Multiple Comparisons Distortions of Parameter Estimates

By NEAL O. JEFFRIES

MSC 1430, 10 Center Drive

Office of the Clinical Director

National Institute of Neurological Disorders and Stroke

National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

neal.jeffries@nih.gov

Summary

In experiments involving many variables investigators typically use multiple comparisons procedures to determine differences that are unlikely to be the result of chance. However, investigators rarely consider how the magnitude of the greatest observed effect sizes may have been subject to bias resulting from multiple testing. These questions of bias become important to the extent investigators focus on the magnitude of the observed effects. As an example, such bias can lead to problems in attempting to validate results if a biased effect size is used to power a follow-up study. Further, such factors may give rise to conflicting findings in comparing two independent samples – e.g. the variables with strongest effects in one study may predictably appear much less so in a second study. An associated important consequence is that confidence intervals constructed using standard distributions may be badly biased. A bootstrap approach is used to estimate and correct the bias in the effect sizes of those variables showing strongest

differences. This bias is not always present; some principles showing what factors may lead to greater bias are given and a proof of the convergence of the bootstrap distribution is provided.

Key words: Effect size, bootstrap, multiple comparisons

1. Introduction

Most considerations involving multiple comparisons problems focus upon the increased probability of false positive errors when the null hypothesis is true. Here the focus is upon the distorting effects of multiple comparisons in evaluating those variables judged to show the strongest effects or differences between groups. These distortions may be present both when the null hypothesis of no difference is true as well as when it is false.

This bias can be relevant in some circumstances. If a power analysis is employed for a follow-up study the study will likely be underpowered if overestimation bias is present. Further, a follow-up study may be difficult to mount and the preliminary study may provide the best point estimate and this estimate should be deflated if bias is present. In genetic epidemiology marker studies there is sometimes interest in assessing the strength of a marker's association – if the strength is low it may not be worth performing a fine-mapping or other follow-up study. Also, confidence intervals for the point estimate will reflect the degree of bias affecting the estimate.

Further, the bias effect may help to explain apparent disagreement between studies based on similar populations but independent samples, i.e. explaining

some disagreement across studies. As an example the most differentially expressed gene in microarray study A is only the 500^{th} or 1000^{th} most expressed gene in microarray study B. Some may view such results as indicative of unreliable technology though the findings may arise from overestimation bias. Studies of microarray platforms have proposed such ranking measures of comparability (e.g. Irizarry et al., 2004) and not acknowledging this bias may lead to an overly pessimistic assessment. Similar problems in validating effects of genetic markers in confirmatory studies may also be due to multiple comparisons distortions.

Recent work by genetic epidemiologists (e.g. Sun and Bull, 2005 and Siegmund, 2002) has focused upon this bias problem in the context of estimating the presence and magnitude of genetic marker effects in genome-wide scans. The former study examines bootstrap and cross-validation approaches similar to that proposed here though in the present work more attention is paid to estimating the entire distribution of overestimation and determining confidence intervals. The latter reference puts forth an analytical approach that is highly dependent upon the genetic model that is assumed and therefore appears to be restricted largely to genetic marker studies. Also, both of these papers posit the overestimation arises from truncation bias related to the significance threshold for declaring significance – here a different presentation of bias is described and addressed that is given without reference to a significance threshold. The basic idea is that observed outcomes are composed of random and deterministic components and under some circumstances the fact that one outcome performs best may suggest the random component for that outcome was unusually beneficial and this 'good

luck' is associated with overestimation of the true effect. Determining when such circumstances exists and measuring their distortion is the focus of this paper.

The use of the bootstrap to estimate confidence intervals for observed maximal statistics is not new – see section 3.1 in Westfall and Young (1993) for creating simultaneous confidence intervals. What is different here is the approach is used to quantify and reduce overestimation bias.

2. Illustrations of the Problem

An elementary two-group $t-$test applied to a number variables will be used to illustrate some principles. A simple examination of gene expression differences between healthy and diseased individuals could give rise to such a design.

We assume each of two groups has n individuals, a total of G variables (e.g. genes) are measured, and for variable j we denote the n response measures as X_{ij} in group 1 and Y_{ij} in group 2 for $i \in \{1 \ldots n\}$ and $j \in \{1 \ldots G\}$. Let $d_j = \bar{x}_j - \bar{y}_j$, σ_j denote the standard deviation under the assumption of common variability in the two groups, and s_j denote an estimate of σ_j, i.e.

$$s_j = \sqrt{.5\left(\frac{\sum(x_{ij} - \bar{x}_j)^2}{n-1}\right) + .5\left(\frac{\sum(y_{ij} - \bar{y}_j)^2}{n-1}\right)} \tag{2.1}$$

where $\bar{x}_j$ and $\bar{y}_j$ denote the two sample averages for variable j. If $\mu_j = EX_{ij} - EY_{ij}$ denotes the average difference for the j^{th} variable then the $t-$statistic may be

written as

$$t_j = \frac{\sqrt{n}(\bar{x}_j - \bar{y}_j)}{\sqrt{2}\, s_j} \tag{2.2}$$

$$= \frac{\sqrt{n}\,(\bar{d}_j - \mu_j)}{\sqrt{2}\, s_j} + \frac{\sqrt{n}\,\mu_j}{\sqrt{2}\, s_j} \tag{2.3}$$

$$= \tau_j + \frac{\sqrt{n}\,\mu_j}{\sqrt{2}\, s_j} \text{ where } \tau_j = \frac{\sqrt{n}\,(\bar{d}_j - \mu_j)}{\sqrt{2} s_j}. \tag{2.4}$$

τ_j is a realization of a random variable with a $t-$distribution having $2n-2$ degrees of freedom. The distinction between t and τ is that the latter has a $t-$distribution (centered about 0) regardless of whether the null hypothesis, $\mu_j = 0$, is true. One sees in (2.4) that the degree to which the sample difference, $\bar{d}_j$, exceeds the true difference, μ_j, is associated with the magnitude of τ_j. This degree of overestimation expressed by τ_j is of primary interest in this paper. Now consider the collection of $j \in \{1 \ldots G\}$ variables and associated values t_j, τ_j, $\bar{d}_j$, and μ_j. Of interest is the distribution of τ_j when it corresponds to a gene with an extreme t_j value. Let $r_1, r_2, \ldots, r_G$ denote the indices associated with the smallest to largest $t-$statistics so that

$$t_{r_1} \leq t_{r_2} \leq \cdots \leq t_{r_G} \tag{2.5}$$

It is difficult to know in general the distribution of

$$\tau_{r_1} = \frac{\sqrt{n}\,(\bar{d}_{r_1} - \mu_{r_1})}{\sqrt{2}\, s_{r_1}} \text{ or } \tau_{r_G} = \frac{\sqrt{n}\,(\bar{d}_{r_G} - \mu_{r_G})}{\sqrt{2}\, s_{r_G}} \tag{2.6}$$

and hence the degree to which $\bar{d}_{r_1}$ underestimates μ_{r_1} (or $\bar{d}_{r_G}$ overestimates μ_{r_G}) though some simplified situations can reveal these distributions' dependence on different factors.

For example, consider a genechip with 30000 genes of which 2% are differentially expressed between the two groups. For simplicity suppose that for these 2 percent of the genes the same μ and s values apply and the measures are independent across genes. Then because

$$t_j = \tau_j + \frac{\sqrt{n}\,\mu}{\sqrt{2}\,s} \tag{2.7}$$

we see that t_{r_G} corresponds to that gene with the highest τ_j value. In this simplified example and assuming the highest $t-$statistic corresponds to one of the 2% with true differential expression, then this maximal value of τ_j corresponds to the maximum value among 600 independent and identically distributed $t-$statistics. For n, the number of biological samples in each group, equal to 10 the expected value of the maximum is about 3.64, i.e.

$$E\left[\tau_{r_G}\right] = E\left[\frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}\,s_{r_G}}\right] = 3.64. \tag{2.8}$$

If the true difference μ in the 600 genes (perhaps measured on a $\log_2$ scale) is 1 and the estimated standard deviation is also 1, then by manipulating (2.8) we see that for the gene with greatest t_j value the expected degree of overestimation is on the order of

$$E\left[\bar{d}_{r_G} - \mu_{r_G}\right] \approx 3.64\frac{1 \cdot \sqrt{2}}{\sqrt{10}} = 1.63 \tag{2.9}$$

where we have treated s_{r_G} and τ_{r_G} as if they were independent. Instead of the true fold difference of 1 (doubling of gene expression), the reported fold difference on average would closer to 2.63 (gene expression more than 6 times higher). In this case the bias is profound.

The preceding example has a simple structure that makes it easy to present but has a number of approximations and assumptions that limit its generality and produce an extreme example of bias. A more realistic assessment of distortion may be presented using real microarray data. Affymetrix hgu133A chips were used in a spike-in experiment to assess different methods of generating expression measures from probe level data (see http://affycomp.biostat.jhsph.edu/). The 28 arrays (corresponding to the first 28 CEL files in hgu133spikein.tgz) were all hybridized to a common mRNA source and the 42 spiked genes were excluded leaving 22258 genes for analysis. An additional 26 genes were excluded that showed high variability and may have been correlated with the spiked genes. Preprocessing involved running the justRMA procedure developed as part of the Bioconductor suite of microarray analysis tools (see the Affy package at http://www.bioconductor.org).

The 28 arrays were randomly split into two groups of 14. By construction, $t-$tests should show no differential expression for any genes. To investigate the bias in effect size a random selection of 1.5% of the genes were chosen to receive nonzero effect sizes and different effect sizes were allocated to these $.015 * 22232 \approx 333$ genes. The effect sizes were asymmetrically distributed about zero. This was done because in many microarray studies differences will tend to more heavily represent over- or under-expression. Further, the different patterns of over- and under-estimation create different patterns of bias. Figure 1 shows the pattern of effect sizes that was used. There were 55 effect sizes less than 0 with differences of .20 units between the 6 most negative effect sizes (the remainder evenly spaced

between -1 and 0) and there were 278 positive effect sizes (evenly distributed between 0 and 2).

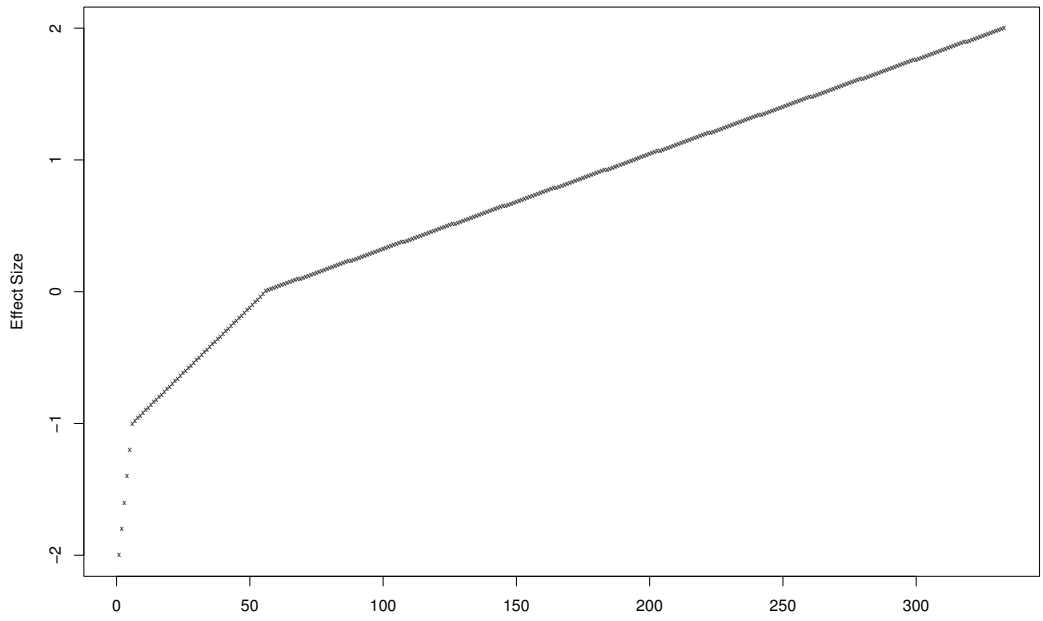

FIGURE 1. Distribution of Nonzero Effect Sizes

After these nonzero effect sizes were imposed upon 333 randomly chosen genes, 22232 $t-$tests were then calculated for the adjusted arrays with to determine the degree to which the genes with most extreme $t-$statistics overestimate the true effect sizes. In other words, $t_{r_1}, t_{r_2}, \ldots, t_{r_G}$ were calculated and $\frac{\mu_{r_1}}{\sigma_{r_1}}$ was compared to $\frac{\bar{d}_{r_1}}{s_{r_1}}$ and $\frac{\mu_{r_G}}{\sigma_{r_G}}$ was compared to $\frac{\bar{d}_{r_G}}{s_{r_G}}$. Then a second simulation was performed by rerandomizing the 28 arrays in 2 groups and choosing a new group of 333 to receive the fixed pattern nonzero effect sizes and new estimates of overestimation of $\frac{\mu_{r_G}}{\sigma_{r_G}}$ and underestimation of $\frac{\mu_{r_1}}{\sigma_{r_1}}$ are obtained. These simulations were performed 1000 times and distributions of over- and underestimation are obtained and shown in Figure 2. It is important to note that different genes (and hence different true

effect sizes) corresponding to τ_{r_G} and τ_{r_1} may be selected in different simulations.

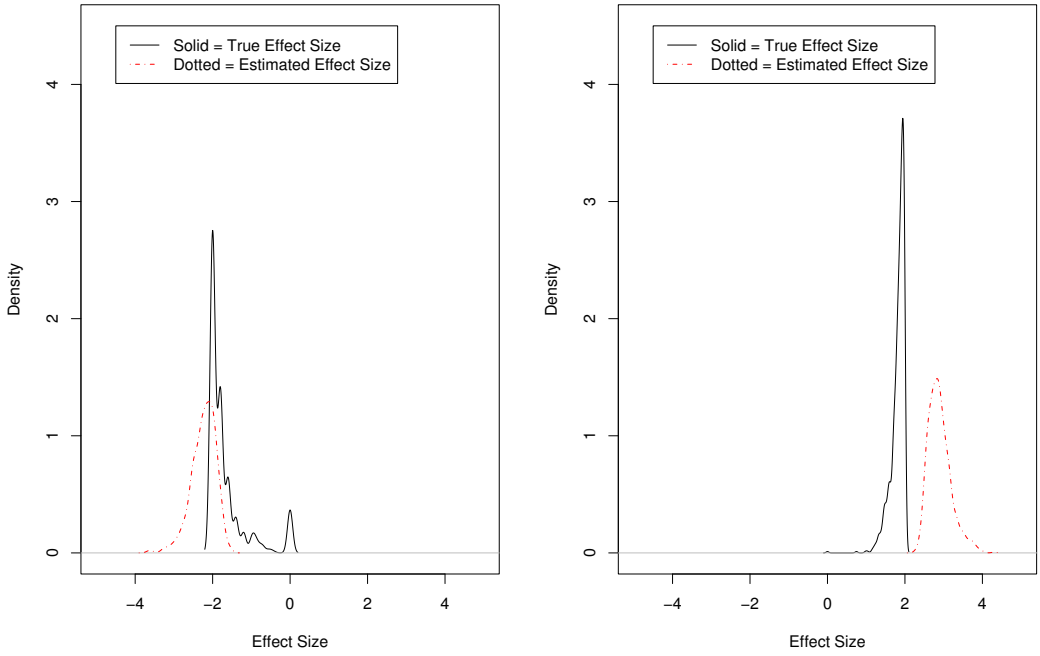

FIGURE 2. Left Panel: Distribution of True Effect Size $\frac{\mu_{r_1}}{\sigma_{r_1}}$ and Estimated Effect Size $\frac{\bar{d}_{r_1}}{s_{r_1}}$ of Gene with Smallest $t-$statistic. Right Panel: Distributions of True Effect Size $\frac{\mu_{r_G}}{\sigma_{r_G}}$ and Estimated Effect Size $\frac{\bar{d}_{r_G}}{s_{r_G}}$ of Gene with Largest $t-$statistic.

We see the estimated effect sizes are more extreme than the true effect sizes. This bias is more pronounced for $\frac{\mu_{r_G}}{\sigma_{r_G}}$ than it is for $\frac{\mu_{r_1}}{\sigma_{r_1}}$. The average estimated effect size is -2.24 for the most negative $t-$statistic while the average estimate is 2.90 for the most positive – both estimates are biased as the true effect sizes all lie between -2 and 2. Another factor to notice is that the distribution of true effect sizes for the smallest $t-$statistic is more broad and includes some genes with 0

effect size, i.e. in some simulations the gene producing the smallest $t-$statistic was one of the 22232-333 = 21899 with no differential expression.

Associated with this distribution of true and estimated effect sizes for the most extreme $t-$statistics are the related distribution for τ_{r_1} and τ_{r_G} where

$$\tau_{r_1} = \frac{\sqrt{n}\left(\bar{d}_{r_1} - \mu_{r_1}\right)}{\sqrt{2}s_{r_1}} \text{ and } \tau_{r_G} = \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}s_{r_G}}. \tag{2.10}$$

These distributions (as derived from the simulations) are shown in Figure 3 along with the usual t-distribution that is commonly used for inference about the genes showing most extreme differences.

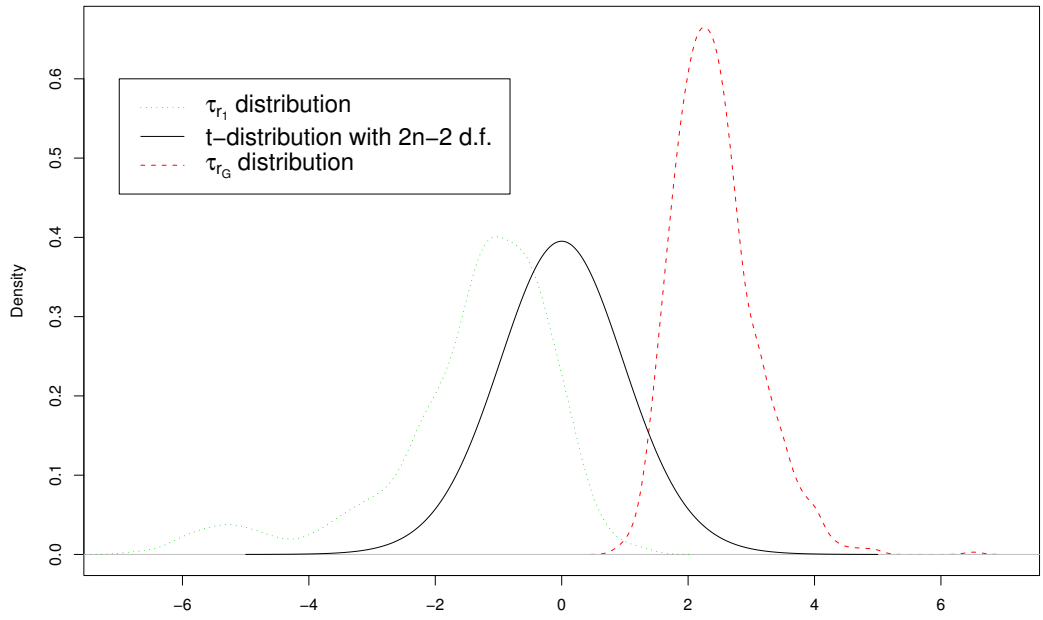

FIGURE 3. Distribution of τ_{r_1}, t with $2n-2$ d.f., and τ_{r_G}

The figure shows the three distributions are quite distinct and suggests that inference for μ_{r_1} or μ_{r_G} would be misleading if a standard $t-$distribution were used. To illustrate, attention will be focused upon μ_{r_G}. For these simulations

the median magnitude of the greatest positive estimated effect size, $\frac{\bar{d}_{r_G}}{s_{r_G}}$ was 2.85. Suppose in a particular fixed simulation $\mu_{r_G} = 2.85$ and $s_{r_G} = 1$ so an $2^{2.85} = 7.21$ change was associated with the largest t-statistic. The usual 95% confidence interval for μ_{r_G} would be given by

$$\bar{d}_{r_G} \pm t^{-1}_{.975,2n-2} \frac{\sqrt{2} s_{r_G}}{\sqrt{n}} = [2.07, 3.63] \tag{2.11}$$

where $t^{-1}_{.975,2n-2}$ satisfies $P[T \leq t^{-1}_{.975,2n-2}] = .975$ where T has a $t-$distribution with $2n-2$ degrees of freedom. This interval is highly suspect as by construction the true effect size lies in $[-2, 2]$ and is likely near 2. To contrast this with the distribution of τ_{r_G} let F denote the distribution of τ_{r_G} and let F^{-1}_{α} satisfy $Pr\left[\tau_{r_G} \leq F^{-1}_{\alpha}\right] = \alpha$. To proceed further consider $F^{-1}_{.975}$ and $F^{-1}_{.025}$ to generate a 95% confidence interval, i.e. find μ_{r_G} satisfying

$$\left[F^{-1}_{.025} \leq \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2} s_{r_G}} \leq F^{-1}_{.975} \right] \tag{2.12}$$

In this case the necessary percentiles may be estimated from the simulations, $F^{-1}_{.975} = 3.91$, and $F^{-1}_{.025} = 1.41$. Given the other parameters $s_{r_G} = 1, \bar{d}_{r_G} = 2.85, n = 14$ the confidence interval for μ_{r_G} may be given as

$$\left[\bar{d}_{r_G} - s_{r_G} \frac{\sqrt{2}}{\sqrt{n}} F^{-1}_{.975},\ \bar{d}_{r_G} - s_{r_G} \frac{\sqrt{2}}{\sqrt{n}} F^{-1}_{.025} \right] = [1.37, 2.32] \tag{2.13}$$

Further, a point estimate may be computed using the median value $F^{-1}_{.50}$ as

$$\bar{d}_{r_G} - s_{r_G} \frac{\sqrt{2}}{\sqrt{n}} F^{-1}_{.50} = 1.96. \tag{2.14}$$

One may compare this point estimate, 1.96, and confidence interval, [1.37,2.32], with those estimates associated with the standard 95% t-distribution: point estimate of 2.85 and confidence interval of [2.07,3.63]. From the simulations one can determine the true median effect size for the largest $t-$statistic is 1.87 so the estimate and confidence interval based on τ_{r_G} comes relatively close to the truth while the usual, naïve estimate is badly biased.

In Figure 3 the distributions of τ_{r_1} and τ_{r_G} are quite different. That of τ_{r_1} appears similar in shape to that of a standard $t-$statistic with $2n - 2 = 26$ degrees of freedom, though shifted somewhat to illustrate the underestimation and with a block of data near -5 or -6 corresponding to those occasions in which a gene with no true differential expression was chosen. The distribution of τ_{r_G} is substantially different from the $t-$distribution, both in location and shape.

As an aside it should be noted (data not provided) that for all but one of the 1000 simulations the p-value associated with the most positive $t-$statistic passed Bonferroni criteria, i.e. $p < .05/22232$ and that the most extreme statistic did correspond to one of the 333 genes with non-zero effect sizes. Given the conservative nature of the Bonferroni threshold the associated genes would be selected as differentially expressed by common multiple corrections criteria. This shows that multiple comparisons are a problem not only when there is no differential expression but can also create difficulty when differences exist.

One may raise the question of whether this overestimation and biased confidence intervals matter. While this may be a qualitatively unimportant distinction the

bias can be important in at least two situations: when a follow-up study is considered and a power analysis is based upon the initial estimates, and when one compares results from one study with those obtained from an independent group of arrays.

With respect to a power analysis, consider a second set of simulations of a similar structure except the effect sizes are chosen to lie between $[-1, 1]$ and 3% of genes are differentially expressed (the first set of simulations used $[-2, 2]$ as the effect size region and only 1.5% of genes were differentially expressed). The same pattern of differences were used except a) the nonzero effect sizes were all divided by 2 and b) there were two genes having each of the distinct effect sizes creating 666 genes with differential expression instead of 333.

In this case the average value of τ_{r_G} was 2.06 and the average true effect size generating τ_{r_G} was 0.78. A power analysis based on an effect size of 2.06 would suggest only 5 arrays/assays per group to achieve power of 80% using a two-sided t-test with $\alpha = 0.05$ without any correction for multiple comparisons (this may be appropriate in the context of follow-up study of a targeted gene). However, even if the investigator chose to double this estimate and use 10 per group due to parameter uncertainty, the true average absolute effect size of 0.78 suggests that at least 27 per group would be needed to obtain power of 80%. In such circumstances the failure of a follow-up study to show significant results is largely a result of overestimation bias rather than lack of a true effect.

3. Estimation of the τ_{r_G} Distribution

As μ_{r_G} is unknown, one cannot directly estimate the distribution of

$$\tau_{r_G} = \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}\, s_{r_G}}. \tag{3.15}$$

The bootstrap techniques popularized by Efron (e.g. Efron and Tibshirani, 1998) will be used to develop confidence intervals that compensate for the multiple comparisons bias described in the previous section. To proceed one first constructs a bootstrap sample from the X_{ij}, Y_{ij} by sampling arrays with replacement from these data in a stratified manner, i.e. sampling from the X_{ij} and Y_{ij} separately. To preserve the within-array correlation structure, one samples entire arrays, not the individual genes. From here one obtains bootstrap samples X^*_{ij} and Y^*_{ij} and can compute associated bootstrap statistics $\bar{d}^*_j, s^*_j,$ and t^*_j. For a particular bootstrap sample designated by the * superscript, let $r^*_1, r^*_2, \ldots, r^*_G$ order the t statistics, t^*_j, i.e.

$$t^*_{r^*_1} \le t^*_{r^*_2} \le \cdots \le t^*_{r^*_G}. \text{ Then compute } \tau^*_{r^*_G} = \frac{\sqrt{n}\left(\bar{d}^*_{r^*_G} - \bar{d}_{r^*_G}\right)}{\sqrt{2}\, s^*_{r^*_G}} \tag{3.16}$$

or $\tau^*_{r_{1^*}}$ or any other ordered τ^* of interest. One may produce and process many bootstrap samples in this way and obtain an empirical distribution of $\tau^*_{r^*_G}$. The hope is that the unknown distribution of τ_{r_G} may be approximated by that of $\tau^*_{r^*_G}$. In considering the terms

$$\tau_{r_G} = \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}\, s_{r_G}} \text{ and } \tau^*_{r^*_G} = \frac{\sqrt{n}\left(\bar{d}^*_{r^*_G} - \bar{d}_{r^*_G}\right)}{\sqrt{2}\, s^*_{r^*_G}} \tag{3.17}$$

the idea is that the degree to which $\bar{d}_{r_G}$ exceeds μ_{r_G} can be approximated by the degree to which $\bar{d}^*_{r^*_G}$ exceeds $\bar{d}_{r^*_G}$. In other words, r^*_G is treated like r_G, $\bar{d}_{r^*_G}$ like

μ_{r_G}, and $\bar{d}^*_{r^*_G}$ like $\bar{d}_{r_G}$. Once an empirical distribution of $\tau^*_{r^*_G}$ is obtained, denoted by F^*, one may use the percentiles, F^{*-1} to create confidence intervals for μ_{r_G} as before, e.g.

$$\mu_{r_G} \text{ satisfying } \left[F^{*-1}_{.025} \leq \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}s_{r_G}} \leq F^{*-1}_{.975} \right] \tag{3.18}$$

where $\bar{d}_{r_G}$ and s_{r_G} are observed from the original data. For this procedure a second order bootstrap is also employed to improve the approximation of F^{-1} by F^{*-1}. This involves creating a further number of bootstrap samples and associated statistics from each first level bootstrap sample. Details of this nested percentile approach and an R program implementing it are available at http://krisa.ninds.nih.gov/multcomps.

Table 1 indicates how this approach works in terms of confidence intervals for μ_{r_G} in a simulation context. As before there are 14 individuals of each type. Instead of considering 22232 genes the simulations involve only $2\% \approx 444$ in an effort to keep the run-time manageable. Each of the 444 genes has an effect size chosen from the set $\{2/444, 4/444, \ldots, 886/444, 888/444\}$ and the genes/variables were constructed as independent. Values of $\bar{d}_{r_G}$ and μ_{r_G} were obtained from each simulation. Further the two-stage bootstrap algorithm was implemented and confidence intervals of varying nominal coverage were constructed. Table 1 gives characteristics of the coverage of these bootstrap intervals and intervals constructed using the naïve t-statistic approach. The results show that the naïve t-statistic intervals fail to cover very often and the bootstrap approach is better. Also worth noting is that the bootstrap intervals are about 20% longer. Though wider, this is not the primary reason the bootstrap covers better – instead it is

Nominal Interval	Bootstrap Coverage %	t-statistic Coverage %	Average Bootstrap Length	Average t-statistic Length
$25^{th} - 75^{th}$	50%	3%	.48	.39
$10^{th} - 90^{th}$	81%	12%	.90	.75
$5^{th} - 95^{th}$	91%	22%	1.18	.97
$2.5^{th} - 97.5^{th}$	96%	36%	1.42	1.17

TABLE 1. Confidence Interval Characteristics for μ_{r_G} with n=14, G=444, Effect Sizes Evenly spaced in $(0, 2]$, Variables Independent, 1000 simulations

due to the overestimation correction as expanding the $t-$statistic regions by 20% will increase the coverage probabilities to only 5%, 19%, 36%, and 55% for the four different intervals.

A second set of simulations was run with a much smaller number of variables/genes. Here the interest is in demonstrating to what extent this remains a problem in other applications when a more modest number of comparisons are involved. For these simulations $n = 14$ for each group, there are $G = 10$ variables (genes in the microarray context), the effect sizes are chosen at evenly spaced intervals between 0 and 1. Table 2 provides confidence interval characteristics for the bootstrap and t-statistic approaches. Here we see the naïve approach performs better though some distortion is still present. In this case the bootstrap intervals are of comparable length with good coverage characteristics.

Nominal Interval	Bootstrap Coverage %	t-statistic Coverage %	Average Bootstrap Length	Average t-statistic Length
$25^{th} - 75^{th}$	49%	34%	0.49	0.48
$10^{th} - 90^{th}$	77%	54%	0.96	0.93
$5^{th} - 95^{th}$	88%	74%	1.25	1.21
$2.5^{th} - 97.5^{th}$	93%	84%	1.53	1.45

TABLE 2. Confidence Interval Characteristics for μ_{r_G} with n=14, G=10, Effect Sizes evenly spaced in $(0, 1]$, Variables Independent, 1000 simulations

While the coverage probabilities of the bootstrap procedure appear accurate in these two tables, some problems remain as the bootstrap estimates continue to underestimate the bias. As an example, consider the interval

$$\left[\bar{d}_{r_G} - s_{r_G}\frac{\sqrt{2}}{\sqrt{n}}F^{*-1}_{.75},\ \bar{d}_{r_G} - s_{r_G}\frac{\sqrt{2}}{\sqrt{n}}F^{*-1}_{.25}\right] \tag{3.19}$$

in for $G = 444$ in Table 1. While this has good empirical coverage of 50% the coverage for the complementary intervals illustrates significant asymmetry. Specifically, ideal coverage probabilities for

$$\left(-\infty, \bar{d}_{r_G} - s_{r_G}\frac{\sqrt{2}}{\sqrt{n}}F^{*-1}_{.75}\right) \text{ and } \left(\bar{d}_{r_G} - s_{r_G}\frac{\sqrt{2}}{\sqrt{n}}F^{*-1}_{.25}, \infty\right) \tag{3.20}$$

are 25% but empirical coverage figures are asymmetric with empirical coverage of 39% and 11% respectively. For $G = 10$ these subintervals are less asymmetric with coverage of 28% and 23%, respectively. A more complete examination of

Nominal Interval	Bootstrap Coverage %	t-statistic Coverage %	Average Bootstrap Length	Average t-statistic Length
$25^{th} - 75^{th}$	48%	50%	0.52	0.51
$10^{th} - 90^{th}$	79%	78%	1.00	1.00
$5^{th} - 95^{th}$	89%	89%	1.32	1.27
$2.5^{th} - 97.5^{th}$	94%	94%	1.61	1.54

TABLE 3. Confidence Interval Characteristics for μ_{r_G} with n=14, G=10, Effect Sizes are $\{3, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$, Variables Independent, 1000 simulations

asymmetry for these two and other estimation methods (basic bootstrap and bias-corrected bootstrap) are given in supplementary material on the website.

A third set of simulations (see Table 3) was run to show when overestimation is not a problem the bootstrap approach yields coverage and confidence interval lengths comparable to those produced by the naïve approach. In these simulations one of the 10 effect sizes was chosen to be 3 and the other 9 were set to 0. In all 1000 simulations the variable with the large effect size generated the largest $t-$statistic and in this case there was no overestimation problem.

From the data in Tables 1, 2, and 3 some generalizations may be drawn. First, the naïve estimate often performs badly – particularly as the number of variables/genes grows. The coverage probabilities in Table 1 give some indication of how poorly the common naïve approach performs under circumstances that may not be atypical in a microarray context. Table 2 shows problems still remain for

the naïve approach when the number of variables is more manageable. In Table 3 one sees that in some circumstances when there is little overestimation the bootstrap and naïve approaches perform appropriately and similarly.

These tables also indicate that while the bootstrap is an improvement, it is not perfect. This evidence that in small samples some substantial underestimation still exists may call into question whether the bootstrap is appropriate even when the sample sizes are larger. Appendix A provides a justification for the procedure in an asymptotic sense.

4. Factors That Contribute to Bias

Tables 1, 2, and 3 indicate varying degrees of bias as illustrated by the performance of the coverage characteristics of the naïve $t-$statistic estimator. Some of the factors that are most important in determining the extent of the problem are 1) n, the sample size, 2) G, the number of variables tested, and 3) the distribution of the true effect sizes. Some would also include the level of $p-$value threshold that is used to declare significance (e.g. Sun and Bull, 2005) but attention here will be focused upon the first 3 factors. While truncating the distribution of $\tau^*_{r^*_G}$ will change its characteristics, the bias principally arises from multiple comparisons rather than this truncation.

A more complete derivation of how these factors influence the bias is given in Appendix B – here some of the conclusions are given. All else equal, 1) a smaller sample size will be associated with more bias on average, 2) as the distance between the most positive effect sizes declines, the more bias (for estimating μ_{r_G})

can be expected, and 3) the effect of increasing the number of genes is more ambiguous, depending upon the distribution of effect sizes.

Below is illustrated the idea that bias tends to occur when there is more 'competition' for the gene that can produce the most extreme $t-$statistic. In other words, those situations when a large number of genes could conceivably produce the largest $t-$statistic in repeated sampling from the underlying distribution are associated with bias. To get a sense for why this should be the case we define the terms $r_1^0, r_2^0, \ldots, r_G^0$ to satisfy

$$\frac{\mu_{r_1^0}}{\sigma_{r_1^0}} \leq \frac{\mu_{r_2^0}}{\sigma_{r_2^0}} \leq \cdots \leq \frac{\mu_{r_G^0}}{\sigma_{r_G^0}} \tag{4.21}$$

so the $r_1^0, r_2^0, \ldots, r_G^0$ order the true effect sizes (recall $r_1, r_2, \ldots, r_G$ order the observed effect sizes).

Suppose in a particular sample at hand the largest $t-$statistic is generated by the variable with the 10^{th} largest effect size, i.e. $r_G = r_{G-9}^0$. Since $t_{r_{G-9}^0}$ exceeds all other $t-$statistics this implies that

$$\tau_{r_{G-9}^0} > \tau_{r_j} + \frac{\sqrt{n}}{\sqrt{2}}\left(\frac{\mu_j}{s_j} - \frac{\mu_{r_{G-9}^0}}{s_{r_{G-9}^0}}\right) \text{ for all } j = r_{G-8}^0, r_{G-7}^0, \ldots, r_{G-1}^0, r_G^0 \tag{4.22}$$

where we have decomposed the $t-$statistic as was shown in (2.4). To the extent that the s terms approximate the σ terms we see

$$\frac{\sqrt{n}}{\sqrt{2}}\left(\frac{\mu_j}{s_j} - \frac{\mu_{r_{G-9}^0}}{s_{r_{G-9}^0}}\right) \approx \frac{\sqrt{n}}{\sqrt{2}}\left(\frac{\mu_j}{\sigma_j} - \frac{\mu_{r_{G-9}^0}}{\sigma_{r_{G-9}^0}}\right) \geq 0 \text{ for } j = r_{G-8}^0, \ldots, r_G^0 \tag{4.23}$$

and the τ terms all have an unconditional $t-$distribution centered about 0. Consequently, for $\tau_{r_{G-9}^0}$ to satisfy the inequalities expressed in (4.22) it is likely $\tau_{r_{G-9}^0}$

is positive if most of the inequalities in (4.23) are strict. Given that

$$\tau_{r^0_{G-9}} = \frac{\sqrt{n}\left(\bar{d}_{r^0_{G-9}} - \mu_{r^0_{G-9}}\right)}{\sqrt{2}s_{r^0_{G-9}}} \tag{4.24}$$

we see a positive value of $\tau_{r^0_{G-9}}$ is associated with $\bar{d}_{r^0_{G-9}}$ overestimating $\mu_{r^0_{G-9}}$. Thus, occasions in which genes other than r^0_G are associated with r_G are times when bias is more likely to occur.

5. Conclusion

There exists a bias related to multiple comparisons that arises because, in the simple models given here, the observed $t-$statistics, t_j, are partially composed of a random component τ_j, a variable with a true t-distribution regardless of the truth of the null hypothesis. When one chooses the variable with large observed effect size, it is more likely there exists a large random component that leads to an overestimation of the associated difference. This particular problem is not alleviated by corrections made to address the number or proportion of Type I errors. While multiple comparisons corrections and false discovery rate approaches affect the choice of which variables may reflect significant changes, they do not address distortions in the associated magnitudes of change. Such problems become important when 1) an estimate of effect size is used to power a follow-up study, 2) comparing results across different studies and finding discrepancies in the strength of those variables showing greatest differences, 3) contemplating further action based on initial study (e.g. follow-up fine-mapping study). Further, the traditional confidence intervals may be badly biased in such circumstances.

The hypothesis tests presented here are particularly simple – two-group $t-$tests. More sophisticated discrimination methods may be less prone to these problems – in particular those types of methods that pool error information across genes (e.g. Limma approach of Smyth, 2004) may be associated with less overestimation because there is less chance the s_j terms in (2.4) are strongly underestimated. The simple tests used here facilitated the use of simulations – especially those that employed two levels of bootstrap resampling. However, it is likely that the general phenomenon of overestimation and bias are present, though perhaps muted, when other methods or statistics are used to detect differences.

The examples within this paper are microarray related; however as Table 2 demonstrates, the phenomenon is present in much more widespread applications where the small sample size/many test problem is less pronounced. The results indicate the bootstrap approach is able to distinguish instances when such bias does an does not exist. Discussion of Tables 1 and 2 indicated that the bootstrap approach used here could likely be improved – perhaps by considering transformations or alternative bootstrap methods. Other avenues of research may include non-bootstrap approaches to the problem; however there is heavy dependence upon the entire distribution of true effect sizes and as such any parametric approach would need to allow a great deal of flexibility in this respect.

References

Bickel, P.J. and Freedman, D.A.(1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196-1217.

Efron, B. and Tibshirani, R.J. (1998). *An Introduction to the Bootstrap.* Boca Raton, Florida: Chapman and Hall/CRC.

Irizarry et al. (2004). Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2(5)**, 345-350.

Siegmund, D. (2002). Upward bias in estimation of genetic effects. *American Journal of Human Genetics* **71**, 1183-1188.

Smythe, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**,(1):3.

Sun, L. and Bull, S.B. (2005). Reduction of selection bias in genome-wide studies by resampling. *Genetic Epidemiology* **28**, 352-367.

Westfall, P.H. and Young, S.S. (1993). *Resampling-based Multiple Testing.* New York: Wiley.

Appendix A: Justification of the Bootstrap

A classic example of bootstrap failure occurs when one considers the maximum of a series of observations and one may question whether something similar is occurring in this context. Let $Z_1, Z_2, \ldots, Z_n$ be independently and identically distributed uniform random variables on the interval given by $[0, \theta]$ with θ unknown and suppose confidence intervals for θ are sought. Let $T = max\{Z_1, \ldots, Z_n\}$. Then it may be shown that

$$Q = n(\theta - T)/\theta \rightarrow \text{ standard exponential distribution.} \qquad \text{(A.1)}$$

Now consider an obvious bootstrap procedure to elicit information regarding θ: let $t =$ observed T, T^* denote the maximum value observed in a bootstrap sample of $Z_1, Z_2, \ldots, Z_n$ and $Q^* = n(t - T^*)/t$. For the bootstrap procedure to work in any meaningful way it should be the case that Q^* also converge to a standard exponential distribution. However it is easy to see that if $Z_{(n)}$ is in the bootstrap sample we have $Q^* = 0$. Further, bootstrap sampling with replacement implies

$$Prob(Q^* = 0) = Prob(Z_{(n)} \text{ in bootstrap sample }) = 1 - \left(\frac{n-1}{n}\right)^n \rightarrow .632 \quad \text{(A.2)}$$

Consequently the limiting distribution of Q^* contains a point mass at 0 with probability .632 so clearly the limiting distribution of Q^* cannot be exponential.

This problem does not arise in the context presented here (investigating the distribution of τ_{r_G}) as the maximum is taken over G variables, the number of which is fixed. In the failing example given above the maximum is taken over n observations – an index that increases asymptotically.

To give a more formal justification of why the bootstrap is appropriate for this overestimation problem we demonstrate that

$$\tau_{r_G} = \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}\, s_{r_G}} \text{ and } \tau^*_{r^*_G} = \frac{\sqrt{n}\left(\bar{d}^*_{r^*_G} - \bar{d}_{r^*_G}\right)}{\sqrt{2}\, s_{r^*_G}} \qquad \text{(A.3)}$$

have the same asymptotic distribution, i.e. the bootstrap procedure works in at least an asymptotic sense. As in (4.21) let $r^0_1, r^0_2, \ldots, r^0_G$ order the true effect sizes. To proceed further we will focus upon τ_{r_G} in a simple and the common situation that there exists a single variable/gene with maximal true effect size greater than

all other effect sizes,

$$\text{i.e. } \frac{\mu_{r_G^0}}{\sigma_{r_G^0}} > \frac{\mu_j}{\sigma_j} \text{ for all } j \neq r_G^0. \tag{A.4}$$

Recall from (2.4) that

$$t_j = \tau_j + \frac{\sqrt{n}\,\mu_j}{\sqrt{2}\,s_j} \text{ where } \tau_j = \frac{\sqrt{n}\left(\bar{d}_j - \mu_j\right)}{\sqrt{2}s_j} \tag{A.5}$$

so for $j \neq r_G^0$ we have

$$Pr\left[t_{r_G^0} > t_j\right] = Pr\left[\tau_{r_G^0} + \frac{\sqrt{n}\,\mu_{r_G^0}}{\sqrt{2}\,s_{r_G^0}} > \tau_j + \frac{\sqrt{n}\,\mu_j}{\sqrt{2}\,s_j}\right]. \tag{A.6}$$

From (A.4) and the realization that the τ terms are governed by a $t-$distribution it is clear the term $\frac{\sqrt{n}\,\mu_{r_G^0}}{\sqrt{2}\,s_{r_G^0}}$ will dominate the inequality above so that with probability 1 $t_{r_G^0}$ will exceed t_j as $n \to \infty$. As this holds for all $j \neq r_G^0$ this implies $r_G \to r_G^0$ with probability 1. From similar reasoning one can deduce that this behavior also occurs in the bootstrap sample so that $r_G^* \to r_G \to r_G^0$ with probability 1 (assuming the number of bootstrap replications increases to ∞ with n). So, because $r_G^* \to r_G^0$ this implies

$$\frac{\sqrt{n}\left(\bar{d}^*_{r^*_G} - \bar{d}_{r^*_G}\right)}{\sqrt{2}s^*_{r^*_G}} - \frac{\sqrt{n}\left(\bar{d}^*_{r_G^0} - \bar{d}_{r_G^0}\right)}{\sqrt{2}s^*_{r_G^0}} \to 0 \text{ with probability 1, or} \tag{A.7}$$

$$\tau^*_{r^*_G} \to \tau^*_{r_G^0} = \frac{\sqrt{n}\left(\bar{d}^*_{r_G^0} - \bar{d}_{r_G^0}\right)}{\sqrt{2}s^*_{r_G^0}}. \tag{A.8}$$

For a fixed index, e.g. r_G^0, it is well known that under general conditions the bootstrap has appropriate asymptotic behavior (Bickel and Freedman, 1981), i.e.

$$\tau^*_{r_G^0} = \frac{\sqrt{n}\left(\bar{d}^*_{r_G^0} - \bar{d}_{r_G^0}\right)}{\sqrt{2}s^*_{r_G^0}} \text{ and } \tau_{r_G^0} = \frac{\sqrt{n}\left(\bar{d}_{r_G^0} - \mu_{r_G^0}\right)}{\sqrt{2}s_{r_G^0}} \tag{A.9}$$

both converge weakly to a Gaussian distribution. From (A.8) and (A.9) this means

$$\tau^*_{r^*_G} = \frac{\sqrt{n}\left(\bar{d}^*_{r^*_G} - \bar{d}_{r^*_G}\right)}{\sqrt{2}s^*_{r^*_G}} \text{ converges in distribution to } N(0,1). \tag{A.10}$$

Similarly, because $r_G \to r^0_G$ with probability 1 this means

$$\tau_{r_G} \text{ converges with probability 1 to } \tau_{r^0_G} = \frac{\sqrt{n}\left(\bar{d}_{r^0_G} - \mu_{r^0_G}\right)}{\sqrt{2}s_{r^0_G}} \tag{A.11}$$

where the right hand term has a $t-$distribution and hence also converges to a $N(0,1)$ distribution. Consequently it has been demonstrated that both $\tau^*_{r^*_G}$ and τ_{r_G} have the same limiting distribution and thus the use of the bootstrap is justified in an asymptotic sense.

Appendix B: Calculation of the Bias

Here an effort is made to sketch the degree of bias that may be expected and link this magnitude to some factors such as sample size, distribution of true effect sizes, and the number of tests. Simplifying assumptions will be employed as necessary. Here attention will be focused upon τ_{r_G} though analogous results hold for τ_{r_1}. One may write

$$E\left[\tau_{r_G}\right] = \sum_{j=1}^{G} E\left[\tau_{r_G} | r_G = j\right] P\left[r_G = j\right]. \tag{B.1}$$

$$\text{Then } E\left[\tau_{r_G} | r_G = j\right] P\left[r_G = j\right] = \left(\int \tau f_{\tau_j | r_G = j}(\tau)\, d\tau\right) P\left[r_G = j\right] \tag{B.2}$$

$$= \frac{\int \tau f_{\tau_j}(\tau, r_G = j)\, d\tau}{P\left[r_G = j\right]} P\left[r_G = j\right] \tag{B.3}$$

$$= \int \tau f_{\tau_j}(\tau, r_G = j)\, d\tau \tag{B.4}$$

where $f_{\tau_j|r_G=j}$ is the conditional distribution of τ_j given $r_G = j$ and $f_{\tau_j}(\tau, r_G = j)$ describes the joint distribution of τ_j and the event $r_G = j$.

$$\text{Now } r_G = j \text{ if and only if } t_j > \max_{k \neq j} t_k \tag{B.5}$$

$$\text{if and only if } \tau_{r_j} > \max_{k \neq j} \left(\tau_k + \frac{\sqrt{n}}{\sqrt{2}} \left(\frac{\mu_k}{s_k} - \frac{\mu_j}{s_j} \right) \right). \tag{B.6}$$

To simplify we will approximate the s_j and s_k terms by σ_j and σ_k. Then we obtain

$$E[\tau_{r_G}] = \sum_{j=1}^{G} \int \tau f_{\tau_j}(\tau, r_G = j)\, d\tau \tag{B.7}$$

$$= \sum_{j=1}^{G} E\left[\int_{M_{-j}}^{\infty} \tau f_\tau(\tau)\, d\tau \right] \tag{B.8}$$

$$\text{where } M_{-j} = \max_{k \neq j} \left(\tau_k + \frac{\sqrt{n}}{\sqrt{2}} \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_j}{\sigma_j} \right) \right) \tag{B.9}$$

f_τ denotes a $t-$distribution with $2n-2$ degrees of freedom and the expectation in (B.8) is necessary because M_{-j} contains random elements τ_k. To simplify further we will approximate f_τ by a standard Gaussian distribution and assume the G variables are independent. Then we may rewrite terms as

$$E[\tau_{r_G}] \approx \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{G} E\left[e^{\frac{-M_{-j}^2}{2}} \right] \tag{B.10}$$

From (B.10) one sees that bias is inversely related to the absolute value of the M_{-j} terms. Some consequences of this derivation are as follows.

Consider the effect of increasing the sample size holding all else constant. It is worthwhile to examine M_{-j} for the case when $j = r_G^0$ and $j \neq r_G^0$ separately where we assume only one variable (with index r_G^0) has the most positive effect

size, i.e. there are no ties. Then

$$\lim_{n\to\infty} M_{-j} = \lim_{n} \max_{k\neq j} \left(\tau_k + \frac{\sqrt{n}}{\sqrt{2}} \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_j}{\sigma_j} \right) \right) \tag{B.11}$$

$$= -\infty \text{ if } j = r_G^0 \tag{B.12}$$

$$= \infty \text{ if } j \neq r_G^0. \tag{B.13}$$

In either case we have that $M_{-j}^2 \to \infty$ so from (B.10) one sees $E[\tau_{r_G}] \approx 0$.

The case for expanding the differences among effect sizes is similar – at least for the simplified example below. For a given pattern of effect sizes among the G variables (again with no ties for the most extreme effect size), consider a new pattern of effect sizes given by multiplying each original effect by a constant $c > 0$. Then if r_G^0 designates the most positive effect size

$$\lim_{c\to\infty} M_{-j} = \lim_{c} \max_{k\neq j} \left(\tau_k + c\frac{\sqrt{n}}{\sqrt{2}} \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_j}{\sigma_j} \right) \right) \tag{B.14}$$

$$= -\infty \text{ if } j = r_G^0 \tag{B.15}$$

$$= \infty \text{ if } j \neq r_G^0. \tag{B.16}$$

Consequently the same conclusion of no bias follows. If one reverses the limiting action of c so that $c \to 0$ from above then

$$\lim_{c\downarrow 0} M_{-j} = \max_{k\neq j} \tau_k \tag{B.17}$$

where the τ_k are identically and independently distributed $t-$statistics and the bias is then positive. This situation corresponds to the situation of no variables showing differential expression.

The case for increasing G, the number of variables is less clear cut as it depends upon the combination of effect sizes. As an example, suppose originally, all true effect sizes are equal (either zero or not) – then there will be non-trivial overestimation. If one additional variable is added that has a much larger effect size then as demonstrated in Table 3 this may reduce or eliminate the bias. Then if an additional variable is added with the same larger effect size some degree of overestimation will then by reintroduced. Empirically it seems that adding variables with effect sizes at or near the size of the largest preexisting effect sizes exaggerates the bias effects for μ_{r_G}. In terms of figuring the change of M_{-j} terms as above there is more ambiguity as some terms M_{-j}^2 terms will likely increase, others decrease, and some new terms will be introduced.