Yu and Zelterman *Journal of Statistical Distributions and Applications* (2020) 7:9
https://doi.org/10.1186/s40488-020-00109-6

Journal of Statistical
Distributions and Applications

RESEARCH Open Access

Distributions associated with simultaneous multiple hypothesis testing

Chang Yu[1] and Daniel Zelterman[2*]

*Correspondence:
daniel.zelterman@yale.edu
[2]Department of Biostatistics, Yale University, 06520 New Haven, CT, United States
Full list of author information is available at the end of the article

Abstract

We develop the distribution for the number of hypotheses found to be statistically significant using the rule from Simes (Biometrika 73: 751–754, 1986) for controlling the family-wise error rate (FWER). We find the distribution of the number of statistically significant *p*-values under the null hypothesis and show this follows a normal distribution under the alternative. We propose a parametric distribution $\Psi_I(\cdot)$ to model the marginal distribution of *p*-values sampled from a mixture of null uniform and non-uniform distributions under different alternative hypotheses. The Ψ_I distribution is useful when there are many different alternative hypotheses and these are not individually well understood. We fit Ψ_I to data from three cancer studies and use it to illustrate the distribution of the number of notable hypotheses observed in these examples. We model dependence in sampled *p*-values using a latent variable. These methods can be combined to illustrate a power analysis in planning a larger study on the basis of a smaller pilot experiment.

Keywords: Bonferroni correction, Simes criteria, False discovery rate, *p*-values

1 Introduction

Much work in informatics is concerned with identifying and classifying statistically significant biological markers. In this work we develop methods for describing the distribution of the numbers of such events. Informatics methods often summarize experiments resulting in a large number of p-values, usually through multiple comparisons of gene expression data. Typically, the number of tests m, is much greater than the number of subjects, N. There are several important rules for identifying statistically significant p-values while maintaining the significance level below a pre-specified level α $(0 < \alpha < 1)$. Benjamini (2010) provides a review of recent advances.

A commonly cited rule to control the FWER is the Bonferroni correction. Given a sample of ordered p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, the Bonferroni rule finds the smallest value of $B = 0, 1, \ldots, m-1$ for which

$$p_{(\mathrm{B}+1)} > \alpha / m \ . \tag{1}$$

The Simes (1986) rule chooses the smallest value $S = 0, 1, \ldots, m$ such that

$$p_{(S+1)} > (S+1)\, \alpha / m \tag{2}$$

to control the FWER $\leq \alpha$.

A similar rule developed by Benjamini and Hochberg (1995) to maintain the false discovery rate (FDR) $\leq \alpha$ finds the largest value of BH such that

$$p_{(BH)} < BH\,\alpha/m\ .$$

This reference shows procedures controlling the FWER also control the FDR, but procedures controlling FDR only control FWER in a weaker sense.

We will concentrate on the distribution of B and S in this report. We describe the probability distribution of B and S under null hypotheses where each p-value has an independent marginal uniform distribution as well as an approximating distribution under the alternative hypothesis with density function $\psi_I(p)$ expressible as a polynomial in $\log p$ of order I.

There has been limited research on parametric distributions for the p-values generated from data under a mixture of the null and different distributions under multiple alternative hypotheses. The mixed p-values are mainly modeled using non-parametric methods (Genovese and Wasserman 2004; Broberg 2005; Langaas et al. 2005; Tang et al. 2007) or alternatively, the p-values are converted into normal quantiles and modeled thereafter (Efron et al. 2001; Efron 2004; Jin and Cai 2007). Another common approach is to approximate the distribution of sampled p-values using a mixture of beta distributions (Pounds and Morris 2003; Broberg 2005; Tang et al. 2007). Other parametric models have been described by Kozoil and Tuckwell (1999); Genovese and Wasserman (2004); Yu and Zelterman (2017, 2019).

Of interest is the fraction π_0 of p-values sampled from the uniform distribution under the null hypothesis. Langaas et al. (2005) and Tang et al. (2007) suggest the estimated density of p-values at $p = 1$ be used to estimate the fraction π_0. Estimating π_0 is of practical importance: The BH statistic controls the FDR no more than $\alpha\pi_0$. Consequently, Benjamini and Hochberg (2000) recommend we perform tests with significance level α/π_0 and still maintain the FDR below α. We found $\psi_I(1 \mid \hat{\boldsymbol{\theta}})$ to be a useful estimator of π_0 in the examples of Section 5, where $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate.

The p-values are usually not independent. In microarray studies, for example, a small number of clusters of p-values in the same biological pathway may have high mutual correlations. Methods for modeling such dependencies are developed by Sun and Cai (2009), Friguet et al. (2009), and Wu (2008) for examples.

In Section 2, we describe the probability distribution of S in (2) when the p_i are independently sampled from an unspecified distribution Ψ. In Section 3 we examine p-values sampled from a uniform distribution under the null hypothesis. Section 4 provides elementary properties of the proposed distribution Ψ_I. The parameters $\boldsymbol{\theta}$ of Ψ_I depend on the specific application and are estimated for two examples in Section 5. In Section 6, we model the distribution of dependent p-values using a latent variable. We combine these methods in Section 7 to illustrate approximate power in planning a proposed study. We provide mathematical details of Sections 2 and 3 in Appendix A. Appendix B examines the behavior of B and S under a close sequence of alternative hypotheses. Appendix C examines the parameter space for the Ψ_I distribution.

2 Simultaneous multiple testing

Let $p_1, p_2, \ldots, p_m$ denote m randomly sampled $p-$values with ordered values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. We will initially assume all p-values are independent and have the same distribution function denoted by $\Psi(\cdot)$ with corresponding density function $\psi(\cdot)$. In Section 6, we return to the assumption of independence. We propose a non-uniform approximation for Ψ in Section 4.

If we follow the Bonferroni rule (1) then the distribution of the number B of statistically significant p-values at FWER $\leq \alpha$ follows a binomial distribution with index m and probability parameter equal to $\Psi(\alpha/m)$.

The distribution of S can be obtained by writing

$$\begin{aligned}\Pr[\, S = k \,] &= \Pr\left[\bigcap_{j=1}^{k} \{ p_{(j)} < j\alpha/m \} \quad \text{and} \quad p_{(k+1)} > (k+1)\alpha/m \right] \\ &= \frac{m!}{(m-k)!} \, [\, 1 - \Psi((k+1)\alpha/m) \,]^{m-k} \quad U_k \,, \end{aligned} \tag{3}$$

where $U_0 = 1$ and

$$U_k \;=\; \int_{p_1=0}^{\alpha/m} \int_{p_2=p_1}^{2\alpha/m} \cdots \int_{p_k=p_{k-1}}^{k\alpha/m} \psi(p_1) \cdots \psi(p_k) \, \mathrm{d}p_k \ldots \mathrm{d}p_2 \, \mathrm{d}p_1 \tag{4}$$

for $k = 1, 2, \ldots, m$.

In Appendix A we prove

$$U_k = \sum_{i=1}^{k} (-1)^{i+1} \, \Psi^i\{(k-i+1)\alpha/m\} \, U_{k-i}/\, i! \;. \tag{5}$$

The p-values are typically sampled from a mixture of a uniform distribution under the null hypothesis and several distributions under different alternative hypotheses. Similarly, the distribution of S will be a mixture of a mass near zero and a normal distribution, described next. This mixture distribution is illustrated in Figs. 2 and 4 for two examples of Section 5.

Specifically, for values of S near zero we have

$$\begin{aligned}\Pr[\, S = 0 \,] &= \{1 - \Psi(\alpha/m)\}^m \,, \\ \Pr[\, S = 1 \,] &= m \, \{1 - \Psi(2\alpha/m)\}^{m-1} \, \Psi(\alpha/m) \,, \\ \Pr[\, S = 2 \,] &= \binom{m}{2} \{1 - \Psi(3\alpha/m)\}^{m-2} \, \Psi(\alpha/m) \, \{2\Psi(2\alpha/m) - \Psi(\alpha/m)\} \,. \end{aligned}$$

To describe the behavior of S away from zero, begin with the quantile function $\Psi^{-1}(i/(m+1))$ giving the approximate expected value of the order statistic $p_{(i)}$. If m is large and i/m is not too close to either zero or one, then $p_{(i)}$ will be approximately normally distributed. In (2), S is the smallest value of k for which $p_{(k+1)} > (k+1)\alpha/m$. This should occur for values of S with mean μ solving

$$\Psi^{-1}((\mu+1)/(m+1)) = (\mu+1)\alpha/m \,,$$

or equivalently,

$$\Psi((\mu+1)\alpha/m) = (\mu+1)/(m+1) \;. \tag{6}$$

If we write $S = mp_{(\mu)}/\alpha$ for integer μ and use the large sample approximation to an order statistic, then the approximate variance of S is

$$\frac{\mu(m-\mu)}{\alpha^2\, m\, [\psi(\alpha(\mu+1)\,/\,m)]^2}\ .$$

If the null (uniform) and alternative hypotheses are not very different from each other, then the solution to μ in (6) will be close to zero and Appendix B describes a different approximation to the behavior of B and S.

3 Behavior under the null hypothesis

Let us next examine the special case where all p-values are independently sampled under the null hypothesis. When the distribution of the p_i are independent and marginally uniformly distributed then (3) and (5) are expressible as

$$\begin{aligned}\Pr[\ \mathrm{S}=0\,] &= (1-\alpha/m)^m\ ,\\ \Pr[\ \mathrm{S}=1\,] &= \alpha\,(1-2\alpha/m)^{m-1}\ ,\\ \Pr[\ \mathrm{S}=2\,] &= 3/2\ \{(m-1)/m\}\,\alpha^2\,(1-3\alpha/m)^{m-2}\ ,\end{aligned}$$

and in general,

$$\Pr[\ \mathrm{S}=k\,] = \binom{m}{k}\ (k+1)^{k-1}\ (\alpha/m)^k\ \{1-(k+1)\alpha/m\}^{m-k}\ . \tag{7}$$

Details of the derivation of (7) appear in Appendix A.

Useful results can be obtained if we also assume the number of hypotheses m is large. The limiting distribution (7) of S, is

$$\Pr[\ \mathrm{S}=k\mid\alpha\,] = \{(k+1)^{k-1}/\,k!\ \}\ \alpha^k\ e^{-(k+1)\alpha} \tag{8}$$

for $k = 0, 1, \ldots$.

The probabilities in (8) sum to unity using equation (130) in Jolley (1961, p. 24). The mean of this distribution is $\alpha/(1-\alpha)$ and the variance is $\alpha/(1-\alpha)^3$. The distribution of $S+1$ in (8) is known as the Borel distribution with applications in queueing theory (Tanner, 1961). Similarly, for large values of m, the number of identified p-values at FWER $\leq \alpha$ for the Bonferroni criteria (1) will follow a Poisson distribution with mean α when sampling p-values under the null hypothesis.

4 Distributions for P—Values

We next propose a marginal distribution Ψ for p-values, independent of the choice of test statistic. We continue to assume the p-values are mutually independent and have the same marginal distributions. We must have Ψ concave (Genovese and Wasserman 2004; Sun and Cai 2009), otherwise the underlying test will have power smaller than its significance level for some α. Similarly, the corresponding density function ψ must be monotone decreasing. We next propose a flexible distribution for modeling the distribution of p-values under alternative hypotheses.

Consider a distribution with a density function expressible as a polynomial in $\log p$ up to degree $I = 0, 1, 2, \ldots$. The uniform (0–1) distribution is obtained for $I = 0$. The marginal density function we propose for p-values is

$$\psi_I(p\mid\boldsymbol{\theta}) = \sum_{i=0}^{I}\ \theta_i\,(-\log p)^i \tag{9}$$

for real-valued parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_I\}$ with $I \geq 1$ where

$$\theta_0 = 1 - \sum_{i=1}^{I} i!\,\theta_i \ , \tag{10}$$

so the densities $\psi_I(p)$ integrate to one over $0 < p \leq 1$. Similarly, θ_0 is not an independent parameter.

The corresponding cumulative distribution function is

$$\Psi_I(p \mid \boldsymbol{\beta}) = p \sum_{i=0}^{I} \beta_i \, (-\log p)^i \ , \tag{11}$$

where $\beta_0 = 1$.

The relationship between these parameters is linear:

$$\beta_j = \sum_{i=j}^{I} \theta_i \, i! \, / j!$$

for $j = 1, 2, \ldots, I$ and $\theta_i = \beta_i - (i+1)\beta_{i+1}$ for $i = 1, 2, \ldots, I-1$. Throughout, we will interchangeably refer to either the $\boldsymbol{\theta}$ or $\boldsymbol{\beta}$ parameterizations for simplicity.

The moments of distribution $\psi_I(p \mid \boldsymbol{\theta})$ are

$$\mathrm{E}(p^j \mid \boldsymbol{\theta}) = \sum_{i=0}^{I} i! \ \theta_i \ / \ (j+1)^{i+1} \ , \tag{12}$$

for $j = 1, 2, \ldots$.

We must have $\theta_I > 0$ in order to have $\psi_I(p) > 0$ for values of p close to zero. Values of θ_0 are restricted in (10) in order for $\psi_I(p)$ to integrate to unity. Since $\psi_I(1 \mid \boldsymbol{\theta}) = \theta_0$ we must also require $\theta_0 \geq 0$. Requiring $\psi_I(p)$ to be decreasing at $p = 1$ gives $\theta_1 \geq 0$.

These restrictions alone on θ_0, θ_1, and θ_I are not sufficient to guarantee $\psi_I(p \mid \boldsymbol{\theta})$ is monotone decreasing or positive valued for all values of $0 \leq p \leq 1$. The necessary conditions for achieving these properties are difficult to describe in general, but sufficient conditions are all $\theta_i \geq 0$. Specific cases are examined in Appendix C for values of I up to $I = 4$. Models for larger values of I could be fitted by maximizing the penalized likelihood, such that $\psi_I(p \mid \boldsymbol{\theta})$ is positive valued and monotone decreasing at the observed, sorted p-values.

In practice, the choice of I is found by fitting a sequence of models. Successive values of I represent nested models so twice the differences of the respective log-likelihoods will behave as χ^2 (1 df) when the underlying additional parameter value is zero. In practice, we found $I = 3$ or 4 were adequate for the three examples in this work.

The ψ_I density function is specially suited for modeling the marginal distribution of a uniform and a variety of non-uniform distributions for p-values. If each p_i $(i = 1, \ldots, m)$ is sampled from a different distribution with density function $\psi_I(p \mid \boldsymbol{\theta}_i)$, then the marginal density of all p_i satisfies

$$m^{-1} \sum_{i}^{m} \psi_I(p \mid \boldsymbol{\theta}_i) = \psi_I(p \mid \overline{\boldsymbol{\theta}}), \tag{13}$$

where $\overline{\boldsymbol{\theta}}$ is the arithmetic average of all $\boldsymbol{\theta}_i$. A similar result holds if the values of I vary across distributions of p_i.

This mixing of distributions includes the uniform as a special case. Specifically, suppose $100\pi_0$−percent of the p-values are sampled from a uniform (0, 1) distribution $(0 \leq \pi_0 \leq$

1) and the remaining $100(1-\pi_0)-$percent are sampled from $\psi_I(p \mid \boldsymbol{\theta})$. Then the marginal distribution has density function

$$\pi_0 + (1-\pi_0)\,\psi_I(p \mid \boldsymbol{\theta}) \;=\; \psi_I(p \mid (1-\pi_0)\boldsymbol{\theta})\,, \tag{14}$$

demonstrating π_0 is not identifiable in this model.

Equations (13) and (14) illustrate the utility of ψ_I in modeling p-values sampled from a mixture of the null hypothesis and different distributions under alternative hypotheses, yet retaining the same parametric distribution form. Donoho and Jin (2004) also describe the value of such a mixture of heterogeneous alternative hypotheses in multiple testing settings. Following Langaas et al. (2005); Tang et al. (2007) we use $\psi_I(p=1 \mid \hat{\boldsymbol{\theta}}) = \hat{\theta}_0$, the estimated density at $p=1$, to estimate π_0, the proportion of p-values sampled from the null hypothesis.

5 Two examples

For each of the examples in this work, we fitted the density function ψ_I described in Section 4 and then used this model to examine the distribution of S given in (3). The fitted parameter values $\hat{\boldsymbol{\theta}}$ for these examples are given for successive values of I. We maximized the likelihoods using standard optimization routine *nlm* in R. This routine also provides estimates of the Hessian used to estimate standard errors of parameter estimates.

The evaluation of U_k in (5) involves adding and subtracting many nearly equal values resulting in numerical instability. We computed U_k using multiple precision arithmetic with the *Rmpfr* package in R (Maechler 2019). A third example will be introduced in Section 7, to illustrate estimation of power for multiple hypothesis testing problems.

5.1 Breast cancer

This microarray dataset was originally described by Hedenfalk et al. (2001) and also analyzed by Storey and Tibshirani (2003). These data summarize marker expressions of $m=3226$ genes in seven women with the BRCA1 mutation and in eight women with the BRCA2 mutation. The objective was to determine differentially-expressed genes between these two groups. Earlier analyses used a two-sample t-test to compare the two groups for each gene, giving rise to m p-values. Efron (2004) and Jin and Cai (2007) model the z-scores corresponding to the p-values.

Fitted parameters are given in Table 1. The fitted model for $I=2$ represents a big improvement over the model with $I=1$ parameter. The model with $I=3$ parameters has a modest improvement over the model with $I=2$ and $I=4$ demonstrates negligible change in the likelihood over $I=3$. Fitted densities ψ_I for $I=2$ and 3 are plotted in Fig. 1 along with the observed data. There is only a small difference between the fitted models in this figure, and both exhibit a good fit to the data. Our estimate of π_0 given by $\hat{\theta}_0$ is .65 for $I=2$ and .62 for $I=3$. An estimate of .67 for π_0 is described in Storey and Tibshirani (2003).

There are $S=29$ statistically significant markers at FWER $=$ .05 using the adjustment for multiplicity given in (2). The fitted distribution of S is displayed in Fig. 2 using $\psi_3(\cdot \mid \hat{\boldsymbol{\theta}})$. The mean of this fitted distribution is 22.75. The distribution in Fig. 2 appears as a mixture of a distribution concentrated near $k=0$ and a left-truncated normal distribution with a local mode at 24. The observed value $S=29$ is indicated in this figure.

Table 1 Maximum likelihood estimated parameter values of ψ_I for the breast cancer data

Model parameters				Log-	2×	$\hat{\theta}_0$ to
I	Symbol	Estimate	Std Err	Likelihood	Difference	estimate π_0
1	θ_1	0.531	0.018	482.04	—	0.469
2	θ_1	0.0	0.049	569.04	174.0	0.649
	θ_2	0.177	0.015			
3	θ_1	0.158	0.084	573.27	8.47	0.623
	θ_2	0.0492	0.0506			
	θ_3	0.0201	0.0075			
4				Same as $I = 3$		

The point mass at $S = 0$ is about 0.1 and values of $S \leq 3$ account for about 20% of the distribution with $I = 3$ and fitted $\hat{\boldsymbol{\theta}}$. This distribution is approximately a mixture of the distribution near zero and 80% of a normal with mean 26.1 and standard deviation 17.9 using (6).

5.2 The cancer genome atlas: lung cancer

This dataset contains the summary of an extensive database collected on tumors from $N = 178$ patients with squamous cell lung carcinoma. A full description of these data and

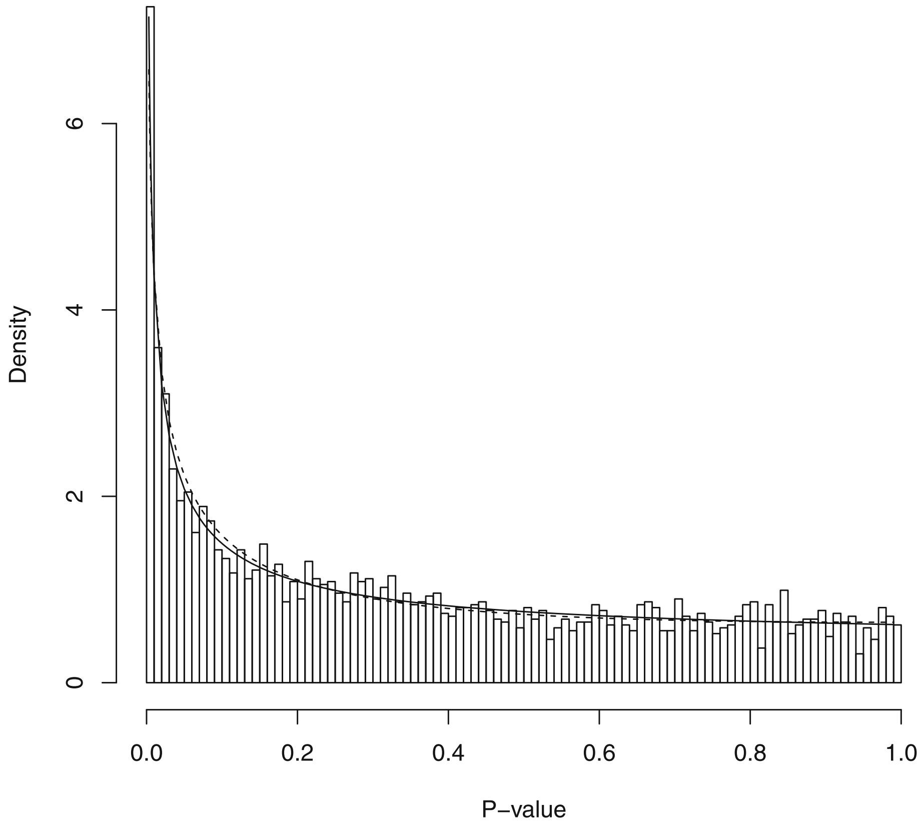

Fig. 1 Histogram and fitted density function ψ_I for the 3226 biomarker *p*-values from the breast cancer data with $I = 2$ (dashed line) and $I = 3$ (solid line). A histogram for these data also appears in Storey and Tibshirani (2003)

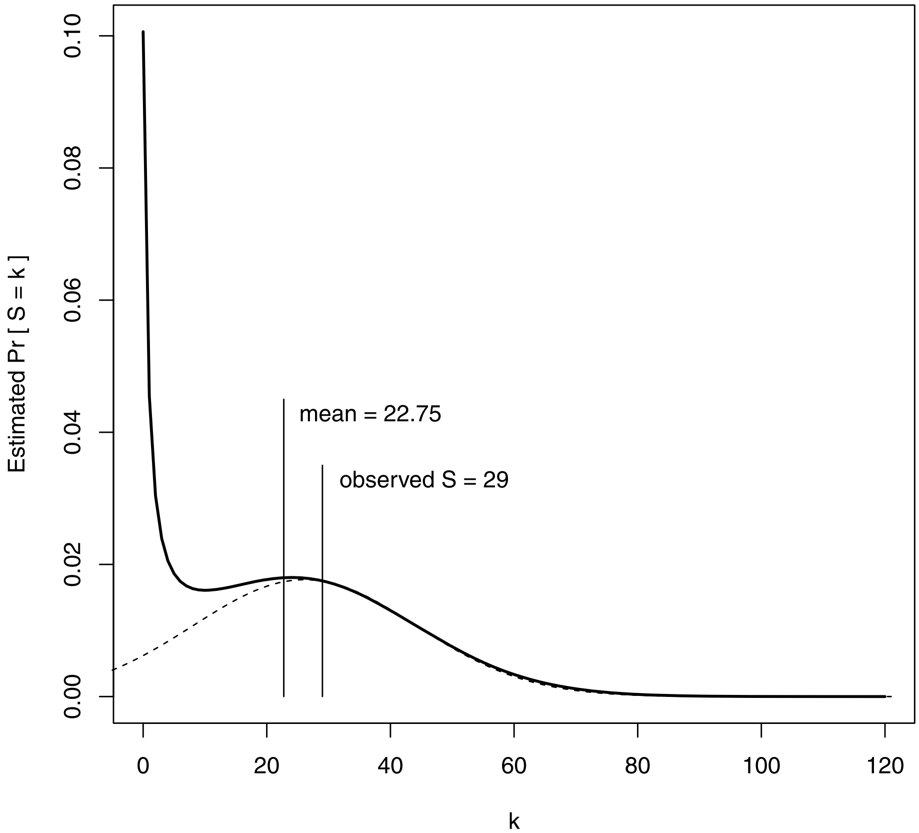

Fig. 2 Fitted distribution of S for breast cancer p-values with $I = 3$. The mean and observed $S = 29$ p-values at FWER $= .05$ are indicated. The dashed line is the normal approximation described by (6)

the analyses performed are summarized in the Cancer Genome Atlas (2012). The data values were downloaded from the website `https://tcga-data.nci.nih.gov/`. We choose to examine p-values representing summaries of statistical comparisons of smokers and non-smokers across the genetic markers. We identified $m = 20,068$ observed p-values after omitting about 2% missing values.

Using the Simes procedure, $S = 173$ p-values are identified with FWER $= .05$. The fitted parameter values $\hat{\boldsymbol{\theta}}$ are given in Table 2. Distributions up to $I = 4$ showed statistically significant improvement in the log-likelihood but larger values of I failed to change it. The fitted density function $\psi_4(\cdot \mid \hat{\boldsymbol{\theta}})$ given in Fig. 3 demonstrates good agreement with the observed data. The estimate $\hat{\theta}_0$ of π_0 is about .70 for $I = 4$.

The fitted distribution of S given in (3) is plotted in Fig. 4. There is close agreement between the observed value (173), the mean (176.35) of the fitted distribution, and the local mode (177). As with Fig. 2, the fitted distribution of S appears as a mixture of a distribution concentrated near zero and a normal distribution. The local mode at zero gives a fitted $\Pr[\, S \leq 2\,]$ of .012. The density mass away from zero is approximately that of a normal distribution with mean 178.8 and standard deviation 39.1 using (6).

6 Sampling dependent *P*-values

In this section we describe a method for sampling of dependent p-values by conditioning on an unobservable, latent variable. Greater dependence among the p-values results

Table 2 Maximum likelihood estimated parameter values of ψ_I for lung cancer example

Model parameters				Log-	2×	$\hat{\theta}_0$ to
I	Symbol	Estimate	Std Err	Likelihood	Difference	estimate π_0
1	θ_1	0.448	0.007	2147.48	—	0.552
2	θ_1	0.0	0.020	2579.47	863.98	0.684
	θ_2	0.158	0.006			
3	θ_1	0.174	0.034	2641.32	123.70	0.684
	θ_2	0.0008	0.020			
	θ_3	0.0233	0.0028			
4	θ_1	0.100	0.0497	2643.49	4.33	0.698
	θ_2	0.0761	0.0423			
	θ_3	0.000493	0.0119			
	θ_4	0.00195	0.0010			
5				Same as $I = 4$		

in greater means and variances for the distribution of p-values. This behavior is also described by Owen (2005). Greater dependence also contributes to a larger point mass at zero. We will use the fitted breast cancer example of Section 5.1 to illustrate these methods.

Let $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ denote I−tuples such that both $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and $\boldsymbol{\theta} - \boldsymbol{\epsilon}$ are valid parameters for the density ψ_I described in Section 4. Let Y denote a Bernoulli random variable with

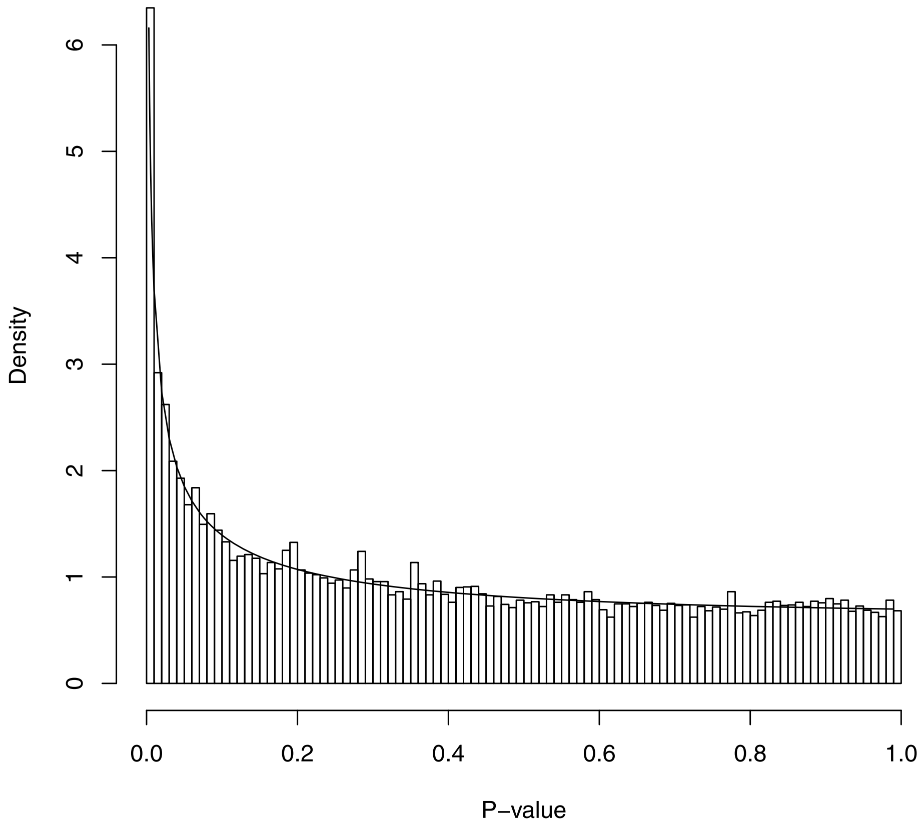

Fig. 3 Observed and fitted ψ_4 density for $m = 20,068$ TCGA lung cancer biomarkers

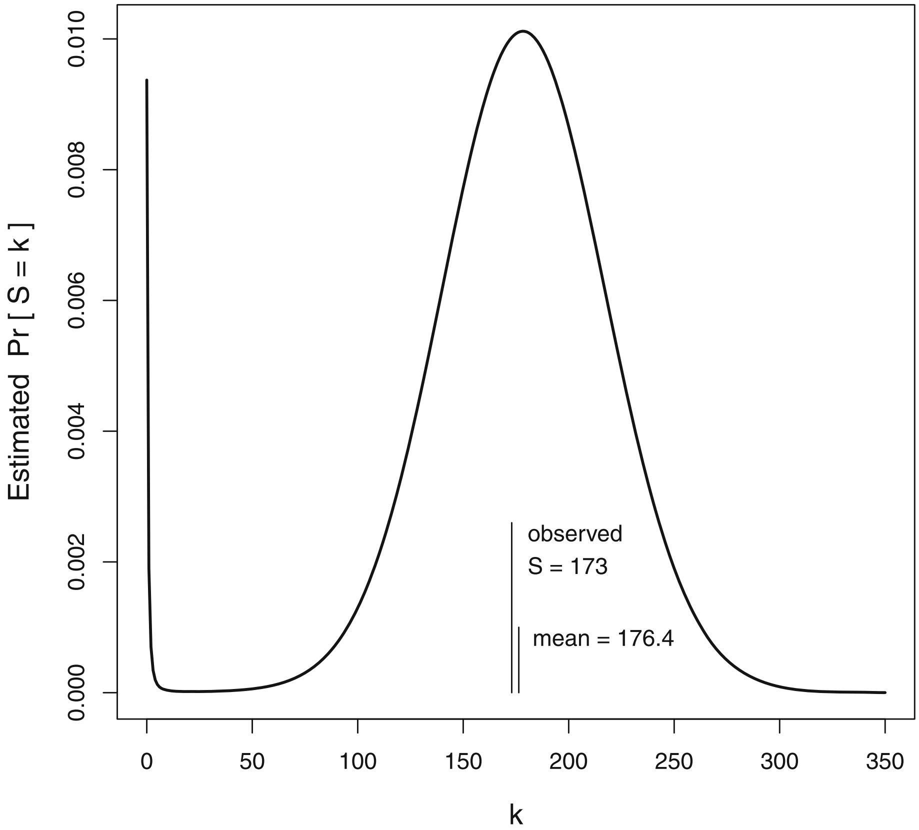

Fig. 4 Fitted distribution of S for TCGA lung cancer biomarker p-values with $\alpha \leq .05$

parameter equal to 1/2. Conditional on the (unobservable) value of Y, assume all p-values are sampled from either $\psi_I(\cdot \mid \boldsymbol{\theta} + \boldsymbol{\epsilon})$ or $\psi_I(\cdot \mid \boldsymbol{\theta} - \boldsymbol{\epsilon})$. The marginal distribution of these exchangeable p-values is then $\psi_I(\cdot \mid \boldsymbol{\theta})$ using (13).

To demonstrate the correlation among the p-values induced by this latent model, let Q_1, Q_2 denote a random sample from ψ_I, both with parameters either $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ or $\boldsymbol{\theta} - \boldsymbol{\epsilon}$, conditional on Y. The Q_i are conditionally independent given Y and have marginal covariance

$$\mathrm{Cov}(Q_1, Q_2) \;=\; \{\mathrm{E}(p \mid \boldsymbol{\theta} + \boldsymbol{\epsilon})\}^2/2 \;+\; \{\mathrm{E}(p \mid \boldsymbol{\theta} - \boldsymbol{\epsilon})\}^2/2 \;-\; \{\mathrm{E}(p \mid \boldsymbol{\theta})\}^2 \,,$$

where $\mathrm{E}(p \mid \boldsymbol{\theta})$ is the expected value of $\psi_I(p \mid \boldsymbol{\theta})$ calculated using (12). This covariance is never negative.

Continuing to sample in this fashion, we then have the marginal distribution

$$\Pr[\, S \;=\; k \,] \;=\; \Pr[\, S = k \mid \boldsymbol{\theta} - \boldsymbol{\epsilon} \,] \;/2 \;+\; \Pr[\, S = k \mid \boldsymbol{\theta} + \boldsymbol{\epsilon} \,] \;/2 \,. \tag{15}$$

As an illustration, we used $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\epsilon} = z\hat{\boldsymbol{\sigma}}$ where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\sigma}}$ are the fitted parameters and their estimated standard errors respectively given in Table 1 for the breast cancer example with $I = 3$. The distributions given by (15) for $z = 0$, .25, .5, and .75 are plotted in Fig. 5. Summaries of these four distributions and the mutual correlations of the p-values are given in Table 3. As we see in Fig. 2, all distributions in Fig. 5 appear as mixtures of distributions concentrated near zero and a truncated normal distribution, away from

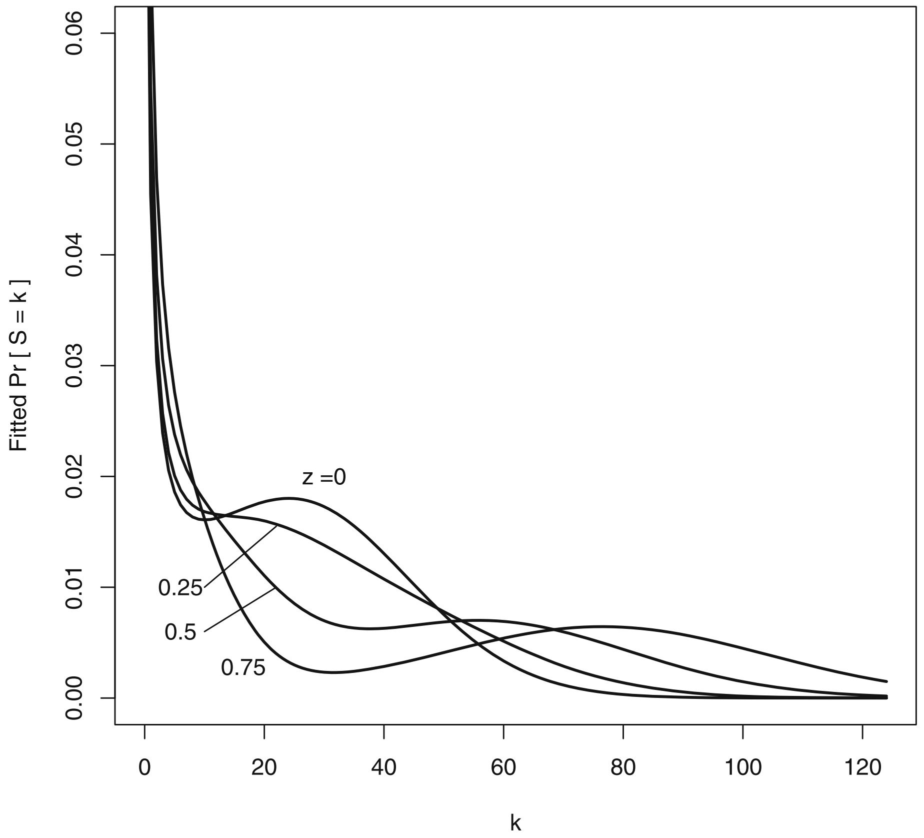

Fig. 5 Distributions of S under dependent sampling for fitted parameters from the breast cancer example. Values of z are given and control the dependence among the underlying *p*-values. Summaries of these distributions are given in Table 3

zero. Greater dependence results in a larger point mass at zero, as well as larger means and variances of S.

7 Power for planning studies

In this final section we describe how to plan for a larger project using data from a smaller pilot study. Huang et al. (2015) report on a study of $N = 78$ patients with lung cancer and examined $m = 48,803$ markers to determine if any of these are related to patient survival. None of these markers were identified as statistically significant at $\alpha = .05$ using the Bonferroni method. A link to their data appears in our References.

Table 3 Properties of the distributions of S when sampling correlated *p*-values using (15) with the fitted breast cancer data

	Correlation			
z	**of *p*-values**	**Mean**	**SD**	**Pr[S = 0]**
0	0	22.75	18.13	.101
.25	.004	24.43	21.44	.104
.5	.017	29.40	29.50	.116
.75	.037	37.18	39.85	.136

These distributions are plotted in Fig. 5

Table 4 Maximum likelihood estimated parameter values of ψ_I for survival of lung cancer patients

Model parameters				**Log-**	**2×**	**$\hat{\theta}_0$ to**
I	**Symbol**	**Estimate**	**Std Err**	**Likelihood**	**Difference**	**estimate π_0**
1	θ_1	0.1366	0.0048	461.09	—	0.863
2	θ_1	0.00863	0.0115	541.72	161.26	0.921
	θ_2	0.03507	0.0030			
3	θ_1	0.0524	0.020	545.44	7.45	0.908
	θ_2	0.00983	0.010			
	θ_3	0.00327	0.0013			
4				Same as $I = 3$		

We examined their data and the parameter estimates for our fitted models ψ_I appear in Table 4. We found the model with $I = 3$ provided the best fit and worked with that maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ to model power. We estimate more than 90% of the p-values were sampled from the null hypothesis in these data.

In order to describe power we will assume the magnitude of the effect, as measured by $\boldsymbol{\theta}$, is proportional to the square root of the subject sample size, as is often the case with parameters whose estimates are normally distributed. This assumption will also require values of $\boldsymbol{\theta}$ to lie near the center of the valid parameter space and wouldn't be valid for extrapolating to extremely large sample sizes. That is, we computed power estimates in Table 5 setting

$$\boldsymbol{\theta} = \boldsymbol{\theta}(N) = (N/78)^{1/2}\,\hat{\boldsymbol{\theta}}$$

where N is the proposed patient sample size and used $\boldsymbol{\epsilon} = z\boldsymbol{\theta}$ in (15) to vary the dependence among p-values for values of $z = 0, .4$, and .8.

A variety of sample sizes and correlations are summarized in Table 5. This table summarizes the power as the probability of identifying at least one marker with $\alpha = .05$. The expected number of identified findings using S is also given in this table.

We estimate the published study by Huang et al. (2015) had about a 50% chance of detecting at least one marker with $\alpha = .05$. Table 5 suggests increasing sample sizes from 78 to $N \geq 450$ patients to achieve power greater than 80% under a model of independent

Table 5 Estimated power based on pilot data from Huang et al. (2015) with $m = 48,803$ markers

Sample	**Dependence**		**Estimated**	
size *N*	***z***	**Correlation**	**Expected *S***	**Pr[S > 0]**
78	0	0	1.5	0.517
	0.4	.001	1.7	0.499
	0.8	.006	2.7	0.444
300	0	0	6.5	0.748
	0.4	.006	11.4	0.712
	0.8	.002	30.9	0.592
450	0	0	12.6	0.813
	0.4	.008	26.2	0.772
	0.8	.034	75.0	0.631
600	0	0	21.7	0.855
	0.4	.011	49.0	0.812
	0.8	.045	90.8	0.657

sampling. Even small mutual correlations result in greater point masses at zero, reducing the power of detecting at least one statistically significant p-values. Another factor is the estimated high proportion of p-values sampled from the null hypothesis ($\hat{\pi}_0 = .908$). Subsequent studies should restrict sampling to those markers showing promise in the pilot, as the case in Haynes et al. (2012).

Appendix A: Details of Sections 2 and 3

We define $U_0 = 1$ in Eq. (4) and

$$U_k = \int_{p_1=0}^{\alpha/m} \int_{p_2=p_1}^{2\alpha/m} \cdots \int_{p_k=p_{k-1}}^{k\alpha/m} \psi(p_1) \cdots \psi(p_k)\, \mathrm{d}p_k \ldots \mathrm{d}p_2\, \mathrm{d}p_1 \;,$$

for $k = 1, 2, \ldots, m$.

To demonstrate (5), we integrate one term at a time to show

$$\begin{aligned} U_k &= \int_{p_1=0}^{\alpha/m} \int_{p_2=p_1}^{2\alpha/m} \cdots \int_{p_{k-1}=p_{k-2}}^{(k-1)\alpha/m} \{\Psi(k\alpha/m) - \Psi(p_{k-1})\}\, \psi(p_1) \cdots \psi(p_{k-1})\, \mathrm{d}p_{k-1} \cdots \mathrm{d}p_1 \\ &= \Psi(k\alpha/n)\, U_{k-1} - \int_{p_1=0}^{\alpha/m} \int_{p_2=p_1}^{2\alpha/m} \cdots \int_{p_{k-2}=p_{k-3}}^{(k-2)\alpha/m} \{\Psi^2((k-1)\alpha/m) - \Psi^2(p_{k-2})\}/2! \\ &\quad \times\; \psi(p_1) \cdots \psi(p_{k-2})\, \mathrm{d}p_{k-2} \ldots \mathrm{d}p_2\, \mathrm{d}p_1 \\ &= \Psi(k\alpha/n)\, U_{k-1} - \Psi^2\{(k-1)\alpha/m\}\, U_{k-2}/2! \\ &\quad + \frac{1}{2!} \int_{p_1=0}^{\alpha/m} \int_{p_2=p_1}^{2\alpha/m} \cdots \int_{p_{k-2}=p_{k-3}}^{(k-2)\alpha/m} \Psi^2(p_{k-2})\, \psi(p_1) \cdots \psi(p_{k-2})\, \mathrm{d}p_{k-2} \ldots \mathrm{d}p_2\, \mathrm{d}p_1 \;, \end{aligned}$$

and continue in this manner to demonstrate the recursive relation

$$U_k = \sum_{i=1}^{k} (-1)^{i+1}\, \Psi^i\{(k-i+1)\alpha/m\}\, U_{k-i}/\, i! \;, \tag{16}$$

given by (5).

To demonstrate (7) for the specific case of $\Psi(p) = p$ we need to show

$$U_k = (k+1)^{k-1}\, (\alpha/m)^k\, /\, k! \;. \tag{17}$$

We will prove (17) by induction on k.

In Section 3 we demonstrate (17) is true for $k = 0, 1, 2$. Next, we demonstrate if (17) is valid for any $k = 0, 1, \ldots, m-1$ then it is also true for $k+1$.

Begin by using the recursive relation (16) with $\Psi(p) = p$ and (17) for k giving

$$\begin{aligned} U_{k+1} &= \sum_{i=1}^{k+1} (-1)^{i+1} \left\{ \frac{(k-i+2)\alpha}{m} \right\}^i \left\{ \frac{(k-i+2)^{k-i} \alpha^{k-i+1}}{(k-i+1)!\; i!\; m^{k-i+1}} \right\} \\ &= (\alpha/m)^{k+1} \sum_{i=1}^{k+1} (-1)^{i+1} \frac{(k-i+2)^k}{(k-i+1)!\; i!} \;. \end{aligned}$$

It remains to show

$$\sum_{i=1}^{k+1} (-1)^{i+1} (k-i+2)^k/(k-i+1)!\; i! \;=\; (k+2)^k/(k+1)! \;,$$

or equivalently

$$\sum_{i=0}^{k+1} (-1)^{i+1} \binom{k+1}{i} (k-i+2)^k = 0 \,.$$

Continue by writing $\binom{k+1}{i} = \binom{k}{i} + \binom{k}{i-1}$ and set $j = i - 1$ giving

$$\sum_{i=0}^{k+1} (-1)^{i+1} \binom{k+1}{i} (k-i+2)^k = \sum_{i=0}^{k} (-1)^{i+1} \binom{k}{i} (k-i+2)^k + \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j+1)^k .$$

The proof of (17) is completed by two applications of the Ruiz Identity (Ruiz, 1996). Specifically,

$$\sum_{i=0}^{k} (-1)^i \binom{k}{i} (x-i)^k = k! \ ,$$

for all integers $k \geq 0$ and all real numbers x.

Appendix B: A close alternative hypothesis

Here we demonstrate the distribution of B and S when a large number of p-values are independently sampled from $\Psi_I(p \mid \boldsymbol{\beta})$ for $I \geq 1$ for values of $\boldsymbol{\beta}$ close to zero. That is, the null and alternative hypotheses are not very different. Specifically, consider a sequence of parameter values $\boldsymbol{\beta}_m = \boldsymbol{\beta}/(\log m)^I$ shrinking to zero. Following (11), we always have $\beta_0 = 1$.

Begin by writing

$$\begin{aligned} m\Psi_I(\gamma/m \mid \boldsymbol{\beta}_m) &= \gamma \left\{ 1 + \frac{\beta_1}{(\log m)^I} (\log m - \log \gamma) + \cdots + \frac{\beta_I}{(\log m)^I} (\log m - \log \gamma)^I \right\} \\ &= \gamma(\beta_I + 1) + O(1/\log m) \ , \end{aligned} \qquad (18)$$

for any fixed $\gamma > 0$.

When sampling from $\Psi_I(\cdot \mid \boldsymbol{\beta}_m)$ using the Bonferroni rule (1), set $\gamma = \alpha$ in (18) to demonstrate the number of statistically significant p-values B will have an approximate Poisson distribution with mean $\alpha(\beta_I + 1)$.

In order to describe the distribution of S we can also use (18) to show

$$\{1 - \Psi_I((k+1)\alpha/m \mid \boldsymbol{\beta}_m)\}^{m-k} = \exp\{-(k+1)\alpha(\beta_I + 1)\} + O(1/\log m) \ ,$$

demonstrating

$$\Pr[\ S = 0 \mid \boldsymbol{\beta}_m\ ] = \exp\{-\alpha(\beta_I + 1)\} + O(1/\log m) \ ,$$

and

$$\Pr[\ S = 1 \mid \boldsymbol{\beta}_m\ ] = \alpha(\beta_I + 1)\ \exp\{-2\alpha(\beta_I + 1)\} + O(1/\log m) \ .$$

More generally, if m p-values are independently sampled from $\Psi_I(\cdot \mid \boldsymbol{\beta}/(\log m)^I)$ then

$$\Pr[\ S = k\ ] = (k+1)^{k-1}/k!\ \ \{\alpha(\beta_I+1)\}^k\ \exp\{-(k+1)\alpha(\beta_I+1)\} + O(1/\log m) \ , \qquad (19)$$

for moderate values of $k = 0, 1, \ldots$ which is the Borel distribution (8) with parameter $\alpha(\beta_I + 1)$. The proof of (19) closely follows the proof by induction of (17) in Appendix A.

Appendix C: Parameter space for $\psi_I(p)$

In this Appendix we describe the limits of parameter values for the density function $\psi_I(p \mid \boldsymbol{\theta})$ defined in (9) for small values of I. Specifically, we must have $\psi_I(p)$ non-negative and monotone decreasing for all $0 < p < 1$.

For all values of I we must have $\theta_I > 0$ in order for $\psi_I(p) > 0$ for values of p close to zero. We must have $\psi_I(1) = \theta_0$ non-negative so $\theta_0 \geq 0$.

Since $\psi_I'(1) = -\theta_1$, in order for ψ_I to be monotone decreasing, we must have $\theta_1 \geq 0$ for all values of I. The condition of all $\theta_i \geq 0$ is sufficient (but may not be neccessary) for ψ to be monotone decreasing because the Descartes Rule of Signs shows the derivative $\psi_I'(p)$ of $\psi_I(p)$ will have no positive roots in p.

$\boldsymbol{I = 1}$**:** If $0 \leq \theta_1 \leq 1$ then $\psi_1(p \mid \theta_1)$ is a valid density and monotone decreasing.

$\boldsymbol{I = 2}$**:** We must have $(\theta_0, \theta_1, \theta_2)$ all non-negative so

$$0 < \theta_2 \leq 1/2 \text{ and } 0 \leq \theta_1 \leq 1 - 2\theta_2 \ .$$

For larger values of I, define $x = -\log p$ and set $g(x) = \sum \theta_i x^i$. It is sufficient for $g(x) \geq 0$ and $g'(x) \geq 0$ for all $x \geq 0$ to show ψ is positive and monotone decreasing. For $\theta_1 \geq 0$ we have $g'(0) \geq 0$ and $g'(x) \geq 0$ for all x sufficiently large because $\theta_I > 0$. To demonstrate $g' > 0$ we need to show $g''(x)$ has no real, positive roots.

$\boldsymbol{I = 3}$**:** We must have $\theta_3 > 0$ and $\theta_1 \geq 0$. The slope of $g(x)$ does not change sign provided its second derivative $g'' = 6\theta_3 x + 2\theta_2$ is never negative for all $x \geq 0$. This shows $\theta_2 > 0$. The restriction $0 \leq \theta_0 \leq 1$ gives

$$0 < \theta_3 \leq 1/6; \qquad 0 \leq \theta_2 \leq 1/2 - 3\theta_3; \quad \text{and} \quad 0 \leq \theta_1 \leq 1 - 2\theta_2 - 6\theta_3 \ .$$

$\boldsymbol{I = 4}$**:** We have $\theta_1 \geq 0$ and $\theta_4 > 0$. If the larger, real root of $g'' = 12\theta_4 x^2 + 6\theta_3 x + 2\theta_2$ is negative then

$$(36\theta_3^2 - 96\theta_2\theta_4)^{1/2} < 6\theta_3$$

showing $\theta_3 > 0$. Squaring both sides of this inequality shows $\theta_2 > 0$.

If g'' has imaginary roots then $36\theta_3^2 - 96\theta_2\theta_4 < 0$ so $\theta_2 > 0$ and g'' is never negative. With imaginary roots, if the minimum of $g''(x)$ occurs at $x > 0$ then $\psi_4(p)$ will be decreasing but not concave. The minimum of $g''(x)$ occurs at $x = -\theta_3/4\theta_4$ which is negative leading to $\theta_3 > 0$.

In either real or imaginary roots, for $I = 4$ we have

$$\begin{aligned} 0 < \theta_4 \leq 1/24; \quad 0 \leq \theta_3 \leq 1/6 - 4\theta_4; \\ 0 \leq \ \theta_2 \ \leq 1/2 - 3\theta_3 - 12\theta_4; \\ \text{and} \quad 0 \leq \ \theta_1 \ \leq 1 - 2\theta_2 - 6\theta_3 - 24\theta_4 \ . \end{aligned}$$

Abbreviations

BRCA: Breast cancer gene; FDR: False discovery rate; FWER: Family-wise error rate; TCGA: The National Institutes of Health Cancer Genome Atlas Program

Acknowledgements

The authors thank J. Jin and T. Cai for providing the data analyzed in Section 5.1 and Beth Nichols for a careful reading of the manuscript.

Authors' contributions

The authors shared equally in all aspects of the creation of the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was supported in part by Vanderbilt CTSA grant 1ULTR002243 from NIH/NCATS, R01 CA149633 from NIH/NCI, R01 FD004778 from FDA, R21 HL129020, P01 HL108800 from NIH/NHLBI (CY) and grants P50-CA196530, P50-CA121974, P30-CA16359, R01-CA177719, R01-ES005775, R01-CA223481, R41-A120546, U48-DP005023, U01-CA235747, R35CA197574, and R01-CA168733 awarded by the NIH (DZ).

Availability of data and materials

The data from Section 5.1 is available from the authors with permission from J. Jin and T. Cai. The data for Section 5.2 is available at `https://tcga-data.nci.nih.gov/`. The data from Section 7 is available at `www.biomedcentral.com/content/supplementary/s12859-015-0463-x-s1.xls`.

Competing interests

The authors declare they have no competing interests.

Author details

[1]Department of Biostatistics, Vanderbilt University Medical Center, 37232 Nashville, TN, United States. [2]Department of Biostatistics, Yale University, 06520 New Haven, CT, United States.

Received: 16 April 2020 Accepted: 13 September 2020
Published online: 19 October 2020

References

Benjamini, Y.: Discovering the false discovery rate. J. R. Stat. Soc. B. **72**, 405–16 (2010). https://doi.org/10.1111/j.1467-9868.2010.00746.x

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B. **57**, 289–300 (1995). http://www.jstor.org/stable/2346101

Benjamini, Y., Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. **25.1**, 60–83 (2000). https://doi.org/10.3102/10769986025001060

Broberg, P.: A comparative review of estimates of the proportion unchanged genes and the false discovery rate. BMC Bioinformatics. **6**, 199–218 (2005). https://doi.org/10.1186/1471-2105-6-199

Cancer Genome Atlas Research Network: Comprehensive genomic characterization of squamous cell lung cancers. Nature. **489**, 519–25 (2012). https://doi.org/10.1038/nature11404

Donoho, D., Jin, J.: Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat. **32**, 962–94 (2004). https://doi.org/10.1214/009053604000000265

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V.: Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. **96**, 1151–60 (2001). https://doi.org/10.1198/016214501753382129

Efron, B.: Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. J. Am. Stat. Assoc. **99**, 96–104 (2004). https://doi.org/10.1198/016214504000000089

Friguet, C., Kloareg, M., Causeur, D.: A factor model approach to multiple testing under dependence. J. Am. Stat. Assoc. **104**, 1406–15 (2009). https://doi.org/10.1198/jasa.2009.tm08332

Genovese, C., Wasserman, L.: A stochastic process approach to false discovery control. Ann. Stat. **32**, 1035–61 (2004). https://doi.org/10.1214/009053604000000283

Haynes, B. F., Gilbert, P. B., McElrath, M. J., Zolla-Pazner, S., Tomaras, G. D., Alam, S. M., et al.: Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N. Engl. J. Med. **366**, 1275–1286 (2012). https://doi.org/10.1056/NEJMoa1113425

Hedenfalk, I., Duggan, D., Chen, Y., et al.: Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. **344**, 539–48 (2001). https://doi.org/10.1056/NEJM200102223440801

Huang, H.-L., Wu, Y.-C., Su, L.-J., *et al*: Discovery of prognostic biomarkers for predicting lung cancer metastasis using microarray and survival data. BMC Bioinformatics. **16**, 54 (2015). https://doi.org/10.1186/s12859-015-0463-x. Their data is available at `www.biomedcentral.com/content/supplementary/s12859-015-0463-x-s1.xls`

Jin, J., Cai, T. T.: Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. J. Am. Stat. Assoc. **102**, 495–506 (2007). https://doi.org/10.1198/016214507000000167

Jolley, L. B. W.: Summation of Series. Second edition. Dover, New York (1961). ASIN: B01K3IQJ08

Kozoil, J. A., Tuckwell, H. C.: A Bayesian method for combining statistical tests. J. Stat. Plan. Infer. **78**, 317–23 (1999). https://doi.org/10.1016/S0378-3758(98)00222-5

Langaas, M., Lindqvist, B. H., Ferkingstad, E.: Estimating the proportion of true null hypotheses, with application to DNA microarray data. J. R. Stat. Soc. B. **67**, 555–72 (2005). https://doi.org/10.1111/j.1467-9868.2005.00515.x

Maechler, M.: Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable (2019). R package version 0.7-2. `https://CRAN.R-project.org/package=Rmpfr`

Owen, A. B.: Variance of the number of false discoveries. J. R. Stat. Soc. Ser. B. **67**, 411–26 (2005). https://doi.org/10.1111/j.1467-9868.2005.00509.x

Pounds, S., Morris, S. W.: Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. Bioinformatics. **19**, 1236–42 (2003). https://doi.org/10.1093/bioinformatics/btg148

Ruiz, S. M.: An algebraic identity leading to Wilson's Theorem. Math. Gaz. **80.489**, 579–82 (1996). https://doi.org/10.2307/3618534

Simes, R. J.: An improved Bonferroni procedure for multiple tests of significance. Biometrika. **73**(3), 751–754 (1986). https://doi.org/10.1093/biomet/73.3.751

Storey, J. D., Tibshirani, R.: Statistical significance for genomewide studies. Proc Natl Acad Sci USA. **100**, 9440–5 (2003). https://doi.org/10.1073/pnas.1530509100

Sun, W., Cai, T. T.: Large-scale multiple testing under dependence. J. R. Stat. Soc. Ser. B. **71**, 393–424 (2009). https://doi.org/10.1111/j.1467-9868.2008.00694.x

Tang, Y., Ghosai, S., Roy, A.: Nonparametric Bayesian estimation of positive false discovery rates. Biometrics. **63**, 1126–34 (2007). https://doi.org/10.1111/j.1541-0420.2007.00819.x

Tanner, J. C.: A derivation of the Borel distribution. Biometrika. **48**, 222–4 (1961). https://doi.org/10.1093/biomet/48.1-2.222

Wu, W.: On false discovery control under dependence. Ann. Stat. **36**, 364–80 (2008). https://doi.org/10.1214/009053607000000730

Yu, C., Zelterman, D.: A parametric model to estimate the proportion from true null using a distribution for *p*-values. Comput Stat Data Anal. **114**, 105–18 (2017). https://doi.org/10.1016/j.csda.2017.04.008

Yu, C., Zelterman, D.: A parametric meta-analysis. Stat. Med. **38**, 4013–25 (2019). https://doi.org/10.1002/sim.8278

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.