Methods for Coherent and Exact Inference

Spencer Hansen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Kenneth M. Rice, Chair

Robyn L. McClelland

Eardi Lila

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Methods for Coherent and Exact Inference

Spencer Hansen

Chair of the Supervisory Committee:
Professor Kenneth M. Rice
University of Washington, Department of Biostatistics

This dissertation provides methods for coherent and exact inference for two types of problems in statistics. The first problem provides coherent criteria for the testing of nested interval null hypothesis. The second and third problems examine exact inference for meta-analysis of proportions and 2×2 tables, respectively.

Criticism of using p-values as measures of support is well-documented in the literature. In particular, one setting where uniformly most powerful unbiased (UMPU) test p-values for null hypotheses that are nested intervals can be incoherent; we may accept a smaller null but reject a larger one in which it is nested (see Schervish, P values: what they are and what they are not, 1996). In order to avoid this incoherence, the Bayesian paradigm offers guarantees of certain forms of coherence. Using Bayesian decision theory, we establish straightforward conditions that ensure coherence. From these, we establish novel frequentist criteria - different to Type I error rate, that give tests that are coherent.

Meta-analysis is a practice that utilizes multiple studies and seeks inference on some overall effect. Meta-analysis of proportions, for example, seeks inference on some overall proportion of successes-failures, where the multiple studies estimate a binary outcome. Under common effect models, which assume each study is estimating the same underlying truth, exact inference has long been available. However, under a more reasonable fixed-effects models, exact inference is not readily available. Instead, non-exact methods are used which can be challenging to interpret. We present methods for exact tests and confidence

intervals for fixed-effects meta-analysis of proportions that maintain interpretability of the parameter of interest and are easily implemented.

Another area of meta-analysis examines 2x2 tables. These are common when synthesizing the results of multiple placebo-controlled trials for a binary outcome. In these analyses, we seek inference on some overall comparison of the outcome between two groups, such as the odds ratio. We provide an approach that is exact by extending the method we used in meta-analysis of proportions that provides inference on an overall odds ratio.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I want to express my gratitude to my advisor, Doctor Kenneth Rice, for all the patience and guidance he provided me. I would also like to thank the other members of my committee, Doctor Robyn McClelland, Doctor Eardi Lila, and Doctor Brandon Guthrie, for all of their comments and suggestions. I would also like thank Doctor Robyn McClelland for advising during my time as a research assistant. To the Department of Biostatistics, thank you for the opportunity when they admitted me to this program. Lastly, I would like to thank my fellow students for reassuring me that I was not the only one struggling with the coursework.

DEDICATION

to my dear wife, Ye Tian

Chapter 1

INTRODUCTION

Statistical inference - using data to infer something about the true, underlying probability distribution[99] - is a cornerstone of statistics. Indeed, while data alone is helpful in describing the sample, inference allows us to draw conclusion from that data about the population. These conclusions have sizeable impacts on decision making in many different arenas - e.g. public policy, economics, and health care. Given the sensitive nature of some of these areas, health care especially, it is important that inference is done with caution. In certain instances, applying a degree of conservatism to the inference is warranted to ensure safety to the public. Furthermore, we do not want inference to provide contradictory results.

In chapter 2, we will describe coherence in testing interval null hypotheses. In a celebrated paper, Mark J Schervish showed that, for testing interval null hypotheses, tests typically viewed as optimal can be logically incoherent [88]. Specifically, one may fail to reject a specific interval null, but nevertheless – testing at the same level with the same data – reject a larger null, in which the original one is nested. Schervish's result is a strong argument against the widespread practice of viewing p-values as measures of evidence. In the current work we approach tests of interval nulls using simple Bayesian decision theory, and establish straightforward conditions that ensure coherence in Schervish's sense. From these, we go on to establish novel frequentist criteria – different to Type I error rate – that, when controlled at fixed levels, give tests that are coherent in Schervish's sense. The results suggest that exploring frequentist properties beyond the familiar Neyman-Pearson framework may ameliorate some of statistical testing's well-known problems.

In chapter 3, we turn our attention to meta-analysis and exact inference, meaning we always have at least the nominal coverage (i.e. 95%). We begin with meta-analysis of proportions which is conceptually simple. Faced with a binary outcome in multiple studies, we seek inference on some overall proportion of successes/failures. Under common effect

models exact inference has long been available, but is not when we more realistically allow for heterogeneity of the proportions. Instead a wide range of non-exact fixed-effects methods are used, the interpretation of some of which is challenging. In this paper we present methods for exact statistical tests and confidence intervals for fixed-effects meta-analysis of proportions. These methods retain the interpretability of the underlying parameter of interest, and can be implemented in straightforward software. We also show how our inference on the overall proportion is compatible with exact inference on heterogeneity of proportions. An illustrative example from a recent kidney disease study shows how the method's performance can be assessed in practice.

In chapter 4, we move to working with 2×2 tables. Meta-analyses of 2×2 contingency tables are common in practice, for example when synthesizing the results of multiple placebo-controlled trials for a binary outcome. In such analyses, we seek inference on some overall comparison of the outcome between two groups, where the comparison summarizes multiple strata. A typical overall comparison is the overall odds ratio, which describes the association between the row and column variables. Under homogeneity, exact inference is straightforward. However, under heterogeneity, exact inference for testing the weak null is not currently available. We propose a method for meta-analysis of 2×2 tables that provides exact (albeit conservative) inference on the conditional maximum likelihood estimate of the odds ratio. We also provide a measure of the degree of conservatism to provide an inclination of the extent of the excess coverage under heterogeneity compared to a homogenous case. We apply our approach to the controversial Avandia meta-analysis, of the effect of rosiglitazone treatment on myocardial infarction and death.

Chapter 5 will add a brief conclusion to the dissertation.

Chapter 2

COHERENT TESTS FOR INTERVAL NULL HYPOTHESES

2.1 Introduction

Testing and the use of p-values remains a controversial area [10]. Even putting aside incorrect interpretations [104], there remain specific problems. Arguably p-values and tests do not provide a single coherent theory of statistical inference [42] and standard tests can be formally incoherent [28]. A particularly striking form of incoherence was noted by Mark J Schervish [88], when testing interval null hypotheses. Specifically, under a straightforward model and using widely-preferred uniformly most powerful unbiased (UMPU) tests, one may fail to reject one null hypothesis, but nevertheless reject a null that encompasses it – despite testing at the same level with the same data. As well as providing practical difficulties for applications of these tests (also known as 'minimum-effect tests [67, 105, 47]), this result calls into question the widespread use of p-values as measures of evidence: if there is enough evidence to reject the larger set of values then logically there has to be enough to reject the smaller set nested inside.

Coherence – meaning the ability to make multiple non-contradictory statements about a parameter – has long been recognized as a natural consequence of adopting Bayesian inference ([56]) which describes uncertainty via the mathematical 'language' of probability [91]. Complete class theorems [103] go further, showing that if criteria for answering a question about a parameter can be expressed in functional form – i.e. as a loss function – then any admissible rule is Bayes for some prior, or the limiting case of a prior. Both properties favor use of Bayesian decision theory, yet do not guarantee coherence in the sense explored by Schervish, in which optimal-but-contradictory *decisions* may be returned despite coherence of the posterior itself as a description of parameter uncertainty.

This chapter explores how Bayesian decision-theoretic tests can be constructed so as to guarantee what we call *Schervish coherence*. After reviewing the literature in Section 2.2,

Section 2.3 extends the loss functions explored by Kenneth Rice [81] to interval-valued nulls. We show how Schervish coherence corresponds to a simple constraint on those loss functions – which can easily be met, although not by default approaches that closely resemble use of UMPU tests. More generally, we show that the criterion of controlling Type I error rate is not compatible with obtaining Schervish coherence. This motivates Section 2.4, in which we consider alternative frequentist criteria that could be controlled, and show how calibrating tests in those terms allows us to recover Schervish coherence with non-Bayesian methods. We conclude with a short discussion.

2.2 Review

2.2.1 Incoherence in Interval Tests

When testing an interval null hypothesis, we assess whether real-valued parameter θ lies between endpoints θ_B and θ_A, where $\theta_B < \theta_A$. Formally the null hypothesis is denoted

$$H_0 : \theta \in (\theta_B, \theta_A).$$

In the simple setting where we have a single observation Y from $N(\theta, 1)$, the Uniformly Most Powerful Unbiased (UMPU) test of level α for H_0 [21, §8.3] can be performed by computing the p-value defined as

$$p_{\theta_B,\theta_A} = \min\left(\Phi(Y-\theta_B) + \Phi(Y-\theta_A), 1 - \Phi(Y-\theta_B) + 1 - \Phi(Y-\theta_A)\right), \tag{2.1}$$

where Φ denotes the cumulative distribution function of a standard Normal random variable, and rejecting H_0 only when $p_{\theta_B,\theta_A} < \alpha$.

Schervish [88] illustrates how testing in this way can be incoherent. Specifically, with $Y = 2.18$ and $\alpha = 0.05$, then testing

$$H_{01} : \theta \in (-.5, .5),$$

we do not reject ($p = 0.0502$), but this null is nested inside that of

$$H_{02} : \theta \in (-.82, .52),$$

which is rejected ($p = 0.0498$).

The incoherence is clear: if we believe that θ lies outside $(-.82, .52)$ then it must also lie outside $(-.5, .5)$, yet this is not the first test's result. Incoherence can hold quite generally, well beyond Schervish's deliberately simple example. In particular it is not due to having small samples, nor testing at level $\alpha = 0.05$, nor any special properties unique to Normal data [28].

Incoherence for general multiple-testing decision problems is considered by Gabriel [30], who defines *coherence* as agreement between rejecting a hypothesis and all hypotheses that would imply it, and *consonance* as agreement between non-rejection of a hypothesis and other hypotheses that it implies, before giving monotonicity conditions for them to hold. Our term 'Schervish coherence' reflects our focus on coherence of decisions for interval nulls, and distinguishes it from the distinct notion of Bayesian coherence [57, §9].

2.2.2 Bayesian Decision Theory and Coherence

Given the automatic coherence of Bayesian descriptions of uncertainty, the Bayesian paradigm is a natural choice when trying to ensure Schervish coherence. Several authors have taken such an approach [76, 93, 45, 25, 26, 28] but almost all the literature addresses problems in some way different from Schervish's assessments of coherence for standard tests of interval nulls.

Several of these authors [25, 45, 26] address other forms of logical consistency in addition to Schervish coherence. Izbicki and Esteves [45] seek methods that also provide *invertibility* (in which the labels of 'null' and 'alternative' hypotheses can be switched without affecting inference), and *consonance*, in which rejecting/not rejecting both H_{0A} and H_{0B} leads respectively to rejecting/not rejecting their union. While da Silva *et al* [25] shows that it is impossible to achieve coherence, invertibility and consonance in a single Bayesian hypothesis testing method. However, invertibility is arguably not essential: formal significance tests (in which a null is either rejected or no conclusions at all are drawn [6, §5.2]) do not treat the null and alternative equally, yet are hardly without logical foundation.

Decision theory has long been suggested as a motivation for testing: Neyman and Pearson's accept/reject framework was developed with behavioral decisions in mind [69, pg 258].

Developing tests using Bayesian decision theory, losses that feature only the truth, or otherwise, of a binary statement about parameters are simple to construct [72, §7.5]. (For our setting, the relevant binary statement would be whether $\theta \in (\theta_B, \theta_A)$, or not.) A full exploration of their properties is beyond the scope of this paper, but versions of them have been assessed with regard to Schervish coherence. Specifically, both da Silva *et al* [25] and Izbicki and Esteves [45] assess coherence (and invertibility and consonance) of accept/reject losses, in single and multiple testing problems. The former shows that coherent Bayesian tests of this sort can be constructed, while the latter prove that they can be achieved in conjunction with at most one of invertibility and consonance. Esteves *et al* [26] expands this framework to also include an 'agnostic' decision, meaning no firm conclusions are made. With this comes the ability (using region based estimators) to satisfy coherence, invertibility, and consonance together. However, this is done only considering the truth as binary, with no account permitted of whether the truth lies above or below the null interval. Lastly, Fossaluza *et al* [28] use posterior measures of support for coherence. They discuss how a Bayesian decision framework can lead to coherence in Schervish's sense, but also show that not every use of posterior support provides it.

Recent explorations of decision-theoretic testing suggest that slightly more detailed loss functions may be useful for connecting Bayesian methods to long-studied frequentist testing properties. Specifically, Rice *et al* [81] shows that considering the sign of a parameter, either side of a point null, under mild conditions must lead to Bayesian analogs of two-sided tests. (Sign errors have been considered extensively in the literature [92, 19, 32], but taking a decision-theoretic approach, and in particular considering the space of all *possible* loss functions is novel.) Extensions to multiple testing, credible intervals, and *post hoc* test assessment all follow [51]. However, assessment of Bayesian tests in terms of simple sign decisions relative to an interval null has not been studied.

More complex Bayesian approaches have, however, been considered, with application to testing interval nulls. Specifically, Stern and Pereira [93] compute the e-value (epistemic value) of a model's *truth function*, the integral (up to a certain cut-point) of the ratio of the posterior relative to a reference density, that can be an uninformative prior. Another approach by the same authors [76] provide coherent Bayesian measures of support for hy-

potheses, but do so by—unconventionally—maximizing the posterior density under the null. In related non-Bayesian work, Zhang and Zhang [109] use axioms forcing coherence to motivate a generalized likelihood ratio, dividing the supremum likelihood over parameters in an 'alternative' subspace by the supremum likelihood over parameters in a 'null' subspace. Further non-Bayesian approaches to coherent tests are given by Bickel and Patriota [75, 14].

2.3 Main Results

2.3.1 Bayesian Interval Testing

We now extend Rice *et al*'s development (of tests as decisions on the sign of θ) to interval nulls. Specifically, we consider losses that only depend on the univariate parameter θ through indicators that $\theta > \theta_A, \theta \in (\theta_B, \theta_A)$ and $\theta < \theta_B$ for some specified values $\theta_A > \theta_B$. (For simplicity, we assume the posterior does not contain point masses at $\theta = \theta_A$ or θ_B and so it suffices to only consider strict inequalities.)

We consider three corresponding mutually-exclusive decisions, to report $(d = A)$ that θ is above θ_A, or $(d = B)$ that it is below θ_B, or $(d = C)$ that it lies in the central null interval (θ_A, θ_B). With three parameter indicators and three decisions, we need consider at most nine truth-decision combinations to fully specify a loss function. For simplicity, however, we will assume symmetry with regard to A and B, with equal loss (denoted l_1) for incorrectly stating that $d = A$ or B when θ is below θ_B or above θ_A respectively, equal loss (l_2) for stating $d = A$ or B when $\theta \in (\theta_B, \theta_A)$, and equal loss ($l_3$) for incorrectly stating $d = C$ when θ is below θ_B or above θ_A. We further assume that all incorrect decisions incur greater loss than correct decisions (making the loss *proper*, as discussed in Hwang *et al* [44]) and without loss of generality assume that correct decisions incur zero loss. In tabular form, the resulting loss function can be written as

		Decision		
		A	C	B
Truth	$\theta < \theta_B$	l_1	l_2	0
	$\theta \in (\theta_B, \theta_A)$	l_3	0	l_3
	$\theta > \theta_A$	0	l_2	l_1

. (2.2)

The expected loss under each decision is therefore

		Expected Loss
Decision	A	$(l_1 - l_3)\mathbb{P}[\,\theta < \theta_B \vert Y\,] + l_3(1 - \mathbb{P}[\,\theta > \theta_A \vert Y\,])$
	C	$l_2(\mathbb{P}[\,\theta < \theta_B \vert Y\,] + \mathbb{P}[\,\theta > \theta_A \vert Y\,])$
	B	$(l_1 - l_3)\mathbb{P}[\,\theta > \theta_A \vert Y\,] - l_3(1 - \mathbb{P}[\,\theta < \theta_B \vert Y\,])$

We see directly that whether the Bayes rule sets $d = A, C$ or B depends on $\mathbb{P}[\,\theta > \theta_A|Y\,]$ and $\mathbb{P}[\,\theta < \theta_B|Y\,]$ alone. Moreover, the expected losses are linear combinations of these quantities, for any pair of decisions (e.g. A versus B) the regions of the $(\mathbb{P}[\,\theta > \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,])$ plane where one decision is preferred (e.g. A has less expected loss than B) must be bounded by straight lines. Combining preferences for all three possible decisions (e.g. preferring A to C and A to B) we find that regions of the $(\mathbb{P}[\,\theta > \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,])$ plane for which the the Bayes rule sets $d = A$, B or C must also be bounded by straight lines.

An example of these lines and regions is given in Figure 2.1, with $(l_1, l_2, l_3) \propto (2, 0.4, 0.6)$. The rule may seem broadly intuitive; each hypothesis is accepted only when posterior support for it is large relative to the alternatives. However, this rule is subject to Schervish incoherence. For a $N(-0.83, 1)$ posterior and where we compare tests T_1 and T_2 with the nulls $(\theta_B, \theta_A) = (-0.25, 0.25)$ and $(-0.4, 1.1)$ respectively. Test T_1 has Bayes rule $d = C$ but T_2 – with an encompassing null – returns $d = B$.

Directly looking at tail areas, as in Figure 2.1a, does not seem to make this intuitive, but we can instead plot the posterior tail areas, i.e. $(\mathbb{P}[\,\theta > \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,])$ against each other, as in Figure 2.1b. Here, the curved lines denote possible posterior tail areas under the Normal location problem. Incoherence follows because despite expanding the null interval (moving down and left) we can still cross from $d = C$ to $d = B$.

From Figure 2.1b, we see that the only way to avoid this incoherence happening, for at least some combination of posterior and nested null intervals, is to insist that the line between the critical region for C and other decisions has negative slope.

Expressed in terms of our loss function, the Schervish coherence condition states that $l_1 < l_2 + l_3$, i.e. that incorrect sign decisions ($d = A$ when $d = B$ is correct, for example)

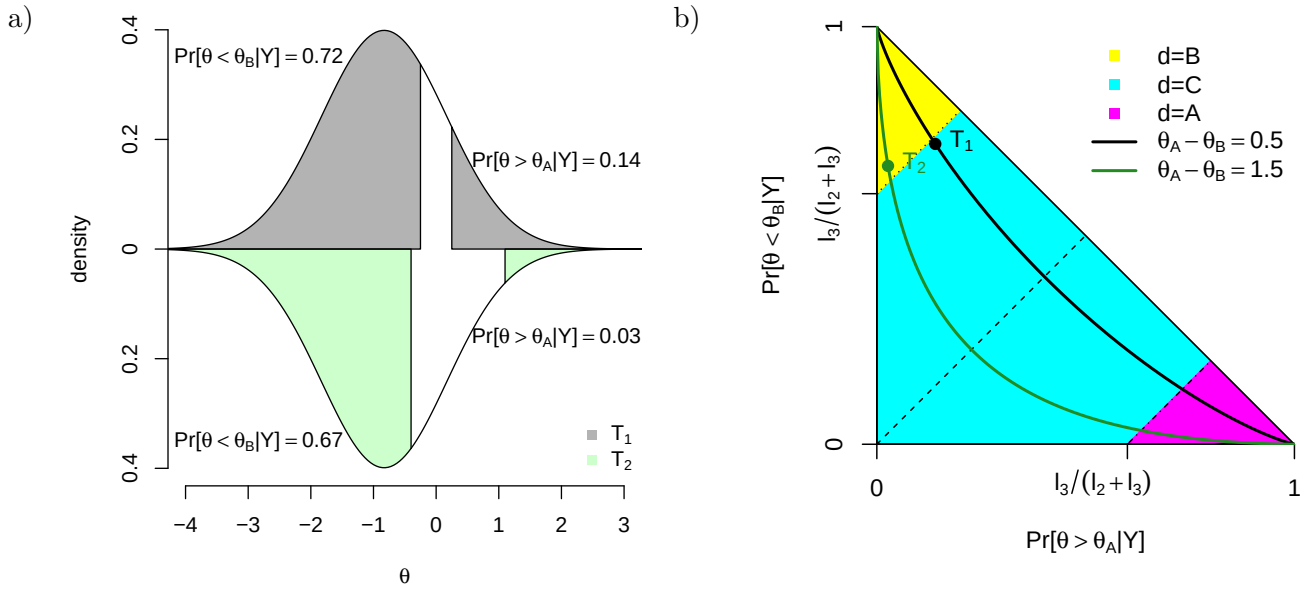


Figure 2.1: a) Posterior tail areas when $Y = -0.83$ and $(\theta_B, \theta_A) = (-0.25, 0.25)$ versus $(\theta_{B'}, \theta_{A'}) = (-0.4, 1.1)$. For T_1 we find $\mathbb{P}[\,\theta < \theta_B|Y\,] < 0.6 + \mathbb{P}[\,\theta > \theta_A|Y\,]$, so the null is accepted: for T_2 the opposite holds and it is rejected. b) The same example plotted in terms of posterior tail areas, i.e. $(\mathbb{P}[\,\theta > \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,])$. The curved lines denote possible posterior tail areas under the Normal location problem, with T_1 and T_2 from a) specified. Incoherence follows because despite expanding the null interval (moving down and left) we still cross from $d = C$ to $d = B$.

are not worse than the combined losses of incorrectly choosing $d = C$ and making a sign decision when the null holds.

To ensure coherence, another reasonable restriction on the Bayes rules is that they should not report $d = C$ when the posterior probability is zero, i.e. we should not accept the null when it has zero posterior support. We shall call this the *null support* property. In terms of the loss function it means constraining $l_2 > l_1/2$, i.e. the cost of incorrectly choosing $d = C$ is at least half the cost of an incorrect sign decision.

Graphical representations of the losses that do and do not meet these two constraints are given in Figure 2.2. We see directly that Schervish coherence (and the null support property) are compatible with some Bayesian tests of interval nulls, but not all. This result is general, and not limited to the special models considered by Schervish or Fossaluza *et al* [88, 28].

2.3.2 Incompatibility of Coherence and Unbiased Type I Error Rate Control

Having shown that some Bayes rules for testing provide coherence for general posteriors, we now address whether for the Normal location problem these coherent Bayes rules can approximate any of the family of UMPU tests considered by Schervish – the family indexed by α, the level at which the Type I error rate is controlled.

Specifically, for the Normal location problem where $Y \sim N(\theta, 1)$ we consider frequentist tests that are symmetric with respect to the interval null – meaning they reject the null only when $|Y - \frac{\theta_B+\theta_A}{2}| > r$ for some $r > 0$ – and Bayes rules under the limiting case of a $N(\mu, \tau^2)$ prior for θ. As is well-known (see e.g. [102, §3.7.1]) for any fixed finite μ, under the limit $\tau \to \infty$ where the prior becomes improper and flat we have

$$(\mathbb{P}[\,\theta > \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,]) = (1 - \Phi(\theta_A - Y), \Phi(\theta_B - Y)) = (\Phi(Y - \theta_A), \Phi(\theta_B - Y)).$$

This setting leads to the following result:

Lemma 1. *Frequentist tests of the form specified that control the Type 1 error rate, that is,*

$$\mathbb{P}[\,d = A; \theta\,] + \mathbb{P}[\,d = B; \theta\,] \leq \alpha$$

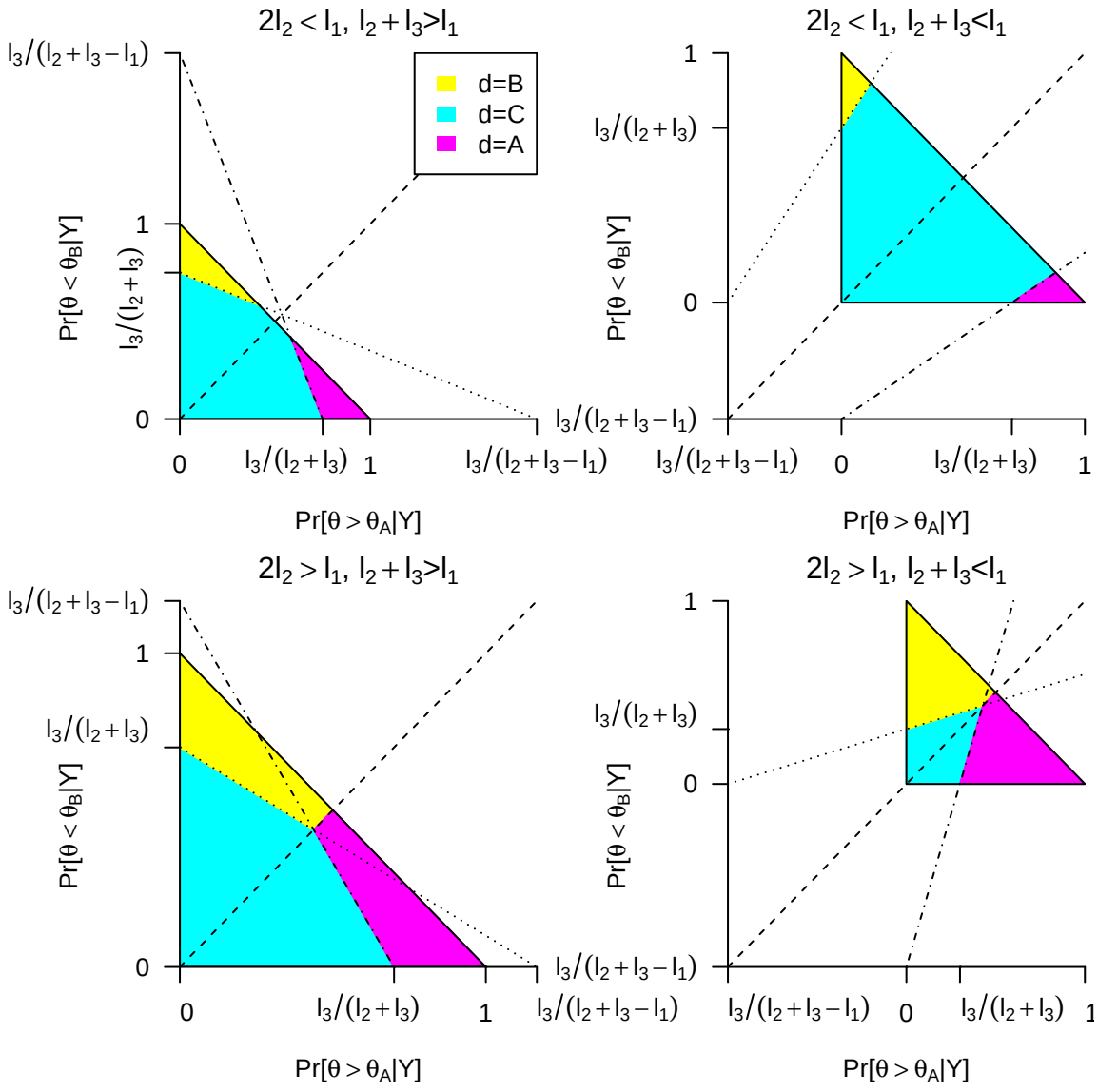


Figure 2.2: Possible decision rules for positive l_1, l_2, l_3. The x-axis is $\mathbb{P}[\,\theta > \theta_A|Y\,]$ and the y-axis is $\mathbb{P}[\,\theta < \theta_B|Y\,]$. The left/right columns represent losses for which the Bayes rule is Schervish incoherent/coherence, respectively. The top/bottom rows represent losses for which the Bayes rule provides the null support property, of not making a decision for which there is zero posterior support.

for fixed α and where the inequality is strict at some $\theta \in (\theta_B, \theta_A)$, results in the Bayes rule for loss (2.2) (in the limiting case of a flat prior on θ) only with $(l_1, l_2, l_3) \propto (2, \alpha, 1-\alpha)$, and are thus incoherent in Schervish's sense.

For proof see Appendix A.1. The solution, $(l_1, l_2, l_3) \propto (2, \alpha, 1-\alpha)$, is in fact the UMPU test. To see this directly, we note that the proper hypothesis-testing loss

		Decision		
		A	C	B
Truth	$\theta < \theta_B$	2	α	0
	$\theta \in (\theta_B, \theta_A)$	$1-\alpha$	0	$1-\alpha$
	$\theta > \theta_A$	0	α	2

is, without loss of generality, equivalent in practice to using loss

		Decision		
		A	C	B
Truth	$\theta < \theta_B$	2	α	0
	$\theta \in (\theta_B, \theta_A)$	1	α	1
	$\theta > \theta_A$	0	α	2

, (2.3)

which generalizes the significance testing loss of Rice et al [81]. Writing the loss this way we can see that decision $d = C$ is rejected only if

$$\tilde{P} = \min\left(\mathbb{P}[\,\theta > \theta_B | Y\,] + \mathbb{P}[\,\theta > \theta_A | Y\,], \mathbb{P}[\,\theta < \theta_B | Y\,] + \mathbb{P}[\,\theta < \theta_A | Y\,]\right) < \alpha,$$

in which case the decision is $d = A$ or $d = B$ depending on which alternative has more posterior support. But with limiting flat priors $\tilde{P}$ is

$$\min\left(\Phi(y-\theta_B) + \Phi(y-\theta_A), 1 - \Phi(y-\theta_B) + 1 - \Phi(y-\theta_A)\right),$$

in other words the UMPU test's p-value p_{θ_B,θ_A}, given in Equation (2.1).

Lemma 1 reiterates the incoherence of the UMPU test, but more importantly shows that it is not even close to coherence: with $(l_1, l_2, l_3) \propto (2, \alpha, 1-\alpha)$ the cost of incorrect sign decisions exceeds the sum of the losses for the other two incorrect decisions – indeed,

the incorrect sign decision has exactly twice the cost of the sum of the other losses. Thus, it clearly violates the coherence condition that $l_1 < l_2 + l_3$. The result strongly suggests that coherence of frequentist tests is only available if those tests are calibrated in terms of something other than Type I error rate, the familiar default.

2.4 Bayesian Conditions Compatible with Novel Frequentist Control Measures

When connecting the general loss function from Section 2.3 to frequentist criteria, it is helpful to reparameterize it, setting $l_2 + l_3 = 1$, and writing $l_1 = 1 - \delta$ for $0 < \delta < 1$. This gives loss

		Decision		
		A	C	B
Truth	$\theta < \theta_B$	$1-\delta$	γ	0
	$\theta \in (\theta_B, \theta_A)$	$1-\gamma$	0	$1-\gamma$
	$\theta > \theta_A$	0	γ	$1-\delta$

, (2.4)

where the positivity constraints on l_1, l_2, l_3 become conditions that $0 < \gamma < 1$ and $\delta < 1$. Schervish coherence corresponds to requiring $\delta \geq 0$, and the additional constraint precluding $d = C$ when the null has no support requires $\gamma > (1 - \delta)/2$.

We will establish a corresponding set of frequentist methods, in which γ and δ respectively describe the level of error control, and the precise error measure that is controlled. To do this we consider frequentist sign-tests that, as with the Bayes rules, return decisions $d = A$, $d = B$ or $d = C$, for the Normal location problem with $Y \sim N(\theta, 1)$. We assume symmetry of the tests around $Y = (\theta_A + \theta_B)/2$, meaning that for some r chosen to calibrate the test, the test returns $d = A$ for $Y < (\theta_A + \theta_B)/2 + r$, $d = B$ for $Y < (\theta_A + \theta_B)/2 - r$, and $d = C$ otherwise.

Lemma 2. *Frequentist tests of the form specified by controlling*

$$s_\delta : (1 - \delta) \max(\mathbb{P}[\, d = A; \theta\,], \mathbb{P}[\, d = B; \theta\,]) + \delta |\mathbb{P}[\, d = A; \theta\,] - \mathbb{P}[\, d = B; \theta\,]| \leq \gamma,$$

for fixed $0 < \gamma < 1$ *and* $0 \leq \delta \leq 1$ *and where the inequality is strict at some* $\theta \in (\theta_B, \theta_A)$*, are Schervish-coherent. In the limiting case of a flat prior on* θ *they are also the Bayes rule for loss (2.2) with* $(l_1, l_2, l_3) \propto (1 - \delta, \gamma, 1 - \gamma)$.

For proof see Appendix A.2. The results shows that a weighted average of two quantities – the maximum of two sign error rates and the absolute difference between those sign error rates – can be used to give a general frequentist analog of the Bayesian tests that are Schervish coherent. Both quantities differ from Type I error rate, control of which in our notation is stated as $\mathbb{P}[\, d = A; \theta\,] + \mathbb{P}[\, d = B; \theta\,] \leq \gamma$, corresponding to use of $\delta = -1$.

We also find it useful to generalize the familiar notion of a p-value to the p_δ-value, with $p_\delta(Y)$ being the largest γ under which decision $d = C$ would still result for data Y. Under this notation, $p_{-1}(Y)$ denotes the standard UMPU p-value, while values of δ between 0 and 1 are guaranteed to behave coherently. (We note that Peskun [77] previously reported the special case of the p_0-value and its coherence, but without giving the frequentist error control shown here.) Plots of $p_\delta(Y)$ for δ = -1, 0, and 1 are given in Figures 2.3a/b/c respectively. The Schervish coherence of tests with $\delta \geq 0$ is immediately apparent from the later plots: moving left or up only increases the p_δ-value, unlike in Figure 2.3a's depiction of the UMPU test's p-value.

Figure 2.3d shows the p_δ-values for $\delta = 1/2$, illustrating how the contours combine the characteristics of those in Figures 2.3b and c. The null support property of Section 2.3 is also apparent on this plot: the contours corresponding to $p_\delta(Y) = \gamma > (1 - \delta)/2$ are those that do not intercept the 45-degree line $\theta_A - Y = \theta_B - Y$.

2.4.1 Examples

We first revisit the example from Figure 2.1, but now with a frequentist interpretation. For the model with Y $N(\theta, 1)$ and observation $Y = -0.83$, we test the hypotheses that $\theta \in (-0.25, 0.25)$, and $\theta \in (-0.40, 1.10)$, in which the two null intervals are nested. Using the default UMPU test, the smaller null has $p_{-1} = 0.42$ while the larger null has $p_{-1} = 0.36$. This means that testing with (say) level $\alpha = 0.4$, the results will be incoherent: the larger null is not rejected but not the smaller one within it. Figure 2.4a shows how using values of δ between 0 and 1, this does not occur. The p_δ-value for the larger null is larger, and at any α for which the larger null is rejected the smaller null is also rejected. (Coherence for other values of δ follows by Lemma 2.)

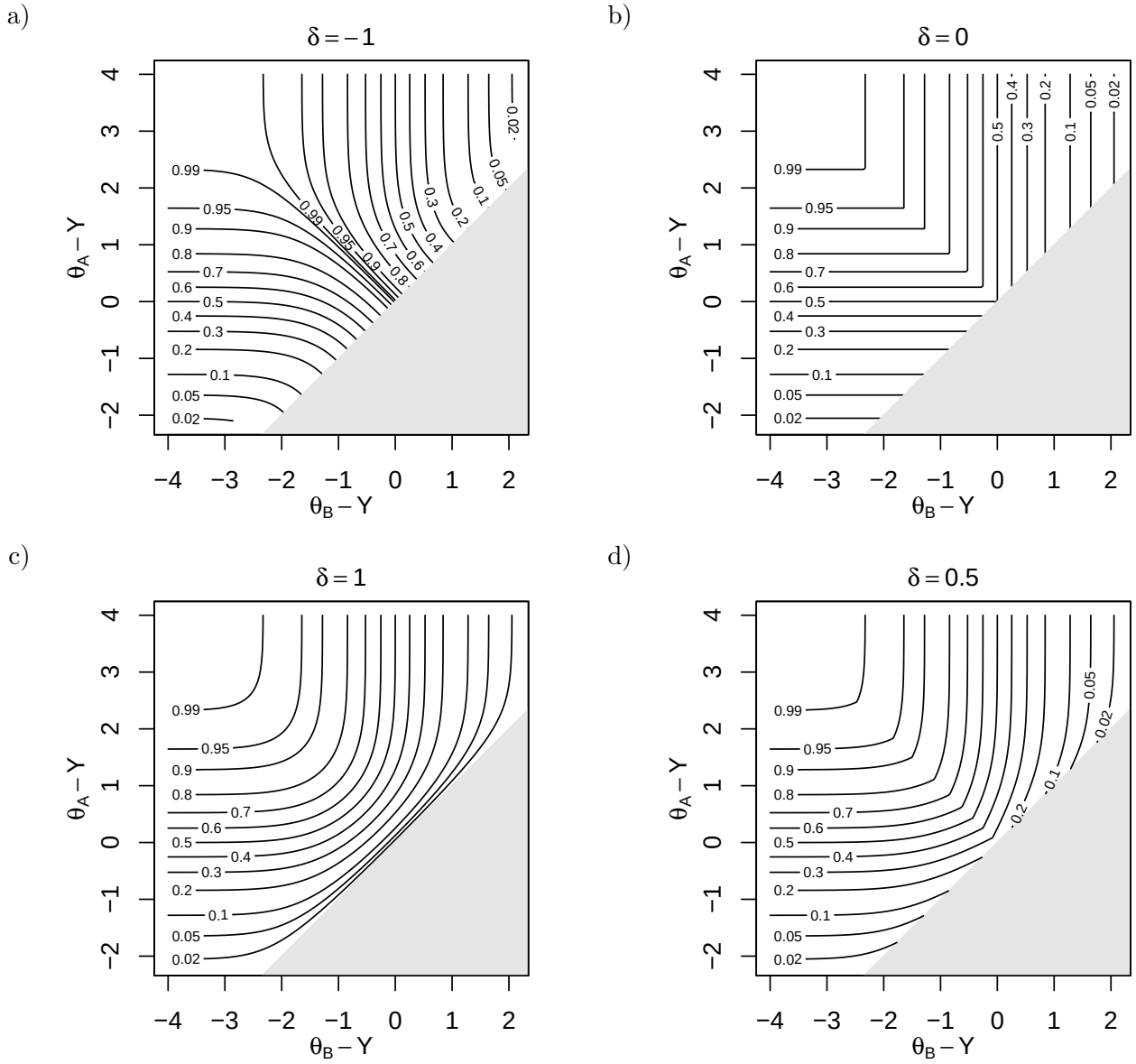


Figure 2.3: For the Normal location problem with $Y \sim N(\theta, 1)$ and interval null (θ_B, θ_A), contour plots of p_δ-values, for selected values of δ. For $\delta = -1$ (panel a) p_δ is the usual p-value from the UMPU test. For δ between 0 and 1 the corresponding tests are Schervish coherent. Contours that do not intercept $\theta_A - Y = \theta_B - Y$ (i.e. the grey shaded region) have $p_\delta(Y) > (1-\delta)/2$, i.e. they correspond to tests with the null support property.

Our second example is that of Schervish's paper [88], discussed in Section 2.2.1, in which $Y = 2.18$ and we test nested null intervals $(-0.5, 0.5)$ and $(-0.82, 0.52)$. As Figure 2.4b shows, the corresponding UMPU tests give p_{-1}=0.0502 and 0.0498 respectively and so with $\alpha = 0.05$ (or in a small range around this value) are incoherent. After the lines on the plot cross, tests using p_0 to p_1 are however coherent.

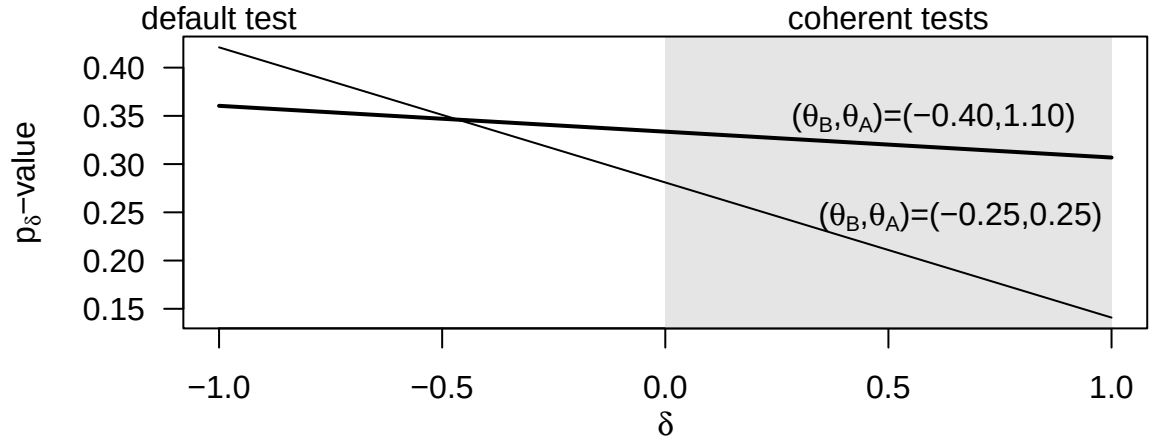
2.5 Discussion

Schervish's example of incoherence – for an optimal test – is a striking argument against the use of classical frequentist tests, in general, as assessments of evidence. Their use as a foundation for statistical inference is also therefore questioned. To consider the consequences of these results, it may help to distinguish continuous-valued p-values (or their generalizations) from discrete-valued tests.

For p-values, Schervish's example shows that p-values from tests with widely-accepted UMPU optimality, do not behave as measures of evidence. We have shown that the Bayesian $\tilde{P}$ value – a posterior summary sufficient for determining whether the Bayes rule (which is optimal for its loss function) rejects the null – also fails in this regard. One might question the optimality criteria; taking a different frequentist approach (e.g. that of [14]) it appears possible to retain forms of coherence. The standard Bayesian approach affords less flexibility: with the utility described by the sign-decision loss function (2.3) one can have optimality, or coherent summary measures, but not in general both.

Of course, p-values can be retained if viewed in other ways. Decoupling them from tests, one may view the p-value as summarizing information; in general as a measure of surprise in the data, or compatibility of the data with a null value [79]. For many widely-used tests, we may similarly view p-values as transformed measure of signal-to-noise. This alternative view can also be Bayesian and decision theoretic: quantifying 'signal' and 'noise' via the posterior mean and standard deviation respectively, a Bayesian analog of two-sided p-values is optimal for a loss that trades estimation accuracy for discrepancy between the true θ and a null value [80, §3]. Clearly, if extended to interval nulls in the same way as here the p-value cannot be Schervish coherent. Nevertheless, coherence-like properties of the standard p-value can hold if we view the p-value as measuring data-model conflict; see e.g. [9].

a) Y=2.18 (Figure 1 example)



b) Y=-0.83 (Schervish's example)

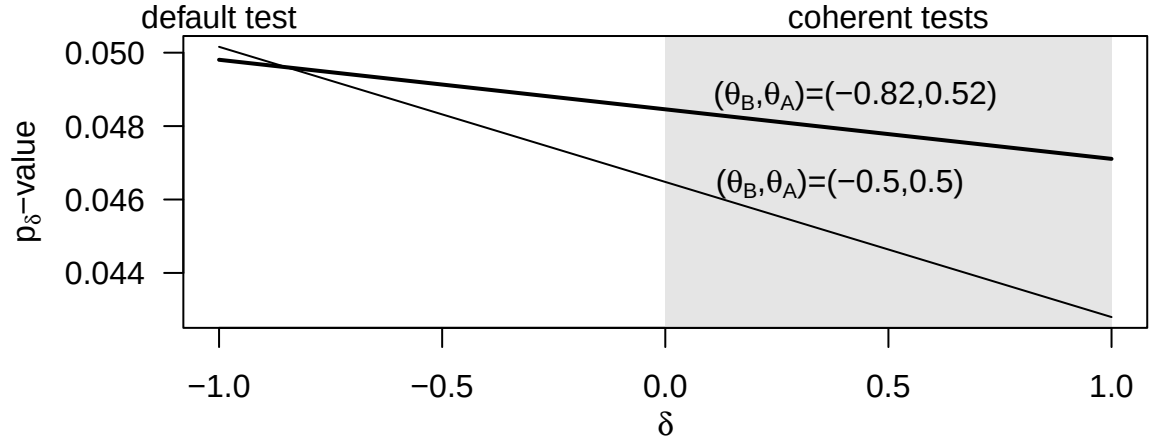


Figure 2.4: Plots of the p_δ-value, as a function of δ, for the examples of Section 3.1. The value at $\delta = -1$ gives the standard p-value from the UMPU test. Values of δ between 0 and 1 give coherent tests. The incoherence of the UMPU test is shown, in each case, by the crossing of the two lines.

In this chapter we have shown that, perhaps surprisingly, coherence of decisions is not automatic in the Bayesian framework. Nevertheless, for the interval nulls considered Schervish coherence of sign-decisions can be achieved through application of simple linear constraints, that apply regardless of the sampling model or prior. In short, using Bayesian decision theory makes this desirable property easy to obtain.

The flexibility of the Bayesian approach provides not one coherent test but a natural set of them, all of which have frequentist analogs. While we do not claim that these are the only frequentist approaches that yield Schervish coherence, we have shown that unbiased control of Type I error rate is incompatible with it for interval nulls, requiring that other criteria be considered. The set of other criteria for testing interval nulls seems to date unexplored, both in terms of their theoretical properties, but also practical issues of what quantities should be controlled to calibrate the scientific testing situation at hand.

The potential of the Bayesian decision framework could potentially be extended to consider overlapping as opposed to nested intervals, i.e. if we test

$$H_{01} : \theta \in (-1, 1),$$

and also test

$$H_{02} : \theta \in (-0.5, 1.5).$$

Here, we would need to extend the decisions beyond A, B, and C. A potential starting point would set up cut-points at $-1, -0.5, 1$, and 1.5, and constraining loss functions in some way to illustrate when one could make an incoherent decision regarding H_{01} and H_{02}. We will not consider this type of problem further here.

We hope that the connections we have provided with Bayesian decision theory make this process easier: our loss functions directly state that a sign-decision question is addressed, with quantified measures of the value of different correct and incorrect results. These may or may not match up with scientific goals in performing the analysis, but their clear statement should facilitate critical discussions connecting those scientific goals with the statistical methods to be used.

This work has been accepted for publication in *The American Statistician*.

Chapter 3

EXACT INFERENCE FOR FIXED-EFFECTS META-ANALYSIS OF PROPORTIONS

3.1 Introduction

Meta-analysis of proportions, in which counts of independent binary successes/failures are aggregated in some way over multiple studies, is widely-used in practice. The overall proportion may be of direct scientific interest [35], or may represent a data summary that usefully frames a subsequent comparison [108], obtained perhaps using regression methods.

In the simplest case of homogeneity (i.e. when each study has the same underlying true proportion), meta-analysis of proportions can just add study-specific binomial outcomes to give a binomially-distributed total, for which inference is well-studied (see e.g. [2, §1.4]) and exact inference is readily available [24, 17]. Inference based around the overall total also has known optimality properties—for example tests that are uniformly most powerful unbiased—and so is a compelling default approach. Alternative exact methods are also available under homogeneity, via user-specified weighted combinations of exact confidence intervals [96], p-values [58] or more generally confidence distributions [106] for the same parameter estimated across multiple studies.

Under heterogeneity of the study-specific proportions there is no such default available, and more complex methods are often introduced. One popular approach is the Freeman-Tukey double arcsine transformation [29, 66], which stabilizes the variance of study-specific contributions, which are then meta-analyzed using standard fixed- or random-effects approaches. However, as well as the obvious difficulty of interpreting estimates on this unfamiliar scale, its results can be misleading [90] in meta-analysis of differently-sized studies. Taking a similar approach with alternative log, logistic and square-root arcsine transformations – or no transformation at all – difficulties persist in small sample settings, as instability in the variance estimates used in the weights for standard meta-analysis methods is propa-

gated to the corresponding results. Direct representation of the binomial outcomes within a single model is available via logistic generalized linear mixed models or beta-binomial models [98, 61], but with them comes the challenge [38] of motivating a sampling model for parameters that are not truly stochastic. Exact inference for beta-binomial models has been provided by Gronsbell *et al* [33], exploiting an underlying common hyper-parameter similarly to the approach of Tian *et al* [96], but requiring a user-chosen grid of evaluation points and use of a continuity correction. Except Gronsbell *et al*'s method, none of the approaches above allowing for heterogeneity are exact, and instead rely on large-sample approximations. Use of continuity corrections is also common [98], but analyses can be sensitive to exactly how these are implemented, particularly when some study outcomes are all successes or all failures [61].

In this chapter we address these problems, providing a simple exact method for fixed-effects meta-analysis of proportions. Our approach, which requires no continuity correction, provides inference on the overall proportion estimated by simply pooling all study-specific results. We stress, and will show, that its confidence intervals and p-values are valid under heterogeneity.

The rest of the chapter is structured as follows: in Section 3.1.1 we define notation, before Section 3.1.2 provides the main theoretical results and Section 3.1.3 describes how they complement assessments of heterogeneity. The confidence interval method provided is evaluated in Section 3.2, before we give an applied example in Section 3.3. We conclude the chapter with a short discussion.

3.1.1 Description of the problem

We assume that independent studies $i = 1, ..., k$ each randomly sample n_i participants from their corresponding population, and hence study-specific number of successes Y_i is independent and distributed as $\text{Bin}(n_i, p_i)$ where p_i denotes the probability of success in study i. Denoting $\hat{p}_i$ as each study's estimated proportion $\hat{p}_i = Y_i/n_i$, and the total sample size as $n_+ = \sum_{i=1}^{k} n_i$, we define the meta-analytic estimate of proportion as

$$\hat{p}_F = \frac{\sum_{i=1}^{k} Y_i}{n_+} = \sum_{i=1}^{k} \frac{n_i}{n_+} \hat{p}_i. \tag{3.1}$$

We note that overall proportion $\hat{p}_F$, using weights proportional to n_i, is similar but not identical to a precision-weighted average of the $\hat{p}_i$, in which the weights are proportional to $n_i/(\hat{p}_i(1-\hat{p}_i))$.

For an overall population comprising k subpopulations, each of relative size η_i proportional to that observed in the samples (i.e. η_i proportional to n_i), the overall population proportion of successes can be written as

$$p_F = \sum_{i=1}^{k} \eta_i p_i,$$

which is the quantity estimated by $\hat{p}_F$. Under homogeneity, all p_i are equal and hence any form of weights η_i would give the same overall proportion. But even under heterogeneity, $\hat{p}_F$ estimates the proportion of successes in a population that is a simple amalgamation of all the study-specific populations being considered. This straightforward interpretation makes p_F an appealing starting point for inference, but it is particularly relevant when testing. If the 'strong' null is that all $p_i = p_0$, for some null value p_0, then rejection of the 'weak' null hypothesis $p_F = p_0$ implies rejection of the strong null: at least one of the p_i must be non-null for $p_F \neq p_0$ to occur. While rejecting $p_F = p_0$ is not the only way to invalidate the strong null, in situations with limited data it may well be an efficient way to use the data that is available.

The mean and variance of $\hat{p}_F$ can be shown to be

$$\mathbb{E}[\,\hat{p}_F\,] = \sum_{i=1}^{k} \frac{n_i}{n_+} p_i = p_F, \qquad \mathrm{Var}[\,\hat{p}_F\,] = \frac{1}{n_+} \sum_{i=1}^{k} \frac{n_i}{n_+} p_i(1-p_i). \tag{3.2}$$

We note that $\mathrm{Var}[\,\hat{p}_F\,]$ can also be written as

$$\frac{p_F(1-p_F)}{n_+} - \frac{1}{n_+} \sum_{i=1}^{k} \frac{n_i}{n_+} (p_i - p_F)^2, \tag{3.3}$$

meaning that, for a given p_F, the estimate's distribution is most diffuse under homogeneity [68].

Inference for p_F in *large-strata* settings (i.e. when each n_i is large and the p_i are fixed) is straightforward. By the Central Limit Theorem each $\hat{p}_i$'s distribution is approximately Normal with mean p_i and variance $p_i(1-p_i)/n_i$. Hence $\hat{p}_F$ is also approximately Normal,

with the mean and variance given above. For *sparse-data* settings, in which the number of strata k grows while the n_i are restricted, the situation is more complex but approximate Normality can follow under the Lyapunov or Lindeberg central limit theorems [15, §27].

With large strata, plugging-in the $\hat{p}_i$ to estimate the variance, its square root gives an approximate standard error estimate. Intervals a fixed number — $\Phi^{-1}(1-\alpha/2)$ — of estimated standard errors either side of $\hat{p}_F$ are approximate $(1-\alpha)\times 100\%$ confidence intervals, and seeing whether p_0 lies outside this form of interval gives an approximate level α test of the null hypothesis that $p_F = p_0$. With sparse data, the situation is again more complex, but with appropriate estimates of the standard error, the same results will hold under asymptotic Normality.

In the rest of this chapter our focus is on situations where the numbers of successes (or non-successes) is small, i.e. where some or all of the Y_i or n_i-Y_i are small. The formal inference we provide is exact, for any number of strata of any size, and does not rely on asymptotic limits.

3.1.2 Theoretical results

The observation above, that the distribution of Y_+ is most diffuse under homogeneity, does not only apply when variance is used to measure diffusion. As shown by [41, Theorem 5], for any two integers b and c such that $0 \le b \le n_+p_F \le c \le n_+$,

$$\sum_{r=b}^{c} \binom{n_+}{r} p_F^r (1-p_F)^{(n_+-r)} \le \mathbb{P}[\, b \le Y_+ \le c\,] \le 1.$$

In other words, for a fixed overall proportion p_F the tails of Y_+ are heaviest under homogeneity. Appendix B.1 shows how this result leads to formally 'exact' tests and confidence intervals – meaning that coverage is guaranteed to be at least the nominal level.

The actual extent of the excess coverage of this approach under heterogeneity can be approximated informally. Defining

$$\delta^2 = \sum_{i=1}^{k} \frac{n_i}{n_+}(p_i - p_F)^2, \tag{3.4}$$

we see from Equations (3.2) and (3.3), that in large samples the effect of tests that are exact under homogeneity is equivalent to using a Wald test based on $\hat{p}_F$, in which the

variance of this estimate is assumed to be $p_F(1-p_F)/n_+$, instead of the appropriate term $p_F(1-p_F)/n_+ - \delta^2/n_+$. In this case, regardless of the relative size of the individual n_i, the corresponding standard errors are therefore too large by a factor we denote as

$$\xi(p_F, \delta) = \sqrt{1/\left(1 - \frac{\delta^2}{p_F(1-p_F)}\right)},$$

and instead of the confidence interval having level $1-\alpha$, its approximate level is $1 - 2\Phi\left(\Phi^{-1}(1-\alpha/2)\xi(p_F,\delta)\right)$.

Hoeffding's result apply for any test that is exact under homogeneity, so long as it rejects only for $Y_+ \notin (b, c)$, i.e. in some central interval. It also applies to any exact confidence interval, for which p_F lies outside the interval only when Y_+ does not lie in some acceptance region (b, c). In what follows we shall consider the exact intervals suggested by Blaker [17], the good properties of which are noted by Agresti [2, §1.4.4], and which are easily computed using R's `exactci` package [27]. Details of the method are given in Appendix B.2, which also describes how the corresponding Blaker p-values, while valid, can have counterintuitive properties in some cases [101]. Should testing and p-values be the focus of the analysis we would instead recommend the widely-used Clopper-Pearson method [24], which are exact, do not have the noted non-inuitive properties but are slightly more conservative than those of Blaker.

3.1.3 Assessing heterogeneity

As Section 3.1.1 describes, the fixed-effects analysis estimates an overall proportion, by averaging over the proportions in the different studies. Parameter δ, defined via Equation (3.4), defines a weighted standard deviation of the p_i, which may be a useful quantitative summary of how much the p_i differ. Formal inference on the heterogeneity of study-specific proportions is also available, under the same assumption of fixed effects in each of the studies. Specifically, we can use Fisher's exact test [2, §3.5.1] on the $2 \times k$ contingency table of success/failures across the k studies, which controls the Type I error rate at the nominal level under the null hypothesis that all unknown p_i are equal. As with our exact tests for p_F, the actual Type I error rate may be conservative.

Following the testing strategies described by Rice *et al* [82, §5.1], the test of homogeneity may be a useful first step when determining how the underlying proportions differ by subpopulation, or may be used secondarily after inference on overall rate p_F. Extending the approach of Section 3.1.1, the test may also be used as part of assessments of the 'strong' null, that all p_i equal the same null value – which can be overturned if the p_i differ, as well as if weighted average p_F is not equal to that null value.

We caution, however, that while Fisher's test has optimality properties [97] and can be considered a reasonable approach in our rare-event setting with multiple strata [63], the power of any test in such settings may be modest unless heterogeneity of the p_i is substantial. In practical terms, this means that despite the presence of heterogeneity – in some settings large enough to be of practical significance – the data may also be insufficient to rule out homogeneity. Consequently, using the test of homogeneity to "pre-test" whether one should account for heterogeneity is not recommended. A prudent default approach is to always account for heterogeneity, as our methods permit.

3.2 Evaluation

We evaluate the coverage of the Blaker intervals under homogeneity and heterogeneity of various forms. In keeping with our focus on small samples we keep $n_+ = 50$. Complete enumeration is used for all calculations.

In the evaluations, the p_i are chosen by forcing p_F and δ equal to specified value. For $k = 2$ (in Figure 3.1) we consequently choose

$$\begin{aligned} p_1 &= p_F - \sqrt{\frac{n_2}{n_1}}\delta \\ p_2 &= p_F + \sqrt{\frac{n_1}{n_2}}\delta. \end{aligned}$$

For higher values of k (in Figure 3.2) we choose the p_i to be equally-spaced values around p_F, with location and spread chosen to fix p_F and δ. Hence for example with $k = 5$ and all $n_i = 10$, we set

$$(p_1, p_2, p_3, p_4, p_5) = p_F + (-2, -1, 0, 1, 2) \times \delta/\sqrt{2},$$

and throughout we only evaluate values of p_F which, for the given δ, result in all p_i lying between 0 and 1.

Throughout Figures 3.1 and 3.2, we see that with moderate heterogeneity of $\delta = 0.1$, the heterogeneity induces extra conservatism of only up to approximately one percentage point. For strong heterogeneity ($\delta = 0.2$) the conservatism is stronger, with excess coverage of three percentage points possible, although most values are smaller. These values should be compared to the "chatter" observed in Figure 3.1a, where the exact coverage of the Blaker intervals may vary by several percentage points depending on the exact value of p_F.

Figures 3.1 and 3.2 also show the accuracy of the Wald test-based approximation, even at this sample size. For both larger and smaller strata it effectively acts as an upper bound on how much extra coverage the Blaker interval will have under heterogeneity, over and above the conservatism it has under homogeneity. This is useful in practice: if the sum of the coverage of the Blaker interval under homogeneity with plausible p_F plus the approximate excess coverage with plausible δ gives an overall coverage that is acceptable, then the Blaker interval should be deemed acceptable for use even under heterogeneity.

Finally, while not immediately apparent from Figures 3.1 and 3.2, the true coverage of the Blaker intervals at specified δ and p_F is extremely similar regardless of whether $k = 2, 5$ or 10, or the variability of the n_i. Were Figures 3.1b,c and and 3.2a,b,c superimposed, the lines for excess coverage under $\delta = 0.1$ would all be visually indistinguishable, as would those for $\delta = 0.2$. This strongly suggests that with adequate knowledge of p_F, δ and n_+, we need no further information to assess the coverage of the Blaker intervals under heterogeneity.

Figure 3.3 shows, for the same settings, the power of Fisher's Exact Test with nominal $\alpha = 0.05$. In line with previous findings [87, 73] we see that in almost all settings with $\delta = 0.1$, the power is so low that significant findings would not be expected, and heterogeneity would most likely not be detected. With $\delta = 0.2$ – when meta-analysis contrasting study results with all events and zero events would be typical – then power is greater. Across all values of p_F and δ, power is notably lower when multiple small studies are present. In the most extreme sparse-data setting, when all $n_i = 5$, power is below 50% for all simulation settings, even those with $\delta = 0.2$.

For further evaluation of the methods, Figures 3.4 and 3.5 show, for $n_1 = n_2 = 25$

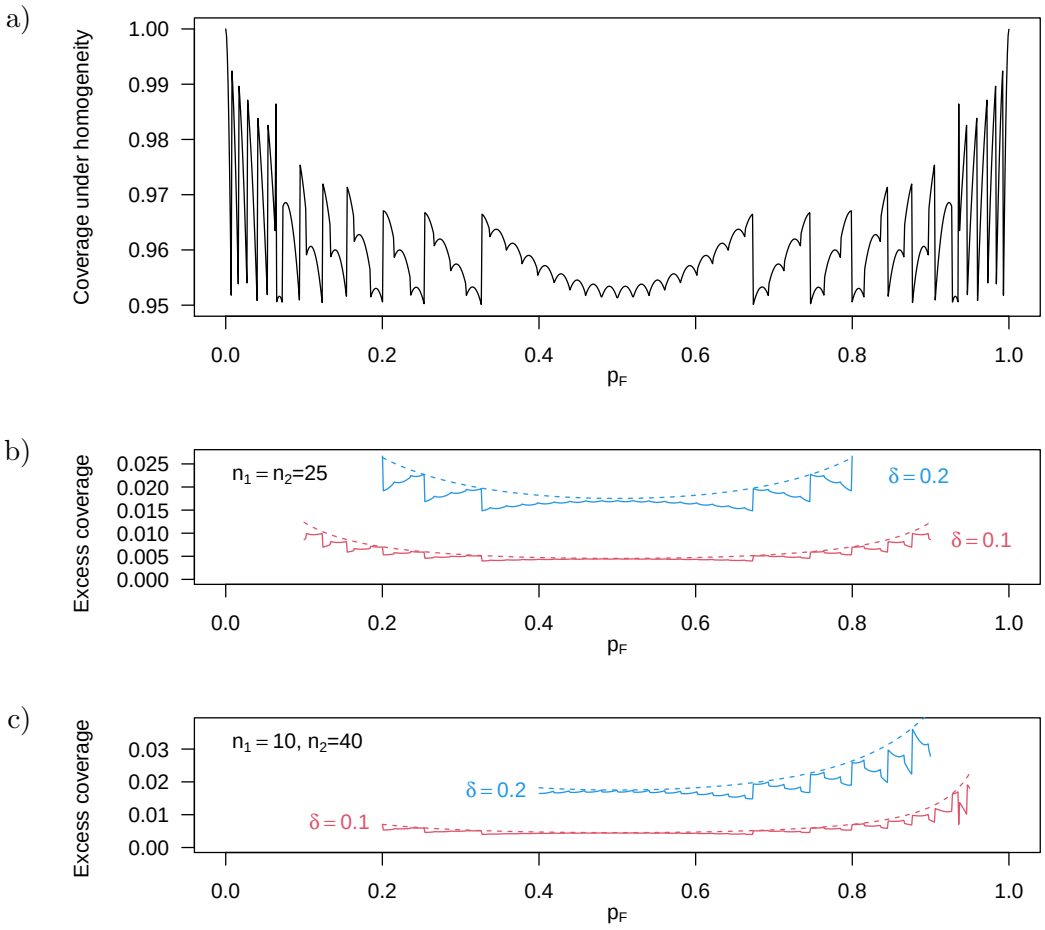


Figure 3.1: a) Coverage of the 95% Blaker intervals for p_F under homogeneity, for $n_+ = 50$ b) Excess coverage of the same intervals for $k = 2$ under mild ($\delta = 0.1$) and severe ($\delta = 0.2$) heterogeneity, with balanced groups. c) Excess coverage of the same intervals for $k = 2$ under mild and severe heterogeneity, with unbalanced groups. In b) and c) solid lines depict true excess coverage, dotted lines give its Wald test-based approximation.

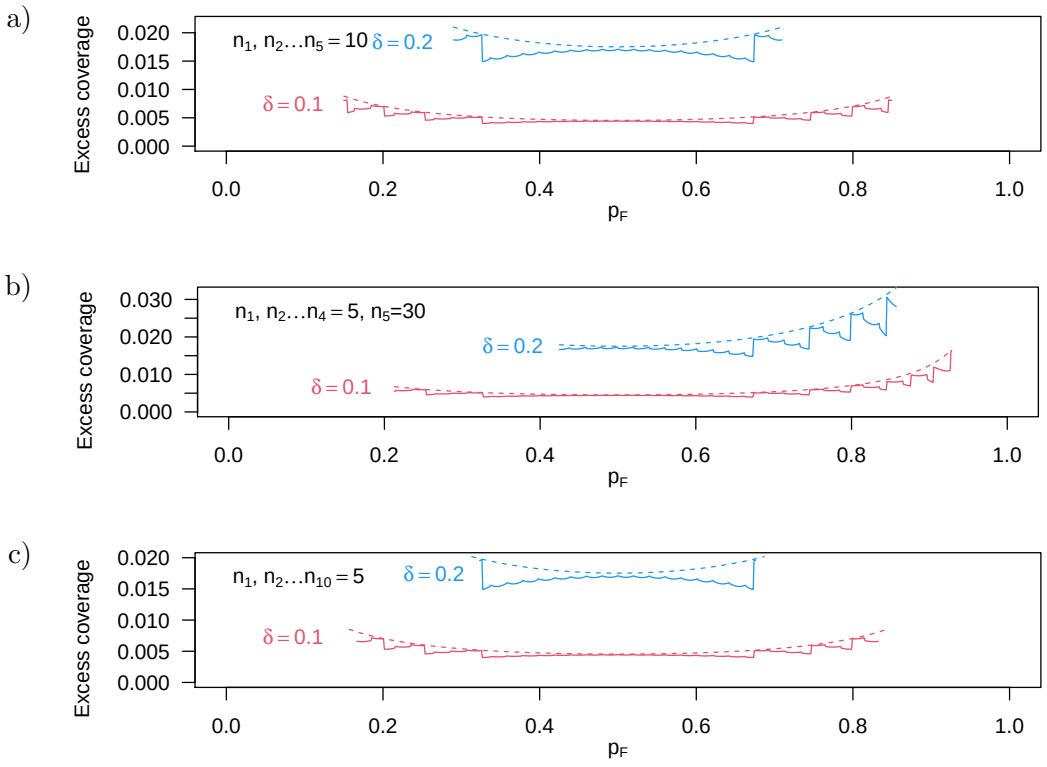


Figure 3.2: a) Excess coverage of the 95% Blaker intervals for p_F under homogeneity, for $k = 5$ equal groups, under mild ($\delta = 0.1$) and severe ($\delta = 0.2$) heterogeneity. b) Excess coverage for $k = 5$ unequal groups c) Excess coverage for $k = 10$ equal groups. Solid lines depict true excess coverage, dotted lines give its Wald test-based approximation.

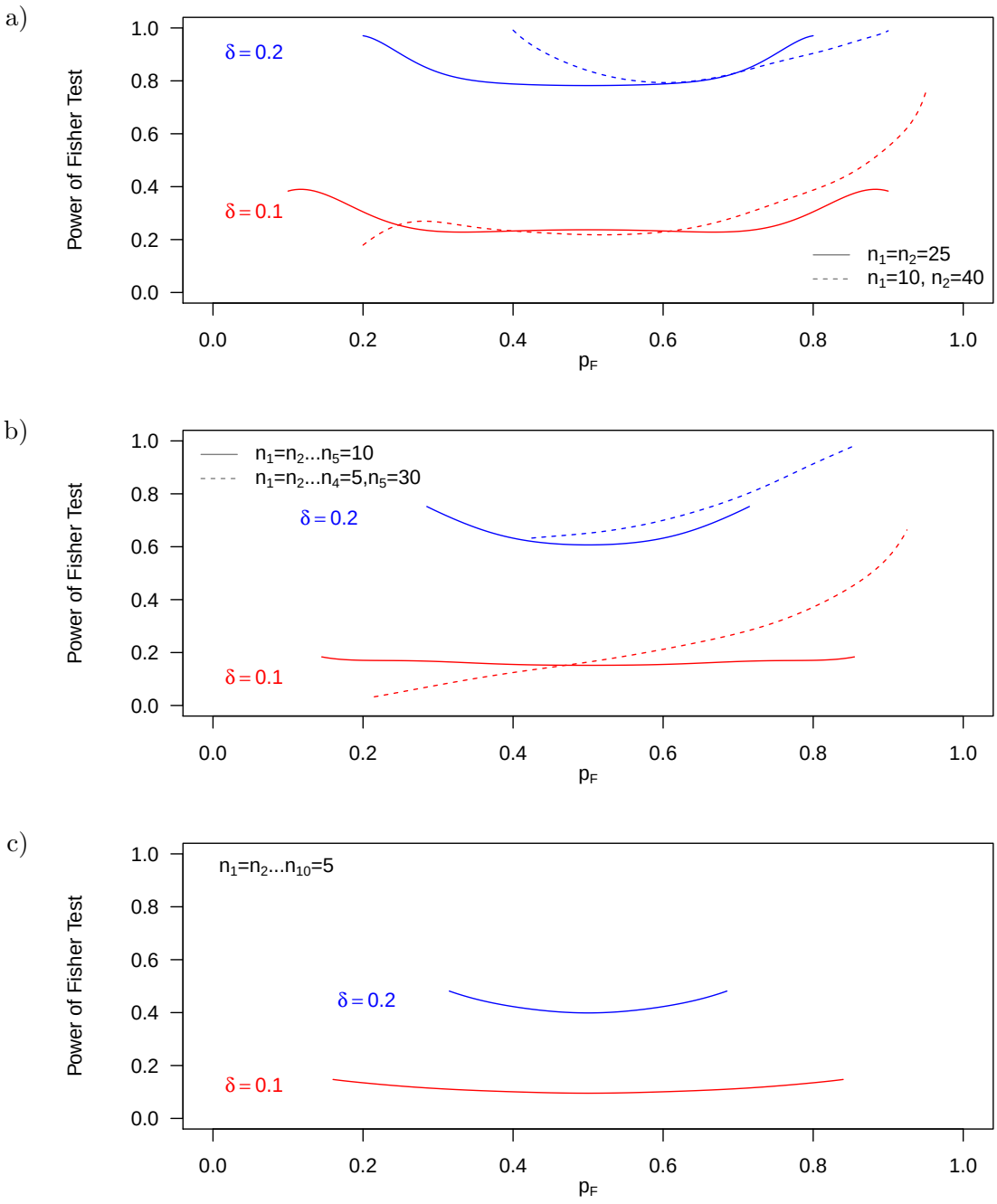


Figure 3.3: Power of Fisher's Exact Test with $\alpha = 0.05$ for $n_+ = 50$, with a) $k = 2$ b) $k = 5$ c) $k = 10$. Simulation settings are those of Figures 1 and 2.

and $n_1 = 10, n_2 = 40$ respectively, performance for all values of p_1, p_2. This includes the Blaker 95% intervals true coverage of the corresponding p_F, excess coverage of p_F under heterogeneity of the p_i versus homogeneity, and the accuracy of the approximation given in Section 3.1.3. The results are broadly as here: at extreme p_i the Blaker interval can be conservative – as can be expected for any exact method: see e.g. [2], §3.5.4 and §3.6.3. The extra conservatism induced by heterogeneity is mild unless the p_i are close to opposite extremes of the [0,1] range – which seems unlikely in practice. The approximation of the extra coverage works well at all values, with least accurate performance again when the p_i diverge strongly.

3.3 Application to meta-analysis of rare events data

We consider an example from the Kidney Precision Medicine Project (KPMP) [78], that recently meta-analysed proportions of various events, in follow-up of subjects who had undergone native kidney biopsies. We focus on the outcome of having pain at the biopsy site, which was collected for $k = 18$ studies. Pain is an outcome with clear heterogeneity, while the definition of pain may not have been consistent across studies, it is still important to have an overall estimated proportion (and CI) to convey risk to patients. The data, with study-specific 95% Blaker confidence intervals are plotted in Figure 3.6a.

Fixed effects analyses using transformations of the study-specific $\hat{p}_i$ are given in Figure 3.6b; specifically we consider inverse-variance combinations of the raw proportions, log-transformed, logistic-transformed (i.e. log odds), the square-root arcsine transform and double arcsine transform. All results are re-transformed back to the original proportion scale; for the double arcsine transform we use the inverse function due to [66]. (As noted in Section 3.1 these are all standard approaches, and are available in standard software such as R's `metafor` package [100].) Figure 3.6b also shows the exact interval obtained using the Blaker method, with total 194 events from total 4413 biopsies.

While it is not possible to assess the true coverage rate of this interval – as the true p_i cannot be known – we can obtain some idea of its conservatism by i) evaluating the coverage of the Blaker interval under homogeneity, for $n = 4413$ and a range of p_F, as in Figure 3.6c and ii) calculating the approximate excess coverage, using an estimate of δ obtained by

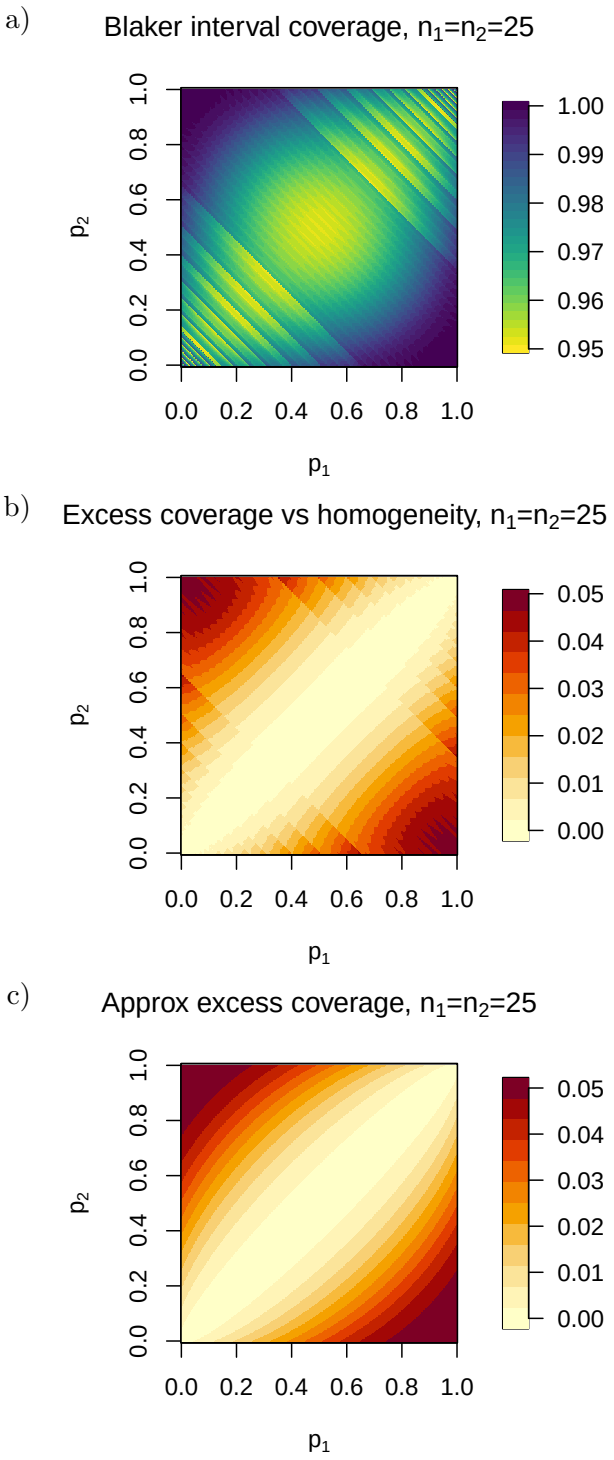


Figure 3.4: a) Coverage of the 95% Blaker intervals for $n_1 = n_2 = 25$, for all p_1, p_2 and corresponding $p_F = (p_1+p_2)/2$. b) Excess coverage of the Blaker interval for heterogenenous p_1, p_2 compared to using that interval with $p_1 = p_2 = p_F$. c) Approximation of that difference obtained from $\lambda(p_F, \delta)$: the maximum discrpancy from the true excess coverage is 1.3%, which occurs only when $|p_1 - p_2|$ approaches 1.

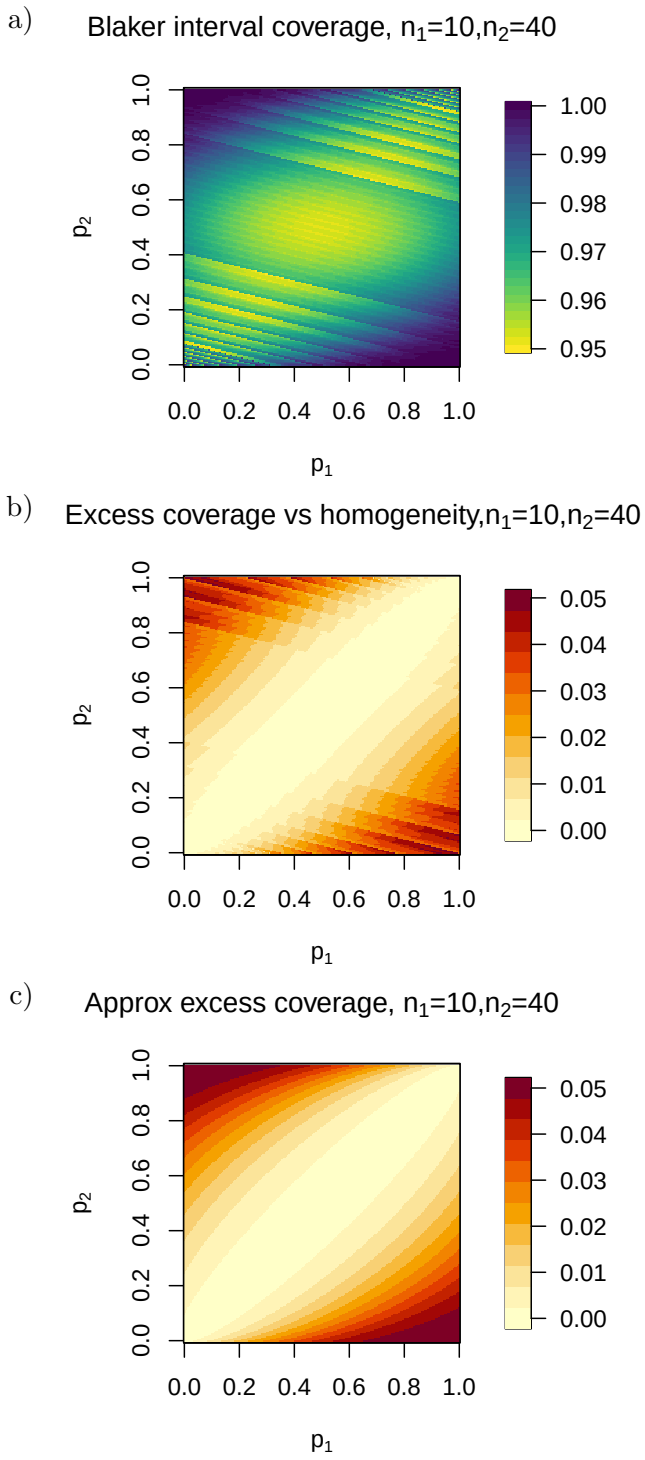


Figure 3.5: a) Coverage of the 95% Blaker intervals for $n_1 = 10, n_2 = 40$, for all p_1, p_2 and corresponding $p_F = (p_1 + 4p_2)/5$. b) Excess coverage of the Blaker interval for heterogenenous p_1, p_2 compared to using that interval with $p_1 = p_2 = p_F$. c) Approximation of that difference obtained from $\lambda(p_F, \delta)$: the maximum discrpancy from the true excess coverage is 2.1%, which occurs only when $|p_1/4 - p_2|$ approaches 1.

plugging in the $\hat{p}_i$ and $\hat{p}_F$ to Equation (3.4), to give $\hat{\delta} = 0.06$. (This plug-in estimate is biased, but as seen in Appendix B.3 it is biased upwards, i.e. conservatively, and the bias decays with the reciprocal of the overall sample size.)

To assess sensitivity to this value we also consider $\delta = 0.05$ and $\delta = 0.07$, which is the central 95% of values of $\hat{\delta}$ that are seen when $Y_i \sim \text{Bin}(n_i, \hat{p}_i)$ independently. Figure 3.6d shows the corresponding approximate excess coverages.

From the evaluation of Figures 3.6c and 3.6d it is therefore reasonable to assume that the Blaker interval for the KPMP pain example is a realization of an interval with no more than 97.5% coverage. The differences between the Blaker intervals and the alternatives shown in Figure 3.6b are therefore not plausibly due to chance alone; those methods must be seen as estimating notably different parameters, and doing so with confidence intervals (except Blaker's) that do not provide exact coverage. The various transformations and corresponding weights lead to very different estimated overall effects, with an order of magnitude difference in the overall estimates, and zero overlap between all but two pairs of 95% confidence intervals we could choose.

Constructing a confidence interval for the overall proportion leaves open the question of heterogeneity between studies. The weighted standard deviation δ, estimated at $\hat{\delta} = 0.06$ indicates how different the underlying p_i from pairs of contributing studies might be. For a formal analysis, Fisher's exact test (see Section 3.1.3) gives a strongly significant result ($p < 10^{-8}$) as might be expected from inspection of Figure 3.6a. It would be reasonable to accompany the overall estimate $\hat{p_F} = 0.044(0.038, 0.05)$ with a statement noting that statistically significant heterogeneity is present. The practical significance of the heterogeneity is context-specific, depending on which sub-studies (or comparisons of sub-studies) – if any – are of particular interest.

Faced with the diversity of fixed-effects estimates, the interpretability of using overall proportion $\hat{p}_F$, the exact coverage property and the compatibility with analyses of heterogeneity make it a compelling choice for what one might report.

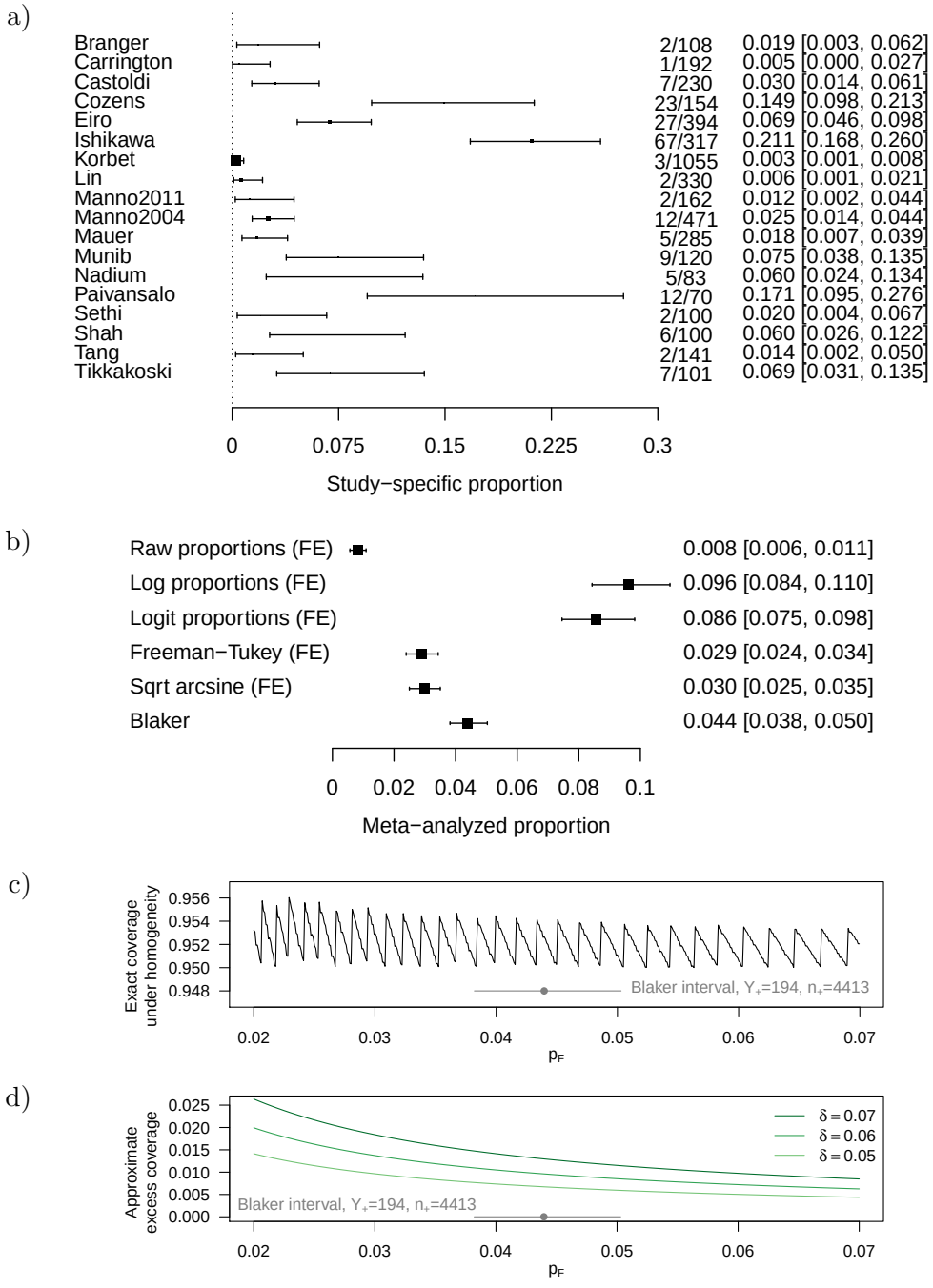


Figure 3.6: Results from the KPMP pain example. a) Study-specific results, given as number of events/number of biopsies and proportion, each with exact 95% Blaker intervals: point sizes are proportional to n_i. The right columns give the observed proportion and its interval. b) Comparison of Blaker interval for meta-analysis with various standard fixed-effects analysis, all of which use different transformations of the observed proportions then re-transform to the original scale. c) Exact coverage of the Blaker interval under homogeneity for a range of plausible p_F d) Approximate excess coverage of the Blaker intervals for plausible values of δ, using Wald test argument

3.4 Discussion

We have presented exact methods for meta-analyzing proportions, showing how exact intervals derived under homogeneity in fact provide exact coverage under general patterns of heterogeneity. This enables us to construct exact confidence intervals for a simple fixed-effects analysis, describing the uncertainty in an unweighted overall proportion. The exact coverage and straightforward interpretation both seem to be of considerable practical appeal in settings where there is currently no default method, and the various options available can easily lead to conflicting results. We anticipate that our methods will be particularly helpful in meta-analysis of rates of rare events, where asymptotic results do not work well but also, with knowledge that all p_i must be small, we can be certain that their weighted standard deviation δ must be small as well.

More generally, our results also provide a small-sample version of the general behavior described by e.g. [54], in which a fixed-effects meta-analysis of association estimates provides large-sample equivalent inference to pooling the data and performing estimates that adjusts for study. For meta-analysis of proportions, the result is somewhat stronger: regardless of heterogeneity of the p_i, the exact interval one obtains from directly pooling all the data is exactly the same as we obtain via meta-analysis, where the meta-analysis focuses on the proportion obtained by amalgamating the various sample populations, proportionally to the sample sizes n_i that were observed.

Our methods focus on overall proportion p_F, the scientific relevance of which merits careful thought—as it would for any overall parameter we estimate, or around which we calculate measures of heterogeneity. As should be transparent from the definition of $\hat{p}_F$ in Equation (3.1), using p_F corresponds to amalgamating the k study populations, in proportion to their observed sample sizes. Considering this overall population is natural — and likely already familiar in practice — as investigators compile the studies they will meta-analyze (see e.g. the methods of [36, Chapters 4–6]) and assess whether they meet the scope of the review being performed. Should another population be of more relevance, re-weighting could be considered but exact methods for meta-analysis of proportions would remain to be developed. We expect, therefore, that a compromise may be needed between

the relevance of the parameter being estimated, and the accuracy with which we can state its desired statistical properties. Compromising in this way is not new: Bailye [3] describes how, pragmatically, meta-analysts may choose to answer questions about the studies at hand, in lieu of generalizing beyond them. Konstantopoulos and Hedges [49] similarly note how 'conditional' inference — on the studies at hand — is a reasonable choice when data are limited, as occurs in our meta-analysis of rare events.

Due perhaps to confusion between assumptions of fixed-effects (plural) versus a single common effect, there seems to be uncertainty about whether analysis based around any population's overall proportion is compatible with assessments of heterogeneity. As discussed at some length by Rice *et al* [82], estimating overall proportion p_F as we have done does *not* preclude or invalidate analyses of variability of the p_i around p_F. Indeed, as well as answering scientific questions about how the p_i vary, as shown in Section 3.3 the variability can be of statistical interest: by estimating δ we can get some idea of how conservative the coverage of our exact intervals for p_F may be. While not discussed here, focusing on p_F does also not rule out making *a priori* statements of exchangeability, that often underpin random effects analyses [38] and which—through exchangeability of future studies with those at hand—can motivate prediction methods. (Furthermore, as noted by an Associate Editor, the fixed-effects assumptions can be viewed as a submodel of the random effects in which we condition on the observed p_i — though nonlinear transformation biases would affect use of corresponding unconditional estimates.) Despite these connections, particularly with the limited data setting on which we have focused, the assumed form of the required prior distributions can be expected to be a source of sensitivity.

Meta-analysis can be used to summarize other types of data. The approach of this chapter will be extended to meta-analysis of 2×2 tables in the next chapter.

In conclusion, while alternative methods are available, we believe that the combination of very mild assumptions, straightforward interpretability and guaranteed statistical properties make our exact methods appealing choices when meta-analyzing sample proportions that are small.

This work has been published in *Research Synthesis Methods*. Code and data to reproduce all the examples is available at

http://faculty.washington.edu/kenrice/kpmpmeta.R.

Chapter 4

EXACT INFERENCE FOR FIXED-EFFECTS META-ANALYSIS OF 2×2 TABLES

4.1 Introduction

Meta-analysis of 2×2 contingency tables has a long history, going back at least to Mantel and Haenszel [62]. A typical aim is estimation of some overall odds ratio, describing association between the row and column variables. The literature on this area is extensive and we shall not attempt to review all of it; see Agresti's textbook [2] and the Cochrane Handbook [37, §10.4] for accessible and comprehensive summaries of prior work.

For practical meta-analysis of 2×2 tables, two important factors are i) whether to provide *exact* inference, meaning that tests control Type I error rates at levels strictly at or below the nominal α and/or confidence intervals that cover the truth in no less that $1-\alpha \times 100\%$ of replicate studies, versus approximate methods that only achieve this in large samples, and ii) whether to assume a *constant effect* across all tables, also known as *homogeneity* of the odds ratios, or instead assume *fixed effects* within each table and its population, of which we estimate some form of average. While these effects may differ, i.e. there may be *heterogeneous* odds ratios, the fixed effects approach still has a useful interpretation and can provide a useful reference point for further analysis of the heterogeneity [83].

Under homogeneity, exact inference for the overall odds ratio is available [16]. However, exact inference under heterogeneity is not. The aim of this paper is to provide an exact method for fixed effects meta-analysis of 2×2 tables, by extending an approach that enables exact meta-analysis of proportions (see previous chapter) [34]. This approach, while conservative, provides exact inference and exact confidence intervals around the conditional maximum likelihood estimate of the odds ratio.

In Section 4.1.1, we discuss the existing approaches for meta-analysis of 2×2 tables. Section 4.2 shares important results for our method; that the hypergeometric distribution

can be written as the convolution of heterogeneous Bernoulli trials, and how to leverage this to provide exact inference for a convolution of heterogeneous hypergeometric disributions. Section 4.3 evaluates our approach and Section 4.4 applies it to the controversial Avandia meta-analysis considered by Nissen and Wolski [70] and also Tian *et al* [96]. We conclude with a discussion.

4.1.1 Existing Approaches

Exploiting large sample approximations, inference is straightforward under both the constant effect and fixed effects assumptions. When each subtable is large, the log odds ratio and its variance are easily estimated in each subtable and can be used in standard inverse-variance weighted meta-analysis methods. Under homogeneity, this approach estimates the common log odds ratio as efficiently as a logistic regression analysis of individual-level data that adjusts for study [53] – which is often a feasible alternative, albeit not a widely-used one. Under heterogeneity, both the meta-analysis and the direct logistic regression instead estimate the same weighted average effect. For inference on that average effect, the large-sample standard error calculations require slight modification: for the meta-analysis corrected large-sample standard error estimates are given by Li and Rice [52]. Direct logistic regression analysis can instead use 'robust' standard error estimates, that are valid in large samples even when the mean model is mis-specified [18].

However, when the subtables are small but the number of them is large – i.e. for *highly stratified* data – the number of nuisance parameters grows proportionally to the dataset, and standard large-sample approximations can fail. This failure can be non-trivial; for example with *pair-matched* designs, where each table contains outcomes for only two observations, under homogeneity the direct logistic regression approach leads to an estimate that is consistent for the square of the true odds ratio – so can be grossly biased away from the null [84]. This problem is typically addressed using one of two approaches, the Mantel-Haenszel (MH) estimate and the conditional likelihood.

The MH estimate [62] gives a form of precision-weighted average of the subtable-specific odds ratio estimates. A widely-used large-sample estimate of the variance of the logarithm

of the MH estimate was given by Robins et al [85]; it is valid with highly stratified data and, conveniently, also with a fixed number of subtables that are all large. As noted by Noma and Nagashima [71], standard sandwich-type large sample approximations are valid under homogeneity and heterogeneity. Noma and Nagashima [71] also note that ordinary nonparametric bootstrapping provides another valid large-sample approximation.

The second approach, the use of the conditional likelihood, conditions on the row and column totals in each 2×2 table, which are approximately ancillary for the odds ratio. The resulting conditional likelihood is free of subtable-specific nuisance parameters. Under homogeneity, standard likelihood theory provides large-sample approximations that can be used for approximate tests and confidence intervals; with some recent advances in numerical manipulation of this conditional likelihood [7] none of these is particularly challenging to implement in practice. Under homogeneity, exact inference is also available. By using the same approach as Fisher's exact test, evaluating the probability of datasets with the same marginal totals but equally or more extreme results under a specified common odds ratio, we obtain exact tests [16], and these can be inverted to give exact confidence intervals.

Other exact methods are available under homogeneity that provide inference on the same parameter estimated across multiple studies. Tian *et al* [96] achieves this using only a user-specified weighted combinations of exact confidence intervals. Similar approaches use weighted combinations of p-values [58] or more generally confidence distributions [106] for the same parameter, estimated across multiple studies. Liu *et al* [59] uses a weighted sign-test to provide exact confidence intervals under the common effect assumption, and also for a random-effects assumption. Of interest in this paper is the method from Tian *et al* [96], since it examined the Avandia data we consider in Section 4.4. Their methods rely on selecting tuning parameters, which can present as a limitation when one has no knowledge of what to select as a tuning parameter. Tuning parameters in their method include weights given to each study, and an option to select multiple confidence interval levels (also choosing corresponding weights) for each study.

Under heterogeneity, exact inference is not currently available, except in the following limited sense. In hypothesis tests of the *strong null hypothesis*, the null states that all subtable-specific odds ratios have a single unique null value, and hence that any overall

odds ratio (e.g. a weighted average) must also have that value. Testing the strong null, it would be appropriate to reject the null under any form of heterogeneity, and hence under forms where the overall odds ratios has some non-null value. This means that exact tests of the strong null are not invalidated by heterogeneity *under the alternative*. However, for testing the weak null hypothesis or constructing confidence intervals for the overall odds ratio, currently there are no exact methods available.

4.2 Analytic results

4.2.1 Known connections between hypergeometric and Poisson-Binomial distributions

We now introduce an important result that is key to our method, following the approach of Barrett [7], that in turn draws on results from Kou and Ying [50]. We define the entries in each 2×2 table as

		Outcome		
		1	0	Total
Group	1	X_k	$M_k - X_k$	M_k
Group	0	$T_k - X_k$	$N_k - T_k + X_k$	N_k
Total		T_k	$M_k + N_k - T_k$	$M_k + N_k$

,

for $k = 1, ..., K$. It is typical to assume that the count of outcomes in the two groups is binomially distributed, with row totals M_k and N_k in each subtable being fixed either by design or by invoking the conditionality principle [86]. Further conditioning on the column totals, i.e. on the sum of the outcomes over both groups, we obtain a hypergeometric distribution for X_k, with

$$\mathbb{P}[\, X_k = x | M_k = m, N_k = n, T_k = t \,] = \frac{\binom{m}{x}\binom{n}{t-x}\psi_k^x}{\sum_{\max(0,t-n)\leq x' \leq \min(t,m).} \binom{m}{x'}\binom{n}{t-x'}\psi_k^{x'}}$$

where

$$\psi_k = \frac{\mathbb{P}[\,\text{Outcome} = 1 | \text{Group} = 1\,]}{\mathbb{P}[\,\text{Outcome} = 0 | \text{Group} = 1\,]} \frac{\mathbb{P}[\,\text{Outcome} = 0 | \text{Group} = 0\,]}{\mathbb{P}[\,\text{Outcome} = 1 | \text{Group} = 0\,]}$$

is the odds ratio describing association of outcome and group in population k. While somewhat controversial [1] this second conditioning step removes nuisance parameters, and for

example in Fisher's exact test with randomization leads to optimal inference—specifically, uniformly most powerful unbiased tests [97].

Working directly with the hypergeometric distribution can be very challenging. Fortunately, as shown by Kou and Ying [50], the hypergeometric distribution is exactly equal to the convolution of a series of non-identical Bernoulli trials, with success probabilities

$$p_{jk}(\boldsymbol{\psi}) = \frac{1}{1 - \lambda_{jk}/\psi_k}$$

where $\boldsymbol{\psi}$ denotes the k-vector of subtable-specific odds ratios and the λ_{jk}, for $j = 1, 2, \ldots \min(t_k, m_k)$, are the roots of the k'th *hypergeometric polynomial*

$$\phi_k(z) = \sum_{\max(0, t_k - n_k) \leq x' \leq \min(t_k, m_k).} \binom{m_k}{x'} \binom{n_k}{t_k - x'} z^{x'}.$$

We note that all λ_{jk} are non-positive, which leads to valid success probabilities. Furthermore, root-finding for these hypergeometric polynomials is straightforward and stable. Combined with the result on convolution of Bernoulli trials, this means it is practical to evaluate the hypergeometric distribution as a specific *Poisson-binomial distribution* [94, 8]. The same representation also simplifies calculation of the conditional maximum likelihood estimate (cMLE) of the odds ratio [7]; for a single 2×2 table, and single odds ratio ψ_k, the cMLE solves

$$X_k = \sum_j p_{jk}(\psi_k)$$

for ψ_k, i.e. equates the observed outcome to the mean of the relevant distributions. The cMLE can therefore be viewed as a method of moments approach. For a sequence of 2×2 tables under assumed homogeneity – so that vector $\boldsymbol{\psi} = \psi \mathbf{1}_k$, i.e. the k-vector of ones scaled by assumed-common odds ratio ψ, the cMLE similarly solves

$$\sum_k X_k = \sum_k \sum_j p_{jk}(\psi \mathbf{1}_k) \tag{4.1}$$

for ψ. While further analytic steps can be made to narrow down the solution, (exploiting e.g. existence of a single root, and bounds on its location) even without them this one-dimensional problem is tractable with standard root-finding algorithms.

Finally, a further benefit of viewing the hypergeometric distribution as a Poisson-binomial draws on the work of Hoeffding [40], in which it is established that, among all Poisson-binomial distributions with a given mean, the binomial distribution has heaviest tails. Consequently, inference that is exact under binomial assumptions (in which all p_{jk} are identical) will also be exact under heterogeneity of the p_{jk}, albeit mildly conservative. The degree of conservatism can be usefully approximated by considering the variance of the p_{jk} [34].

4.2.2 Novel Approaches

We consider use of $X^+ = \sum_{k=1}^{K} X_k$ as the test statistic for assessing the weak null hypothesis, that an overall odds ratio (to be defined below) has specific value ψ_0. Conditioning the row and column totals, the distribution of X^+ is a convolution of multiple Bernoulli trials, the set of success probabilities across them all being

$$\bigcup_{k=1}^{K} \{p_j(\psi_k; m_k, n_k, t_k), j = 1, 2, ... \min(t_k, m_k)\}.$$

To use this representation together with the result from Hoefding [40], we need to specify a mean value for X^+ under the null. To do this, we draw on the moment-matching argument that defines the cMLE in Equation (4.1) above. We say the weak null holds, with specified value ψ_0, if and only if

$$\sum_{jk} p_{jk}(\boldsymbol{\psi}) = \sum_{jk} p_{jk}(\psi_0 \mathbf{1}_k),$$

or in other words that the conditional mean of X^+ under heterogeneity is the same as would occur under homogeneity. This definition can also be reversed, saying that the overall odds ratio summarizing the $\boldsymbol{\psi}$ vector is the one for which

$$\sum_{jk} p_{jk}(\boldsymbol{\psi}) = \sum_{jk} p_{jk}(\psi \mathbf{1}_k).$$

A consequence of this definition is that the estimating equation for the cMLE, i.e.

$$X^+ = \sum_{jk} p_{jk}(\psi \mathbf{1}_k)$$

is therefore consistent for the overall odds ratio, under the weak null hypothesis.

To provide inference on the overall odds ratio, we apply the result (as we did in the previous chapter) of Hoefding's paper [40], which proves that the approximation whereby, with $U_k = min(t_k, m_k)$,

$$X^+ \sim \text{Bin}(\sum_k U_k, \sum_{k,j} p_{jk}(\psi_0) / \sum_k U_k)$$

is at worst conservative, i.e. has heavier tails than the true distribution of X^+. The binomial distribution in question is the one where the index is the maximum value of X^+ conditional on the subtable-specific margins, with mean equal to the conditional mean of the hypergeometric convolution when all $\psi_k = \psi_0$. Our approach is to reject the null $\psi_F = \psi_0$ if X^+ is extreme relative to this Binomial distribution. By Hoeffding's result, this exact test is then conservative (but also exact) for Poisson-Binomial distributions with the same index and mean, i.e. under any formulation of the ψ_k that has the property that $\sum_{k,j} p_{jk}(\psi_0) = \sum_{k,j} p_{jk}(\psi_k)$. We obtain exact $(1-\alpha)$ confidence intervals by inverting the exact level α tests, i.e. finding the set of null values ψ_0 that would not be rejected by the level α exact test.

Recall in the previous chapter (and its appendix) that we discussed the different approaches to exact tests for the Binomial distribution. Blaker's [17] method for exact confidence intervals combines the probability of the smaller observed tail with the smallest probability of the opposite tail that does not exceed that observed tail probability. The method inverts the test to obtain confidence intervals. While the method is valid, it has been noted that the corresponding Blaker p-values can have counter-intuitive properties, not being bimonotonic with respect to the parameter value [101]. Therefore, if the focus of one's analysis is testing, we would instead recommend the Clopper-Pearson method [24], which is widely available in R and also relies on inverting the test to obtain confidence intervals. This method is also exact, and does not have the unintuitive properties above. However, Clopper-Pearson intervals are slightly more conservative than those of Blaker's method.

Given that the degree of conservatism of the approximation of the Poisson-Binomial by the Binomial depends largely on the variance of the $\{p_{jk}\}$, then the Binomial approximation will work well when all the p_{jk} are very similar. If the $p_{jk}(\psi_0)$ are very different, then it

will be conservative under homogeneity, but under heterogeneity the procedure will be even more conservative as there will be heterogeneity in the $p_{jk}(\psi_k)$. However, in the setting where this method is likely more needed (e.g. with rare events) the conservatism may not be a practical impediment. We illustrate this in Section 4.4's Avandia example.

4.3 Evaluation

To evaluate the method of Section 4.2.2, we consider complete enumeration for a set number of 2×2 tables. Following the approach of the previous chapter [34], we evaluate coverage when the strata-specific odds ratios are homogeneous, i.e. all ψ_k are equal to some ψ_F. This indicates the magnitude of the inevitable conservatism of using exact methods ([1], [46]), before evaluating how much more excess coverage occurs under heterogeneity of odds ratios.

We measure heterogeneity by examining the sum of the squared difference between study effects and the underlying parameter. Specifically we define

$$\delta^2 = \sum_{k,j} \frac{1}{\sum_k U_k} \left(p_{jk}(\psi_k) - \frac{\sum_{k,j} p_{jk}(\psi_F)}{\sum_k U_k} \right)^2,$$

a measure of the excess variability in the convolution of Bernoullis, compared to a Binomial distribution with the same mean. Defining, $p_F = \frac{\sum_{k,j} p_{jk}(\psi_F)}{\sum_k U_k}$, the impact of using tests that are exact under homogeneous p_{jk} (when heterogeneity holds instead) is approximately the same as using a Wald test based on $\hat{p}_F$ in which the standard errors are too large by a factor we introduced in the previous chapter:

$$\xi(p_F, \delta) = \sqrt{1/\left(1 - \frac{\delta^2}{p_F(1-p_F)}\right)}.$$

Which yields confidence intervals with approximate level of $1-2\Phi\left(\Phi^{-1}(1-\alpha/2) \times \xi(p_F, \delta)\right)$.

We first consider a situation with two strata, with $K = 2$, and row margins $M_k = (500, 500)$ and $N_k = (500, 500)$. We consider a balanced setting where the column totals are $T_k = (25, 25)$, and an unbalanced setting with $T_k = c(15, 35)$. (Appendix C.1 gives results with $K = 2$ but varying row totals M_k and N_k.) In the enumerations, we consider a grid of values of the two strata log odds ratios. We calculate ψ_F using the relation in $\sum_{k,j} p_{jk}(\psi_F) = \sum_{k,j} p_{jk}(\psi_k)$. From there, we can calculate δ by calculating p_F using the

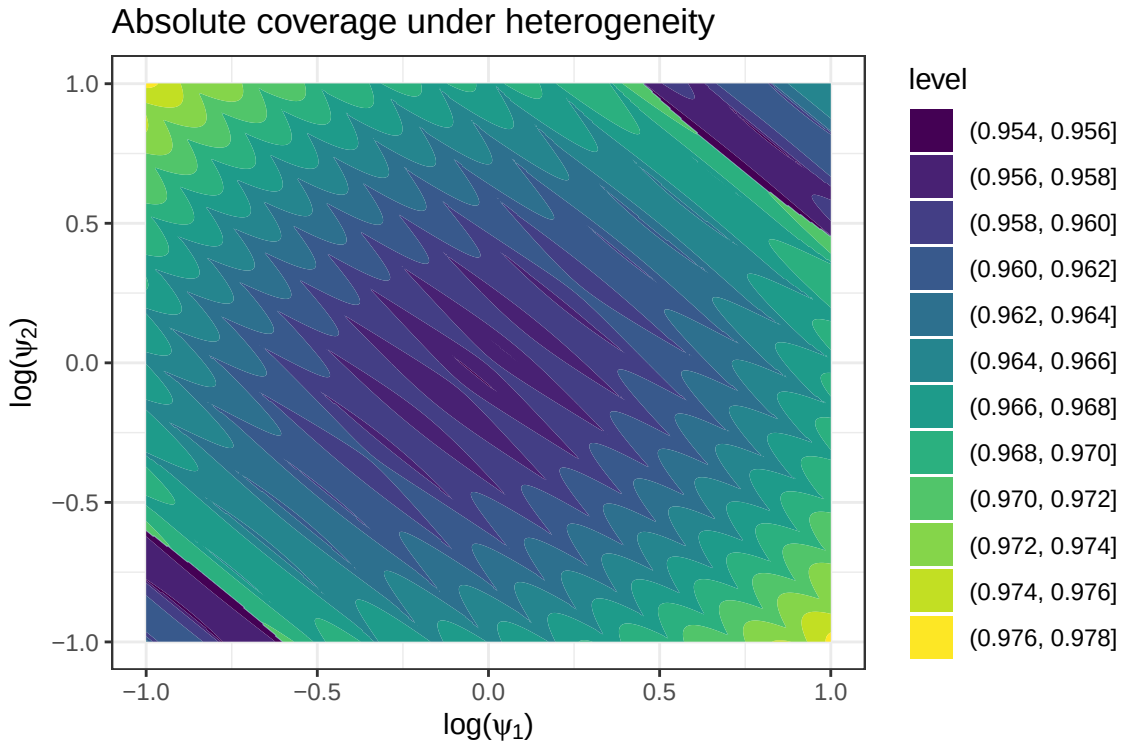


Figure 4.1: Absolute coverage under heterogeneity of the 95% Blaker intervals using our method for a grid of values with table margins $M_k = N_k = (500, 500)$, $T_k = (25, 25)$.

relation introduced earlier. Calculating p_F as the mean of $p_{jk}(\psi_k)$ for $k = 1, 2$ makes computation easier.

For the balanced setting, Figure 4.1 displays the absolute coverage of 95% confidence intervals for a grid of $\log(\psi_1)$ and $\log(\psi_2)$ values, confirming that we have at least 95% coverage and that our method is exact. Figure 4.2 compares the excess coverage for the same grid of values to our measure of heterogeneity, δ, and we note the similarity; δ is in large part determining the conservatism of the exact interval. Lastly, Figure 4.3 provides, for $\log(\psi_1) < \log(\psi_2)$, the coverage under homogeneity (top panel) and, for some fixed values of δ, the excess coverage under heterogeneity (bottom panel). It also shows the high accuracy of the Wald-test based approximation of excess coverage determined by δ—again stressing that δ determines the conservatism, essentially alone.

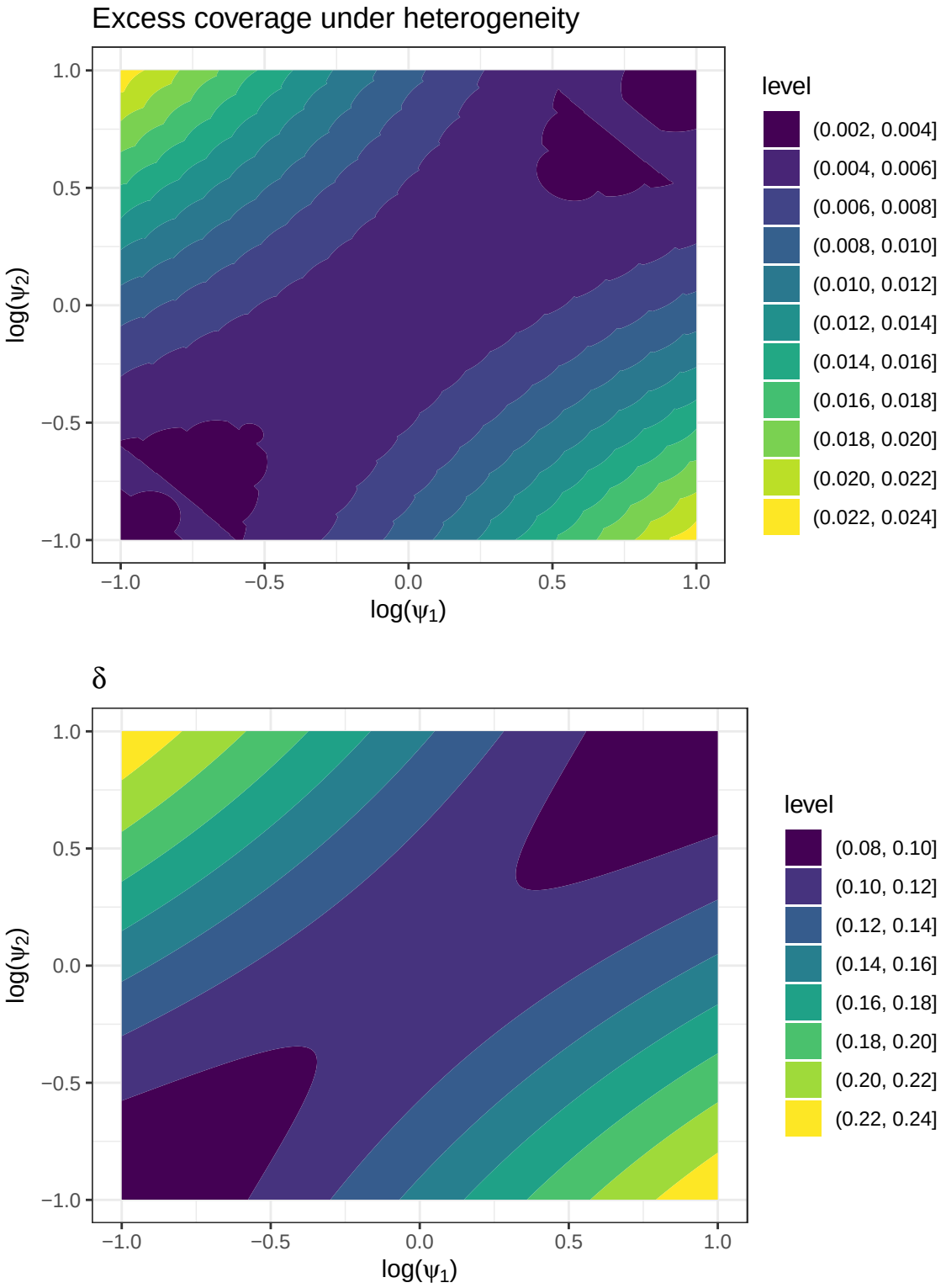


Figure 4.2: Top panel: Excess coverage under heterogeneity of the 95% Blaker intervals using our method for a grid of values with table margins of $M_k = N_k = (500, 500)$, $T_k = (25, 25)$. Bottom panel: Measure of heterogeneity, δ, for a grid of values with same table margins.

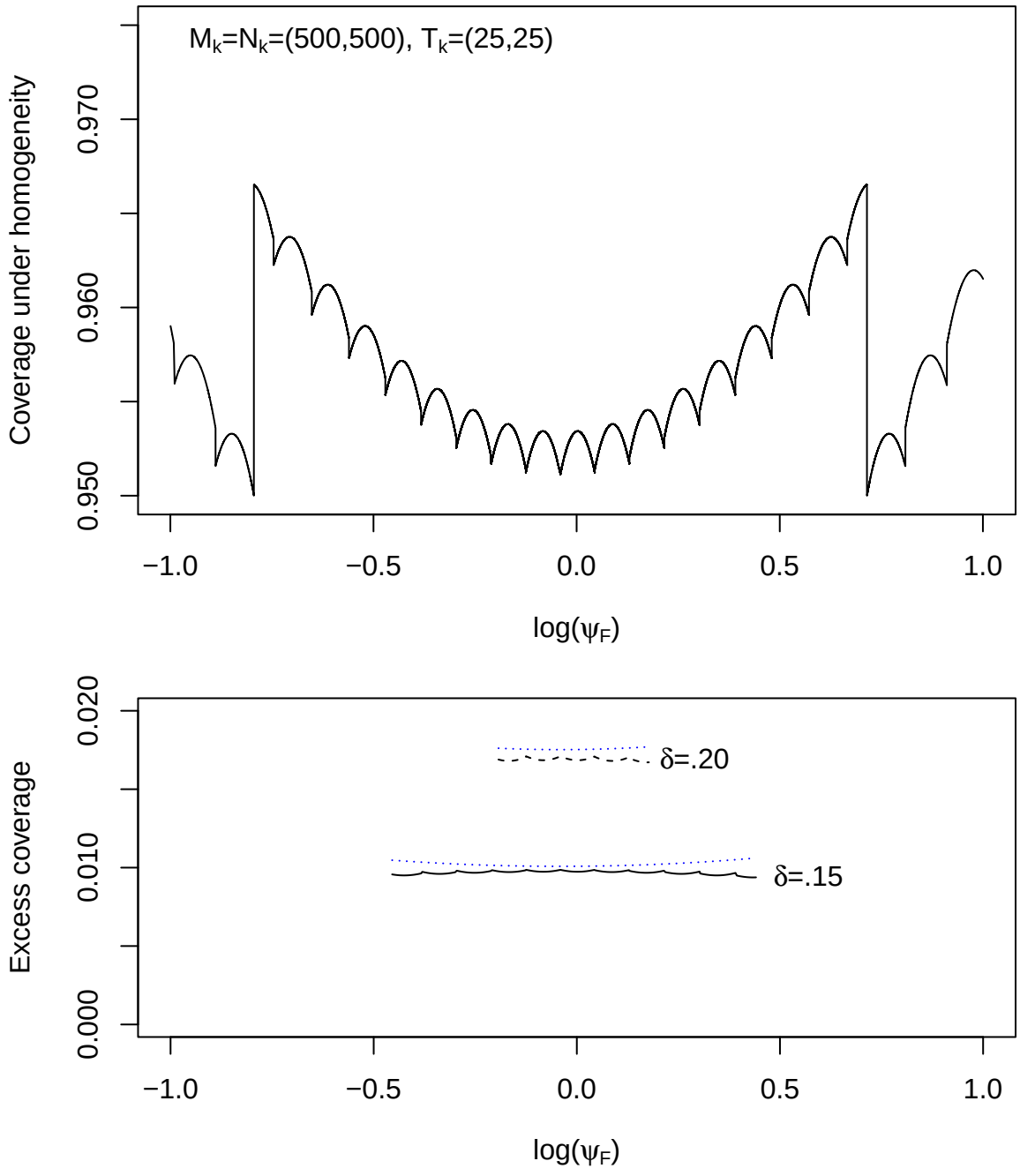


Figure 4.3: Top panel: Coverage of the 95% Blaker intervals using our method for varying values of the ψ_F under homogeneity and for $\psi_1 < \psi_2$, for table margins of $M_k = N_k = (500, 500)$, $T_k = (25, 25)$. Bottom panel: Excess coverage of the same intervals under heterogeneity for certain values of δ (blue, small dotted lines represent Wald-test based approximations).

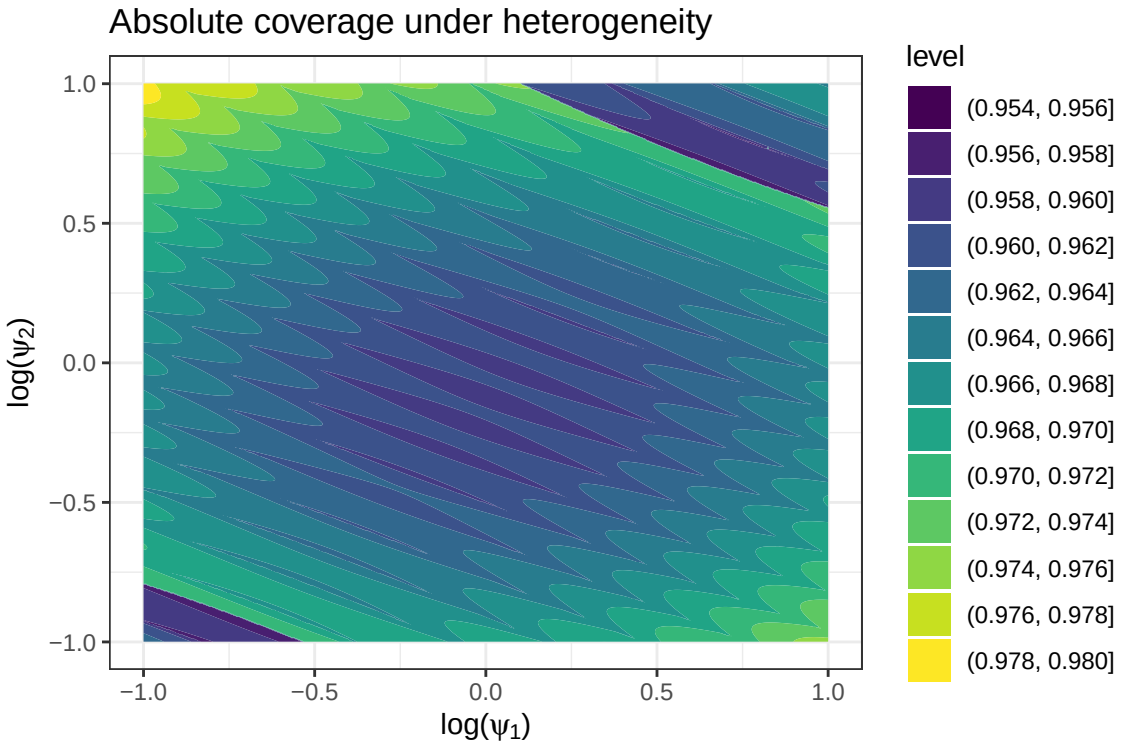


Figure 4.4: Absolute coverage under heterogeneity of the 95% Blaker intervals using our method for a grid of values with table margins $M_k = N_k = (500, 500)$, $T_k = (15, 35)$.

Figures 4.4, 4.5, and 4.6 consider the situation where $T_k = (15, 35)$. Again, we confirm that the method is exact, that our measure of heterogeneity is useful for estimating excess coverage, and that the excess coverage even for moderate heterogeneity is not too extreme.

For the five strata setting setting, we consider row totals of $M_k = N_k = (200, 200, 200, 200, 200)$ with column totals $T_k = (5, 10, 20, 10, 5)$. We vary the stratum odds ratios by having them uniformly spread on the log scale, around $\log(\psi_3) = 0$ and with the spread (i.e. $\log(\psi_5) - \log(\psi_1)$) chosen to vary the value of δ. In this setting, we calculate p_F by the mean of all $p_{jk}(\psi_k)$, $k = 1, ..., 5$. Using complete enumeration we calculate our method's excess coverage at the nominal 95% level, shown in Figure 4.7. The positive slope between δ and the excess coverage gives an indications that δ can be used to accurately capture the excess heterogeneity. The dashed blue line represents Wald-test based approximations, and

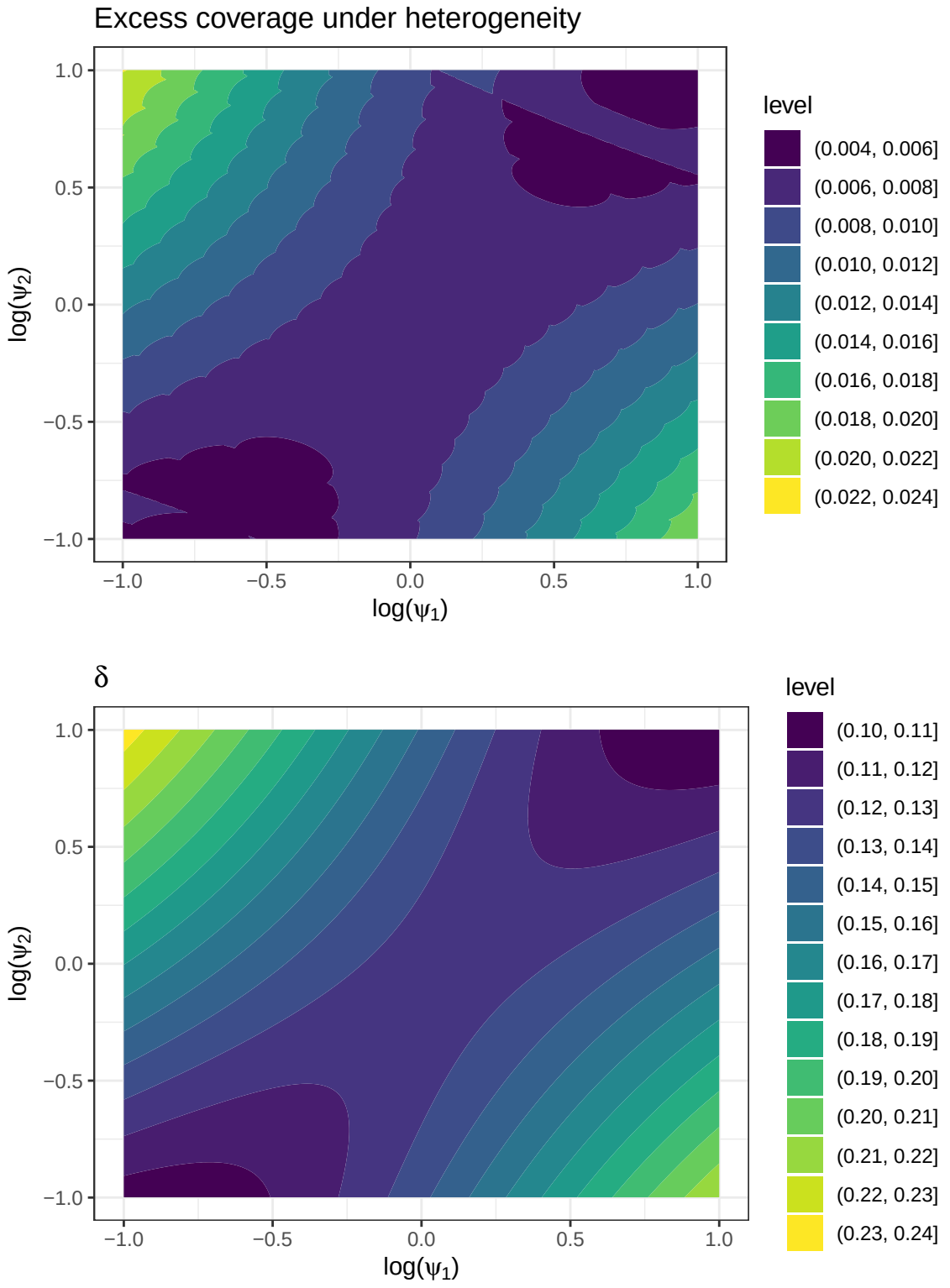


Figure 4.5: Top panel: Excess coverage under heterogeneity of the 95% Blaker intervals using our method for a grid of values with table margins of $M_k = N_k = (500, 500)$, $T_k = (15, 35)$. Bottom panel: Measure of heterogeneity, δ, for a grid of values with same table margins.

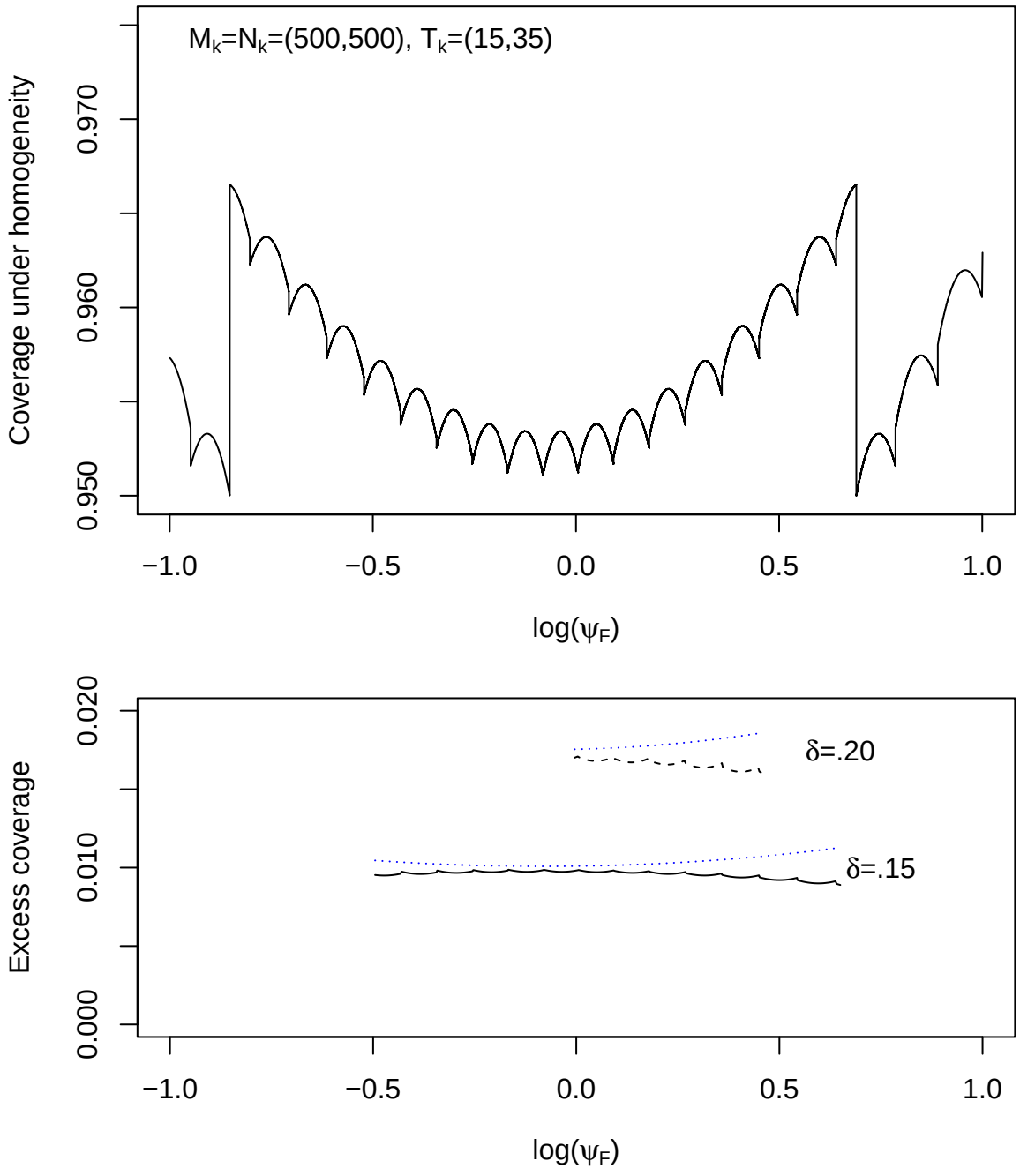


Figure 4.6: Top panel: Coverage of the 95% Blaker intervals using our method for varying values of the ψ_F under homogeneity and for $\psi_1 < \psi_2$, for table margins of $M_k = N_k = (500, 500)$, $T_k = (15, 35)$. Bottom panel: Excess coverage of the same intervals under heterogeneity for certain values of δ (blue, small dotted lines represent Wald-test based approximations).

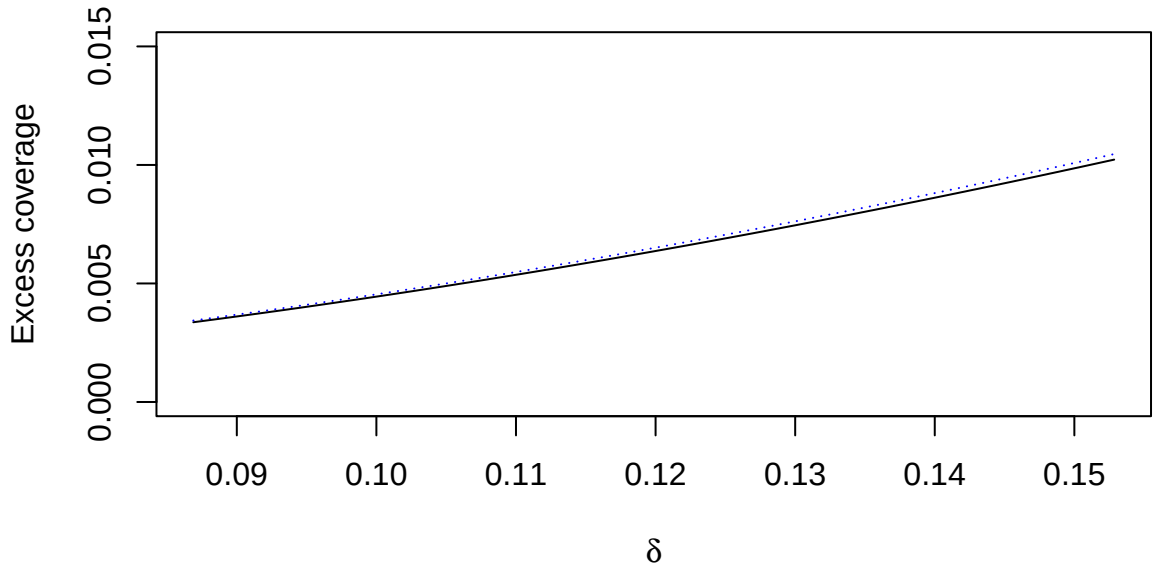


Figure 4.7: Excess coverage of our method for table margins of $M_k = N_k = (200, 200, 200, 200, 200)$, $T_k = (5, 10, 20, 10, 5)$ as a function of δ (blue, small dotted lines represent Wald-test based approximations).

is very similar, albeit slightly larger, to the excess coverage we calculated.

4.4 Application

Nissen and Wolski [70] conducted a meta-analysis of 48 trials comparing rosiglitazone, a drug used in treating Type 2 diabetes mellitus, to a control—an amalgamation of different common drugs and the placebo. The goal was to determine if the drug had an effect on the risk of myocardial infarction (MI) and cardiovascular (CVD) related death. Using standard large sample approximate fixed effects methods (Peto's method), a statistically significant increase in MI on rosiglitazone was observed, a result which eventually led to the drug being removed from use. Subsequently, Tian *et al* [96] re-analyzed the data using their method to provide exact inference under an assumed common effect for the risk difference (they did not consider odds ratios) and in their analysis the result was not significant. However, the estimated effects were in the direction of the drug having higher risk, for both endpoints. In our implementation of this method, we measure the effect using the log odds ratio, and

selecting a grid of 19 η values (confidence interval levels) that range from 0.05 to 0.95 with weights proportional to $(\eta \times (1-\eta))^{-1}$. Furthermore, study weights are equal to sample size and similarly to Tian *et al*'s analysis of risk differences, we used the mid-p value method to construct the individual study confidence intervals [39].

We compare our analysis (cMLE with Blaker) of the log odds ratio in the Avandia data for both endpoints to the Tian *et al* method, and to the conditional MLE with confidence intervals calculated using the Fisher information approach (denoted as 'plain cMLE'), the Mantel-Haenszel method for the common odds ratio ([62]), the Peto method ([107]), and the higher-order asymptotic methods of Li and Rice [52].

Continuity correction is often used when meta-analyzing odds ratios. For the "plain" Mantel-Haenszel, Woolf, and Peto methods, we added the default 0.5 to all cells of the tables in which a zero appeared. In these methods, and also the higher order methods from Li and Rice's method which have their own continuity control, we excluded the double zero entries – studies where the treatment and control are both zero. In our method, the double zero studies contribute no information since there is only one way in which the convolution can occur (i.e. the sub-table for a double zero study can only happen in one way, and it occurs with 100% probability).

Our novel Poisson-binomial approach requires no continuity correction for inference on ψ_F and so none was used. We note, as we did in the previous paragraph, that double zero entries contribute no information to the conditional distribution, so are omitted for calculation purposes. We did not continuity-correct the method of Tian *et al*, as one can still calculate the required one-sided confidence intervals when zeroes are present. Figure 4.8 displays the results.

To approximate the unknown δ – used to ascertain how conservative the exact inference might be, a form of diagnostic check – in the heterogeneous case standard cross-ratio estimates $\frac{X_k(N_k-T_k+X_k)}{(M_k-X_k)(T_k-X_k)}$ were used for each ψ_k. Here we did correct zero cell entries by adding 0.5, to avoid implausible unbounded log odds ratios. Under homogeneity we use the overall cMLE as the common estimate for all ψ_k.

For the MI endpoint, the results are in broad agreement, with the refinements due to Li and Rice resulting in slightly wider intervals. This slight difference is practically

important here, as it nudges the results into non-significance at the 0.05 level. The Tian *et al* result has a smaller interval. The extent to which the large sample methods (or their refinements) are inaccurate is unknown, which makes the significance of their results difficult to assess. Tian *et al*'s method is exact, but depends on multiple user-chosen tuning parameters. Furthermore, it is difficult to justify the use of certain tuning parameters over others. Together with its common effect assumptions – not needed for the refined large sample methods – Tian *et al*'s interval is also difficult to interpret. The exact interval around the cMLE avoids these difficulties: no homogeneity assumption is required, and the significance of the result at the specified level is not in question, as the inference is conservative at worst. Nevertheless, some information on the degree of conservatism is available; a simple plug-in estimate for δ gives $\hat{\delta} = 0.177$ for this example, representing moderate heterogeneity across studies. This is approximately the same as a Wald test based on $\hat{p_F}$ with standard errors too large by a factor of $\hat{\xi} = 1.070$. Thus, the approximate level is 96.4% instead of 95%. Some of this conservatism appears due to heterogeneity, but the non-binomial nature of the test statistic even under homogeneity plays a role. Specifically, plugging in the cMLE for all ψ_k, the variability in the $p_{jk}(\hat{\psi})$ instead gives $\hat{\delta} = 0.138$, with standard errors too-wide by $\hat{\xi} = 1.041$ and a coverage of 95.9%. With the caveat that the degree of conservatism cannot be known exactly, based on these measures neither source of conservatism seems of major concern.

For the mortality endpoint, with fewer events, all intervals are wider than for MI. Also, the effect sizes vary more, possibly due to a lack of data. Again, with the selection of tuning parameters, Tian *et al*'s approach is difficult to assess. We see the extra conservatism of the approach we have introduced in this paper. The evaluation of heterogeneity is very similar to that for MI; from the data we get $\hat{\delta} = 0.175$, compared to $\hat{\delta} = 0.107$ under homogeneous ψ_i. The approximate coverage under heterogeneity of study specific odds ratios is 95.5% ($\hat{\xi} = 1.073$), while under homogeneity it is 95.6% ($\hat{\xi} = 1.026$), against suggesting conservatism that is not a major concern.

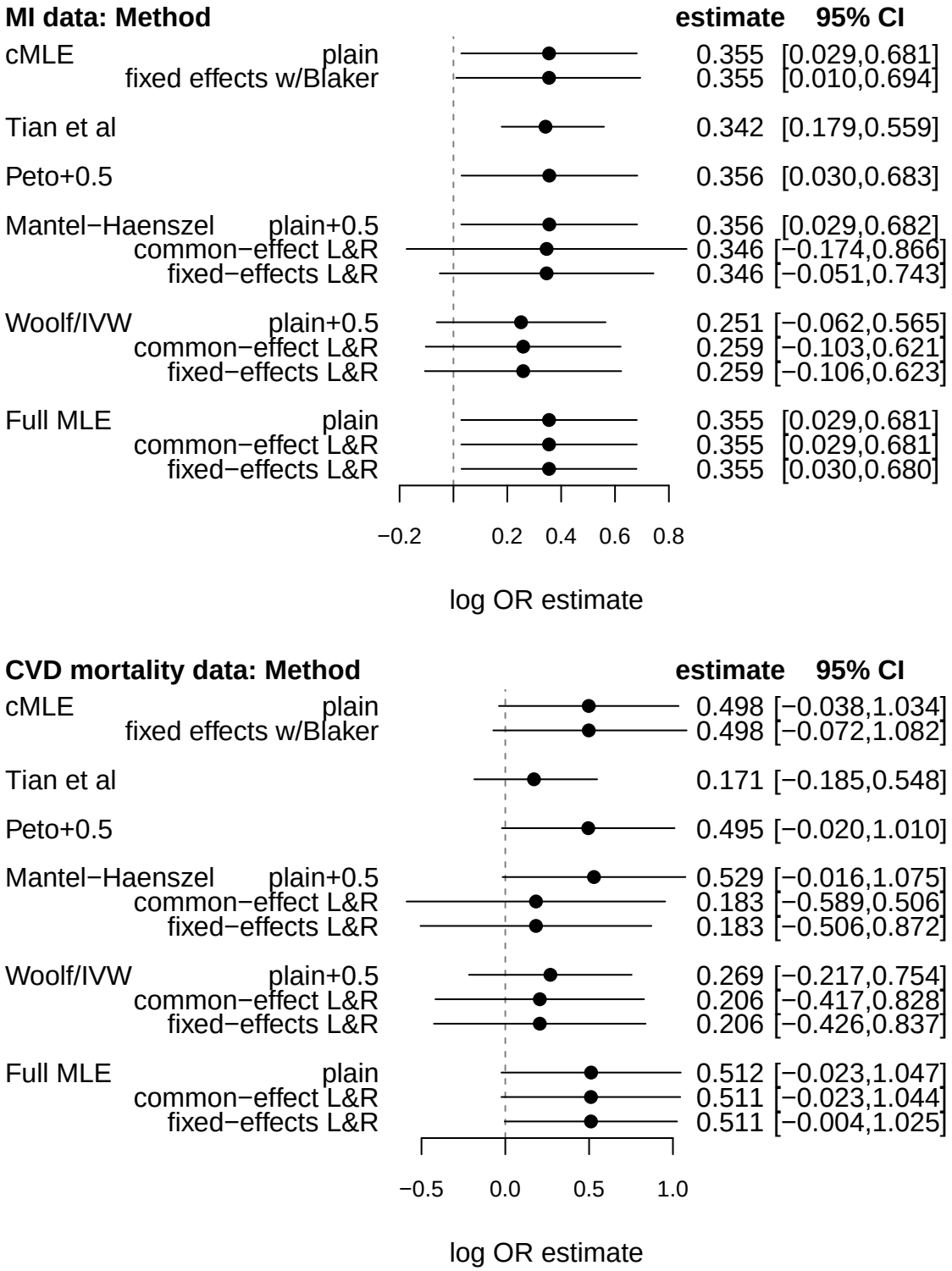


Figure 4.8: Top panel: Approximate and exact 95% confidence intervals for overall log odds ratio using the Avandia data on MI (top panel) and cardiovascular-related death (bottom panel). Standard large sample methods derived under common effect assumptions (cMLE, Peto, Woolf, Mantel-Haenszel) are shown beside (where available) large-sample versions that account for heterogeneity. The exact method assume either homogeneity (Tian *et al*) or heterogeneity (cMLE with Blaker). For details please see the main text.

4.5 Discussion

We have presented novel exact methods for meta-analysis of 2×2 contingency tables, giving inference on an overall odds ratio – the quantity estimated by the widely-studied conditional MLE. Our method would be appropriate for any application where events are rare and heterogeneity is expected (or at least plausible) – and where inference centers on an overall comparison of an outcome between two groups, across the studies, or more generally across strata. The exact inference provided, albeit conservative under some settings, can provide a useful default, producing inference around a well-studied estimate – the cMLE – without concerns about accuracy of large-sample approximations, or typically-implausible assumptions of homogeneity. Furthermore, we have provided a metric, δ^2, to assess the degree of conservatism.

As seen in the Avandia example, our novel approach appears conservative, but not at a level that is likely to outweigh the reassurance provide by having access to exact inference. However, there could be cases when the conservatism calculated from our metrics may be unacceptably high. In these instances, one might obtain intervals nearer the 95% level – though sacrificing their 'exact' property – by using the estimated value of δ to indicate what nominal level of coverage would provide actual coverage under homogeneity of approximately 95% (as indicated using by the Wald test-like argument of Section 4.3) and then using intervals with this nominal level. This would not, however eliminate conservatism due to heterogeneity. With much more involved calculation, one might also re-calibrate the level of coverage by considering the infimum coverage (under homogeneity) with respect to ψ_F, and adjusting the nominal level so that this infimum is at the desired level. (For similar calculations in this spirit, see e.g. [63].) With few tables this might be achieved by complete enumeration, but would otherwise require Monte Carlo approximation. The approaches discussed here will be left to future work.

In conclusion, while alternative methods are available, we believe that the combination of very mild assumptions, straightforward interpretability and guaranteed statistical properties make our exact method an appealing choice. In the Avandia example, while no single analysis can be considered uniquely "right", our exact approach provides considerable re-

assurance that the originally-published results – relying on large-sample properties when their accuracy is far from obvious – were not substantially misleading.

Code and data to reproduce all the examples is available at `https://github.com/slh789/exactmeta2x2`.

Chapter 5

CONCLUSION

In this dissertation, we added to the vast area of inference. First, we provided a Bayesian framework that yields coherent decisions for nested interval null hypotheses. This framework is connected to a frequentist type control that differs from the more common type of control. Then, we provided exact inference under the more reasonable fixed-effects paradigm in meta-analysis. This method of inference can be used when either the effect being meta-analyzed is proportions or odds ratios from multiple 2×2 tables.

We hope that the framework in the second chapter provides flexibility in making multiple decisions. Additionally we hope the methods in the third and fourth chapter appeal to those who want a straightforward estimate that requires few assumptions.

BIBLIOGRAPHY

[1] Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.

[2] Alan Agresti. *Categorical Data Analysis*, volume 482. John Wiley & Sons, Hoboken, NJ, 2003.

[3] Kent R Bailey. Inter-study differences: How should they influence the interpretation and analysis of results? *Statistics in Medicine*, 6(3):351–358, 1987.

[4] Ilyas Bakbergenuly and Elena Kulinskaya. Beta-binomial model for meta-analysis of odds ratios. *Statistics in Medicine*, 36(11):1715–1734, 2017.

[5] Jan J Barendregt, Suhail A Doi, Yong Yi Lee, Rosana E Norman, and Theo Vos. Meta-analysis of prevalence. *J Epidemiol Community Health*, 67(11):974–978, 2013.

[6] Vic Barnett. *Comparative Statistical Inference*, volume 522. John Wiley & Sons, Chichester, United Kingdom, 1999.

[7] Bruce Barrett. Simple, exact inference for 2×2 tables. *Communications in Statistics–Theory and Methods*, 46(1):221–233, 2017.

[8] Bruce E Barrett and J Brian Gray. Efficient computation for the Poisson binomial distribution. *Computational Statistics*, 29(6):1469–1479, 2014.

[9] M Jésus Bayarri and James O Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

[10] Yoav Benjamini, Richard De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry I Graubard, Xuming He, Xiao-Li Meng, Nancy Reid, Stephen M Stigler, Stephen B Vardeman, Christopher K Wikle, Tommy Wright, Linda J Young, and Karen Kafader. ASA President's task force statement on statistical significance and replicability. *Annals of Applied Statistics*, 15(3):1084–1085, 2021.

[11] James O Berger. Could fisher, jeffreys and neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.

[12] James O Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, 82(397):112–122, 1987.

[13] David R Bickel. Coherent checking and updating of Bayesian models without specifying the model space: A decision-theoretic semantics for possibility theory. *International Journal of Approximate Reasoning*, 142:81–93, 2022.

[14] David R Bickel and Alexandre G Patriota. Self-consistent confidence sets and tests of composite hypothesis applicable to restricted parameters. *Bernoulli*, 25(1):47–74, 2019.

[15] P Billingsley. *Probability and Measure*. WileyInterscience, New York, 1995.

[16] MW Birch. The detection of partial association, i: the 2×2 case. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):313–324, 1964.

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

[18] Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, Linda Zhao, et al. Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019.

[19] J. S. Butler and Peter Jones. Theoretical and empirical distributions of the p value. *Metron*, 76(1):1–30, April 2018.

[20] George Casella and Roger L Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2002.

[21] George Casella and Roger L Berger. *Statistical Inference (2nd Edition)*. Duxury, Pacific Grove, California, 2002.

[22] Ivan Chan and Zhongxin Zhang. Test-based exact confidence interval for the difference of two binomial proportions. *Biometrics*, 55:1202–1209, 1999.

[23] Douglas B Clarkson, Yuan-An Fan, and Harry Joe. A network algorithm for performing fisher's exact test in r x c contingency tables. *ACM Transactions on Mathematical Software*, 19(4):484–488, 1993.

[24] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[25] Gustavo Miranda da Silva, Luis Gustavo Esteves, Victor Fossaluza, Rafael Izbicki, and Sergio Wechsler. A Bayesian decision-theoretic approach to logically-consistent hypothesis testing. *Entropy*, 17:6534–6559, 2015.

[26] Luis G Esteves, Rafael Izbicki, Julio M Stern, and Rafael B Stern. The logical consistency of simultaneous agnostic hypothesis tests. *Entropy*, 18(256):1–22, 2016.

[27] Michael P. Fay. Two-sided exact tests and matching confidence intervals for discrete data. *R Journal*, 2(1):53–58, 2010.

[28] Victor Fossaluza, Rafael Izbicki, Gustavo Miranda da Silva, and Luís Gustavo Esteves. Coherent hypothesis testing. *The American Statistician*, 71(3):242–248, 2017.

[29] Murray F Freeman and John W Tukey. Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, pages 607–611, 1950.

[30] K R Gabriel. Simultaneous test procedures – some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pages 224–250, 1969.

[31] Mitchell Gail and Nathan Mantel. Counting the number of r x c contingency tables with fixed margins. *Journal of the American Statistical Association*, 72(360):859–862, 1977.

[32] Anjana Grandhi, Wenge Guo, and Joseph P. Romano. Control of directional errors in fixed sequence multiple testing. *Statistica Sinica*, 29(2):1047–1064, 2019.

[33] Jessica Gronsbell, Chuan Hong, Lei Nie, Ying Lu, and Lu Tian. Exact inference for the random-effect model for meta-analyses with rare events. *Statistics in Medicine*, 39(3):252–264, 2020.

[34] Spencer Hansen and Kenneth Rice. Exact inference for fixed-effects meta-analysis of proportions. *Research Synthesis Methods*, 13(2):204–213, 2022.

[35] Jingjing He, Yifei Guo, Richeng Mao, and Jiming Zhang. Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. *Journal of Medical Virology*, 93(2):820–830, 2021.

[36] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.

[37] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane handbook for systematic reviews of interventions*, volume 6.2. Cochrane, Available from www.training.cochrane.org/handbook, 2021.

[38] Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

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

[40] Wassily Hoeffding. On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics*, 27(3):713–721, 1956.

[41] Wassily Hoeffding et al. On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics*, 27(3):713–721, 1956.

[42] Raymond Hubbard and M J Bayarri. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, 57(3):171–178, 2003.

[43] J T Gene Hwang and Ming-Chung Yang. An optimality theory for mid p-values in 2×2 contingency tables. *Statistica Sinica*, 11(3):807–826, 2001.

[44] Jiunn Tzon Hwang, George Casella, Christian Robert, Martin T Wells, and Roger H Farrell. Estimation of accuracy in testing. *The Annals of Statistics*, pages 490–509, 1992.

[45] Rafael Izbicki and Luis Gustavo Esteves. Logical consistency in simultaneous statistical test procedures. *Logic Journal of the IGPL*, 23(5):732–758, 2015.

[46] Donguk Kim and Alan Agresti. Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association*, 90(430):632–639, 1995.

[47] Jae H Kim and Andrew P Robinson. Interval-based hypothesis testing and its applications to economics and finance. *Econometrics*, 7(2):21, 2019.

[48] Bas Kluitenberg, Marienke van Middelkoop, Ron Diercks, and Henk van der Worp. What are the differences in injury proportions between different populations of runners? a systematic review and meta-analysis. *Sports Medicine*, 45(8):1143–1161, 2015.

[49] Spyros Konstantopoulos and Larry V. Hedges. *Analyzing effect sizes: fixed-effects models*, pages 279–294. Russell Sage Foundation, 2009.

[50] SG Kou and Z Ying. Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica*, pages 809–829, 1996.

[51] Chloe Krakauer and Kenneth Rice. Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):452–453, 2021.

[52] Kendrick Qijun Li and Kenneth Rice. Improved inference for fixed-effects meta-analysis of 2×2 tables. *Research Synthesis Methods*, 11(3):387–396, 2020.

[53] Dan-Yu Lin and Daniel Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010.

[54] DY Lin and D Zeng. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(1):60–66, 2010.

[55] Dennis V Lindley. Decision making. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 18(4):313–326, 1968.

[56] Dennis V Lindley. Theory and practice of Bayesian statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):1–11, 1983.

[57] Dennis V Lindley. The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):293–337, 2000.

[58] Dungang Liu, Regina Y Liu, and Min-ge Xie. Exact meta-analysis approach for discrete data and its application to 2× 2 tables with rare events. *Journal of the American Statistical Association*, 109(508):1450–1465, 2014.

[59] Sifan Liu, Lu Tian, Steve Lee, and Min-ge Xie. Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostatistics & Epidemiology*, 2(1):1–22, 2018.

[60] Sifan Liu, Lu Tian, Steve Lee, and Min-ge Xie. Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostatistics & Epidemiology*, 2(1):1–22, 2018.

[61] Yan Ma, Haitao Chu, and Madhu Mazumdar. Meta-analysis of proportions of rare events–a comparison of exact likelihood methods with robust variance estimation. *Communications in Statistics-Simulation and Computation*, 45(8):3036–3052, 2016.

[62] Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

[63] Cyrus R Mehta and Joan F Hilton. Exact power of conditional and unconditional tests: going beyond the 2× 2 contingency table. *The American Statistician*, 47(2):91–98, 1993.

[64] Cyrus R Mehta and Nitin R Patel. A network algorithm for performing fisher's exact test in r x c contingency tables. *Journal of the American Statistical Association*, 78(382):427–434, 1983.

[65] Cyrus R Mehta and Nitin R Patel. A network algorithm for performing fisher's exact test in r x c contingency tables. *ACM Transactions on Mathematical Software*, 12(2):154–161, 1986.

[66] John J Miller. The inverse of the Freeman–Tukey double arcsine transformation. *The American Statistician*, 32(4):138–138, 1978.

[67] Kevin Murphy and Brett Myors. Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, 84(2):234–248, 1999.

[68] Jerry Nedelman and Ted Wallenius. Bernoulli trials, Poisson trials, surprising variances, and Jensen's inequality. *The American Statistician*, 40(4):286–289, 1986.

[69] Jerzy Neyman. *First course in probability and statistics*. Henry Holt and Company, New York, 1950.

[70] Steven E Nissen and Kathy Wolski. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*, 356(24):2457–2471, 2007.

[71] Hisashi Noma and Kengo Nagashima. A note on the Mantel-Haenszel estimators when the common effect assumptions are violated. *Epidemiologic Methods*, 5(1):19–35, 2016.

[72] Giovanni Parmigiani and Lurdes Inoue. *Decision theory: Principles and approaches*. John Wiley & Sons, Chichester, United Kingdom, 2009.

[73] Cynthia G Parshall and Jeffrey D Kromrey. Tests of independence in contingency tables with small samples: a comparison of statistical power. *Educational and Psychological Measurement*, 56(1):26–44, 1996.

[74] W M Patefield. Algorithm as 159: An efficient method of generating r x c tables with given row and column totals. *Applied Statistics*, 30(1):91–97, 1981.

[75] Alexandre G Patriota. A classical measure of evidence for general null hypotheses. *Fuzzy Sets and Systems*, 223:74–78, 2013.

[76] Carlos A De Braganca Pereira and Julio M Stern. Evidence and credibility: Full Bayesian significance test for precise hypothesis. *Entropy*, 1(4):99–110, 1999.

[77] Peter H Peskun. Two-tailed p-values and coherent measures of evidence. *The American Statistician*, 74(1):80–86, 2020.

[78] Emilio D Poggio, Robyn L McClelland, Kristina N Blank, Spencer Hansen, Shweta Bansal, Andrew S Bomback, Pietro A Canetta, Pascale Khairallah, Krzysztof Kiryluk, Stewart H Lecker, et al. Systematic review and meta-analysis of native kidney biopsy complications. *Clinical Journal of the American Society of Nephrology*, 15(11):1595–1602, 2020.

[79] Zad Rafi and Sander Greenland. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC medical research methodology*, 20(1):1–13, 2020.

[80] Kenneth Rice. A decision-theoretic formulation of fisher's approach to testing. *The American Statistician*, 64(4):345–349, 2010.

[81] Kenneth Rice, Tyler Bonnett, and Chloe Krakauer. Knowing the signs: a direct and generalizable motivation of two-sided tests. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):411–430, 2020.

[82] Kenneth Rice, Julian PT Higgins, and Thomas Lumley. A re-evaluation of fixed effect (s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1):205–227, 2018.

[83] Kenneth Rice, Julian PT Higgins, and Thomas Lumley. A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1):205–227, 2018.

[84] Kenneth M Rice. Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association*, 99(466):510–522, 2004.

[85] James Robins, Norman Breslow, and Sander Greenland. Estimators of the mantel-haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, pages 311–323, 1986.

[86] James Robins and Larry Wasserman. Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association*, 95(452):1340–1346, 2000.

[87] Rick Routledge. Fisher's exact test including power and mid-p value. *Wiley StatsRef: Statistics Reference Online*, 2014.

[88] Mark J Schervish. P values: what they are and what they are not. *The American Statistician*, 50(3):203–206, 1996.

[89] M.J. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer New York, 2012.

[90] Guido Schwarzer, Hiam Chemaitelly, Laith J Abu-Raddad, and Gerta Rücker. Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Research Synthesis Methods*, 10(3):476–483, 2019.

[91] David Spiegelhalter and Kenneth Rice. Bayesian statistics. *Scholarpedia*, 4(8):5230, 2009.

[92] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[93] Julio M Stern and Carlos A De Braganca Pereira. Bayesian epistemic values: focus on surprise, measure probability! *Logic Journal of IGPL*, 22(2):236–254, 2014.

[94] Marlin A Thomas and Audrey E Taub. Calculating binomial probabilities when the trial probabilities are unequal. *Journal of Statistical Computation and Simulation*, 14(2):125–131, 1982.

[95] Mans Thulin. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, 8(1):817–840, 2014.

[96] Lu Tian, Tianxi Cai, Marc A Pfeffer, Nikita Piankov, Pierre-Yves Cremieux, and LJ Wei. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2×2 tables with all available data but without artificial continuity correction. *Biostatistics*, 10(2):275–281, 2009.

[97] Keith D Tocher. Extension of the neyman-pearson theory of tests to discontinuous variates. *Biometrika*, 37(1/2):130–144, 1950.

[98] Thomas A Trikalinos, Paul Trow, and Christopher H Schmid. *Simulation-based comparison of methods for meta-analysis of proportions and rates*. Agency for healthcare research and quality (US), 2013.

[99] Graham Upton and Ian Cook. *A dictionary of statistics*, volume 3. Oxford University Press, 2014.

[100] Wolfgang Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

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

[102] Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Science & Business Media, 2013.

[103] Abraham Wald. An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics*, pages 549–555, 1947.

[104] Ronald L Wasserstein and Nicole A Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[105] René Weber and Lucy Popova. Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, 6(3):190–213, 2012.

[106] Guang Yang, Dungang Liu, Junyuan Wang, and Min-ge Xie. Meta-analysis framework for exact inferences with application to the analysis of rare events. *Biometrics*, 72(4):1378–1386, 2016.

[107] S Yusuf, R Peto, J Lewis, R Collins, and P Sleight. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases*, 27(5):335–371, 1985.

[108] Melvyn WB Zhang, Roger CM Ho, Mike WL Cheung, Erin Fu, and Anselm Mak. Prevalence of depressive symptoms in patients with chronic obstructive pulmonary disease: a systematic review, meta-analysis and meta-regression. *General Hospital Psychiatry*, 33(3):217–223, 2011.

[109] Zhiwei Zhang and Bo Zhang. A likelihood paradigm for clinical trials. *Journal of Statistical Theory and Practice*, 7:155–177, 2013.

Appendix A

CHAPTER 2 APPENDINCES

A.1 Proof of Lemma 1

Proof. Controlling the Type I error rate corresponds to requiring

$$\mathbb{P}[\, d = A; \theta\,] + \mathbb{P}[\, d = B; \theta\,] \leq \alpha.$$

In the limiting case of the Bayes rules for the Normal location problem results in

$$(\mathbb{P}[\, \theta > \theta_A | Y\,], \mathbb{P}[\, \theta < \theta_B | Y\,]) = (1 - \Phi(\theta_A - Y), \Phi(\theta_B - Y)) = (\Phi(Y - \theta_A), \Phi(\theta_B - Y)).$$

Hence, for any null interval (θ_B, θ_A) the posterior can be represented as a point on a single parameterized curve on the unit simplex; examples are shown in Figure 2.1. For any null interval the representations lie along a curve given by

$$\mathbb{P}[\, \theta < \theta_B | Y\,] = \Phi(\theta_B - \theta_A - \Phi^{-1}(\mathbb{P}[\, \theta > \theta_A | Y\,])),$$

which is symmetric when reflected in the line $\mathbb{P}[\, \theta > \theta_A | Y\,] = \mathbb{P}[\, \theta < \theta_B | Y\,]$, that it intersects when $Y = (\theta_B + \theta_A)/2$.

For any specific θ_B, θ_A, the critical values of Y occur where this curve hits the boundary between where $d = A$ versus C and $d = B$ versus C. Because of the assumed symmetry of the loss with respect to A and B, this is where $Y = \frac{\theta_B + \theta_A}{2} \pm r$ for some radius r, which can be found by solving

$$\frac{l_3}{l_2 + l_3} + \frac{l_1 - l_2 - l_3}{l_2 + l_3} \Phi(\frac{\theta_B - \theta_A}{2} - r) = \Phi(\frac{\theta_B - \theta_A}{2} + r).$$

The Type I error rate at both $\theta = \theta_A$ and $\theta = \theta_B$ is

$$\Phi(\frac{\theta_A - \theta_B}{2} - r) + \Phi(\frac{\theta_B - \theta_A}{2} - r),$$

and is lower for any θ within the null interval. Hence we obtain an unbiased test of level α by choosing r such that

$$\alpha = \Phi(\frac{\theta_B - \theta_A}{2} - r) + 1 - \Phi(\frac{\theta_B - \theta_A}{2} + r),$$

Using the expression that solves for r from before, the Bayes rules can only be unbiased if

$$\frac{l_3}{l_2+l_3}+\left(\frac{l_1}{l_2+l_3}-1\right)\Phi(\frac{\theta_B-\theta_A}{2}-r)=1-\alpha+\Phi(\frac{\theta_B-\theta_A}{2}-r).$$

Ruling out solutions where l_1, l_2, l_3 change depending on the null hypothesis being considered, the only solution for l_1, l_2, l_3 occurs by equating terms in $\Phi(\frac{\theta_B-\theta_A}{2}-r)$, and the remaining constants. Doing this we obtain $(l_1, l_2, l_3) \propto (2, \alpha, 1-\alpha)$, which is therefore a condition for the Bayes rules to be compatible with unbiased testing at level α. This solution is in fact the UMPU test, as used by Schervish, and results in incoherence as it violates $l_1 < l_2 + l_3$. □

A.2 Proof of Lemma 2

Proof. Criterion s_δ specifies controlling the weighted sum of the maximum of the sign error and the absolute difference of sign error rates. That is, for fixed $0 < \gamma < 1$ and $0 \leq \delta \leq 1$,

$$(1-\delta)\max\left(\mathbb{P}[\,d=A;\theta\,],\mathbb{P}[\,d=B;\theta\,]\right)+\delta|\mathbb{P}[\,d=A;\theta\,]-\mathbb{P}[\,d=B;\theta\,]|\leq\gamma.$$

Denoting the left hand side as $t(\theta)$, in the Normal location problem under the assumed symmetry we have

$$t(\theta)=\begin{cases}1-\Phi\left(\theta-\frac{\theta_B+\theta_A}{2}+r\right)-\delta\Phi\left(\theta-\frac{\theta_B+\theta_A}{2}-r\right), & \text{if } \theta<\frac{\theta_B+\theta_A}{2}\\ 1-\Phi\left(\frac{\theta_B+\theta_A}{2}-\theta+r\right)-\delta\Phi\left(\frac{\theta_B+\theta_A}{2}-\theta-r\right), & \text{if } \theta>\frac{\theta_B+\theta_A}{2}\\ \Phi\left(r\right)-\delta\Phi\left(-r\right), & \text{if } \theta=\frac{\theta_B+\theta_A}{2}.\end{cases}$$

For unbiasedness, we consider $\theta \in [\theta_B, \theta_A]$ $t(\theta)$ at which $t(\theta)$ is maximized. We separately consider θ each side of $\frac{\theta_B+\theta_A}{2}$, the lowest point.

For $\theta < \frac{\theta_B+\theta_A}{2}$ the derivative of $t(\theta)$ is

$$t'(\theta)=-\frac{1}{\sqrt{2\pi}}e^{-\frac{(\theta-\frac{\theta_B+\theta_A}{2}+r)^2}{2}}-\frac{\delta}{\sqrt{2\pi}}e^{-\frac{(\theta-\frac{\theta_B+\theta_A}{2}-r)^2}{2}}.$$

Since the exponential function is non-negative, this derivative is always negative, meaning from θ_B to $\frac{\theta_B+\theta_A}{2}$ the function $t(\theta)$ is decreasing.

For $\theta > \frac{\theta_B+\theta_A}{2}$ we get

$$t'(\theta)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(\theta-\frac{\theta_B+\theta_A}{2}+r)^2}{2}}+\frac{\delta}{\sqrt{2\pi}}e^{-\frac{(\theta-\frac{\theta_B+\theta_A}{2}-r)^2}{2}}.$$

This expression is always positive, meaning from $\frac{\theta_B+\theta_A}{2}$ to θ_A the function $t(\theta)$ is increasing.

Taking the results from the two sides together, we see that $t(\theta)$ is maximized at either θ_B or θ_A. Since $t(\theta_B) = t(\theta_A)$, forcing unbiasedness means findings r that solves

$$\begin{aligned} \gamma &= \left[1 - \Phi\left(\frac{\theta_B - \theta_A}{2} + r\right)\right] - \delta\Phi\left(\frac{\theta_B - \theta_A}{2} - r\right) \\ \Leftrightarrow \gamma &= \Phi\left(\frac{\theta_A - \theta_B}{2} - r\right) - \delta\Phi\left(\frac{\theta_B - \theta_A}{2} - r\right). \end{aligned}$$

From the second equation, in which the RHS is decreasing in r, we see that to have a solution with positive r we must have

$$\gamma > \Phi\left(\frac{\theta_A - \theta_B}{2}\right) - \delta\Phi\left(\frac{\theta_B - \theta_A}{2}\right).$$

Defining the p_δ-value as the greatest γ under which the null would not be rejected, we therefore have

$$p_\delta(Y) = (1-\delta)\min(\Phi(Y-\theta_A), \Phi(\theta_B - Y)) + \Phi(\theta_A - Y) - \Phi(\theta_B - Y).$$

In the Bayesian setting with a limiting flat prior, this can be instead stated as

$$p_\delta(Y) = (1-\delta)\min\left(\mathbb{P}[\,\theta > \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,]\right) + \mathbb{P}[\,\theta_B < \theta < \theta_A|Y\,]$$

and the test is the Bayes rule under the loss where

		Decision		
		A	C	B
Truth	$\theta < \theta_B$	$1-\delta$	γ	0
	$\theta \in (\theta_B, \theta_A)$	1	γ	1
	$\theta > \theta_A$	0	γ	$1-\delta$

.

This is Schervish coherent, corresponding to $(l_1, l_2, l_3) \propto (1-\delta, \gamma, 1-\gamma)$.

While this concludes the proof, we note two special circumstances, one of which was illustrated in Peskun's paper [77]. We can extend the criterion s_δ to s_0 and s_1 by setting $\delta = 0$ and $\delta = 1$, respectively. For $\delta = 0$, the criterion becomes

$$\max\left(\mathbb{P}[\,d = A; \theta\,], \mathbb{P}[\,d = B; \theta\,]\right) \leq \gamma,$$

i.e. controlling only the maximum of the sign error rates. The p_0-value becomes

$$p_0(Y) = \min(\Phi(Y - \theta_A), \Phi(\theta_B - Y)) + \Phi(\theta_A - Y) - \Phi(\theta_B - Y).$$

This can be rewritten as

$$p_{\delta=0}(Y) = \begin{cases} \Phi(Y - \theta_B), & \text{if } Y < (\theta_B + \theta_A)/2 \\ \Phi(\theta_A - Y), & \text{if } Y > (\theta_B + \theta_A)/2 \end{cases}$$

This is the e-value (evidence value) from Peskun's paper [77]. Viewed as the Bayes rule under a limiting flat prior, the p_0-value can be interpreted as

$$p_0(Y) = \min(\mathbb{P}[\,\theta > \theta_A|Y\,] + \mathbb{P}[\,\theta_B < \theta < \theta_A|Y\,], \mathbb{P}[\,\theta < \theta_B|Y\,] + \mathbb{P}[\,\theta_B < \theta < \theta_A|Y\,]),$$

and hence tests that reject when $p_0(Y) < \gamma$ correspond to the loss $(l_1, l_2, l_3) \propto (1, \gamma, 1 - \gamma)$.

If we instead set $\delta = 1$, the criterion becomes

$$|\mathbb{P}[\,d = A; \theta\,] - \mathbb{P}[\,d = B; \theta\,]| \leq \gamma,$$

i.e. controlling the absolute difference in sign error rates.

In this case, p_1-value can be written as

$$p_1(Y) = \Phi(\theta_A - Y) - \Phi(\theta_B - Y),$$

and in the case of a limiting flat prior, can be instead stated as

$$p_1(Y) = \mathbb{P}[\theta_B < \theta < \theta_A|Y],$$

which corresponds to a loss of $(l_1, l_2, l_3) \propto (0, \gamma, 1 - \gamma)$.

□

Appendix B

CHAPTER 3 APPENDICES

B.1 Theoretical Results from Hoeffding et al

Using Hoeffding's result, suppose we had a level α test of the null value $p_F = p_0$, that was exact under homogeneity and rejected the null for values of $Y_+ \notin (b, c)$. As the test is exact we must have

$$1 - \alpha \leq \sum_{r=b}^{c} \binom{n_+}{r} p_0^r (1 - p_0)^{(n_+ - r)},$$

i.e. it controls the Type I error rate at a level strictly no greater than α [20, §8.3]. But Hoeffding's result means that even under heterogeneity of the p_i, so long as the 'weak null' that $p_F = p_0$ holds and $b \leq n_+ p_0 \leq c$, we must have

$$1 - \alpha \leq \sum_{r=b}^{c} \binom{n_+}{r} p_0^r (1 - p_0)^{(n_+ - r)} \leq \mathbb{P}[\, b \leq Y_+ \leq c\,],$$

meaning that the test is still exact. (While not explored here, Hoeffding's result further implies that heterogeneity within the contributing studies does not invalidate the test, so long as the outcomes are independent.) As is typical in exact tests, conservative control is expected in some circumstances – in this case when the p_i are heterogeneous. And when the p_i are close to homogeneous, as one might often expect in meta-analytic practice, then the conservatism under heterogeneity should be negligibly less than that under true homogeneity.

B.2 Exact inference for binomial proportions

We consider the Clopper-Pearson [24] and Blaker [17] methods that give exact confidence intervals for binomial proportions, meaning that there coverage is guaranteed to be at least the nominal level, regardless of the underlying proportion.

The two-sided Clopper-Pearson confidence interval, for a given observation y from

$\text{Bin}(n, p)$, constructs confidence interval (p_L, p_U) that satisfies

$$\sum_{r=y}^{n} \binom{n}{r} p_L^r (1-p_L)^{n-r} = \alpha/2,$$
$$\sum_{r=0}^{y} \binom{n}{r} p_U^r (1-p_U)^{n-r} = \alpha/2.$$

This interval is widely available in software, for example the `binom.test()` function in base R.

Blaker's method for exact confidence intervals combines the probability of the smaller observed tail with the smallest probability of the opposite tail that does not exceed that observed tail probability. Formally, it defines the function

$$a(p, y) = \begin{cases} P(Y \geq y) + P(Y \leq y^*), & \text{if } P(Y \geq y) < P(Y \leq y) \\ P(Y \leq y) + P(Y \geq y^{**}), & \text{if } P(Y \geq y) > P(Y \leq y) \\ 1, & \text{otherwise,} \end{cases}$$

where

$$y^* \equiv \text{largest u such that } P(Y \leq u) \leq P(Y \geq y),$$
$$y^{**} \equiv \text{smallest v such that } P(Y \geq v) \leq P(Y \leq y),$$

and gives the $1 - \alpha$ confidence interval as the set of p such that $a(p, y) \geq \alpha$. This interval is less well-known than Clopper-Pearson, but is readily available in e.g. R's `exactci` package [27].

Either confidence interval can of course be inverted to give tests, rejecting a null value when it does not lie in the confidence interval. Corresponding p-values can similarly be obtained by finding the minimum coverage level (i.e. maximum α) such that the null is not rejected. However, the resulting Blaker p-values can have counterintuitive properties [101]. Specifically, they need not be strictly monotonic with respect to the null value on either side of the maximum, and are also discontinuous due to the discrete nature of observed binomial proportions. Figure B.1 illustrates this for various situations.

In contrast, Clopper-Pearson p values (shown as dashed lines on Supplemental Figure S4 have neither problem, although they derive from intervals that tend to be slightly wider than Blaker's and hence give slightly more conservative p-values.

For the KPMP example, Figure B.2 shows how the differences between the intervals are minor, with Clopper-Pearson intervals as expected being slightly wider for each study.

B.3 Bias of $\hat{\delta^2}$

To calculate the expectation of the general case of the plug-in estimator from Section 3.3, we first note the following definitions:

$$\begin{aligned}
\delta^2 &= \sum_{i=1}^{k} \frac{n_i}{n_+}(p_i - p_F)^2 = \sum_{i=1}^{k} \frac{n_i}{n_+} p_i^2 - \left(\sum_{i=1}^{k} \frac{n_i}{n_+} p_i \right)^2 \\
\hat{p}_i &= Y_i / n_i \\
\hat{\delta}^2 &= \sum_{i=1}^{k} \frac{Y_i^2}{n_i n_+} - \left(\sum_{i=1}^{k} \frac{Y_i}{n_+} \right)^2 .
\end{aligned}$$

It therefore suffices to use standard Binomial moments to provide

$$\begin{aligned}
\mathbb{E}[\, Y_i^2 \,] &= n_i(p_i(1-p_i) + n_i p_i^2), \\
\mathbb{E}\left[(\sum_{i=1}^{k} Y_i)^2 \right] &= \left(\sum_{i=1}^{k} n_i p_i (1-p_i) \right) + \left(\sum_{i=1}^{k} n_i p_i \right)^2 ,
\end{aligned}$$

from which we find that

$$\begin{aligned}
\mathbb{E}[\, \hat{\delta}^2 \,] &= \left(\sum_{i=1}^{k} \frac{p_i(1-p_i) + n_i p_i^2}{n_+} \right) - \left(\sum_{i=1}^{k} \frac{n_i}{n_+^2} p_i(1-p_i) \right) - \left(\sum_{i=1}^{k} \frac{n_i}{n_+} p_i \right)^2 \\
&= \delta^2 + \frac{1}{n_+} \sum_{i=1}^{k} \left(1 - \frac{n_i}{n_+} \right) p_i(1-p_i).
\end{aligned}$$

We see that the plug-in estimate is biased slightly high, by a term that is bounded above by $\frac{k-1}{4n_+}$, but will be notably less than this when the p_i are either all small or all large. In practice this positive bias will make the approximated true coverage of the nominal 95% Blaker intervals slightly overstated, i.e. it will be conservative.

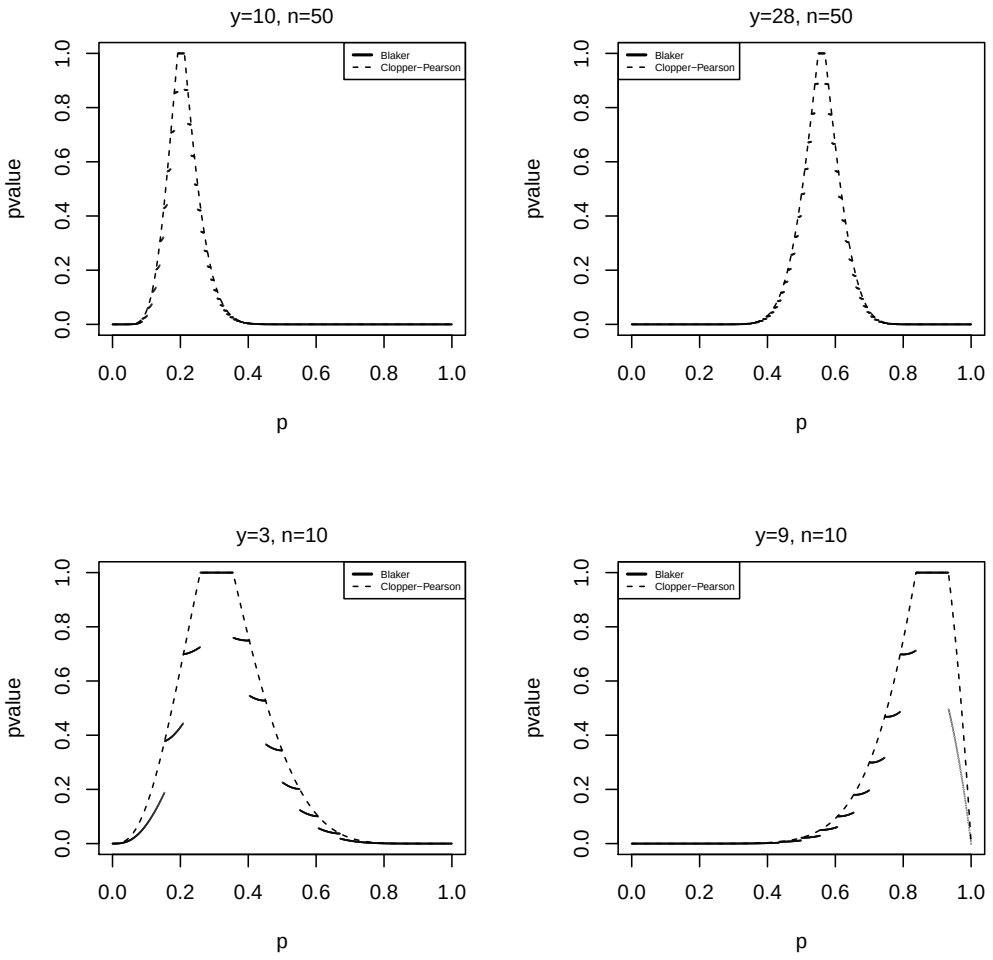


Figure B.1: Behaviour of Blaker and Clopper-Pearson p-values for varying null values, given fixed y and n. The Blaker p-values are discontinuous and slightly violate monotonicity, even away from the maximum p-value, while the Clopper-Pearson intervals have neither problem.

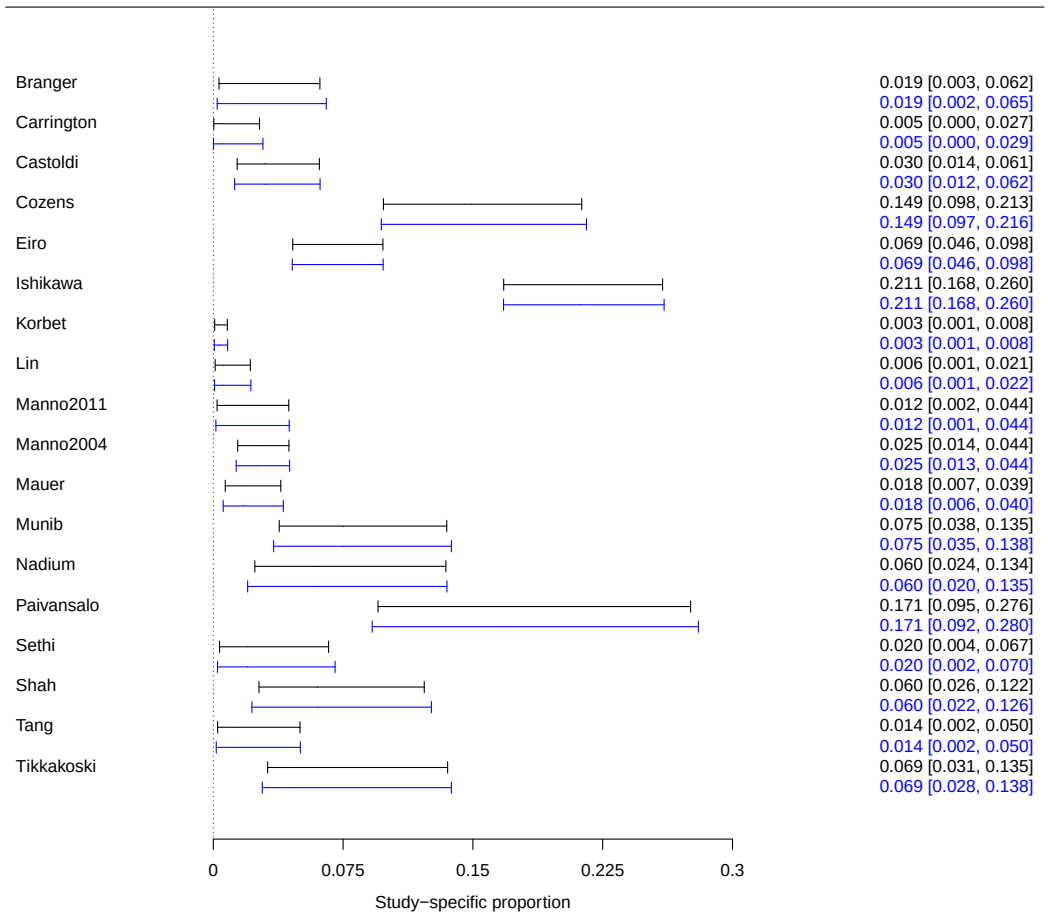


Figure B.2: Blaker (black) and Clopper-Pearson (blue) 95% confidence intervals for the KPMP study-specific data. The Clopper-Pearson intervals are slightly wider, for each study.

For an unbiased estimate of δ^2, we can subtract

$$\frac{1}{n_+}\sum_{i=1}^{k}\left(1-\frac{n_i}{n_+}\right)\frac{n_i}{n_i-1}\hat{p}_i(1-\hat{p}_i)$$

to the simpler plug-in approach described in Section 3.1.3. In the KPMP example this correction takes the plug-in $\hat{\delta}^2 = 0.0035$ to a bias-corrected 0.0033.

Appendix C

CHAPTER 4 APPENDICES

C.1 Coverage for the two strata setting with unequal row totals

Figures C.1, C.2, and C.3 consider the situation where $M_k = (400, 600)$, $N_k = (350, 650)$, and $T_k = (25, 25)$. Again, we confirm that the method is exact, that our measure of heterogeneity is useful for estimating excess coverage, and that the excess coverage—even when there is moderate heterogeneity—is not too extreme.

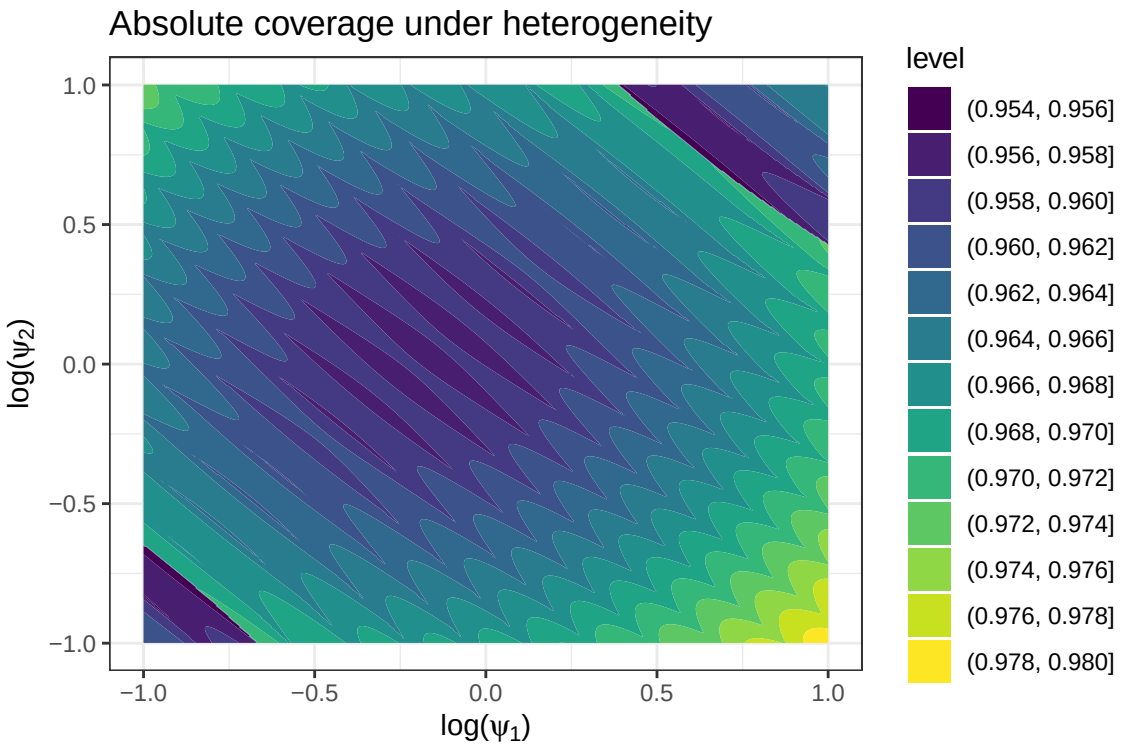


Figure C.1: Absolute coverage under heterogeneity of the 95% Blaker intervals using our method for a grid of values with table margins $M_k = (400, 600)$, $N_k = (350, 650)$, $T_k = (25, 25)$.

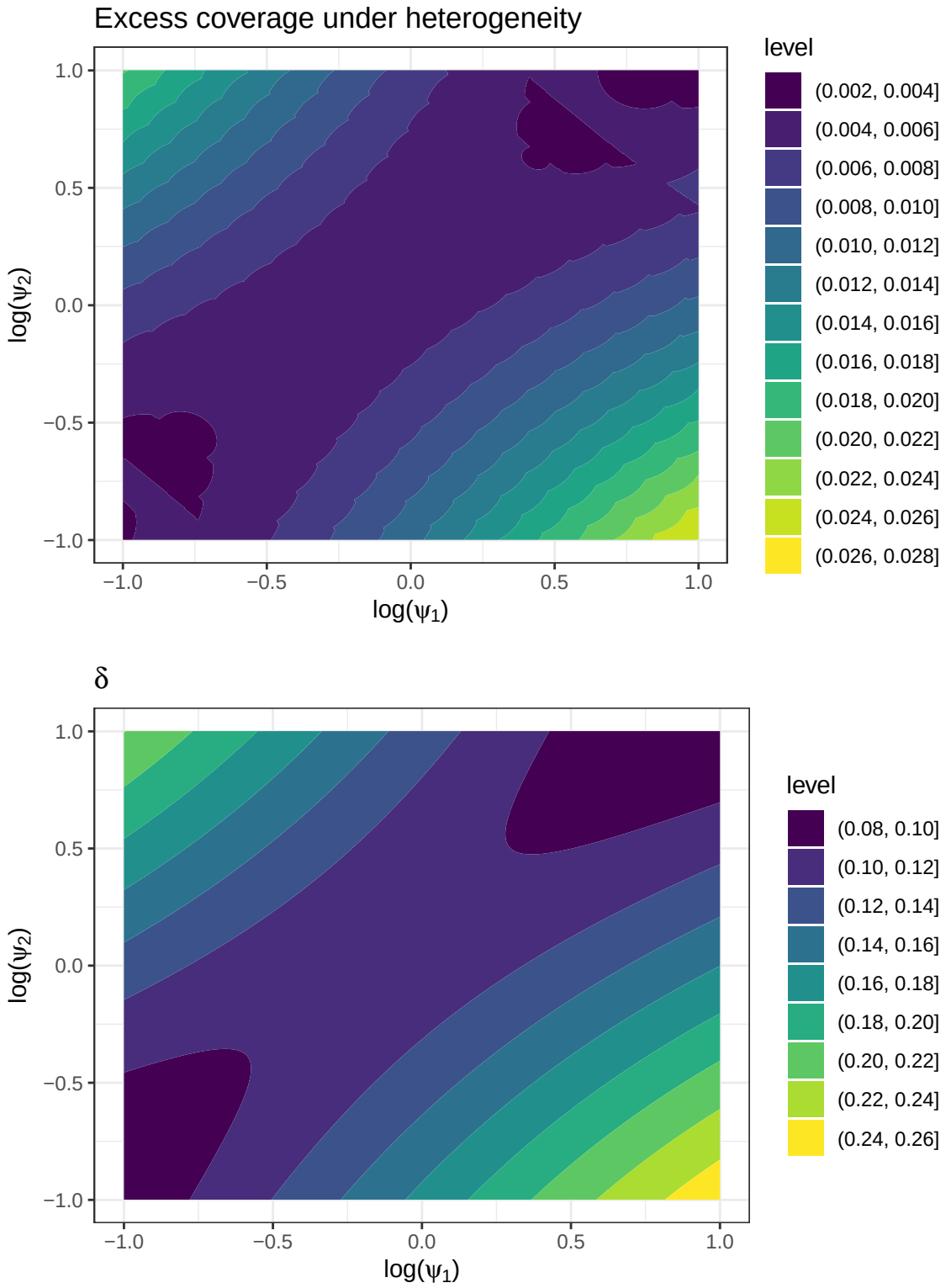


Figure C.2: Top panel: Excess coverage under heterogeneity of the 95% Blaker intervals using our method for a grid of values with table margins of $M_k = (400, 600)$, $N_k = (350, 650)$, $T_k = (25, 25)$. Bottom panel: Measure of heterogeneity, δ, for a grid of values with same table margins.

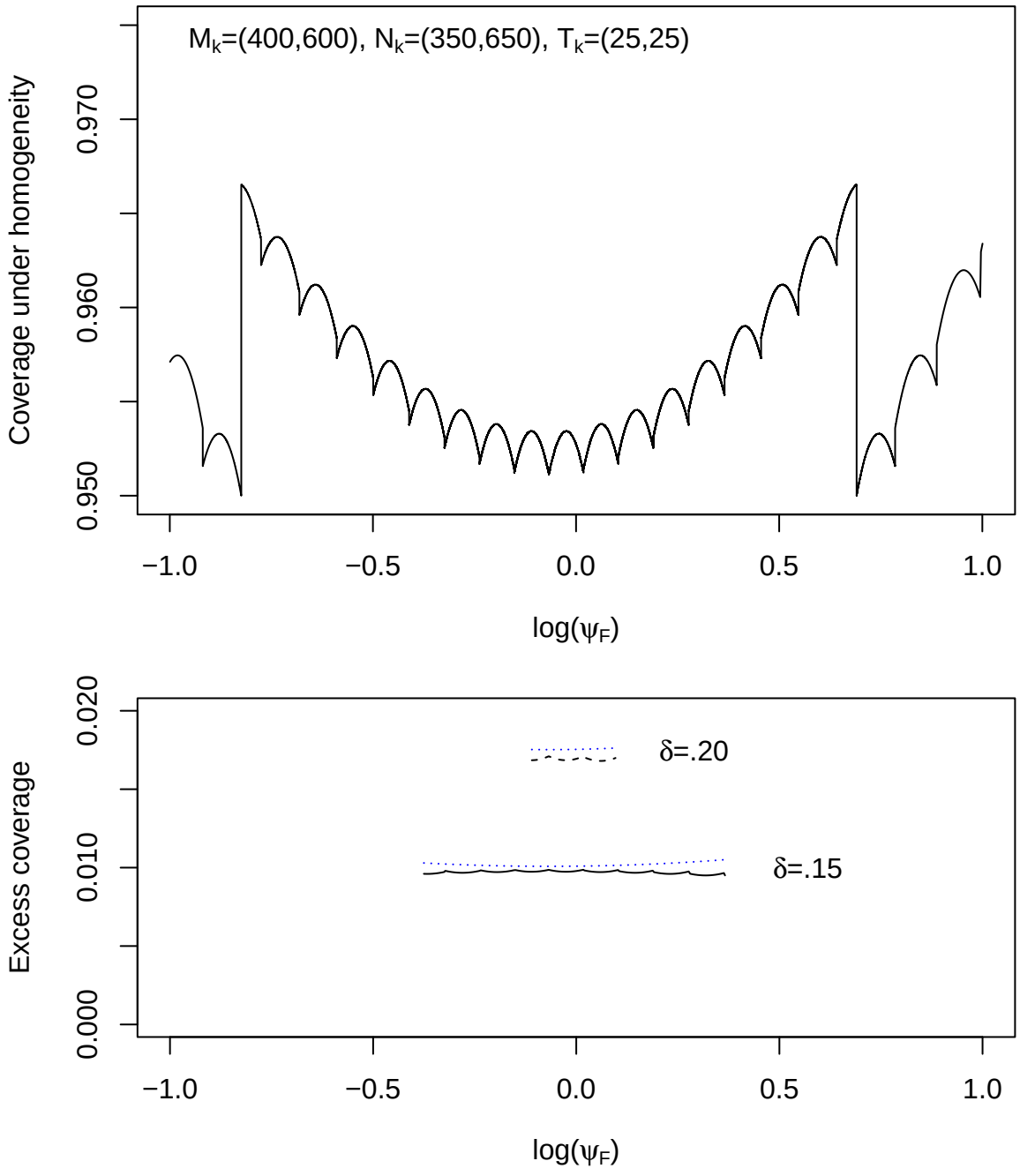


Figure C.3: Top panel: Coverage of the 95% Blaker intervals using our method for varying values of the ψ_F under homogeneity and for $\psi_1 < \psi_2$, for table margins of $M_k = (400, 600)$, $N_k = (350, 650)$, $T_k = (25, 25)$. Bottom panel: Excess coverage of the same intervals under heterogeneity for certain values of δ (blue, small dotted lines represent Wald-test based approximations).

VITA

Spencer Hansen was born in Salt Lake City, Utah. He attended the United State Military Academy in West Point, New York, and graduated with a Bachelor's degree in Mathematical Sciences (with Honors) in 2012. Spencer left the army after serving five years and entered the PhD program in Biostatistics at the University of Washington in 2017.