

Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test

Katherine S. Pollard
Division of Biostatistics
University of California
School of Public Health
Earl Warren Hall #7360
Berkeley, CA 94720-7360

Mark J. van der Laan
Division of Biostatistics
University of California
School of Public Health
Earl Warren Hall #7360
Berkeley, CA 94720-7360

Keywords: multiple testing, strong control, permutation, bootstrap, asymptotics, gene expression

1 Introduction

Gene expression studies produce data with which inferences can be made for thousands of genes simultaneously, allowing researchers to answer questions such as “Which genes are significantly differently expressed between two (or more) conditions?” or “Which genes have a significant association with an outcome or covariate?”. In order to make statements about the statistical significance of thousands of genes at once, it is essential to appropriately account for multiple tests. Multiple testing methods are hypothesis testing procedures designed to simultaneously test $p > 1$ hypotheses while controlling an error rate. Traditional approaches to the multiplicity problem are reviewed by [1]. More recent developments in the field include resampling methods [2], step-wise procedures, and the false discovery rate [3]. Standard practice in multiple testing with gene expression data is to use t-statistics as test statistics and to control a type I error rate under the permutation distribution. In this paper, we revisit the rationale behind such choices and suggest situations in which alternatives are more sensible.

2 Multiple Hypothesis Testing Procedures

In this paper, we focus on the two sample problem and refer the reader to [4] for more general results.

2.1 Data and Null Hypotheses

Let X_1, \dots, X_n be i.i.d. $X \sim P \in \mathcal{P}$, where \mathcal{P} is a model and X is a p -dimensional vector of gene expression measurements, plus possibly covariates or outcomes. Typically, the dimension of the data p far exceeds the sample size n . Suppose that the $n = n_1 + n_2$ samples come from two different populations, with n_1 from population 1 and n_2 from population 2. We can think of the data as (X_i, L_i) , where $L_i \in \{1, 2\}$ is a label indicating subject i 's group membership. Let $\mu_{1,j}$ and $\mu_{2,j}$ denote the means of gene j in populations 1 and 2, respectively. Suppose we are interested in testing for every gene $j = 1, \dots, p$ if mean expression is the same in the two populations:

$$H_{0,j} : \mu_j \equiv \mu_{2,j} - \mu_{1,j} = 0. \quad (1)$$

More generally, one could also test if the difference in mean expression is equal to some null value: $H_{0,j} : \mu_j = \mu_j^0$. Other gene-specific parameters of interest are differences in medians or regression parameters β_j from a model

$E(Y) = m(X_j | \beta_j)$ for every gene $j = 1, \dots, p$ and an outcome Y .

2.2 Models

Consider the following data generating models for this two sample problem:

1. \mathcal{P}_1 : $X|L = 1 \sim P_1$ and $X|L = 2 \sim P_2$, where P_1, P_2 can be arbitrary distributions,
2. \mathcal{P}_2 : $X|L = 1 \sim P_0(\cdot - \mu_1)$ and $X|L = 2 \sim P_0(\cdot - \mu_2)$, for a common non-parametric distribution P_0 with mean zero.

Model \mathcal{P}_2 makes a much stronger assumption, specifically that under the null hypotheses, the data are identically distributed in the two populations. If we were testing the hypothesis $H_0 : P_1 = P_2$, then this would clearly be a good choice of model, but it may be a poor choice for testing Equations (1). Other choices of models, which might be more parametric, could also be considered.

2.3 Test Statistics

Let μ_{jn} be an estimator of the difference in means $\mu_j = \mu_j(P)$ based on X_1, \dots, X_n , $j = 1, \dots, p$. Typical choices of test statistics include:

$$D_{jn} = \mu_{jn} - 0 = \bar{X}_{2,j} - \bar{X}_{1,j},$$

$$T_{jn} = \frac{\mu_{jn} - 0}{sd(\mu_{jn})} = \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}_{1,j}^2/n_1 + \hat{\sigma}_{2,j}^2/n_2}}.$$

2.4 Null Distributions

Multiple testing procedures allow one to compare the observed test statistics to a test statistic null distribution in order to assess whether or not there is sufficient evidence to reject any of the null hypotheses and declare the corresponding genes significantly differently expressed between the two populations. It is interesting to note that the asymptotically correct test statistic null distribution can be viewed as the Kullback-Leibler projection of

the asymptotic distribution of the test statistics onto the space of multivariate distributions with mean zero [4]. In practice, we use an estimated test statistic null distribution, because we do not know the underlying data generating distribution (which determines the correct test statistic null distribution). We propose the following simple bootstrap estimator. Other estimators are discussed and compared in [4].

2.4.1 Bootstrap Approach

Let \tilde{P}_n be an estimator of the true data generating distribution P (e.g.: the empirical distribution or a parametric distribution with estimated parameters). Let $\tilde{\mu}_n = \mu(\tilde{P}_n)$ be the estimated parameter (e.g.: the observed difference in means), and let $\mu_n^\#$ be the estimator μ_n but now applied to n i.i.d. copies $X_1^\#, \dots, X_n^\#$ of $X^\# \sim \tilde{P}_n$. Let $Z_n^\# = (\mu_n^\# - \tilde{\mu}_n)$. Then, the distribution of $Z_n^\#$ is a bootstrap estimated null distribution for the test statistics $D_n = (\mu_n - \mu^0) = (\mu_2 - \mu_1)$. Similarly, for the test statistics T_n , use the distribution of $Z_n^\# = (\mu_n^\# - \tilde{\mu}_n)/sd(\mu_n^\#)$, where $sd(\mu_n^\#)$ is an estimate of the standard error of $\mu_n^\#$. And for the test statistics $\sqrt{n} * (\mu_2 - \mu_1)$, use the distribution of $Z_n^\# = \sqrt{n}(\mu_n^\# - \tilde{\mu}_n)$.

2.4.2 Bootstrap Estimation Under Different Models

We now explain how one would implement the bootstrap method for the two choices of data generating model. Under both models, we first estimate μ_1, μ_2 with the sample means μ_{1n_1}, μ_{2n_2} .

- \mathcal{P}_1 : \tilde{P}_n is the empirical distribution of (X_i, L_i) , and we resample n_1 observations from population 1 and n_2 observations from population 2 *separately* to form the bootstrap samples $X_1^\#, L_1^\#, \dots, X_n^\#, L_n^\#$. Then, the empirical distribution of $Z_n^\# = (\mu_{2n_2}^\# - \mu_{1n_1}^\# - (\mu_{2n_2} - \mu_{1n_1}))$ provides an estimated joint null distribution for the test statistics $D_n = (\mu_2 - \mu_1)$.

- \mathcal{P}_2 : We first estimate P_0 by making centered observations $X_i - \mu_{1n_1}$ if $L_i = 1$ and $X_i - \mu_{2n_2}$ if $L_i = 2$ and forming the empirical distribution P_{0n} of the *combined* sample of centered observations. Then, we re-sample n_1 observations from P_{0n} and add μ_{1n_1} and n_2 observations from P_{0n} and add μ_{2n_2} to form the bootstrap samples $X_1^\#, L_1^\#, \dots, X_n^\#, L_n^\#$. Again, the empirical distribution of $Z_n^\#$ provides an estimated null distribution for $D_n = (\mu_2 - \mu_1)$.

We note that the procedure for \mathcal{P}_2 is equivalent to forming a combined empirical distribution of the X_i ($i = 1, \dots, n$) and using the distribution of the difference in the sample means when we draw n_1 samples and set $L_i = 1$ and n_2 samples and set $L_i = 2$. This is the re-sampling (with replacement) analogue of the commonly used permutation test. Hence, we will refer to the method for \mathcal{P}_1 as the non-parametric bootstrap and the method for \mathcal{P}_2 as permutations. Similar procedures can be derived for different choices of test statistics.

2.5 Error Control

We assume the reader is familiar with the distinction between type I (false positive) and type II (false negative) errors in the standard univariate setting, where the typical approach is to control the type I error rate at a pre-specified level α and compare different procedures with type I error rate α based on their type II error rates (or power). Dudoit *et al.* compare different generalizations of type I error to the multiple testing setting [5]. Let R be the total number of rejected hypotheses, let V be the (unobservable) number of false rejections, and let k be a user supplied constant. Some error rates include:

- PCER = $E(V)/p$: per-comparison error rate,
- PFER = $E(V)$: per-family error rate,
- gFWER = $Pr(V \geq k)$: generalized family-wise error rate,

Note that when $k = 1$, the gFWER is the usual family-wise error rate (FWER). These error rates are defined under the true data generating distribution P , so that they depend on which hypotheses are in fact true. In practice, we do not know which hypotheses are true since we do not know P , so we have to choose a way to compute the expectations and probabilities in the error rates (*i.e.*: choose an estimated null distribution). It is desirable to use multiple testing methods which control an error rate under the true data generating distribution, at least asymptotically. We refer to the latter as asymptotic strong control. Weak control means that the error rate is controlled under a particular choice of distribution for the test statistics.

2.6 Cut-off Rule

Given test statistics and a target error rate α , a two-sided multiple testing procedure $MT(c)$ can then be defined by: Reject $H_{0,j}$ whenever $|T_{jn}| > c_j$, $j = 1, \dots, p$ (and similarly for D_n). The vector function c is a cut-off rule such that the error rate of $MT(c)$ is α under the asymptotically correct null distribution. Since we do not know the correct null distribution, we use a vector function cut-off rule c_n based on an estimated null distribution, chosen such that if $c_n \rightarrow c$ in probability as $n \rightarrow \infty$ then asymptotically the error rate of $MT(c_n)$ is at most α (*i.e.*: $MT(c_n)$ has asymptotic strong control under the correct data generating model). In [4], we show that the bootstrap method described in Section 2.4 provides an estimated null distribution which controls the error rate asymptotically.

One particular method for computing c_n is to select a common quantile of each marginal bootstrap resampling-based test statistic null distribution. With the resampled null distribution in hand, these common quantiles can be fine-tuned to control the chosen error rate exactly under this estimated distribution. In traditional testing settings, a common threshold is used to make the testing decision for every variable. The common quantile method is

\mathcal{P}_1	$Var(D_j)$	$\frac{\sigma_{1,j}^2}{n_1} + \frac{\sigma_{2,j}^2}{n_2}$
\mathcal{P}_2		$\frac{\sigma_{1,j}^2}{n_2} + \frac{\sigma_{2,j}^2}{n_1}$
\mathcal{P}_1	$Cov(D_1, D_2)$	$\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}$
\mathcal{P}_2		$\frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}$

Table 1: Formulas for the variance and covariance of the difference in means statistic under two different models.

a generalization of this approach, which corresponds with a common threshold *only* if the marginal distributions have identical tail probabilities, which is not the case in many applications.

3 When is the Permutation Null Distribution Correct?

3.1 Covariance

For simplicity, we suppose that $p = 2$, but note that conclusions about the covariance of two genes can be applied to any pairwise covariance when p is much larger. For gene j , denote the variance of X_i by $\sigma_{1,j}^2$ in population 1 and by $\sigma_{2,j}^2$ in population 2. Let ϕ_1 be the covariance between the two genes in population 1 and ϕ_2 be the covariance between the two variables in population 2. We have derived formulas for the variance of D_j ($j = 1, 2$) and the covariance of the two test statistics D_1, D_2 under both models \mathcal{P}_1 and \mathcal{P}_2 (Table 1, derivations in [4]).

These expressions show us that under most values of the underlying parameters, the non-parametric bootstrap and permutation distributions of D_j are not equivalent. But, when (i) $n_1 = n_2$ or (ii) $\sigma_{1,j}^2 = \sigma_{2,j}^2 \equiv \sigma_j^2$ ($j = 1, 2$) and $\phi_1 = \phi_2 \equiv \phi$, then they are the same. Thus, when a study is “balanced” ($n_1 = n_2$), these results suggest that one should use the permutation distribution, because the variances and covariances are the same for both populations and estimates of these “pooled” values (which make use of all n subjects) are more efficient. Notice that if we were to use the usual stan-

dardized t-statistics $T_{jn} = (\mu_{jn} - \mu_j^0)/sd(\mu_{jn})$, despite the fact that the variances are equal under both models, the covariances are still not equivalent unless $n_1 = n_2$ or the correlation structures are the same in the two populations.

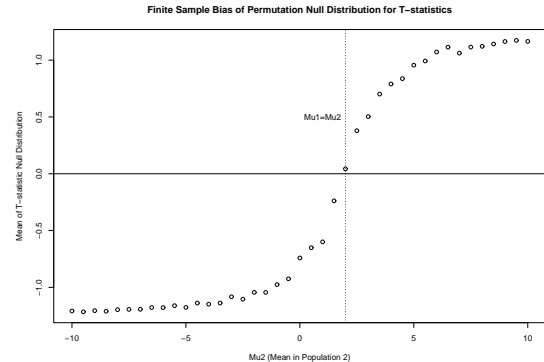


Figure 1: Mean of the permutation null distribution of the standardized two sample t-statistic for simulated data. Population 1 consists of $n_1 = 2$ subjects with observed values 1 and 3. Population 2 consists of $n_2 = 50$ subjects with observed values distributed $N(\mu_2, 0.1)$ for different choices of μ_2 . The mean of the null distribution is plotted versus μ_2 (*i.e.*: as a function of the difference in means, since μ_1 is constant). The vertical line marks where the difference in means is truly zero. The mean of the null distribution is close to zero there, but increases in magnitude with the difference in means. The mean of the null distribution should be zero for all data sets. All 1326 possible permutations were performed exactly.

3.2 Bias

We have found that bootstrap resampling-based estimated null distributions of standardized t-statistics do not have mean zero whenever $n_1 \neq n_2$, unless the observed difference in means is zero. For the permutation method, this bias depends on the observed difference in means (Figure 1), while for the non-parametric bootstrap method the bias is independent of the observed difference. This finite sample bias arises from using a variance estimate in the denominator of the t-statistics, and disappears

in simulations when the estimate is replaced by the true variance. In small, heavily unbalanced samples, one should be aware that this bias could be relatively quite large. We found that there is also a bias in the estimation of the variance of both D_n and T_n in unbalanced designs whenever the two groups have unequal observed means.

4 Simulations

We have conducted simulations to understand the performance of different multiple testing procedures, including choice of test statistics and null distribution [4]. We summarize the results here.

4.1 Choice of Test Statistic

We compared D_n and T_n based on the ease with which their null distributions can be estimated for reasonable sample sizes. First, we observed the finite sample bias of the estimated null distributions of T_n noted in Section 3, while null distributions of both test statistics had observed means close to zero when the observed difference in means between the two samples was close to zero. Second, the variance of T_n 's null distribution is usually much too large with the non-parametric bootstrap estimator (resulting in conservative error rate control). Third, whenever $n_1 \neq n_2$ the permutation estimates of the variance and correlation of the null distribution of D_n and the correlation (but not the variance) of the null distribution of T_n are far from the truth, as predicted by the formulas in Section 3. Thus, it is certainly interesting to do multiple testing with D_n in addition to T_n . We suggest that D_n may be a better choice at small sample sizes and with non-parametric data generating models, whereas T_n is often preferable with larger sample sizes or more parametric models. In other words, pivoting (*i.e.*: dividing by $sd(\mu_n)$) only helps when the estimate $sd(\mu_n)$ is close to a constant (*e.g.*: asymptotically). How fast it becomes beneficial to pivot (as $n \rightarrow \infty$) is determined by the variance of $sd(\mu_n)$, which de-

pends on (i) the data generating model (*i.e.*: model-based estimation versus non-parametric estimation) and (ii) the variance of the data.

4.2 Comparison of Null Distributions

For both D_n and T_n , we compared the finite sample performance of the two choices of test statistic null distribution estimators presented in Section 2.4: (i) non-parametric bootstrap and (ii) permutations. The most striking finding is that when $n_1 = n_2$, the permutation method performs very well even when the covariance structures are unbalanced, as predicted by the algebraic results in Section 3. The non-parametric bootstrap generally performs better for D_n than for T_n for two reasons. First, the non-parametric bootstrap method estimates $sd(\mu_{jn})$ non-parametrically. Second, ties in the resampling can result in very small estimates of $sd(\mu_{jn})$. Smoothing the empirical distribution does reduce this problem. Both of these factors contribute to the non-parametric bootstrap method producing highly variable and unrealistically large resampled t-statistics. In contrast, the permutation test statistic (which uses the data from both populations to estimate the pooled estimate $sd(\mu_{jn})$) is much less variable, so that the asymptotic results of [6] will apply. In terms of error rate control, the non-parametric bootstrap method tend to be conservative for T_n and anti-conservative for D_n , whereas the permutation method tends to be anti-conservative for both statistics (but particularly for D_n).

5 Data Analysis

We applied resampling-based multiple testing methods to a publicly available data set [7]. Expression levels of 13,412 clones (relative to a pooled control) were measured in the blood samples of 40 diffuse large B-cell lymphoma (DLBCL) patients using cDNA arrays. According to [7], the patients belong to two molecularly distinct disease groups, 21 Activated and 19 Germinal Center (GC).

5.1 Testing for a Difference in Means

Our goal was to identify clones with significantly different mean expression levels between the Activated and GC groups. We computed standardized t-statistics T_{jn} for each gene. We chose to control the usual family-wise error and compared the clones identified as having significantly different means between the two groups using several methods. Table 2 contains the multiple testing results. Based on the results of Sections 3 and 4, we believe that the permutation subset is likely to be closer to the true subset, since it makes use of a pooled variance estimate in T_n and $n_1 \approx n_2$. The non-parametric bootstrap subset is likely to be slightly smaller than the true subset.

Null Distribution	Rejections
Non-parametric bootstrap	186
Permutations	287
Student’s t-distribution	32

Table 2: Difference in Means. Number of rejected null hypotheses (out of $p = 13,412$) for three different choices of multiple testing procedure controlling the FWER at level $\alpha = 0.05$. Standardized t-statistics were used to test for a difference in means between the Activated and GC disease groups. All 32 of the genes in the t-distribution subset are in both the permutation and the bootstrap subset, and the bootstrap and permutation subsets have 156 genes in common. Data are from [7].

5.2 Testing for Association with Disease Group Using Logistic Regression

One might also be interested in testing for an association between gene expression and an outcome Y of interest, such as survival or disease group. In this case, a regression model $E(Y) = m(X_j | \beta_j)$ (*e.g.*: linear or logistic regression) is fit for every gene $j = 1, \dots, p$, producing a vector of observed regression coef-

ficients β_n which measure the association between gene expression and the outcome. The usual test statistics can be used (with $\mu_j = \beta_j$ as the parameter) to test the hypotheses $H_{0,j} : \beta_j = 0, j = 1, \dots, p$ (or more generally, $H_{0,j} : \beta_j = \beta_j^0$). The bootstrap method of Section 2.4 can then be used to estimate the test statistic null distribution, using appropriate resampled random variables (*e.g.*: $Z_n^\# = \sqrt{n}(\beta_n^\# - \beta_n)$ for test statistics $\sqrt{n}(\beta_n - 0)$).

We applied the non-parametric bootstrap method to the data set of [7], with disease group (Activated versus GC) as a binary outcome and a logistic regression model. This is an example of a case that illustrates the simplicity of the bootstrap method. Despite the fact that the outcome is not a linear function of gene expression and the error may not be independent of gene expression, the bootstrap can be applied directly without any thought about the form of the test statistic distribution. In contrast, the usual resampling-based multiple testing methods (*e.g.*: permutations or resampling residuals as proposed by [2]) do not work, because the assumptions under which they are appropriate do not hold. Table 3 contains the number of genes that were significantly associated with disease group. The finding that the number of rejected null hypotheses is the same for $k = 1, 10, 50$ is partially due to the discreteness of the resampled null distribution (with $B = 1000$ resamples). By resampling more times (*e.g.*: $B = 10000$), a sharper bound can be achieved.

6 Conclusions

We have learned that the common practice of using standardized t-statistics T_n and a permutation null distribution is not always the best multiple testing procedure. In particular, we found in our limited simulation study that the null distribution of T_n was harder to estimate than that of the difference in means D_n when $sd(\mu_n)$ was variable (*i.e.*: with a non-parametric model and/or a small sample size).

We have proposed a simple bootstrap

$k =$	1	10	50	100	200
Rejections	303	303	303	471	553

Table 3: Logistic Regression Parameters. Number of rejected null hypotheses (out of $p = 13,412$) using the non-parametric bootstrap estimated null distribution and controlling the gFWER $P(V > k)$ for different choices of k , where V is the number of false positives. The test statistics used were $\sqrt{n} * (\beta_n - 0)$. Fine-tuned common quantiles $\{c_j : j = 1, \dots, p\}$ were computed from the estimated null distribution in order to control the gFWER at level $\alpha = 0.05$. Data are from [7].

method for estimating the null distribution of test statistics based on any parameter of interest (*e.g.*: differences in means or regression parameters). Using this method, we derived a resampling null distribution for the two sample problem which is nearly equivalent to the permutation test. This choice of null distribution is correct under a model for which the data from the two populations are identically distributed. But, if one is interested in testing the equality of the means without assuming similar covariance structures in both populations, then the model \mathcal{P}_2 under which permutations are the correct null distribution may not hold. In this case, the non-parametric bootstrap (correct under model \mathcal{P}_1) is a better choice, *unless* the sample sizes are equal. Interestingly, when $n_1 = n_2$ the permutation estimator is not only correct under model \mathcal{P}_1 , but is also more efficient than the non-parametric bootstrap estimator. We illustrate the application of both methods in the analysis of data from a lymphoma gene expression study.

7 Acknowledgment

We thank Sandrine Dudoit and Peter Westfall for the insightful discussions and helpful comments. This research has been supported by a grant from the Life Sciences Informatics Pro-

gram with industrial partner biotech company Chiron Corporation.

References

- [1] Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, 1987.
- [2] P.H. Westfall and S.S. Young. *Resampling-based Multiple Testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.
- [3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stistical Society*, 57:289–300, 1995.
- [4] K.S. Pollard and M.J. van der Laan. Resampling-based multiple testing: asymptotic strong control of type i error and applications to gene expression data. Technical Report 121, Group in Biostatistics, University of California, 2002. Submitted.
- [5] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. Technical Report 110, Group in Biostatistics, University of California, 2002. Submitted.
- [6] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, 1992.
- [7] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, T.Jr. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenberger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.