BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 18 2006, pages 2244–2248
doi:10.1093/bioinformatics/btl383

Gene expression

To permute or not to permute

Yifan Huang[1], Haiyan Xu[2], Violeta Calian[3] and Jason C. Hsu[4]

[1]H. Lee Moffitt Cancer Center & Research Institute, The University of South Florida, Tampa, FL 33612, USA, [2]Department of Clinical Biostatistics, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., USA, [3]Science Institute, University of Iceland, Dunhaga 3, 107 Reykjavik, Iceland and [4]Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

Received on December 9, 2005; revised on June 14, 2006; accepted on July 6, 2006
Advance Access publication July 26, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Permutation test is a popular technique for testing a hypothesis of no effect, when the distribution of the test statistic is unknown. To test the equality of two means, a permutation test might use a test statistic which is the difference of the two sample means in the univariate case. In the multivariate case, it might use a test statistic which is the maximum of the univariate test statistics. A permutation test then estimates the null distribution of the test statistic by permuting the observations between the two samples.

We will show that, for such tests, if the two distributions are not identical (as for example when they have unequal variances, correlations or skewness), then a permutation test for equality of means based on difference of sample means can have an inflated Type I error rate even when the means are equal. Our results illustrate permutation testing should be confined to testing for non-identical distributions.

Contact: calian@raunvis.hi.is

1 INTRODUCTION

A popular technique for testing hypotheses of no effect, when the distribution of the test statistic is unknown, is to resample the data. Permutation testing is a version of this technique.

For example, to test the equality of two means $H_0^\mu : \mu_X = \mu_Y$, one might use a test statistic which is the difference of the two sample means, and estimate its null distribution by permuting the observations in the combined X and Y samples. The basis for permutation testing is if the X data are sampled from distribution P_X and the Y data from distribution P_Y, then under the null hypothesis of identical distributions $H_0^P : P_X = P_Y$ all permutations of the observations are equally probable.

We will show, using simple examples, that if the identical distributions hypothesis H_0^P is false, as for example when P_X and P_Y have unequal variances, skewness or (in the multivariate case) unequal correlations, then a permutation test for H_0^μ based on difference of sample means can have an inflated Type I error rate even when H_0^μ is true. Our results thus illustrate the appropriateness of permutation testing may depend on whether the purpose of testing is to detect differences in means, or non-identical distributions.

This purpose will depend on the intended application. Take the analysis of gene expression levels for instance. For discovering regulatory networks, it may be useful to detect groups of genes with non-identically distributed expression levels between normal and disease subjects. On the other hand, for selecting genes to train a prognostic algorithm using supervised machine learning, as in the re-analysis of the data in van't Veer *et al*. (2002) by Ein-Dor *et al*. (2005), detecting differences in gene expression levels (i.e. testing $H_0^\mu : \mu_X = \mu_Y$) would be of primary interest if the prognostic algorithm is based on differences.

2 MAIN RESULTS AND SIMULATED EXAMPLES

In this section, we examine the distribution of test statistics for the difference of two means. In the normal distribution case, we show that if the sample sizes are unequal, then a permutation test will pick up signals from unequal variances and (in the multivariate case) unequal correlations, leading potentially to an inflated Type I error rate. In the case of arbitrary distributions, we show even if the sample sizes are equal, a permutation test will pick up signals from unequal skewness and higher order cumulants, leading potentially to an inflated Type I error rate. We illustrate this for a univariate lognormal distribution.

For the univariate case, Romano (1990) showed that the difference of sample means test statistic asymptotically has the issues described above, while Janssen (1997) showed the Welch t-test is asymptotically valid. Our results are based on the exact distribution of the difference in sample mean test statistic for finite samples.

Pollard and van der Laan (2005) proposed a general multiple testing method using asymptotically linear statistics, estimating the null distribution of the test statistic by re-sampling (with replacement), centering the data or statistics appropriately. They prove that, under regularity conditions, asymptotically (as sample size approaches infinity) their method controls the probability of incorrect rejection under the true distribution. [While this error rate control is weaker than strong control of the Familywise error rate (FWER), it should suffice in practice.] In the case of testing for the equality of means when the distributions do not differ by a location shift only, their method will re-sample mean-centered data within each group of the populations to be compared. (In other situations such as testing the equality of correlations or in tests parameters from non-linear models, they re-sample uncentered data then center the test statistics distribution.) Pollard and van der Laan (2003, 2005) then show, in testing the equality of the mean vectors of two bivariate distributions, estimating the null distribution of the test statistic by re-sampling the pooled

*To whom correspondence should be addressed.

mean-centered data will result in too small a critical value, resulting in an inflated error rate, if the sample sizes are unequal and the covariances are not the same. This is very similar to our result that permutation test has inflated error rate in that situation. But our method of proof differs from theirs in that they compare the inappropriately re-sampled distribution with the asymptotic distribution of the appropriate non-parametric boot-strap distribution, whereas we compare the permutation distribution with the exact distribution for finite samples.

To gain some intuition towards what might go wrong, before considering permutation tests, let us consider the simpler situation of testing by resampling with replacement, in the case of normal distributions.

Let $X_i \overset{\text{i.i.d.}}{\sim} N(\mu_X, \sigma_X^2)$ and $Y_j \overset{\text{i.i.d.}}{\sim} N(\mu_Y, \sigma_Y^2)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. Consider testing the null hypothesis $H_0^\mu : \mu_X = \mu_Y$ using the test statistic $T = \bar{X} - \bar{Y}$. The distribution of T under the null hypothesis is

$$N\left(0, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right) \tag{1}$$

If we re-sample $X_i, i = 1, \ldots, m$ and $Y_j, j = 1, \ldots, n$, from the pooled sample $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$, recomputing T each time we re-sample, it turns out the distribution of T under the null hypothesis H_0^μ is

$$N\left(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right) \tag{2}$$

This can be seen as follows. Each re-sampled observation has chance $m/(m+n)$ of being an X with variance σ_X^2, and chance $n/(m+n)$ of being a Y with variance σ_Y^2. So the variance of $\bar{X} - \bar{Y}$ is

$$\begin{aligned}
&\frac{\frac{m}{m+n}\sigma_X^2 + \frac{n}{m+n}\sigma_Y^2}{m} + \frac{\frac{m}{m+n}\sigma_X^2 + \frac{n}{m+n}\sigma_Y^2}{n} \\
&= \sigma_X^2\left(\frac{1}{m+n} + \frac{m/n}{m+n}\right) + \sigma_Y^2\left(\frac{n/m}{m+n} + \frac{1}{m+n}\right) \\
&= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}
\end{aligned}$$

Thus, if $\sigma_X^2 \neq \sigma_Y^2$, then the re-sampled distribution (2) equals the true null distribution (1) only if $m = n$.

2.1 Effect of different correlations

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ig}) \overset{\text{i.i.d.}}{\sim} MVN_g(\mu_X, \mathbf{\Sigma}_X)$ and $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jg}) \overset{\text{i.i.d.}}{\sim} MVN_g(\mu_Y, \mathbf{\Sigma}_Y)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, where $MVN_g(\mu, \mathbf{\Sigma})$ denotes a g-dimentional multivariate normal distribution with mean vector μ and variance–covariance matrix $\mathbf{\Sigma}$.

For inference on the difference of means $\mu_X - \mu_Y$, consider the statistics

$$T_l = \frac{\bar{X}_l - \bar{Y}_l}{\sqrt{\frac{1}{m} + \frac{1}{n}}}, l = 1, \ldots, g, \tag{3}$$

where $\bar{X}_l = \sum_{i=1}^m X_{il}/m$ and $\bar{Y}_l = \sum_{j=1}^n Y_{jl}/n$. Under the null hypothesis of equality of means $H_0^\mu : \mu_X = \mu_Y$, the statistic $\mathbf{T} = (T_1, T_2, \ldots, T_g) = \frac{1}{\sqrt{\frac{1}{m}+\frac{1}{n}}}(\bar{\mathbf{X}}_g - \bar{\mathbf{Y}}_g) = \frac{1}{\sqrt{\frac{1}{m}+\frac{1}{n}}}(\bar{X}_1 - \bar{Y}_1, \ldots, \bar{X}_g - \bar{Y}_g)$ is distributed as:

$$\mathbf{T} \sim MVN_g\left(\mathbf{0}, \frac{\frac{\Sigma_X}{m} + \frac{\Sigma_Y}{n}}{\frac{1}{m} + \frac{1}{n}}\right) \tag{4}$$

However, the permutation distribution of $\mathbf{T}$ may be different.

THEOREM 2.1. *Let $m \leq n$. Under $H_0^\mu : \mu_X = \mu_Y$, the permutation distribution of $\mathbf{T} = \frac{1}{\sqrt{\frac{1}{m}+\frac{1}{n}}}(\bar{\mathbf{X}}_g - \bar{\mathbf{Y}}_g)$ is*

$$\sum_{r=0}^{m} \frac{\binom{m}{r}\binom{n}{r}}{\binom{m+n}{m}} MVN\left(\mathbf{0}, \frac{\frac{(m-r)\Sigma_X + r\Sigma_Y}{m^2} + \frac{r\Sigma_X + (n-r)\Sigma_Y}{n^2}}{\frac{1}{m} + \frac{1}{n}}\right) \tag{5}$$

The proof of Theorem 2.1 is given in Appendix A.

A direct consequence of Theorem 2.1 is that

(1) If $m = n$ even though $\mathbf{\Sigma}_X \neq \mathbf{\Sigma}_Y$, or if $m \neq n$ but $\mathbf{\Sigma}_X = \mathbf{\Sigma}_Y$, the permutation distribution of $\mathbf{T}$ coincides with the distribution under the null hypothesis (4).

(2) If $m \neq n$ and $\mathbf{\Sigma}_X \neq \mathbf{\Sigma}_Y$, the permutation and null - hypothesis distributions of $\mathbf{T}$ are different.

A common test for $H_0^\mu : \mu_X = \mu_Y$ is

$$\text{Reject } H_0^\mu \text{ if } \max_{i=1,\ldots,g} |T_i| > c.$$

To control the familywise error rate (FWER) at α, the threshold c should be chosen to be the upper α quantile of the distribution of the maximum of $|T_1|, |T_2|, \ldots, |T_g|$, where $\mathbf{T}$ is distributed according to (4).

To assess the potential effect of different correlations on computing critical value based on the permutation distribution (5) instead of the true distribution (4), we conducted a simple simulation. We sampled 10 000 datasets with $g = 50$ from the permutation distribution (4) given by Theorem 2.1, a mixture of $MVN_g(\mu_X, \mathbf{\Sigma}_X)$ and $MVN_g(\mu_Y, \mathbf{\Sigma}_Y)$ where $\mu_X = \mu_Y = 0$, $\mathbf{\Sigma}_X$ has all the diagonal elements equal to 1 and all the off-diagonal elements equal to zero, while $\mathbf{\Sigma}_Y$ has all the diagonal elements equal to 1 and all the off-diagonal elements equal to 0.9. For $m = 2$ and $n = 4$, the permutation test proves to be liberal (Fig. 1), i.e. it does not control the Type I error rate of testing $H_0^\mu : \mu_X = \mu_Y$, as the distribution of $\max_{i=1,\ldots,g}|T_i|$ sampled from the permutation distribution (5) turns out to be stochastically smaller than its distribution sampled from the true distribution (4). (If $m = 4$ and $n = 2$, then the permutation test would be conservative.)

To compare the critical values of the permutation test with the correct critical value, we drew sets of 10 000 sample data from the permutation distribution (5) and from the true test statistic distribution (4), for $g = 500, 600, \ldots, 1000$. Figure 2 shows the extent the permutation test critical values are too small for $m = 2$ and $n = 8$.

As the $\max|T|$ test is often used to test component hypotheses in multiple testing, our result has implication in multiple testing as well.

2.2 Effect of different variances

We conducted a simple simulation to show that different variances can cause permutation test for equality of means to be either liberal or conservative.

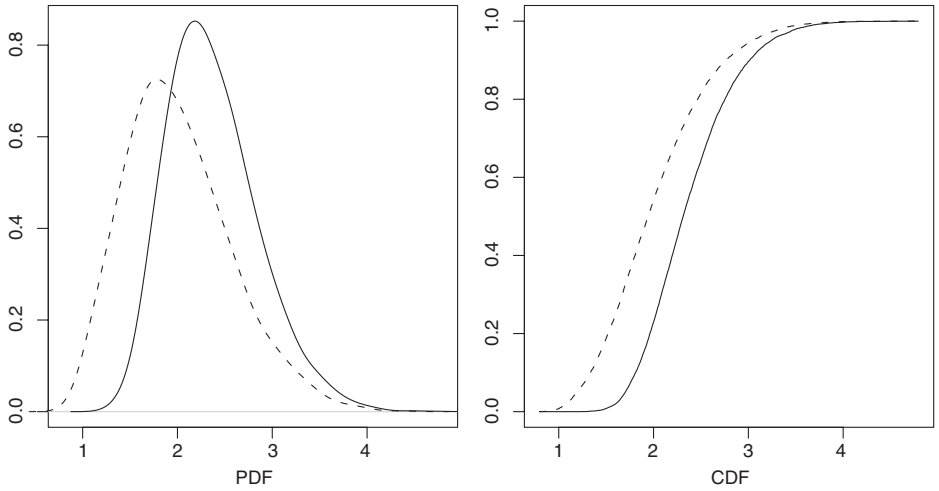

Fig. 1. Density and cumulative density plots of $\max i = 1, \ldots, g|T_i|$ for $g = 50$ with $m = 2$ and $n = 4$. The permutation distribution is stochastically smaller than the true null distribution.

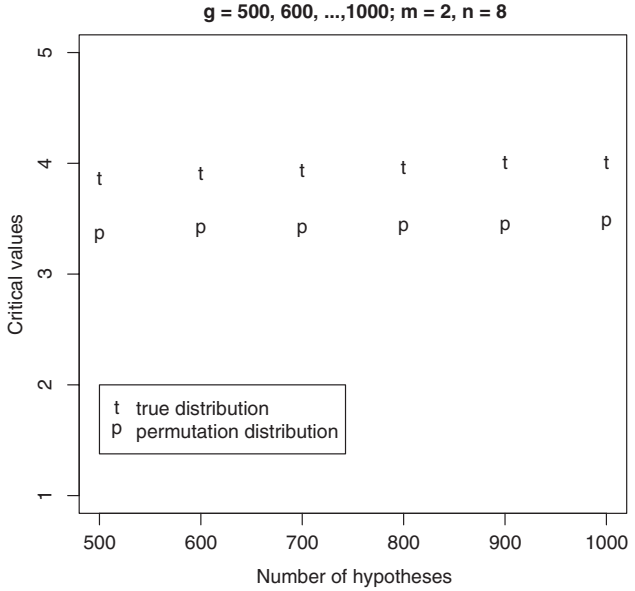

Fig. 2. Critical values of the test statistic $\max i = 1, \ldots, g|T_i|$ for $g = 500, 600, \ldots, 1000$ with $m = 2$ and $n = 8$. Critical values based on permutation distribution are smaller than the ones based on true null distribution.

Let $X_i \overset{\text{i.i.d.}}{\sim} N(\mu_X, \sigma_X^2)$ and $Y_j \overset{\text{i.i.d.}}{\sim} N(\mu_Y, \sigma_Y^2)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, and consider testing the null hypothesis $H_0^\mu : \mu_X = \mu_Y$ using the test statistic $T = (\bar{X} - \bar{Y})/\sqrt{1/m + 1/n}$.

In the first simulation study, 10 000 sets of random samples are drawn from $N(\mu_X = 0, \sigma_X^2 = 1)$ with $m = 6$ and from $N(\mu_Y = 0, \sigma_Y^2 = 3)$ with $n = 4$, independently. For each random sample, the p-value of the permutation test is computed by complete enumeration. Figure 3a compares the cumulative distribution of the simulated p-values with that of the Uniform(0, 1) distribution (a straight line along the 45° diagonal), showing that for a typical significance level the permutation test is liberal in this case. For instance, if the nominal significance level is 0.05, the actual significance level is ∼0.09.

In the second simulation study, 10 000 sets of random samples are drawn from $N(\mu_X = 0, \sigma_X^2 = 1)$ with $m = 4$ and from $N(\mu_Y = 0, \sigma_Y^2 = 3)$ with $n = 6$, independently. For each random sample, the p-value of the permutation test is computed by complete enumeration. Figure 3b compares the cumulative distribution of the simulated p-values with that of the Uniform(0, 1) distribution (a straight line along the 45° diagonal), showing that for a typical significance level the permutation test is conservative in this case. For instance, if the nominal significance level is 0.05, the actual significance level is 0.04.

2.3 Effect of differences in higher order cumulants

So far, we have shown that if $\mathbf{X}$ and $\mathbf{Y}$ are multivariate normal, then the permutation distribution of the test statistic happens to be the same as the true distribution of the test statistic if the sample sizes are equal ($m = n$). This lucky coincidence only holds when $\mathbf{X}$ and $\mathbf{Y}$ are multivariate normal, as we now show.

Suppose $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} F_X$ and $\mathbf{Y}_j \overset{\text{i.i.d.}}{\sim} F_Y$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where F_X and F_Y are arbitrary multivariate distributions. The permutation distribution of the test statistic and the true distribution of the test statistic can be described in terms of cumulants $k_a(F_X)$ and $k_a(F_Y), a = 1, 2, 3, \ldots,$ of F_X and F_Y (assuming they exist):

THEOREM 2.2. (1) *The true distribution of the test statistic* $\mathbf{T} = \bar{\mathbf{X}} - \bar{\mathbf{Y}}$ *has cumulants*

$$k_a(\mathbf{T}) = m^{1-a} k_a(F_X) + (-1)^a n^{1-a} k_a(F_Y). \tag{6}$$

(2) *For a given permutation with r elements relabeled, the distribution* (P_r) *of the test statistic* $\mathbf{T}^r = \bar{\mathbf{X}}^r - \bar{\mathbf{Y}}^r$ *obtained by permutation has cumulants*

$$k_a(\mathbf{T}^r) = k_a(\mathbf{T}) - r\left(\frac{1}{m^a} - \frac{(-1)^a}{n^a}\right)(k_a(F_X) - k_a(F_Y)) \tag{7}$$

The proof is given in Appendix B.

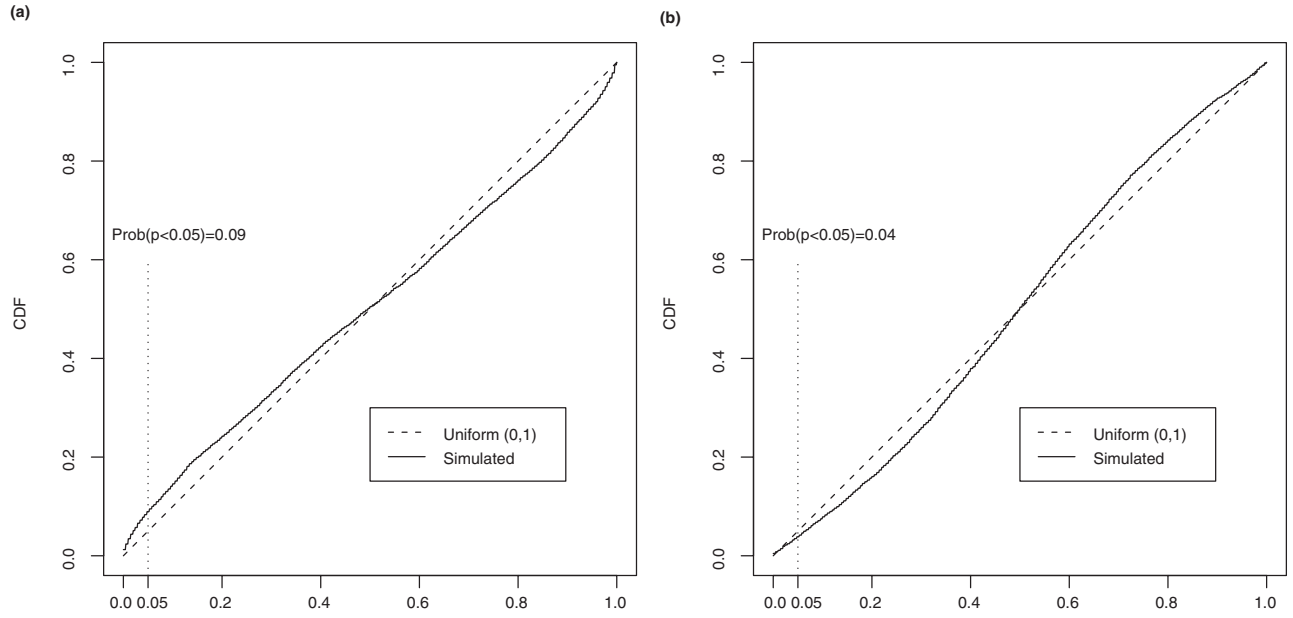

Fig. 3. (**a**) Cumulative distribution plot of simulated p-value. $m = 6$ random samples are drawn from $N(\mu_X = 0, \sigma_X^2 = 1)$ and $n = 4$ random samples are drawn from $N(\mu_Y = 0, \sigma_Y^2 = 3)$, independently. (**b**) Cumulative distribution plot of simulated p-value. $m = 4$ random samples are drawn from $N(\mu_X = 0, \sigma_X^2 = 1)$ and $n = 6$ random samples are drawn from $N(\mu_Y = 0, \sigma_Y^2 = 3)$, independently.

For the test statistic $\mathbf{T} = 1/(\sqrt{1/m + 1/n})(\bar{\mathbf{X}} - \bar{\mathbf{Y}})$, results are similar except there are constants $\left(1/(\sqrt{1/m + 1/n})\right)^a$ multiplying $k_a(F_X)$ and $k_a(F_Y)$ in Equations (6) and (7).

We thus have

COROLLARY 2.3. (1) *The true and permutation distributions of the test statistic* $\mathbf{T}$ *will have the same even-order cumulants if* $m = n$.

(2) *The true and permutation distributions of the test statistic will not necessarily have the same ath order cumulants for a odd, regardless of whether* $m = n$, *unless* $k_a(F_X) = k_a(F_Y)$.

So, if $\mathbf{X}$ and $\mathbf{Y}$ are not multivariate normal, then differences in cumulants of order higher than two can cause permutation test for equality of means to be liberal even if $m = n$. We use a simulation with different skewness to demonstrate this, even when $\mathbf{X}$ and $\mathbf{Y}$ are univariate.

Let $X_i \overset{\text{i.i.d.}}{\sim} \text{Lognormal}(\mu_X, \sigma_X^2)$ and $Y_j \overset{\text{i.i.d.}}{\sim} \text{Lognormal}\ (\mu_Y, \sigma_Y^2)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, and consider testing the null hypothesis $H_0^E : E(X_i) = e^{\mu_X + \sigma_X^2/2} = E(Y_j) = e^{\mu_Y + \sigma_Y^2/2}$, using the test statistic

$$T = (\bar{X} - \bar{Y})/\sqrt{1/m + 1/n}.$$

We generated 10 000 sets of random samples from $\mu_X = -0.25$, $\sigma_X^2 = 1, m = 5$ and independently from $\mu_Y = 0.125, \sigma_Y^2 = 0.25$, $n = 5$. The two distributions have the same mean ($e^{\mu_X + \sigma_X^2/2} = e^{\mu_Y + \sigma_Y^2/2} = 1.284$) but different skewness, as shown in Figure 4. For each random sample, the p-value of the permutation test is computed by complete enumeration.

The cumulative distribution of the simulated p-values is shown in Figure 5. The actual significance level is higher than the nominal significance level. For instance, for a nominal significance level of 0.05, the estimated actual significance level is 0.154, with a 95% confidence interval of (0.150, 0.157). For a nominal significance level of 0.10, the estimated actual significance level is 0.233, with a 95% confidence interval of (0.228, 0.237). Apparently, the permutation test can be liberal.

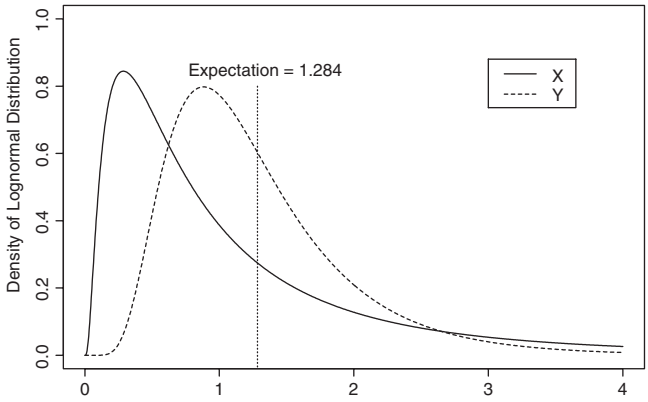

Fig. 4. Density plot of $X \overset{\text{i.i.d.}}{\sim} \text{Lognormal}(\mu_X = -0.25, \sigma_X^2 = 1)$ and $Y \overset{\text{i.i.d.}}{\sim} \text{Lognormal}(\mu_Y = 0.125, \sigma_Y^2 = 0.25)$. X and Y have the same expectation (1.284) but different skewness.

3 EXAMPLES AND RECOMMENDATIONS

An advantage of permutation testing is no knowledge of the distribution of the observations is required. Its control of error rate, however, only holds under the condition of identical distribution among groups to be compared. If the purpose of testing is to detect differences in means, then permutation testing may pick up unintended signals, rejecting an equality hypothesis for the wrong reason.

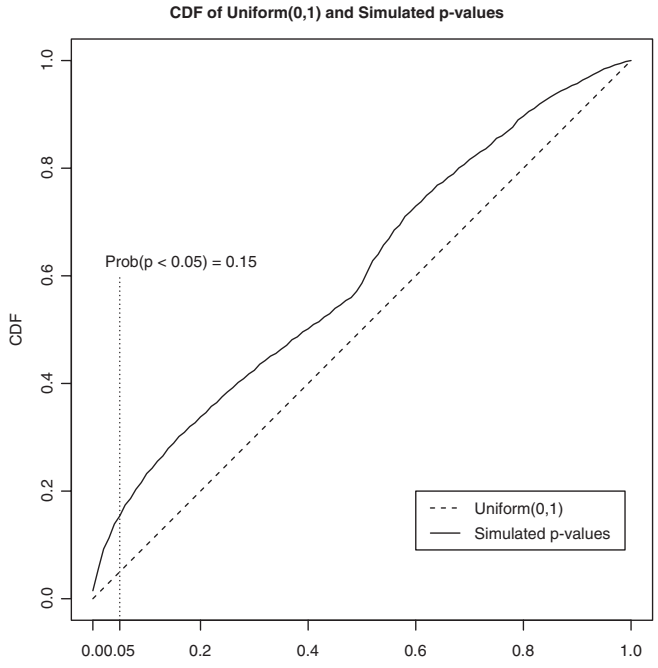

Fig. 5. Cumulative distribution plot of simulated p-value. $m = 5$ random samples are drawn from Lognormal($\mu_X = -0.25, \sigma_X^2 = 1$) and $n = 5$ random samples from Lognormal($\mu_Y = 0.125, \sigma_Y^2 = 0.25$), independently.

For example, van't Veer *et al*. (2002) reported a 70 gene signature as strongly predictive of short interval to distant metastasis of breast cancer. Their data was reanalyzed by Ein-Dor *et al*. (2005), in a way similar to the original data analysis. First, based on sample fold changes, (non-statistical) filtering was applied to select 5852 genes from the 24 481 genes probed on the microarrays. Then 1234 genes were selected for more detailed study from these 5852 genes, those genes corresponding to the rejection of the null hypothesis that its expression levels across patients is uncorrelated with their prognosis (a dichotomized outcome of either metastasis-free for >5 years or not, controlling FDR at 10%). Turns out it can be shown that permutation testing of sample correlation between gene expression profile and dichotomized prognoses is equivalent to permutation testing of difference in mean gene expression levels between good and poor prognosis groups. Since $m = 51 \neq n = 45$ in Ein-Dor *et al*. (2005), such permutation testing will detect, in addition to equality of means, unequal variance, correlations and skewness. Whether detecting such differences is useful or not may depend on the intended use of permutation testing. As the purpose of selecting genes in van't Veer *et al*. (2002) and Ein-Dor *et al*. (2005) was to train prognostic algorithms using machine learning, it seems to us the hypotheses of primary interests are $H_0^\mu : \mu_X = \mu_Y$ and not $H_0^P : P_X = P_Y$ (at least for algorithm based on distances measured by differences).

Instead of permutation testing, we recommend the following alternative approaches.

If gene expression data from microarray experiments can be modeled, and modeling diagnostics of the residuals (observed values minus values predicted from the model) show the errors have reasonable i.i.d. structure, then setting critical values for testing by appropriately re-sampling the residuals controls the error rate asymptotically. For example, Hsu *et al*. (2006) describe a statistically designed microarray experiment whose (logarithm of the) expression levels can be modeled linearly. After identifying error vectors that can be reasonably assumed to be i.i.d, they then bootstrapped the corresponding residuals vectors to set critical values for multiple testing.

If modeling of the data is difficult and the sample sizes are reasonably large, then we recommend the non-parametric bootstrap method of Pollard and van der Laan (2005) which re-samples data within each group and then centers the re-sampled test statistics to obtain a test null distribution. This re-sampling method is implemented as the MTP function in the multtest package of bioconductor (Pollard *et al*., 2005). They used both real and simulated data to study the behavior of permutation methods and their bootstrap methods, controlling generalized FWER. They found the error rates of permutation tests to be systematically higher than the target level, with the exception being the equal sample size case for difference of sample means test statistics. They also found that bootstrap methods can be either liberal or conservative in terms of error rates, depending on the test statistics used.

ACKNOWLEDGEMENTS

The authors thank Professor Yoon Lee for useful discussions. Research of J.C.H. is supported by NSF Grant No. DMS-0505519.

Conflict of Interest: none declared.

REFERENCES

Hsu,J.C., Chang,J., Wang,T., Steingrimsson,E., Magnusson,M.K. and Bergsteinsdottir,K. (2006) Statistically designing microarrays and microarray experiments to enhance sensitivity and specificity. *em Technical Report 771,* Department of Statistics, The Ohio State University, OH.

Ein-Dor,L. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Janssen,A. (1997) Studentized permutation tests for non - i.i.d. hypotheses and the generalized Behrens-Fisher. *Stat. Probab. Lett*., **36**, 9–21.

Pollard,K.S. and van der Laan,M. (2003) Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test. In *Proceedings of the 2003 International MultiConference in Computer Science and Engineering,* METMBS'03 Conference, pp. 3–9.

Pollard,K.S. and van der Laan,M. (2005) Resampling-based multiple testing: asymptotic control of type I error and applications to gene expression data. *J. Stat. Plan. Infer*., **125**, 85–100.

Pollard,K.S., Dudoit,S. and van der Laan,M.J. (2005) Multiple testing procedures: R multtest package and applications to genomics. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer-Verlag, New York.

Romano,J. (1990) On the behavior of randomization tests without a group—symmetry assumption. *J. Am. Stat. Assoc*., **85**, 686–692.

van't Veer,L.J. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.