

The L_1 -version of the Cramér-von Mises Test for Two-Sample Comparisons in Microarray Data Analysis

Yuanhui Xiao*, Alexander Gordon and Andrei Yakovlev
Department of Biostatistics and Computational Biology
University of Rochester, 601 Elmwood Avenue, Box 630
Rochester, New York 14642, USA

Abstract

Distribution-free statistical tests offer clear advantages in situations where the exact unadjusted p -values are required as input for multiple testing procedures. Such situations prevail when testing for differential expression of genes in microarray studies. The Cramér-von Mises two-sample test, based on a certain L_2 -distance between two empirical distribution functions, is a distribution-free test that has proven itself as a good choice. A numerical algorithm is available for computing quantiles of the sampling distribution of the Cramér-von Mises test statistic in finite samples. However, the computation is very time-and-space consuming. An L_1 counterpart of the Cramér-von Mises test represents an appealing alternative. In this work, we present an efficient algorithm for computing exact quantiles of the L_1 -distance test statistic. The performance and power of the L_1 -distance test is compared with those of the Cramér-von Mises and two other classical tests, using both simulated data and a large set of microarray data on childhood leukemia. The L_1 -distance test appears to be nearly as powerful as its L_2 counterpart. The lower computational intensity of the L_1 -distance test allows computation of exact quantiles of the null distribution for larger sample sizes than is possible for the Cramér-von Mises test.

*Corresponding author. E-mail: yxiao@bst.rochester.edu

Key words: L_1 -distance; two-sample tests; Cramér-von Mises test; exact distribution; non-parametric inference; differential expression; microarray data

1 Introduction

As larger sets of microarray gene expression data become readily available, nonparametric methods for microarray data analysis are beginning to be more appreciated ([22, 10, 23, 11, 14, 17], to name a few). This is attributable in part to serious concerns about the widely invoked distributional assumptions, such as log-normality of gene expression levels, in parametric inference from microarray data. It is well recognized that, in general, when the assumption of normality is violated the normal theory based statistical inference loses validity or becomes highly inefficient in terms of power [21]. In particular, Student's t -test can perform very poorly under arbitrarily small departures from normality [24]. Computer-assisted permutation tests employing resampling techniques cannot remedy this problem when the exact unadjusted p -values are needed as input for multiple testing procedures. Indeed, the small p -values required by procedures controlling the familywise error rate (FWER, see Dudoit et al. [7] for definition), such as the Bonferroni or Holm methods, cannot be estimated with sufficient accuracy by resampling, because the required number of permutations is astronomical [13] and cannot be accomplished with present-day hardware.

There are two properties of distribution-free methods that hamper their wide use in microarray studies. First, they are believed to have low power with small to moderate sample sizes, which property is attributable to their discrete nature. This common belief comes from computer simulations conducted for normally distributed data under location (shift) alternatives, conditions under which the t -test is known to be optimal. However, depending on the choice of a test statistic, the power of a given distribution-free test may be quite close to that of the t -test even under such ideal (for the t -test) conditions, with the gap between the two methods diminishing as the sample size increases. For example, the Cramér-von Mises test appears to be quite competitive when its power is assessed by simulating normally distributed log-expression levels under location alternatives [17] and it can provide a substantial gain in power under some other types of alternative hypotheses. Since one never knows the relevant class of alternative hypotheses, the

virtues of distribution-free tests are clear when a pertinent test statistic is judiciously chosen. The second problem with distribution-free test statistics is that they all have an attainable maximum. This property represents a serious obstacle to simultaneous testing of multiple hypotheses in small sample studies because it may make the adjusted p -values too large to declare even a single gene differentially expressed, even in the case where the empirical distributions pertaining to the two phenotypes under comparison do not overlap for many genes ([14, 13]).

Both problems are alleviated by increasing the sample size. Our experience suggests that the nonparametric inference based on distribution-free tests does not appear to be stymied (because of the second property) in genome-wide microarray studies when the number of subjects per group is greater than 20. We are convinced that samples of such or much larger sizes will be routinely used in microarray analysis in the not so distant future.

The implementation of distribution-free tests in microarray studies is also hampered by the fact that efficient numerical algorithms for computing p -values in finite samples are not readily available. The sampling distributions of such statistics do not depend upon which distribution generated the observed data under the null hypothesis. However, explicit analytical formulas for these distributions have been derived only in some special cases. Relevant asymptotic results are of limited utility in microarray analysis, because the accuracy of approximation in the tail region of the limiting distribution one is interested in (the region of very small p -values) is inevitably poor. Consider the example discussed in Section 3 of the present paper, where $m = n = 43$ and 12558 hypotheses are tested. For the Cramér-von Mises statistic value equaling $A = 2.2253921$, the exact and asymptotic p -values are equal to 2.115×10^{-6} and 3.994×10^{-6} , respectively. The Bonferroni-adjusted p -values are, therefore, equal to .02656 and .05015, respectively. Similarly, for the the statistic value equaling $B = 2.1193889$, the exact and asymptotic Bonferroni-adjusted p -values are .0493 and .0866, respectively. As a result, all the genes with values of the test statistic falling in the interval $[B, A]$ will be declared differentially expressed when using exact p -values, but they will not be selected if asymptotic p -values are used. This example shows that the development of universal numerical algorithms for computing exact p -values has no sound alternative. Such an algorithm for the Cramér-von Mises test with equal sample sizes was suggested by Burr [3]. While the predecessor of Burr's algorithm, which looked over all ordered arrangements of the two samples under comparison, was exponential-time in the sample sizes, the al-

gorithm of Burr is polynomial-time [3]. However, the computation is still quite time- and space-consuming, which limits its feasibility when the sample size increases. What is needed is a distribution-free test which is competitive with the Cramér-von Mises test in terms of power and stability of gene selection, while being more computationally efficient. Such a test was proposed by Schmid and Trede [19]. The test is based on a certain L_1 -distance between two empirical distribution functions. No explicit analytical expression is available for the sampling distribution of the L_1 -distance statistic, but its exact quantiles can be computed using a numerical algorithm described in the present paper. This algorithm shares many common features with the aforementioned algorithm of Burr for the Cramér-von Mises test [3] (see also Hájek and Šidák [12]) and builds on the idea which was first explored by Anderson in conjunction with the latter test [2]. The properties of the L_1 -distance test are studied below in applications to real and simulated data.

2 Methods

2.1 The L_1 -distance test and its relation to the Cramér-von Mises (L_2 -distance) test

Consider two independent samples x_1, x_2, \dots, x_m and y_1, y_2, \dots, y_n from continuous distributions $F(x)$ and $G(x)$, respectively; let F_m and G_n be their respective empirical distribution functions. Two-sample statistical tests are designed to test the null hypothesis $\mathbf{H}_0: F(x) = G(x)$ for all x versus the alternative: $F \neq G$.

The Cramér-von Mises statistic is defined as follows:

$$W_2 = \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^m [F_m(x_i) - G_n(x_i)]^2 + \sum_{j=1}^n [F_m(y_j) - G_n(y_j)]^2 \right\}. \quad (1)$$

This statistic and the test based on it (rejecting H_0 if the value of W_2 is “too large”) were introduced by Anderson [2] as a 2-sample variant of the goodness-of-fit test of Cramér [5] and von Mises [15].

Several authors tabulated the exact distribution of W_2 for small sample sizes under \mathbf{H}_0 [2, 3, 4, 25].

The L_1 -variant of W_2 introduced by Schmid and Tiede [19] is given by

$$W_1 = \frac{(mn)^{1/2}}{(m+n)^{3/2}} \left\{ \sum_{i=1}^m |F_m(x_i) - G_n(x_i)| + \sum_{j=1}^n |F_m(y_j) - G_n(y_j)| \right\}. \quad (2)$$

Let H_{m+n} be the empirical distribution function associated with the pooled sample of x_1, x_2, \dots, x_m and y_1, y_2, \dots, y_n . Then both statistics (1) and (2) can be represented similarly in the form

$$W_p = \left(\frac{mn}{m+n} \right)^{p/2} \int_{-\infty}^{\infty} |F_m(w) - G_n(w)|^p dH_{m+n}(w), \quad p = 1, 2. \quad (3)$$

Statistics (3) have a simple meaning. Move the $m+n$ points x_1, x_2, \dots, x_m and y_1, y_2, \dots, y_n , without changing their mutual order, to new positions, which are $1/(m+n), 2/(m+n), \dots, (m+n)/(m+n) = 1$. Let $\{\xi_1, \dots, \xi_m\}$ and $\{\eta_1, \dots, \eta_n\}$ be two subsets of the set $\{1/(m+n), 2/(m+n), \dots, 1\}$ coming from the x_i 's and y_j 's, respectively, and let F_m^* and G_n^* be the corresponding empirical distribution functions. Then W_p equals, up to a constant factor (depending only on m, n , and p), the p th power of the L_p -distance between F_m^* and G_n^* . In particular, W_1 is proportional to the area of the region between the graphs of F_m^* and G_n^* .

The discrete statistic W_1 has fewer possible values than the Cramér-von Mises statistic W_2 , its atoms are generally more “massive”, thus leading to a less powerful test. However, as evidenced by our simulations, the losses in power appear to be light and well compensated by substantial gains in computational efficiency (see Section 3).

2.2 An algorithm for computing the distribution of W_1

The algorithm described below uses the idea utilized earlier by Burr [3]. The formulas (4), (5), (6) on which the algorithm is based are close to those in Hájek and Šidák [12, pp. 143-144].

Let G be a directed graph with set of vertices $V(G) = \{(j, k) \in \mathbf{Z}^2 : 0 \leq j \leq m, 0 \leq k \leq n\}$ and with all possible edges of two types: from (j, k) to $(j+1, k)$ and from (j, k) to $(j, k+1)$, so that G has $(m+1)(n+1)$ vertices and $2mn - (m+n)$ edges.

A pair of samples x_1, \dots, x_m and y_1, \dots, y_n generates a few objects: the set X of all x_j 's; the set Y of all y_k 's; the pooled and ordered sample

z_1, \dots, z_{m+n} ; the sequence $h_i := F_m(z_i) - G_n(z_i)$, $i = 1, 2, \dots, m+n$ (we also put $h_0 := 0$); and, finally, a path $w = (w_0, w_1, \dots, w_{m+n})$ in the graph G defined as follows: $w_0 = (0, 0)$ and for $i = 1, 2, \dots, m+n$

$$w_i = \begin{cases} w_{i-1} + (1, 0), & \text{if } z_i \in X; \\ w_{i-1} + (0, 1), & \text{if } z_i \in Y, \end{cases}$$

so that w leads from $(0, 0)$ to (m, n) . The sequence $(h_i)_{i=0}^{m+n}$ satisfies equations $h_0 = 0$ and

$$h_i = \begin{cases} h_{i-1} + 1/m, & \text{if } z_i \in X; \\ h_{i-1} - 1/n, & \text{if } z_i \in Y, \end{cases}$$

$i = 1, 2, \dots, m+n$; it is, therefore, completely determined by the path w . More precisely, if $w_i = (j, k)$, then $h_i = j/m - k/n$. Note that under the null hypothesis (x_1, \dots, x_m and y_1, \dots, y_n are independent samples from the same continuous distribution) all paths w in G from $(0, 0)$ to (m, n) are equally likely.

The statistic W_1 equals

$$\frac{(mn)^{1/2}}{(m+n)^{3/2}} \sum_{i=0}^{m+n} |h_i|.$$

Let L be the least common multiple of m and n ; put $u := L/m$, $v := L/n$ and $g_i := Lh_i$, $i = 0, 1, \dots, m+n$, so that all g_i belong to \mathbf{Z} and W_1 equals $(mn)^{1/2}(m+n)^{-3/2}L^{-1}\eta$, where

$$\eta := \sum_{i=0}^{m+n} |g_i|.$$

Finding the null distribution of W_1 is, therefore, equivalent to finding that of η . If we introduce a function H on $V(G)$, putting

$$H(j, k) := |ju - kv|$$

(which quantity, up to a constant factor, equals the Euclidean distance in \mathbf{R}^2 from (j, k) to the line segment that connects $(0, 0)$ and (m, n)), then the value of η on the path $w = (w_i)_{i=0}^{m+n}$ equals

$$\eta(w) = \sum_{i=0}^{m+n} H(w_i).$$

For any $q = (j, k) \in V(G)$, define the frequency function $N(q; s) \equiv N(j, k; s)$, $s \in \mathbf{Z}_+ = \{0, 1, 2, \dots\}$, as the number of paths $(w_i)_{i=0}^{j+k}$ from $(0, 0)$ to (j, k) in G , such that

$$\sum_{i=0}^{j+k} H(w_i) = s.$$

In the special case $j = m$, $k = n$, knowledge of this frequency function yields the distribution of $\eta(w)$, since

$$\begin{aligned} Pr\{\eta(w) = s\} &= N(m, n; s) \left(\sum_{s' \geq 0} N(m, n; s') \right)^{-1} \\ &= N(m, n; s) \binom{m+n}{m}^{-1}. \end{aligned}$$

The problem becomes: to find the frequency function $N(m, n; s)$, $s \geq 0$. This can be achieved by finding the frequency functions $N(j, k; s)$ for all pairs $(j, k) \in V(G)$, which can be done recursively as follows.

First, assume $k = 0$. There is only one path $(w_i)_{i=0}^j$ from $(0, 0)$ to $(j, 0)$; the corresponding sum of $H(w_i)$ equals $\sum_{i=0}^j lu = j(j+1)u/2$, so that

$$N(j, 0; s) = \begin{cases} 1, & \text{if } s = j(j+1)u/2, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Similarly,

$$N(0, k; s) = \begin{cases} 1, & \text{if } s = k(k+1)v/2, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Furthermore, if $j, k > 0$, then for every path $(w_i)_{i=0}^{j+k}$ from $(0, 0)$ to (j, k) we have either $w_{i-1} = (j-1, k)$ or $w_{i-1} = (j, k-1)$, so that

$$\begin{aligned} N(j, k; s) &= N(j-1, k; s - H(j, k)) + N(j, k-1; s - H(j, k)) \\ &= N(j-1, k; s - |ju - kv|) + N(j, k-1; s - |ju - kv|). \end{aligned} \quad (6)$$

(Note that the right hand side equals 0 if $s < |ju - kv|$.) The recursive formula (6) and the boundary conditions (4), (5) allow one to compute the frequency functions $N(j, k; s)$, $s \geq 0$, in the lexicographic (dictionary) order of pairs (j, k) .

Here are some remarks on the computer implementation of the algorithm. First of all, every function $N(j, k; s)$ vanishes if $s \geq R_{m,n} := m(m+1)u/2 + n(n+1)v/2 + 1 = L(m+n+2)/2 + 1$, so that no more than $R_{m,n}$ values should be stored for every pair $(j, k) \in V(G)$.

There are $|V(G)| = (m+1)(n+1)$ such frequency functions, but all of them do not need to be stored simultaneously. Once such functions $N(j, k; s)$ have been computed for $j = j^*$ ($1 \leq j^* \leq m$) and all $k = 0, 1, \dots, n$, the functions with $0 \leq j < j^*$ are not needed any more, and the memory they occupy can be freed. Therefore, at any time we need to store such functions for only two neighboring values of j . For large m, n the required memory M is, therefore, of order $L(m+n)n$; re-organizing the computation appropriately, with the use of the symmetry with respect to m and n , we can improve the estimate to

$$M = O(L(m+n) \min(m, n)) = O(Lmn). \quad (7)$$

We remind the reader that L is the least common multiple of m and n , and the symbol $O(X)$, for large X , means any quantity Y that satisfies an inequality $|Y| < AX + B$ with some fixed constants A and B .

Assuming that $m \leq n$, the two extreme cases are $m = n - 1$ and $m = n$, where (7) gives $M = O(n^4)$ and $M = O(n^3)$, respectively.

The time (or, more precisely, the number of computer operations), T , required for the computation, satisfies the inequality: $T \leq C(m+1)(n+1)L(m+n+2)/2$ with a certain constant C . (Indeed, we need to calculate each value $N(j, k; s)$, which is a sum of at most two previously computed values.) This implies that

$$T = O(mnL \max(m, n)).$$

Assuming, as above, that $m \leq n$, we obtain the general estimate $T = O(n^5)$, while in the special case $m = n$ we have $T = O(n^4)$.

These estimates should be compared with those for the corresponding algorithm for computing the distribution of the Cramér-von Mises statistic. The estimated number of stored values $N(j, k; s)$ for each pair (j, k) is approximately L times more than for the algorithm described above. This multiplies both required memory and time by a factor of L , which factor, assuming $m \leq n$, may vary from n (the case $m = n$) to $n(n-1)$ (the case $m = n-1$).

The exact quantiles of the sampling distribution of W_1 resulted from the above algorithm are in complete agreement with the corresponding quantiles given by Schmid and Trede [19] for small and moderate balanced samples.

3 Results

3.1 Computational Efficiency of the Algorithm

We compared the computational efficiency of the proposed algorithm for computing the null distribution of the L_1 -distance test statistic W_1 to that for the Cramér-von Mises test statistic W_2 . We studied the time requirements of both algorithms, as well as their respective maximum sample sizes for which the computation is still feasible. All our computation experiments were carried out on a UNIX workstation (Sunfire V480) with 16.3GB RAM, 4×8.0 MB Cache and 4×1200 MHz CPU.

Table 1 presents the time it takes the computer to find the distribution function of each of the two statistics W_1 and W_2 . (More precisely, the table shows the CPU time, i.e., the processor time, needed for the computation.) For simplicity of representation of the results, only two extreme cases with $n = m$ and $n = m + 1$ are shown. For each test, the computing time increases as a power of the sample size. However, the difference in the corresponding exponents leads to a significant difference in the computing time. Because of the design of the algorithm presented in Section 2.2, the case $n = m + 1$ is the least favorable so that the difference in computing time for the two methods becomes evident even in small samples. For $n = m = 40$, the computing time for the Cramér-von Mises test is about 12 times longer than that for the L_1 -distance test. The divergence is more dramatic for larger sample sizes. For $n = m = 150$, the computing time increases to almost half an hour for the Cramér-von Mises test, while it is less than 20 seconds for the L_1 -distance test.

The difference in memory requirements leads to a difference in the maximum sample sizes for which the computation is still feasible. With the above mentioned computer, in the case of equal sample sizes ($m = n$), the maximum sample sizes are approximately 800 and 200 for the test statistics W_1 and W_2 , respectively.

3.2 Power of the L_1 -distance test

To assess the power of the proposed test, we designed our simulation study as follows:

1. In each sample, data are generated from a normal distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2 . In the context of microarray data analysis, this design implies that the original gene expression levels are log-transformed.
2. One of the two samples under comparison is generated from the distribution with $\mu = 0$ and $\sigma = 1$. To generate the other sample, either the parameter μ or the parameter σ^2 is set at different values keeping the other parameter constant.
3. The resultant pair of samples is used to compute the observed values of the test statistics under study.
4. Steps 1-3 are repeated 10,000 times. The number of times when the null hypothesis gets rejected at a significance level of 0.05 is divided by 10,000 and plotted as a function of each parameter.

Under the above described design, we compared the power of the L_1 -distance test with that of the Cramér-von Mises, Kolmogorov-Smirnov, and Student's t -test. Figure 1 presents the power curves for the four tests at significance level $\alpha = 0.05$ under the location (shift) alternatives. As expected, the t -test outperforms the other three because of its optimality under these conditions. For the balanced case $m = n = 20$ and the unbalanced case $m = 20$ and $n = 21$, the gap between the power curves for the Cramér-von Mises and L_1 -distance tests is almost undetectable. The Kolmogorov-Smirnov test is the least powerful among the four tests in both cases.

Figure 2 presents the results of testing differences in the variance. In this simulation study, the samples were drawn from two normal distributions with equal means ($\mu_1 = \mu_2 = 0$) but different variances. It comes as no surprise that the power curve for the t -test is practically flat in this case. For the cases $m = n = 20$ and $m = 20, n = 21$, the simulated power curves for the Cramér-von Mises and L_1 -distance tests agree closely. Both tests outperform the Kolmogorov-Smirnov test.

Figure 3 shows the power curves for the four tests at the same significance level with the samples drawn from exponential distributions. In this case, the power curve is plotted as a function of the ratio of the means of the two exponential distributions under comparison. Again the Kolmogorov-Smirnov is the least powerful among the four tests while the L_1 -distance test and the Cramér-von Mises test are highly competitive with each other.

3.3 Analysis of Biological Data

For the purposes of this study, we used the publicly available St. Jude Children’s Research Hospital (SJCRH) Database on childhood leukemia (<http://www.stjuderesearch.org/data/ALL1/>). The whole SJCRH Database contains gene expression data on 335 subjects, each represented by a separate array (Affymetrix, Santa Clara, CA) reporting measurements on the same set of $m = 12558$ genes. We selected two groups of patients with hyperdiploid (Hyperdip) and T-cell acute lymphoblastic leukemia (TALL), respectively. The groups were balanced to include 43 patients in each group. The microarray data were background corrected and normalized using the Bioconductor RMA software. The raw (background corrected but not normalized) expression data were generated by the output of the RMA procedure when choosing the option: *normalization=false*. The L_1 -distance test was compared with Student’s t and the Cramér-von Mises tests in this application. The three tests were applied to select differentially expressed genes by testing two-sample hypotheses with the Hyperdip and TALL data. The FWER was controlled by resorting to either the Bonferroni or the Westfall-Young method.

The stability of gene selection was assessed by resampling as described in [17]. We used a subsampling variant of the delete- d -out jackknife method (with $d = 7$) for estimation of the variance of the number of selected genes [18]. This method is technically equivalent to the leave- d -out cross-validation technique. The general recommendation is to leave out more than $d = \sqrt{n}$ but much fewer than the available n arrays ([8, 18]). We followed this recommendation when selecting $d = 7$ and checked the results obtained with slightly larger values of d . The results were largely similar. For the Bonferroni adjustment, the number of subsamples was equal to 1000, while for the Westfall-Young step-down permutation algorithm we used only 200 subsamples because the latter procedure is much more time-consuming. We used 10,000 permutations to estimate adjusted p -values with the Westfall-Young algorithm.

Tables 2 and 3 present the numbers of genes selected by the three tests combined with the Bonferroni adjustment or the Westfall-Young algorithm for normalized and raw data. The tables also present the mean numbers of genes selected across the leave-7-out subsamples and their jackknife standard deviations (in parentheses). The t -test appears to be the most conservative one among the three tests in this particular analysis. The results obtained

by the Cramér-von Mises test and its L_1 variant agree quite closely. This is especially true for the Westfall and Young method. With the Bonferroni adjustment, the Cramér-von Mises test appears to be slightly more conservative than the L_1 -distance test in terms of the mean (over subsamples) number of selected genes. The stability of gene selection appears to be similar for the three tests.

4 DISCUSSION

The Cramér-von Mises nonparametric test has received much attention in the literature. The bulk of theoretical work in this field has been focused on the Cramér-von Mises goodness-of-fit test [1, 6]. The two-sample Cramér-von Mises test was considered by Anderson [2], Burr [4], Zajta and Pandikow [25]. Among other things, some limited tables of quantiles for the two-sample Cramér-von Mises test were presented in these works. The tables were generated by a simple but extremely time consuming (exponential-time) algorithm looking over all ordered arrangements of the two samples and treating them (under the null hypothesis) as equally likely. Burr [3] proposed a much more efficient polynomial-time algorithm for computing such quantiles. His algorithm was designed for the case of equal sample sizes. In theory this idea was extended to arbitrary sample sizes by Hájek and Šidák [12]. However, the computation is still quite time and space consuming.

Schmid and Trede [19] proposed a new distribution-free test for the two-sample problem, namely, an L_1 -variant of the Cramér-von Mises test [19]. They also generated limited tables of quantiles for that test (in the case of equal sample sizes), using a simple exponential-time algorithm based on rearrangements, and studied the power of this L_1 -distance test in comparison with the Cramér-von Mises (L_2 -distance) and some other tests. In another paper [20], Schmid and Trede considered the utility of an L_1 -variant of the Cramér-von Mises goodness-of-fit test.

The present paper further explores the L_1 -distance test. We present a time- and space-efficient algorithm and software for computing its exact quantiles. The polynomial-time algorithm is based on the idea of Burr [3] mentioned above and uses formulas similar to those in Hájek and Šidák [12]. The sample sizes are not necessarily equal. The algorithm enables an investigator to compute exact tail probabilities, no matter how small they are. Using a standard design of power studies, we have found, based on simu-

lated data, that the L_1 -distance two-sample test is almost as powerful as the original Cramér-von Mises test based on the L_2 -distance between two empirical distribution functions. This observation is consistent with the results of a simulation study by Schmid and Trede [19]. The results of computer simulations reported in Section 3.2 cannot be taken as evidence that the Cramér-von Mises test is always superior, even if slightly, to the L_1 -distance test in terms of power. It is conceivable that, under real-world alternatives, the power of the L_1 -test may be even higher than that of the Cramér-von Mises test. At the same time, the L_1 -distance test is computationally much less intensive than its L_2 counterpart. In particular, this allows one to compute exact quantiles for the L_1 test with larger sample sizes than for the L_2 test. In an application to actual biological data both tests have generated lists of differentially expressed genes having almost equal sizes.

In summary, we recommend the L_1 -variant of the Cramér-von Mises test as a good alternative to the original Cramér-von Mises test for selecting differentially expressed genes in microarray studies.

ACKNOWLEDGEMENTS

The work was supported in part by NIH grant GM075299.

References

- [1] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain “goodness of fit” criterion based on stochastic processes. *The Annals of Mathematical Statistics*, **23**, 193-212.
- [2] Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *The Annals of Mathematical Statistics*, **33**, 1148-1159.
- [3] Burr, E. J. (1963). Small-sample distribution of the two-sample Cramér-von Mises criterion for small equal samples. *The Annals of Mathematical Statistics*, **34**, 95-101.
- [4] Burr, E. J. (1964). Distribution of the two-sample Cramér-von Mises W^2 and Watson’s U^2 . *The Annals of Mathematical Statistics*, **35**, 1091-1098.

- [5] Cramér, H. (1928). On the composition of elementary errors: II. Statistical applications. *Skand. Akt.*, **11**, 141-180.
- [6] Csorgo, S. and Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-von Mises statistics. *Journal of the Royal Statistical Society. Series B*, **58**, 221-234.
- [7] Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**, 71-103.
- [8] Efron, B. and Tibshirani, R (1993). An Introduction to the Bootstrap, Chapman & Hall/CRC, New York.
- [9] Fisz, M. (1960). On a result by M. Rosenblatt concerning the von Mises Smirnov test, *The Annals of Mathematical Statistics*, **31**, 427-429.
- [10] Grant, G. R., Manduchi, E. and Stoeckert, C. J. (2002). Using non-parametric methods in the context of multiple testing to determine differentially expressed genes. In Lin SM, Johnson KF (eds), *Methods of Microarray Data Analysis: Papers from CAMDA '00*, Kluwer Academic Publishers, Norwell, MA, 37-55.
- [11] Guan, Z. and Zhao, H. (2005). A semiparametric approach for marker gene selection based on gene expression data. *Bioinformatics*, **21**, 529-536.
- [12] Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.
- [13] Klebanov, L., Gordon, A., Xiao, Y., Land, H. and Yakovlev, A. (2005). A permutation test motivated by microarray data analysis. *Comp. Stat. Data Anal.*, in press.
- [14] Lee, M-L. T., Gray, R. J., Björkbacka, H. and Freeman, M.W. (2005). Generalized rank tests for replicated microarray data. *Statistical Applications in Genetics and Molecular Biology*, **4(1)**, Article 3.
- [15] von Mises, R. (1931). *Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theoretischen Physik*. Deuticke, Leipzig.
- [16] Rosenblatt, M. (1952). Limit theorems associated with variants of the von Mises statistic. *The Annals of Mathematical Statistics*, **23**, 617-623.

- [17] Qiu, X., Xiao, Y., Gordon, A. and Yakovlev, A. (2005). Assessing stability of gene selection in microarray data analysis. Technical Report 05/10,
<http://www.urmc.rochester.edu/smd/biostat/people/techreports.html>.
- [18] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics, Springer, New York.
- [19] Schmid, F. and Trede, M. (1995). A distribution free test for the two sample problem for general alternatives. *Comp. Statistics and Data Analysis*, **20**, 409-419.
- [20] Schmid, F. and Trede, M. (1996). An L_1 -variant of the Cramér-von Mises test. *Stat. Probab. Letters*, **26**, 91-96.
- [21] Srivastava, D. K. and Mudholkar, G. S. (2003). Goodness-of-fit tests for univariate and multivariate normal models. In: *Handbook of Statistics*, Khattree, R. and Rao, C. R., eds., **22**, 869-906.
- [22] Stamey, T. A., Warrington, J. A., Caldwell, M. C., Chen, Z., Fan, Z., Mahadevappa, M., McNeal, J. E., Nolley, R. and Zhang, Z. (2001). Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia. *J. Urol.*, **166**, 2171-2177.
- [23] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D. and Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-1461.
- [24] Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. Springer, New York.
- [25] Zajta, A. J. and Pandikow, W. (1977). A table of selected percentiles for the Cramér-von Mises-Lehmann test: equal sample sizes. *Biometrika*, **64**, 165-167.

Table 1: CPU time used for finding the distribution function for W_1 and its L_2 -counterpart W_2 under the null hypothesis \mathbf{H}_0 . The CPU time was measured in units of 10^{-3} seconds. The computing time is too small to be observable for $m < 40$ if $n = m$ and for $m < 10$ if $n = m + 1$.

$m = n$	W_1	W_2	$m = n$	W_1	W_2	m, n	W_1	W_2
40	80	1000	100	3120	160930	10, 11	10	10
50	190	3210	110	4690	282000	20, 21	120	1190
60	400	9290	120	6790	476170	30, 31	1050	23630
70	750	21940	130	9800	774070	40, 41	5920	193250
80	1270	45980	140	13950	1212940	50, 51	21750	833790
90	2050	87580	150	18890	1792010	60, 61	63080	$> 2^{31} \cdot 10^{-3}$

Table 2: Numbers of genes selected by L_1 -distance test, Cramér-von Mises test and t test combined with Bonferroni adjustment. The family-wise error rate was controlled at the level 0.05. The numbers in parentheses are jackknife standard deviations.

Statistical test	L_1 -test	L_2 -test	t -test
Normalized data			
Original Sample	1029	1031	951
Mean ($d = 7$)	1371(153)	1092(134)	779(98)
Raw data			
Original Sample	516	545	458
Mean ($d = 7$)	704(317)	572(219)	388(141)

Table 3: Numbers of genes selected by L_1 -distance test, Cramér-von Mises test and t test combined with Westfall-Young algorithm. The family-wise error rate was controlled at the level 0.05. The numbers in parentheses are jackknife standard deviations.

Statistical test	L_1 -test	L_2 -test	t -test
Normalized data			
Original Sample	1091	1092	1058
Mean ($d = 7$)	882(122)	885(119)	876(109)
Raw data			
Original Sample	870	866	790
Mean ($d = 7$)	743(379)	752(325)	675(317)

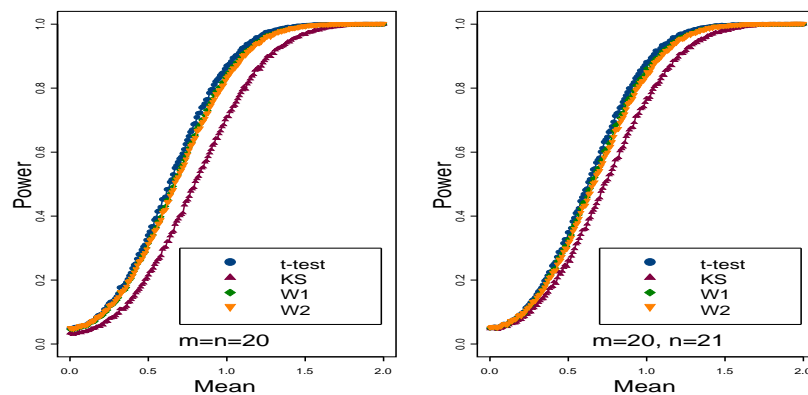


Figure 1: Power curves for t , Kolmogorov-Smirnov (KS), L_1 -distance and Cramér-von Mises tests against location (shift) alternatives at significance level 0.05. Samples were drawn from normal distributions with the same variance 1 but unequal means.

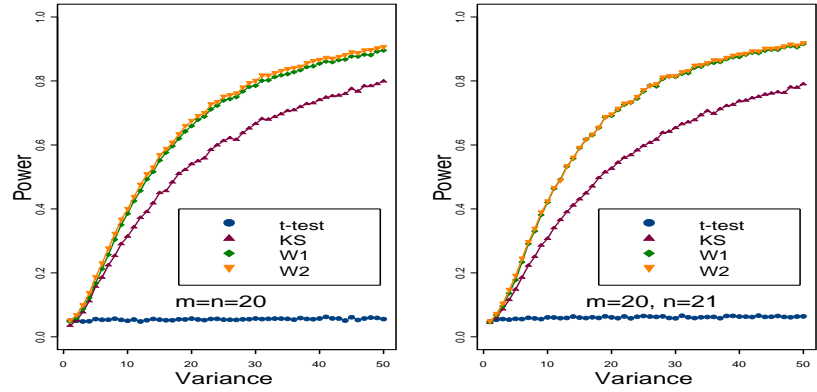


Figure 2: Power curves for t , Kolmogorov-Smirnov (KS), L_1 -distance and Cramér-von Mises tests at significance level 0.05. Samples were drawn from normal distributions with equal means but different variances.

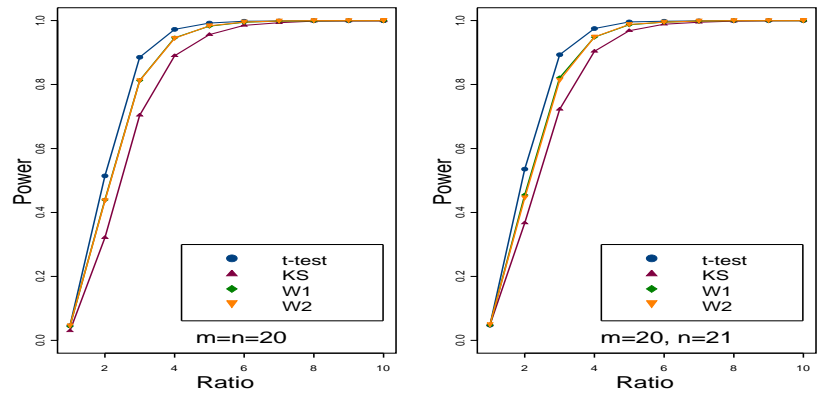


Figure 3: Power curves for t , Kolmogorov-Smirnov (KS), L_1 -distance and Cramér-von Mises tests at significance level $\alpha = 0.05$. Samples were drawn from exponential distributions with different means. X -axis is the ratio of the means of the two exponential distributions from which the samples were drawn.