W_{L2}]$ for the lower limit, but $[W_{L1}, W_{U2}]$ for the upper limit. Fay et al. (2015) conjectured that if the one-sample confidence interval procedures are valid, central, and nested, and $\beta(\theta)$ is monotonic within each parameter, then the melded confidence interval is valid, nested and central. Some mathematical results, simulations in several situations, and extensive numeric calculations in the binomial case supported this conjecture. A rigorous proof of the conjecture is still needed.

Let $p_{Um}(\mathbf{x}, \beta_0)$ and $p_{Lm}(\mathbf{x}, \beta_0)$ be the one-sided melded p-values, the p-values that match with the one-sided melded confidence limits. Then for the binomial case, Fay et al. (2015) showed that the one-sided melded p-values equal the exact one-sided conditional p-values when testing the null with margin β_0 which implies $\theta_1 = \theta_2$. For example, for testing $H_0 : \beta_d \geq 0$, we have $p_{Um}(\mathbf{x}, 0) = p_{Uc}(\mathbf{x}, 0)$, and for testing $H_0 : \beta_r \geq 1$, we have $p_{Um}(\mathbf{x}, 1) = p_{Uc}(\mathbf{x}, 1)$. This means that the melded confidence intervals match the p-values from the one-sided Fisher's exact test.

The melded CIs for β_{or} are very close to the exact conditional ones, but the melded CIs for β_d are more efficient (lower are larger, and upper are smaller) than the exact conditional ones (see Figure 6).

7 Noninferiority and Equivalence Hypotheses

Two other types of hypotheses are noninferiority and equivalence hypotheses. Suppose we are comparing two treatments, and larger β means that the new treatment is better than the standard one. Let β_0 denote $\theta_1 = \theta_2$ and define an equivalence region as $\beta_{M_L} < \beta_0 < \beta_{M_U}$. For $\beta \in (\beta_{M_L}, \beta_0)$, although the standard treatment is better, the difference between the two is not substantial enough to be of practical importance. When we reject the one-sided hypothesis, $H_0 : \beta \leq \beta_{M_L}$ versus $H_1 : \beta > \beta_{M_L}$, we declare the new treatment noninferior. This is just an "alternative is greater" one-sided hypothesis already discussed in Section 1.1. The equivalence hypothesis, however, is qualitatively different,

$$\begin{aligned} H_0 : &\quad \beta \leq \beta_{M_L} \text{ or } \beta_{M_U} \leq \beta \\ H_1 : &\quad \beta_{M_L} < \beta < \beta_{M_U}. \end{aligned}$$

Just as the two-sided hypothesis is often treated as a three decision rule (see Section 3.3),

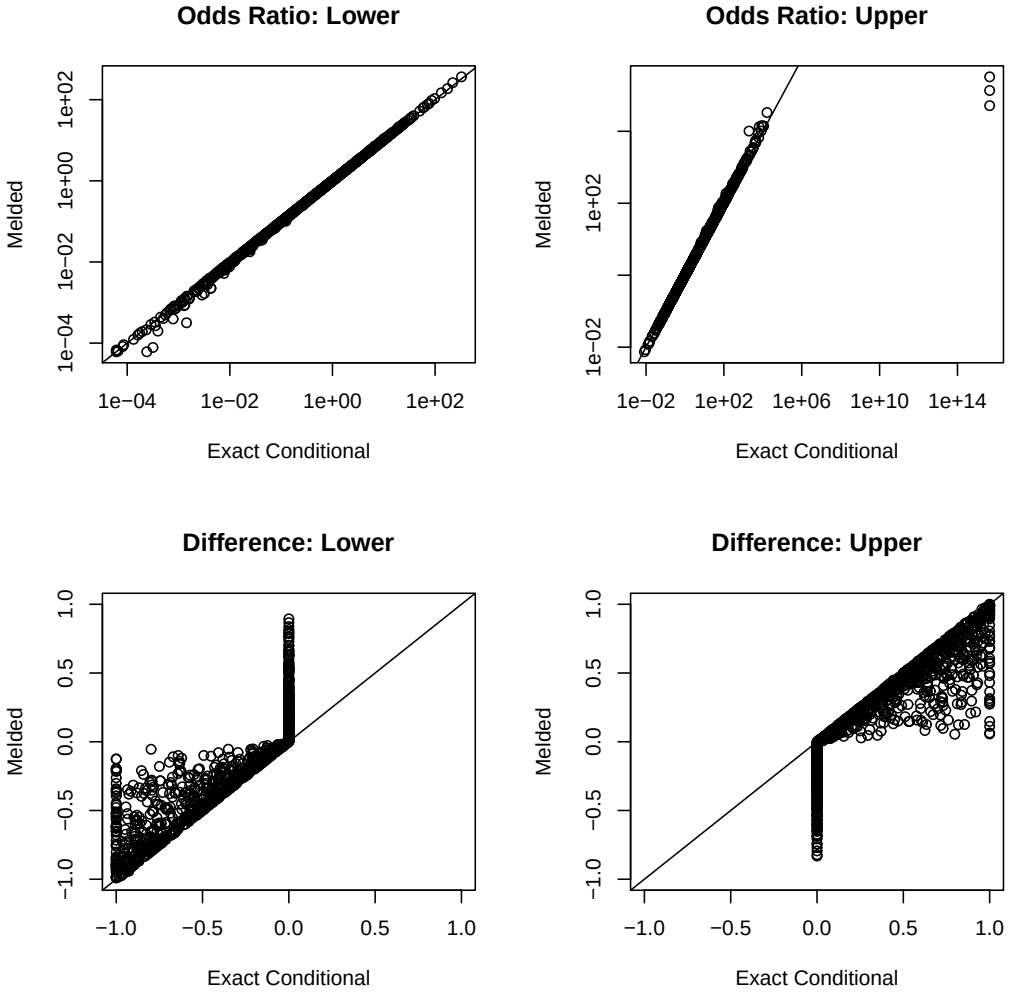

Figure 6: Lower and Upper limits associated with 95% central confidence intervals by exact conditional method and melding method. Simulated data where n_a is simulated from uniform on 1 to 100, and x_a is uniform on 0 to n_a, 1000 replications. Calculation used the exact2x2 R package for melded confidence limits and fisher.test from the stats package for the exact conditional limits. The limits for β_{or} agree well, except for some extreme data (e.g., $x_1/n_1 = 1/68$ and $x_2/n_2 = 57/61$) perhaps caused by numeric issues in the computation, while the limits for β_d show that the melded are smaller intervals (lower is larger, upper is smaller).

the statement after testing an equivalence hypothesis may be more expansive than either reject or fail to reject the non-equivalence. Let (L, U) be a valid central nested $100(1-\alpha)\%$ CI. Then we can make the following declarations based on the relationship between (L, U) and β_{M_L} and β_{M_U}:

- if $\beta_{M_L} < L < U < \beta_{M_U}$ declare equivalence at level α,
- if $\beta_{M_L} < L < \beta_{M_U} < U$ declare noninferiority at level $\alpha/2$,
- if $\beta_{M_L} < \beta_{M_U} < L < U$ declare (substantial) superiority at level $\alpha/2$, or
- if $L < U < \beta_{M_L} < \beta_{M_U}$ declare (substantial) inferiority at level $\alpha/2$.

The last three statements are valid because of the centrality (see e.g., Goeman et al., 2010, for a similar statement).

8 Non-central Confidence Intervals and Associated Tests

Let $T_{ts}(\mathbf{x}) \equiv T_{ts}(\mathbf{x}, \alpha, \beta_0)$ be an ordering distribution for testing the two-sided null $H_0 : \beta = \beta_0$, with smaller values suggesting β further away from the null. Then we can create exact unconditional two-sided p-values using

$$p_{ts}(\mathbf{x}, \beta_0) \quad = \quad \sup_{\theta \in \Theta(\beta_0)} P_\theta \left[T_{ts}(\mathbf{X}) \leq T_{ts}(\mathbf{x}) \right]$$

and exact conditional two-sided p-values using

$$p_{ts}(\mathbf{x}, \beta_0) \quad = \quad \sup_{\theta \in \Theta(\beta_0)} P_\theta \left[T_{ts}(\mathbf{X}) \leq T_{ts}(\mathbf{x}) | S = s \right].$$

which simplifies to

$$p_{ts}(\mathbf{x}, \beta_0) = P_{\beta_0} \left[T_{ts}(\mathbf{X}) \leq T_{ts}(\mathbf{x}) | S = s \right], \tag{7}$$

if $\beta = \beta_{or}$.

For example, consider $T_{ts}(\mathbf{x}, \beta_0) = f(\mathbf{x}, \beta_0)$, where f is the probability mass function for the extended hypergeometric distribution with parameter $\beta_{or} = \beta_0$. Then the associated exact conditional p-value is the usual Fisher's exact test, which we call the Fisher-Irwin test since it was proposed by Irwin (1935) and to distinguish it from the central Fisher's exact test created by doubling the minimum of the one-sided Fisher's exact p-values. Using Fisher's exact p-values (either Fisher-Irwin or central version) as an ordering function in an unconditional exact test gives a version of Boschloo's test. Boschloo (1970) showed that using the Fisher-Irwin p-values in this way is uniformly more powerful than the Fisher-Irwin test. This superiority in power holds for both one-sided tests and central tests Lydersen et al. (2009).

Blaker (2000) studied non-central confidence sets that always are subsets of the central confidence sets in one parameter distributions. To translate into this problem, we consider only the conditional distribution based on $S = s$ and $\beta = \beta_{or}$. Start with $T(\mathbf{x}) = x_2$, a one-sided ordering function for the conditional problem (see Section 5). Define

$$\gamma(\mathbf{x}, \beta) = \min\left\{P_\beta[X_2 \leq x_2 | S = s], P_\beta[X_2 \geq x_2 | S = s]\right\}.$$

Let the two-sided ordering function be

$$T_{ts}(\mathbf{x}, \beta) = P_\beta\left[\gamma(\mathbf{X}, \beta) \leq \gamma(\mathbf{x}, \beta) | S = s\right].$$

Then the two-sided p-value is $p_{ts}(\mathbf{x}, \beta_0)$ from equation 7, and the associated $100(1-\alpha)\%$ confidence region is

$$C_{ts}(\mathbf{x}, 1-\alpha) = \left\{\beta : p_{ts}(\mathbf{x}, \beta) > \alpha\right\}.$$

Then Blaker (2000) showed that this gives smaller confidence sets than the central CIs. Specifically, $C_{ts}(\mathbf{x}, 1-\alpha) \subset C_c(\mathbf{x}, 1-\alpha)$, where C_c is the exact conditional central CI

using the one-sided ordering function $T(\mathbf{x}) = x_2$. Let the $100(1-\alpha)\%$ matching confidence interval to p_{ts} be the smallest interval that contains C_{ts}.

Agresti and Min (2001) showed that if one wants to create two-sided CIs with smaller length, it is generally better to invert p-values from two-sided hypothesis tests that are not central. This makes sense because centrality is a restriction, and two-sided tests without that restriction will leave room for improving CI length. For the two-sample binomial problem, basing $T_{ts}(\mathbf{x}, \beta_0)$ on score tests gives good CI length; see Chan and Zhang (1999) for β_d and Agresti and Min (2002) for β_{or}. Despite this apparent improvement, if directional inferences are needed then central confidence intervals are recommended (see Section 3.3).

9 Mid-p Methods: Improving Accuracy by Giving Up Validity

The mid-p value is a modification of a p-value for discrete data. Instead of calculating the probability of observing equal or more extreme responses, the mid-p value is 0.5 times the probability of equality plus the probability of more extreme. For example, the conditional exact test of equation 6 becomes

$$P_{\beta_0}\left[X_2 > x_2 | S\right] + \frac{1}{2} P_{\beta_0}\left[X_2 = x_2 | S\right].$$

Hwang and Yang (2001) gave some optimality criteria for the mid-p approach applied to one parameter situations, which applies to the conditional test using β_{or} since the conditional probability is completely described by only the β_{or} parameter. They show that for one-sided or two-sided hypothesis tests, the loss based on squared error between an indicator that $\beta \in \{b(\theta) : \theta \in \Theta_0\}$ and the p-value function, and shows that for all $\beta \in \{b(\theta) : \theta \in \Theta_1\}$ (and $\beta = \beta_0$) the expected loss is less than or equal to (strictly less than) the expected

loss from any randomized exact p-value function (Theorem 3.3 and 4.3 with Yang et al. (2004)). Fellows (2010) showed the minimaxity under the squared error loss and linear loss, and also showed that of all non-randomized ordered decision rules, the mid-p version is the only one that has expectation 1/2 under point null.

10 Computational Issues

Overall, the conditional p-values are much easier to calculate than the unconditional ones, since they do not require taking the supremum over the null space. The melded confidence intervals allow matching CIs to conditional tests of $\theta_1 = \theta_2$, and are very quick to calculate, since they use numeric integration. There may be some precision issues in the numeric integration for extreme data sets.

The main computational speed issues are mostly with respect to the unconditional tests, since they require estimating the supremum. Röhmel and Mansmann (1999, p. 161) showed that for ordering statistics, T, that meet the BC conditions, the supremum in the p-value calculation is on the boundary between the hypotheses. For example,

$$\sup_{\theta \in \Theta_0} \quad P_\theta \left[T(\mathbf{X}) \geq T(\mathbf{x}) \right] = \sup_{\theta : b(\theta) = \beta_0} P_\theta \left[T(\mathbf{X}) \geq T(\mathbf{x}) \right].$$

For example, the score statistic on β_d (Farrington and Manning, 1990), has been shown to follow the BC conditions for fixed β_0 (Röhmel, 2005). Further, if T meets the BC conditions and does not depend on β_0, then Theorem 3.1 of Kabaila (2005) shows that the exact unconditional one-sided p-values based on T, are either nonincreasing (for $p_U(\mathbf{x}, \beta_0)$) or nondecreasing (for $p_L(\mathbf{x}, \beta_0)$ in β_0 for fixed $\mathbf{x}$. This property means that for these p-values, the associated $100(1 - \alpha/2)$ one-sided confidence intervals can be easily calculated by finding the value β_0 where the p-value equals $\alpha/2$.

Calculation using Barnard's CSM p-value ordering can be very slow, because determining the ordering itself requires p-value calculation. Röhmel and Kieser (2013) discussed one-sided exact unconditional test using Barnard's CSM p-value ordering, except with breaking ties in a manner that does not worry about symmetry equivariance. Their additional contribution was to not worry about the exact ordering for very small p-values. This can speed up the calculations substantially.

Table 1 gives a review of the different methods, their properties of centrality and unified inferences, as well as approximate ranking of computational speed and power. The last column gives some software availability for the methods; it is not a comprehensive list, and only considers SAS 9.4, R (with packages), and StatXact 11.

11 Connection to Causal Inference

Suppose there is a population of interest with N individuals. The jth individual has two potential binary outcomes of interest, $Y_j(1)$ would be the outcome if the individual were to get treatment 1, and $Y_j(2)$ would be the outcome if the individual were to get treatment 2. Let $\mathbf{Y}_j = [Y_j(1), Y_j(2)]$. Then there are 4 types of individuals with respect to these potential outcomes, those with:

$\mathbf{Y}_j = [0, 0]$ (always fail),

$\mathbf{Y}_j = [1, 1]$ (always succeed),

$\mathbf{Y}_j = [1, 0]$ (succeed on treatment 1 only), or

$\mathbf{Y}_j = [0, 1]$ (succeed on treatment 2 only).

Let the number of individuals in each of the 4 types be respectively, N_{00}, N_{11}, N_{10}, and N_{01}. Let $\theta_1 = (N_{11} + N_{10})/N$ and $\theta_2 = (N_{11} + N_{01})/N$. Presenting the data this way implies

that the treatment one subject receives does not affect the responses of other subjects, and there is only one treatment effect for each type of treatment. (This is the stable unit treatment value assumption, see e.g., Imbens and Rubin, 2015, Section 1.6).

Consider the following type of study.

Step 1: Define the study population as a simple random sample of size $n = n_1 + n_2$ from the population of interest (of size N). Let $i_1, \ldots, i_n$ be the indices for the individuals in the study population.

Step 2: Randomly assign n_1 of the study subjects to treatment 1, and n_2 to treatment 2. Let w_{i_h} be the treatment assigned to the hth individual in the study.

Step 3: Apply assigned treatments and observe responses; for the hth individual in the study observe $Y_{i_h}(w_{i_h})$.

Let $X_a = \sum_{h=1}^{n} Y_{i_h}(a) I(w_{i_h} = a)$. If we treat N as infinity, then we can treat $X_1 \sim Binomial(n_1, \theta_1)$ and independently $X_2 \sim Binomial(n_2, \theta_2)$. Further, the parameters β_d, β_r and β_{or} have causal interpretation. For example, β_d in this situation is called the average causal difference (or average causal effect). Thus, all the previous results can be interpreted as causal inferences.

Randomized clinical trials typically use a convenience study population based on some inclusion criteria based on ethical risks to study subjects and other practical considerations, and they rarely if ever take a simple random sample from the population of interest (i.e., they rarely do Step 1). Because of this, some suggest basing causal inferences on study specific parameters that are defined only for the individuals included in the study (Robins, 1988; Rigdon and Hudgens, 2015; Li and Ding, 2016; Ding and Dasgupta, 2016). Let the individuals selected (not necessarily randomly) for inclusion into the jth study be $i_{1j}, \ldots, i_{nj}$. Let $N_{00j}, N_{11j}, N_{10j}$ and N_{01j} be the number of individuals in that study in

each of the 4 types of potential outcomes. Then the study specific parameters of interest are $\theta_{1j} = (N_{11j} + N_{10j})/n$ and $\theta_{2j} = (N_{11j} + N_{01j})/n$. The finite population average causal difference for the jth study is $\beta_{dj} = \theta_{2j} - \theta_{1j}$. If we had randomized individuals to treatment, then we can get confidence intervals for study specific parameters (such as β_{dj} and the related ones for ratios, β_{rj}, and odds ratios, β_{orj}) using only assumptions about the randomization. This is called randomization inference or Neymanian inference (Rigdon and Hudgens, 2015; Li and Ding, 2016; Ding and Dasgupta, 2016).

Scientifically, we are usually interested in two aspects of the study (see e.g., Kempthorne and Doerfler, 1969; Fay and Proschan, 2010). First, is there a treatment effect on the study population itself (internal inferences)? And second, is there a similar treatment effect on the population of interest (external inferences)? The advantage of the randomization inference is that it requires no assumptions about how the study sample was obtained in order to make valid internal inferences. The disadvantage is that those inferences are study specific inferences (e.g., inferences about β_{dj}). Alternatively, we can make the convenience assumption that the study population acts similarly to a simple random sample from the population of interest, and use our study data to make inferences about the population parameters (e.g., β_d). This has the advantage that our inferences are more generally applicable, but has the disadvantage that we have essentially assumed away the problem of generalizing the study specific inference to the external population of interest. For more discussion on this issues see Robins (1988) (for observational studies) and Imbens and Rubin (2015, Chapter 6) (for randomized experiments).

12 Power and Efficiency Comparisons

A comprehensive simulation or calculation comparing the different methods with respect to power or efficiency is beyond the score of this review. Here we review a few of the best of those types of papers and add an example and a couple of graphical calculation results to supplement the previous literature on the topic. In essence this section gives some detailed justification for the rough power/efficiency classifications listed in Table 1.

In general conditional tests (e.g., Fisher's exact tests) are less powerful than the best of the unconditional tests, because the latter tests are less discrete (Lydersen et al., 2009). Martín Andrés and Silva Mato (1994) provide a very comprehensive power comparison of several valid unconditional tests (including tests based on either an ordering function of the difference in sample proportions, or on some test-based ordering functions related to Fisher's exact p-value, the unpooled Z test, or Barndard's CSM test). They only considered ordering functions that do not depend on α or β_0 (since they only consider power to show $\theta_2 > \theta_1$ [i.e., with $\beta_0 = 0$ for the difference or $\beta_0 = 1$ for the ratio or odds ratio] the ordering functions automatically do not depend on β_0). Martín Andrés and Silva Mato (1994) based power comparisons on expected power assuming bivariate uniformly distributed (θ_1, θ_2). They found that Barnard's CSM test was the most powerful on average, and that ordering by either the unpooled Z statistics for the difference in means or the Fisher's exact p-values (i.e., a Boschloo-type test) gave the next best power. Martín Andrés and Silva Mato (1994) did not include a pooled Z test, but Mehrotra et al. (2003) did, and they showed that the pooled Z test can have much better power with unequal sample sizes. So in general we can recommend ordering by the pooled Z instead of the unpooled Z. Since Barndard's CSM test is difficult to calculate, Martın Andrés et al. (2002) compared many approximations to that value. They concluded that the mid-p Fisher's p-value was the best approximation to

the CSM test, although it could be conservative for very small samples. Hirji et al. (1991) did extensive calculations finding the type I error rate for the exact conditional mid-p one-sided and two-sided (Fisher-Irwin-type) tests. They found that out of 3125 sample size and parameter situations (all with $\theta_1 = \theta_2$), typically 90-95% of both types of the mid-p p-value when used to test at a 5% significance level, had type I error rates less than or equal to 5%. Further, Lydersen et al. (2009) stated that the mid-p version of the Fisher-Irwin test approximates the Fisher-Boschloo test well, and the latter test (or the exact unconditional test on Pearson's chi-squared test) was their recommendation.

For confidence intervals, we focus on two papers. Chan and Zhang (1999) compared unconditional confidence intervals based on estimates of or tests on the difference: the difference in proportions, the unpooled Z statistic, the score statistic (which they called the δ-Projected Z statistic), and the likelihood ratio statistic. They tried all with and without the Berger and Boos (1994) adjustment. They showed the score statistic with no adjustment generally gave smaller average confidence interval length. Santner et al. (2007) did a very comprehensive set of calculations for β_d confidence intervals, calculating the expected coverage and confidence interval length for a 100×100 grid of values of (θ_1, θ_2). They compared three valid methods and two approximate methods, including the unconditional method based on a two-sided score test, the unconditional method based on two one-sided score tests, and an approximate method of Coe and Tamhane (1993). The results show that of the valid methods, the unconditional method based on the two-sided score test statistic had the lowest expected length, while the central unconditional method based on two one-sided score tests had larger expected length. However, if directional inferences are important, then the proper comparison should be the former method using $100(1 - 2\alpha)\%$ intervals compared to the latter method using $100(1 - \alpha)\%$ intervals (see Section 3.3). Further, the score tests may lack coherence (see Figure 3). Santner et al.

(2007) ended up recommending the approximate method of Coe and Tamhane (1993), which had smaller length confidence intervals and gave coverage above the nominal except in less than 0.6% of the cases. Fagerland et al. (2015), also recommends for small samples the exact unconditional confidence intervals with the ordering function the two-sided score test statistic. Fagerland et al. (2015) mentions using one-sided tests if direction is important.

We now compare the score tests to some of the tests introduced in this paper. Between the unconditional tests applied to β_r and β_{or}, the ordering based on score tests or the ordering based on one-sided mid-p Fisher's exact p-values perform much better than ordering by the estimates with tie breaks as in Section 4.3. For example, with $n_1 = n_2 = 20$, $\theta_1 = 0.4$, $\theta_2 = 0.8$, and a one-sided 0.025 significance level, the power is 73% for score-based or mid-p Fisher-based tests of both β_r and β_{or}, but it is very small for the test that orders by estimates with tie breaks (power ≈ 0 for β_r and power $\approx 1\%$ for β_{or}). We get a slight increase in power for the latter tests when we use Berger and Boos adjustment with $\gamma = 10^{-6}$ (power is 11% for β_r and 16% for β_{or}). In contrast, for β_d in that example all three methods of ordering with or without the Berger-Boos adjustment give 73% power.

In Figure 7 we compare powers on the two-sided 0.05 level central tests that $\beta_d = 0$. Powers are calculated on a grid 99×99 grid of values of (θ_1, θ_2). We plot the difference in powers between all pairs of three tests: two unconditional exact tests (one based on the score test for the difference in proportions, and one based on the difference in proportions with a tie break) and the conditional test (the central Fisher's exact test). We find, as expected, that the unconditional tests do better, and that the simple method with a tie break does well when the sample sizes are not equal (see e.g., Mehrotra et al., 2003, for a different set of simulations showing a similar result for the two-sided test).

In Figure 8 we compare the unconditional exact tests ordered by score statistics (on either $\beta_d = 0, \beta_{or} = 1$, or $\beta_r = 1$) compared to the unconditional exact tests based on the

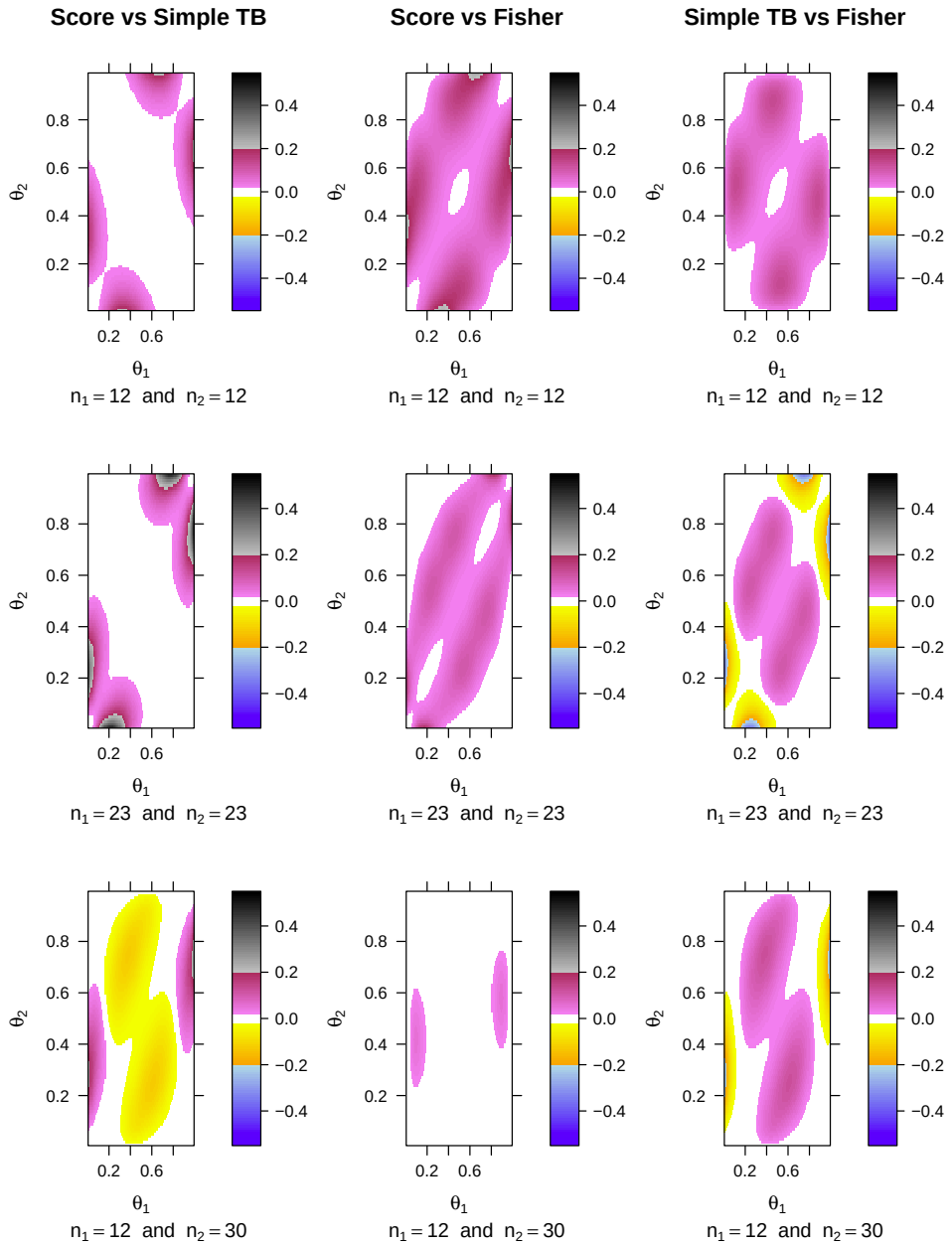

Figure 7: Comparison of powers for testing $\theta_1 = \theta_2$ using central tests at the two-sided 0.05 level. The three tests compared are "score"= unconditional exact test based on the score test of the difference in proportions, "simple TB"= unconditional exact test based on the difference in proportions using a simple tie-break (see Section 4.2), and "Fisher"= tests based on central Fisher's exact test. For columns labeled Test 1 vs Test 2, the result is power of Test 1 minus Power of Test 2, so that positive values (pink and gray) indicate that Test 1 is more powerful. White indicates that powers are within 0.025 of each other.

mid p-values from the one-sided Fisher's exact test. We find that the latter test is generally more powerful.

13 Recommendations

Some recommendations:

1. We should almost always use central confidence intervals with either a central p-value, or the minimum of the one-sided p-values. Although using non-central two-sided CIs can slightly decrease CI length, that advantage comes at a cost in terms of allowable one-sided inferences. Since after rejecting a two-sided test we usually care about the direction of effect, non-central CIs are not routinely recommended.

2. It is usually not useful to maximize the power or minimize the confidence interval. It comes at the price of increased computational burden and will lead to incoherent p-values and non-nested CIs.

3. For fast calculations use the one-sided conditional exact tests and the melded confidence intervals.

4. For more power use the unconditional one-sided valid p-values and associated central CIs. For inferences on β_d we can order based on the difference in sample proportions, except break ties while maintaining the BC conditions, and do not let the ordering function depend on β_0 or α. This will ensure monotonicity of the p-values as a function of β_0, allowing for relatively fast calculations, and avoiding incoherence and non-nestedness. For inferences on β_r and β_{or}, using the simple function with a tie breaking ordering will have a much smaller power than the score method or ordering based

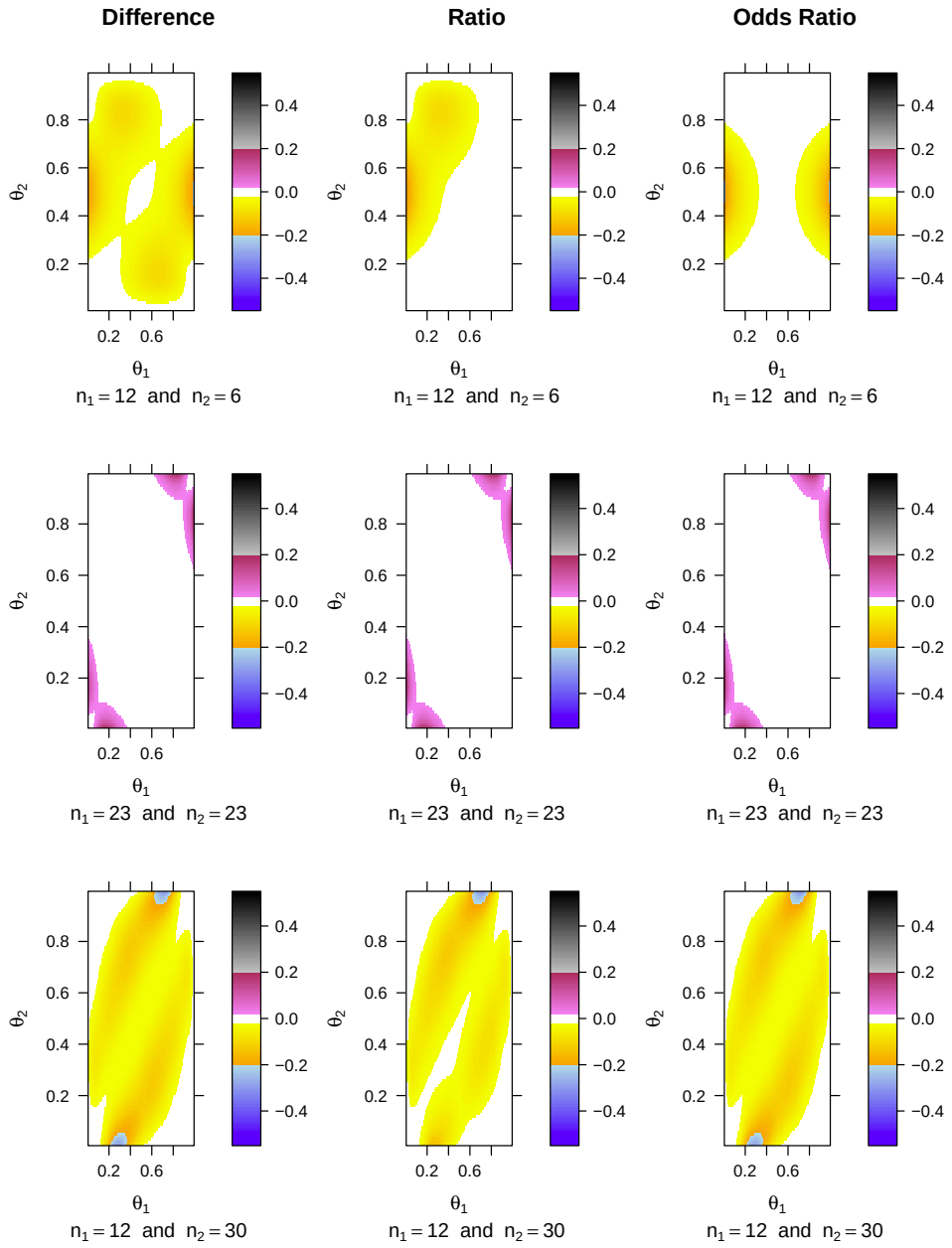

Figure 8: Power of unconditional exact score test minus power of unconditional exact test based on ordering by Fisher's exact test one-sided mid p-value. Negative values (yellow and blue) denote parameter values in which the latter test is more powerful. The unconditional exact score tests are defined based on testing either $H_0 : \beta_d = 0$ (first column), $H_0 : \beta_{or} = 1$ (second column), or $H_0 : \beta_r = 1$ (third column). White indicates that the two powers are within 0.025 of each other. Additional calculations with $n_1 = 12, n_2 = 12$ showed nearly equal powers (all white) for all three columns and are not plotted.

on one-sided mid-p Fisher's exact p-values. The score method introduces problems with incoherence or non-nestedness, while the mid-p Fisher p-value ordering does not. Because the latter method only uses the mid p-values for ordering within the exact unconditional test framework, the resulting p-values are valid. Further, for inferences on β_d, the mid-p ordering meets the BC conditions and is relatively fast to calculate.

5. If validity is not vital, then the mid-p conditional tests are a good approximation to the more powerful of the unconditional exact ones. Additionally, with a large proportion of situations with $\theta_1 = \theta_2$, the mid-p conditional tests still have type I error rates less than the nominal value.

Table 1: Valid (and Mid-p adjusted) Methods for Two-Sample Binomial Problem, and some Properties, References, and Software

Method	Central	Unified Infer.	Comput. Speed*	Power/ Efficiency**	References	Sect.	Software***
Smallest CI	yes	no	3	1	Wang (2010) (for β_d) Wang and Shan (2015) (for β_r, β_{or})	4.5	Rpkg:ExactCIdiff (for β_d CI only)
Barnard's CSM	both	?	3	1	Barnard (1947)	4.1	Rpkg: Exact(p-value only)
Boschloo Test	both	both	2	2	Boschloo (1970)	8	Rpkg: Exact (p-value only), exact2x2
Uncond Exact Score Stat (square T)	no	no	2	2	Chan and Zhang (1999) (for β_d) Agresti and Min (2001) (for β_d, β_r) Agresti and Min (2002) (for β_{or})	8	StatXact-11 (only β_d, β_r), SAS 9.4 (only β_d, β_r), Rpkg: exact2x2 (tsmethod="square")
Uncond Exact β Estimates with tie break	yes	yes	2	2(β_d) 5(β_r) 5(β_{or})		4.2 4.3	Rpkg: exact2x2 (tsmethod="central")
Uncond Exact Wald Stat (T^2)	no	no	2	2	Mehrotra et al. (2003)	8	StatXact-11 (only β_d) Rpkg: exact2x2 (tsmethod="square")
Uncond Exact β Estimates	yes	yes	2	3(β_d) 5(β_r, β_{or})	Barnard (1945) Mehrotra et al. (2003)	4.1 4.3	Rpkg: exact2x2 (tsmethod="central")
Cond Exact with Fisher-Irwin Exact Test	no	no	1	3	Fisher (1934) (for p-value) Fay (2010a) (for CI)	5	Rpkg:exact2x2
Cond Exact with Blaker Method	no	no	1	3	Blaker (2000) Fay (2010a)	8	Rpkg: exact2x2
Cond Exact with Melded CIs	yes	yes	1	4	Fisher (1934) (for p-value) Fay et al. (2015) (for CI)	6	Rpkg: exact2x2
Cond exact with tail approach CI (only for β_{or})	yes	yes	1	4	Agresti and Min (2001) Fay (2010a)	5	SAS 9.4 (use double one-sided Fisher's exact p-values) StatXact-11, Rpkg: exact2x2

Adjustment	Notes	Sect.	Software
Berger-Boos	Adjustment by Berger and Boos (1994) applies to unconditional exact tests and generally increases power	4.4	StatXact-11, Rpkg: exact2x2 Rpkg: Exact(p-values only)
E+M	Adjustment by Lloyd (2008) applies to unconditional exact tests and generally increases power	4.4	Rpkg: exact2x2
Mid-p	Applies to any method, increases power at the cost of validity	9	Rpkg: exact2x2 SAS 9.4 (not all tests)

* Approximate computation speed: 1=fast, 2=moderate, 3=slow. ** Approximate power/efficiency: 1=higher power/smaller CI, . . ., 5=lower power/larger CI.
*** Software (not comprehensive, only considered R, SAS and StatXact): R packages available at `https://cran.r-project.org/`.
For SAS the methods are available in PROC FREQ using exact option. The value "both" denotes there could be versions with and without the property, and "?" denotes that it is not clear if the matching confidence intervals are unified because confidence intervals have not been studied with that test (although it is likely the method will not be unified because it is similar to the smallest CI method).

Appendix: Proof of Theorem 3.1

Proof of statement 1 :

(Unified Inferences) $\Rightarrow$ ($C_I = C$): If the confidence region associated with a p-value is not an interval, then there must be an α and β_0 such that $p(\mathbf{x}, \beta_0) \leq \alpha$ and $\beta_0 \in C_I(\mathbf{x}, 1-\alpha)$, which contradicts the unified inferences, therefore $C_I(\mathbf{x}, 1-\alpha) = C(\mathbf{x}, 1-\alpha)$.

($C_I = C$) $\Rightarrow$ (Unified Inferences): If the confidence region associated with the p-value is the matching confidence interval, then the inferences are unified by definition (equation 1).

Proof of statement 2 , (unified inferences) $\Rightarrow$ (nested CI): We show the contrapostive. If a method has non-nested CIs, then there exists some $\alpha_1 < \alpha_2$ and some β_0 such that $\beta_0 \notin C_I(\mathbf{x}, 1-\alpha_1)$ and $\beta_0 \in C_I(\mathbf{x}, 1-\alpha_2)$. If the method had unified inferences, then $p(\mathbf{x}, \beta_0) \equiv p \leq \alpha_1$ and $p > \alpha_2$. This leads to the contradiction, $p \leq \alpha_1 < \alpha_2 < p$, so the method must not have unified inferences, and we have proven the result.

Proof of statement 3 , (Unified Inferences) $\Rightarrow$ (Coherence): From statement 2, the unified inferences imply nested CIs. For one-sided p-values, the unified inferences with the nested CIs imply that the p-values are non-decreasing as the null space expands (e.g., β_0 gets larger when $H_0 : \beta \leq \beta_0$), and hence are coherent by definition. For two-sided p-values, because of unified inferences and nested CIs, the p-values are increasing (i.e., non-decreasing) as $1-\alpha$ decreases. This is directional coherence by definition.

References

Agresti, A. and Y. Min (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics 57*(3), 963–971.

Agresti, A. and Y. Min (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics 3*(3), 379–386.

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

Barnard, G. (1947). Significance tests for 2× 2 tables. *Biometrika 34*(1/2), 123–138.

Berger, R. L. and D. D. Boos (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association 89*(427), 1012–1016.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics 28*(4), 783–798.

Boschloo, R. (1970). Raised conditional level of significance for the 2× 2-table when testing the equality of two probabilities. *Statistica Neerlandica 24*(1), 1–9.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association 91*(433), 14–28.

Chan, I. S. (2003). Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Statistical methods in medical research 12*(1), 37–58.

Chan, I. S. and Z. Zhang (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics 55*(4), 1202–1209.

Coe, P. R. and A. C. Tamhane (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics-Simulation and Computation 22*(4), 925–938.

Coulibaly, Y. I., B. Dembele, A. A. Diallo, E. M. Lipner, S. S. Doumbia, S. Y. Coulibaly, S. Konate, D. A. Diallo, D. Yalcouye, J. Kubofcik, et al. (2009). A randomized trial of doxycycline for mansonella perstans infection. *New England Journal of Medicine 361*(15), 1448–1458.

Cytel (2010). *StatXact 9*.

Ding, P. and T. Dasgupta (2016). A potential tale of two-by-two tables from completely randomized experiments. *Journal of the American Statistical Association 111*(513), 157–168.

Fagerland, M. W., S. Lydersen, and P. Laake (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical methods in medical research 24*(2), 224–254.

Farrington, C. P. and G. Manning (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in medicine 9*(12), 1447–1454.

Fay, M. (2010a). Confidence intervals that match fisher's exact or blaker's exact tests. *Biostatistics 11*(2), 373–374.

Fay, M. P. (2010b). Two-sided exact tests and matching confidence intervals for discrete data. *R journal 2*(1), 53–58.

Fay, M. P. and M. A. Proschan (2010). Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys 4*, 1–39.

Fay, M. P., M. A. Proschan, and E. Brittain (2015). Combining one-sample confidence procedures for inference in the two-sample case. *Biometrics 71*(1), 146–156.

Fellows, I. (2010). The minimaxity of the mid p-value under linear and squared loss functions. *Communications in Statistics-Theory and Methods 40*(2), 244–254.

Finner, H. and K. Strassburger (2001). Ump (u)-tests for a binomial parameter: A paradox. *Biometrical journal 43*(6), 667–675.

Fisher, R. A. (1934). *Statistical Methods for Research Workers, 5th edition.* Edinburgh: Oliver and Boyd.

Fog, A. (2008). Sampling methods for wallenius' and fisher's noncentral hypergeometric distributions. *Communications in StatisticsSimulation and Computation® 37*(2), 241–257.

Freedman, L. S. (2008). An analysis of the controversy over classical one-sided tests. *Clinical Trials 5*(6), 635–640.

Goeman, J. J., A. Solari, and T. Stijnen (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in medicine 29*(20), 2117–2125.

Hirji, K. (2006). *Exact Analysis of Discrete Data.* New York: Chapman and Hall/CRC.

Hirji, K. F., S.-J. Tan, and R. M. Elashoff (1991). A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine 10*(7), 1137–1153.

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

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Irwin, J. (1935). Tests of significance for differences between percentages based on small numbers. *Metron 12*(2), 84–94.

Johnson, N. L., A. W. Kemp, and S. Kotz (2005). *Univariate Discrete Distributions, 3rd edition*. New York: John Wiley & Sons.

Kabaila, P. (2005). Computation of exact confidence limits from discrete data. *Computational Statistics 20*(3), 401–414.

Kabaila, P. and C. J. Lloyd (2003). The efficiency of buehler confidence limits. *Statistics & probability letters 65*(1), 21–28.

Kabaila, P. and C. J. Lloyd (2006). Improved buehler limits based on refined designated statistics. *Journal of statistical planning and inference 136*(9), 3145–3155.

Kempthorne, O. and T. Doerfler (1969). The behaviour of some significance tests under experimental randomization. *Biometrika 56*(2), 231–248.

Lehmann, E. and J. Romano (2005). *Testing Statistical Hypotheses, third edition*. Springer.

Li, X. and P. Ding (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in medicine 35*(6), 957–960.

Lloyd, C. J. (2008). Exact p-values for discrete models obtained by estimation and maximization. *Australian & New Zealand Journal of Statistics 50*(4), 329–345.

Lloyd, C. J. and P. Kabaila (2003). On the optimality and limitations of buehler bounds. *Australian & New Zealand Journal of Statistics 45*(2), 167–174.

Lydersen, S., M. W. Fagerland, and P. Laake (2009). Recommended tests for association in 2× 2 tables. *Statistics in medicine 28*(7), 1159–1175.

Martın Andrés, A., M. Sánchez Quevedo, and A. Silva Mato (2002). Asymptotical tests in 2× 2 comparative trials:(unconditional approach). *Computational statistics & data analysis 40*(2), 339–354.

Martín Andrés, A. and A. Silva Mato (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational statistics & data analysis 17*(5), 555–574.

Mehrotra, D. V., I. S. Chan, and R. L. Berger (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics 59*(2), 441–450.

Mehta, C. R., N. R. Patel, and R. Gray (1985). Computing an exact confidence interval for the common odds ratio in several 2× 2 contingency tables. *Journal of the American Statistical Association 80*(392), 969–973.

Rigdon, J. and M. G. Hudgens (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in medicine 34*(6), 924–935.

Robins, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine 7*(7), 773–785.

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. *Biometrical Journal 47*(1), 37–47.

Röhmel, J. and M. Kieser (2013). Investigations on non-inferioritythe food and drug administration draft guidance on treatments for nosocomial pneumonia as a case for exact tests for binomial proportions. *Statistics in medicine 32*(14), 2335–2348.

Röhmel, J. and U. Mansmann (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal 41*(2), 149–170.

Santner, T. J., V. Pradhan, P. Senchaudhuri, C. R. Mehta, and A. Tamhane (2007). Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions. *Computational Statistics & Data Analysis 51*(12), 5791–5799.

Santner, T. J. and M. K. Snell (1980). Small-sample confidence intervals for p 1–p 2 and p 1/p 2 in 2× 2 contingency tables. *Journal of the American Statistical Association 75*(370), 386–394.

Sas (2012). *SAS/STAT 12.1 User's Guide, Proc Freq.*

Vos, P. W. and S. Hudson (2008). Problems with binomial two-sided tests and the associated confidence intervals. *Australian & New Zealand Journal of Statistics 50*(1), 81–89.

Wang, W. (2010). On construction of the smallest one-sided confidence interval for the difference of two proportions. *The Annals of Statistics 38*(2), 1227–1243.

Wang, W. and G. Shan (2015). Exact confidence intervals for the relative risk and the odds ratio. *Biometrics 71*(4), 985–995.

Xie, M.-g. and K. Singh (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review 81*, 3–77.

Yang, M.-C., D.-W. Lee, and J. G. Hwang (2004). The equivalence of the mid p-value and the expected p-value for testing equality of two balanced binomial proportions. *Journal of statistical planning and inference 126*(1), 273–280.

Yates, F. (1984). Tests of significance for 2× 2 contingency tables. *Journal of the Royal Statistical Society. Series A (General) 147*, 426–463.