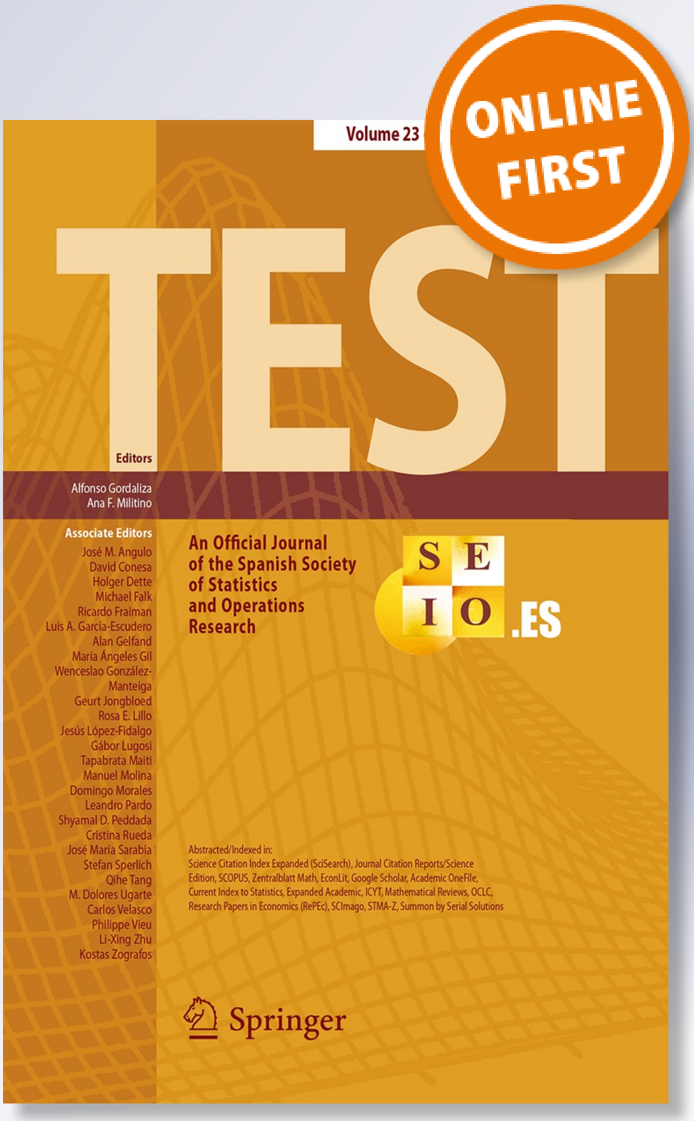
Bayesian tests for composite alternative hypotheses in cross-tabulated data

Daniel Yekutieli

TEST
An Official Journal of the Spanish
Society of Statistics and Operations
Research

ISSN 1133-0686

TEST
DOI 10.1007/s11749-014-0407-1



Volume 23

ONLINE FIRST

TEST

Editors
Alfonso Gordaliza
Ana F. Militino

Associate Editors
José M. Angulo
David Conesa
Holger Dette
Michael Falk
Ricardo Fraiman
Luis A. García-Escudero
Alan Gelfand
María Ángeles Gil
Wenceslao González-Manteiga
Geurt Jongbloed
Rosa E. Lillo
Jesús López-Fidalgo
Gábor Lugosi
Tapabrata Maiti
Manuel Molina
Domingo Morales
Leandro Pardo
Shyamal D. Peddada
Cristina Rueda
José María Sarabia
Stefan Sperlich
Qihe Tang
M. Dolores Ugarte
Carlos Velasco
Philippe Vieu
Li-Xing Zhu
Kostas Zografos

An Official Journal
of the Spanish Society
of Statistics
and Operations
Research

SE
IO .ES

Abstracted/Indexed in:
Science Citation Index Expanded (SciSearch), Journal Citation Reports/Science Edition, SCOPUS, Zentralblatt Math, EconLit, Google Scholar, Academic OneFile, Current Index to Statistics, Expanded Academic, ICYT, Mathematical Reviews, OCLC, Research Papers in Economics (RePEc), SCImago, STMA-Z, Summon by Serial Solutions

Springer

Springer

Your article is protected by copyright and all rights are held exclusively by Sociedad de Estadística e Investigación Operativa. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

Springer

Author's personal copy

TEST
DOI 10.1007/s11749-014-0407-1

ORIGINAL PAPER

Bayesian tests for composite alternative hypotheses in cross-tabulated data

Daniel Yekutieli

Received: 12 January 2014 / Accepted: 2 October 2014
© Sociedad de Estadística e Investigación Operativa 2014

Abstract We present a methodology for constructing significance tests for "difficult" composite alternative hypotheses that have no natural test statistic. We apply our methodology to construct exact tests for cross-tabulated data, and our motivating example is constructing a test for discovering Simpson's Paradox. Our tests are Bayesian extensions of the likelihood ratio test; they are optimal with respect to the prior distribution and are also closely related to Bayes factors and Bayesian FDR controlling testing procedures.

Keywords Hypotheses testing · Simpson's Paradox · Composite alternative hypotheses · Exact tests · Likelihood ratio tests · Bayes rules · Bayes factors

Mathematics Subject Classification 62C10 · 62G10

1 Introduction

We present a methodology for constructing significance tests for composite alternative hypotheses. To apply our tests, it is necessary to elicit a prior distribution on the parameters and to specify the subset of the parameter space that corresponds to the composite alternative hypothesis, which we call the discovery event. Our tests are frequentist significance tests that use a Bayesian machinery to induce an order to the data sample space. The significance level for our test is the probability under the null hypothesis of observing a data set whose posterior probability of the discovery event

D. Yekutieli (✉)
Department of Statistics and OR, Tel Aviv University, Tel Aviv 6997801, Israel
e-mail: yekutiel@post.tau.ac.il

Published online: 17 October 2014

Author's personal copy

is larger than the posterior probability of the discovery event for the observed data. Our level α tests reject the null hypothesis if this significance level is $\leq \alpha$.

As our tests are complicated and computationally intensive, we suggest applying our methodology for testing difficult alternative hypotheses that have no natural scalar test statistic for ordering the data sample space. In this paper, we apply our methodology to construct exact tests for cross-tabulated data, in which use of exact tests is well established and for which applying the Bayesian machinery is relatively straightforward.

In the job satisfaction example (Agresti 2002), we test the association between two ordinal variables in a 4-by-4 contingency table. This is an instructive example in which there are natural scalar statistics for the composite alternative hypotheses. We use this to illustrate the difference between our statistics and the commonly used statistics, and we show, that in this case, even though the statistics are derived very differently the order they induce to the data sample space is very similar.

In the death penalty example (also taken from Agresti 2002), the composite alternative hypothesis is the Simpson's Paradox that conditional on victim's race, black defendants are more likely to receive death sentence than white defendants, while marginally black defendants are less likely to receive death sentence than white defendants. In this case (the data is a 2-by-2-by-2 contingency table) the sample odds ratio is the natural statistic for comparing the conditional and marginal risks that black and white defendants receive death sentence. However, since tables with larger conditional odds ratio also have larger marginal odds ratio, it is not clear how to construct a scalar statistic that orders all the 2-by-2-by-2 tables according to the property that the conditional and marginal odds ratios have opposite signs. In contrast, the discovery event for this example is that the conditional parameter odds ratio is positive and the marginal parameter odds ratio is negative, and our statistic for ordering the data sample space is the posterior probability that this event occurs.

In Sect. 2, we present our general testing methodology and its conditional variant we use for constructing exact tests for cross-tabulated data, phrase and prove their optimality property, and explain the relation between our tests and Bayesian FDR methodology, Bayes factors, and likelihood ratio tests. The job satisfaction example is given in Sect. 3, the death penalty example is given in Sect. 4 and we end the paper with a discussion.

2 Mean most powerful tests

We denote the parameter by $\mathbf{p} \in \mathcal{P}$, $\pi(\mathbf{p})$ is the prior distribution, the data are $\mathbf{N} \in \Omega$, and the likelihood is $\Pr(\mathbf{n}|\mathbf{p})$. The alternative hypothesis is $H_1 : \mathbf{p} \in \mathcal{P}_1$, for $\mathcal{P}_1 \subseteq \mathcal{P}$. Following Benjamini and Hochberg (1995) who referred to rejecting the null hypothesis as making a statistical discovery, we call $\mathcal{P}_1$ the discovery event. For constructing our test, we need to specify another subset of the parameter space, $\mathcal{P}_0 \subseteq \mathcal{P} - \mathcal{P}_1$, which we call the non-discovery event. The role of $\mathcal{P}_0$ is to determine the optimality property of the test, given in Definition 2.1. We explain how to set $\mathcal{P}_0$ in Remark 2.3. The null hypothesis H_0 does not have to correspond to an explicit subset of $\mathcal{P}_0$; all we will need is that the null hypothesis specifies a null distribution $\Pr_{H_0}(\mathbf{N} = \mathbf{n})$

Author's personal copy

on Ω. Tests are mappings $\mathcal{T}: \ \Omega \to \{0, 1\}$. For $S \subseteq \Omega$, let $\mathcal{T}(S) := I(\mathbf{n} \in S)$, where $\mathcal{T}(S) = 1$ corresponds to declaring that $\mathbf{p} \in \mathcal{P}_1$. Thus, the significance level of $\mathcal{T}(S)$ is $Pr_{H_0}(\mathbf{N} \in S)$.

Our tests are Bayes rules for discriminating between $\mathcal{P}_0$ and $\mathcal{P}_1$ that minimize the average risk for the following loss function:

$$L(S; \lambda_1, \lambda_2) = \lambda_1 \cdot I(\mathbf{N} \in S,\ \mathbf{P} \in \mathcal{P}_0) + \lambda_2 \cdot I(\mathbf{N} \notin S,\ \mathbf{P} \in \mathcal{P}_1), \tag{1}$$

where $0 < \lambda_1$ is the loss incurred by a type I error and $0 < \lambda_2$ is the loss incurred by a type II error. As the marginal distribution of $\mathbf{N}$ is

$$\Pr(\mathbf{N} = \mathbf{n}) = \int_{\mathbf{p}} \pi(\mathbf{p}) \cdot \Pr(\mathbf{N} = \mathbf{n}|\ \mathbf{p})\ d\mathbf{p},$$

and the conditional distribution of $\mathbf{p}$ given $\mathbf{N} = \mathbf{n}$ is

$$\pi(\mathbf{p}|\ \mathbf{n}) = \Pr(\mathbf{N} = \mathbf{n}|\ \mathbf{p}) \cdot \pi(\mathbf{p}) / \Pr(\mathbf{N} = \mathbf{n}),$$

the average risk can be expressed as

$$\begin{aligned} &\sum_{\mathbf{n} \in \Omega} \Pr(\mathbf{n}) \cdot \int_{\mathbf{p}} \pi(\mathbf{p}|\ \mathbf{n}) \cdot [\lambda_1 \cdot I(\mathbf{n} \in S,\ \mathbf{P} \in \mathcal{P}_0) + \lambda_2 \cdot I(\mathbf{n} \notin S,\ \mathbf{P} \in \mathcal{P}_1)]\ d\mathbf{p} \\ &\quad = \sum_{\mathbf{n} \in S} \Pr(\mathbf{n}) \cdot \lambda_1 \cdot \Pr(\mathbf{P} \in \mathcal{P}_0|\ \mathbf{n}) + \sum_{\mathbf{n} \notin S} \Pr(\mathbf{n}) \cdot \lambda_2 \cdot \Pr(\mathbf{P} \in \mathcal{P}_1|\ \mathbf{n}). \end{aligned} \tag{2}$$

Thus for $\delta = \lambda_1/\lambda_2$, S that minimizes the average risk in (2) is

$$S^{Bayes}(\delta) = \left\{ \mathbf{n} :\ \delta \le \frac{\Pr(\mathbf{P} \in \mathcal{P}_1|\ \mathbf{n})}{\Pr(\mathbf{P} \in \mathcal{P}_0|\ \mathbf{n})} \right\}. \tag{3}$$

Similarly, the Bayes rule can be specified according to its significance level. For $\alpha \in [0, 1]$, let $S^{Bayes}(\alpha) := S^{Bayes}(\delta_\alpha)$ for

$$\delta_\alpha = \min\{\delta :\ Pr_{H_0}(\mathbf{N} \in S^{Bayes}(\delta)) \le \alpha\}.$$

Definition 2.1 1. The *mean significance level* of $\mathcal{T}(S)$ is $Pr(\mathbf{N} \in S|\ \mathbf{p} \in \mathcal{P}_0)$.
2. The *mean power* of $\mathcal{T}(S)$ is $Pr(\mathbf{N} \in S|\ \mathbf{p} \in \mathcal{P}_1)$.
3. $\mathcal{T}(S)$ is a *mean most powerful* test if all tests with less or equal mean significance level have less or equal mean power.

Proposition 2.2 *$\forall \delta$, $\mathcal{T}(S^{Bayes}(\delta))$ is a mean most powerful test.*

Proof Let $\mathcal{T}(\tilde{S})$ be a test with less or equal mean significance than $\mathcal{T}(S^{Bayes})$,

$$\Pr(\mathbf{N} \in \tilde{S}|\ \mathbf{P} \in \mathcal{P}_0) \le Pr(\mathbf{N} \in S^{Bayes}|\ \mathbf{P} \in \mathcal{P}_0). \tag{4}$$

Author's personal copy

We begin by expressing

$$\Pr(\mathbf{N} \in \tilde{S} |\, \mathbf{p} \in \mathcal{P}_0) = \sum_{\mathbf{n} \in \tilde{S}} \Pr(\mathcal{P}_0 |\, \mathbf{n}) \cdot \Pr(\mathbf{n}) / \Pr(\mathcal{P}_0) \tag{5}$$

and

$$\Pr(\mathbf{N} \in S^{Bayes} |\, \mathbf{p} \in \mathcal{P}_0) = \sum_{\mathbf{n} \in S^{Bayes}} \Pr(\mathcal{P}_0 |\, \mathbf{n}) \cdot \Pr(\mathbf{n}) / \Pr(\mathcal{P}_0). \tag{6}$$

Subtracting the summands in $S^{Bayes} \cap \tilde{S}$ from the sums in (5) and (6) and multiplying by $\Pr(\mathcal{P}_0)$, Inequality (4) implies that

$$\sum_{\mathbf{n} \in \tilde{S} - (S^{Bayes} \cap \tilde{S})} \Pr(\mathcal{P}_0 |\, \mathbf{n}) \cdot \Pr(\mathbf{n}) \leq \sum_{\mathbf{n} \in S^{Bayes} - (S^{Bayes} \cap \tilde{S})} \Pr(\mathcal{P}_0 |\, \mathbf{n}) \cdot \Pr(\mathbf{n}). \tag{7}$$

According to the construction of S^{Bayes}, $\forall \mathbf{n}_1 \in \tilde{S} - (S^{Bayes} \cap \tilde{S})$ and $\forall \mathbf{n}_2 \in S^{Bayes} - (S^{Bayes} \cap \tilde{S})$

$$\Pr(\mathcal{P}_1 |\, \mathbf{n}_1) / \Pr(\mathcal{P}_0 |\, \mathbf{n}_1) \leq \Pr(\mathcal{P}_1 |\, \mathbf{n}_2) / \Pr(\mathcal{P}_0 |\, \mathbf{n}_2). \tag{8}$$

Next, we express

$$\Pr(\mathbf{N} \in \tilde{S} |\, \mathbf{p} \in \mathcal{P}_1) = \sum_{\mathbf{n} \in S^{Bayes} \cap \tilde{S}} \Pr(\mathcal{P}_1 |\, \mathbf{n}) \cdot \Pr(\mathbf{n}) / \Pr(\mathcal{P}_1) \tag{9}$$

$$+ \sum_{\mathbf{n} \in \tilde{S} - (S^{Bayes} \cap \tilde{S})} \left(\Pr(\mathcal{P}_0 |\, \mathbf{n}) \cdot \frac{\Pr(\mathcal{P}_1 |\, \mathbf{n})}{\Pr(\mathcal{P}_0 |\, \mathbf{n})} \right) \cdot \frac{\Pr(\mathbf{n})}{\Pr(\mathcal{P}_1)} \tag{10}$$

and

$$\Pr(\mathbf{N} \in \tilde{S} |\, \mathbf{p} \in \mathcal{P}_1) = \sum_{\mathbf{n} \in S^{Bayes} \cap \tilde{S}} \Pr(\mathcal{P}_1 |\, \mathbf{n}) \cdot \Pr(\mathbf{n}) / \Pr(\mathcal{P}_1) \tag{11}$$

$$+ \sum_{\mathbf{n} \in S^{Bayes} - (S^{Bayes} \cap \tilde{S})} \left(\Pr(\mathcal{P}_0 |\, \mathbf{n}) \cdot \frac{\Pr(\mathcal{P}_1 |\, \mathbf{n})}{\Pr(\mathcal{P}_0 |\, \mathbf{n})} \right) \cdot \frac{\Pr(\mathbf{n})}{\Pr(\mathcal{P}_1)}. \tag{12}$$

Note that Expression (10) is the left hand side of (7) and Expression (12) is the right hand side of (7), divided by $\Pr(\mathcal{P}_1)$ and multiplied by a factor, which according to (8) is larger in each summand of (12) than in all of the summands of (10). Therefore, the sum in (12) is larger than the sum in (10), and as the sums in the right hand side of (9) and (11) are the same,

$$\Pr(\mathbf{N} \in \tilde{S} |\, \mathbf{p} \in \mathcal{P}_1) \leq \Pr(\mathbf{N} \in S^{Bayes} |\, \mathbf{p} \in \mathcal{P}_1).$$

□

Remark 2.3 Determining $\mathcal{P}_1$, $\mathcal{P}_0$, and $\pi(\mathbf{p})$ produces a family of mean most powerful tests. Per construction, $\mathcal{T}(S^{Bayes}(\alpha))$ has significance level α and has more mean

Author's personal copy

power than all mean most powerful tests with significance level $< \alpha$. According to Proposition 2.2, $\mathcal{T}(S^{Bayes}(\alpha))$ also has more mean power than all tests with smaller or equal mean significance level.

Ideally, the prior distribution captures the knowledge regarding the parameters that is available prior to the study. In the examples of the paper, we use conjugate non-informative priors that provide easy test statistic computation and yield general optimal tests for each alternative hypothesis. While the choice of $\mathcal{P}_1$ is usually dictated by the application, $\mathcal{P}_0$ can be any subset of $\mathcal{P} - \mathcal{P}_1$. If $\mathcal{P}_0 = \{\mathbf{p}_0\}$, then the mean significance level would equal the significance level; thus $\mathcal{T}(S^{Bayes}(\alpha))$ would have more mean power than all tests with significance level $\leq \alpha$. In the case that the choice of priors assigns zero probability to $\{\mathbf{p}_0\}$, we suggest setting $\mathcal{P}_0$ to be a "small" set containing $\mathbf{p}_0$ that would produce a very similar family of mean most powerful tests. In the examples of the paper, in which constructing a non-discovery event $\mathcal{P}_0$ that is disjoint from $\mathcal{P}_1$ is impossible because $\mathbf{p}_0$ is on the boundary $\mathcal{P}_1$, we set $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$. Note that setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$ yields

$$\frac{\Pr(\mathbf{P} \in \mathcal{P}_1 |\, \mathbf{n})}{\Pr(\mathbf{P} \in \mathcal{P}_0 |\, \mathbf{n})} = \frac{\Pr(\mathbf{P} \in \mathcal{P}_1 |\, \mathbf{n})}{1 - \Pr(\mathbf{P} \in \mathcal{P}_1 |\, \mathbf{n})},$$

which means that sorting the data points according to $Pr(\mathcal{P}_1 |\, \mathbf{n})$ is equivalent to sorting the data points according to $Pr(\mathcal{P}_1 |\, \mathbf{n}) / \Pr(\mathcal{P}_0 |\, \mathbf{n})$. Thus to construct our test, for each data point, we only need to assess the posterior probability of $\mathcal{P}_1$. In Sect. 2.2, we will connect the choice of $\mathcal{P}_0$ to the relation between our tests and likelihood ratio tests.

2.1 Conditional mean most powerful tests

In this section, we present the mean most powerful tests for the conditional analysis of contingency tables, in which the sample space is partitioned according to the row and column sums and a separate level α test is conducted in each partition.

Let a be the statistic that partitions the sample space $\Omega = \cup_{a \in \mathcal{A}} \Omega_a$, for $\mathcal{A} = \{a(N) : \; N \in \Omega\}$ the set of statistic values.

Definition 2.4 A conditional level α test is $\mathcal{T}(\mathcal{S}_{\mathcal{A}}(\alpha))$, such that $\forall a \in \mathcal{A}$, $\Pr_{H_0}(\mathbf{N} \in \mathcal{S}_{\mathcal{A}}(\alpha) | \mathbf{N} \in \Omega_a) \leq \alpha$.

To construct $\mathcal{S}_{\mathcal{A}}^{Bayes}(\alpha)$, the rejection region of the conditional mean most powerful test, we repeat the following for each $a \in \mathcal{A}$: sort the data points $\mathbf{N} \in \Omega_a$ according to $\Pr(\mathbf{P} \in \mathcal{P}_1 | \mathbf{N}) / \Pr(\mathbf{P} \in \mathcal{P}_0 | \mathbf{N})$ and, then following that order, as long as $\Pr_{H_0}(\mathbf{N} \in \mathcal{S}_{\mathcal{A}}^{Bayes}(\alpha) |\, \mathbf{N} \in \Omega_a) \leq \alpha$, sequentially add data points into $\mathcal{S}_{\mathcal{A}}^{Bayes}(\alpha)$.

Remark 2.5 Per construction, $\mathcal{T}(S_{\mathcal{A}}^{Bayes}(\alpha))$ is a conditional level α test and, for all a, $\mathcal{T}(S_{\mathcal{A}}^{Bayes}(\alpha) \cap \Omega_a)$ is a mean most powerful test on Ω_a. Conditional level α tests are also level α tests:

Author's personal copy

$$\Pr_{H_0}(\mathbf{N} \in \mathcal{S}_{\mathcal{A}}(\alpha)) = \sum_{a \in \mathcal{A}} \Pr_{H_0}(\mathbf{N} \in \mathcal{S}_{\mathcal{A}}(\alpha), \mathbf{N} \in \Omega_a)$$
$$= \sum_{a \in \mathcal{A}} \Pr_{H_0}(\mathbf{N} \in \mathcal{S}_{\mathcal{A}}(\alpha) | \mathbf{N} \in \Omega_a) \cdot \Pr_{H_0}(\mathbf{N} \in \Omega_a) \leq \sum_{a \in \mathcal{A}} \alpha \cdot \Pr_{H_0}(N \in \Omega_a) = \alpha.$$

when a assumes a single value, then $\mathcal{S}_{\mathcal{A}}^{Bayes}(\alpha) = \mathcal{S}^{Bayes}(\alpha)$. But in general, $\mathcal{T}(S_{\mathcal{A}}^{Bayes}(\alpha))$ is not a mean most powerful test and there may even be other conditional level α tests with smaller mean significance level and larger mean power. However, if $\mathcal{P}_0 = \{\mathbf{p}_0\}$ and $\Pr_{H_0}(\mathbf{N} \in \mathcal{S}_{\mathcal{A}}^{Bayes}(\alpha)|\ \mathbf{N} \in \Omega_a) = \alpha$ for all a, then as $\mathcal{T}(S_{\mathcal{A}}^{Bayes}(\alpha) \cap \Omega_a)$ is a mean most powerful test on Ω_a and the mean significance level identifies with the significance level, any other conditional level α test, $\mathcal{T}(S_{\mathcal{A}}(\alpha))$, would have smaller mean significance level than $\mathcal{T}(S_{\mathcal{A}}^{Bayes}(\alpha))$ on Ω_a and thus it would also have smaller mean power on Ω_a. Summing over all Ω_a, $\mathcal{T}(S_{\mathcal{A}}(\alpha))$ would have smaller mean power than $\mathcal{T}(S_{\mathcal{A}}^{Bayes}(\alpha))$.

2.2 Relation between our tests and Bayesian FDR controlling tests, Bayes factors, and likelihood ratio tests

Our methodology is closely related to the Bayesian FDR multiple hypotheses testing methodology. Bayesian FDR methodology assumes a two group mixture model, in which the parameter vector $\theta = (\theta_1 \cdots \theta_m)$ consists of iid dichotomous components $\theta_i \in \{0, 1\}$, with corresponding null hypotheses $H_i : \theta_i = 0$. The data $\mathbf{X} = (X_1 \cdots X_m)$ consists of independent components: for $\theta_i = 0$, X_i has null cdf G_0; for $\theta_i = 1$, X_i has non-null CDF for G_1. Storey (2007) and Sun and Cai (2007) show that the most powerful tests in the two group model are of the form: reject the null hypothesis if $\Pr(\theta_i = 0|X_i = x_i)$, the local FDR Efron et al. (2001), is smaller than some suitably selected threshold δ. Heller and Yekutieli (2014) extend this result to the case that θ_i are iid samples from a non-dichotomous distribution $\pi(\theta_i)$ and the null hypothesis is $H_i : \theta_i \in \mathcal{P}_0$, for $\mathcal{P}_0$ an arbitrary subset of the parameter space.

In Proposition 2.2, we further extend this result to the case that $\mathcal{P}_0$ is a subset $\mathcal{P} - \mathcal{P}_1$. For $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$, our statistic $\Pr(\theta_i = 1|X_i = x_i)$ is equal to one minus the local FDR. However, unlike the Bayesian FDR approach in which the marginal distribution of θ_i is used to determine a threshold δ that ensures Bayes FDR control, in our tests the parameter $\mathbf{p}$ is a single dependent multivariate realization, we do not assume that $\pi(p)$ is its marginal distribution, and we determine a threshold δ that ensures significance level $\leq \alpha$.

Expressing the statistic in (3)

$$\frac{\Pr(\mathbf{P} \in \mathcal{P}_1|\ \mathbf{N} = \mathbf{n})}{\Pr(\mathbf{P} \in \mathcal{P}_0|\ \mathbf{N} = \mathbf{n})} = \frac{\frac{\Pr(\mathbf{N}=\mathbf{n}|\ \mathbf{P}\in\mathcal{P}_1)\cdot\Pr(\mathbf{P}\in\mathcal{P}_1)}{\Pr(\mathbf{N}=\mathbf{n})}}{\frac{\Pr(\mathbf{N}=\mathbf{n}|\ \mathbf{P}\in\mathcal{P}_0)\cdot\Pr(\mathbf{P}\in\mathcal{P}_0)}{\Pr(\mathbf{N}=\mathbf{n})}} \propto \frac{\Pr(\mathbf{N} = \mathbf{n}|\ \mathbf{P} \in \mathcal{P}_1)}{\Pr(\mathbf{N} = \mathbf{n}|\ \mathbf{P} \in \mathcal{P}_0)} \tag{13}$$

reveals that we actually order the data points according to the Bayes factor between “model” $\mathcal{P}_1$ and “model” $\mathcal{P}_0$. However, note that in our tests the cutoff threshold

Author's personal copy

of the rejection region is not a nominal Bayes factor value (cf. Kass and Raftery 1995).

Our tests are also closely related to likelihood ratio tests. For simple hypotheses, $H_0 : \mathbf{p} = \mathbf{p}_0$ for $\mathbf{p}_0 \in \mathcal{P}_0$ versus $H_1 : \mathbf{p} = \mathbf{p}_1$ for $\mathbf{p}_1 \in \mathcal{P}_1$, our test reduces to the likelihood ratio test if $\mathcal{P}_0 = \{\mathbf{p}_0\}$ and $\mathcal{P}_1 = \{\mathbf{p}_1\}$, or if the prior distribution assigns all its probabilities to the two hypotheses: $\pi(\mathbf{p}_0) = \pi_0$ and $\pi(\mathbf{p}_1) = 1 - \pi_0$, for $0 < \pi_0 < 1$. The likelihood ratio statistic (Casella and Berger 2001) for testing the composite hypotheses $H_0 : \mathbf{p} \in \mathcal{P}_{null}$ versus $H_1 : \mathbf{p} \notin \mathcal{P}_{null}$ is

$$\Lambda(\mathbf{n}) = \frac{\sup_{\mathbf{p}\in\mathcal{P}_{null}} \Pr(\mathbf{N} = \mathbf{n}|\mathbf{p})}{\sup_{\mathbf{p}\in\mathcal{P}} \Pr(\mathbf{N} = \mathbf{n}|\mathbf{p})}.$$

For $\mathcal{P}_1 = \mathcal{P} - \mathcal{P}_{null}$, setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$ yields $\mathcal{P}_0 = \mathcal{P}_{null}$ and thus $\Lambda(\mathbf{n})$ orders the data points similarly to one minus our statistic, except that in our statistic we consider the average rather than the supremum of the likelihood, which according to our theoretical results yield tests with more power with respect to the prior distribution. However for $\mathcal{P}_1 \subset \mathcal{P} - \mathcal{P}_{null}$ and setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$, our statistic that orders the data points according to $\mathcal{P}_1$ yields considerably more powerful tests than $\Lambda(\mathbf{n})$ that orders the data points according to the null hypothesis, especially for the case that $\mathcal{P}_1$ is a "small" subset of $\mathcal{P} - \mathcal{P}_{null}$. We illustrate this in the following example and it occurs in the job satisfaction example where our tests yield considerably smaller p values than the χ^2 statistic, which is the likelihood ratio statistic for testing independence for cross-tabulated data.

Example 2.6 The parameter is $\mu = (\mu_1 \cdots \mu_K)$. The data are $\mathbf{Y} = (Y_1 \cdots Y_K)$ with $Y_k \sim N(\mu_k, 1)$ for $k = 1 \cdots K$. The null hypothesis is $H_0 : \mu = 0$ and $\mathcal{P}_1 = \{\mu : 0 \leq \mu_1\}$. The test statistic for the composite hypotheses likelihood ratio test for $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ is

$$\Lambda(\mathbf{y}) = \frac{\Pr(\mathbf{Y} = \mathbf{y}|\mu = 0)}{\sup_\mu \Pr(\mathbf{Y} = \mathbf{y}|\mu)} = \frac{\exp\{-\sum_{i=1}^K \frac{(y_i-0)^2}{1}\}}{\sup_\mu \exp\{-\sum_{i=1}^K \frac{(y_i-\mu_i)^2}{1}\}} = \exp\left(-\sum_{i=1}^K y_i^2\right).$$

Thus in the composite hypotheses likelihood ratio test, the data points are ordered according to their l_2 norm $\|\mathbf{y}\|$. Setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$ and using a flat prior for μ, as

$$\Pr(\mu \in \mathcal{P}_1|\mathbf{Y} = \mathbf{y}) = \Pr(0 \leq \mu_1|Y_1 = y_1) = \Phi(y_1),$$

our test sorts the data points according to y_1.

For $K = 100$ and $\mu_1 = (3.2, 0 \cdots 0)$, we compare the power of the level 0.05 tests based on the two statistics. But first of all, note that

$$\frac{\Pr(\mathbf{Y} = \mathbf{y}|\mu = \mu_1)}{\Pr(\mathbf{Y} = \mathbf{y}|\mu = 0)} = \frac{\exp\{-(y_1 - 3.2)^2\}}{\exp\{-(y_1 - 0)^2\}} = \exp(-3.2^2 + 2 \cdot 3.2 \cdot y_1).$$

Author's personal copy

In this case, our test is identical to the optimal likelihood ratio test for the simple null hypotheses $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_1$, which also sorts the data points according to y_1.

The level 0.05 composite hypotheses likelihood ratio test rejects the null hypothesis for $\mathbf{y}$ with $\chi^2_{100,0.95} \leq \|\mathbf{y}\|^2$, where $\chi^2_{100,0.95} = 124.34$ is the 0.95 quantile of the 100 degree of freedom χ^2 distribution. The power of this test is 0.179. For comparison, our $\alpha = 0.05$ test rejects the null hypothesis for $\mathbf{y}$ with $1.645 \leq y_1$, where 1.645 is the 0.95 quantile of the standard normal distribution. The power of our test is 0.940.

3 Job satisfaction example

3.1 Analysis of the job satisfaction data

The data in Table 1, taken from Agresti (2002, Table 2.8), correspond to a sample of 96 black males that were classified by income ("<1,500", "15,000–25,000", "25,000–40,000", ">40,000") and job satisfaction ("very dissatisfied", "little dissatisfied", "moderately satisfied", "very satisfied"). For $i = 1 \cdots 4$ and $j = 1 \cdots 4$, π_{ij} is the probability that a respondent has income level i and job satisfaction level j. We assume that the number of respondents $\mathbf{N} = (N_{11} \cdots \; N_{44})$ is multinomial with probabilities $\pi_{11} \cdots \; \pi_{44}$. n_{ij} is the observed number of respondents recorded in Table 1. The null hypothesis is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, for $\pi_{i+} = \pi_{i1} + \cdots + \pi_{i4}$ and $\pi_{+j} = \pi_{1j} + \cdots + \pi_{4j}$. A pair of respondents is concordant if they have different income and job satisfaction, and the respondent with higher income has higher job satisfaction. The probability that a pair of respondents is concordant is

$$\Pi_C = 2 \sum_i \sum_j \pi_{ij} \left(\sum_{i<h} \sum_{j<k} \pi_{hk} \right). \tag{14}$$

A pair of respondents are discordant if they have different incomes and job satisfaction, and the respondent with a higher income has lower job satisfaction. The probability that a pair of respondents are discordant is

Table 1 Job satisfaction data

Income (dollars)	Job satisfaction			
	Very dissatisfied	Little dissatisfied	Moderately satisfied	Very satisfied
<15,000	1	3	10	6
15,000–25,000	2	3	10	7
25,000–40,000	1	6	14	12
>40,000	0	1	9	11

Author's personal copy

$$\Pi_D = 2\sum_i \sum_j \pi_{ij} \left(\sum_{i<h} \sum_{k<j} \pi_{hk} \right). \tag{15}$$

The degree of concordance is measured by Kendall's gamma rank correlation coefficient, $\gamma = (\Pi_C - \Pi_D)/(\Pi_C + \Pi_D)$, which is the difference between the conditional probability of concordance and discordance given that the pair of respondents have different incomes and different job satisfaction.

We begin by testing H_0 with tests implemented in *R*, whose significance levels are based on parametric approximations of the test statistics' distribution under the null hypothesis. The first test is Pearson's Chi-squared test for count data implemented in the *chisq.test* function. The test statistic value for the observed data was 5.97 with 9 degrees of freedom and p value 0.743. The second test is a test for positive correlation between two ordinal variables implemented in the *cor.test* function. The statistic for this test is Kendall's tau. This is a sample value of γ, defined $\tau = (\hat{\Pi}_C - \hat{\Pi}_D)/\hat{\Pi}_{C+D}$, for $\hat{\Pi}_C$ and $\hat{\Pi}_D$ computed by replacing π_{ij} with $\hat{\pi}_{ij} = N_{ij}/96$ in (14) and (15) and an estimator of $\Pi_C + \Pi_D$, $\hat{\Pi}_{C+D}$. The statistic value for the observed data was $\tau = 0.1524$ with p value 0.0430.

To construct the exact tests, we condition on n_{i+} and n_{+j} the row and column sums of Table 1. There are 90,208,550 possible 4-by-4 tables with the same row and columns sums as in Table 1. Under the null hypothesis, the distribution of these tables is multivariate hypergeometric. The first exact test is based on Kendall's tau. In 21,101,151 tables, τ was greater than the value of τ in Table 1. The sum of the probabilities under H_0 of these tables was 0.0415.

For our Bayesian statistics, we use a Dirichlet prior distribution with concentration parameters $(0.5 \cdots\ 0.5)$ for $(\pi_{11} \cdots\ \pi_{44})$, for which the posterior probability is a Dirichlet distribution with concentration parameters $(N_{11} + 0.5 \cdots\ N_{44} + 0.5)$. For the two statistics, we set $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$. To compute the posterior probability of $\mathcal{P}_1$ for a given table, we sample $(\pi_{11}, \cdots \pi_{44})$ from the posterior probability and record the proportion of times $\mathcal{P}_1$ occurred.

Our first Bayesian statistic is the posterior probability of the concordance event, $\mathcal{P}_1^{Cncrd} = \{(\pi_{11} \cdots \pi_{44}) :\ 0 \leq \gamma\}$. The probability of concordance for $N_{ij} = n_{ij}$, based on a sample of 10^7 draws from the posterior, was 0.9564 (s.e. $<$ 0.0001). Computing this statistic for all 4-by-4 tables is too time consuming. Thus to assess the significance level for this statistic, we generated a sample of 50,000 4-by-4 contingency tables from the multivariate hypergeometric null distribution, and for each contingency table we sampled 10,000 $(\pi_{11}, \cdots \pi_{44})$ from the posterior probability and recorded the proportion of times the concordance event occurred. The estimated significance level was 0.036 (s.e. $<$ 0.001), which is the proportion of contingency tables with estimated probability of concordance $\geq$ 0.9564.

Our second Bayesian statistic is the posterior probability that income and job satisfaction are positively dependent. This is a stronger property than concordance that corresponds to the event

$$\mathcal{P}_1^{Pos} = \{(\pi_{11}, \ldots, \pi_{44}) :\ \Pr(\pi_{j|i} \leq t) \geq \Pr(\pi_{j|i+1} \leq t)\ \ \forall t, \forall j, \forall i\}, \tag{16}$$

Author's personal copy

for $\pi_{j|i} = \pi_{ij}/\pi_{i+}$. Based on a sample of 10^7 draws, the posterior probability of positive dependence for the observed table is 0.0118 (s.e. < 0.0001). And again, to assess the significance level for this statistic, we sampled 50,000 4-by-4 contingency tables from the multivariate hypergeometric null distribution and for each contingency table we sampled 10,000 $(\pi_{11}, \cdots \pi_{44})$ from the posterior probability. The estimated significance level was 0.0093 (s.e. < 0.001), which is the proportion of contingency tables with posterior probability of positive dependence ≥ 0.0118.

3.2 Comparison between the different test statistics

To apply the exact tests in the previous section, we computed the test statistic values on a null sample of 4-by-4 contingency tables. The similar significance levels of the tests suggest that the statistics induce similar orderings on the data sample space. To verify this, we generated a null sample of 2,000 of 4-by-4 contingency tables. For each contingency table, we compute Kendall's tau statistic, the posterior probability of the positive dependence event $\mathcal{P}_1^{Pos}$ for a Dirichlet prior with concentration parameter $(0.5 \cdots 0.5)$, and the posterior probability of the concordance event $\mathcal{P}_1^{Cncrd}$ for a Dirichlet prior with concentration parameter $(0.5 \cdots 0.5)$ and for a Dirichlet prior with concentration parameter $(5 \cdots 5)$.

Figure 1 is a scatter plot of the posterior probability of concordance and Kendall's tau statistic. The plot reveals that the two statistics that measure the degree of concordance induce almost identical ordering to the data sample space. Figure 2 is a scatter plot of the posterior probability of concordance and the posterior probability of dependence. Notice that positive dependence is a relatively rare event—in more than 90 % of the null contingency tables the estimated probability of positive dependence was 0 (recall: we assess this probability by counting the number of occurrences of $\mathcal{P}_1^{Pos}$ in 10,000 posterior samples of $(\pi_{11} \cdots \pi_{44})$). Nonetheless, the plot is consistent with the fact that $\mathcal{P}_1^{Pos} \subset \mathcal{P}_1^{Cncrd}$: the tables with small posterior probability of concordance also have small posterior probability of positive dependence, tables with large posterior probability of positive dependence have large posterior probability of concordance, but there are also tables with small posterior probability of positive dependence and large posterior probability of concordance. Figure 3 compares the posterior probability of concordance for a Dirichlet prior with concentration parameter $(0.5 \cdots 0.5)$ and for a Dirichlet prior with concentration parameter $(5 \cdots 5)$. Interestingly, the median posterior probability of concordance is greater than 0.5: 0.515 for concentration parameter 0.5 and 0.553 for concentration parameter 5. The plot also reveals that even though the concordance probabilities for the two priors are very different (the shrinkage toward the median is much stronger for concentration parameter 5), the two statistics induce very similar ordering on the data sample space. Lastly, for the observed data the posterior of concordance for concentration parameter 5 is 0.8844 corresponding to an exact p value of 0.0412, which is a slightly less significant result than the concentration parameter 0.5 result: posterior probability 0.9564 for the observed data corresponding to a p value of 0.036.

Author's personal copy

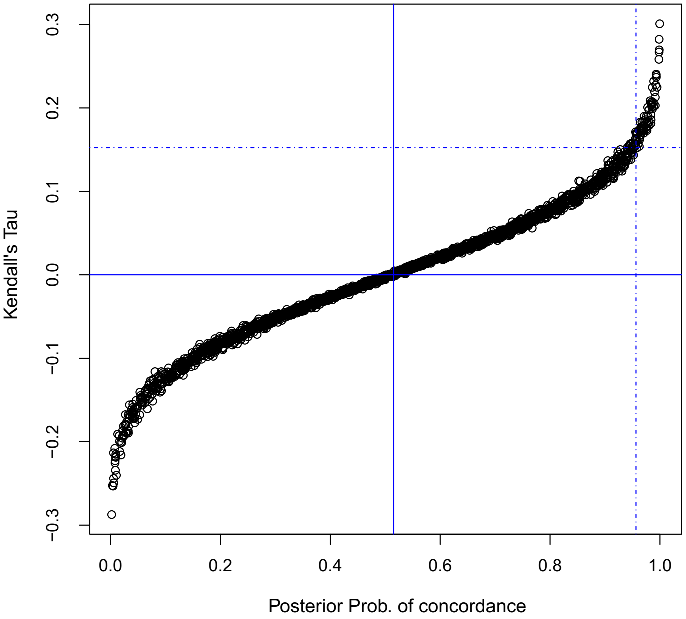


Fig. 1 Comparison between Kendall's tau statistic and the posterior probability of concordance. The plot displays Kendall's tau statistic (Y-axis) and the posterior probability of concordance for Dirichlet prior distribution with concentration parameter 0.5 (X-axis) for 2,000 contingency tables sampled under the null hypothesis. The *horizontal* and *vertical solid blue lines* are drawn at the median of Kendall's tau statistic values and the median posterior probability of concordance. The *horizontal* and *vertical dotted-dashed blue lines* are drawn at the Kendall's tau statistic value and posterior probability of concordance for the observed data (color figure online)

3.3 Job satisfaction simulation

The simulation compares the power of the conditional exact test based on Kendall's tau statistic with the conditional exact test whose test statistic is the posterior probability of concordance on

$$\Omega_a = \{(N_{11} \cdots \ N_{44}) : \ N_{1+} = n_{1+}, N_{2+} = n_{2+}, \ldots, N_{+4} = n_{+4}\}, \qquad (17)$$

for which the null distribution of $\mathbf{N}$ is the multivariate hypergeometric considered in the previous sections. The alternative distribution is that $\mathbf{N}$ is $multinomial$ $(\hat{\pi}_{11} \cdots \ \hat{\pi}_{44})$, with $\hat{\pi}_{ij} = n_{ij}/96$ truncated to Ω_a in (17).

We use the following importance sampling scheme to generate samples of $\mathbf{N}$ from the alternative distribution. We sample 10^6 proposal realizations of $\mathbf{N}$ from the multivariate hypergeometric null distribution; for each proposal realization, we compute a sampling weight that is the probability of observing this realization under the alternative multinomial distribution divided by the probability of observing this realization under the multivariate hypergeometric null distribution. We use weighted with-replacement sampling of the 10^6 proposal values to generate a sample of 10^5 realizations from the alternative distribution.

Author's personal copy

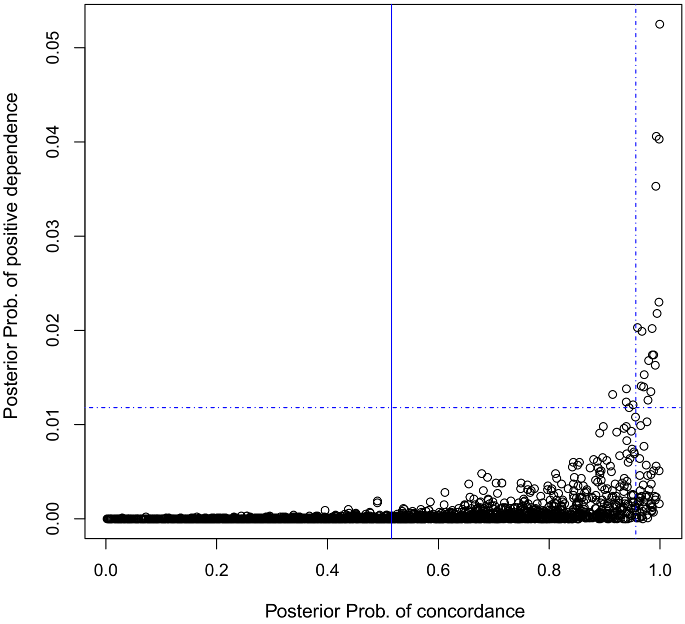


Fig. 2 Comparison between the posterior probability of positive dependence and the posterior probability of concordance. The plot displays the posterior probability of positive dependence (Y-axis) and the posterior probability of concordance (X-axis) for the same Dirichlet prior distribution with concentration parameter 0.5 for 2,000 contingency tables sampled under the null hypothesis. The *vertical solid blue line* is drawn at the median posterior probability of concordance. The *horizontal* and *vertical dotted-dashed blue lines* are drawn at the posterior probabilities of positive dependence and concordance for the observed data (color figure online)

We compute the two test statistic values for each of the 10^5 realizations of $\mathbf{N}$ from the alternative distribution. To assess the significance level of each alternative distribution realization for the two test statistics, we generate another sample of 10^5 realizations of $(N_{11} \cdots N_{44})$ from the null distribution and compute the two test statistic value for each null realization. The p values assigned to each alternative distribution realization is the proportion of null realizations for which the statistic values were larger than the alternative realization's statistic values.

Recall that for the Table 1 data, the p value for the exact test based on Kendall's tau statistic was 0.0415 and the p value for the exact test for the probability of concordance was 0.036. In the 10^5 simulated alternative distribution realizations, the p values computed for the probability of concordance statistic were also slightly smaller than the p values computed for Kendall's tau statistic. For Kendall's tau statistic, the mean p value was 0.0988 and the median p value was 0.0399; 0.679 (s.e. < 0.005) of the p values were smaller than 0.10, and 0.537 (s.e. < 0.005) of were smaller than 0.05. However, for the p values computed based on the probability of concordance statistics, the mean p value was 0.0947 and the median p value was 0.0370; 0.701 (s.e. < 0.005) of the p-values were smaller than 0.10 and 0.550 (s.e. < 0.005) were smaller than 0.05.

Author's personal copy

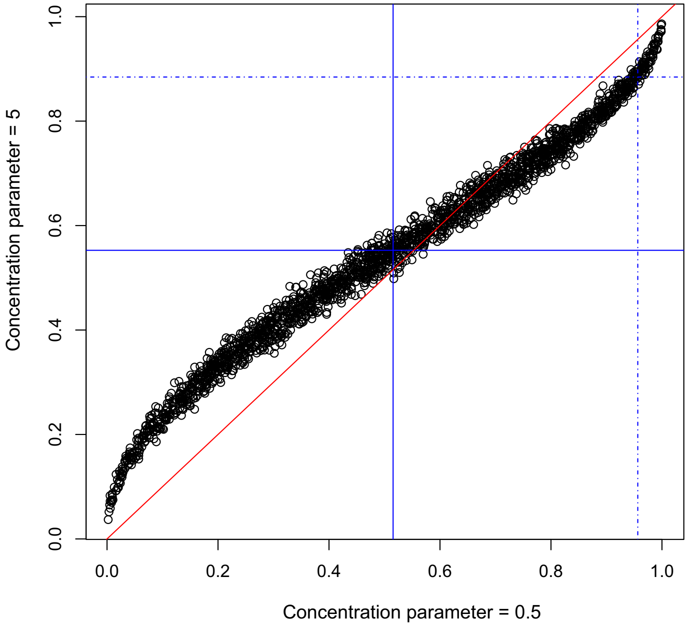


Fig. 3 The effect of the prior distribution on the posterior probability concordance. The plot displays the posterior probability of concordance for Dirichlet prior distribution with concentration parameter 5 (Y-axis) and the posterior probability of concordance for Dirichlet prior distribution with concentration parameter 0.5 (X-axis), for 2,000 contingency tables sampled under the null hypothesis. The *horizontal* and *vertical solid blue lines* are drawn at the median posterior probabilities. The *horizontal* and *vertical dotted-dashed blue lines* are the posterior probabilities for the observed data. The *red diagonal* is $X = Y$ (color figure online)

Table 2 Death penalty data

Victim	Defendant	Death penalty	No death penalty
White	White	19	132
	Black	11	52
Black	White	0	9
	Black	6	97

4 Death penalty example

Table 2 displays data from a study on death penalty in Florida (Agresti 2002, Table 2.13). The 326 subjects classified in Table 2 were the defendants in indictments involving cases with multiple murders in Florida. The goal of the analysis is to determine whether the probability of receiving death sentence depends on the defendant's race.

The variables are X—race of victim (“white”, “black”); Y—race of defendant (“white”, “black”)’), and z—death penalty verdict (“yes”, “no”). π_{ijk} is the probability that X takes on its ith value, Y takes on its jth value, and Z takes on its kth value. The conditional odds ratio between the defendant's race and death penalty for white

Author's personal copy

victims is $\theta_{YZ|X=1} = (\pi_{111} \cdot \pi_{122})/(\pi_{112} \cdot \pi_{121})$ and for black victims it is $\theta_{YZ|X=2} = (\pi_{211} \cdot \pi_{222})/(\pi_{212} \cdot \pi_{221})$. The marginal odds ratio between defendant's race and death penalty is $\theta_{YZ} = (\pi_{+11} \cdot \pi_{+22})/(\pi_{+12} \cdot \pi_{+21})$, for $\pi_{+jk} = \pi_{1jk} + \pi_{2jk}$. Similarly, θ_{XZ} is the marginal odds ratio between the victim's race and death penalty and θ_{XY} is the marginal odds ratio between the defendant's race and death penalty.

We used the *R fisher.test* function to test dependency between the pairs of variables. Defendant race and victim race are highly dependent, $\hat{\theta}_{XY} = 27.1$ with 0.95 CI [12.7, 64.8]; and risk of receiving death penalty is higher for white victims than for black victims, $\hat{\theta}_{XZ} = 2.87$ with 0.95 CI [1.13, 8.73]. Thus the victim's race is a confounder: white defendants have higher probability of receiving death penalty just because they are more likely to kill a white victim. Indeed, we see that $\hat{\theta}_{YZ} = 1.18$ with 0.95 CI [0.56, 2.52]. The null hypothesis we consider is that conditional on victim's race, defendant's race and death penalty are independent, $H_0 : \ \theta_{YZ|X=1} = 1, \ \theta_{YZ|X=2} = 1$. The alternative hypothesis is that the following Simpson's Paradox occurs, $H_1 : \ \theta_{YZ|X=1} < 1, \ \theta_{YZ|X=2} < 1, \ 1 < \theta_{YZ}$.

To test the null hypothesis, for white victims we further condition on the observed values $N_{11+} = 151$, $N_{12+} = 63$, $N_{1+1} = 30$, $N_{1+2} = 184$, and for black victims on the observed values $N_{21+} = 9$, $N_{22+} = 103$, $N_{2+1} = 6$, $N_{2+2} = 106$. We form a conditional sample space with 217 points that can be expressed as

$$\Omega_a = \{(N_{111}, N_{211}) : \ N_{111} \in (0, 1, \ldots, 30), \ N_{211} \in (0, 1, \ldots, 6)\}.$$

The observed data point is ($N_{111} = 19, N_{211} = 0$). Under H_0, N_{111} and N_{211} are independent hypergeometric random variables. Applying the R *fisher.test* function to the observed 2-by-2 tables corresponding to white and black victims yields $\hat{\theta}_{YZ|X=1} = 0.68$ with 0.95 CI [0.28, 1.70] and $\hat{\theta}_{YZ|X=2} = 0$ with 0.95 CI [0, 10.72]. To construct an exact test, the 217 data sample points are ordered according to a statistic that quantifies their strength of evidence in favor of Simpson's Paradox, and then the exact significance level of the observed table is the sum of the probabilities of the data points with greater or equal test statistic values. However, as Simpson's Paradox involves effects having conflicting signs, determining the strength of evidence in favor of Simpson's Paradox is difficult. For example, does data point (20, 0) with larger or equal conditional associations ($\hat{\theta}_{YZ|X=1} = 0.810, \hat{\theta}_{YZ|X=2} = 0$) and larger marginal ($\hat{\theta}_{YZ} = 1.34$) association offer more evidence in favor of Simpson's Paradox than the observed data point?

The statistic we propose for ordering the points in the data sample space is the posterior probability of the event corresponding to H_1

$$\mathcal{P}_1 = \{(\pi_{111} \cdots \pi_{222}) : \ \theta_{YZ|X=1} < 1, \ \theta_{YZ|X=2} < 1, \ 1 < \theta_{YZ}\}.$$

For our analysis we use a Dirichlet prior with concentration parameters $(0.5 \cdots \ 0.5)$. Thus for data point $(N_{111} \cdots \ N_{222})$, the posterior distribution of $(\pi_{111} \cdots \ \pi_{222})$ is Dirichlet with concentration parameters $(N_{111} + 0.5 \cdots \ N_{222} + 0.5)$. To compute the probability of $\mathcal{P}_1$ for a given data point, we sample $(\pi_{111}, \cdots \pi_{222})$ from the posterior probability and count the proportion of samples that either events occurred.

Author's personal copy

Based on 2×10^6 samples from the posterior distribution, data point $(20, 0)$ with $\Pr_{H_0}(20, 0) = 0.087$ has the largest posterior probability of $\mathcal{P}_1$, 0.085954 (s.e. $<$ 0.0001); the observed table with $\Pr_{H_0}(19, 0) = 0.064$ has the second largest posterior probability, 0.0797 (s.e. < 0.0001); data point $(21, 0)$ with $\Pr_{H_0}(21, 0) = 0.101$ has the third largest posterior probability, 0.0795 (s.e. < 0.0001). Thus the significance level of the observed table is $0.151 = 0.087 + 0.064$.

5 Discussion

We applied our methodology to construct exact tests for cross-tabulated data in which the Bayesian machinery is relatively straightforward and the computational burden was not too heavy. In the job satisfaction example, computing the test statistic value by sampling 100,000 posterior realizations takes 2–3 s. Determining whether the test rejects the null hypothesis at level 0.05 requires a few hundred null samples and takes several minutes.

Our main message in this paper is that in cases where it is not clear how to construct a test statistic for a composite alternative hypothesis, instead of using the likelihood ratio test for composite hypotheses, try implementing our methodology even though it will require setting up new Bayesian machinery and may be computationally difficult.

An aspect of our methodology that we had not explored that may be relevant in other cases is experimenting with different choices of $\mathcal{P}_0$. Our suggestion is to begin with setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$ and, similarly to what we did in Example 26, specify a non-null parameter value $\mathbf{p}_1 \in \mathcal{P}_1$ and compare our mean most powerful test to the simple hypotheses likelihood ratio test. Consider a different choice of $\mathcal{P}_0$ only if the mean most powerful test is very different from the likelihood ratio test.

References

Agresti A (2002) Categorical data analysis. Wiley, New Jersey

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Casella G, Berger RL (2001) Statistical inference. Duxbury Press, Belmont

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Heller R, Yekutieli D (2014) Replicability analysis for genome-wide association studies. Ann Appl Stat 8:481–498

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Storey JD (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc B 69:347–368

Sun W, Cai TT (2007) Oracle and adaptive compound decision rules for false discovery rate control. J Am Stat Assoc 102:901–912