Finite sample inference for empirical Bayesian methods

Hien Nguyen, Mayetri Gupta

Finite sample inference for empirical Bayesian methods

Hien Duy Nguyen[1,2] Mayetri Gupta[3]

[1] School of Mathematics and Physics, University of Queensland, St. Lucia 4067

[2] Department of Mathematics and Statistics, La Trobe University, Bundoora 3086

[3] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ

Abstract

In recent years, empirical Bayesian (EB) inference has become an attractive approach for estimation in parametric models arising in a variety of real-life problems, especially in complex and high-dimensional scientific applications. However, compared to the relative abundance of available general methods for computing point estimators in the EB framework, the construction of confidence sets and hypothesis tests with good theoretical properties remains difficult and problem specific. Motivated by the universal inference framework of Wasserman et al. [2020], we propose a general and universal method, based on holdout likelihood ratios, and utilizing the hierarchical structure of the specified Bayesian model for constructing confidence sets and hypothesis tests that are finite sample valid. We illustrate our method through a range of numerical studies and real data applications, which demonstrate that the approach is able to generate useful and meaningful inferential statements in the relevant contexts.

1 Introduction

Let $\mathbf{D}_n = (\boldsymbol{X}_i)_{i\in[n]}$ be our data, presented as a sequence of $n \in \mathbb{N} = \{1, 2, \dots\}$ random variables $\boldsymbol{X}_i \in \mathbb{X}$ ($i \in [n] = \{1, \dots, n\}$). For each $i \in [n]$, let $\boldsymbol{\Theta}_i \in \mathbb{T}$ be a random variable with probability density function (PDF) $\pi(\boldsymbol{\theta}_i; \boldsymbol{\psi})$, where $\boldsymbol{\psi} \in \mathbb{P}$ is a hyperparameter. Furthermore, suppose that $[\boldsymbol{X}_i|\boldsymbol{\Theta}_i = \boldsymbol{\theta}_i]$ arises from a family of data generating processes (DGPs) with conditional PDFs

$$f(\boldsymbol{x}_i|\boldsymbol{\Theta}_i = \boldsymbol{\theta}_i) = f(\boldsymbol{x}_i|\boldsymbol{\theta}_i),$$

and that the sequence $((\boldsymbol{X}_i, \boldsymbol{\Theta}_i))_{i\in[n]}$ is independent.

Suppose that $(\boldsymbol{\Theta}_i)_{i\in[n]}$ is realized at $\boldsymbol{\vartheta}_n^* = (\boldsymbol{\theta}_i^*)_{i\in[n]}$, where each realization $\boldsymbol{\theta}_i^*$ ($i \in [n]$) is unknown, and where $\boldsymbol{\psi}$ is also unknown. Let $\mathbb{I} \subset [n]$, and write $\boldsymbol{\vartheta}_{\mathbb{I}}^* = (\boldsymbol{\theta}_i^*)_{i\in\mathbb{I}}$. When $\mathbb{I} = \{i\}$, we shall use the

shorthand $\mathbb{I} = i$, where it causes no confusion.

Under this setup, for significance level $\alpha \in (0,1)$, we wish to draw inference regarding the realized sequence $\boldsymbol{\vartheta}_n^*$ by way of constructing $100\,(1-\alpha)\,\%$ confidence sets $\mathcal{C}_i^\alpha\,(\mathbf{D}_n)$ that satisfy:

$$\Pr_{\boldsymbol{\theta}_i^*}\left[\boldsymbol{\theta}_i^* \in \mathcal{C}_i^\alpha\,(\mathbf{D}_n)\right] \geq 1-\alpha, \tag{1}$$

and p-values $P_{\mathbb{I}}\,(\mathbf{D}_n)$ for testing null hypotheses $\mathrm{H}_0 : \boldsymbol{\vartheta}_{\mathbb{I}}^* \in \mathbb{T}_{\mathbb{I},0} \subset \mathbb{T}^{|\mathbb{I}|}$ that satisfy:

$$\sup_{\boldsymbol{\vartheta}_{\mathbb{I}}^* \in \mathbb{T}_{\mathbb{I},0}} \Pr_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}\left[P_{\mathbb{I}}\,(\mathbf{D}_n) \leq \alpha\right] \leq \alpha, \tag{2}$$

where $\Pr_{\boldsymbol{\theta}_i^*}$ and $\Pr_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}$ denote probability measures consistent with the PDF $f\,(\boldsymbol{x}_i|\boldsymbol{\theta}_i^*)$, for each $i \in [n]$, and for all $i \in \mathbb{I}$, respectively. That is, for a measurable set $\mathcal{A} \subset \mathbb{X}^n$, and assuming absolute continuity of $\Pr_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}$ with respect to some measure $\mathfrak{m}$ (typically the Lebesgue or counting measure), we can write

$$\Pr_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}(\mathcal{A}) = \int_{\mathcal{A}} \prod_{i \in \mathbb{I}} f\,(\boldsymbol{x}_i|\boldsymbol{\theta}_i^*) \prod_{j \notin \mathbb{I}} f\,(\boldsymbol{x}_j|\boldsymbol{\theta}_j)\,\mathrm{d}\mathfrak{m}\,(\mathbf{d}_n)\,, \tag{3}$$

where $\boldsymbol{\theta}_j$ is an arbitrary element of $\mathbb{T}$, for each $j \notin \mathbb{I}$.

The setup above falls within the framework of empirical Bayesian (EB) inference, as exposited in the volumes of Maritz and Lwin [1989], Ahmed and Reid [2001], Serdobolskii [2008], Efron [2010], and Bickel [2020]. Over the years, there has been a sustained interest in the construction and computation of EB point estimators for $\boldsymbol{\vartheta}_n^*$, in various contexts, with many convenient and general computational tools now made available, for instance, via the software of Johnstone and Silverman [2005], Leng et al. [2013], Koenker and Gu [2017], and Narasimhan and Efron [2020]. Unfortunately, the probabilistic properties of $\boldsymbol{\vartheta}_n^*$ tend to be difficult to characterize, making the construction of confidence sets and hypothesis tests with good theoretical properties relatively less routine than the construction of point estimators. When restricted to certain classes of models, such constructions are nevertheless possible, as exemplified by the works of Casella and Hwang [1983], Morris [1983a], Laird and Louis [1987], Datta et al. [2002], Tai and Speed [2006], Hwang et al. [2009], Hwang and Zhao [2013], and Yoshimori and Lahiri [2014], among others.

In this work, we adapt the universal inference framework of Wasserman et al. [2020] to produce valid confidence sets and p-values with properties (1) and (2), respectively, for arbitrary estimators of $\boldsymbol{\vartheta}_n^*$. As with the constructions of Wasserman et al. [2020], the produced inferential methods are all valid for finite sample size n and require no assumptions beyond correctness of model specification. The confidence sets and p-values arise by construction of holdout likelihood ratios that can be demonstrated to have the e-value property, as described in Vovk and Wang [2021] (see also the s-values of Grunwald

et al., 2020 and the betting values of Shafer, 2021). Here, we are able to take into account the hierarchical structure of the Bayesian specified model by using the fact that parameterized e-values are closed when averaged with respect to an appropriate probability measure (cf. Vovk, 2007 and Kaufmann and Koolen, 2018). Due to the finite sample correctness of our constructions, we shall refer to our methods as finite sample EB (FSEB) techniques.

Along with our methodological developments, we also demonstrate the application of our FSEB techniques in numerical studies and real data applications. These applications include the use of FSEB methods for constructing confidence intervals (CIs) for the classic mean estimator of Stein [1956], and testing and CI construction in Poisson–gamma models and Beta–binomial models, as per Koenker and Gu [2017] and Hardcastle and Kelly [2013], respectively. Real data applications are demonstrated via the analysis of insurance data from Haastrup [2000] and differential methylation data from Cruickshanks et al. [2013]. In these real and synthetic applications, we show that FSEB methods, satisfying conditions (1) and (2), are able to generate useful and meaningful inferential statements.

We proceed as follows. In Section 2, we introduce the confidence set and p-value constructions for drawing inference regarding EB models. In Section 3, numerical studies of simulated data are used to demonstrate the applicability and effectiveness of FSEB constructions. In Section 4, FSEB methods are applied to real data to further show the practicality of the techniques. Lastly, in Section 5, we provide discussions and conclusions regarding our results.

2 Confidence sets and hypothesis tests

We retain the notation and setup from Section 1. For each subset $\mathbb{I} \subset [n]$, let us write $\mathbf{D}_{\mathbb{I}} = (\boldsymbol{X}_i)_{i\in\mathbb{I}}$ and $\overline{\mathbf{D}}_{\mathbb{I}} = (\boldsymbol{X}_i)_{i\in[n]\setminus\mathbb{I}}$.

Suppose that we have available some estimator of $\boldsymbol{\psi}$ that only depends on $\overline{\mathbf{D}}_{\mathbb{I}}$ (and not $\mathbf{D}_{\mathbb{I}}$), which we shall denote by $\hat{\boldsymbol{\psi}}_{\mathbb{I},n}$. Furthermore, for fixed $\boldsymbol{\psi}$, write the integrated and unintegrated likelihood of the data $\mathbf{D}_{\mathbb{I}}$, as

$$L_{\mathbb{I}}(\boldsymbol{\psi}) = \prod_{i\in\mathbb{I}} \int_{\mathbb{T}} f(\boldsymbol{X}_i|\boldsymbol{\theta}_i)\, \pi(\boldsymbol{\theta}_i;\boldsymbol{\psi})\, \mathrm{d}\mathfrak{n}(\boldsymbol{\theta}_i) \tag{4}$$

and

$$l_{\mathbb{I}}(\boldsymbol{\vartheta}_{\mathbb{I}}) = \prod_{i\in\mathbb{I}} f(\boldsymbol{X}_i|\boldsymbol{\theta}_i), \tag{5}$$

respectively, where $\boldsymbol{\vartheta}_{\mathbb{I}} = (\boldsymbol{\theta}_i)_{i\in\mathbb{I}}$ (here, $\boldsymbol{\vartheta}_{\{i\}} = \boldsymbol{\theta}_i$). We note that in (4), we have assumed that $\pi(\cdot;\boldsymbol{\psi})$ is a density function with respect to some measure on $\mathbb{T}$, $\mathfrak{n}$.

Define the ratio statistic:

$$R_{\mathbb{I},n}\left(\boldsymbol{\vartheta}_{\mathbb{I}}\right)=L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)/l_{\mathbb{I}}\left(\boldsymbol{\vartheta}_{\mathbb{I}}\right), \tag{6}$$

and consider sets of the form

$$\mathcal{C}_{i}^{\alpha}\left(\mathbf{D}_{n}\right)=\left\{\boldsymbol{\theta}\in\mathbb{T}:R_{i,n}\left(\boldsymbol{\theta}\right)\leq 1/\alpha\right\}.$$

The following Lemma is an adaptation of the main idea of Wasserman et al. [2020] for the context of empierical Bayes estimators, and allows us to show that $\mathcal{C}_{i}^{\alpha}\left(\mathbf{D}_{n}\right)$ satisfies property (1).

Lemma 1. *For each $\mathbb{I}\subset[n]$ and fixed sequence $\boldsymbol{\vartheta}_{n}^{*}\in\mathbb{T}^{n}$, $\mathrm{E}_{\boldsymbol{\vartheta}_{\mathbb{I}}^{*}}\left[R_{\mathbb{I},n}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^{*}\right)\right]=1$.*

Proof. Let $\mathbf{d}_{\mathbb{I}}$ and $\bar{\mathbf{d}}_{\mathbb{I}}$ be realizations of $\mathbf{D}_{\mathbb{I}}$ and $\overline{\mathbf{D}}_{\mathbb{I}}$, respectively. Then, using (3), write

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}_{\mathbb{I}}^{*}}\left[R_{\mathbb{I},n}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^{*}\right)\right] &= \int_{\mathbb{X}^{n}} R_{\mathbb{I},n}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^{*}\right)\prod_{i\in\mathbb{I}}f\left(\boldsymbol{x}_{i}|\boldsymbol{\theta}_{i}^{*}\right)\prod_{j\notin\mathbb{I}}f\left(\boldsymbol{x}_{j}|\boldsymbol{\theta}_{j}\right)\mathrm{d}\mathfrak{m}\left(\mathbf{d}_{n}\right)\\
&\underset{\text{(i)}}{=}\int_{\mathbb{X}^{n-|\mathbb{I}|}}\int_{\mathbb{X}^{|\mathbb{I}|}}\frac{L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)}{l_{\mathbb{I}}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^{*}\right)}\prod_{i\in\mathbb{I}}f\left(\boldsymbol{x}_{i}|\boldsymbol{\theta}_{i}^{*}\right)\mathrm{d}\mathfrak{m}\left(\mathbf{d}_{\mathbb{I}}\right)\prod_{j\notin\mathbb{I}}f\left(\boldsymbol{x}_{j}|\boldsymbol{\theta}_{j}\right)\mathrm{d}\mathfrak{m}\left(\bar{\mathbf{d}}_{\mathbb{I}}\right)\\
&\underset{\text{(ii)}}{=}\int_{\mathbb{X}^{n-|\mathbb{I}|}}\int_{\mathbb{X}^{|\mathbb{I}|}}L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)\mathrm{d}\mathfrak{m}\left(\mathbf{d}_{\mathbb{I}}\right)\prod_{j\notin\mathbb{I}}f\left(\boldsymbol{x}_{j}|\boldsymbol{\theta}_{j}\right)\mathrm{d}\mathfrak{m}\left(\bar{\mathbf{d}}_{\mathbb{I}}\right)\\
&\underset{\text{(iii)}}{=}\int_{\mathbb{X}^{n-|\mathbb{I}|}}\prod_{j\notin\mathbb{I}}f\left(\boldsymbol{x}_{j}|\boldsymbol{\theta}_{j}\right)\mathrm{d}\mathfrak{m}\left(\bar{\mathbf{d}}_{\mathbb{I}}\right)\\
&\underset{\text{(iv)}}{=}1.
\end{aligned}$$

Here, (i) is true by definition of (6), (ii) is true by definition of (5), (iii) is true by the fact that (4) is a probability density function on $\mathbb{X}^{|\mathbb{I}|}$, with respect to $\mathfrak{m}$, and (iv) is true by the fact that $\prod_{j\notin\mathbb{I}}f\left(\boldsymbol{x}_{j}|\boldsymbol{\theta}_{j}\right)$ is a probability density function on $\mathbb{X}^{n-|\mathbb{I}|}$, with respect to $\mathfrak{m}$. □

Proposition 1. *For each $i\in[n]$, $\mathcal{C}_{i}^{\alpha}\left(\mathbf{D}_{n}\right)$ is a $100\left(1-\alpha\right)\%$ confidence set, in the sense that*

$$\mathrm{Pr}_{\boldsymbol{\theta}_{i}^{*}}\left[\boldsymbol{\theta}_{i}^{*}\in\mathcal{C}_{i}^{\alpha}\left(\mathbf{D}_{n}\right)\right]\geq 1-\alpha.$$

Proof. For each i, Markov's inequality states that

$$\mathrm{Pr}_{\boldsymbol{\theta}_{i}^{*}}\left[R_{i,n}\left(\boldsymbol{\theta}_{i}^{*}\right)\geq 1/\alpha\right]\leq\alpha\mathrm{E}_{\boldsymbol{\theta}_{i}^{*}}\left[R_{i,n}\left(\boldsymbol{\theta}_{i}^{*}\right)\right]=\alpha,$$

which implies that

$$\mathrm{Pr}_{\boldsymbol{\theta}_{i}^{*}}\left[\boldsymbol{\theta}_{i}^{*}\in\mathcal{C}_{i}^{\alpha}\left(\mathbf{D}_{n}\right)\right]=\mathrm{Pr}_{\boldsymbol{\theta}_{i}^{*}}\left[R_{i,n}\left(\boldsymbol{\theta}_{i}^{*}\right)\leq 1/\alpha\right]\geq 1-\alpha$$

by Lemma 1. □

Next, we consider the testing of null hypotheses H_0: $\boldsymbol{\vartheta}_{\mathbb{I}}^* \in \mathbb{T}_{\mathbb{I},0}$ against an arbitrary alternative H_1: $\boldsymbol{\vartheta}_{\mathbb{I}}^* \in \mathbb{T}_{\mathbb{I},1} \subseteq \mathbb{T}^{|\mathbb{I}|}$. To this end, we define the maximum unintegrated likelihood estimator of $\boldsymbol{\vartheta}_{\mathbb{I}}^*$, under H_0 as

$$\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}} \in \left\{ \tilde{\boldsymbol{\vartheta}}_{\mathbb{I}} \in \mathbb{T}_{\mathbb{I},0} : l_{\mathbb{I}}\left(\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\right) = \sup_{\boldsymbol{\vartheta}_{\mathbb{I}} \in \mathbb{T}_{\mathbb{I},0}} l_{\mathbb{I}}\left(\boldsymbol{\vartheta}_{\mathbb{I}}\right) \right\}. \tag{7}$$

Using (7), and again letting $\hat{\boldsymbol{\psi}}_{\mathbb{I},n}$ be an arbitrary estimator of $\boldsymbol{\psi}$, depending only on $\overline{\mathbf{D}}_{\mathbb{I}}$, we define the ratio test statistic

$$T_{\mathbb{I}}\left(\mathbf{D}_n\right) = L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right) / l_{\mathbb{I}}\left(\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\right).$$

The following result establishes the fact that the p-value $P_{\mathbb{I}}\left(\mathbf{D}_n\right) = 1/T_{\mathbb{I}}\left(\mathbf{D}_n\right)$ has the correct size, under H_0.

Proposition 2. *For any* $\alpha \in (0,1)$ *and* $\boldsymbol{\vartheta}_{\mathbb{I}}^* \in \mathbb{T}_{\mathbb{I},0}$, $\mathrm{Pr}_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}\left[P_{\mathbb{I}}\left(\mathbf{D}_n\right) \leq \alpha\right] \leq \alpha$.

Proof. Assume that $\boldsymbol{\vartheta}_{\mathbb{I}}^* \in \mathbb{T}_{\mathbb{I},0}$. By Markov's inequality, we have

$$\begin{aligned}
\mathrm{Pr}_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}\left[T_{\mathbb{I}}\left(\mathbf{D}_n\right) \geq 1/\alpha\right] &\leq \alpha \mathrm{E}_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}\left[T_{\mathbb{I}}\left(\mathbf{D}_n\right)\right] \\
&= \alpha \mathrm{E}_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}\left[\frac{L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)}{l_{\mathbb{I}}\left(\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\right)}\right] \underset{\text{(i)}}{\leq} \alpha \mathrm{E}_{\boldsymbol{\vartheta}_{\mathbb{I}}^*}\left[\frac{L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)}{l_{\mathbb{I}}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^*\right)}\right] \underset{\text{(ii)}}{=} \alpha,
\end{aligned}$$

where the (i) is true due to the fact that $l_{\mathbb{I}}\left(\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\right) \geq l_{\mathbb{I}}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^*\right)$, by the definition of (7), and the (ii) is true due to Lemma 1. □

We note that Propositions 1 and 2 are empirical Bayes analogues of Theorems 1 and 2 from Wasserman et al. [2020], which provide guarantees for universal inference confidence set and hypothesis test constructions, respectively. Furthermore, the use of Lemma 1 in the proofs also imply that the CIs constructed via Proposition 1 are e-CIs, as defined by Xu et al. [2022], and the p-values obtained via Proposition 2 can be said to be e-value calibrated, as per the definitions of Wang and Ramdas [2022].

3 FSEB examples and some numerical results

To demonstrate the usefulness of the FSEB results from Section 2, we shall present a number of synthetic and real world applications of the confidence and testing constructions. All of the computation is conducted in the R programming environment (R Core Team, 2020) and replicable scripts are made available at `https://github.com/hiendn/Universal_EB`. Where unspecified, numerical optimization

is conducted using the `optim()` or `optimize()` functions in the case of multivariate and univariate optimization, respectively.

3.1 Stein's problem

We begin by studying the estimation of normal means, as originally considered in Stein [1956]. Here, we largely follow the exposition of Efron [2010, Ch. 1] and note that the estimator falls within the shrinkage paradigm exposited in Serdobolskii [2008]. We consider this setting due to its simplicity and the availability of a simple EB-based method to compare our methodology against.

Let $((X_i, \Theta_i))_{i\in[n]}$ be IID and for each $i \in [n]$, $\Theta_i \sim \mathrm{N}\left(0, \psi^2\right)$ $(\psi^2 > 0)$ and $[X_i|\Theta_i = \theta_i] \sim \mathrm{N}\left(\theta_i, 1\right)$, where $\mathrm{N}\left(\mu, \sigma^2\right)$ is the normal law with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. We assume that ψ^2 is unknown and that we observe data $\mathbf{D}_n$ and wish to construct CIs for the realizations θ_n^*, which characterize the DGP of the observations X_n.

Following Efron [2010, Sec. 1.5], when ψ^2 is known, the posterior distribution of $[\Theta_n|X_n = x_n]$ is $\mathrm{N}\left(g\left(\psi^2\right)x_n, g\left(\psi^2\right)\right)$, where $g\left(\psi^2\right) = \psi^2/\left(1+\psi^2\right)$. Using the data $\mathbf{D}_n$, we have the fact that $\sum_{i=1}^{n-1} X_i^2 \sim \left(\psi^2+1\right)\chi^2_{n-1}$, where χ^2_ν is the chi-squared distribution with ν degrees of freedom. This implies a method-of-moment estimator for g of the form: $\bar{g}_n = 1 - (n-2)/\sum_{i=1}^n X_i^2$, in the case of unknown ψ^2.

We can simply approximate the distribution of $[\Theta_n|\mathbf{D}_n]$ as $\mathrm{N}\left(\bar{g}_n X_n, \bar{g}_n\right)$, although this approximation ignores the variability of $\bar{g}_n$. As noted by Efron [2010, Sec. 1.5], via a hierarchical Bayesian interpretation using an objective Bayesian prior, we may instead deduce the more accurate approximate distribution:

$$\mathrm{N}\left(\bar{g}_n X_n, \bar{g}_n + 2\left[X_n\left(1-\bar{g}_n\right)^2\right]/\left[n-2\right]\right). \tag{8}$$

Specifically, Efron [2010] considers the hyperparameter ψ^2 as being a random variable, say Ψ^2, and places a so-called objective (or non-informative) prior on Ψ^2. In particular, the improper prior assumption that $\Psi^2 + 1 \sim \mathrm{Uniform}\left(0, \infty\right)$ is made. Then, it follows from careful derivation that

$$\mathrm{E}\left[\Theta_n|\mathbf{D}_n\right] = \bar{g}_n X_n \text{ and } \mathrm{var}\left[\Theta_n|\mathbf{D}_n\right] = \bar{g}_n + \frac{2X_n\left(1-\bar{g}_n\right)^2}{n-2},$$

and thus we obtain (8) via a normal approximation for the distribution of $[\Theta_n|\mathbf{D}_n]$ (cf. Morris 1983b, Sec. 4).

The approximation then provides $100\,(1-\alpha)\,\%$ posterior credible intervals for Θ_n of the form

$$\bar{g}_n X_n \pm \zeta_{1-\alpha/2}\sqrt{\bar{g}_n + \frac{2\left[X_n\,(1-\bar{g}_n)^2\right]}{n-2}}, \tag{9}$$

where $\zeta_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. This posterior result can then be taken as an approximate $100\,(1-\alpha)\,\%$ confidence interval for θ_n^*.

Now, we wish to apply the FSEB results from Section 2. Here, $\mathbb{I}=\{n\}$, and from the setup of the problem, we have

$$f\,(x_n|\theta_n)=\phi\,(x_n;\theta_n,1)\ \text{ and }\ \pi\,(\theta_n;\psi)=\phi\left(\theta_n;0,\psi^2\right),$$

where $\phi\left(x;\mu,\sigma^2\right)$ is the normal PDF with mean μ and variance σ^2. Thus,

$$L_{\mathbb{I}}\,(\psi)=\int_{\mathbb{R}}\phi\,(X_n;\theta,1)\,\phi\left(\theta;0,\psi^2\right)\mathrm{d}\theta=\phi\left(X_n;0,1+\psi^2\right)$$

and $l_{\mathbb{I}}\,(\theta_n)=\phi\,(x_n;\theta_n,1)$, which yields a ratio statistic of the form

$$\begin{aligned}R_{\mathbb{I},n}\,(\theta_n)&=L_{\mathbb{I}}\,(\psi_{-n})\,/l_{\mathbb{I}}\,(\theta_n)\\&=\phi\left(X_n;0,1+\hat{\psi}^2_{-n}\right)/\phi\,(X_n;\theta_n,1)\,,\end{aligned}$$

when combined with an appropriate estimator $\hat{\psi}^2_{-n}$ for ψ^2, using only $\bar{\mathbf{D}}_{\mathbb{I},n}=\mathbf{D}_{n-1}$. We can obtain the region $\mathcal{C}^{\alpha}_{\mathbb{I}}\,(\mathbf{D}_n)$ by solving $R_{\mathbb{I},n}\,(\theta_n)\leq 1/\alpha$ to obtain:

$$(X_n-\theta)^2\leq 2\log\,(1/\alpha)+2\log\left(1+\hat{\psi}^2_{-n}\right)+\frac{X_n^2}{\left(1+\hat{\psi}^2_{-n}\right)},$$

which, by Proposition 1, yields the $100\,(1-\alpha)\,\%$ CI for θ_n^*:

$$X_n\pm\sqrt{2\log\,(1/\alpha)+2\log\left(1+\hat{\psi}^2_{-n}\right)+\frac{X_n^2}{\left(1+\hat{\psi}^2_{-n}\right)}}. \tag{10}$$

We shall consider implementations of the CI of form (10) using the estimator

$$\hat{\psi}^2_{-n}=\max\left\{0,s^2_{-n}-1\right\},$$

where s^2_{-n} is the sample variance of the $\bar{\mathbf{D}}_{\mathbb{I},n}$, and $s^2_{-n}-1$ is the method of moment estimator of ψ^2. The maximum operator stops the estimator from becoming negative and causes no problems in the computation of (10).

Table 1: Stein's problem simulation results reported as average performances over 1000 replications.

n	ψ^2	α	Coverage of (9)	Coverage of (10)	Relative Width
10	1^2	0.05	0.948*	1.000*	1.979*
		0.005	0.988*	1.000*	1.738*
		0.0005	0.993*	1.000*	1.641*
	5^2	0.05	0.943	1.000	1.902
		0.005	0.994	1.000	1.543
		0.0005	0.999	1.000	1.388
	10^2	0.05	0.947	1.000	2.058
		0.005	0.994	1.000	1.633
		0.0005	0.999	1.000	1.455
100	1^2	0.05	0.937	0.999	2.068
		0.005	0.997	1.000	1.806
		0.0005	1.000	1.000	1.697
	5^2	0.05	0.949	1.000	1.912
		0.005	0.995	1.000	1.540
		0.0005	1.000	1.000	1.395
	10^2	0.05	0.947	1.000	2.068
		0.005	0.995	1.000	1.635
		0.0005	0.999	1.000	1.455
1000	1^2	0.05	0.949	0.999	2.087
		0.005	0.991	1.000	1.815
		0.0005	1.000	1.000	1.705
	5^2	0.05	0.963	1.000	1.910
		0.005	0.997	1.000	1.544
		0.0005	1.000	1.000	1.399
	10^2	0.05	0.942	1.000	2.066
		0.005	0.995	1.000	1.632
		0.0005	0.999	1.000	1.455

*The results on these lines are computed from 968, 967, and 969 replicates, respectively, from top to bottom. This was due to the negative estimates of the standard error in the computation of (9).

We now compare the performances of the CIs of forms (9) and (10). To do so, we shall consider data sets of sizes $n \in \{10, 100, 1000\}$, $\psi^2 \in \left\{1^2, 5^2, 10^2\right\}$, and $\alpha \in \{0.05, 0.005, 0.0005\}$. For each triplet $\left(n, \psi^2, \alpha\right)$, we repeat the computation of (9) and (10) 1000 times and record the coverage probability and average relative widths of the intervals (computed as the width of (10) divided by that of (9)). The results of our experiment are presented in Table 1.

From Table 1, we observe that the CIs of form (9) tended to produce intervals with the desired levels of coverage, whereas the FSEB CIs of form (10) tended to be conservative and contained the parameter of interest in almost all replications. The price that is paid for this conservativeness is obvious when viewing the relative widths, which implies that for 95% CIs, the EB CIs of form (10) are twice as wide, on average, when compared to the CIs of form (9). However, the relative widths decrease as α gets smaller, implying that the intervals perform relatively similarly when a high level of confidence is required. We further observe that n and ψ^2 had little effect on the performances of the intervals except in the case when $n = 10$ and $\psi^2 = 1$, whereupon it was possible for the intervals of form (9) to not be computable in some cases.

From these results we can make a number of conclusions. Firstly, if one is willing to make the necessary hierarchical and objective Bayesian assumptions, as stated in Efron [2010, Sec. 1.5], then the intervals of form (9) provide very good performance. However, without those assumptions, we can still obtain reasonable CIs that have correct coverage via the FSEB methods from Section 2. Furthermore, these intervals become more efficient compared to (9) when higher levels of confidence are desired. Lastly, when n is small and ψ^2 is also small, the intervals of form (9) can become uncomputable and thus one may consider the use of (10) as an alternative.

3.2 Poisson–gamma count model

The following example is taken from Koenker and Gu [2017] and was originally studied in Norberg [1989] and then subsequently in Haastrup [2000]. In this example, we firstly consider IID parameters $(\Theta_i)_{i\in[n]}$ generated with gamma DGP: $\Theta_i \sim \text{Gamma}(a, b)$, for each $i \in [n]$, where $a > 0$ and $b > 0$ are the shape and rate hyperparameters, respectively, which we put into $\boldsymbol{\psi}$. Then, for each i, we suppose that the data $\mathbf{D}_n = (X_i)_{i\in[n]}$, depending on the covariate sequence $\mathbf{w}_n = (w_i)_{i\in[n]}$, has the Poisson DGP: $[X_i|\Theta_i = \theta_i] \sim \text{Poisson}(\theta_i w_i)$, where $w_i > 0$. We again wish to use the data $\mathbf{D}_n$ to estimate the realization of Θ_n: θ_n^*, which characterizes the DGP of X_n.

Under the specification above, for each i, we have the fact that (X_i, Θ_i) has the joint PDF:

$$f(x_i, \theta_i; \boldsymbol{\psi}) = \frac{b^a}{\Gamma(a)} \theta_i^{a-1} \exp(b\theta_i) \frac{(\theta_i w_i)^{x_i} \exp(-\theta_i w_i)}{x_i}, \tag{11}$$

which we can marginalize to obtain

$$f(x_i; \boldsymbol{\psi}) = \binom{x_i + a + 1}{x_i} \left(\frac{b}{w_i + b}\right)^a \left(\frac{w_i}{w_i + b}\right)^{x_i}, \tag{12}$$

and which can be seen as a Poisson–gamma mixture model. We can then construct the likelihood of $\mathbf{D}_n$ using expression (12), from which we may compute maximum likelihood estimates $\hat{\boldsymbol{\psi}}_n = \left(\hat{a}_n, \hat{b}_n\right)$ of $\boldsymbol{\psi}$. Upon noting that (11) implies the conditional expectation $\mathrm{E}[\Theta_i|X_i = x_i] = (x_i + a)/(w_i + b)$, we obtain the estimator for θ_n^*:

$$\hat{\theta}_n = \frac{X_i + \hat{a}_n}{w_i + \hat{b}_n}. \tag{13}$$

3.2.1 Confidence intervals

We again wish to apply the general result from Section 2 to construct CIs. Firstly, we have $\mathbb{I} = \{n\}$ and

$$f(x_n|\theta_n) = \frac{(\theta_n w_n)^{x_n} \exp(-\theta_n w_n)}{x_n} \text{ and } \pi(\theta_n; \boldsymbol{\psi}) = \frac{b^a}{\Gamma(a)} \theta_n^{a-1} \exp(b\theta_n).$$

As per (12), we can write

$$L_{\mathbb{I}}(\boldsymbol{\psi}) = \binom{X_n + a + 1}{X_n} \left(\frac{b}{w_n + b}\right)^a \left(\frac{w_n}{w_n + b}\right)^{X_n}.$$

Then, since $l_{\mathbb{I}}(\theta_n) = f(X_n|\theta_n)$, we have

$$\begin{aligned} R_{\mathbb{I},n}(\theta_n) &= L_{\mathbb{I}}(\boldsymbol{\psi}) / l_{\mathbb{I}}(\theta_n) \\ &= \binom{X_n + \hat{a}_{-n} + 1}{X_n} \left(\frac{\hat{b}_{-n}}{w_n + \hat{b}_{-n}}\right)^{\hat{a}_{-n}} \left(\frac{w_n}{w_n + \hat{b}_{-n}}\right)^{X_n} \frac{X_n}{(\theta_n w_n)^{X_n} \exp(-\theta_n w_n)}, \end{aligned}$$

when combined with an estimator $\hat{\boldsymbol{\psi}}_{-n} = \left(\hat{a}_{-n}, \hat{b}_{-n}\right)$ of $\boldsymbol{\psi}$, using only $\bar{\mathbf{D}}_{\mathbb{I},n} = \mathbf{D}_{n-1}$.

For any $\alpha \in (0,1)$, we then obtain a $100(1-\alpha)\%$ CI for θ_n by solving $R_{\mathbb{I},n}(\theta_n) \leq 1/\alpha$, which can be done numerically. We shall use the MLE of $\boldsymbol{\psi}$, computed with the data $\bar{\mathbf{D}}_{\mathbb{I},n}$ and marginal PDF (12), as the estimator $\hat{\boldsymbol{\psi}}_{-n}$.

To demonstrate the performance of the CI construction, above, we conduct the following numerical experiment. We generate data sets consisting of $n \in \{10, 100, 1000\}$ observations characterized by hyperparameters $\boldsymbol{\psi} = (a, b) = \{(2,2), (2,5), (5,2)\}$, and we compute intervals using significance levels $\alpha \in \{0.05, 0.005, 0.0005\}$. Here, we shall generate $\mathbf{w}_n$ IID uniformly between 0 and 10. For each triplet $(n, \boldsymbol{\psi}, \alpha)$, we repeat the construction of our CIs 1000 times and record the coverage probability and average width for each case. The results of the experiment are reported in Table 2.

From Table 2, we observe that the empirical coverage of the CIs are higher than the nominal value and are thus behaving as per the conclusions of Proposition 1. As expected, we also find that increasing the nominal confidence level also increases the coverage proportion, but at a cost of increasing the lengths of the CIs. From the usual asymptotic theory of maximum likelihood estimators, we anticipate that increasing n will decrease the variance of the estimator $\hat{\boldsymbol{\psi}}_{-n}$. However, as in Section 3.1, this does not appear to have any observable effect on either the coverage proportion nor lengths of the CIs.

3.2.2 Hypothesis tests

Next, we consider testing the null hypothesis H_0: $\theta^*_{n-1} = \theta^*_n$. To this end, we use the hypothesis testing framework from Section 2. That is, we let $\mathbb{I} = \{n-1, n\}$ and estimate $\boldsymbol{\psi}$ via the maximum likelihood estimator $\hat{\boldsymbol{\psi}}_{\mathbb{I},n} = (a_{\mathbb{I},n}, b_{\mathbb{I},n})$, computed from the data $\bar{\mathbf{D}}_{\mathbb{I},n} = \mathbf{D}_{n-2}$.

We can write

$$L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right) = \prod_{i=n-1}^{n} \binom{X_i + a_{\mathbb{I},n} + 1}{X_i} \left(\frac{b_{\mathbb{I},n}}{w_i + b_{\mathbb{I},n}}\right)^{a_{\mathbb{I},n}} \left(\frac{w_i}{w_i + b_{\mathbb{I},n}}\right)^{X_i},$$

Table 2: Experimental results for CIs constructed for Poisson–gamma count models. The Coverage and Length columns report the coverage proportion and average lengths in each scenario, as computed from 1000 replications.

n	$\boldsymbol{\psi}$	α	Coverage	Length
10	$(2, 2)$	0.05	0.998	3.632
		0.005	1.000	5.484
		0.0005	1.000	6.919
	$(2, 5)$	0.05	0.999	2.976
		0.005	0.999	3.910
		0.0005	1.000	5.481
	$(5, 2)$	0.05	0.997*	5.468*
		0.005	0.999*	7.118*
		0.0005	1.000*	8.349*
100	$(2, 2)$	0.05	0.998	3.898
		0.005	0.999	5.277
		0.0005	1.000	6.883
	$(2, 5)$	0.05	0.999	2.958
		0.005	1.000	3.914
		0.0005	1.000	5.374
	$(5, 2)$	0.05	1.000	5.628
		0.005	1.000	7.124
		0.0005	1.000	8.529
1000	$(2, 2)$	0.05	1.000	4.070
		0.005	1.000	5.424
		0.0005	1.000	6.344
	$(2, 5)$	0.05	0.999	3.049
		0.005	1.000	3.960
		0.0005	1.000	5.479
	$(5, 2)$	0.05	0.998	5.297
		0.005	1.000	7.205
		0.0005	1.000	8.714

*The results on these lines are computed from 999, 999, and 998 replicates, respectively. This was due to there being no solutions to the inequality $R_{\mathbb{I},n}(\theta_n) \leq 1/\alpha$, with respect to $\theta_n > 0$ in some cases.

$$l_{\mathbb{I}}\left(\boldsymbol{\vartheta}_{\mathbb{I}}^{*}\right)=\prod_{i=n-1}^{n}\frac{\left(\theta_{i}^{*}w_{i}\right)^{X_{n}}\exp\left(-\theta_{i}^{*}w_{i}\right)}{X_{i}},$$

and $\boldsymbol{\vartheta}_{\mathbb{I}}^{*}=\left(\theta_{n-1}^{*},\theta_{n}^{*}\right)$. We are also required to compute the maximum likelihood estimator of $\boldsymbol{\vartheta}_{\mathbb{I}}^{*}$, under H_0, as per (7), which can be written as

$$\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\in\left\{\tilde{\boldsymbol{\theta}}=(\theta,\theta):l_{\mathbb{I}}\left(\tilde{\boldsymbol{\theta}}\right)=\sup_{\theta>0}\prod_{i=n-1}^{n}\frac{\left(\theta w_{i}\right)^{X_{n}}\exp\left(-\theta w_{i}\right)}{X_{i}}\right\}.$$

Using the components above, we define the test statistic $T_{\mathbb{I}}\left(\mathbf{D}_{n}\right)=L_{\mathbb{I}}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)/l_{\mathbb{I}}\left(\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\right)$, from which we can derive the p-value $P_{\mathbb{I}}\left(\mathbf{D}_{n}\right)=1/T_{\mathbb{I}}\left(\mathbf{D}_{n}\right)$ for testing H_0.

To demonstrate the application of this test, we conduct another numerical experiment. As in Section 3.2.1, we generate data sets of sizes $n\in\{10,100,1000\}$, where the data $\mathbf{D}_{n-1}$ are generated with parameters $(\Theta_i)_{i\in[n-1]}$ arising from gamma distributions with hyperparameters $\boldsymbol{\psi}=(a,b)=\{(2,2),(2,5),(5,2)\}$. The final observation X_n, making up $\mathbf{D}_n$, is then generated with parameter $\Theta_n=\Theta_{n-1}+\Delta$, where $\Delta\in\{0,1,5,10\}$. As before, we generate the covariate sequence $\mathbf{w}_n$ IID uniformly between 0 and 10. For each triplet $(n,\boldsymbol{\psi},\Delta)$, we test H_0: $\theta_{n-1}^{*}=\theta_{n}^{*}$ 1000 times and record the average number of rejections under at the levels of significance $\alpha\in\{0.05,0.005,0.0005\}$. The results are then reported in Table 3.

The results for the $\Delta=0$ cases in Table 3 show that the tests reject true null hypotheses at below the nominal sizes α, in accordance with Proposition 2. For each combination of n and $\boldsymbol{\psi}$, as Δ increases, the proportion of rejections increase, demonstrating that the tests become more powerful when detecting larger differences between θ_{n-1}^{*} and θ_{n}^{*}, as expected. There also appears to be an increase in power due to larger sample sizes. This is an interesting outcome, since we can only be sure that sample size affects the variability of the estimator $\boldsymbol{\psi}_{\mathbb{I},n}$. Overall, we can be confident that the tests are behaving as required, albeit they may be somewhat underpowered as they are not achieving the nominal sizes.

3.3 Beta–binomial data series

Data from genome-level biological studies, using modern high-throughput sequencing technologies [Krueger et al., 2012], often take the form of a series of counts, which may be modelled through sets of non-identical (possibly correlated) binomial distributions, with beta priors, in a Bayesian framework. The question of interest may vary, for example, from assessing the range of likely values for the binomial parameter in a particular region of the data, to comparing whether two sections of one or more data series are generated from identical distributions. For purposes of demonstrating the performance of the FSEB method in these scenario, we will make the simplifying assumption that all data points are

Table 3: Experimental results for testing the hypothesis H_0: $\theta^*_{n-1} = \theta^*_n$ for Poisson–gamma count models. The Rejection Proportion columns report the average number of rejections, from 1000 tests, at levels of significance $\alpha \in \{0.05, 0.005, 0.0005\}$.

			Rejection Proportion at level α		
n	$\boldsymbol{\psi}$	Δ	0.05	0.005	0.0005
10	$(2, 2)$	0	0.000	0.000	0.000
		1	0.004	0.000	0.000
		5	0.280	0.193	0.128
		10	0.413	0.363	0.317
	$(2, 5)$	0	0.000	0.000	0.000
		1	0.007	0.002	0.000
		5	0.143	0.096	0.064
		10	0.222	0.192	0.170
	$(5, 2)$	0	0.001	0.000	0.000
		1	0.001	0.000	0.000
		5	0.177	0.107	0.052
		10	0.389	0.320	0.254
100	$(2, 2)$	0	0.000	0.000	0.000
		1	0.014	0.003	0.000
		5	0.401	0.289	0.194
		10	0.562	0.489	0.427
	$(2, 5)$	0	0.000	0.000	0.000
		1	0.015	0.000	0.000
		5	0.208	0.127	0.074
		10	0.296	0.235	0.179
	$(5, 2)$	0	0.000	0.000	0.000
		1	0.004	0.000	0.000
		5	0.264	0.150	0.090
		10	0.500	0.425	0.344
1000	$(2, 2)$	0	0.001	0.000	0.000
		1	0.021	0.001	0.000
		5	0.423	0.300	0.216
		10	0.576	0.513	0.450
	$(2, 5)$	0	0.000	0.000	0.000
		1	0.012	0.000	0.000
		5	0.185	0.108	0.061
		10	0.321	0.254	0.197
	$(5, 2)$	0	0.000	0.000	0.000
		1	0.003	0.001	0.000
		5	0.276	0.168	0.088
		10	0.507	0.428	0.354

independently distributed, within, as well as across, any of G data series that may be observed.

3.3.1 Confidence Sets

First, let us assume that we only have a single series, i.e. $G = 1$. Then, we can assume $X_i \sim \text{Bin}(m_i, \theta_i)$, and propose a common prior distribution for Θ_i $(i = 1, \ldots, n)$: Beta(γ, β). Using the techniques described in Section 2, we can find confidence sets for θ_i^*, $(i = 1, \ldots, n)$. For each i, we define, as previously, a subset $\mathbb{I} = \{i\}$, so that $\mathbf{D}_\mathbb{I} = X_i$ and $\overline{\mathbf{D}}_\mathbb{I} = (X_i)_{i \in [n] \setminus \{i\}}$. We then have,

$$R_{\mathbb{I},n}\left(\boldsymbol{\vartheta}_\mathbb{I}\right) = \frac{L_\mathbb{I}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right)}{l_\mathbb{I}\left(\boldsymbol{\vartheta}_\mathbb{I}\right)},$$

where

$$l_\mathbb{I}\left(\boldsymbol{\vartheta}_\mathbb{I}\right) = \binom{m_i}{x_i} \theta_i^{x_i} (1 - \theta_i)^{m_i - x_i}$$

and

$$L_\mathbb{I}\left(\hat{\boldsymbol{\psi}}_{\mathbb{I},n}\right) = \int_{\theta_i} f(x_i|\theta_i) \pi(\theta_i;\, \hat{\gamma}_{-n}, \hat{\beta}_{-n}) \mathrm{d}\theta_i,$$

which gives the ratio

$$R_{\mathbb{I},n}\left(\boldsymbol{\vartheta}_\mathbb{I}\right) = \frac{B(x_i + \hat{\gamma}_{-n}, m_i - x_i + \hat{\beta}_{-n})}{B(\hat{\gamma}, \hat{\beta}_{-n}) \theta_i^{x_i} (1 - \theta_i)^{m_i - x_i}}. \tag{14}$$

Here, $\hat{\gamma}_{-n}$ and $\hat{\beta}_{-n}$ are the empirical Bayes estimates of γ and β, given by

$$\hat{\gamma}_{-n} = (\hat{\phi}_{\text{EB}}^{-1} - 1) \hat{\mu}_{\text{EB}}$$

and

$$\hat{\beta}_{-n} = (\hat{\phi}_{\text{EB}}^{-1} - 1)(1 - \hat{\mu}_{\text{EB}}),$$

where

$$\hat{\mu}_{\text{EB}} = \frac{1}{n-1} \sum_{j \in [n] \setminus i} \frac{x_j}{m_j},$$
$$\hat{\phi}_{\text{EB}} = \left[\frac{\bar{m} \hat{V}_x}{\mu(1-\mu)} - 1 \right] \Big/ (\bar{m} - 1),$$

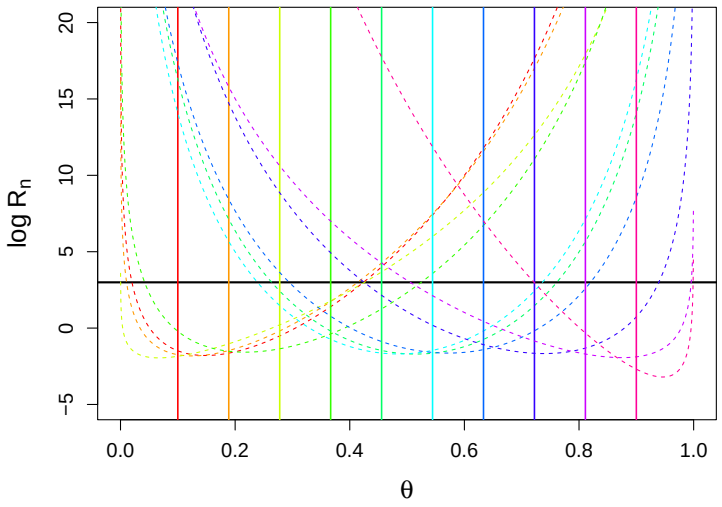

Figure 1: Plots of 95% confidence regions for θ_i^* when true values of θ_i^* span the interval 0.1 to 0.9 ($n = 10$). Here, the 95% CIs are given by the points where the curves for $\log R_{\mathbb{I},n}(\boldsymbol{\vartheta}_{\mathbb{I}})$ intersect with the horizontal line (black), representing a confidence level of $1 - \alpha = 0.95$. Each CI can be seen to contain the corresponding true value of θ_i^*, represented by a vertical line of the same colour as the interval.

$\bar{m} = \frac{1}{n-1}\sum_{j\in[n]\setminus i} m_j$, and $\hat{V}_x = \frac{1}{n-1}\sum_{j\in[n]\setminus i}(\frac{x_j}{m_j} - \hat{\mu}_{\text{EB}})^2$. Further, $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$ is the Beta function, taking inputs $a > 0$ and $b > 0$.

We simulated data from the binomial model under two cases: (a) setting beta hyperparameters $(\alpha, \beta) = (10, 10)$, and hierarchically simulating θ_i^*, $i \in [n]$, and then x_i from a binomial distribution; and (b) setting a range of θ_i^* ($i \in [n]$) values equidistantly spanning the interval $(0.1, 0.9)$ for $n = 10, 100$. Here, m_i ($i \in [n]$) were given integer values uniformly generated in the range $[15, 40]$. In all cases, it was seen that the CIs had perfect coverage, always containing the true value of θ_i^*. An example of the $n = 10$ case is shown in Figure 1.

3.3.2 Hypothesis testing

Aiming to detect genomic regions that may have differing characteristics between two series, a pertinent question of interest may be considered by testing the hypotheses: H_0: $\theta_{i1}^* = \theta_{i2}^*$ vs. H_1: $\theta_{i1}^* \neq \theta_{i2}^*$, for every $i \in [n]$ (with $G = 2$ series). Then, $\mathbf{D}_n = (\boldsymbol{X}_i)_{i\in[n]}$, where $\boldsymbol{X}_i = (X_{i1}, X_{i2})$. From Section 2, the ratio test statistic takes the form

$$T_{\mathbb{I}}(\mathbf{D}_n) = L_{\mathbb{I}}\left(\hat{\gamma}_{\mathbb{I},n}, \hat{\beta}_{\mathbb{I},n}\right) / l_{\mathbb{I}}\left(\tilde{\boldsymbol{\vartheta}}_{\mathbb{I}}\right),$$

where $\hat{\gamma}_{\mathbb{I},n}$ and $\hat{\beta}_{\mathbb{I},n}$ are EB estimators of γ and β, depending only on $\bar{\mathbf{D}}_{\mathbb{I},n} = \mathbf{D}_n \backslash \{X_{i1}, X_{i2}\}$. With $\tilde{\vartheta}_{\mathbb{I}} = \frac{x_{i1}+x_{i2}}{m_{i1}+m_{i2}} = \tilde{\theta}_i$, write $l_{\mathbb{I}}\left(\tilde{\vartheta}_{\mathbb{I}}\right) = f(x_{i1}, x_{i2}|\tilde{\theta}_i)$, and

$$\begin{aligned} L_{\mathbb{I}}\left(\hat{\gamma}_{\mathbb{I},n}, \hat{\beta}_{\mathbb{I},n}\right) &= \int_{\mathbb{T}} f(x_{i1}|\boldsymbol{\theta}_i) f(x_{i2}|\boldsymbol{\theta}_i) \pi(\boldsymbol{\theta}_i;\ \hat{\gamma}_{\mathbb{I},n}, \hat{\beta}_{\mathbb{I},n}) \mathrm{d}\boldsymbol{\theta}_i \\ &= \binom{m_{i1}}{x_{i1}} \binom{m_{i2}}{x_{i2}} \frac{B(x_{i1}+\hat{\gamma}_{\mathbb{I},n}, m_{i1}-x_{i1}+\hat{\beta}_{\mathbb{I},n}) B(x_{i2}+\hat{\gamma}_{\mathbb{I},n}, m_{i2}-x_{i2}+\hat{\beta}_{\mathbb{I},n})}{\left[B(\hat{\gamma}_{\mathbb{I},n}, \hat{\beta}_{\mathbb{I},n})\right]^2}, \end{aligned}$$

which gives

$$T_{\mathbb{I}}\left(\mathbf{D}_n\right) = \frac{B(x_{i1}+\hat{\gamma}_{\mathbb{I},n}, m_{i1}-x_{i1}+\hat{\beta}_{\mathbb{I},n}) B(x_{i2}+\hat{\gamma}_{\mathbb{I},n}, m_{i2}-x_{i2}+\hat{\beta}_{\mathbb{I},n})}{[B(\hat{\gamma}_{\mathbb{I},n}, \hat{\beta}_{\mathbb{I},n})]^2 \tilde{\theta}_i^{x_{i1}+x_{i2}} (1-\tilde{\theta}_i)^{m_{i1}+m_{i2}-x_{i1}-x_{i2}}},$$

where $\hat{\gamma}_{\mathbb{I},n}$ and $\hat{\beta}_{\mathbb{I},n}$ are calculated in a similar fashion to Section 3.3.1 except that data from both sequences should be used to estimate $\hat{\mu}_{\mathrm{EB}}$ and $\hat{\phi}_{\mathrm{EB}}$, in the sense that

$$\begin{aligned} \hat{\mu}_{\mathrm{EB}} &= \frac{1}{2n-2} \sum_{k \neq i} \sum_{g=1}^{2} \frac{x_{kg}}{m_{kg}}, \text{ and} \\ \hat{\phi}_{\mathrm{EB}} &= \left[\frac{\bar{m} V_{xy}}{\hat{\mu}_{\mathrm{EB}}(1-\hat{\mu}_{\mathrm{EB}})} - 1 \right] \Big/ (\bar{m}-1), \end{aligned}$$

where

$$\begin{aligned} \bar{m} &= \frac{1}{2n-2} \sum_{k \neq i} \sum_{g=1}^{2} m_{kg}, \text{ and} \\ V_{xy} &= \frac{1}{2n-2} \sum_{k \neq i} \sum_{g=1}^{2} \left(\frac{x_{kg}}{m_{kg}} - \hat{\mu}_{\mathrm{EB}} \right)^2. \end{aligned}$$

In our first simulation, we assessed the performance of the test statistic in terms of the Type I error. Assuming a window size of $n = 20$, realized data (x_{i1}, x_{i2}) $(i \in [n])$, were simulated from independent binomial distributions with $\theta^*_{i1} = \theta^*_{i2} = \theta^*_i$ $(i = 1, \ldots, n)$, with θ^*_i ranging between 0.1 and 0.9, and $m_{i1}, m_{i2} \in \mathbb{N}$ uniformly and independently sampled from the range $[15, 40]$. The first panel of Figure 2 shows the calculated test statistic values $T_{\mathbb{I}}\left(\mathbf{D}_n\right)$ for the 20 genomic indices on the logarithmic scale, over 100 independently replicated datasets, with horizontal lines displaying values $\log(1/\alpha)$, for significance levels $\alpha \in \{0.01, 0.02, 0.05\}$. No points were observed above the line corresponding to $\alpha = 0.01$, indicating that the Type I error of the test statistic does not exceed the nominal level. Next, we assessed the power of the test statistic at three levels of significance ($\alpha \in \{0.01, 0.02, 0.05\}$) and differing effect sizes. For each i $(i \in [n])$, θ^*_{i1} was set to be a value between 0.05 and 0.95, and $\theta^*_{i2} = \theta^*_{i1} + \Delta$, where $0.1 < \Delta < 0.9$ (with $\theta^*_{i2} < 1$). A sample of 20 replicates were simulated under each possible set of values of (θ^*_1, θ^*_2). The second panel of Figure 2 shows that the power functions

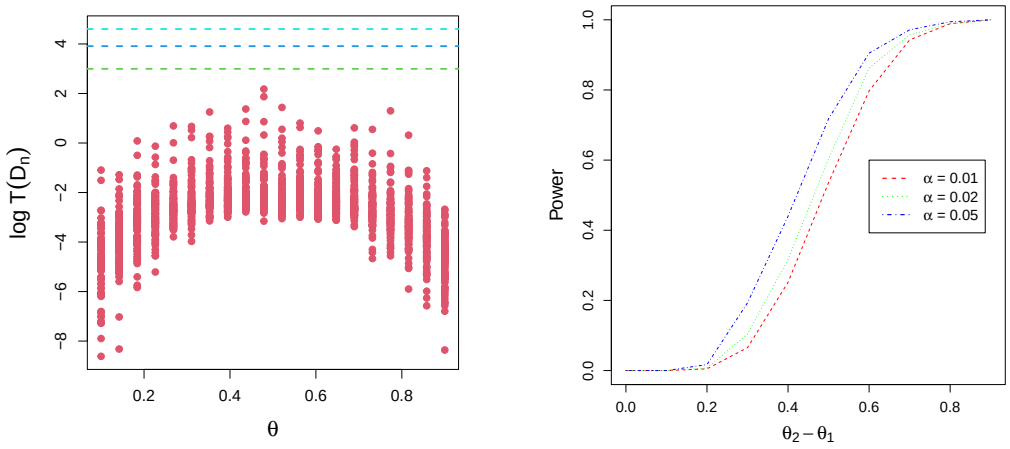

Figure 2: Panel (a): Test statistic for 100 replications of the beta–binomial example under the null hypothesis of equality of proportions. The three horizontal lines correspond to cutoffs according to significance levels of $\alpha = 0.05$ (green), $\alpha = 0.02$ (blue), and $\alpha = 0.01$ (turquoise). Panel (b): Power function over different values of $\Delta = \theta_2^* - \theta_1^*$ at three levels of significance: $\alpha \in \{0.01, 0.02, 0.05\}$.

increased rapidly to 1 as the difference Δ was increased.

In our next numerical experiment, we generated data sets of sizes $n \in \{10, 100, 1000\}$, where realized observations x_{i1}, and x_{i2} are simulated from independent binomial distributions with parameters θ_{i1}^* and θ_{i2}^*, respectively ($i \in [n]$). For each i, θ_{i1*} was generated from a beta distribution, in turn, with hyperparameters $\boldsymbol{\psi} = (\gamma, \beta) \in \{(2,2), (2,5), (5,2)\}$; and $\theta_{i2}^* = \theta_{i1}^* + \Delta$, where $\Delta \in \{0, 0.2, 0.5, 0.9\}$. We generated 100 instances of data under each setting and assessed the power of the FSEB test statistic through the number of rejections at levels $\alpha \in \{0.0005, 0.005, 0.05\}$. The results are shown in Table 4.

Similarly to the Poisson–gamma example, it can be seen that the tests reject true null hypotheses at below the nominal sizes α, in each case. For each combination of n and $\boldsymbol{\psi}$, as Δ increases, the rejection rate increases, making the tests more powerful as expected, when detecting larger differences between θ_{i1}^* and θ_{i2}^*, frequently reaching a power of 1 even when the difference was not maximal. There did not appear to be a clear increase in power with the sample size, within the settings considered. Overall, we may conclude, as previously, that the tests are behaving as expected, although both this example and the Poisson–gamma case show that the tests may be underpowered as they do not achieve the nominal size for any value of α.

As an additional assessment of how FSEB performs in comparison to other tests in a similar setting, we carried out a number of additional simulation studies, in which FSEB was compared with Fisher's exact test and a score test, over various settings of n, $\boldsymbol{\psi}$ and Δ, as well as for different ranges of m_i ($i = 1 \in [n]$). Comparisons were made using the p-values as well as false discovery rate (FDR) corrected p-values arising from FDR control methods [Wang and Ramdas, 2022], and are presented

Table 4: Experimental results for testing the hypothesis H_0: $\theta^*_{i1} = \theta^*_{i2}$ for Beta–binomial count series models. The Rejection proportion columns report the average number of rejections, from 100 test replicates, at levels of significance $\alpha \in \{0.05, 0.005, 0.0005\}$.

			Rejection proportion at level α		
n	$\boldsymbol{\psi}$	Δ	0.0005	0.005	0.05
10	$(2,2)$	0	0.000	0.000	0.000
		0.2	0.000	0.004	0.039
		0.5	0.305	0.471	0.709
		0.9	0.980	1.000	1.000
	$(2,5)$	0	0.000	0.000	0.000
		0.2	0.000	0.001	0.025
		0.5	0.249	0.464	0.692
		0.9	0.995	1.000	1.000
	$(5,2)$	0	0.000	0.000	0.000
		0.2	0.000	0.006	0.052
		0.5	0.281	0.459	0.690
		0.9	0.993	0.993	1.000
100	$(2,2)$	0	0.000	0.000	0.000
		0.2	0.000	0.004	0.037
		0.5	0.272	0.459	0.700
		0.9	0.996	0.998	1.000
	$(2,5)$	0	0.000	0.000	0.000
		0.2	0.000	0.003	0.032
		0.5	0.267	0.459	0.693
		0.9	0.994	0.999	1.000
	$(5,2)$	0	0.000	0.000	0.000
		0.2	0.000	0.004	0.047
		0.5	0.269	0.459	0.697
		0.9	0.987	0.998	0.999
1000	$(2,2)$	0	0.000	0.000	0.000
		0.2	0.000	0.003	0.031
		0.5	0.280	0.476	0.707
		0.9	0.982	0.992	0.998
	$(2,5)$	0	0.000	0.000	0.000
		0.2	0.000	0.003	0.030
		0.5	0.264	0.459	0.693
		0.9	0.989	0.996	1.000
	$(5,2)$	0	0.000	0.000	0.000
		0.2	0.000	0.005	0.047
		0.5	0.279	0.474	0.706
		0.9	0.986	0.995	0.999

in the online Supplementary Materials (Tables S1–S8 and Figures S1–S8). It is evident in almost all cases (and especially in case C, which most closely resembles the real life application scenario) that (i) the power levels are very similar across methods, especially as values of n, m_i ($i \in [n]$) and effect sizes increase, and (ii) in every case, there are some settings in which Fisher's test and the score test are anti-conservative (even after FDR correction), with their Type I error greatly exceeding the nominal levels of significance, while this never occurs for FSEB, even without FDR correction.

4 Real-data applications

4.1 The `Norberg` data

We now wish to apply the FSEB CI construction from Section 3.2.1 to produce CIs in a real data application. We shall investigate the `Norberg` data set from the REBayes package of Koenker and Gu [2017], obtained from Haastrup [2000]. These data pertain to group life insurance claims from Norwegian workmen. Here, we have $n = 72$ observations $\mathbf{D}_n$, containing total number of death claims X_i, along with covariates $\mathbf{w}_n$, where w_i is the number of years of exposure, normalized by a factor of 344, for $i \in [n]$. Here each i is an individual occupation group.

To analyze the data, we use the Poisson–gamma model and estimate the generative parameters $\boldsymbol{\vartheta}_n^*$ using estimates of form (13). Here, each θ_i^* can be interpreted as an unobserved multiplicative occupation specific risk factor that influences the number of claims made within occupation group i. To obtain individually-valid 95% CIs for each of the n estimates, we then apply the method from Section 3.2.1. We present both the estimated risk factors and their CIs in Figure 3.

From Figure 3, we notice that most of the estimates of $\boldsymbol{\vartheta}_n^*$ are between zero and two, with the exception of occupation group $i = 22$, which has an estimated risk factor of $\theta_{22}^* = 2.59$. Although the risk factors are all quite small, the associated CIs can become very large, as can be seen in the top plot. This is due to the conservative nature of the CI constructions that we have already observed from Section 3.1.

We observe that wider CIs were associated with observations where $X_i = 0$, with w_i being small. In particular, the largest CI, occurring for $i = 55$, has response $X_{55} = 0$ and the smallest covariate value in the data set: $w_{55} = 4.45$. The next largest CI occurs for $i = 5$ and also corresponds to a response $X_5 = 0$ and the second smallest covariate value $w_5 = 11.30$.

However, upon observation of the bottom plot, we see that although some of the CIs are too wide to be meaningful, there are still numerous meaningful CIs that provide confidence regarding the lower limits as well as upper limits of the underlying risk factors. In particular, we observe that the CIs for occupation groups $i = 26$ and $i = 54$ are remarkably narrow and precise. Of course, the preceding

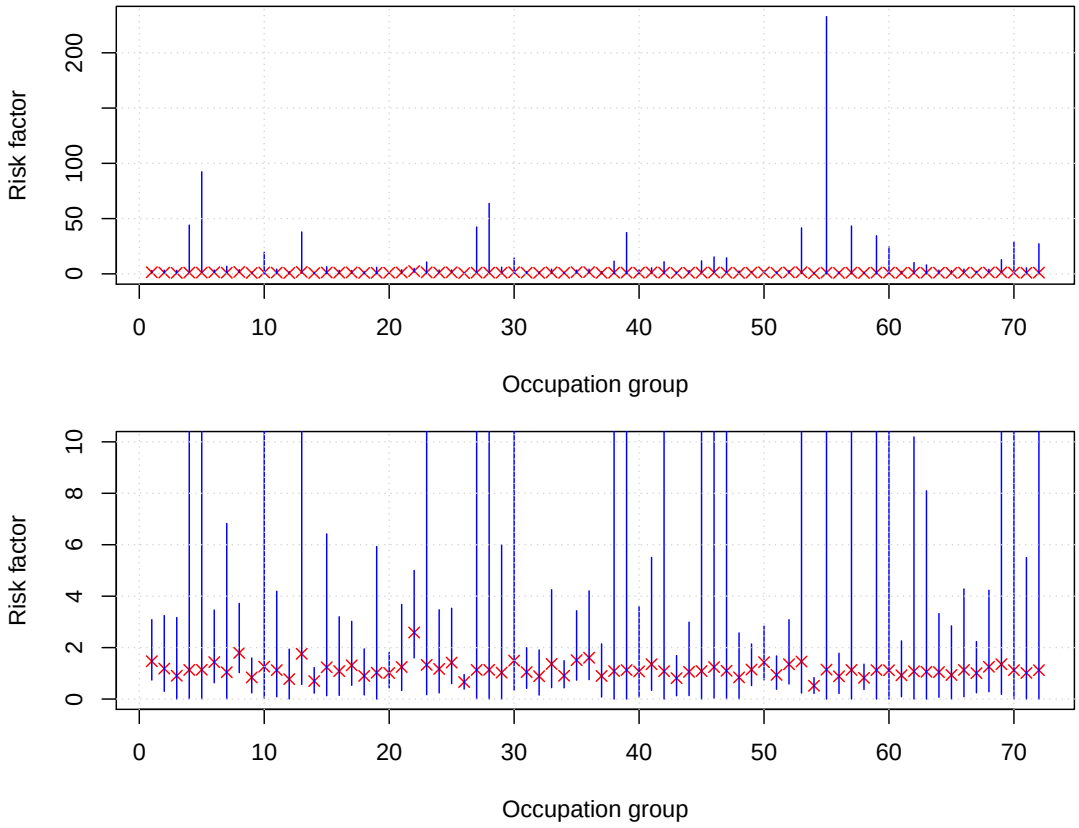

Figure 3: Estimates of risk factors $\boldsymbol{\vartheta}_n^*$ for the `Norberg` data set along with associated 95% CIs. The estimated risk factor for each occupation group is depicted as a cross and the associate (individually-valid) CI is depicted as a line. The top plot displays the CIs at their entire lengths, whereas the bottom plot displays only the risk factor range between 0 and 10.

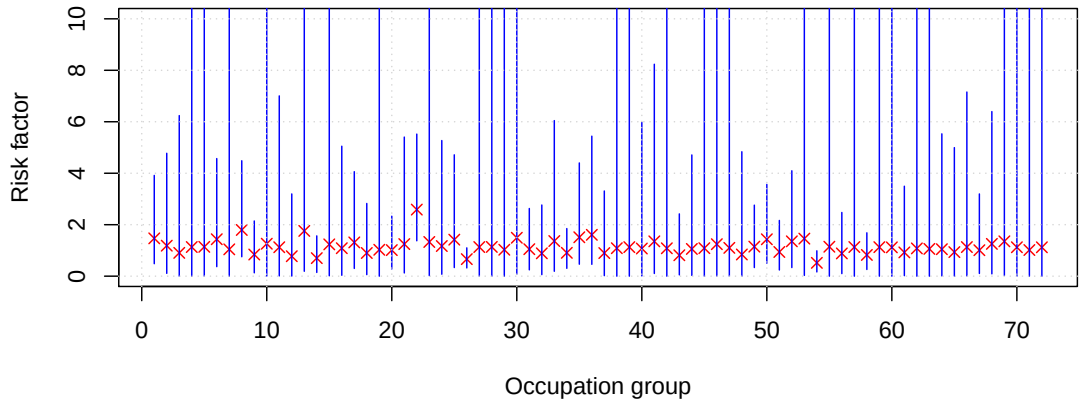

Figure 4: Estimates of risk factors $\boldsymbol{\vartheta}_n^*$ for the `Norberg` data set along with the associated simultaneous 95% confidence set. The estimated risk factors for each occupation group is depicted as a cross and the simultaneous confidence set can be constructed via the cartesian product of the adjusted CIs, depicted as lines. The plot is focused on the risk factor range between 0 and 10.

inferential observations are only valid when considering each of the n CIs, individually, and under the assumption that we had chosen to draw inference regarding the corresponding parameter of the CI, before any data are observed.

If we wish to draw inference regarding all n elements of $\boldsymbol{\vartheta}_n^*$, simultaneously, then we should instead construct a $100\,(1-\alpha)\,\%$ simultaneous confidence set $\bar{\mathcal{C}}^\alpha\,(\mathbf{D}_n)$, with the property that

$$\Pr_{\boldsymbol{\vartheta}_n^*}\left[\boldsymbol{\vartheta}_n^* \in \bar{\mathcal{C}}^\alpha\,(\mathbf{D}_n)\right] \geq 1-\alpha.$$

Using Bonferroni's inequality, we can take $\bar{\mathcal{C}}^\alpha\,(\mathbf{D}_n)$ to be the Cartesian product of the individual $100\,(1-\alpha/n)\,\%$ (adjusted) CI for each parameter θ_i^*:

$$\bar{\mathcal{C}}^\alpha\,(\mathbf{D}_n) = \prod_{i=1}^{n} \mathcal{C}_i^{\alpha/n}\,(\mathbf{D}_n)\,.$$

Using the $\alpha = 0.05$, we obtain the 95% simultaneous confidence set that appears in Figure 4. We observe that the simultaneous confidence set now permits us to draw useful inference regarding multiple parameters, at the same time. For example, inspecting the n adjusted CIs, we observe that the occupations corresponding to indices $i \in \{8, 22, 50\}$ all have lower bounds above 0.5. Thus, interpreting these indices specifically, we can say that each of the three adjusted confidence intervals, which yield the inference that the risk factors $\theta_i^* > 0.5$ for $i \in \{8, 22, 50\}$, contains the parameter θ_i^* with probability 0.95, under repeated sampling.

Since our individual CI and adjusted CI constructions are e-CIs, one can alternatively approach the problem of drawing simultaneously valid inference via the false coverage rate (FCR) controlling techniques of Xu et al. [2022]. Using again the parameters θ_i^* corresponding to $i \in \{8, 22, 50\}$, as an

example, we can use Theorem 2 of Xu et al. [2022] to make the statement that the three adjusted CIs $\mathcal{C}_i^{3\alpha/n}(\mathbf{D}_n)$, for $i \in \{8, 22, 50\}$, can be interpreted at the FCR controlled level $\alpha \in (0, 1)$, in the sense that

$$\mathrm{E}_{\boldsymbol{\vartheta}^*_{\mathbb{I}(\mathbf{D}_n)}}\left[\frac{\sum_{i\in\mathbb{I}}\left[\!\left[\theta_i^* \notin \mathcal{C}_i^{|\mathbb{I}(\mathbf{D}_n)|\alpha/n}(\mathbf{D}_n)\right]\!\right]}{\max\{1, |\mathbb{I}(\mathbf{D}_n)|\}}\right] \leq \alpha,$$

where $\mathbb{I}(\mathbf{D}_n)$ is a data-dependent subset of parameter indices. In particular, we observe the realization $\{8, 22, 50\}$ of $\mathbb{I}(\mathbf{D}_n)$, corresponding to the data-dependent rule of selecting indices with adjusted CIs $\mathcal{C}_i^{\alpha/n}(\mathbf{D}_n)$ with lower bounds greater than 0.5. Here, $[\![\mathsf{A}]\!] = 1$ if statement A is true and 0, otherwise.

Clearly, controlling the FCR at level α yields narrower CIs for each of our the three assessed parameters than does the more blunt simultaneous confidence set approach. In particular, the 95% simultaneous adjusted CIs obtained via Bonferroni's inequality are $(0.775, 4.485)$, $(1.375, 5.520)$, and $(0.505, 3.565)$, and the 0.05 level FCR controlled adjusted CIs are $(0.810, 4.300)$, $(1.430, 5.390)$, and $(0.555, 3.390)$, for the parameters θ_i^* corresponding to the respective parameters $i \in \{8, 22, 50\}$. Overall, these are positive results as we do not know of another general method for generating CIs in this EB setting, whether individually or jointly.

4.2 Differential methylation detection in bisulphite sequencing data

DNA methylation is a chemical modification of DNA caused by the addition of a methyl (CH_3-) group to a DNA nucleotide – usually a C that is followed by a G – called a CpG site, which is an important factor in controlling gene expression over the human genome. Detecting differences in the methylation patterns between normal and ageing cells can shed light on the complex biological processes underlying human ageing, and hence has been an important scientific problem over the last decade [Smith and Meissner, 2013]. Methylation patterns can be detected using high-throughput bisulphite sequencing experiments [Krueger et al., 2012], in which data are generated in the form of sequences of numbers of methylated cytosines, x_{ig}, among the total counts of cytosines, m_{ig}, for n CpG sites on a genome ($i \in [n]$), for G groups of cell types $g \in [G]$. Often, there are $G = 2$ groups, as in our example that follows, for which the question of interest is to detect regions of differential methylation in the DNA of normal and ageing cells. Based on the setup above, a set of bisulphite sequencing data from an experiment with G groups might be considered as G series of (possibly correlated) observations from non-identical binomial distributions. The degree of dependence between adjacent CpG sites typically depends on the genomic distance between these loci, but since these are often separated by hundreds of bases, for the moment it is assumed that this correlation is negligible and is not incorporated into our model.

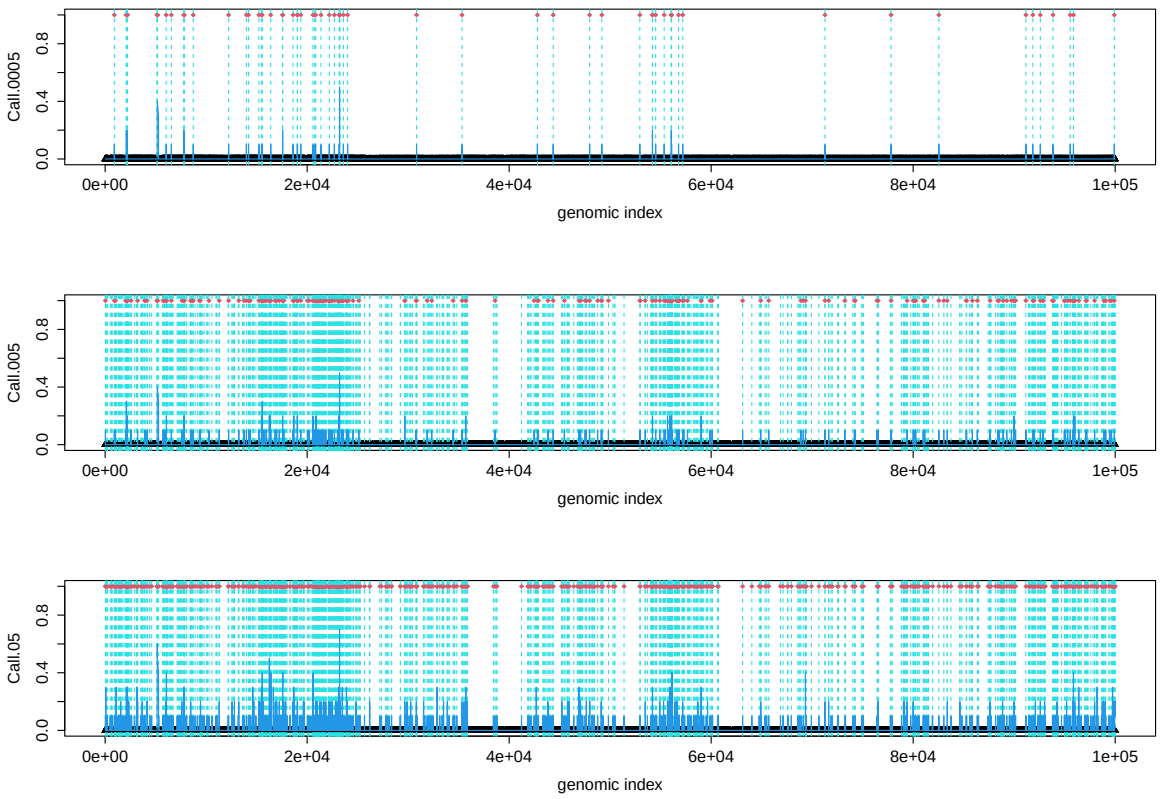

Figure 5: FSEB test statistics over a segment of methylation data. The panels show the demarcation of loci into differentially methylated (coded as "1") and non-differentially methylated sites (coded as "0") with an overlay of a moving average with a window size of 10 CpG sites, at significance level cutoffs of 0.0005, 0.005, and 0.05.

4.2.1 Application to Methylation data from Human chromosome 21

We evaluated the test statistic $T_{\mathbb{I}}(\mathbf{D}_n)$ over a paired segment of methylation data from normal and ageing cells, from $100,000$ CpG sites on human chromosome 21 [Cruickshanks et al., 2013]. After data cleaning and filtering (to remove sites with too low or too high degrees of experimental coverage, that can introduce errors), $58,361$ sites remained for analysis. Figure 5 shows the predicted demarcation of the data into differentially and non-differentially methylated sites over the entire region, at three cutoff levels of significance, overlaid with a moving average using a window size of 10 sites. It was observed that large values of the test statistic were often found in grouped clusters, which would be biologically meaningful, as loss of methylation in ageing cells is more likely to be highly region-specific, rather than randomly scattered over the genome. The overall rejection rates for the FSEB procedure corresponding to significance levels of $\alpha = 0.0005, 0.05, 0.02$ and 0.01 were found to be 0.0012, 0.0154, 0.0092, and 0.0064, respectively.

As a comparison to other methods for detecting differential methylation, we also applied site-by-site

Fisher tests and score tests as implemented for bisulphite sequencing data in the R Bioconductor package DMRcaller [Catoni et al., 2018]. For purposes of comparison, we used two significance level cutoffs of 0.05 and 0.0005 for our FSEB test statistic, along with the same cutoffs subject to a Benjamini–Hochberg FDR correction for the other two testing methods. Figure 6 shows the comparison between the calculated site-specific p-values of the Fisher and score tests with the calculated FSEB test statistic (all on the logarithmic scale) over the entire genomic segment, which indicates a remarkable degree of overlap in the regions of differential methylation. There are, however, significant differences as well, in both the numbers of differential methylation calls and their location. In particular, the FSEB test statistic appeared to have stronger evidence for differential methylation in two regions, one on the left side of the figure, and one towards the centre. The Fisher test, being the most conservative, almost missed this central region (gave a very weak signal), while the score test gave a very high proportion of differential methylation calls compared to both other methods – however, the results from the score test may not be as reliable as many cells contained small numbers of counts which may render the test assumptions invalid. Table 5 gives a summary of the overlap and differences of the results from the different methods at two levels of significance, indicating that with FDR corrections, the Fisher test appears to be the most conservative, the score test the least conservative, and the FSEB procedure in-between the two. We also calculated, for each pair of methods, the proportion of matching calls, defined as the ratio of the number of sites predicted by both methods as either differentially methylated, or non-differentially methylated, to the total number of sites. These proportions indicated a high degree of concordance, especially between FSEB and Fisher tests, with the score test showing the least degree of concordance at both levels of significance. As expected, the degree of concordance decreased with an increase in α, but only slightly so, between the FDR-corrected Fisher test and FSEB.

5 Conclusion

EB is a powerful and popular paradigm for conducting parametric inference in situations where the DGP can be assumed to possess a hierarchical structure. Over the years, general frameworks for point estimation have been developed for EB, such as via the shrinkage estimators of Serdobolskii [2008] or the various method of moments and likelihood-based methods described in Maritz and Lwin [1989, Sec. 3]. Contrastingly, the construction of interval estimators and hypothesis tests for EB parameters rely primarily on bespoke derivations and analysis of the specific models under investigation.

In this paper, we have adapted the general universal inference framework for finite sample valid interval estimation and hypothesis testing of Wasserman et al. [2020] to construct a general framework within the EB setting, which we refer to as the FSEB technique. In Section 2, we proved that these

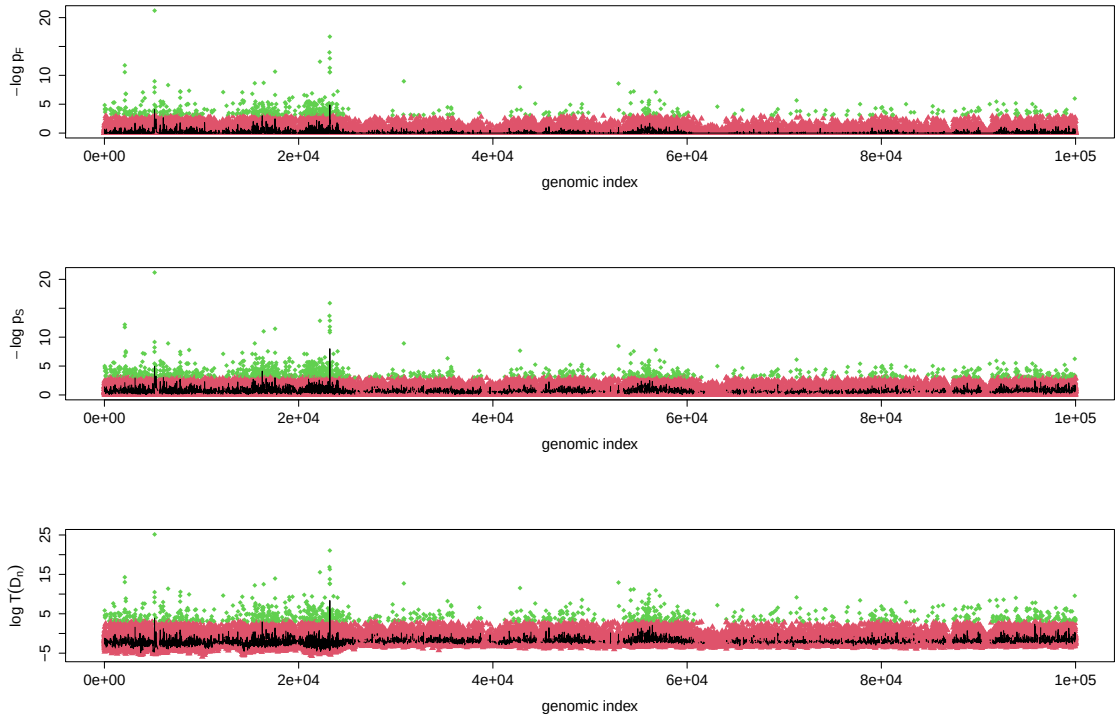

Figure 6: Results of three testing procedures to detect sites of differential methylation over a segment of methylation data. The first two panels show the negative logarithms of the FDR-corrected p-values for the (i) Fisher test ($-\log p_F$) and (ii) score test ($-\log p_S$), while the third panel shows the logarithm of the FSEB test statistic ($\log T(D_n)$). The black curve in each plot corresponds to a moving average with a window size of 10. The points are coloured by differential methylation state call: green if differentially methylated, and red if not, at test size 0.05.

Table 5: Comparison of differential methylation calling results between different methods: (i) FSEB (ii) Fisher tests with FDR-adjusted p-values (FF) (iii) Fisher tests, unadjusted (F) (iv) score tests with FDR-adjusted p-values (SF) and (v) score tests, unadjusted (S). The upper table gives the proportions of sites called to be differentially expressed under the tests of sizes $\alpha \in \{0.0005, 0.05\}$. The lower table gives the proportion of overlaps between differential methylation calls from each pair of methods at a fixed level $\alpha \in \{0.0005, 0.05\}$.

	Proportion of rejections at level	
	$\alpha = 0.0005$	$\alpha = 0.05$
FSEB	0.0012	0.0154
FF	0.0003	0.0097
F	0.0098	0.1102
SF	0.1333	0.1528
S	0.1457	0.2926

	Proportion of overlap in matching calls at level								
	$\alpha = 0.0005$					$\alpha = 0.05$			
Method	FF	F	SF	S	Method	FF	F	SF	S
FSEB	0.999	0.991	0.866	0.856	FSEB	0.992	0.905	0.860	0.723
FF		0.991	0.867	0.855	FF		0.900	0.857	0.717
F			0.858	0.864	SF			0.777	0.818
SF				0.988	S				0.860

FSEB techniques generate valid confidence sets and hypothesis tests of the correct size. In Section 3, we demonstrated via numerical simulations, that the FSEB methods can be used in well-studied synthetic scenarios. There, we highlight that the methods can generate meaningful inference for realistic DGPs. This point is further elaborated in Section 4, where we also showed that our FSEB approach can be usefully applied to draw inference from real world data, in the contexts of insurance risk and the bioinformatics study of DNA methylation.

We note that although our framework is general, due to it being Markov inequality-based, it shares the same general criticism that may be laid upon other universal inference methods, which is that the confidence sets and hypothesis tests can often be conservative, in the sense that the nominal confidence level or size is not achieved. The lack of power due to the looseness of Markov's inequality was first mentioned and discussed in Wasserman et al. [2020], where it is also pointed out that, in the universal inference setting, the logarithm of the analogous ratio statistics to (6) have tail probabilities that scale, in α, like those of χ^2 statistics. The conservativeness of universal inference constructions is further discussed in the works of Dunn et al. [2021], Tse and Davison [2022], and Strieder and Drton [2022], where the topic is thoroughly explored via simulations and theoretical results regarding some classes of sufficiently regular problems. We observe this phenomenon in the comparisons in Sections 3.1 (and further expanded in the Supplementary Materials). We also explored subsampling-based tests within the FSEB framework, along the lines proposed by Dunn et al. [2021], which led to very minor increases in power in some cases with small sample sizes without affecting the Type I error. With such an outcome not entirely discernible from sampling error, and with the substantial increase to computational cost, it does not seem worthwhile to employ the subsampling-based approach here. A possible reason for the lack improvement in power observed, despite subsampling, can be attributed to the fact that the sets $\mathbb{I}$, and their complements, are not exchangeable; since the indices fundamentally define the hypotheses and parameters of interest.

However, we note that since the methodology falls within the e-value framework, it also inherits desirable properties, such as the ability to combine test statistics by averaging [Vovk and Wang, 2021], and the ability to more-powerfully conduct false discovery rate control when tests are arbitrarily dependent [Wang and Ramdas, 2022].

Overall, we believe that FSEB techniques can be usefully incorporated into any EB-based inference setting, especially when no other interval estimators or tests are already available, and are a useful addition to the statistical tool set. Although a method that is based on the careful analysis of the particular setting is always preferable in terms of exploiting the problem specific properties in order to generate powerful tests and tight intervals, FSEB methods can always be used in cases where such careful analyses may be mathematically difficult or overly time consuming.

References

S E Ahmed and N Reid, editors. *Empirical Bayes and Likelihood Inference*. Springer, New York, 2001.

D R Bickel. *Genomics Data Analysis: False Discovery Rates and Empirical Bayes Methods*. CRC Press, Boca Raton, 2020.

G Casella and J T Hwang. Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *Journal of the American Statistical Association*, 78:688–698, 1983.

M Catoni, J M Tsang, A P Greco, and N R Zabet. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Research*, 46:e114, 2018.

H A Cruickshanks, T McBryan, D M Nelson, N D Vanderkraats, P P Shah, J van Tuyn, T S Rai, C Brock, G Donahue, D S Dunican, M E Drotar, R R Meehan, J R Edwards, S L Berger, and P D Adams. Senescent cells harbour features of the cancer epigenome. *Nature Cell Biology*, 15: 1495–1506, 2013.

G S Datta, M Ghosh, D D Smith, and P Lahiri. On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scandinavian Journal of Statistics*, 29:139–152, 2002.

R Dunn, A Ramdas, S Balakrishnan, and L Wasserman. Gaussian universal likelihood ratio testing. *arXiv:2104.14676*, 2021.

B Efron. *Large-scale Inference*. Cambridge University Press, Cambridge, 2010.

P Grunwald, R de Heide, and W M Koolen. Safe testing. In *Information Theory and Applications Workshop (ITA)*, 2020.

S Haastrup. Comparison of some Bayesian analyses of heterogeneity in group life insurance. *Scandinavian Actuarial Journal*, 2000:2–16, 2000.

T J Hardcastle and K A Kelly. Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics*, 14:135, 2013.

J T G Hwang and Z Zhao. Empirical Bayes confidence intervals for selected parameters in high-dimensional data. *Journal of the American Statistical Association*, 108:607–618, 2013.

J T G Hwang, J Qiu, and Z Zhao. Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society B*, 71:265–285, 2009.

I M Johnstone and B W Silverman. EbayesThresh: R programs for empirical Bayes thresholding. *Journal of Statistical Software*, 12:1–38, 2005.

E Kaufmann and W M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv:1811.11419v1*, 2018.

R Koenker and J Gu. REBayes: Empirical Bayes mixture methods in R. *Journal of Statistical Software*, 82:1–26, 2017.

F Krueger, B Kreck, A Franke, and S R Andrews. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9:145–151, 2012.

N M Laird and T A Louis. Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82:739–750, 1987.

N Leng, J A Dawson, J A Thomson, V Ruotti, A I Rissman, B M G Smits, J D Haag, M N Gould, R M Stewart, and C Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29:1035–1043, 2013.

J S Maritz and T Lwin. *Empirical Bayes Methods*. CRC Press, Boca Raton, 1989.

C N Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78:47–55, 1983a.

C N Morris. Parametric empirical bayes confidence intervals. In *Scientific inference, data analysis, and robustness*. Elsevier, 1983b.

B Narasimhan and B Efron. deconvolveR: a G-modeling program for deconvolution and empirical Bayes estimation. *Journal of Statistical Software*, 94:1–20, 2020.

R Norberg. Experience rating in group life insurance. *Scandinavian Actuarial Journal*, 1989:194–224, 1989.

V I Serdobolskii. *Multiparametric Statistics*. Elsevier, Amsterdam, 2008.

G Shafer. Testing by betting: a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society B*, 184:407–431, 2021.

Z D Smith and A Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14:204–220, 2013.

C Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1956.

D Strieder and M Drton. On the choice of the splitting ratio for the split likelihood ratio test. *arXiv:2203.06748*, 2022.

Y C Tai and T P Speed. A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics*, 34:2387–2412, 2006.

T Tse and A C Davison. A note on universal inference. *Stat*, to appear, 2022.

V Vovk. Strong confidence intervals for autoregression. *arXiv:0707.0660v1*, 2007.

V Vovk and R Wang. E-values: calibration, combination, and applications. *Annals of Statistics*, 49: 1736–1754, 2021.

R Wang and A Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society B*, 84:822–852, 2022.

L Wasserman, A Ramdas, and S Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117:16880–16890, 2020.

Z Xu, R Wang, and A Ramdas. Post-selection inference for e-value based confidence intervals. *arXiv:2203.12572*, 2022.

M Yoshimori and P Lahiri. A second-order efficient empirical Bayes confidence interval. *Annals of Statistics*, 42:1233–1261, 2014.

Supplementary Tables and Figures for

"Finite sample inference for empirical Bayesian methods"

Hien Duy Nguyen[1] Mayetri Gupta[2]

[1] School of Mathematics and Physics, University of Queensland, St. Lucia 4067

[2] Department of Mathematics and Statistics, La Trobe University, Bundoora 3086

[3] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ

This document contains details of some further simulation studies comparing the power and type I errors of tests with and without FDR control in the Beta-Binomial series example (Section 3.3.2), under varying settings for:

1. Sample size ($N : 4,\ 10, 100$);
2. Total binomial counts (m_i) per site i, ($i = 1, \ldots, N$) ranges:
   (a) Case A: (4, 20)
   (b) Case B: (15, 40)
   (c) Case C: (5, 200)
   (d) Case D: (50, 300);
3. $\Delta_i = |\theta_{i1} - \theta_{i2}| : 0.6,\ 0.75$; and
4. $\boldsymbol{\psi} = (\gamma, \beta) : (2, 2),\ (2, 5),\ (5, 2)$.

FDR adjustments for FSEB were done using the Benjamini–Hochberg method while the Benjamini–Yekutieli method was used for the Fisher and Score tests, following suggestions in Wang and Ramdas [2022]. Here, the authors prove that Benjamini–Hochberg method is equivalent to the Benjamini–Yekutieli method, when applied to conduct FDR control of p-values obtained from e-value test statistics. That is, because FSEB test statistics are e-values, the FDR control of arbitrarily dependent tests can be conducted using the Benjamini–Hochberg method, which is more powerful than the Benjamini–Yekutieli method.

Table S1: **Case A.** Power comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Numbers averaged over $5-10$ data sets under each setting, within each data set, 50% of points generated with $\theta_{i1}-\theta_{i2}=\Delta$, and the rest with $\theta_{i1}=\theta_{i2}$. Total counts for each binomial series range between 4 and 20.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.000	0.045	0.136	0.000	0.045	0.091	0.136	0.318	0.636	0.045	0.136	0.318	0.182	0.455	0.818	0.136	0.273	0.636
(2,2)	0.75	4	0.100	0.300	0.400	0.000	0.200	0.400	0.400	0.400	0.800	0.200	0.400	0.400	0.400	0.400	0.900	0.300	0.400	0.700
(2,5)	0.60	4	0.000	0.050	0.200	0.000	0.000	0.100	0.250	0.450	0.650	0.100	0.250	0.500	0.300	0.650	0.700	0.250	0.450	0.650
(2,5)	0.75	4	0.100	0.200	0.500	0.000	0.100	0.300	0.300	0.500	0.600	0.200	0.300	0.500	0.500	0.600	0.600	0.200	0.500	0.600
(5,2)	0.60	4	0.000	0.000	0.222	0.000	0.000	0.111	0.222	0.389	0.611	0.056	0.222	0.389	0.278	0.500	0.778	0.056	0.278	0.500
(5,2)	0.75	4	0.167	0.500	0.667	0.167	0.333	0.667	0.333	0.667	1.000	0.333	0.333	0.667	0.333	0.833	1.000	0.333	0.667	1.000
(2,2)	0.60	10	0.050	0.100	0.250	0.000	0.050	0.150	0.200	0.300	0.700	0.050	0.150	0.250	0.250	0.550	0.900	0.050	0.250	0.550
(2,2)	0.75	10	0.240	0.520	0.800	0.160	0.360	0.640	0.600	0.800	0.920	0.280	0.640	0.840	0.680	0.840	1.000	0.360	0.680	0.880
(2,5)	0.60	10	0.120	0.220	0.380	0.040	0.160	0.280	0.240	0.460	0.640	0.120	0.240	0.440	0.320	0.560	0.820	0.200	0.280	0.580
(2,5)	0.75	10	0.150	0.350	0.400	0.000	0.250	0.400	0.350	0.600	0.750	0.200	0.350	0.500	0.450	0.650	0.950	0.200	0.450	0.650
(5,2)	0.60	10	0.080	0.260	0.400	0.020	0.080	0.300	0.300	0.480	0.700	0.100	0.240	0.460	0.360	0.600	0.840	0.140	0.340	0.620
(5,2)	0.75	10	0.171	0.314	0.629	0.086	0.200	0.486	0.429	0.714	0.857	0.200	0.400	0.714	0.514	0.829	1.000	0.257	0.543	0.829
(2,2)	0.60	100	0.075	0.193	0.412	0.007	0.047	0.193	0.253	0.465	0.735	0.031	0.138	0.364	0.345	0.601	0.867	0.044	0.195	0.518
(2,2)	0.75	100	0.315	0.475	0.675	0.190	0.375	0.560	0.495	0.680	0.920	0.275	0.470	0.605	0.540	0.840	0.985	0.310	0.515	0.800
(2,5)	0.60	100	0.051	0.163	0.357	0.000	0.017	0.157	0.223	0.434	0.731	0.009	0.089	0.297	0.294	0.577	0.843	0.009	0.154	0.483
(2,5)	0.75	100	0.251	0.434	0.646	0.106	0.277	0.509	0.454	0.669	0.894	0.194	0.394	0.609	0.551	0.789	0.957	0.234	0.500	0.766
(5,2)	0.60	100	0.085	0.210	0.390	0.007	0.048	0.198	0.268	0.445	0.715	0.007	0.115	0.340	0.328	0.580	0.875	0.018	0.200	0.490
(5,2)	0.75	100	0.250	0.450	0.680	0.160	0.265	0.535	0.495	0.695	0.895	0.230	0.435	0.655	0.585	0.815	0.970	0.250	0.520	0.755

Table S2: **Case A.** Type I error comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Same settings as Table S1.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.000	0.000	0.000	0.000	0.050	0.050	0.000	0.000	0.050
(2,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(5,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(5,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.000	0.000	0.000	0.000	0.000	0.050	0.000	0.000	0.000
(2,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000	0.000	0.000	**0.120**	0.000	0.000	0.000
(2,5)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(5,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(5,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.057**	0.000	0.000	0.000	0.000	0.000	**0.057**	0.000	0.000	0.000
(2,2)	0.60	100	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.002	0.022	0.000	0.000	0.000	0.000	**0.006**	**0.055**	0.000	0.000	0.004
(2,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.037	0.000	0.000	0.000
(2,5)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.027	0.000	0.000	0.000
(2,5)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.026	0.000	0.000	0.000
(5,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.029	0.000	0.000	0.000
(5,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.047	0.000	0.000	0.000

Table S3: **Case B.** Power comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Numbers averaged over $5-10$ data sets under each setting, within each data set, 50% of points generated with $\theta_{i1} - \theta_{i2} = \Delta$, and the rest with $\theta_{i1} = \theta_{i2}$. Total counts for each binomial series range between 15 and 40.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.591	0.818	0.864	0.455	0.773	0.864	0.864	0.909	1.000	0.864	0.864	0.955	0.864	0.955	1.000	0.864	0.909	1.000
(2,2)	0.75	4	0.833	1.000	1.000	0.667	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.60	4	0.250	0.550	0.800	0.150	0.350	0.700	0.650	0.900	1.000	0.550	0.700	0.950	0.750	0.950	1.000	0.650	0.850	0.950
(2,5)	0.75	4	0.833	0.917	1.000	0.750	0.917	1.000	1.000	1.000	1.000	0.917	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.60	4	0.350	0.550	0.800	0.350	0.400	0.750	0.750	0.950	1.000	0.650	0.850	1.000	0.850	1.000	1.000	0.750	0.950	1.000
(5,2)	0.75	4	0.625	1.000	1.000	0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.60	10	0.533	0.711	0.911	0.422	0.622	0.867	0.867	0.956	1.000	0.644	0.911	0.956	0.911	0.978	1.000	0.756	0.933	0.978
(2,2)	0.75	10	0.867	1.000	1.000	0.867	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.60	10	0.473	0.655	0.873	0.291	0.564	0.745	0.745	0.909	0.982	0.618	0.800	0.927	0.818	0.945	1.000	0.709	0.909	0.945
(2,5)	0.75	10	0.833	0.933	0.967	0.833	0.933	0.967	0.967	0.967	1.000	0.933	0.967	1.000	0.967	1.000	1.000	0.933	0.967	1.000
(5,2)	0.60	10	0.540	0.680	0.860	0.420	0.600	0.820	0.840	0.920	0.980	0.660	0.840	0.960	0.840	0.960	0.980	0.700	0.880	0.980
(5,2)	0.75	10	0.900	0.950	0.950	0.850	0.900	0.950	0.950	1.000	1.000	0.950	0.950	1.000	0.950	1.000	1.000	0.950	0.950	1.000
(2,2)	0.60	100	0.551	0.729	0.891	0.445	0.633	0.849	0.831	0.945	0.991	0.622	0.825	0.945	0.891	0.969	0.996	0.675	0.878	0.969
(2,2)	0.75	100	0.894	0.966	0.983	0.860	0.940	0.983	0.980	0.994	1.000	0.934	0.980	0.991	0.983	1.000	1.000	0.946	0.983	1.000
(2,5)	0.60	100	0.576	0.760	0.898	0.451	0.664	0.851	0.833	0.951	0.998	0.649	0.822	0.949	0.884	0.971	1.000	0.707	0.873	0.971
(2,5)	0.75	100	0.887	0.940	0.990	0.860	0.920	0.983	0.980	0.993	1.000	0.917	0.973	0.993	0.993	1.000	1.000	0.937	0.993	1.000
(5,2)	0.60	100	0.506	0.662	0.826	0.362	0.588	0.770	0.768	0.890	0.976	0.586	0.736	0.886	0.818	0.928	0.996	0.624	0.810	0.916
(5,2)	0.75	100	0.824	0.928	0.992	0.764	0.892	0.964	0.968	1.000	1.000	0.896	0.964	1.000	0.992	1.000	1.000	0.912	0.992	1.000

Table S4: **Case B.** Type I error comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Same settings as Table S3.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.083**	0.000	0.000	0.000
(5,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.100**	0.000	0.000	0.000
(5,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.125**	0.000	0.000	0.000
(2,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.022	0.000	0.000	0.000
(2,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.133**	0.000	0.000	**0.067**	0.000	**0.067**	**0.133**	0.000	0.000	**0.067**
(2,5)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.055**	0.000	0.000	0.000
(2,5)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.100**	0.000	0.000	0.000
(5,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.000	0.000
(5,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.100**	0.000	0.000	0.000
(2,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.016	0.000	0.000	0.005	0.000	**0.007**	0.050	0.000	0.000	0.007
(2,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000	0.000	**0.051**	0.000	0.000	0.000
(2,5)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.000	0.002	0.027	0.000	0.000	0.002
(2,5)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.013	0.000	0.000	0.000	0.000	0.000	0.045	0.000	0.000	0.000
(5,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.016	0.000	0.000	0.002	0.000	**0.006**	0.036	0.000	0.000	0.004
(5,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.000	0.000	**0.068**	0.000	0.000	0.000

Table S5: **Case C.** Power comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Numbers averaged over $5-10$ data sets under each setting, within each data set, 50% of points generated with $\theta_{i1}-\theta_{i2}=\Delta$, and the rest with $\theta_{i1}=\theta_{i2}$. Total counts for each binomial series range between 5 and 200.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.909	0.909	0.955	0.909	0.909	0.955	0.955	0.955	1.000	0.955	0.955	0.955	0.955	0.955	1.000	0.955	0.955	1.000
(2,2)	0.75	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.60	4	0.929	0.929	0.929	0.857	0.929	0.929	0.929	0.929	1.000	0.929	0.929	0.929	0.929	0.929	1.000	0.929	0.929	1.000
(2,5)	0.75	4	0.917	1.000	1.000	0.917	0.917	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.60	4	0.700	0.800	0.850	0.700	0.800	0.850	0.850	0.850	0.950	0.850	0.850	0.900	0.850	0.900	1.000	0.850	0.850	0.950
(5,2)	0.75	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.60	10	0.975	1.000	1.000	0.975	0.975	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.75	10	0.950	0.950	0.950	0.925	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.975	0.950	0.950	0.950
(2,5)	0.60	10	0.914	0.914	0.971	0.914	0.914	0.943	0.914	1.000	1.000	0.914	0.943	1.000	0.943	1.000	1.000	0.914	0.971	1.000
(2,5)	0.75	10	0.967	0.967	0.967	0.900	0.967	0.967	0.967	0.967	1.000	0.967	0.967	0.967	0.967	0.967	1.000	0.967	0.967	1.000
(5,2)	0.60	10	0.780	0.840	0.940	0.780	0.820	0.940	0.880	0.960	0.960	0.840	0.900	0.960	0.920	0.960	0.980	0.840	0.940	0.960
(5,2)	0.75	10	0.920	0.960	1.000	0.880	0.960	0.960	1.000	1.000	1.000	0.960	1.000	1.000	1.000	1.000	1.000	0.960	1.000	1.000
(2,2)	0.60	100	0.891	0.917	0.937	0.886	0.909	0.931	0.926	0.949	0.983	0.906	0.923	0.940	0.931	0.963	0.991	0.914	0.931	0.963
(2,2)	0.75	100	0.927	0.933	0.980	0.913	0.933	0.973	0.933	0.980	0.987	0.933	0.933	0.973	0.973	0.987	0.993	0.933	0.973	0.987
(2,5)	0.60	100	0.890	0.920	0.946	0.882	0.908	0.936	0.932	0.958	0.976	0.908	0.932	0.958	0.940	0.968	0.990	0.920	0.938	0.968
(2,5)	0.75	100	0.972	0.980	0.988	0.968	0.976	0.984	0.980	0.988	0.996	0.972	0.980	0.988	0.984	0.992	1.000	0.976	0.984	0.992
(5,2)	0.60	100	0.878	0.900	0.932	0.876	0.886	0.920	0.914	0.938	0.966	0.886	0.914	0.934	0.924	0.948	0.976	0.892	0.922	0.948
(5,2)	0.75	100	0.932	0.944	0.968	0.924	0.944	0.968	0.948	0.976	0.992	0.940	0.948	0.976	0.960	0.980	0.992	0.940	0.956	0.980

Table S6: **Case C.** Type I error comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Same settings as Table S5.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.045	0.000	0.000	0.000	0.000	0.000	0.045	0.000	0.000	0.045
(2,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.167**	0.000	0.000	0.000
(2,5)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.083**	0.000	0.000	0.000	0.000	0.000	**0.083**	0.000	0.000	**0.083**
(5,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.100**	0.000	0.000	0.000	0.000	0.000	**0.200**	0.000	0.000	0.050
(5,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.050**	0.050	0.000	0.000	0.050	0.000	**0.050**	**0.100**	0.000	0.000	0.050
(2,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.025	0.000	0.000	0.000	0.000	0.000	**0.075**	0.000	0.000	0.000
(2,5)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.057**	0.000	0.000	0.000
(2,5)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.033**	**0.067**	0.000	0.000	0.033	0.000	**0.033**	**0.067**	0.000	0.000	0.033
(5,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.120**	0.000	0.000	0.000	0.000	0.000	**0.180**	0.000	0.000	0.000
(5,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000
(2,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.037	0.000	0.000	0.003	0.000	0.003	0.049	0.000	0.000	0.003
(2,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.033	0.000	0.000	0.000
(2,5)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.032	0.000	0.000	0.002	0.000	0.002	0.042	0.000	0.000	0.002
(2,5)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	**0.004**	0.004	0.048	0.000	0.004	0.004	**0.004**	0.004	**0.076**	0.000	0.004	0.004
(5,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000	0.002	**0.058**	0.000	0.000	0.002
(5,2)	0.75	100	0.000	0.004	0.004	0.000	0.000	0.004	**0.004**	**0.008**	0.028	0.000	0.004	0.008	**0.004**	**0.008**	0.041	0.000	0.004	0.008

Table S7: **Case D.** Power comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Numbers averaged over $5-10$ data sets under each setting, within each data set, 50% of points generated with $\theta_{i1}-\theta_{i2}=\Delta$, and the rest with $\theta_{i1}=\theta_{i2}$. Total counts for each binomial series range between 50 and 300.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.75	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.60	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.75	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.60	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.75	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.60	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.75	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.60	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.75	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.60	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.75	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.60	100	0.998	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,2)	0.75	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.60	100	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2,5)	0.75	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.60	100	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(5,2)	0.75	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table S8: **Case D.** Type I error comparisons of tests (with and without FDR corrections) with varying N, Effect size and ψ. Same settings as Table S7.

ψ	Δ	N	FSEB			FSEB-FDR			Fisher			Fisher-FDR			Score			Score-FDR		
			0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05	0.0005	0.005	0.05
(2,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.045	0.000	0.000	0.000
(2,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,5)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.042	0.000	0.000	0.000	0.000	0.000	0.042	0.000	0.000	0.000
(2,5)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(5,2)	0.60	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(5,2)	0.75	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(2,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.057**	0.000	0.000	0.029	0.000	0.000	**0.057**	0.000	0.000	**0.057**
(2,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.120**	0.000	0.000	0.000	0.000	0.000	**0.120**	0.000	0.000	0.000
(2,5)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.025	0.000	0.000	0.000	0.000	0.000	**0.075**	0.000	0.000	0.000
(2,5)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000
(5,2)	0.60	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.057**	0.000	0.000	0.000
(5,2)	0.75	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000	0.000	0.000	**0.080**	0.000	0.000	0.000
(2,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.037	0.000	0.000	0.005	0.000	**0.007**	0.045	0.000	0.000	0.007
(2,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028	0.000	0.000	0.000	0.000	0.000	0.044	0.000	0.000	0.000
(2,5)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.044	0.000	0.000	0.002	0.000	0.002	0.047	0.000	0.000	0.002
(2,5)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.043	0.000	0.000	0.000
(5,2)	0.60	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.033	0.000	0.000	0.002	0.002	0.004	**0.053**	0.000	0.002	0.004
(5,2)	0.75	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.006**	**0.060**	0.000	0.000	0.006	0.000	**0.009**	**0.083**	0.000	0.000	0.009

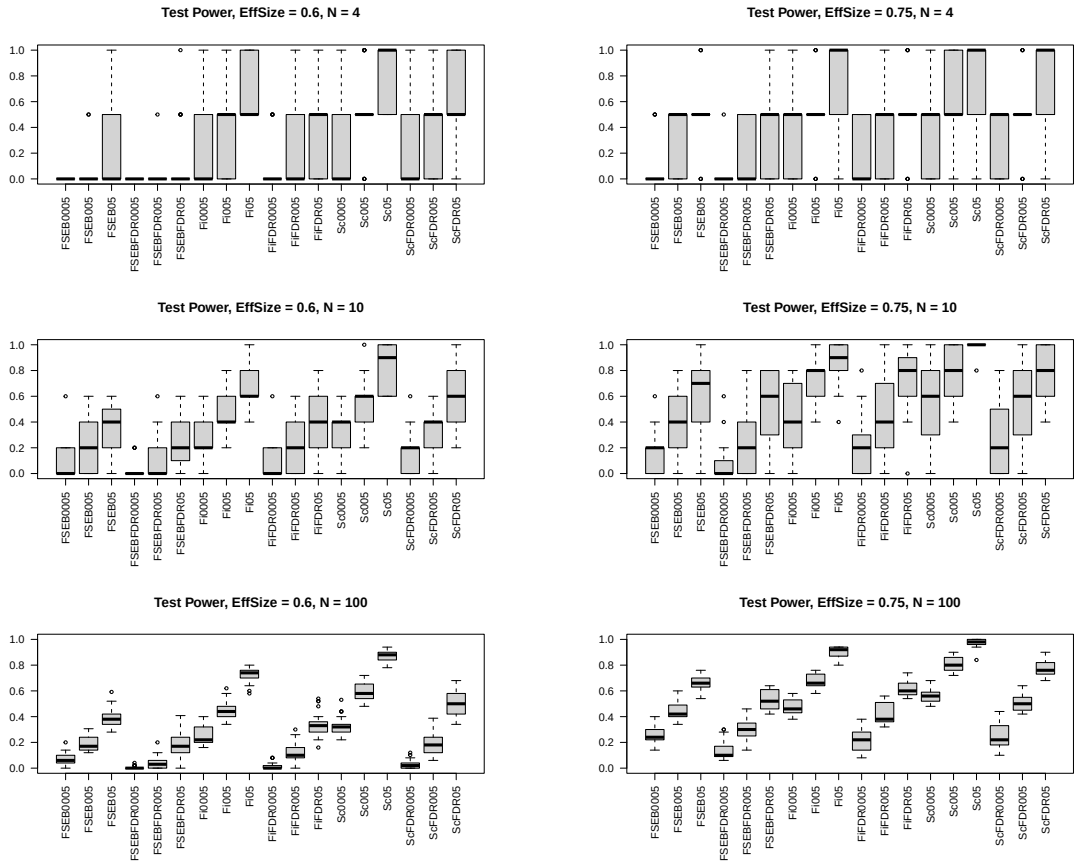

Figure S1: **Case A.** Power of tests compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 4 and 20. P-values calculated are based on the FSEB, Fisher test (Fi), and Score test (Sc), at levels of significance 0.0005, 0.005 and .05. (Values in this figure correspond to Table S1.)

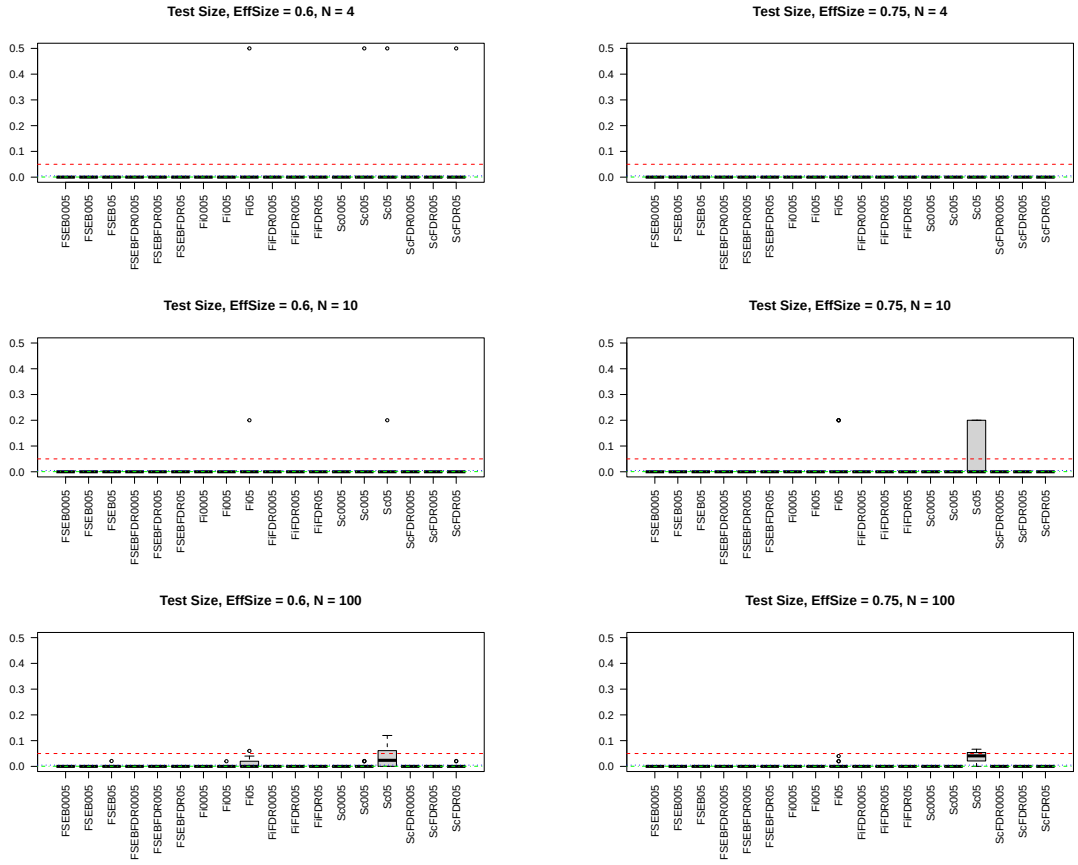

Figure S2: **Case A.** Type I errors compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 4 and 20. The horizontal lines correspond to levels of significance $\alpha = 0.0005$ (green), $\alpha = 0.005$ (blue) and $\alpha = 0.05$ (red). (This figure corresponds to Table S2.)

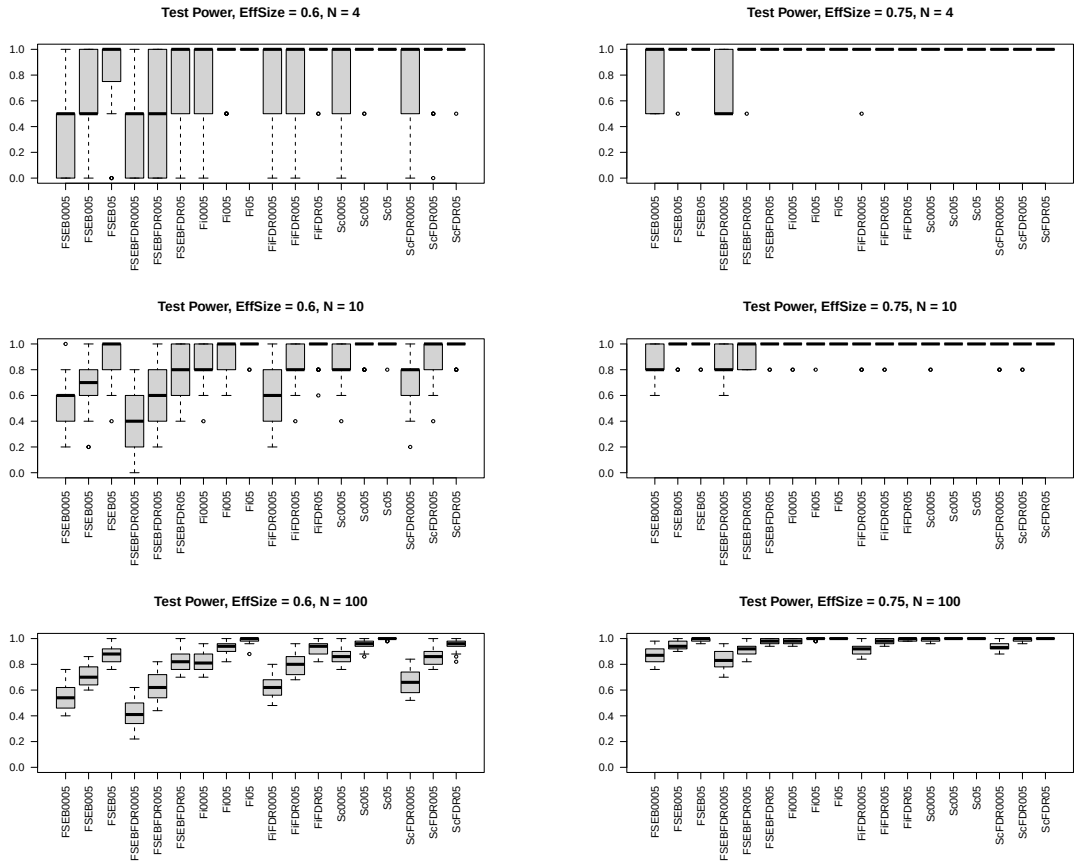

Figure S3: **Case B.** Power of tests compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 15 and 40. (Corresponds to Table S3.)

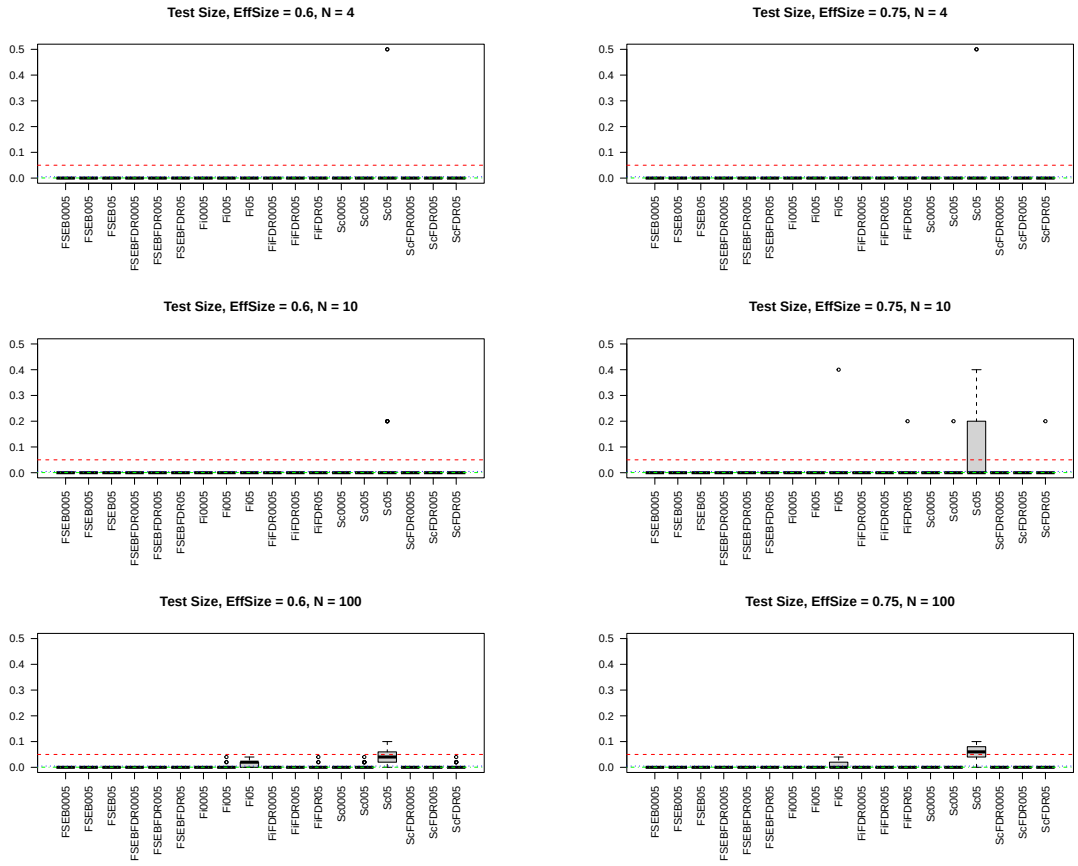

Figure S4: **Case B.** Type I errors compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 15 and 40. (Corresponds to Table S4.)

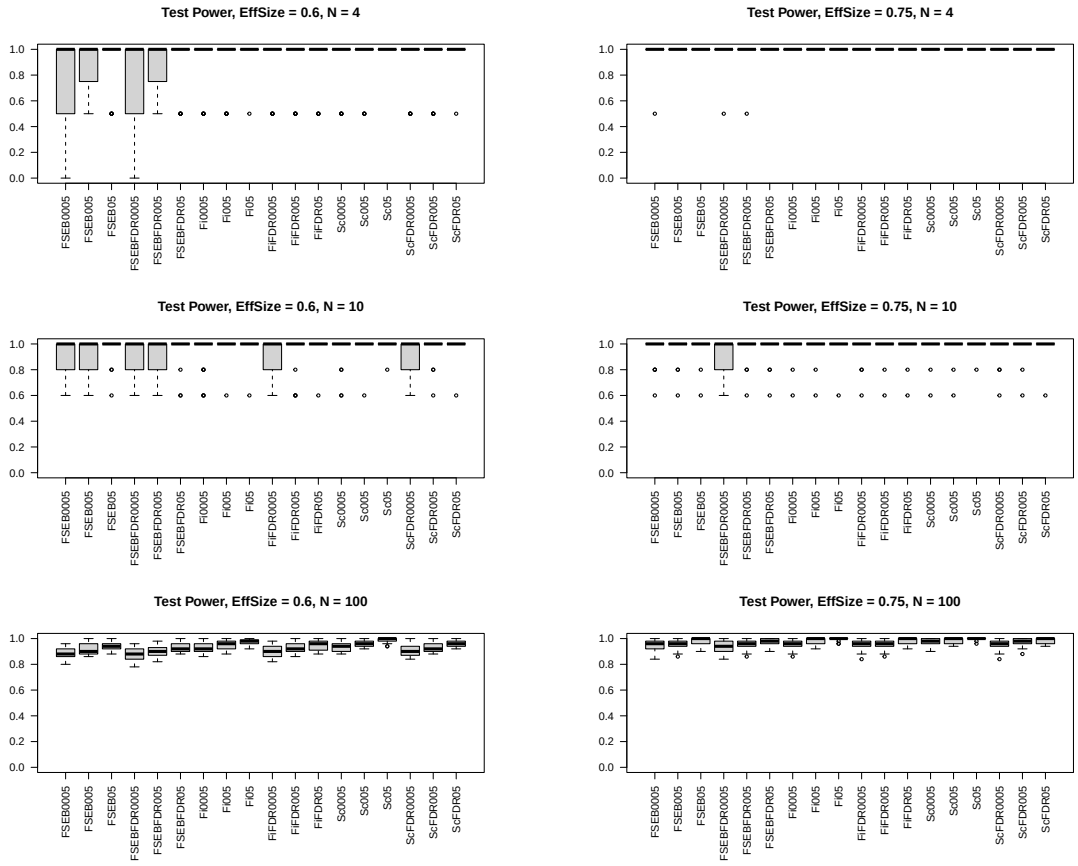

Figure S5: **Case C.** Power of tests compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 5 and 200. (Corresponds to Table S5.)

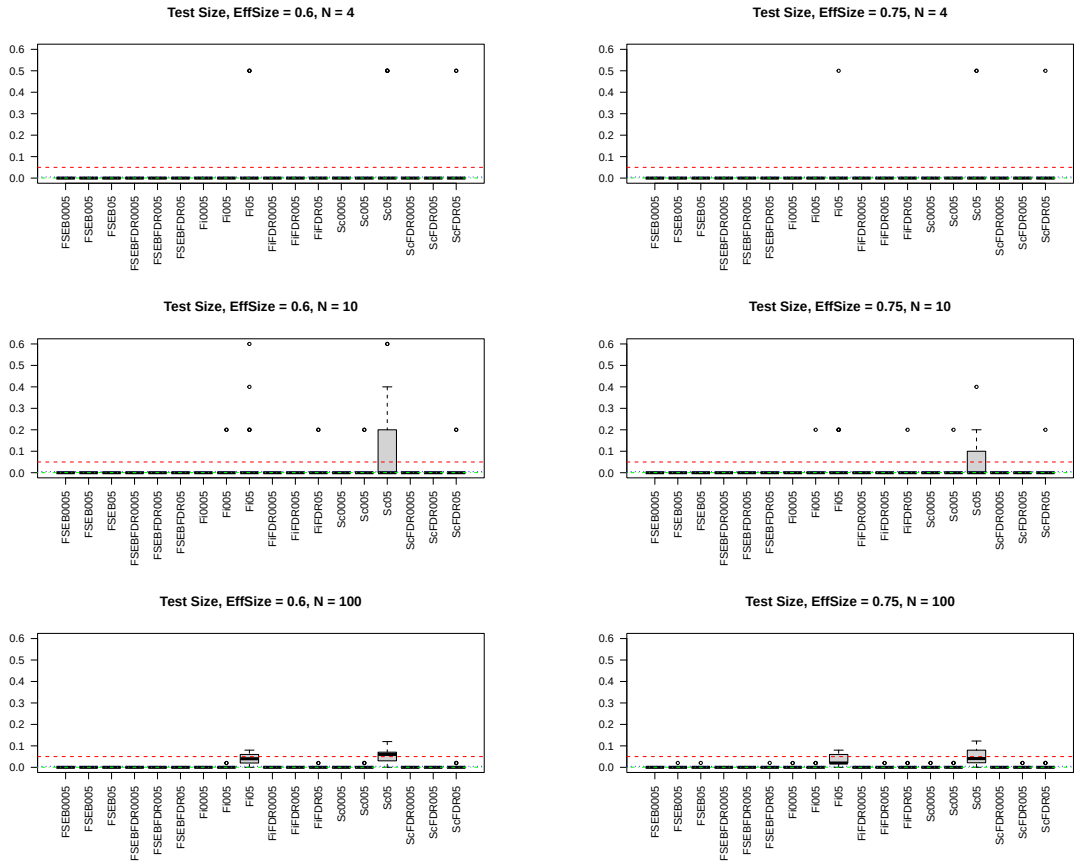

Figure S6: **Case C.** Type I errors compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 5 and 200. (Corresponds to Table S6.)

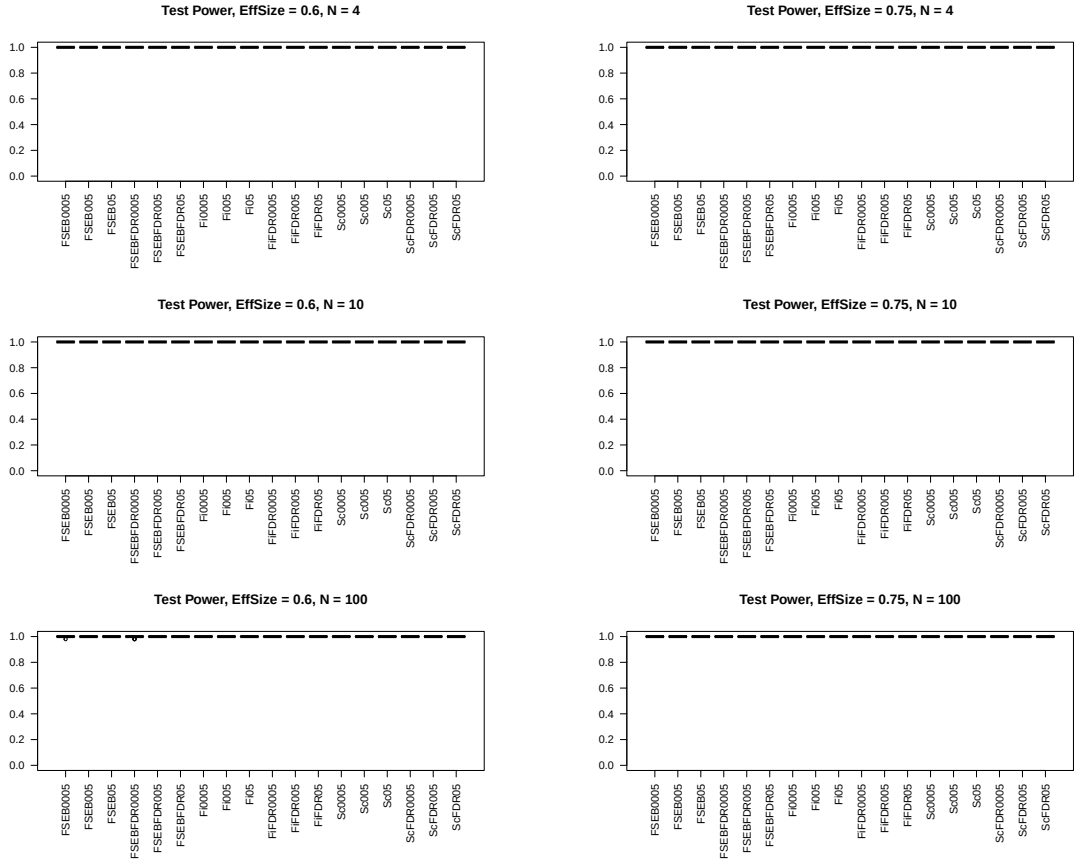

Figure S7: **Case D.** Power of tests compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 50 and 300. (Corresponds to Table S7.)

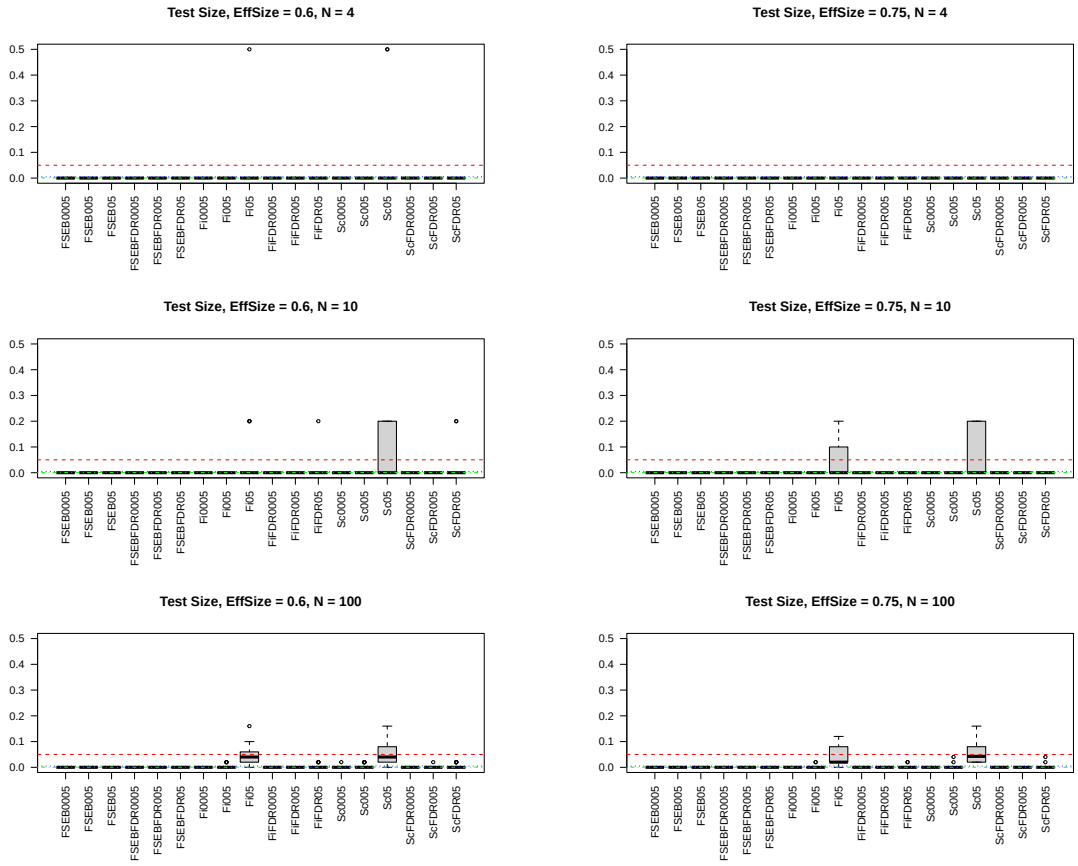

Figure S8: **Case D.** Type I errors compared (with and without FDR corrections) for three methods under varying N and effect sizes. Total counts for each binomial series range between 50 and 300. (Corresponds to Table S8.)

References

R Wang and A Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society B*, 84:822–852, 2022.