A novel characterization of the generalized family wise error rate using empirical null distributions

Jeffrey C. Miecznikowski*

Department of Biostatistics
SUNY University at Buffalo
433 Kimball Tower
3435 Main St.
Buffalo, NY, 14214
716.881.8953
jcm38@buffalo.edu

Daniel P. Gaile

Department of Biostatistics
SUNY University at Buffalo
706 Kimball Tower
3435 Main St.
Buffalo, NY, 14214
716.829.2756
dpgaile@buffalo.edu

Abstract

In this manuscript we present a novel characterization of the generalized family wise error rate: kFWER. The interpretation allows researchers to view kFWER as a function of the test statistics rather than current methods based on p-values. Using this interpretation we present several theorems and methods (parametric and non parametric) for estimating kFWER in various data settings. With this version of kFWER, researchers will have an estimate of kFWER in addition to knowing what tests are significant at the estimated kFWER. Additionally, we present methods that use empirical null distributions in place of parametric distributions in standard p-value kFWER controlling schemes. These advancements represent an improvement over common kFWER methods which are based on parametric assumptions and merely report the tests that are significant under a given value for kFWER.

keywords: empirical null distribution; false discovery rate; family wise error; multiple testing

1. Introduction

With the development of genomic microarrays and related high throughput assays such as mass spectrometry there has been a concomitant development of the multiple testing algorithms designed to work with this type of data. The goal in these algorithms is to determine a subset of genes/proteins/micro ribonucleic acids (RNAs) generally called "cases" that are related to an outcome. On a univariate level this relationship is commonly assessed via a hypothesis test and corresponding test statistic and p-value. A group error rate, such as the false discovery rate (FDR) or the generalized family wise error rate (kFWER), is then used to determine the significant subset of interest. Commonly, the hypothesis test for each case has a null hypothesis of "not significantly related to outcome" and the alternative hypothesis of "significantly related

*Corresponding author

to outcome." Likewise, we consider each case as either a "null case" or an "alternative case" depending on its true (unknown) status. In this general framework, we explore a novel method to study multiple testing errors.

We will assume the two group model for multiple testing as in Efron (2010). We will also assume that there are N cases or tests. In a gene expression microarray setting N may be the number of genes. Each case may be either null or non null with prior probability π_0 or $\pi_1 = 1 - \pi_0$ with test statistics (z values) having density either $f_0(z)$ or $f_1(z)$. In other words, we have the following,

$$\begin{array}{ll} \pi_0 = Pr\{null\}, & f_0(z) = \text{density if null}, \\ \pi_1 = Pr\{non-null\}, & f_1(z) = \text{density if non-null}. \end{array} \tag{1}$$

We will also let F_0 and F_1 denote the probability distributions corresponding to f_0 and f_1 so that for any subset $\mathcal{Z}$ of the real line,

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz \text{ and } F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz. \tag{2}$$

Thus the test statistics z_i follow the mixture density

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z), \tag{3}$$

with the mixture distribution

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z}). \tag{4}$$

With the N cases, we further define the following functions:

$$\begin{array}{ll} N_0(\mathcal{Z}) = & \text{the number of null } z_i \text{ in set } \mathcal{Z}, \\ N_1(\mathcal{Z}) = & \text{the number of alternative } z_i \text{ in set } \mathcal{Z}, \\ N_+(\mathcal{Z}) = & \text{the number of } z_i \text{ in set } \mathcal{Z}. \end{array} \tag{5}$$

With this notation, we have $N_+(\mathcal{Z}) = N_0(\mathcal{Z}) + N_1(\mathcal{Z})$. Further, N_0 and N_1 represent the total of true nulls and true alternative, respectively among N cases. Using similar notation, and recognizing that researchers commonly report $z \in \mathcal{Z}$ as non-null, Efron (2010) defines the false discovery rate quantities,

$$fdr(z_0) = \quad Pr(null|z = z_0) \quad = \pi_0 f_0(z_0)/f(z_0), \tag{6}$$

$$Fdr(\mathcal{Z}) = \quad \pi_0 F_0(\mathcal{Z})/F(\mathcal{Z}), \tag{7}$$

$$\overline{Fdr}(\mathcal{Z}) = \quad \pi_0 F_0(\mathcal{Z})/\overline{F}(\mathcal{Z}), \tag{8}$$

where $\overline{F}(\mathcal{Z})$ is the empirical distribution function of the z_i. The false discovery rate quantity in (6) is considered the local false discovery rate and (6)-(8) are all functions of the unknowns π_0, f_0, F_0, f, F and, as such, need to be estimated. Methods to estimate these quantities, and thus the quantities in (6), are presented in Efron (2010).

In the following subsections we define a novel (functional) version of the generalized family wise error rate with several estimation methods inspired by the FDR methods in Efron (2010). We relate our kFWER method to existing kFWER methods and explore conditions under which it should be preferred. We also develop empirical versions of the traditional parametric p-value based methods to control kFWER.

2. Generalized Family Wise Error Rate

The kFWER error rate is a generalized version of the family wise error rate (FWER). Control of FWER refers to controlling the probability of committing one or more false discoveries. If we let V denote the number of false positives from N tests, then notationally, (according to Lehmann and Romano (2005)) α control of FWER can be expressed as, $Pr(V \geq 1) \leq \alpha$ or equivalently, $Pr(V = 0) \geq 1 - \alpha$. Note that α is

usually chosen to be small, e.g. 0.05. In kFWER the equation becomes,

$$Pr(V \geq k) \leq \alpha, \tag{9}$$

where k and α are usually determined prior to the analysis. Methods designed to control kFWER with derivations for the exact distribution of V are provided in Miecznikowski et al. (2011).

Thinking about $Pr(V \geq k)$ as a functional quantity we propose the following novel definition for a generalized family wise error rate,

$$kFWER(\mathcal{Z}) = Pr(N_0(\mathcal{Z}) \geq k), \tag{10}$$

where $N_0(\mathcal{Z})$ is defined in (5). We commonly consider the sets $\mathcal{Z} = (-\infty, z)$ and $\mathcal{Z} = (z, \infty)$ where tests with z scores in $\mathcal{Z}$ are reported as non-null. For simplicity of notation, for a single point z we define the left hand side (LHS) $kFWER(z) = kFWER(\mathcal{Z})$ where $\mathcal{Z} = (-\infty, z)$ and the right hand side (RHS) $kFWER(z) = kFWER(\mathcal{Z})$ where $\mathcal{Z} = (z, \infty)$. Commonly one sided hypotheses tests examine either the LHS or RHS kFWER, while two sided tests may consider both LHS and RHS kFWER significance.

We expand (10) in light of the two group model (1). By conditioning on the total number of true nulls N_0 we obtain,

$$\begin{aligned} kFWER(\mathcal{Z}) &= Pr(N_0(\mathcal{Z}) \geq k) && (11)\\ &= \sum_{\nu=k}^{N} Pr(N_0(\mathcal{Z}) = \nu) && (12)\\ &= \sum_{\nu=k}^{N} \sum_{\eta=\nu}^{N} Pr(N_0(\mathcal{Z}) = \nu | N_0 = \eta) \cdot Pr(N_0 = \eta) && (13)\\ &= \sum_{\eta=k}^{N} (1 - F_b(k-1|\eta, F_0(\mathcal{Z}))\, f_b(\eta|N, \pi_0), && (14) \end{aligned}$$

where $F_B(a|b,c)$ and $f_B(a|b,c)$ denote the CDF and PDF, respectively, for a binomial distribution evaluated at a with size parameter b and probability parameter c. Note that our transition from (13) to (14) requires an

Independence Assumption: Each z_i follows model (1) independently. (15)

In our functional specification of kFWER in (14) we must estimate π_0 and $F_0(\mathcal{Z})$; quantities related to the null distribution. In Efron (2010), we are presented with several options to estimate these quantities.

2.0.1. Estimating the functional $kFWER$

We consider a parametric (theoretical) and nonparametric estimator for F_0. We will assume that for case i we have the following null hypothesis,

$$H_{0i}: \text{ case } i \text{ is "null".} \tag{16}$$

Commonly, we use a t statistic t_i to examine this null hypothesis for case i. As in Efron (2010), we will transform our t_i to $z_i = \Phi^{-1}(G(t_i))$ where Φ and G are the cumulative distribution functions for the standard normal and (appropriate) t distribution. Then, under the assumption of normal sampling, z_i will have a standard normal distribution if H_{0i} is true,

$$H_{0i} : z_i \sim \mathcal{N}(0,1). \tag{17}$$

We call (17) the parametric null. Thus, the parametric or theoretical estimator $\overline{F_0}$ is given as $\overline{F_0}(\mathcal{Z}) = \int_{\mathcal{Z}} \phi(z)dz$ where $\phi(z)$ is the PDF for a standard normal random variable. The non parametric estimator $\widehat{F_0}$

is obtained using the maximum likelihood estimator (MLE) method described in Efron (2010), see Appendix. In short, the MLE method uses a data driven procedure to estimate the mean and standard deviation for the null (normal) distribution. Note, the central matching method presented in Efron (2010) can also be used to estimate F_0, see Appendix. Reasons why the theoretical null may fail are provided in the Discussion section.

Similarly, we estimate π_0 using a parametric and non parametric estimator. In our two group model, if we believe that $f_1(z)$ is near zero for a subset of $\mathcal{A}_0$ of the sample space, perhaps the points near zero, then the expected value of $N_+(\mathcal{A}_0)$, the observed number of z_i values in $\mathcal{A}_0$ is given as $E\left[N_+(\mathcal{A}_0)\right] = \pi_0 N \cdot F_0(\mathcal{A}_0)$. This suggests the parametric estimator

$$\overline{\pi_0} = N_+(\mathcal{A}_0) / \left(N \cdot \overline{F_0}(\mathcal{A}_0)\right), \tag{18}$$

and the non parametric estimator

$$\widehat{\pi_0} = N_+(\mathcal{A}_0) / \left(N \cdot \widehat{F_0}(\mathcal{A}_0)\right), \tag{19}$$

where

$$\mathcal{A}_0 = \text{Subset of sample space where } f_1(z) \text{ near 0, e.g } (-2, 2). \tag{20}$$

With these estimators for F_0 and π_0 and assuming (15) we can define a **parametric** and **non parametric** estimator of kFWER in (10). The **parametric** estimator is given by,

$$\overline{kFWER}(\mathcal{Z}) = \sum_{\eta=k}^{N} \left(1 - F_b(k-1|\eta, \overline{F_0}(\mathcal{Z}))\right) f_b(\eta|N, \overline{\pi_0}). \tag{21}$$

The **non parametric** estimator is given by

$$\widehat{kFWER}(\mathcal{Z}) = \sum_{\eta=k}^{N} \left(1 - F_b(k-1|\eta, \widehat{F_0}(\mathcal{Z}))\right) f_b(\eta|N, \widehat{\pi_0}). \tag{22}$$

3. Other *k*FWER methods

In this section, we present the methods commonly used to control the traditional kFWER in (9). These kFWER schemes are more fully presented in Lehmann and Romano (2005); Guo and Romano (2007); Miecznikowski et al. (2011); Roquain and Villers (2011) including theorems and proofs that each method controls kFWER in a two group mixture model setting. In the following sections we claim (with proofs provided in the Appendix) that each method fits into our larger functional definition for kFWER in (14). We also present simple methods to develop empirical versions of these traditional parametric methods.

3.1. Adjusted Bonferroni Method

The adjusted Bonferroni adjustment to control kFWER at α specifies $k\alpha/N$ as the p-value cut point for significance where tests with p-values less than the cut point are considered significant. We can define our p-values using parametric or empirical methods. For simplicity we assume one-sided tests. Our parametric estimator uses LHS p-values defined by

$$\overline{p}_i = \overline{F_0}^{-1}(z_i), \tag{23}$$

and RHS p-values defined by

$$\overline{p}_i = 1 - \overline{F_0}^{-1}(z_i). \tag{24}$$

The empirical estimator uses LHS p-values defined

$$\widehat{p}_i = \widehat{F_0}^{-1}(z_i), \tag{25}$$

and RHS p-values defined by

$$\widehat{p}_i = 1 - \widehat{F_0}^{-1}(z_i). \tag{26}$$

The LHS (RHS) adjusted parametric Bonferroni procedure rejects all null hypotheses with LHS (RHS) p-values less than or equal to $k\alpha/N$ where the p-values are estimated by $\overline{p}_i$ as given in (23)-(24). The adjusted empirical Bonferroni procedure estimates the p-values via $\widehat{p}_i$ as given in (25)-(26).

Further, if we let $z_{bon} = F_0^{-1}(k\alpha/N)$, where F_0 is given in (2) then $kFWER\,(z_{bon}) \leq \alpha$ when using the (LHS) $kFWER$ definition where $kFWER(z) = kFWER(\mathcal{Z})$ where $\mathcal{Z} = (-\infty, z)$ (see Appendix for proof). A similar version holds for the (RHS) kFWER.

3.2. Adjusted Šidàk Method

Under independence of the N cases, the generalized Šidàk procedure to control the generalized family wise error rate works by rejecting all hypotheses with a p-value less than p_{sid} where p_{sid} is such that $F_b(k-1|N, p_{sid}) = 1-\alpha$. The proof that this procedure controls kFWER can be found in Miecznikowski et al. (2011); Roquain and Villers (2011).

Let $z_{sid} = F_0^{-1}(p_{sid})$ then under the two group model with independence of the N test cases, we have $(LHS)\,kFWER\,(z_{sid}) \leq \alpha$ when using the (LHS) kFWER definition where $kFWER(z) = kFWER(\mathcal{Z})$ where $\mathcal{Z} = (-\infty, z)$ (see Appendix for proof). A similar version holds for the (RHS) kFWER.

As in the adjusted Bonferroni setting, we can define a parametric and empirical version of the adjusted Šidàk method based on the p-values used in the analysis.

The LHS (RHS) adjusted parametric Šidàk method rejects all null hypothesis with LHS (RHS) p-values less than or equal to p_{sid} where the p-values are estimated with $\overline{p}$. The adjusted empirical Šidàk method uses $\widehat{p}$ to estimate the p-values.

3.3. Adjusted Holm Method

A method to control kFWER using the Holm procedure is given in Lehmann and Romano (2005). This method is an adjustment to the Holm method designed to control the FWER (Holm, 1979). The following procedure describes the Holm method to control FWER at level α for N tests. Let

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_N \tag{27}$$

be constants defined by $\alpha_i = \alpha/(N-i+1)$ and let the ordered p-values be denoted by $p_{(1)} \leq \cdots \leq p_{(N)}$ corresponding to hypotheses, $H_{(1)}, \cdots, H_{(N)}$. If $p_{(1)} > \alpha_1$,then reject no null hypothesis. Otherwise, if

$$p_{(1)} \leq \alpha_1, \ldots, p_{(r)} \leq \alpha_r, \tag{28}$$

then reject hypothesis $H_{(1)}, \ldots, H_{(r)}$ where the largest r satisfying (28) is used. With this framework to control FWER at level α, the Holm method to control kFWER at level α is done by redefining α_i as

$$\alpha_i = \begin{cases} \frac{k\alpha}{N}, & i \leq k, \\ \frac{k\alpha}{N+k-i}, & i > k. \end{cases} \tag{29}$$

We show in the Appendix that the largest r satisfying (28) is such that (LHS) $kFWER\,(z_{holm}) \leq \alpha$ when $z_{holm} = F_0^{-1}\left(p_{(r)}\right)$. A similar version also holds for (RHS) kFWER (see Appendix for proof). The adjusted parametric Holm method uses $\overline{p}$ to estimate the p-values while the adjusted empirical Holm method uses $\widehat{p}$ to estimate the p-values.

4. Software Implementation

The kFWER estimators given in (21),(22) are implemented in R (R Core Team, 2012) and can be found with the authors' technical report (Miecznikowski and Gaile, 2012). Note the code requires the R *locfdr* package (Efron et al., 2011).

5. Simulation and Examples

In this section we provide three simulations and two examples implementing and comparing $\overline{kFWER}$, and $\widehat{kFWER}$ procedures along with empirical and parametric versions of the adjusted Bonferroni, Šidàk, and Holm methods. The simulations are designed to study kFWER in an independence setting, a setting with an overdispersed null distribution, and a setting with correlated test statistics.

5.1. BUM simulation

For our first simulation, we consider control of kFWER when using a beta-uniform model (BUM) for generating p-values (Pounds and Morris, 2003). Specifically, we consider two different BUM model settings, a one sided test setting where we examine the LHS kFWER and a two sided test where we examine both the LHS and RHS kFWER.

The BUM model represents a mixture model for generating p-values. With a BUM model with N tests, the probability density function (PDF) for the i^{th} p-value, p_i, with $i \in \{1, 2, \ldots, N\}$ is

$$g(p_i) = \begin{cases} (1-\pi) + \pi \frac{1}{B(\gamma,\mu)} p_i^{\gamma-1}(1-p_i)^{\mu-1}, & p_i, \in [0,1], \\ 0, & \text{otherwise}, \end{cases} \tag{30}$$

where B represents the Beta function. In other words, with probability π, p_i is an observation generated under the alternative hypothesis which is parameterized as a Beta distribution with shape parameters (γ, μ). Also with probability $(1-\pi)$, p_i is an observation generated under the null hypothesis which is parameterized as a uniform distribution on $[0, 1]$. For the one sided test setting we obtained the z-scores via $z_i = \Phi^{-1}(p_i)$. For the two sided test setting, we assumed the BUM p-values were two sided p-values where $p_i = 2 \cdot \min(\Phi(z_i), 1 - \Phi(z_i))$. To obtain the z scores, we employed a Bernoulli scheme with probability $= 0.50$ and we inverted (roughly) half of p-values using the standard normal distribution $z_i = \Phi^{-1}(p_i/2)$ and the other half of the p-values using $z_i = \Phi^{-1}(1 - p_i/2)$.

We note differences between the estimated empirical (red) and parametric (black) null densities in Figure 1, which contains the histogram of z-scores which were generated under the one sided test setting and with $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$, $\gamma = 0.3$, and $\mu = 5$. Such differences are more pronounced in the presence of an over-dispersed null distribution, and are responsible for differences in the operating characteristics of $\overline{kFWER}(\mathcal{Z})$ and $\widehat{kFWER}$.

Figure 2 displays true $kFWER(\mathcal{Z})$ and estimates $\overline{kFWER}(\mathcal{Z})$ and $\widehat{kFWER}$ values which were generated under the same simulation conditions as Figure 1 and across k values of $1, 5, 20$, and 50. Since F_0 is a standard normal CDF, we observed $kFWER(\mathcal{Z})$ and $\overline{kFWER}(\mathcal{Z})$ that were nearly identical except, with the differences easily attributable to the Monte-Carlo error associated with the estimation of $kFWER(\mathcal{Z})$. The shaded grey regions in Figure 2 demark the region between the 5th and 95th quantiles of $\widehat{kFWER}(\mathcal{Z})$, as estimated via 1000 Monte-Carlo simulations. Note, the grey shaded regions are rather narrow and centered slightly to the left $kFWER(\mathcal{Z})$ indicating that the $\widehat{kFWER}$ estimation procedure was slightly conservative in that simulation setting. Single step kFWER cutpoints for the Bonferroni, Holm, and Šidàk methods designed to control kFWER at 0.20 are also included in Figure 2. The cutpoints indicate that the Šidàk method was the most liberal among the three methods. As the value of k increased, the Šidàk and Holm cutpoints converged (bottom panels in Figure 2).

Figure 3 contains a histogram of z-scores which were generated under the two-sided test setting and with $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$, $\gamma = 0.3$, and $\mu = 5$. Figure 4, illustrates $kFWER(\mathcal{Z}), \overline{kFWER}(\mathcal{Z})$ estimates as a function of z and across k values of $1, 5, 20$, and 50. The shaded grey regions in Figure 4 demark the region between the 5th and 95th quantiles of $\widehat{kFWER}(\mathcal{Z})$, as estimated via 1000 Monte-Carlo simulations. Note, that the regions are rather narrow and centered around $kFWER(\mathcal{Z})$, indicating that the empirical estimation procedure is accurate.

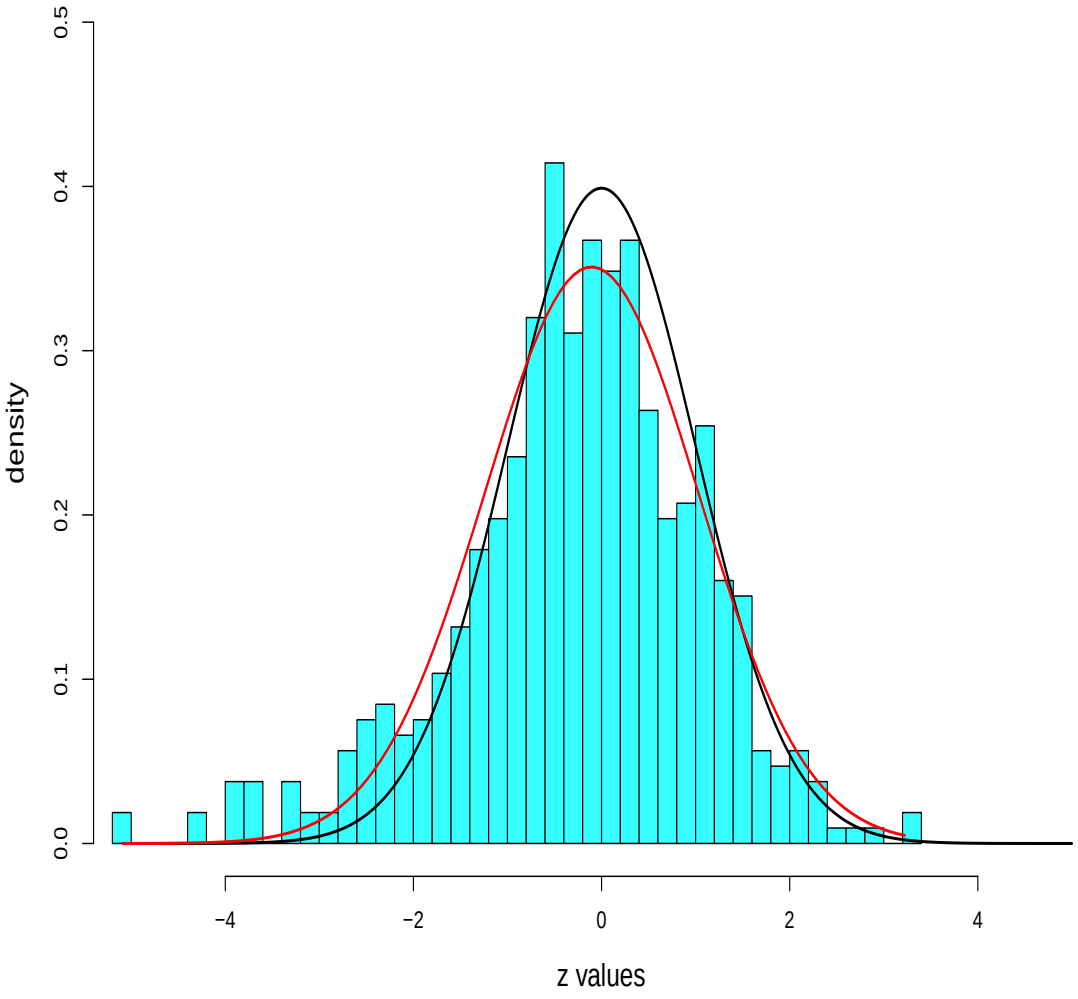

Figure 1: **BUM Simulation (One-Sided)**: The histogram (scaled to a total area of one) of z-values from a BUM model simulation with $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$ and F_0 a distribution function for a standard normal distribution and F_1 a distribution function that yields one sided p-values that follow a beta distribution with parameters $\gamma = 0.3$ and $\mu = 5$. The z-values were obtained via $z_i = \Phi^{-1}(p_i)$. Estimated empirical (red line) and parametric (black line) null densities are superimposed.

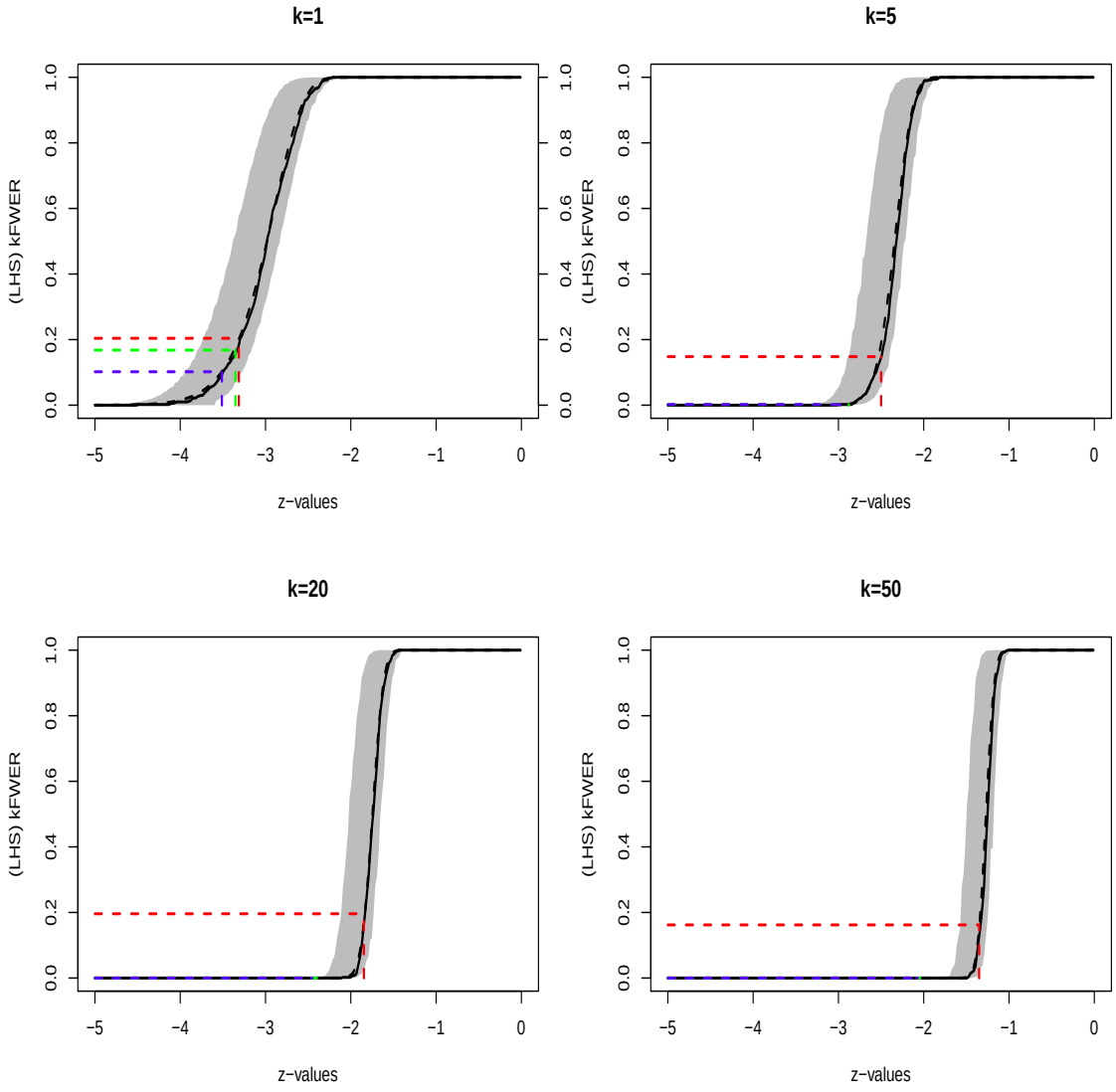

Figure 2: **BUM Simulation (One-Sided)**: A BUM model simulation study with 1000 replicate datasets. Data were simulated with $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$ and F_0 a distribution function for a standard normal distribution and F_1 a distribution function that yields one sided p-values that follow a beta distribution with parameters $\gamma = 0.3$ and $\mu = 5$. The z-values were obtained via $z_i = \Phi^{-1}(p_i)$. Parametric LHS $\overline{kFWER}$ (black broken lines) and the true LHS ($kFWER$ (black solid lines) from 1000 simulations are plotted. Since the BUM model assumes independence, the solid and broken lines are expected to be identical, with the observed differences attributable to Monte Carlo error. The grey region demarks the empirical 5th percentile and 95th percentile of the empirical kFWER ($\widehat{kFWER}$) as given in (22). The empirical kFWER appears very accurate under the given simulation conditions. The parametric Bonferroni, (average) Holm, and Šidàk z value cutoffs for kFWER $= 0.20$ are shown in green, blue, and red, respectively.

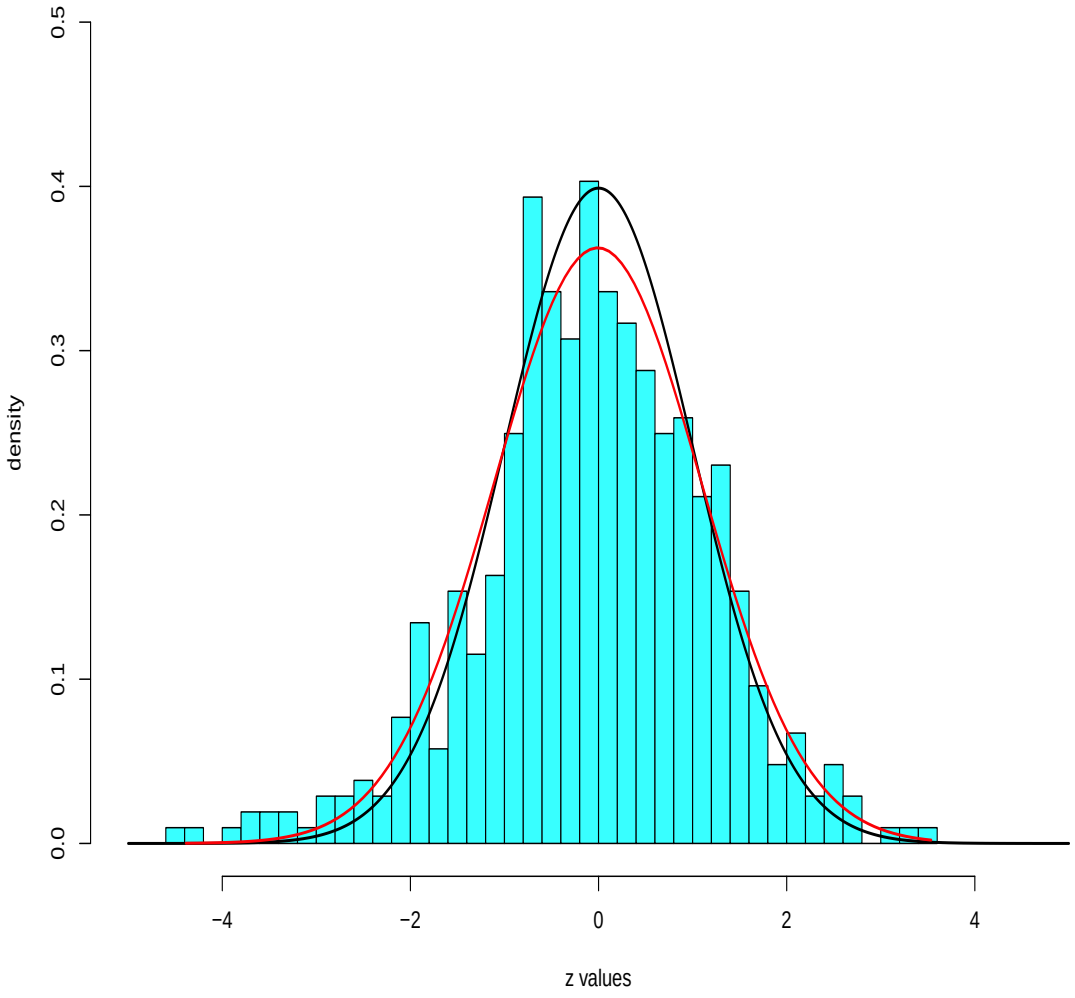

Figure 3: **BUM Simulation (Two-Sided)**: The histogram (scaled to a total area of one) of z-values from a BUM model simulation with $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$ and F_0 a distribution function for a standard normal distribution and F_1 a distribution function that yields one sided p-values that follow a beta distribution with parameters $\gamma = 0.3$ and $\mu = 5$. The z-values are obtained via $p_i = 2 \cdot \min\left(\Phi(z_i), 1 - \Phi(z_i)\right)$.

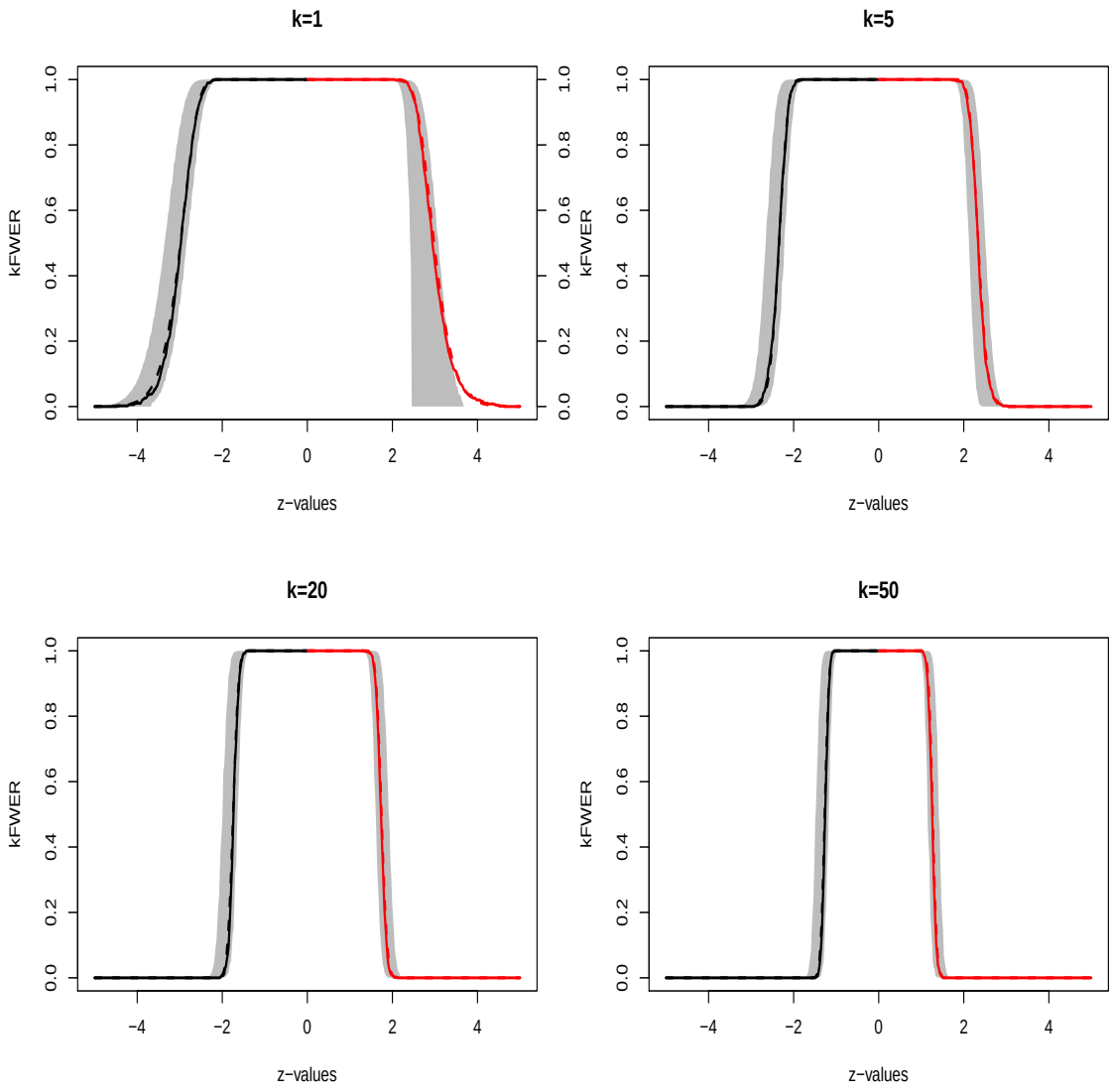

Figure 4: **BUM Simulation (Two-Sided)**: A BUM model simulation study with 1000 replicate datasets. Data were simulated with $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$ and F_0 a distribution function for a standard normal distribution and F_1 a distribution function that yields one sided p-values that follow a beta distribution with parameters $\gamma = 0.3$ and $\mu = 5$. The z-values are obtained via $p_i = 2 \cdot \min(\Phi(z_i), 1 - \Phi(z_i))$ where a probability of 0.50 is applied to choose between the two possible minimum values. Since the BUM model assumes independence, the solid and broken lines are expected to be identical, with the observed differences attributable to Monte Carlo error. The grey region demarks the empirical 5th percentile and 95th percentile of the empirical kFWER ($\widehat{kFWER}$) as given in (22). The empirical kFWER appears very accurate under the given simulation conditions.

5.2. Overdispersed Null Distribution Simulation

Utilizing a simulation study similar to one proposed in Efron (2010), we demonstrate that $\widehat{kFWER(\mathcal{Z})}$ can provides better results than $\overline{kFWER}(\mathcal{Z})$ when the null distribution of the test statistics is overdispersed. Such overdispersion can occur when unobserved covariates influence the cases in an experiment, such as one involving microarrays, and motivates the use of $\widehat{kFWER}$ compared to $\overline{kFWER}$.

Specifically, we considered simulated data of the form:

$$x_{ij} = u_{ij} + \frac{I_j}{2}\beta_i \qquad \begin{cases} u_{ij} \sim N(0,1) \\ \beta_i \sim N(0,\sigma_\beta^2) \end{cases} \tag{31}$$

with $i = 1, \ldots 16$ subjects corresponding to 8 subjects in each of two groups, and $j = 1, \ldots, 200$ 'feature' values (e.g. gene expression values) observed on each subject. Moreover, $u_{i1}, u_{i2}, \ldots, u_{in}, \beta_i$ were simulated to be mutually independent and

$$I_j = \begin{cases} -1 & j = 1, 2, \ldots, 8 \\ 1 & j = 9, 10, \ldots 16. \end{cases} \tag{32}$$

Under these conditions and the null hypothesis of equal means between groups, the standard two sample t-statistic, which we denote t_i, is distributed as a dilated t distribution with 14 degrees of freedom:

$$t_i \sim \left(1 + 4\sigma_\beta^2\right)^{1/2} \cdot t_{14}, \tag{33}$$

and with a dispersion factor of

$$(1 + 4\sigma_\beta^2)^{1/2}. \tag{34}$$

We transformed the t_i values to z_i values using $z_i = \Phi^{-1}(F_{14}(t_i))$ where F_{14} where Φ and F_{14} are the CDF for the standard normal and t_{14} distributions. As shown in Efron (2010), the empirical null distribution as estimated via the MLE method "correctly" estimated the null distribution. This improvement with respect to estimation of the null distribution enabled $\widehat{kFWER}(\mathcal{Z})$ to more accurately estimate kFWER compared to $\overline{kFWER}(\mathcal{Z})$. Figure 5 displays $\widehat{kFWER}(\mathcal{Z})$ and $\overline{kFWER}(\mathcal{Z})$ for a range of overdisperion factors with the number of false discoveries set at 20 ($k = 20$). Regardless of the overdispersion factor, $\widehat{kFWER}$ provided an accurate measure of kFWER, while $\overline{kFWER}$ was only acceptable, at most, at an overdispersion factor of 1.2 (upper left panel Figure 5).

5.3. Correlated Z Simulation

Utilizing another simulation study similar to one proposed in Efron (2010), we demonstrate that the presence of correlated test statistics can manifest notable differences in $\widehat{kFWER(\mathcal{Z})}$ and $\overline{kFWER}(\mathcal{Z})$. In short, correlation can effect our estimates of kFWER by influencing the estimation of the null distribution and the accuracy of our binomial assumption in (14). From the work in Efron (2010), the accuracy of the empirical null distribution depends on the root mean square (RMS) of the $N \cdot (N-1)/2$ pairwise correlations of the N length vector of z-values $\mathbf{z} = (z_1, z_2, \ldots, z_N)$. In the R package *locfdr*, the user can employ the function `simz` to generate an N length vector of z-values with a desired (approximate) RMS value. We simulated (approximate) RMS values of 0.10, 0.30, 0.40, and 0.45, where $N = 500$ with a kFWER setting where $k = 20$ and $\pi_0 = 0.95$. Marginally, each null gene follows a $N(0,1)$ distribution and each alternative gene follows a $N(-3,1)$ distribution. As shown in Figure 6, the standard errors for $\widehat{kFWER}$ were much larger than for $\overline{kFWER}$ and increased sharply as the RMS value increases. However, $\widehat{kFWER}$ appears to provide a relatively unbiased estimator of kFWER across all RMS settings while $\overline{kFWER}$ estimators begin to exhibit substantial bias as RMS increases.

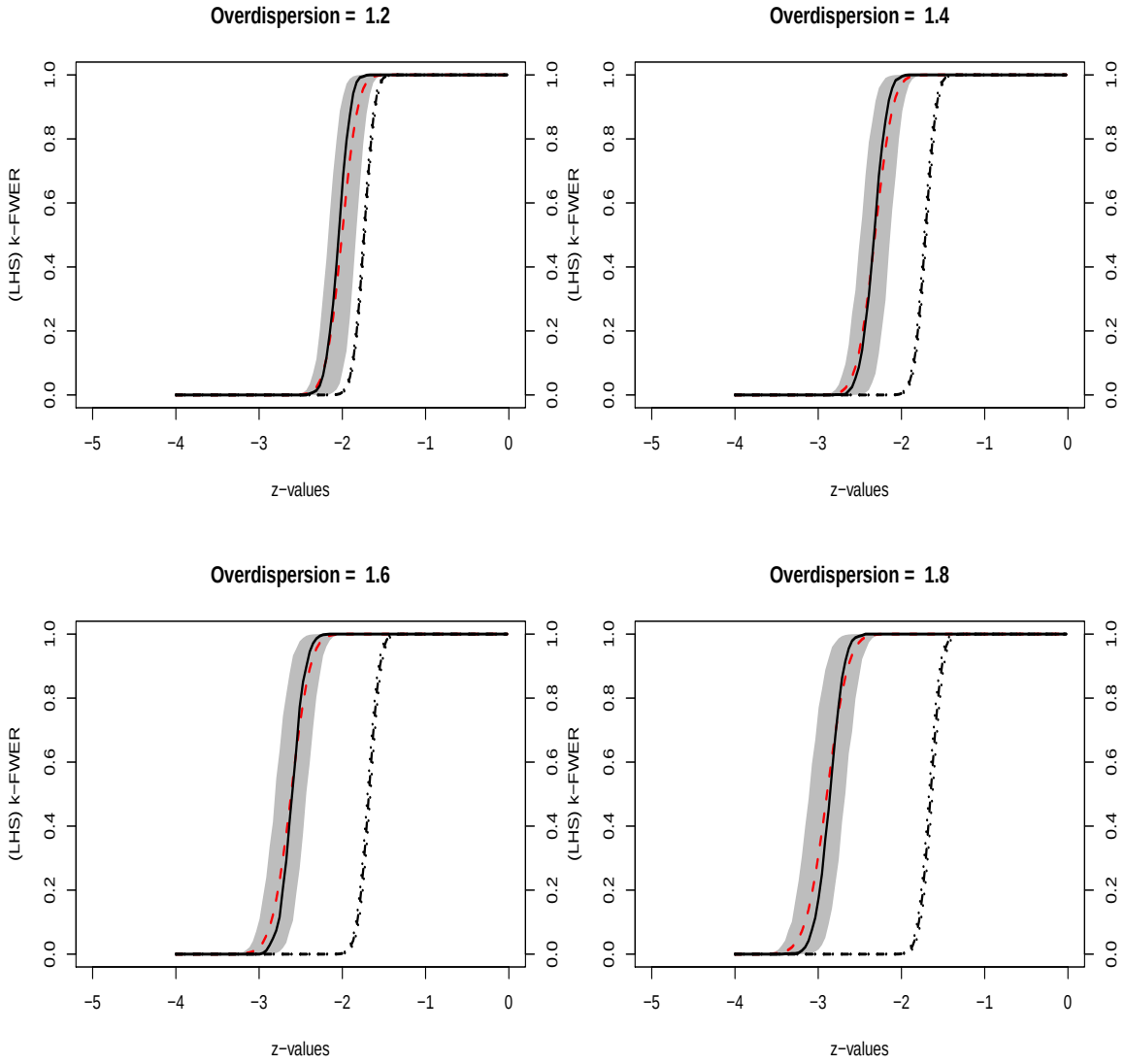

Figure 5: **Overdispersed Simulation (One-Sided)**: An overdispersed t distribution simulation with $N = 500$. The overdispersion factor is given in (34). Using 1000 simulations, we examined the parametric, empirical, and true $kFWER$. The true $kFWER$ is shown with a solid black line. The grey region displays the Monte Carlo based 5th percentile and 95th percentile of the empirical kFWER ($\widehat{kFWER}$) as given in (22) with the mean $\widehat{kFWER}$ displayed in the dashed red line. The dotted black lines represent the Monte Carlo based 5th percentile and 95th percentile of $\widehat{kFWER}$ with the mean shown as a dashed black line. In this simulation we see the empirical kFWER is very accurate, while the parametric kFWER is increasingly inaccurate as the overdispersion factor in (34) becomes larger.

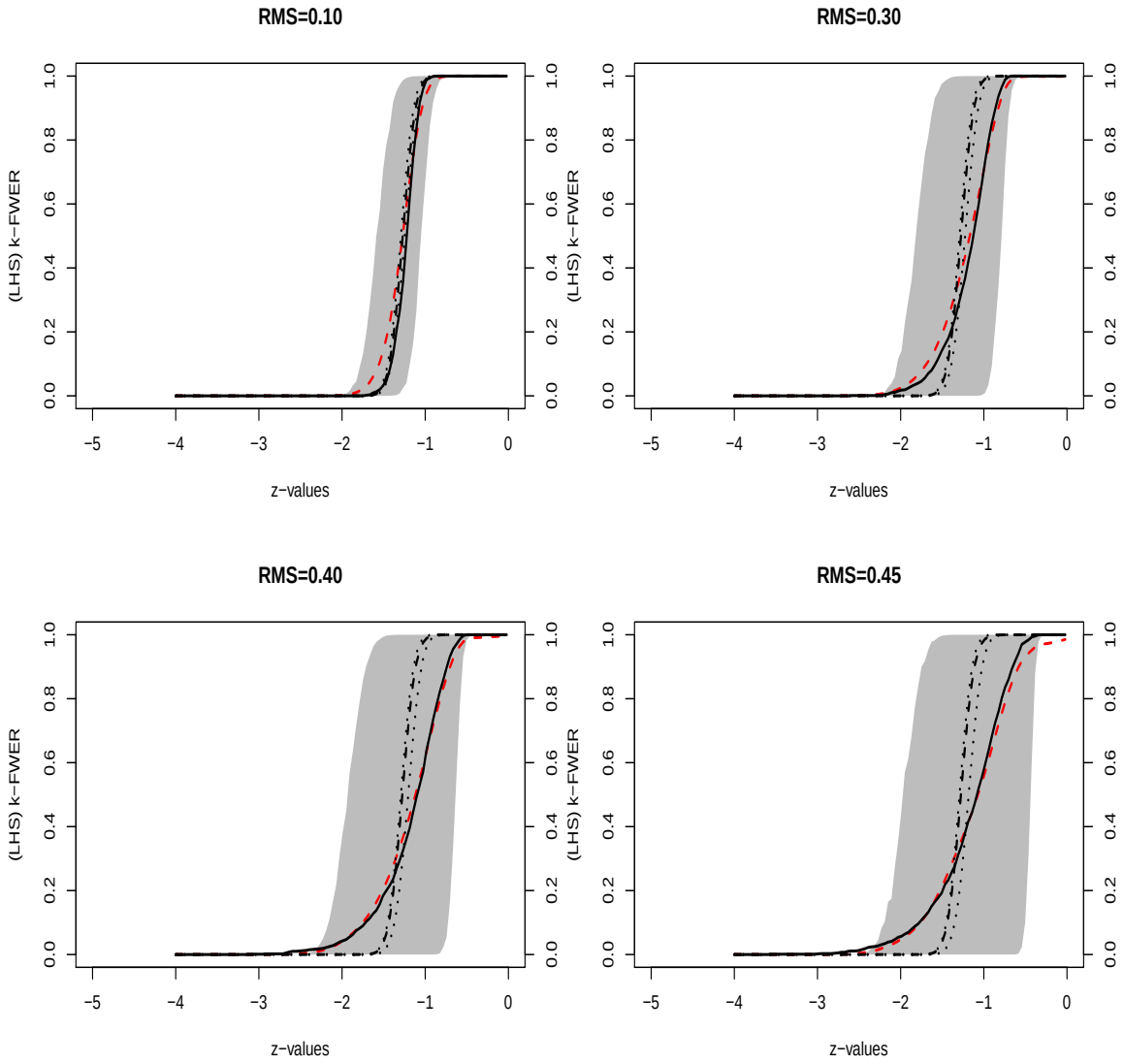

Figure 6: **RMS Simulation (One-Sided)**: A simulation with correlation where $N = 500$, $\pi_0 = 0.95$, $\pi_1 = 0.05$ and F_0 a distribution function for a multivariate normal distribution. Using 1000 simulations, we examined the parametric, empirical, and true $kFWER$. The grey region displays the Monte Carlo based 5th percentile and 95th percentile of the empirical kFWER ($\widehat{kFWER}$) as given in (22), while the red dashed line represents the mean $\widehat{kFWER}$. The dotted black lines represent the Monte Carlo based 5th percentile and 95th percentile of $\overline{kFWER}$ with the mean shown as a dashed black line. In this simulation we see the variability of $\widehat{kFWER}$ increases as the root mean square (RMS) of the correlation of the statistics increases.

5.4. Data Examples

We applied functional $kFWER$ methods to the leukemia and prostate cancer datasets referenced and analyzed in Efron (2010). These examples highlight a setting where the empirical null closely agrees with the parametric null (prostate cancer dataset) and an example where the empirical null is very different than the parametric null (leukemia dataset). Both datasets can be found at *http://www-stat.stanford.edu/∼omkar/monograph/data*.

For the prostate dataset originally presented in Singh et al. (2002), we examined microarray data designed to measure the level of gene expression for 6033 genes. These measurements were obtained for 102 men where 50 men were control subjects and 52 men had prostate cancer. The goal of this experiment was to discover genes associated with prostate cancer. As in Efron (2010), we performed two sample t-tests and transformed the t statistic for each gene to a z statistic via, $z_i = \Phi^{-1}(F_{100}(t_i))$ where F_{100} represents the CDF for a t distribution with 100 degrees of freedom.

For our analysis, we estimated $kFWER(\mathcal{Z})$. Figure 7 (a) shows the histogram of the z-values and Figure 7 (c) shows displays $\overline{kFWER}(\mathcal{Z})$, $\widehat{kFWER}(\mathcal{Z})$ for the prostate dataset. Since the parametric and empirical densities were in fair agreement (Figure 7(a)), our estimates of $\overline{kFWER}(\mathcal{Z})$, $\widehat{kFWER}(\mathcal{Z})$ in Figure 7(c) were similar for a variety of k, e.g. $k = 1, 5, 20$. Setting $k = 5$ and controlling (LHS) $\overline{kFWER}(\mathcal{Z})$ at 0.10 provided 27 discoveries compared to 16 discoveries with (LHS) $\widehat{kFWER}(\mathcal{Z}) \leq 0.10$ (see Table 1). For the RHS versions, there were 27 and 14 discoveries with $\overline{kFWER}(\mathcal{Z})$ and $\widehat{kFWER}(\mathcal{Z})$, respectively (see Table 1). Importantly, the difference in results between the empirical LHS and RHS versions of kFWER demonstrates the potential improvements of our method compared to standard kFWER testing with two sided p-values. Our functional kFWER estimator may have highlighted significance which was obscured when using standard kFWER methods (e.g. adjusted Bonferroni, Šidàk, and Holm) with two sided p-values. With one sided p-values (LHS or RHS), we examined parametric and empirical versions of the adjusted Bonferroni, Šidàk, and Holm procedures (see Table 1). The empirical based approaches provided slightly fewer discoveries when compared to their parametric counterparts, a result that can be attributed to the fact that the empirical distribution was estimated to be slightly wider than the parametric distribution (see Figure 7(a)).

The leukemia dataset originally presented in Golub et al. (1999) contains microarray data with 7128 gene expression levels for 72 patients, 45 with ALL (acute lymphoblastic leukemia) and 27 with AML (acute myeloid leukemia). The two sample t-tests (70 degrees of freedom) comparing AML with ALL patients were transformed to z-values, as described in Efron (2010).

Figure 7(b) provides the histogram (scaled) of z-values and the empirical and parametric null densities. The histogram is highly overdispersed compared to the $N(0, 1)$ (parametric) null distribution and, as a result, the parametric kFWER provides a greater number of "discoveries" when compared to the empirical estimate. Setting $k = 5$ and controlling (LHS) $\overline{kFWER}(\mathcal{Z})$ at 0.10, provided 444 discoveries compared to 83 discoveries when controlling (LHS) $\widehat{kFWER}(\mathcal{Z})$ at 0.10 (see Table 1). For the RHS, there were 302 and 47 discoveries at 0.10 with $k = 5$ when controlling the $\overline{kFWER}$ and $\widehat{kFWER}$, respectively. Importantly, we believe the empirical null is more in line with the expectations of researchers in this experiment. When discussing similar types of experiments with biologists and scientists, we are more likely to believe that there are relatively few changes between the two conditions. As stated in Efron (2010) regarding the parametric (theoretical) null for this example, "it seems more likely that there is something inappropriate about the theoretical null." The contrast between $\overline{kFWER}(\mathcal{Z})$ and $\widehat{kFWER}(\mathcal{Z})$ and the likelihood that the theoretical null is incorrect demonstrates the improvements gained by using the empirical version in conjunction with functional kFWER. We also examined the p-value based methods (Bonferroni, ,Šidàk, and Holm) for the leukemia dataset in Table 1. Since the empirical distribution has a much larger variance compared to the parametric distribution, we observed a large discrepancy in the number of discoveries between the two methods, with the parametric methods most likely having provided a large number of false positives.

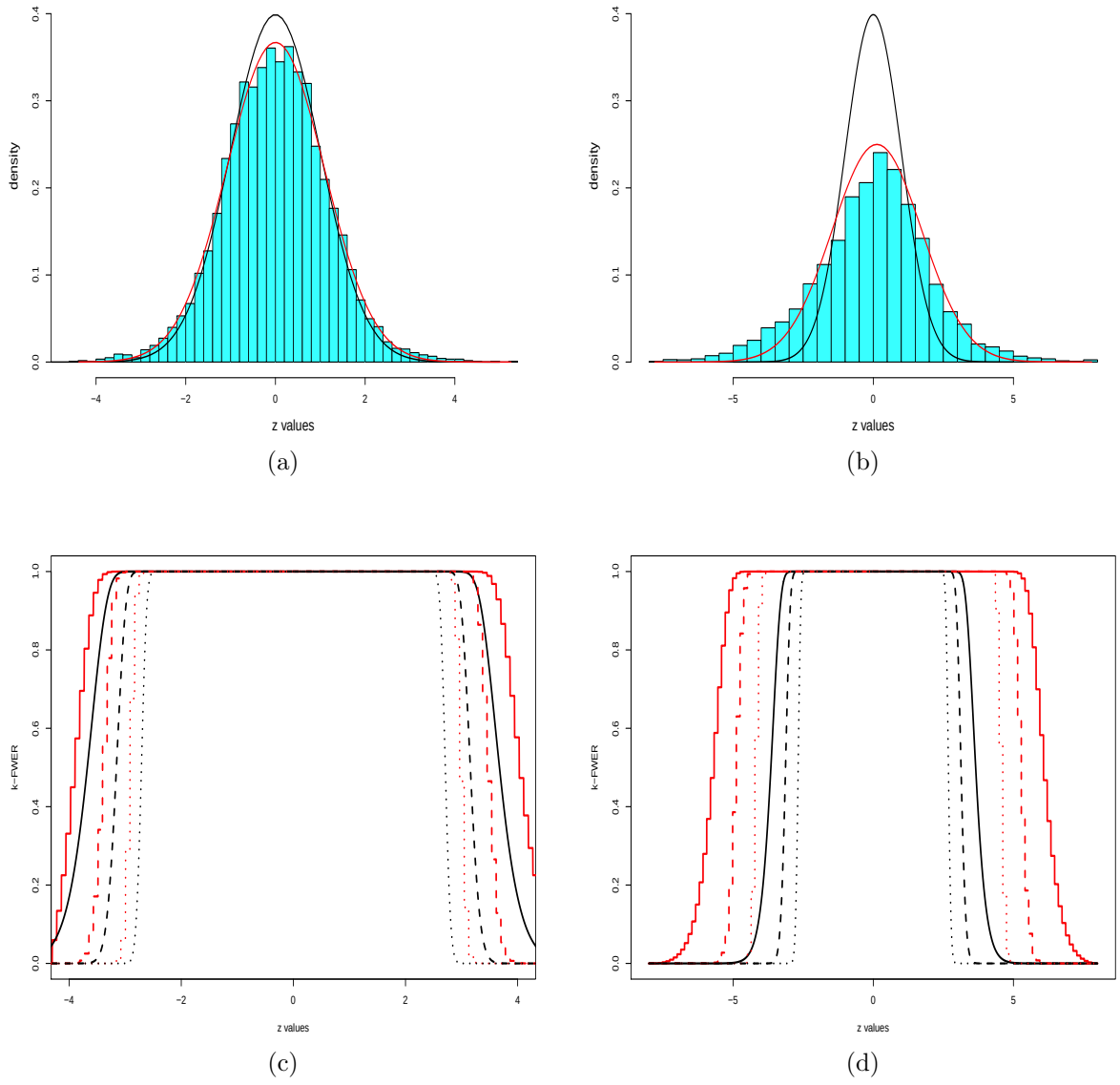

Figure 7: **Data Examples**: The prostate (a) and leukemia (b) dataset z histograms with the parametric null distribution in black and the empirical null distribution in red. The prostate (c) and leukemia (d) $kFWER$ for $k = 1$ in solid line, $k = 5$ in dashed line, and $k = 20$ in dotted line with the LHS versions of kFWER to left of 0 and RHS to right of 0. For (c) and (d), the $\widehat{kFWER}$ is in red and the $\overline{kFWER}$ is in black.

	z-value based		p-value based					
	pF	eF	pSIDAK	eSIDAK	pHOLM	eHOLM	pBON	eBON
LHS prostate	27	16	27	17	8	4	8	4
RHS prostate	27	14	26	14	13	8	13	8
LHS leukemia	444	83	426	75	312	35	308	35
RHS leukemia	302	47	286	43	209	23	207	23

Table 1: **Data Example Results:** Table of counts of significant genes in data examples (prostate and leukemia datasets) using various kFWER methods with $k = 5$ and $\alpha = 0.10$: pF denotes the number of significant genes when $\overline{kFWER}(\mathcal{Z}) = 0.10$, eF denotes the number of significant genes when $\widehat{kFWER}(\mathcal{Z}) = 0.10$, pSIDAK is the Šidàk method using parametric (LHS/RHS) p-values, eSIDAK is the Šidàk method with empirical F_0 ($\hat{F}_0$) to compute p-values, pHOLM is the Holm method using parametric (LHS/RHS) p-values, eHOLM uses Holm method with empirical F_0 ($\hat{F}_0$) to compute p-values, pBON is the adjusted Bonferroni method ($k = 5$) using parametric (LHS/RHS) defined p-values, eBON is the Bonferroni method with empirical F_o ($\hat{F}_0$) to compute p-values.

6. Discussion

Many of the works discussing kFWER controlling methods were designed with p-values in mind, e.g. Holm, Šidàk, and adjusted Bonferroni methods. Also more advanced techniques such as step up and step down procedures often commonly rely on the p-values, (e.g. Finos and Farcomeni (2011)). As an alternative to p-values, the maxT method presented in Dudoit et al. (2004) is a method designed to work on the test statistics rather than the p-values. These methods also propose similar techniques to determine p-values based on empirical distributions. Software to implement these stepwise procedures can be found in the *multtest* and *someKfwer* R packages, see Pollard, Ge, Taylor, and Dudoit (Pollard et al.) and Finos and Farcomeni (2010), respectively. Existing kFWER controlling procedures (e.g. Holm, Bonferroni, and Šidàk) are designed to report significant tests while controlling kFWER, without necessarily directly estimating kFWER. As shown in Miecznikowski et al. (2011), certain kFWER controlling methods can indirectly estimate kFWER more accurately than other controlling methods. Further, these more accurate methods are shown to have increased power in simulations (Miecznikowski et al., 2011; Roquain and Villers, 2011).

We have demonstrated that existing kFWER controlling procedures can be incorporated into our proposed novel functional method to estimate kFWER and that empirical versions of these traditional methods can be implemented using empirical null distributions. We have shown our empirical estimator of kFWER estimates kFWER more accurately in settings where the null distribution may be overdispersed due to unobserved covariates. These improvements in accuracy will lead to fewer false positives in situations with an overdispersed null distribution. Although a comprehensive power study was not within the scope of our present work, we expect our parametric kFWER estimator to have similar power to the Šidàk method in a two group independence setting. (See Figure 2 where the Šidàk method to control kFWER at 0.20 intersects our parametric kFWER estimator at 0.20.)

A major challenge facing many multiple testing schemes is robustness in light of correlated data. Many of the results for correlated data require the positive regression dependency subsets (PRDS) assumption, e.g. Cai and Sarkar (2006) and Sarkar (2008). This assumption is originally presented in Lehmann (1966) and can be difficult to assess in practice. Additionally, there are other common assumptions which may be violated in many situations where multiple testing procedures are desired. Further work with correlated tests are presented in Romano and Wolf (2005, 2010). As discussed in Efron (2010), there are several reasons for the theoretical null distribution to fail in practice, thus necessitating a need for an empirically based null distribution. Namely, failed mathematical assumptions, for example, violations of the independent and identically distributed assumption required for a two sample t statistic. Also, correlation across sampling units (e.g. patients) or correlation across cases (e.g. genes) could cause the theoretical null to be a poor estimate of the null distribution. Lastly, unobserved covariates such as age, weight or body mass index (BMI) can affect the estimation of the null distribution. Importantly, Efron (2010) in Chapter 7 has presented results for the MLE and central matching methods for estimating F_0 that discuss their performance (accuracy)

in light of correlation. Ultimately, for many correlated data settings, both methods to estimate the null distribution work well and it can be shown that their performance is heavily dependent on the root mean square of the correlation of the test statistics. Other methods to estimate null distributions can be found in Jin and Cai (2007) and Muralidharan (2010). Comparison of empirical methods for estimating the null distributions will be pursued as future work.

Note, our derivation for kFWER($\mathcal{Z}$) in (14) is based on the independence assumption, while the accuracy of $\widehat{F_0}$ and, consequently, $\widehat{\pi_0}$ depend on the root mean square of the $N \cdot (N-1)/2$ pairwise gene correlations. To extend our work to correlated settings, our group is developing a correlation based representation for kFWER($\mathcal{Z}$) using a sparse version of Bahadur's representation for correlated Bernoulli trials (Bahadur, 1959).

7. Conclusion

In this manuscript we have introduced a novel functional version of the generalized family wise error rate (kFWER). We have shown parametric and non parametric (empirical) versions to estimate $kFWER(\mathcal{Z})$. We have demonstrated that several popular kFWER controlling methods can be considered as a single point version of our functional estimator. Additionally we have adapted these popular kFWER controlling methods for use with empirical null distributions which work well in settings with an overdispersed null distribution or correlated statistics setting. Our version of kFWER offers improved inference compared to standard kFWER methods based on (two sided) p-values derived from parametric null distributions. We expect our version of kFWER will guide researchers in choosing biomarkers for validation in experiments controlling the generalized family wise error rate.

8. Acknowledgements

The authors are very grateful to David Tritchler for sharing his insights and providing helpful comments on an earlier version of this manuscript.

References

Bahadur, R. (1959). A representation of the joint distribution of responses to N dichotomous items. Technical report, Defense Technical Information Center Document.

Cai, G. and Sarkar, S. (2006). Modified Simes' critical values under positive dependence. *Journal of statistical planning and inference* **136,** 4129–4146.

Dudoit, S., Van Der Laan, M., and Pollard, K. (2004). Multiple testing Part I. single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology* **3,** Article 13.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association* **99,** 96–104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102,** 93–103.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge Univ Pr.

Efron, B., Turnbull, B. B., and Narasimhan, B. (2011). *locfdr: Computes local false discovery rates.* R package version 1.1-7.

Finos, L. and Farcomeni, A. (2010). *someKfwer: Controlling the Generalized Familywise Error Rate*. R package version 1.1.

Finos, L. and Farcomeni, A. (2011). k-fwer control without p-value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics* **67,** 174–181.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Guo, W. and Romano, J. (2007). A generalized Šidàk-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology* **6,** Article 3.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

Jin, J. and Cai, T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102,** 495–506.

Lehmann, E. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics* **37,** 1137–1153.

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33,** 1138–1154.

Miecznikowski, J. and Gaile, D. (2012). The local family wise error rate; a novel approach to examining generalized family wise error. *Technical Report #1203 University at Buffalo, Department of Biostatistics, Buffalo, NY* .

Miecznikowski, J., Gold, D., Shepherd, L., and Liu, S. (2011). Deriving and comparing the distribution for the number of false positives in single step methods to control k-fwer. *Statistics & Probability Letters* **81,** 1695–1705.

Muralidharan, O. (2010). An empirical bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics* **4,** 422–438.

Pollard, K. S., Ge, Y., Taylor, S., and Dudoit, S. *multtest: Resampling-based multiple hypothesis testing*. R package version 1.22.0.

Pounds, S. and Morris, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19,** 1236–1242.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100,** 94–108.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics* **38,** 598–633.

Roquain, E. and Villers, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *The Annals of Statistics* **39,** 584–612.

Sarkar, S. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *The Annals of Statistics* **36,** 337–363.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell* **1,** 203–209.

9. Appendix

In the following subsections we present results related to the adjusted Bonferroni, Šidàk, and Holm procedures. Additionally we present several methods to empirically estimate null distributions.

9.1. Adjusted Bonferroni Method

Theorem 1. *Using the two sample mixture model defined in* (1)*,* $z_{bon} = F_0^{-1}(k\alpha/n)$ *is such that* $kFWER\left(z_{bon}\right) \leq \alpha$ *when using the (LHS)* $kFWER$ *definition where* $kFWER(z) = kFWER(\mathcal{Z})$ *where* $\mathcal{Z} = (-\infty, z)$*.*

Proof. Let $\mathcal{Z} = (-\infty, z_{bon})$. Then,

$$\begin{aligned}
\text{(LHS) } kFWER(z_{bon}) = Pr(N_0(\mathcal{Z}) \geq k) \quad &\leq \tfrac{E(N_0(\mathcal{Z}))}{k} \text{ (by Markov Inequality)} \\
&= E\left(E(N_0(\mathcal{Z})|N_0)\right)/k \text{ (Law of total expectation)} \\
&= E\left(N_0 F_0(z_{bon})\right)/k \text{ (since } N_0(\mathcal{Z})|N_0 \sim Bin(N_0, F_0(z_{bon}))) \\
&= N\pi_0 F_0\left(F_0^{-1}(k\alpha/N)\right)/k \text{ (since } N_0 \sim Bin(N, \pi_0)) \\
&= \pi_0\alpha \leq \alpha.
\end{aligned}$$

□

Corollary 9.1. *Using* $z_{bon} = F_0^{-1}(1 - k\alpha/n)$ *and the (RHS)* $kFWER$ *definition, we show that (RHS)* $kFWER\left(z_{bon}\right) \leq \alpha$ *where (RHS)* $kFWER(z) = kFWER(\mathcal{Z})$ *where* $\mathcal{Z} = (z, \infty)$*.*

Proof. A similar argument to Theorem 1 establishes the (RHS) kFWER result. □

9.2. Šidàk Method

Theorem 2. *Consider the two sample mixture model defined in* (1)*,* $z_{sid} = F_0^{-1}(p_{sid})$ *where* p_{sid} *is such that*

$$F_b(k-1|N, p_{sid}) = 1 - \alpha. \tag{35}$$

Then $(LHS)\, kFWER\left(z_{sid}\right) \leq \alpha$ *when using the (LHS)* $kFWER$ *definition where* $kFWER(z) = kFWER(\mathcal{Z})$ *where* $\mathcal{Z} = (-\infty, z)$*.*

Proof. We can assume that $p_{sid} = F_0(z_{sid})$. Then we have

$$(LHS)\, kFWER(z_{sid}) = \sum_{i=k}^{N} \left(1 - F_b(k-1|i, p_{sid})\right) f_b(i|N, \pi_0). \tag{36}$$

Note from (35) we can assume that $F_b(k-1|N, p_{sid}) = 1 - \alpha$ and so $1 - F_b(k-1|N, p_{sid}) = \alpha$. Hence the last term of the sum in (36) is $\alpha\pi_0^N$. Importantly, we further note that $F_b(k-1|n, p) > F_b(k-1|N, p)$ for any $n < N$. Thus,

$$1 - F_b(k-1, N, p) > 1 - F_b(k-1, n, p). \tag{37}$$

Thus, we can rewrite (36) with the understanding that $(LHS)\, kFWER(z_{sid}) = kFWER(\mathcal{Z})$ with $\mathcal{Z} = (-\infty, z_{sid})$ as follows,

$$\begin{aligned}
(LHS)\, kFWER(z_{sid}) = \quad &\textstyle\sum_{i=k}^{N} \left(1 - F_b(k-1|i, p_{sid})\right) \cdot f_b(i|N, \pi_0) \\
&\leq \alpha\left[f_b(k|N, \pi_0)) + f_b(k+1|N, \pi_0) + \cdots + \right. \\
&\quad \left. f_b(N|N, \pi_0)\right] \\
&\leq \alpha\left(1 - F_b(k-1|N, \pi_0)\right) \\
&\leq \alpha \;\; \text{(since } 1 - F_b(k-1|N, \pi_0) \leq 1).
\end{aligned} \tag{38}$$

□

Corollary 9.2. *Similarly, if we define* $z_{sid} = F_0^{-1}(1 - p_{sid})$ *where* p_{sid} *is such that*

$$F_b(k-1|N, p_{sid}) = 1 - \alpha. \tag{39}$$

Then $(RHS)\, kFWER\,(z_{sid}) \leq \alpha$ *when using the (RHS)* $kFWER$ *definition where (RHS)* $kFWER(z) = kFWER(\mathcal{Z})$ *where* $\mathcal{Z} = (z, \infty)$.

Proof. Similar to the proof for Theorem 2. □

9.3. Holm Method

Theorem 3. *Consider the two sample mixture model defined in* (1) *and* $z_{holm} = F_0^{-1}\left(p_{(r)}\right)$ *where* r *is the largest index satisfying* (28). *Then (LHS)* $kFWER\,(z_{holm}) \leq \alpha$ *when using the (LHS)* $kFWER$ *definition with one sided p-values defined by* $p_i = F_0(z_i)$.

Proof. If $r \leq k$, then our α adjustment is the same as the Bonferroni α adjustment and we can use the Markov inequality as employed in the proof for the adjusted Bonferroni argument.

Assume $r > k$, then we use the technique employed in Lehmann and Romano (2005). Let $y_1, y_2, \ldots, y_{N_0}$ denote the ordered z statistics for the true null hypotheses where $y_1 \leq y_2 \leq y_3 \cdots \leq y_{N_0}$. Then let $z_j = y_k$ where $z_1 \leq z_2 \leq \cdots \leq z_N$ denote the ordered z statistics. Thus, the following probability statements hold,

$$\begin{aligned} (LHS)\, kFWER\,(z_{holm}) &= Pr\left(\{N_0(-\infty, z_{holm}) \geq k\}\right) \\ &= Pr(\#\text{ of null z-values} \in (-\infty, z_{holm}) \geq k). \end{aligned} \tag{40}$$

The event $\{\#$ of null z-values $\in (-\infty, z_{holm}) \geq k\}$ is equal to the event $\{y_K = z_j \leq z_{holm}\}$. In order to reject at least k true nulls, the largest possible value of j is $N - N_0 + k$, namely, the situation where the $N - N_0$ true alternatives are the smallest z statistics. Hence, we have that $y_k = z_j \leq z_{N-N_0+k}$. Hence, we have $Pr(\#$ of null z-values $\in (-\infty, z_{holm}) \geq k) = Pr(\{y_k = z_j \leq z_{N-N_0+k}\})$. Now apply F_0 to the event $\{y_k = z_j \leq z_{N-N_0+k}\}$ which is a non decreasing function in order to obtain, $F_0(y_k) = F_0(z_j) = p_j \leq F_0(z_{N-N_0+k}) = p_{N-N_0+k}$. However, since $z_j \leq z_{holm}$ we must have that $p_j \leq \alpha_j$. Also, $\alpha_j \leq \alpha_{N-N_0+k} = k\alpha/N_0$ since α is an increasing function, see (29). Hence,

$$\begin{aligned} (LHS)\, kFWER\,(z_{holm}) &= Pr(\#\text{ of null z-values} \in (-\infty, z_{holm}) \geq k) \\ &= Pr(\{y_k = z_j \leq z_{N-N_0+k}\}) \\ &= Pr\left(\{F_0(y_k) = F_0(z_j) \leq p_{N-N_0+k}\}\right) \\ &\leq Pr(F_0(y_k) = F_0(z_j) = p_j \leq \alpha_{N-N_0+k} = k\alpha/N_0. \\ &= Pr\left(U_k \leq \tfrac{k\alpha}{N_0}\right) \text{ where } U_k \equiv F_0(y_k);\text{ the } k\text{-th null p-value} \\ &= Pr(W > k) \text{ where } W \sim \text{Bin}(N_0, k\alpha/N_0) \\ &\leq E(W)/K \text{ By Markov inequality} \\ &= \alpha \end{aligned} \tag{41}$$

□

Corollary 9.3. *Consider the two sample mixture model defined in* (1) *and* $z_{holm} = 1 - F_0^{-1}\left(p_{(r)}\right)$ *where* r *is the largest index satisfying* (28). *Then* $kFWER\,(z_{holm}) \leq \alpha$ *when using the (RHS)* $kFWER$ *definition where* $kFWER(z) = kFWER(\mathcal{Z})$ *where* $\mathcal{Z} = (z, \infty)$.

Proof. Similar to the proof for Theorem 3. □

9.4. Methods to estimate the null distribution

In this section, we paraphrase two methods described in Efron (2010) for estimating the null distribution. We assume that $f_0(z)$ is normal but not necessarily $\mathcal{N}(0,1)$ say,

$$f_0(z) \sim \mathcal{N}(\delta_0, \sigma_0^2), \tag{42}$$

and we define $f_{\pi 0}(z) = \pi_0 f_0(z)$. This implies that

$$\log(f_{\pi 0}(z)) = \left[\log(\pi_0) - \frac{1}{2}\left\{\frac{\delta_0^2}{\sigma_0^2} + \log(2\pi\sigma_0^2)\right\}\right] + \frac{\delta_0}{\sigma_0^2}z - \frac{1}{2\sigma_0^2}z^2, \tag{43}$$

is a quadratic function of z.

9.5. The MLE method for empirical null distribution

This method was first introduced in Efron (2007). The maximum likelihood estimator (MLE) method starts with the zero assumption, where we assume that $f_1(z)$ is zero for a certain subset $\mathcal{A}_0$ of the sample space. In other words,

$$f_1(z) = 0 \text{ for } z \in \mathcal{A}_0. \tag{44}$$

We assume N_0 is the number of z_i in $\mathcal{A}_0$ and $\mathcal{I}_0$ their indices, $\mathcal{I}_0 = \{i : z_i \in \mathcal{A}_0\}$ and $N_0 = \#\mathcal{I}_0$. We define $\mathbf{z}_0$ as the corresponding collection of z-values,

$$\mathbf{z}_0 = \{z_i, i \in \mathcal{I}_0\}. \tag{45}$$

Also, let $\varphi_{\delta_0,\sigma_0}(z)$ be the $N(\delta_0, \sigma_0^2)$ density function,

$$\varphi_{\delta_0,\sigma_0}(z) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2}\left(\frac{z-\delta_0}{\sigma_0}\right)^2\right\} \tag{46}$$

and

$$H_0(\delta_0, \sigma_0) \equiv \int_{\mathcal{A}_0} \varphi_{\delta_0,\sigma_0}(z)dz, \tag{47}$$

this being the probability that a $N(\delta_0, \sigma_0^2)$ variate falls in $\mathcal{A}_0$.

We suppose that the N z_i values follow the two-group model (1) with $f_0 \sim N(\delta_0, \sigma_0^2)$ and $f_1(z) = 0$ for $z \in \mathcal{A}_0$. Then $\mathbf{z}_o$ has density and likelihood function

$$f_{\delta_0,\sigma_0,\pi_0}(\mathbf{z}_0) = \left[\binom{N}{N_0}\theta^{N_0}(1-\theta)^{N-N_0}\right]\left[\prod_{\mathcal{I}_0}\frac{\varphi_{\delta_0,\sigma_0}(z_i)}{H_0(\delta_0,\sigma_o)}\right] \tag{48}$$

when $\theta = \pi_0 H_0(\delta_0, \sigma_0) = Pr(\{z_i \in \mathcal{A}_0\})$.

Computations can produce maximum likelihood estimators $(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0)$; $f_{\delta_0,\sigma_0,\pi_0}(\mathbf{z}_0)$ is the product of two exponential families which can be solved separately (the two bracketed terms). The binomial term gives $\hat{\theta} = N_0/N$ while $\hat{\delta}_0$ and $\hat{\sigma}_0$ are the MLEs from a truncated normal family, obtained by familiar iterative calculations, finally yielding

$$\hat{\pi}_0 = \hat{\theta}/H_o(\hat{\delta}_0, \hat{\sigma}_0). \tag{49}$$

The log of (48) is concave in $(\delta_0, \sigma_0, \pi_0)$ guaranteeing that the MLE solutions are unique. This is described more fully in Section 6.3 of Efron (2010).

9.6. The central matching method for empirical null distribution

This method was first introduced in Efron (2004). In this method, we define y_k as the number of observations z_i in the kth bin,

$$y_k = \#\{z_i \in \mathcal{Z}_k\}, \tag{50}$$

where we partition the range $\mathcal{Z}$ of z_i values into K bins of equal width d with

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k. \tag{51}$$

Then, with the central matching method, we estimate $f_0(z)$ and π_0 by assuming that $\log(f(z))$ is quadratic near 0 and equal to (43) with,

$$\log(f(z)) \approx \beta_0 + \beta_1 z + \beta_2 z^2. \tag{52}$$

Estimating $(\beta_0, \beta_1, \beta_2)$ can be done using least squares with the histogram counts y_k around $z = 0$ and matching coefficients between (43) and (52). In other words, via matching, we obtain,

$$\sigma_0^2 = -1/(2\beta_2), \tag{53}$$

$$\delta_0 = -\beta_1/(2\beta_2), \tag{54}$$

$$\log \pi_0 = \beta_0 - \frac{\beta_1^2}{4\beta_2} + \log(-\pi/\beta_2). \tag{55}$$