Statistica Sinica **15**(2005), 841-855

ON THE OPTIMUM NUMBER OF HYPOTHESES TO TEST WHEN THE NUMBER OF OBSERVATIONS IS LIMITED

Andreas Futschik and Martin Posch

University of Vienna and Medical University of Vienna

Abstract: We investigate the problem of deciding on the number of hypotheses to be considered in a multiple hypothesis testing framework when the overall number of observations that can be collected is limited. A natural question in this context is whether the number of hypotheses to be tested should be limited in favor of additional observations per considered hypothesis. We provide guidelines concerning the choice of an optimum number of considered hypotheses in common testing situations. The optimization is with respect to the expected number of correct rejections in the hypothesis testing context. We also briefly discuss the classification setting, where a linear combination of true and false positives is considered. The overall number of observations may be limited for several reasons, such as the number of patients or the amount of probe material available. We demonstrate that considering an appropriate number of hypotheses in this context can lead to a substantial increase in the expected number of correct rejections.

Key words and phrases: Bonferroni rule, classification, Dunnett test, false discovery rate, multiple hypotheses testing.

1. Introduction

One of the goals in statistics is to extract as much information as possible from a limited number of observations. In the context of multiple hypothesis testing, extraction of information is usually considered equivalent to correct rejections of null hypotheses while ensuring some global control of the type I error. Therefore a lot of effort has gone into the development of procedures that permit one to reject as many hypotheses as possible and to control criteria like the familywise error or the false discovery rate. Of course the performance of any such procedure will crucially depend on the number of observations available per hypothesis test. Due to practical constraints the number of observations available is usually limited, leading to some bound m on the total number of observations cumulative over all individual hypothesis tests. Depending on the type of experiment, there may be some choice at the design stage concerning how many hypotheses will be tested and how the overall m observations will be allocated to the individual hypothesis tests.

Multiple test problems where the overall number of observations is limited occur in several situations. Consider for instance a clinical trial involving several possible treatments of potential interest. Here the issue is how many treatment groups to consider, given a limited total number of patients. In agriculture, the total area of the experimental field is usually fixed and a decision has to be made on the number of competing plant varieties to test. If too many varieties are tested against some standard (say), the individual tests will suffer from poor power possibly causing good varieties to remain undetected. On the other hand, when investigating too few varieties, some promising varieties may not even be tested. A similar issue arises in the context of gene expression data, where a frequent goal is to test for differential expression of a large number of genes represented by spots on a microarray. The amount of available probe material (for instance taken from human cancer tissues) to be applied on a microarray is often small. Limiting the number of different genes represented on a microarray permits a larger number of observations for the represented genes. In the context of Stochastic Discrete Event Systems, simulations are often carried out for an extremely large number of combinations of input variables influencing a stochastic event (Rubinstein and Shapiro (1993) and Rubinstein and Melamed (1998)). Such systems arise in a variety of engineering contexts, including manufacturing systems, communication networks, computer systems, logistics, and vehicular traffic. Due to constraints on the computationally feasible total number m of simulation runs, there is a trade off between the number of input combinations considered and the number of simulation runs per input combination.

A worked out example illustrating the practical relevance of our issue in the medical context is given in Section 7.

The question is how many hypotheses should be tested, if it is desired to reject as many incorrect null hypotheses as possible. Suppose there is a total of K hypothesis pairs available for testing, containing a certain proportion of incorrect null hypotheses. Our goal is to investigate the optimum number k to pick out of the K hypotheses, in order to maximize the expected number of rejections of incorrect null hypotheses. Other objective functions, like the probability of at least one correct rejection, might make sense in some situations but we focus on the expected number of correct rejections. We give examples illustrating that sometimes considerably more can be gained in terms of possible rejections of incorrect null hypotheses by choosing an appropriate number k, than by choosing a more sophisticated multiple testing procedure.

More formally, suppose that K hypothesis pairs $H_{0,i}$ versus $H_{1,i}$ $(1 \leq i \leq K)$ are available for testing. We investigate the choice of an optimum k at the design stage, assuming that k hypotheses are picked from some set of possible hypothesis pairs, leading to approximately m/k observations for each test. We assume that the chance of picking a pair where the alternative holds is equal to

some constant q. It will turn out that often the optimal number of hypotheses to pick does not depend on q. Subsequently we focus on z- and t-tests based on $N(\theta, \sigma^2)$ distributed observations and consider the basic one-sample, one-sided multiple testing situation

$$H_{0,i} : \theta_i = 0, \qquad \text{against } H_{1,i} : \theta_i > 0, \quad 1 \le i \le K,$$

but we also discuss multiple comparisons with a control, two-sided problems as well as classical two-sample problems.

We maximize the expected number of correct rejections $\mathbf{E}N_k$ in k, and obtain both asymptotic results for $m \to \infty$, assuming an unlimited supply of hypotheses, and results for fixed m. The optimization is under the assumption that for each hypothesis pair either the null hypothesis is true or a fixed reference alternative $\theta_{(a)}$ holds. The extension to composite alternatives is also discussed.

In Section 2, we consider the optimization of the expected number of possible rejections for the Bonferroni multiple test procedure, and discuss the extension to the Dunnett and Bonferroni−Holm test. In Section 3, we cover optimization for the Benjamini−Hochberg procedure that controls the false discovery rate. In Sections 4 and 5, we discuss the extension of our results for the z-test to t-tests and to composite alternatives. Finally, the optimization problem is addressed for classification problems in Section 6. All proofs are given in the Appendix.

2. Controlling the Familywise Error Rate

The familywise error is defined as the probability of rejecting at least one null hypothesis erroneously. There are several multiple testing procedures that control the familywise error in the strong sense, i.e., regardless of how many null hypotheses are true. The most prominent is the classical Bonferroni rule that implies that the familywise error will stay below α if the individual tests are carried out at level α/k. This approach is easy to implement but in some situations a considerable amount of power can be gained by using more sophisticated multiple testing rules. For example, if a large fraction of hypotheses is false, then the Bonferroni−Holm procedure will usually lead to an increase in the number of possible rejections. In the following, we first concentrate on the Bonferroni procedure and on normally distributed one-sample test statistics T_i. The extensions to the Bonferroni−Holm and the Dunnett procedure, as well as two-sample and two-sided tests, are discussed in Subsection 2.2.

2.1. Bonferroni tests

With $\Phi_{(\mu,\sigma^2)}$ denoting the normal $N(\mu, \sigma^2)$ c.d.f., z_γ denoting the $1 - \gamma$ standard normal quantile, and for

$$\Delta_m := \theta_{(a)} \frac{\sqrt{m}}{\sigma}, \tag{1}$$

the expected number of correctly rejected null hypotheses with the Bonferroni procedure and z-tests is

$$E(N_k) = q\,k\,\left(1 - \Phi_{(\Delta_m/\sqrt{k},1)}(z_{\alpha/k})\right). \tag{2}$$

Notice that independence of the test statistics for different hypotheses is not required for (2) to hold. A plot of $E(N_k)$ in k (see Figure 1) shows that choosing an inadequate number of hypotheses results in a substantial loss in the expected number of rejections. To account for cases where the overall number of observations m could not be split evenly, we used $\lfloor m/k \rfloor + 1$ observations for the first $m \bmod k$ hypotheses and $\lfloor m/k \rfloor$ for the rest in the displayed simulation results. Here $\lfloor x \rfloor$ denotes the largest integer not exceeding x. The maximizer k_m^* of (2) is obviously independent of q and depends only on α and Δ_m. Consequently, $E(N_{k_m^*})$ increases linearly in q. Table 1 gives the optimal k_m^* for typical parameter choices obtained by numerical optimization with the R language (Ihaka and Gentleman (1996)). We now provide an asymptotic approximation to k_m^* optimizing (2), for $m \to \infty$.

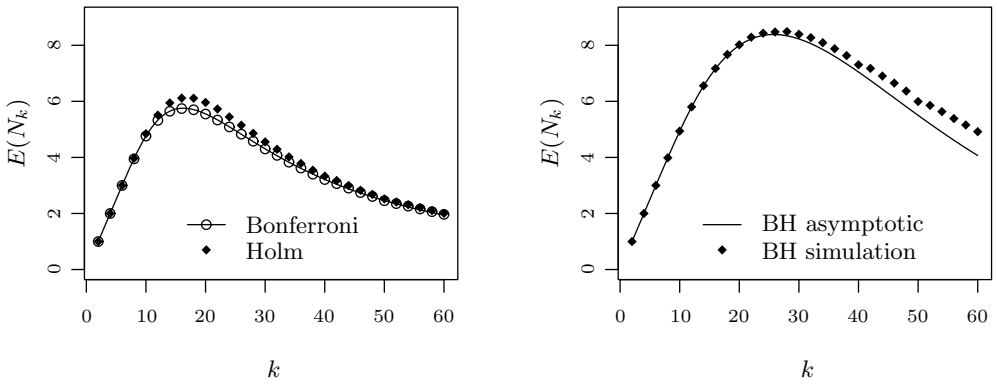

Figure 1. The expected number of correctly rejected null hypotheses for given k with the Bonferroni procedure and the Bonferroni-Holm procedure (left graph), as well as the Benjamini-Hochberg procedure (right graph). The dots give simulation results (all standard errors below 0.012) and the solid lines the theoretical solution for the Bonferroni test (left graph), and the asymptotic solution for the Benjamini-Hochberg procedure (right graph). The parameters are $\theta = 1$, $m = 200$, $\alpha = 0.025$, and $q = 0.5$.

Theorem 1. *Take* $k_m := \Delta_m^2/(2\log(\Delta_m^2))$. *The optimum number of hypotheses to test is* $k_m^* = k_m[1 - c_m(\alpha)]$, *where* $c_m(\alpha) \to 0$, *and* $[\log m]^{1/2} c_m(\alpha) \to \infty$, *as* $m \to \infty$.

For small and moderate m, Theorem 1 gives only a rough estimate of k_m^*, as convergence is slow. However, we see that asymptotically the sample size for

each individual hypothesis, m/k_m^*, tends to infinity at the rate $\log(m)$ as $m \to \infty$, when choosing k_m^* hypotheses. As a consequence, the power for the individual tests at the considered alternative $\theta_{(a)}$ converges to one and the expected number of rejections is $E(N_{k_m^*}) = q\, k_m(1 + o(1))$.

Table 1. The optimum number of hypotheses k_m^* and the power (in %) to reject an individual incorrect null hypotheses. The expected number of correctly rejected null hypotheses is then given by the product of q, k_m^* and the power.

		Δ_m					
		5	10	20	50	100	1000
α	0.01	3 (57)	8 (70)	25 (74)	124 (76)	425 (78)	28908 (82)
	0.025	3 (69)	9 (71)	29 (72)	138 (75)	469 (77)	30883 (81)
	0.05	4 (60)	11 (66)	33 (70)	152 (74)	508 (76)	32564 (81)

2.2. Other multiple testing procedures

The Bonferroni-Holm Procedure. Consider the hypotheses $H_{0,i}$, $i = 1, \ldots, k$, and let p_i denote the p-value for the ith hypothesis, and $p_{[i:k]}$ the ith smallest p-value. Let $i^* = \max\{i \,|p_{[l:k]} < \alpha/(k - l + 1) \text{ for all } l \leq i\}$. Then the Bonferroni−Holm rule rejects all hypotheses i, $1 \leq i \leq k$, such that $p_i < \alpha/(k - i^* + 1)$.

Unlike for Bonferroni tests, the optimum k for the Bonferroni−Holm procedure depends on the probability q of picking hypotheses pairs where the null hypothesis fails to be true. Since q is usually unknown, a natural question is how well the optimum k_m for the Bonferroni rule works as an approximation. Our simulations for several values of q, m and $\theta_{(a)}$ showed a close similarity of the objective functions for both rules, even for q as large as 0.5. See Figure 1 for an example. The figure also suggests that the number of possible rejections can decrease considerably more by an inappropriate choice of the number of hypotheses k than by the use of a less sophisticated multiple testing procedure.

Subsequently we state heuristic bounds for the optimum k that are derived from the Bonferroni procedure. Since the Bonferroni−Holm procedure leads always to at least as many rejections as the Bonferroni procedure, the expected number of correct rejections $E(N_k)$ for the Bonferroni procedure gives a lower bound L_k for the rejections for Bonferroni−Holm rule. A heuristic upper bound can be obtained as follows. If the proportion $(1 - q)$ of correct null hypotheses were known, the Bonferroni procedure could be applied with the level of significance $\alpha/(k(1 - q))$, while still controlling the familywise error. This would lead to a number of possible rejections that is not smaller than that obtained via

the Bonferroni-Holm procedure, except for some rare cases involving erronous rejection of true null hypotheses by the Bonferroni−Holm rule. Neglecting the possibility of data leading to such rare cases, this modified Bonferroni procedure provides us with an upper bound U_k for the expected number of possible correct rejections.

An interval containing the optimum number $k_m^{(H)}$ of hypotheses to test with the Bonferroni−Holm procedure is now given by the set of all k, where $U_k \geq \max_j L_j$. Furthermore since $\max_j U_j / \max_j L_j \to 1$ as $m \to \infty$ (according to Theorem 1), any k between the optimum k for the classical Bonferroni rule and that for the Bonferroni procedure with level $\alpha/(k(1-q))$ will lead to a vanishing relative loss in the optimum number of rejections.

The Dunnett Procedure. The Dunnett (1955) procedure is frequently used when comparing several treatments to a common standard. It is based on the standardized differences $\sqrt{\gamma\, m}\,(\bar{X}_i - \bar{X}_0)/[\sqrt{k}\,\sigma(1+\gamma)]$, $i = 1, \ldots, k$, of the sample means. Here σ^2 is the common variance of the observations, and γ is the common ratio n_i/n_0 between treatment and control sample sizes. Under the alternative of a positive difference $\theta_{(a)}$ between the treatment i and the control, the non-centrality parameter is $\tilde{\Delta}_m/\sqrt{k}$ where

$$\tilde{\Delta}_m := \frac{\sqrt{\gamma\, m}\;\theta_{(a)}}{\sigma(1+\gamma)}.$$

The resulting objective function can be obtained from (2) by plugging in the above defined $\tilde{\Delta}_m$ instead of Δ_m, and replacing the quantile $z_{\alpha/k}$ by the corresponding Dunnett multivariate normal critical values $d_{\alpha,k,\gamma}$. See Somerville and Bretz (2001) for a numerical algorithm to obtain critical values. For obtaining the optimum k, direct numerical optimization is an obvious strategy. Notice that recommendations concerning the optimum choice of γ are available in the literature (see e.g., Bechhofer and Tamhane (1983)).

Two sample and/or two sided tests. We discuss here Bonferroni z-tests. There, for two sided tests, the quantile $z_{\alpha/k}$ in (2) needs to replaced by $z_{\alpha/(2\,k)}$. If directional errors are not considered as correct rejections, this is the only necessary modification. For two sample tests with equal group sizes, the parameter Δ_m in (1) changes to $\theta_{(a)}\,\sqrt{m/[2\,(\sigma_1{}^2 + \sigma_2{}^2)]}$, where $\theta_{(a)}$ denotes the parameter difference, and σ_j, $j = 1, 2$ the standard deviations in the two groups.

3. Controlling the False Discovery Rate

An alternative to the control of the familywise error is the Benjamini−Hochberg (BH) procedure that controls the more liberal false discovery rate.

As proposed by Benjamini and Hochberg (1995), the false discovery rate is defined by $E(V/R)$, where R is the number of rejected null hypotheses and V the number of true null hypotheses that are rejected. If $R = 0$, the fraction V/R is defined to be 0. In the case that all tested null hypotheses are true, the false discovery rate is equivalent to the familywise type I error. For the control of the false discovery rate, Benjamini and Hochberg (1995) propose a sequential p-value method (which dates back to Seeger (1968), see Finner and Roters (2001) for a historical treatise). Consider the hypotheses $H_{0,i}$, $i = 1, \ldots, k$ and let p_i denote the p-value for the ith hypothesis, and $p_{[i:k]}$ the ith smallest p-value. If there exists a j such that $p_{[j:k]} < \alpha_j$ with $\alpha_j = \alpha\, j/k$, then all hypotheses i, $1 \le i \le k$ such that $p_i < \alpha_j$ can be rejected. This procedure controls the false discovery rate at α in particular under the assumption that all test statistics are independent (Benjamini and Hochberg (1995)).

An exact computation of the expected number of correctly rejected hypotheses for the BH procedure is very hard. However, the asymptotic approximations provided by Genovese and Wasserman (2002) for independent test statistics and an increasing number of hypotheses are helpful in our setting. Indeed, according to Genovese and Wasserman (2002) and for a single alternative $\theta_{(a)}$, the Benjamini−Hochberg procedure is asymptotically equivalent to a single−step procedure where each hypothesis is tested at level u, and u solves

$$H_{\Delta_m/\sqrt{k}}(u) = \beta u, \tag{3}$$

with $H_{\Delta_m/\sqrt{k}}(\cdot)$ being the distribution of the p-value under the non-centrality parameter $\Delta_m/\sqrt{k}$ and $\beta = (1/\alpha - (1-q))/q$.

Therefore the optimum k for such a single step procedure using the significance level u should provide a guideline for the choice of an optimum k for the Benjamini−Hochberg procedure. We optimize the above single step test procedure in k and focus, as for Bonferroni tests, on normally distributed test statistics. In the one-sided setting our optimization problem is given by

$$E(N_k) = q\,k\,\left(1 - \Phi_{(\Delta_m/\sqrt{k},1)}(z_u)\right) \to \max_k, \tag{4}$$

where u is defined by (3).

Unlike for Bonferroni tests, it follows from Theorem 2 below that the optimum k grows linearly in the number of available observations m and quadratically in Δ_m defined in (1). Therefore the optimum number of observations per hypothesis test m/k_m^* (and also the power to reject an individual null hypothesis) does not depend on m, and does in particular not tend to infinity when $m \to \infty$. The actual optimization requires maximization of a simple objective function.

Theorem 2. *The solution of* (4) *is given by* $k_m^* = \Delta_m^2/(z_{u_\beta^*} - z_{\beta u_\beta^*})^2$, *where* u_β^* *maximizes* $u/(z_u - z_{\beta u})^2$.

It follows from Theorem 2 that $EN_{k_m^*} \to \infty$ for $m \to \infty$, if the optimum number of hypotheses k_m^* is chosen.

Figure 1 shows a typical example of the relationship between the number of hypotheses considered and the expected number of correct rejections. We plotted the asymptotic estimate based on Theorem 2 together with estimates from a simulation study. The asymptotic results give a good approximation even for the chosen moderate m and θ. Table 2 allows for an easy derivation of the optimal number of hypotheses for different values of α and q.

Table 2. The optimum number of hypotheses k_m^* over Δ_m^2 and the power (in %) to reject an individual incorrect null hypotheses (obtained by numerical optimization). The expected number of correctly rejected null hypotheses is then given by the product of q, k_m^* and the power.

		q				
		0.1	0.25	0.5	0.75	0.9
α	0.01	0.070 (73)	0.084 (70)	0.099 (68)	0.110 (67)	0.116 (66)
	0.025	0.084 (70)	0.105 (68)	0.129 (65)	0.148 (63)	0.158 (63)
	0.05	0.100 (68)	0.130 (65)	0.167 (62)	0.198 (60)	0.215 (59)

4. Student's t-Test

Analogous to the normal case, the expected number of rejections based on one-sample Bonferroni t-tests involving k hypotheses is

$$\mathbf{E}N_k^{(t)} = q\,k[1 - F^{(t)}_{m/k-1,\Delta_m/\sqrt{k}}(t_{\alpha/k,m/k-1})], \tag{5}$$

where $F^{(t)}_{\nu,\delta}$ is the cdf of the non-central t-distribution with ν degrees of freedom and noncentrality parameter δ, and $t_{\gamma,\nu}$ denotes the $1-\gamma$ quantile of the standard t-distribution with ν degrees of freedom. Similarly, we obtain, with u defined by (3),

$$\mathbf{E}N_k^{(t)} = q\,k[1 - F^{(t)}_{m/k-1,\Delta_m/\sqrt{k}}(t_{u,m/k-1})] \tag{6}$$

for the Benjamini–Hochberg procedure.

The following result suggests that the optimum k for z-tests should provide a good approximation also for t-tests, involving small effects and a large overall number of observations m. The result assumes a constant sequence $\Delta_m =: \Delta$.

Theorem 3. *Let* $\theta_{(a)} > 0$, *and take* $\theta_m = \theta_{(a)}/\sqrt{m}$. *Assume that* $\Delta_m = \theta_m\sqrt{m}/\sigma = \theta_{(a)}/\sigma$. *Then, for* $m \to \infty$, *the optimum solution of* (5) *converges to that of* (2) *and the optimum solution of* (6) *to that of* (4).

5. Composite Alternatives

So far we have focused on the optimization of the expected number of possible rejections for a fixed reference alternative. Assuming that hypotheses are chosen at random from a pool of hypotheses at the design stage, we may also optimize the expected number of possible rejections under parameters generated at random from some alternative generating distribution F. We first discuss multiple Bonferroni tests, and then the Benjamini−Hochberg (FDR) setting.

For one-sided, one-sample z-tests, the expected number of correct rejections is

$$EN_k = q\,k \int_0^\infty \left(1 - \Phi(z_{\alpha/k} - \frac{\Delta_m(\theta)}{\sqrt{k}})\right) dF(\theta), \tag{7}$$

where F denotes the conditional c.d.f. of θ for a randomly chosen pair of hypotheses given $\theta > 0$, $q = P(\theta > 0)$, and $\Delta_m(\theta) = \theta(\sqrt{m}/\sigma)$. Let $k_{m,F} := m\,d_F^2/[2\,\sigma^2\,\log(m\,d_F^2/\sigma^2)]$, where d_F maximizes $d^2(1 - F(d))$.

Theorem 4. *Assume that F is continuous and that $d^2(1-F(d)) \to 0$ as $d \to \infty$. The optimum solution $k^*_{m,F}$ to (7) satisfies $k^*_{m,F} = k_{m,F}(1 + o(1))$.*

As for fixed alternatives, arguments by Genovese and Wasserman (2002) turn out to be helpful when optimizing k for the Benjamini−Hochberg procedure in the case of distributed alternatives. Indeed, according to their Theorem 7, the BH-procedure is now asymptotically equivalent to a single-step procedure with significance level u solving

$$\beta\,u = \int_0^\infty H_{\Delta_m(\theta)/\sqrt{k}}(u)\,dF(\theta), \tag{8}$$

where F is defined as above and H and β are as in (3). Analogous to Section 3, we focus on the one-sided normal case. Our optimization problem is given by

$$EN_k = q\,k \int_0^\infty \left(1 - \Phi(z_u - \frac{\Delta_m(\theta)}{\sqrt{k}})\right) dF(\theta) \to \max_k, \tag{9}$$

where u is defined in (8).

According to Theorem 5 below, the solution can be reduced to the maximization of an objective function that does not depend on m and σ. Furthermore, a scale transform $F(\theta/\gamma)$ has γ^2 times the optimum k of $F(\theta)$.

Theorem 5. *Take $k^*_{m,F} = m/(\omega_\beta(u^*_\beta)\,\sigma)^2$ to optimize (9), where $\omega_\beta(u)$ solves $\int_0^\infty(1 - \Phi(z_u - \theta\omega))dF(\theta) = \beta u$, in ω, and u^*_β maximizes $u/\omega_\beta(u)^2$.*

The proof is analogous to the case of a single alternative.

As in the case of a single alternative, it follows now from Theorem 5 that $EN_{k^*_{m,F}} \to \infty$, if the optimum number of hypotheses $k^*_{m,F}$ is chosen.

6. Classification Procedures

We focus on the problem of classifying between $\theta = 0$ (decision D_0) versus $\theta = \theta_{(a)}$ (decision D_1).

With $g_k(\theta)$ denoting the probability that a chosen classification rule decides for D_1 under the true parameter θ, classification rules intend to minimize some linear combination

$$k\left(w_1\, q\,[1 - g_k(\theta_1)] + w_0\,(1-q) g_k(0)\right) \tag{10}$$

of the expected number of false decisions for D_0 and for D_1. The constants w_1 and w_0 can be interpreted as costs associated with the two types of false decisions. The direct minimization of the above objective function in k does not make sense, since obviously doing nothing and choosing $k = 0$ would minimize the expected number of errors.

However, for fixed k, the problem is equivalent to that of finding a rule that maximizes

$$k\left(w_1\, q\, g_k(\theta_1) - w_0\,(1-q) g_k(0)\right). \tag{11}$$

This can be interpreted as maximizing the expected number of correct decisions for D_1 minus the incorrect decisions for D_1, weighted according to the respective gain w_1 and cost w_0. The advantage of this version is that it permits maximization in k.

We consider the optimization of (11) for the Bayes classifier. For any given k, it is well known that the Bayes classifier optimizes both (10) and (11) (see for instance Duda, Hart and Stork (2001, Chap. 2)). The Bayes classifier decides for D_1, if the likelihood ratio exceeds a threshold $r = w_0\,(1-q)/(w_1\, q)$, depending on the ratio of the priors and the ratio of the cost.

It is easily checked that for n i.i.d. normal $N(\mu, \sigma^2)$ observations and $\theta_{(a)} > 0$, the Bayes classifier decides for D_1, if

$$T_i := \sqrt{n}\frac{\bar{X}}{\sigma} > c(r, \Delta_n) := \frac{\log(r)}{\Delta_n} + \frac{\Delta_n}{2},$$

where $\Delta_n = \theta_{(a)}\sqrt{n}/\sigma$. See for instance Das Gupta (1982). In this context, (11) becomes

$$U(k) := k\,\left\{w_1\, q\,\Phi\left(\Delta_n - c(r, \Delta_n)\right) - w_0\,(1-q)\Phi(-c(r, \Delta_n))\right\}. \tag{12}$$

When k classification problems are considered, $n = m/k$ observations are available for each problem, and we set $\Delta_n = \Delta_{m/k} = \Delta_m/\sqrt{k}$ when optimizing in k. The result below follows by differentiating (12) with respect to $\Delta_{m/k}$.

Theorem 6. *If $r > 1$, the maximum of* (12) *in k satisfies $k = (\Delta_m / x_r)^2$, where x_r is the solution of*

$$0 = x\,\varphi[x - c(r,x)]/2 - \Phi[x - c(r,x)] + r\,\Phi[-c(r,x)],$$

with $c(r,x) = \log(r)/x + x/2$, and $\varphi(\cdot)$ denoting the standard normal density.

Notice that the optimum sample size for each individual classification problem, given by $m/k_m^* = (x_r/\theta_{(a)})^2$, is independent of m. Substituting the optimum k into $c(r, \Delta_{m/k})$, it turns out that the critical value of the Bayes classifier is independent of Δ_m, and depends only on r. This is in contrast to the case of a fixed number of classification problems, where m and $\theta_{(a)}$ do influence the Bayes classifier.

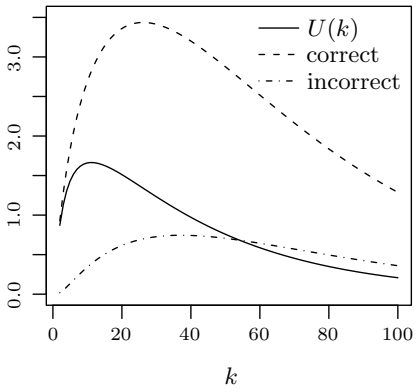

Figure 2. The function $U(k)$ and the expected number of correct and incorrect decisions for D_1 for given k and the parameters $m = 100$, $q = 0.5$, $w_1 = 1$, $w_2 = 3$, and $\theta_{(a)}/\sigma = 1/2$. The optimum is reached at $k = 11.15$.

A typical example displaying the expected number of correct and incorrect decisions D_1 and the value of $U(k)$ is given in Figure 2. Both the number of correct and the number of incorrect decisions D_1 attain a global maximum in k.

7. Example

The following example follows the lines of an actual consulting case for a project investigating gender differences in the expression of metabolizing and membrane transporting enzymes in gall bladders. Gender differences in expression intensity are to be assessed by two sided two-sample t-tests. The determination of expression intensity is performed by polymerase chain reaction (PCR). The investigator originally planned to test for ten different enzymes but the project budget allows PCR's for at most 200 intensity measurements in total. The optimization problem is to asses how many different

enzymes should be investigated. Assuming equal variances σ^2 in both genders, analogous to (5) the expected number of correct rejections is given by $q\,k[1-F^{(t)}_{m/(2\,k)-2,|\theta_{(a)}|\sqrt{m/(2\,\sigma^2 k)}}(t_{\alpha/(2\,k),m/(2\,k)-2})]$, where the Bonferroni multiple testing procedure with two sided multiple level α is used. Setting $\alpha = 0.05$ and assuming an effect size of half a standard deviation, the expected number of correct rejections is maximized if two enzymes are tested, each with 100 patients per group. The expected number of rejected null hypothesis, given for both enzymes the alternative holds, is 1.49. If, instead, the experimenters were testing all ten enzymes, after Bonferroni correction the expected number of rejections is as low as 0.39, given that for all enzymes the alternative holds (see Figure 3).

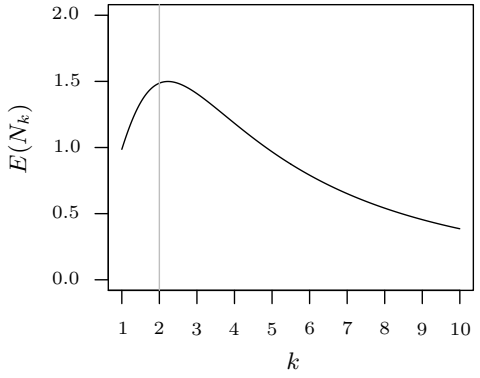

Figure 3. The expected number of rejected hypothesis over k for the example in Section 7. The optimum is reached at $k^* = 2$.

8. Discussion

From our theoretical and simulation results, it turned out that the appropriate choice of the number of hypotheses considered has a considerable influence on the number of rejections, and should thus be considered at the design stage of studies when the overall number of observations is limited.

Traditionally sample sizes are often planned in order to achieve a desired power for the individual tests. Choices of 0.8 or 0.9 for the power are particularly popular and motivated by cost and other considerations. Our results may be used as another guideline to choose an appropriate power, taking into account the expected number of possible rejections.

Comparing the results for Bonferroni testing on the one hand, and classification as well as Benjamini−Hochberg testing on the other, it turns out that there is a qualitative difference in the limit behavior of the optimum k^*_m as $m \to \infty$: while the optimum k^*_m for the Bayes classifier and BH tests increases at a linear

rate, it increases only at the rate $m/\log(m)$ for the Bonferroni procedure. This implies, for the latter, that the optimum sample size for each individual hypothesis tends to infinity, while for both BH tests and Bayes classifier it is independent of m.

We provide R functions to calculate the optimum k in our considered situations at http://www.meduniwien.ac.at/medstat/misc/opt/optimum.htm.

Appendix

Proof of Theorem 1. Note that $(\varphi(x)/x)(1-x^{-2}) \le 1-\Phi(x) \le \varphi(x)/x$ for $x > 0$ (see inequality (61) of Galambos (1987)) and therefore (see also Reiss (1989)) $z_{\alpha/k} = \sqrt{2\log(k/\alpha)} + o(1)$, as $k \to \infty$.

Let now $c_m^{(1)} = a_m/\sqrt{2\log(m)}$ with $a_m \to \infty$ and $a_m = o(\sqrt{2\log m})$ and $k_m^{(1)} = k_m(1-c_m^{(1)})$. Then

$$\begin{aligned} z_{\alpha/k_m^{(1)}} - \sqrt{\frac{\Delta_m^2}{k_m^{(1)}}} &= \sqrt{2\log(m)}(1-[1-c_m^{(1)}]^{-1/2}) + o(1) \\ &= -\frac{a_m}{2} + o(1) \to -\infty, \end{aligned}$$

by Taylor expansion of $(1-c_m^{(1)})^{-1/2}$. Therefore, according to (2), $E(N_{k_m^{(1)}}) = qk_m(1+o(1))$.

It remains to show that $k_m^{(2)} = k_m(1-c_m^{(2)})$ leads an expected number of rejections that is lower than $E(N_{k_m^{(1)}})$ for large m, if $c_m^{(2)}$ does not satisfy both $c_m \to 0$ and $c_m\sqrt{\log m} \to \infty$. Assume first that $c_m^{(2)} > \delta$ for some $\delta > 0$. Then obviously $E(N_{k_m^{(2)}}) \le qk_m(1-\delta)$, and $k_m^{(2)}$ cannot be optimal. If, on the other hand $c_m^{(2)} < -\delta < 0$, then

$$d_m := z_{\alpha/k_m^{(2)}} - \sqrt{\frac{\Delta_m^2}{k_m^{(2)}}} \ge \delta^* \sqrt{2\log m}$$

for some $\delta^* > 0$ and therefore $1-\Phi(d_m) \le (\varphi(d_m)/d_m) \le m^{-(\delta^*)^2}$ for large enough m. Thus $E(N_{k_m^{(2)}}) = o(k_m)$ in this case. Assume finally that $c_m^{(2)} = a_m/\sqrt{\log m}$ and $|a_m| < \delta$ for some constant δ. Then $E(N_{k_m^{(2)}}) \le qk_m[1-\Phi(-\delta)](1+o(1))$, and $k_m^{(2)}$ again is not optimal.

Proof of Theorem 2. By (3), we have that $\beta u = 1 - \Phi_{(|\Delta_m|\sqrt{1/k},1)}(z_u)$, or equivalently $k = \Delta_m^2/(z_u - z_{\beta u})^2$. By substituting for k, (4) becomes $q\,\Delta_m^2\,\beta\,u/(z_u - z_{\beta u})^2$. The result follows now by maximizing the above in u instead of (4) in k.

Proof of Theorem 3. We give a proof for the Bonferroni rule only, since the arguments for the Benjamimi−Hochberg procedure are completely analogous. Since the z-test is uniformly most powerful, the objective function (5) is always below (2). The result follows, since $\mathbf{E}N_k^{(t)} \to \mathbf{E}N_k$ for the fixed k optimizing (2), due to the weak convergence of the t-statistic to the corresponding z-statistic.

Proof of Theorem 4. For simplicity, assume $\sigma = 1$. We set w.l.o.g. $k = k_m = m\,\delta_m^2/(2\log m)$ in (7) and optimize equivalently over $\delta_m \in S_m := [\sqrt{2\log m/m}, \sqrt{2\log m}]$ corresponding to $k_m \in [1, m]$. Define furthermore $a_m := z_{\alpha/k_m}$, $b_m = b_m(\delta_m) := \sqrt{m/k_m} = \sqrt{(2/\delta_m^2)\log m}$, and $c_m = \log\log m$.

We start by considering the case $\delta_m \in S_m^* = [c_m^{-1}, \sqrt{2\log m}] \subset S_m$. In this case, we have that $a_m = \sqrt{2\log m} + O((\log\log m)/\sqrt{\log m})$ uniformly in δ_m (see Reiss (1989, p.161)). Dropping q and multiplying by $\log m/m$, our objective function (7) is

$$\delta_m^2 \int_0^\infty (1 - \Phi(a_m - \theta b_m))dF(\theta) = \delta_m^2\Big(\int_0^{\frac{a_m - c_m}{b_m}} + \int_{\frac{a_m - c_m}{b_m}}^\infty\Big) := I_1 + I_2. \quad (13)$$

Obviously $I_1 \le c_m^2(1 - \Phi(c_m)) = o(1)$ and

$$I_2 \le \delta_m^2[1 - F(\frac{a_m - c_m}{b_m})] = \delta_m^2\Big[1 - F\Big(\delta_m\Big(1 - \frac{\log\log m}{\sqrt{2\log m}} + O(\frac{\log\log m}{\log m})\Big)\Big)\Big].$$

Since $d^2(1 - F(d)) \to 0$ for $d \to \infty$ according to our assumptions, our result for δ_m restricted to S_m^* follows by proving that the inequality for I_2 is asymptotically sharp for δ_m in some open bounded interval (C_1, C_2) containing the maximizer(s) of $u^2(1 - F(u))$. Write

$$I_2 = \delta_m^2\Big(\int_{\frac{a_m - c_m}{b_m}}^{\frac{a_m + c_m}{b_m}} + \int_{\frac{a_m + c_m}{b_m}}^\infty\Big) := I_3 + I_4.$$

Since $(c_m/b_m) \to 0$ on $[C_1, C_2]$, it follows that $I_3 \le C_2^2(2c_m/b_m) = o(1)$ uniformly on (C_1, C_2). Furthermore, since the integrand converges to one on the considered interval, $I_4 = \delta_m^2[1 - F((a_m + c_m)/b_m)] + o(1) = u_m^2[1 - F(u_m)] + o(1)$, for $u_m = \delta_m(1 - (\log\log m)/\sqrt{2\log m} + O((\log\log m)/\log m))$. Together the δ_m maximizing (13) satisfies $\delta_m = u + o(1)$ for the u maximizing $u^2[1 - F(u)]$. The proof is now completed by checking that (13) is bounded from above by $(\log\log m)^{-2}$ for $\delta_m < c_m^{-1}$, and by checking that $k_{m,F} = k_m(1 + o(1))$ for k_m.

Acknowledgement

We wish to thank Peter Bauer for pointing us at the example in Section 7 and for further helpful suggestions.

References

Bechhofer, R. E. and Tamhane, A. C. (1983). Design of experiments for comparing treatments with a control: tables of optimal allocations of observations. *Technometrics* **25**, 87-95.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Das Gupta, S. (1982). Optimum rules for classification into two multivariate normal populations with the same covariance matrix. In *Classification, Pattern Recognition and Reduction of Dimensionality*. **2** *Handbook of Statist.*, 47-60. North-Holland, Amsterdam.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*. 2nd edition. Wiley-Interscience, New York.

Dunnett, D. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* **50**, 1096-1121.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. *Biometrical J.* **43**, 985-1005.

Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. 2nd edition. Krieger, Malabar.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64**, 499-517.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5**, 299-314.

Reiss, R.-D. (1989). *Approximate Distributions of Order Statisitics*. Springer, New York.

Rubinstein, R. Y. and Melamed, B. (1998). *Modern Simulation and Modeling*. John Wiley, New York.

Rubinstein, R. Y. and Shapiro, A. (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization*. John Wiley, New York.

Seeger, P. (1968). A note on a method for the analysis of significance en masse. *Technometrics* **10**, 586-593.

Somerville, P. N. and Bretz, F. (2001). Obtaining critical values for simultaneous confidence intervals and multiple testing. *Biometrical J.* **43**, 657-663.

Department of Statistics, University of Vienna, Universitätsstr. 5, A-1010 Vienna, Austria.

E-mail: Andreas.Futschik@univie.ac.at

Section of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.

E-mail: Martin.Posch@meduniwien.ac.at

(Received October 2003; accepted May 2004)