Published in final edited form as:

Int J Biostat. 2008 ; 4(1): . doi:10.2202/1557-4679.1103.

Exact Calculations of Average Power for the Benjamini-Hochberg Procedure*

Deborah H. Glueck^{*}, Jan Mandel[†], Anis Karimpour-Fard[‡], Lawrence Hunter^{**}, and Keith E. Muller^{††}

^{*}University of Colorado Denver, deborah.glueck@uchsc.edu [†]University of Colorado Denver, jan.mandel@ucdenver.edu [‡]University of Colorado Denver, anis.karimpour-fard@ucdenver.edu ^{**}University of Colorado Denver, larry.hunter@ucdenver.edu ^{††}University of Florida, keith.muller@biostat.ufl.edu

Abstract

Exact analytic expressions are developed for the average power of the Benjamini and Hochberg false discovery control procedure. The result is based on explicit computation of the joint probability distribution of the total number of rejections and the number of false rejections, and expressed in terms of the cumulative distribution functions of the p-values of the hypotheses. An example of analytic evaluation of the average power is given. The result is confirmed by numerical experiments and applied to a meta-analysis of three clinical studies in mammography.

Keywords

hypothesis testing; multiple comparisons; false discovery; distribution of rejections; meta-analysis

1 Introduction

The Benjamini and Hochberg (1995) procedure (BH procedure) for simultaneous testing of multiple independent hypotheses aims to control the False Discovery Rate, defined as the expected ratio of false rejections to the number of total rejections.

The power of a test of a single hypothesis is the probability of rejecting the null hypothesis. This rejection is correct when the alternative hypothesis is true, but is described as a Type I error when the null hypothesis is true. The average power (Benjamini and Liu, 1999) of a multiple hypothesis test is defined as the ratio of the expectation of the number of correct rejections (a random variable) and the number of hypotheses for which the alternative holds (assumed to be known).

In this paper, we provide exact analytic results for the average power for the BH procedure. The results presented here are based on analytic formulas for the joint distribution of the number of total rejections and the number of false rejections, given the number of the tests where the null hypothesis holds.

* Glueck was supported by NCI K07CA88811. Mandel was supported by NSF grants CNS-0325314, CNS-0719641 and DMS-0623983. Muller was supported by NCI P01 CA47 982-04, NCI R01 CA095749-01A1 and NIAID 9P30 AI 50410. Hunter was supported by NLM 5R01LM008111-03 and NCI 5 P30 CA46934-15.

In Glueck et al. (2008b), the average power of the BH procedure was studied by decomposing the rejection region into a large number of elementary regions, given by simple inequalities, and integrating probability densities over the elementary regions. In this paper, a much more efficient method is developed, based on cumulative distribution functions rather than integrals. The results of this paper were made possible by recent work on computing the joint cumulative distribution function of order statistics of samples from two populations (Glueck et al., 2008a).

Many papers describe extensions of the power theory for false discovery rate, or related statistics, e.g., (Curran-Everett, 2000; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001; Efron et al., 2001; Genovese and Wasserman, 2002, 2004; Storey, 2002; Finner and Roters, 2002; Sarkar, 2002, 2004, 2006). Power has been studied almost entirely via simulation (Benjamini and Liu, 1999; Storey, 2002; Keselman et al., 2002; Lee and Whitmore, 2002). Jung (2005) derived analytic power results for the Storey procedure. Ferreira and Zwinderman (2006) used asymptotic methods for an adaptive version of the BH procedure. In Section 2, we review the BH procedure and find its exact average power. In Section 3, we give explicit formulae for a simple example with three hypotheses. In Section 4, we confirm the accuracy of the result by comparison to simulations. In Section 5, we give an application to a meta-analysis of clinical trials in mammography. Section 6 is the discussion and conclusion.

2 Exact power analysis

2.1 The BH procedure

Given $\alpha_* \in [0, 1]$, null hypotheses H_{0i} , $i = 1, 2, \dots, m$, with independent but not necessarily identically distributed p-values P_i , and corresponding order statistics $P_{(i)}$ (the p-values P_i sorted in non-decreasing order), the BH procedure produces a nondecreasing sequence of rejection bounds $b_i = i\alpha_*/m \in (0, 1)$, rejects the k null hypotheses $H_{0(i)}$, $i = 1, 2, \dots, k$, $k \leq m$, such that k is the largest number for which $P_{(k)} \leq b_k$, and does not reject all other hypotheses. Given that the null hypothesis H_{0i} is true, the p-value P_i is defined as the probability of obtaining a realization of the data which is as or more unlikely than the observed data.

2.2 Notation and assumptions

For the purposes of power analyses, we define alternative hypotheses H_{1i} . Assume that the true state of nature is known, i.e., whether the null or the alternative hypothesis is true. Without loss of generality, suppose that H_{0i} is true for $i \leq n$, and that H_{1i} is true for $i > n$. Let F_i be the cumulative distribution function of the p-value P_i . Assume that F_i is continuous.

Denote by K the number of null hypotheses rejected by the BH procedure. Because the data are stochastic, K is a random variable, and the rejection of exactly k out of m hypotheses is the event

$$\{K=k\} = \{P_{(k)} \leq b_k\} \cap \{P_{(k+1)} \geq b_{k+1}\} \cap \dots \cap \{P_{(m)} \geq b_m\}. \quad (1)$$

When the null hypothesis is true, the rejection of the null is a false rejection. In the notation here, false rejection occurs when $i \leq n$ and $P_i \leq b_k$. Denote by J the number of false rejections and by $G(k, j)$ the event The occurrence of exactly j false rejections out of a total of exactly k rejections is the event

$$G(k, j) = \{P_i \leq b_k \text{ for exactly } j \text{ indices } i \text{ out of } i=1, \dots, n\}. \quad (2)$$

The occurrence of exactly j false rejections out of a total of exactly k rejections is the event

$$\{K=k\} \cap \{J=j\} = \{K=k\} \cap G(k, j). \quad (3)$$

2.3 Distribution of the number of total and false rejections

To compute the power, we need to find the probability of the events $\{K\} = k \cap \{J=j\}$, that is, the joint probability distribution of the total number of rejections K and the number of false rejections J . From (1) and (3),

$$\begin{aligned} \Pr(\{K=k\} \cap \{J=j\}) \\ = \Pr[\{P_{(k)} \leq b_k\} \cap \{P_{(k+1)} \geq b_{k+1}\} \cap \cdots \cap \{P_{(m)} \geq b_m\} \cap G(k, j)]. \end{aligned} \quad (4)$$

It is shown in Appendix A that (4) yields

$$\begin{aligned} \Pr(\{J=j\} \cap \{K=k\}) &= F_{(k)}[b_k; G(k, j)] \\ &- \sum_{r=k+1}^m F_{(k)(r)}[b_k, b_r; G(k, j)] \\ &+ \sum_{\substack{r,s=k+1 \\ r \neq s}}^m F_{(k)(r)(s)}[b_k, b_r, b_s; G(k, j)] - \cdots \\ &+ (-1)^{m-k} F_{(k)\dots(m)}[b_k, \dots, b_m; G(k, j)], \end{aligned} \quad (5)$$

where

$$\begin{aligned} F_{(n_1)\dots(n_e)}[p_{n_1}, \dots, p_{n_e}; G(k, j)] \\ = \Pr[\{P_{(n_1)} \leq p_{n_1}\} \cap \cdots \cap \{P_{(n_e)} \leq p_{n_e}\} \cap G(k, j)] \end{aligned} \quad (6)$$

denotes the joint cumulative distribution function of the order statistics, jointly with the event $G(k, j)$.

To evaluate the distribution functions in (5), we use an extension of the method developed by Glueck et al. (2008a). Define $n_1 = k$, so $p_{n_1} = b_k$ and let $p_{n_0} = 0$ and $p_{n_{e+1}} = 1$. Define the index vector $\mathbf{i} = (i_0, i_1, \dots, i_{e+1})$, the summation index set

$$\mathcal{I} = \left\{ \mathbf{i}: \begin{array}{l} 0 = i_0 \leq i_1 \leq \cdots \leq i_e \leq i_{e+1} = m, \\ \text{and } i_h \geq n_h \text{ for all } 1 \leq h \leq e \end{array} \right\}, \quad (7)$$

the index matrix $\boldsymbol{\mu} = [\mu_{ih}]$, $i = 1, \dots, m$, $h = 0, \dots, e$, and the index set $\mathcal{M}$, consisting of all $\boldsymbol{\mu}$ such that

$$\mu_{ih} \in \{0, 1\}, \text{ for all } i \text{ and all } h, \quad (8)$$

$$\sum_{i=1}^n \mu_{i0} = j, \quad (9)$$

$$\sum_{i=1}^m \mu_{ih} = i_{h+1} - i_h, \text{ for all } h=1, \dots, e, \quad (10)$$

$$\sum_{h=0}^e \mu_{ih} = 1. \quad (11)$$

It is proven in Appendix B that

$$F_{(n_1) \dots (n_e)}[p_{n_1}, \dots, p_{n_e}; G(k, j)] = \sum_{\mathbf{i} \in \mathcal{I}} \sum_{\mu \in \mathcal{M}} \sum_{h=0}^e \prod_{i: \mu_{ih}=1} [F_i(p_{n_{h+1}}) - F_i(p_{n_h})]. \quad (12)$$

Note that the sum over $\mathcal{I}$ in (12) can be written as

$$\sum_{\mathbf{i} \in \mathcal{I}} = \sum_{i_e=n_e}^m \sum_{i_{e-1}=n_{e-1}}^{i_e} \dots \sum_{i_1=n_1}^{i_2}, \quad (13)$$

which is readily implemented as a collection of loops. The sum in (12) could be also written as sum over a subset of permutations of the set $\{1, \dots, m\}$, much like the sum over all permutations can be written as the sum over all square zero-one matrices with unit row and column sums.

The sum (12) is in fact a selection of certain terms in the expression of the joint distribution function of order statistics as a sum of permanents, following Bapat and Beg (1989). However, the present form appears to be simpler and thus easier to program. It avoids the repetition of terms that occur in the formulation with permutations.

2.4 Average power of the BH procedure

The following formulation provides a computable expression for average power.

Theorem 2.1—For $m \neq n$, i.e. when the null is not true for every hypothesis, the average power of the BH procedure is

$$E\left(\frac{K-J}{m-n}\right) = \sum_{k=1}^m \sum_{j=0}^n \frac{k-j}{m-n} \Pr\{(K=k) \cap (J=j)\}, \quad (14)$$

where $\Pr\{(K=k) \cap (J=j)\}$ is given by (5) with (12).

When $m = n$, all the hypotheses are null, and the average power is not defined.

It should be noted that the sum over j from 0 to n in (14) can be replaced by the sum from $\max\{0, k-(m-n)\}$ to $\min\{n, k\}$ as the other terms are always zero.

3 Example

Consider the computation of average power for the BH procedure with three hypotheses, so $m = 3$. Suppose that one null and two alternative hypotheses are true, and $n = 1$. From (14), the expected power is

$$E \left(\frac{K-J}{m-n} \right) = \frac{1-0}{3-1} \Pr \{ (K=1) \cap (J=0) \} \\ + \frac{2-0}{3-1} \Pr \{ (K=2) \cap (J=0) \} + \frac{2-1}{3-1} \Pr \{ (K=2) \cap (J=1) \} \\ + \frac{3-0}{3-1} \Pr \{ (K=3) \cap (J=0) \} + \frac{3-1}{3-1} \Pr \{ (K=3) \cap (J=1) \} \quad (15)$$

where, from expansion (5),

$$\Pr \{ (K=3) \cap (J=1) \} \\ = F_{(3)} [b_3; G(3, 1)] \\ = F_1(b_3) F_2(b_3) F_3(b_3), \quad (16)$$

$$\Pr \{ (K=2) \cap (J=0) \} \\ = F_{(2)} [b_2; G(2, 0)] - F_{(2)(3)} [b_2, b_3; G(2, 0)] \\ = F_2(b_2) F_3(b_2) - F_1(b_3) F_2(b_2) F_3(b_2), \quad (17)$$

$$\Pr \{ (K=2) \cap (J=1) \} \\ = F_{(2)} [b_2; G(2, 1)] - F_{(2)(3)} [b_2, b_3; G(2, 1)] \\ = F_1(b_2) F_2(b_2) + F_1(b_2) F_3(b_2) - F_1(b_2) F_2(b_3) F_3(b_2) \\ - F_1(b_2) F_2(b_2) F_3(b_3), \quad (18)$$

$$\Pr \{ (K=1) \cap (J=0) \} \\ = F_{(1)} [b_1; G(1, 0)] - F_{(1)(2)} [b_1, b_2; G(1, 0)] \\ - F_{(1)(3)} [b_1, b_3; G(1, 0)] + F_{(1)(2)(3)} [b_1, b_2, b_3; G(1, 0)] \\ = F_2(b_1) - F_1(b_2) F_2(b_1) + F_3(b_1) \\ - F_1(b_2) F_3(b_1) - F_2(b_2) F_3(b_1) \\ + F_1(b_3) F_2(b_2) F_3(b_1) + F_1(b_2) F_2(b_3) F_3(b_1) - F_1(b_3) F_2(b_3) F_3(b_1) \\ - F_2(b_1) F_3(b_2) + F_1(b_3) F_2(b_1) F_3(b_2) \\ + F_1(b_2) F_2(b_1) F_3(b_3) - F_1(b_3) F_2(b_1) F_3(b_3), \quad (19)$$

$$\Pr \{ (K=1) \cap (J=1) \} \\ = F_{(1)} [b_1; G(1, 1)] - F_{(1)(2)} [b_1, b_2; G(1, 1)] \\ - F_{(1)(3)} [b_1, b_3; G(1, 1)] + F_{(1)(2)(3)} [b_1, b_2, b_3; G(1, 1)] \\ = F_1(b_1) - F_1(b_1) F_2(b_2) - F_1(b_1) F_3(b_2) \\ + F_1(b_1) F_2(b_3) F_3(b_2) + F_1(b_1) F_2(b_2) F_3(b_3) - F_1(b_1) F_2(b_3) F_3(b_3). \quad (20)$$

We now need to evaluate F_1 , F_2 , and F_3 . The cumulative distribution functions of the p-values depend on the choice of hypothesis tests. Here, we test each hypothesis with a two sided one sample z test. Suppose that each test decides between the same null hypothesis $H_0 : \mu = \mu_0$, and the same alternative hypotheses $H_1 : \mu = \mu_1$. For convenience, assume that

the sample size for each hypothesis test is the same, N . For $\theta = 1, 2, \dots, N$, suppose $\epsilon_{i\theta}$ is an observed data point for hypothesis test i . If

$$\bar{\epsilon}_i = N^{-1} \sum_{\theta=1}^N \epsilon_{i\theta}, \quad (21)$$

the test statistics are given by

$$Z_i = \left(\frac{\sigma}{\sqrt{N}} \right)^{-1} (\bar{\epsilon}_i - \mu_0), \quad (22)$$

and the two sided p-values are (Rosner, 2006, p. 244)

$$P_i = \begin{cases} 2\Phi(Z_i) & Z_i \leq 0 \\ 2[1 - \Phi(Z_i)] & Z_i > 0 \end{cases}. \quad (23)$$

Then

$$F_1(p) = p,$$

and

$$\begin{aligned} F_2(p) &= F_3(p) \\ &= \Phi \left[\Phi^{-1} \left(\frac{p}{2} \right) + \frac{\mu_0 - \mu_1}{\sigma / \sqrt{N}} \right] + 1 - \Phi \left[\Phi^{-1} \left(1 - \frac{p}{2} \right) + \frac{\mu_0 - \mu_1}{\sigma / \sqrt{N}} \right], \end{aligned} \quad (24)$$

where Φ is the cumulative distribution function for the standard normal distribution. Equation (24) is a special case of the result given in Ruppert et al. (2007, Equation (5)).

4 Computational results

We numerically examine our exact method by comparison to a simulation. We conducted m independent two sided one sample z tests. As in Section 3, each test decided between the same null hypothesis $H_0 : \mu = \mu_0$, and the same alternative hypotheses $H_1 : \mu = \mu_1$. The null hypothesis holds for the first n hypothesis tests, and the alternative hypothesis holds for the remaining $m - n$ hypothesis tests. For each hypothesis test, we simulated N random variables, formed the sample mean as in (22), and calculated the p-values as in (23). Using the sorted p-values, the BH procedure gave the number of rejections. The process was repeated 100,000 times, and gave empirical estimates of the joint probability distribution of K and J . A subset of the results are shown in Table 1.

The largest difference between the exact and simulated values was < 0.0014 . The half width of a 95% confidence interval for the true value is at most $2[(0.5)^2/100,000]^{0.5} = 0.003$, showing that the difference was within the sampling error. This affirms the result in (4).

The exact average power was calculated from the exact joint distribution of K and J , using (14). The simulated average power was calculated similarly, but using the empirical values of the joint distribution of K and J . The results are shown in Table 2. The error in the

simulated power shown in Table 2 is slightly larger than the error in the simulated probabilities in Table 1, as the error compounds over the sum of different values of k and j .

Finally, we calculated and simulated average power for a range of values of μ_1 . The resulting average power curves are shown in Figure 1. Because the simulated values were so close to the exact values, the lines overlapped. For clarity, only the exact average power curves are shown.

5 Application to mammography

In power analysis for single hypothesis studies, the goal is to calculate the power for a given sample size and design. This is quite different from data analysis, where we observe data, construct a test statistic, test a hypothesis, and draw conclusions. The BH procedure is often used to perform data analysis for studies with multiple comparisons.

A problem that arises often in the clinical trial literature is that many scientists conduct trials to test the same hypothesis. These trials are independent, in the sense that they have different subjects, different investigators, and occur in different geographical locations. Using the results of multiple independent trials to analyze scientific questions is meta-analysis.

Consider three clinical trials comparing full field digital and screen-film mammography (Lewin et al., 2002; Skaane et al., 2003; Pisano et al., 2005). The purpose of each of these trials is to contrast the relative diagnostic accuracy of digital and film mammography. The trials share a similar design. Each woman had mammographic examinations with both digital and film machines. Although the trials were not designed to measure the effect of conducting two mammograms on one woman, in each one of the trials, more cancers were detected than would have been by either mammographic modality alone. This suggests that full-field digital mammography and screen film mammography could be used together to improve breast cancer detection.

In order to examine this question, Glueck et al. (2007) conducted a re-analysis of the trial performed by Lewin et al. (2002). A combined method was the test that declared a woman to have cancer if the cancer was found by either (or both) film and digital mammography. The combined method resulted in the detection of significantly more cancers than the film mammographic method alone (Glueck et al., 2007).

This is an interesting result, but it immediately raises the question as to whether the same result holds true for the other two trials with similar designs, namely those of Skaane et al. (2003) and Pisano et al. (2005). One way to decide whether the combined mammographic method is better than the film mammographic method alone is to test this hypothesis in all three studies, and to make a decision based on the result of all three tests.

Benjamini and Hochberg (1997) discuss two possible ways to draw inference from multiple hypothesis tests. One possibility is to use an intersection hypothesis test, in which one considers a grand null which is true only if the null hypothesis is true for every component test. Another option is to use the multiple hypothesis test, where the goal is to provide multiple inferences about the same hypothesis, while controlling the error rate.

Why is control of the false discovery rate important? With classical, uncorrected single hypothesis testing methods, and three independent hypothesis tests, each with a Type I error rate of 0.05, the probability of making at least one Type I error when all three null hypotheses hold is $1 - (0.95^3) \approx 0.14$. Thus, even for only three hypotheses, the failure to correct for multiple comparisons sharply raises the probability of making at least one Type I error. The probability of making at least one Type I error is the family-wise error rate, $\Pr(J/$

$m > 0$). While the false discovery rate is less than or equal to the family-wise error rate, controlling the false discovery rate does provide some error control.

For this proposed meta-analysis, we use the BH procedure to control the false discovery rate for the multiple hypothesis approach. Recall that the BH procedure guarantees that the false discovery rate, the expected value of the ratio of incorrect rejections J to the total rejections K , is bounded, so that $E(J/K) < \alpha^*$. We calculate the average power using the methods introduced in this paper.

To calculate the expected power for the BH procedure using our method, we need to specify m , the number of hypotheses being tested, the null hypothesis for each test, the alternative hypothesis for each test, the distribution of the p-value under the null and the alternatives, and n , the number of hypotheses for which the null holds. In the paragraphs that follow, we define or derive each one of the inputs for the power analysis.

In each study, cancer was diagnosed by the film mammographic method, or by the combined mammographic method. This gave a data table for each study, organized as in Table 3. By the definition of the combined method, the film method can never detect cancers that the combined method does not, so $n_{i,10} = 0$. The corresponding population probabilities are given in Table 4. In a typical power analysis, although the sample size would be assumed to be known, the data would not yet be observed.

The null hypotheses $i = 1, 2, 3$, for the studies of Lewin et al. (2002), Skaane et al. (2003) and Pisano et al. (2005) respectively, are $H_{0i} : \pi_{i,10} - \pi_{i,01} = 0$. The alternatives are $H_{1i} : \pi_{i,10} - \pi_{i,01} = \delta \neq 0$. The value of δ is a parameter will vary as part of the power analysis.

From previous studies that looked at combining information from two mammographic methods (see, e.g., Sickles et al., 1986; Anttinen et al., 1993; Thurfjell et al., 1994; Ciatto et al., 2003) it is reasonable to believe that the combined test is always better than the film mammographic method, so we set $n = 0$ for the power analysis: we expect that none of the null hypotheses will be true.

We use an unconditional McNemar's test (McNemar, 1947), a common test for paired binary outcome data. There has been great controversy in the statistical literature as to whether to use the conditional or unconditional test (Selicato and Muller, 1998). We use the unconditional test for this application because it is reasonable to believe that neither margin is fixed.

There are many equivalent transformations of the test statistic for the unconditional test. Depending on the form of the statistic chosen, the distribution of the test statistic under the null and the alternative differs. As suggested by Selicato and Muller (1998), we choose McNemar's test statistic computed as an F statistic,

$$F_i = \frac{\bar{D}_i^2}{S_i^2 / (N - 1)}, \quad (25)$$

where

$$\bar{D} = \frac{(n_{i,10} - n_{i,01})}{N_i} \quad (26)$$

and

$$S_i^2 = [(n_{i,10} - n_{i,01}) / N_i] - \bar{D}_i^2. \quad (27)$$

When conducting a single test, and the null hypothesis is true, F_i has approximately a central F distribution, with degrees of freedom 1 and $N_i - 1$. The distribution is approximate, because the tables, and hence the test statistic are discrete. The accuracy of the approximation is adequate, as demonstrated by Selicato and Muller (1998). When the alternative hypothesis is true, F_i is approximately distributed as the non-central F distribution with the same degrees of freedom, and noncentrality parameter $N_i \omega_i$, where

$$\omega_i = \frac{(\pi_{i,10} - \pi_{i,01})^2}{(\pi_{i,10} + \pi_{i,01}) - (\pi_{i,10} - \pi_{i,01})^2}. \quad (28)$$

Thus, the p-value (a random variable) for study i is given by

$$P_i = 1 - F_F(F_i; 1, N_i - 1) \quad (29)$$

where $F_F(x; 1, N_i - 1)$ is the cumulative distribution function of the central F distribution with degrees of freedom 1 and $N_i - 1$, evaluated at a point x . The cumulative distribution function of the p-value under the alternative for study i is approximated by

$$F_i(p) = 1 - F_F[F_F^{-1}(1 - p; 1, N_i - 1); N_i - 1, N_i \omega_i], \quad 0 < p < 1.$$

which is again a special case of the result of Ruppert et al. (2007, Equation (5)). If the null hypothesis is true for study i , $\omega_i = 0$, $F_i(p) = p$, and the p-value P_i has approximately a uniform $[0, 1]$ distribution.

We choose the bound on the false discovery rate $\alpha^* = 0.05$. The number of cancers observed in the Lewin et al. (2002), Skaane et al. (2003) and Pisano et al. (2005) trials were $N_1 = 49$, $N_2 = 31$ and $N_3 = 335$, respectively. Let $F_i = F_{P_i}(p; N_i, \pi_{10}, \pi_{01})$. We can obtain the average power using (14). While (14) gives exact average power, the results for this power analysis will be approximate, since the distribution of the test statistic given in Selicato and Muller (1998) is approximate. When $\delta = 0.2$, the average power is approximately 0.9. When $\delta = 0.3$, the average power is approximately .98. Average power was calculated for those values of δ because the observed difference in probability of cancer detection between the combined mammographic method and the film mammographic method demonstrated in Glueck et al. (2007) was close to 0.2.

The interpretation of average power for the BH procedure is somewhat different from the interpretation of power for a single hypothesis test. For a single hypothesis test, power is the probability of rejecting the null, for a given design, sample size, and alternative. For the BH procedure, average power is the fraction of hypothesis tests for which the alternative is true that are rejected. When no nulls are true, as in this example, average power is simply the fraction of hypotheses that should be rejected that are rejected. As is true for power analysis for single hypotheses, the decision to choose a goal average power for the study depends more on scientific considerations than on statistical ones. In this case, the average power seems adequate to detect a clinically important difference.

6 Conclusions

The method we have proposed in this paper produces exact calculations of average power. We have demonstrated its utility for an example with one sample z tests, and for a meta-analysis using McNemar's test. The method will provide results no matter the distribution of the test statistic, and even for studies which use different statistics to test each hypothesis. It should prove useful for calculations of power for study designs where accuracy is important, such as clinical trials and their meta-analyses.

One major drawback of our technique is the computational complexity. The complexity depends on the number of hypotheses being considered. Equation (5) has complexity of order 2^{m-k} , and the evaluation of the joint probability in (12) involves a number of computations of order $m!$. Thus, our results are easily computable only for the small number of hypotheses likely to be seen in clinical trials. Genomic problems are beyond our reach, with the exact methods described here.

We believe that our methods are important for many reasons. Our formula for average power is an exact result for any number of hypotheses, and is computable for situations with small numbers of hypotheses. Exact results provide a sound theoretical basis for deriving high accuracy, easily computable approximations. Not only do our results suggest approximations, but they will allow testing the approximations by comparison to exact results. In addition, understanding average power for the BH procedure immediately opens windows to understanding average power for other multiple comparisons procedures. Exact power calculations will allow the comparative study of the sensitivity and specificity of multiple comparison procedures.

A Proof of (5)

From (4),

$$\begin{aligned} \{K=k\} \cap G(k, j) \\ = \underbrace{G(k, j) \cap \{P_{(k)} \leq b_k\}}_B \cap \{P_{(k+1)} \geq b_{k+1}\} \cap \cdots \cap \{P_{(m)} \geq b_m\}. \end{aligned} \quad (30)$$

We will change the direction of the inequalities in (4) for $k+1, \dots, m$ by passing to the set complement. Using standard set operations and the fact that the cumulative distribution functions are continuous, we have from (4),

$$\begin{aligned} \Pr(\{K=k\} \cap G(k, j)) &= \Pr\left(\bigcap_{q=k+1}^m (B \cap \{P_{(q)} \geq b_q\})\right) \\ &= \Pr\left(B \setminus \bigcup_{q=k+1}^m (B \cap \{P_{(q)} < b_q\})\right) \\ &= \Pr(B) - \Pr\left(\bigcup_{q=k+1}^m (B \cap \{P_{(q)} \leq b_q\})\right) \end{aligned} \quad (31)$$

Expand the last term as

$$\begin{aligned}
& \Pr \left(\bigcup_{q=k+1}^m (B \cap \{P_{(q)} \leq b_q\}) \right) \\
&= \sum_{q=k+1}^m \Pr(B \cap \{P_{(q)} \leq b_q\}) \\
&- \sum_{\substack{r,s=k+1 \\ r \neq s}}^m \Pr(B \cap \{P_{(r)} \leq b_r\} \cap \{P_{(s)} \leq b_s\}) + \dots \\
&+ (-1)^{m-k+1} \Pr(B \cap \{P_{(k+1)} \leq b_{k+1}\} \cap \dots \cap \{P_{(m)} \leq b_m\}),
\end{aligned} \tag{32}$$

by adding the probabilities of the sets in the union, subtracting probabilities of the intersections of two sets, which were counted twice, adding back the probabilities of the intersections of three sets, and so on. Substituting for B from (30) and using the definition (6) gives (5).

B Proof of (12)

Recall that P_i , $1 \leq i \leq m$, are the unsorted p-values, and $P_{(i)}$ are the sorted p-values. Denote $A = F_{(n_1) \dots (n_e)} [p_{n_1}, \dots, p_{n_e}; G(j, k)]$. Then we have

$$\begin{aligned}
A &= \Pr \left(\bigcap_{h=1}^e \{P_{(n_h)} \leq p_{n_h}\} \cap G(j, k) \right) \\
&= \Pr \left(\bigcap_{h=1}^e \{\text{at least } n_h \text{ of the } P_i \leq p_{n_h}\} \cap G(j, k) \right) \\
&= \Pr \left(\bigcup_{i \in I} \bigcap_{h=1}^e \{\text{exactly } n_h \text{ of the } P_i \leq p_{n_h}\} \cap G(j, k) \right) \\
&= \Pr \left(\bigcup_{i \in I} B_i \right) = \sum_{i \in I} \Pr(B_i),
\end{aligned} \tag{33}$$

where B_i are the disjoint events

$$\begin{aligned}
B_i &= \{P_i \in [p_{n_0}, p_{n_1}] \text{ for exactly } j \text{ indices } i=1, \dots, n\} \\
&\cap \bigcap_{h=0}^e \{P_i \in (p_{n_h}, p_{n_{h+1}}) \text{ for exactly } i_{h+1} - i_h \text{ indices } i=1, \dots, m\}.
\end{aligned} \tag{34}$$

Since F_i are continuous, $\Pr \{P_i \in (p_{n_h}, p_{n_{h+1}}]\} = F_i(p_{n_{h+1}}) - F_i(p_{n_h})$, the random variables P_i have values in the disjoint union

$$P_i \in [0, 1] = \bigcup_{h=0}^e (p_{n_h}, p_{n_{h+1}}], \tag{35}$$

and P_i are independent, (12) now follows from elementary probability. The condition (8) means that μ_{ih} acts as indicator for the presence of the term $F_i(p_{n_{h+1}}) - F_i(p_{n_h})$. Equation (9) describes the conditions on $P_i \in [p_{n_0}, p_{n_1}]$ and (10) describes the conditions on $P_i \in [p_{n_h}, p_{n_{h+1}}]$, $h \geq 1$. Finally, (11) comes from the fact that the union (35) is disjoint.

References

Anttinen I, Pamilo M, Soiva M. Double reading of mammography screening films—one radiologist or two? *Clinical Radiology* 1993;48:414–421. [PubMed: 8293648]

- Bapat RB, Beg MI. Order statistics for non-identically distributed variables and permanents. *Sankhyā Ser. A* 1989;51:79–93.
- Benjamini Y, Liu W. A Step-Down Multiple Hypotheses Testing Procedure That Controls the False Discovery Rate Under Independence. *Journal of Statistical Planning and Inference* 1999;82:163–170.
- Benjamini, Yoav; Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1995;57(1):289–300.
- Benjamini, Yoav; Hochberg, Yosef. Multiple Hypotheses Testing with Weights. *Scandinavian Journal of Statistics* 1997;24(3):407–418.
- Benjamini, Yoav; Yekutieli, Daniel. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 2001;29(4):1165–1188.
- Ciatto S, Rosselli Del Turco M, Burke C, Visioli P, Paci E, Zappa M. Comparison of standard and double reading and computer-aided detection (CAD) of interval cancers at prior negative screening mammograms: blind review. *British Journal of Cancer* 2003;89:1645–1649. [PubMed: 14583763]
- Curran-Everett D. Multiple Comparisons: Philosophies and Illustrations. *American Journal of Physiology. Regulatory Integrative and Comparative Physiology* 2000;279:R1–R8.
- Efron B, Storey J, Tibshirani R. Microarrays, Empirical Bayes Methods, and False Discovery Rates. *Journal of the American Statistical Association* 2001;96:1151–1160.
- Ferreira, José A.; Zwinderman, Aeilko. Approximate Sample Size Calculations with Microarray Data: an Illustration. *Statistical Applications in Genetics and Molecular Biology* 2006;5 Article 25.
- Finner H, Roters M. Multiple Hypothesis Testing and Expected Number of Type I Errors. *The Annals of Statistics* 2002;30(1):220–238.
- Genovese CR, Wasserman L. Operating Characteristics and Extensions of the False Discovery Rate Procedure. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2002;64:499–517.
- Genovese CR, Wasserman L. A Stochastic Process Approach to False Discovery Control. *The Annals of Statistics* 2004;32(3):1035–1061.
- Glueck DH, Lamb MM, Lewin JM, Pisano ED. Two-modality mammography may confer an advantage over either full-field digital mammography or screen-film mammography. *Academic Radiology* 2007;14(6):670–676. [PubMed: 17502256]
- Glueck, Deborah H.; Karimpour-Fard, A.; Mandel, J.; Hunter, L.; Muller, KE. Fast computation by block permanents of cumulative distribution functions of order statistics from several populations. *Communications in Statistics: Theory and Methods* 37:18.
- Glueck, Deborah H.; Muller, Keith E.; Karimpour-Fard, Anis; Hunter, Lawrence. Expected Power for the False Discovery Rate with Independence. *Communications in Statistics: Theory and Methods* 2008b;37(12):1855–1866. [PubMed: 20975846]
- Jung, Sin-Ho. Sample size for FDR-control in microarray data analysis. *Bioinformatics* 2005;21(14):3097–3104. [PubMed: 15845654]
- Keselman HJ, Cribbie R, Holland B. Controlling the Rate of Type I Error Over a Large Set of Statistical Tests. *British Journal of Mathematical and Statistical Psychology* 2002;55:27–39. [PubMed: 12034010]
- Lee M, Whitmore G. Power and Sample Size for DNA Microarray Studies. *Statistics in Medicine* 2002;21:3543–3570. [PubMed: 12436455]
- Lewin, John M.; D’Orsi, Carl J.; Edward Hendrick, R.; Moss, Lawrence J.; Isaacs, Pamela K.; Karellas, Andrew; Cutter, Gary R. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. *American Journal of Roentgenology* 2002;179:671–7. [PubMed: 12185042]
- McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12:153–157. [PubMed: 20254758]
- Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D’Orsi C, Jong R, Rebner M. Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine* 2005;353:1773–1783. [PubMed: 16169887]

- Rosner, B. Fundamentals of Biostatistics. 6th edition. Brooks-Cole; New York: 2006.
- Ruppert D, Nettleton D, Hwang JTG. Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics* 2007;63:483–495. [PubMed: 17715492]
- Sarkar SK. Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures. *The Annals of Statistics* 2002;30(1):239–257.
- Sarkar SK. FDR-Controlling Stepwise Procedures and Their False Negatives Rates. *Journal of Statistical Planning and Inference* 2004;125:119–137.
- Sarkar SK. False Discovery and False Nondiscovery rates in Single-Step Multiple Testing Procedures. *The Annals of Statistics* 2006;34(1):394–415.
- Selicato GR, Muller KE. Approximating power of the unconditional test for correlated binary pairs. *Communications in Statistics: Simulation and Computation* 1998;27:553–564.
- Sickles EA, Weber WN, Galvin HB, Ominsky SH, Sollitto RA. Baseline screening mammography: one vs. two views per breast. *American Journal of Roentgenology* 1986;147:1149–1153. [PubMed: 3490749]
- Skaane P, Young K, Skjennald A. Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading–Oslo I study. *Radiology* 2003;229:877–884. [PubMed: 14576447]
- Storey JD. Direct Approach to the False Discovery Rate. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2002;64(3):479–498.
- Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. *Radiology* 1994;191:241–244. [PubMed: 8134580]

Figure 1.

Average power for the Benjamini and Hochberg (1995) procedure and three independent two sided one sample z tests, with one true null hypothesis, so $m = 3$ and $n = 1$. The null hypothesis for each test is $H_0 : \mu = \mu_0$, and the alternative hypotheses for each test is $H_1 : \mu = \mu_1$. Here $\mu_0 = 0$, $\sigma^2 = 1$, and μ_1 was varied between 0 and 2. The false discovery rate is less than $\alpha^* = 0.05$. The average power is shown as a function of the sample size used for each hypothesis test, N . Average power is undefined when $\mu_1 = 0$, and all hypotheses are null. Thus, the average power curve is undefined at that point. Notice that the average power is smaller for $N = 3$, than for $N = 5$, or $N = 7$.

Table 1

Comparison of calculations of $\Pr(\{K = k\} \cap \{J = j\})$ via simulation and theory. The sample size per hypothesis was $N = 5$, with $\alpha_* = 0.05$, $m = 5$, $n = 3$, $\mu_0 = 0$, $\mu_1 = 1$, and $\sigma^2 = 1$

k	j	Exact	Simulation
0	0	0.37584	0.37718
1	0	0.36883	0.36838
	1	0.00812	0.00788
2	0	0.19646	0.19515
	1	0.02476	0.02540
	2	0.00026	0.00027
3	1	0.02297	0.02286
	2	0.00116	0.00122
4	2	0.00149	0.00157
	3	0.00003	0.00001
5	3	0.00005	0.00008

Table 2

Comparison of calculations of average power via simulation and exact theory. The sample size per hypothesis was $N = 5$, with $\alpha_* = 0.05$, $\mu_0 = 0$, $\mu_1 = 1$, and $\sigma^2 = 1$. The largest difference between the exact and simulated values was < 0.002

<i>m</i>	<i>n</i>	Average Power	
		Exact	Simulated
2	0	0.56539	0.56371
2	1	0.50342	0.50455
3	0	0.54576	0.54466
3	1	0.49842	0.49813
3	2	0.44439	0.44557
4	0	0.53446	0.53352
4	1	0.49584	0.49494
4	2	0.45256	0.45322
4	3	0.40451	0.40684
5	0	0.52712	0.52650
5	1	0.49440	0.49619
5	2	0.45819	0.45931
5	3	0.41837	0.41717
5	4	0.37494	0.37439

Table 3

Observed number of cancers detected by each mammographic technique for study i . Combined is the combined mammographic method, while film is the film mammographic method. N_i is the total number of cancers detected in study i . Here 0 indicates that the method missed the cancer, while 1 indicates that the method found the cancer. Thus $n_{i,00}$ is the number of cancers missed by both methods. Also, $n_{i,0\cdot} = n_{i,00} + n_{i,01}$, $n_{i,1\cdot} = n_{i,10} + n_{i,11}$, $n_{i,\cdot 0} = n_{i,00} + n_{i,10}$, and $n_{i,\cdot 1} = n_{i,01} + n_{i,11}$

		Combined		
		0	1	
Film	0	$n_{i,00}$	$n_{i,01}$	$n_{i,0\cdot}$
	1	$n_{i,10}$	$n_{i,11}$	$n_{i,1\cdot}$
		$n_{i,\cdot 0}$	$n_{i,\cdot 1}$	N_i

Table 4

Population probabilities for study i . Combined is the combined mammographic method, while film is the film mammographic method. $\pi_{i,00}$ is the probability that both methods missed the cancer

		Combined		
		0	1	
Film	0	$\pi_{i,00}$	$\pi_{i,01}$	$\pi_{i,0.}$
	1	$\pi_{i,10}$	$\pi_{i,11}$	$\pi_{i,1.}$
		$\pi_{i,.0}$	$\pi_{i,.1}$	