

Multiple Comparisons Distortions of Parameter Estimates

BY NEAL O. JEFFRIES

MSC 1430, 10 Center Drive

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. An associated important consequence is that confidence intervals constructed using standard distributions may be badly biased. A bootstrap approach is used to estimate and adjust for 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 are 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.

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 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.

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 $i \in \{1 \dots n\}$ and $j \in \{1 \dots G\}$. Let $d_j = \bar{x}_j - \bar{y}_j$, σ_j denote the common standard deviation for X_{ij} and Y_{ij} and s_j denote the pooled estimate of σ_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

$$t_j = \frac{\sqrt{n}(\bar{x}_j - \bar{y}_j)}{\sqrt{2} s_j} = \frac{\sqrt{n} \bar{d}_j}{\sqrt{2} s_j} \quad (2.1)$$

$$= \frac{\sqrt{n} (\bar{d}_j - \mu_j)}{\sqrt{2} s_j} + \frac{\sqrt{n} \mu_j}{\sqrt{2} s_j} \quad (2.2)$$

$$= \tau_j + \frac{\sqrt{n} \mu_j}{\sqrt{2} s_j} \text{ where } \tau_j = \frac{\sqrt{n} (\bar{d}_j - \mu_j)}{\sqrt{2} s_j}. \quad (2.3)$$

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 (2.3) 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. 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, \dots, r_G denote the indices associated with the smallest to largest t -statistics so that $t_{r_1} \leq t_{r_2} \leq \dots \leq t_{r_G}$. While τ_j has a t -distribution marginally, this is generally no longer the case when we condition on j corresponding to an extreme τ_j value. It is impossible to know in general the distribution of

$$\tau_{r_1} = \frac{\sqrt{n} (\bar{d}_{r_1} - \mu_{r_1})}{\sqrt{2} s_{r_1}} \text{ or } \tau_{r_G} = \frac{\sqrt{n} (\bar{d}_{r_G} - \mu_{r_G})}{\sqrt{2} s_{r_G}} \quad (2.4)$$

and hence the degree to which \bar{d}_{r_1} underestimates μ_{r_1} (or \bar{d}_{r_G} overestimates μ_{r_G}). Situations in which $E[\tau_{r_G}] > 0$ are consistent with $E[\bar{d}_{r_G}] > \mu_{r_G}$ and in this way reflect bias in the estimated mean or effect size. As an aside it should be noted that the distribution of τ_{r_G} may also be driven by small values of the s_j terms.

In terms of confidence intervals, naïve application of a t -test distribution can be misleading, e.g. the interval for μ_{r_G} given by $\bar{d}_{r_G} \pm t_{.975, 2n-2}^{-1} \frac{\sqrt{2} s_{r_G}}{\sqrt{n}}$ is likely to systematically overestimate μ_{r_G} ; this will be demonstrated in simulations below.

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

$$\tau_{r_G} = \frac{\sqrt{n} (\bar{d}_{r_G} - \mu_{r_G})}{\sqrt{2} s_{r_G}}. \quad (2.5)$$

The bootstrap techniques popularized by Efron (e.g. Efron, 1979) will be used to develop alternative confidence intervals that compensate for the bias described above. To proceed one first constructs a bootstrap sample from the X_{ij}, Y_{ij} by sampling individuals with replacement from these two-group data in a stratified manner, i.e. sampling from the $X_i = \{X_{i1}, \dots, X_{iG}\}$ and $Y_i = \{Y_{i1}, \dots, Y_{iG}\}$ separately. One samples each individual's entire data, not the individual variables separately. 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 $r_1^*, r_2^*, \dots, r_G^*$ order the t statistics, t_j^* , i.e.

$$t_{r_1^*}^* \leq t_{r_2^*}^* \leq \dots \leq t_{r_G^*}^*. \text{ Then compute } \tau_{r_G^*}^* = \frac{\sqrt{n} (\bar{d}_{r_G^*}^* - \bar{d}_{r_G^*})}{\sqrt{2} s_{r_G^*}^*} \quad (2.6)$$

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

$$\tau_{r_G} = \frac{\sqrt{n} (\bar{d}_{r_G} - \mu_{r_G})}{\sqrt{2} s_{r_G}} \text{ and } \tau_{r_G^*}^* = \frac{\sqrt{n} (\bar{d}_{r_G^*}^* - \bar{d}_{r_G^*})}{\sqrt{2} s_{r_G^*}^*} \quad (2.7)$$

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.

$$\mu_{r_G} \text{ satisfying } \left[F_{.025}^{*-1} \leq \frac{\sqrt{n} (\bar{d}_{r_G} - \mu_{r_G})}{\sqrt{2} s_{r_G}} \leq F_{.975}^{*-1} \right] \quad (2.8)$$

where \bar{d}_{r_G} and s_{r_G} are observed from the original data. For this procedure a second order bootstrap may also be employed to improve the approximation of F^{-1} by F^{*-1} where F^{-1} denotes the cdf of τ_{r_G} . 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://data.ninds.nih.gov/Jeffries/multcomps/index.htm>.

Table 1 indicates how this approach works in terms of confidence intervals for μ_{r_G} in a simulation context. Here $n = 14$ in each group and $G=444$. Each variable has an effect size chosen from the set $\{2/444, 4/444, \dots, 886/444, 888/444\}$. The idea is that this may correspond to about 2% of approximately 22,200 genes being differentially expressed and the genes/variables are constructed as independent for convenience. Rather than perform tests on all 22,200 variables attention was restricted to those differentially expressed to make the second order bootstrap computations feasible. 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 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.

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 $(0, 2]$, 1000 simulations

A second set of simulations was run with a much smaller number of variables/genes. Here the interest is in evaluating overestimation when a more modest number of comparisons are involved. Here $n = 14$, there are $G = 10$ independent variables, and 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.

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

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 $(0, 1]$, 1000 simulations

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 $\{3, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$, 1000 simulations

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 reduced. In Table 3 one sees that in

some circumstances when there is little overestimation the bootstrap and naïve approaches perform appropriately and similarly.

3. CONCLUSION

In multiple testing situations, when one examines the variable/test with largest observed effect size, it is more likely there exists a large random component that leads to an overestimation of the associated effect size. This bias is potentially present whenever attention is focused upon the most extreme results among a number of tests. The 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, or 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 results indicate the bootstrap approach may be able to distinguish instances when such bias does and does not exist and associated confidence intervals are less prone to overestimation than those derived from a naïve t -statistic approach.

Supplemental materials at <http://data.ninds.nih.gov/Jeffries/multcomps/index.htm> provide 1) a more complete description of the two-stage bootstrap approach

used here, 2) asymptotic justification for applying the bootstrap in these situations, and 3) analysis showing which factors exacerbate this bias. Mitigating factors are increased sample size and greater distinction among the most extreme effect sizes.

REFERENCES

- EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1-26.
- SIEGMUND, D. (2002). Upward bias in estimation of genetic effects. *American Journal of Human Genetics* **71**, 1183-1188.
- SUN, L. AND BULL, S.B. (2005). Reduction of selection bias in genome-wide studies by resampling. *Genetic Epidemiology* **28**, 352-367.