University of Hannover

Faculty of Horticulture

Bioinformatics Unit

Small Sample Inference For The Two-Sample Design

By
Donghui Ma

A Thesis Submitted in Partial Fulfilment of the Master of Science Degree in Horticulture (Major in Genetics and Plant Breeding)

SUPERVISORS

Prof. Dr. Ludwig Hothorn

Prof. Dr. Bernhard Hau

Hannover, Germany

September 2004

ABSTRACT

It is common that two populations are to be compared, and one objective is to test the null hypothesis that the two populations have the same response distributions against the alternative that the response distributions are different. Three classes of tests are provided for this situation, depending on the type of alternative hypothesis. One class of tests has good power against shift alternatives (i.e., changes in the location), a second against dispersion alternatives (i.e., changes in the scale), and a third against general alternatives. In this research, the main focus will be the two-sample test for location problem, especially when the data are under nonnormality and heteroscedasticity, since the classical test are based on either of this two assumptions. For example, Welch *t* test (Welch, 1947) improves the simple *t* test ("Student", 1908) when data are heterogeneous, but it works less efficiently than Wilcoxon Rank Sum test (Wilcoxon, 1945) when the data are skewed. Wilcoxon test is powerful under nonnormality but it behaves poorly under heteroscedasticity.

In fact, without the parametric assumption of the underlying distribution of the data, the uniformly most powerful test does not exist. But under certain circumstance, there exists locally most powerful test. Some of the appropriate parametric and nonparametric tests will be introduced in section 4 for different conditions. When there is no prior information about the conditions of the data, there are other statistical procedures available, such as Adaptive test (Bickel, 1982), Maximin efficiency robust test (Gastwirth, 1966), Maximum test (Tarone, 1981; Fleming & Harrington, 1991) etc. At the end a new Maximum test is proposed, when the underlying distribution of the data is a priori unknown. For a good presentation of the Microarray data, the test results can be plotted using the so-called Volcano Plot. There is also an improved Volcano Plot proposed using the concept of confidence interval, which is discussed at the end of Chapter A.

When the data is dichotomized with some priori cut-point, for example maximally selected cut point (Hothorn and Lausen, 2002), the inference for binary data can be also used for continuous data. At the end of Chapter B, a new method for the construction of confidence intervals for the ratio of proportions is also discussed.

All the tests and methods for confidence intervals are compared via Monte-Carlo simulation, the simulation results are shown in chapter A and B respectively. Among all the candidate tests, there are no clear winner in all the conditions, but when the data are under nonnormality and heteroscedasticity, Welch *t* test behaves relatively better than others, although the assumption of the test is violated. The new method for the confidence interval of ratio of proportions is proved to maintain the nominal level of confidence (95 percent).

for mom, dad, Jin and Inger

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GENERAL INTRODUCTION

For the two-sample location problem, Student's *t*-test was developed by "Student" (Gossett, 1908) to deal with the problems associated with inference based on small samples. The classical *t*-test based on the assumption of normally distributed data and variance homogeneity. In the heterogeneous variance case (so called Behrens-Fisher problem), Welch *t*-test (Welch, 1947) was proposed to fill this void. It is very often that doubts on the normality exist, consequently distribution free test, such as Wilcoxon rank sum test (Wilcoxon, 1945) is favorable. But the Wilcoxon test also based on the homogeneous variance. When such condition is violated, the Wilcoxon test will have poor power than expected. Such violation of assumption is very common and serious in the microarray data, in this paper all the examples and simulation conditions will be chosen to mimic the microarray data. But the results and conclusion in this paper are not limited to the microarray data, they can be used to all kinds of experimental data when the data are under certain conditions.

The problems occurred in the analysis of array data is one of the motivations of this research. Firstly some characteristics of microarray data will be characterized in chapter A. The similar problems can be found also in horticulture science, another two data sets from horticulture experiments will also be introduced. Different distribution systems are also introduced to mimic the data generating mechanism, such as some standard probability distributions (Normal, Lognormal, Exponential, etc.) and some other distribution systems which can generate distributions with short tails, long tails, skewness and kurtosis. In this thesis, mixture of normal distributions, Fleishman distribution and Johnson distribution are used for the purpose of generate such nonstandard distributions. When the data with certain characteristics are possible to regenerate, different test can be compared with such regenerated data. This kind of techniques is called Monte-Carlo simulation. In Chapter A, the general idea of the simulation study is reviewed, and the usual method for random number generating is also introduced.

The test results can be presented efficiently using the so-called Volcano plot; Volcano plot is nothing but a scatter plot of the base 2 logarithm of the ratio of means versus the base 10 logarithm of the p value from the statistical test. One disadvantage of the Volcano plot is that the distance of two points is hard to interpret since the x- and y-axis are in two different scales. To make the interpretation easier, a modified version of Volcano plot is proposed, which use the confidence intervals instead of the p values for the y-axis. The confidence interval for the

ratio of means of two samples will be calculated. And the base 2 logarithm of the lower limited of the confidence interval for the ratio will be plotted when the estimated ratio is larger than 1 and the base 2 logarithm of the upper limit of the confidence interval for the ratio will be plotted when the estimated ratio is smaller than 1. In this plot, both x- and y-axis are of the same meaning, namely fold change. The x-axis is the estimated fold change, and the y-axis is how extreme the fold change can be. In the modified Volcano plot, the points with both estimated ratio and lower limit larger than 1 and the points with both estimated ratio and upper limit larger smaller than 1 are interesting, since they are more likely to give the significant result in the corresponding test. The details of both the Volcano plot and the modified Volcano plot will be discussed in Chapter A.

Also when the data is dichotomized with some priori cut-point, for example maximally selected cut point (Hothorn and Lausen, 2002), the inference for binary data can be also used for continuous data. In this research, a review the inference for the ratio is made, especially the method for construction of confidence intervals for the ratio of two proportions. A new method (add-4 asymptotic method) for the construction of confidence intervals for the ratio of proportions is proposed. The proposed new method is proved to maintain the nominal level of confidence (0.95) for the confidence interval via Monte-Carlo simulation.

1. CHAPTER A

Introduction to Microarray Data

It is necessary to know whether the data are really fulfilled the assumption of the statistical tests before conducting the statistical analysis. For example, when the experimental data is nonnormal but variance homogeneous, Wilcoxon test can be used. Thus the distributional characteristics of Microarray data are important for choosing the appropriate statistical test.
Microarray experiments are conducted in such a manner as to profile the behavior patterns of thousands of nucleic acid sequences or protein simultaneously. Plus, they are capable of being automated and run in a high throughput mode. Thus they generate mountains of data and data analysis is necessary for converting data to knowledge. But the microarray data have some characteristics, such as small number of replicates, missing values, variance heterogeneity, nonnormality, bimodal distribution, different shapes in two samples, etc., which make the statistical analysis hard to implement. Several datasets are investigated, which is shown in Table 1:

Table 1 Data sets investigated in this study

Dataset Name	Replications	Gene Number	Type
Armstrong *et al.* (2002)	24 ALL, 20 MLL	12582	Oligo.
Golub *et al*. (1999)	25 ALL, 47AML	7129	Oligo.
Singh *et al*. (2002)	50 Normal, 50 Tumor	12600	Oligo.
Yeoh *et al*. (2002)	27 E2APBX,79 TEL AML	12625	Oligo.
Shipp *et al.* (2002)	19 DLBCL, 19 FL	7129	Oligo.
Garber *et al.* (2001)	29 Ade, 31 others	22115	cDNA
Gruvberger *et al.* (2001)	28 ER+, 30 ER-	3389	cDNA
Khan *et al.* (2001)	23 EWS, 20 RMS	2303	cDNA
Huang *et al.* (2001)	8 Tumor, 8 Normal	12558	cDNA

Summary of the 9 data sets used to study the characteristics of the dataset. The number of genes (or, more precisely, the number of array elements) is indicated. The middle column description of the experiment, sample size and the comparison we studied. For details, see the web supplement. Abbreviations: E2APBx, GIST,

gastrointestinal stromal tumor; ER, estrogen receptor; AML, acute myeloid leukemia; BPH, benign prostate hyperplasia; DLBC, diffuse large B cell lymphoma; FL, follicular lymphoma; EWS, Ewing's sarcoma; RMS, rhabdomyosarcoma; MLL, mixed lineage leukemia; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia.

Characteristics of Microarray data:

In this thesis, the characteristics of Singh *et al*. (2002) data set will be shown as a example, because Singh data are of the largest sample size among all the available datasets. It is more reliable to describe the distribution shapes of Microarray data. The design of this experiment is to determine whether global biological differences underlie common pathological features of prostate cancer and to identify genes that might anticipate the clinical behavior of this disease. The sample sizes of both samples are 50.

To describe the distributions for all the genes, four descriptive statistics are to be used. They are mean, standard deviation, skew and kurtosis. With these four statistics, the distributional information of the Microarray data can be roughly described.

The histograms of two samples are shown in Figure 1, which globally indicates the differences of gene expression.

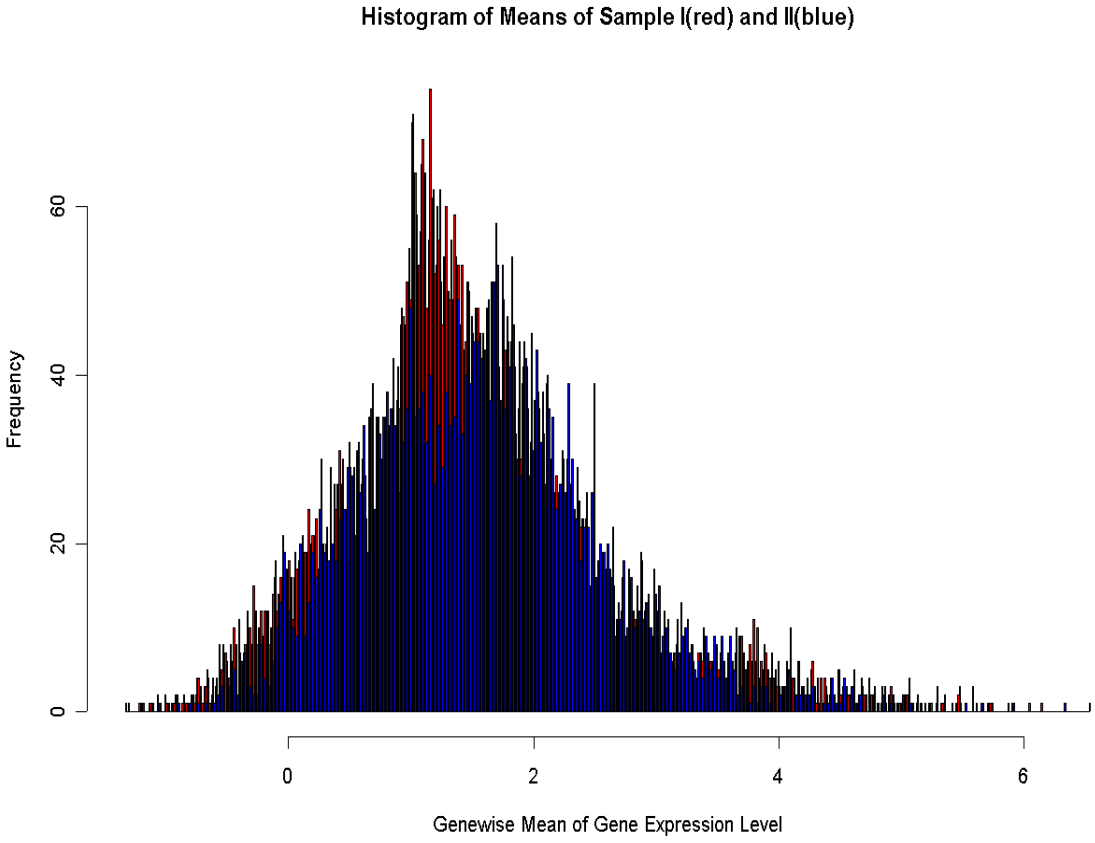

Figure 1 Histograms of tumor (blue) and normal (red) samples.

The Histograms above only show the global profiles of the gene expression differences, i.e. only the means of all the genes in the first group and the second group are plotted separately. We can see the differences between two the two samples, but we cannot find out how many genes are really differently expressed. Thus, the genewise absolute and relative differences for each gene between two groups are calculated, the genewise differences and relative differences (ratios) study is also performed and shown in Figure 2:

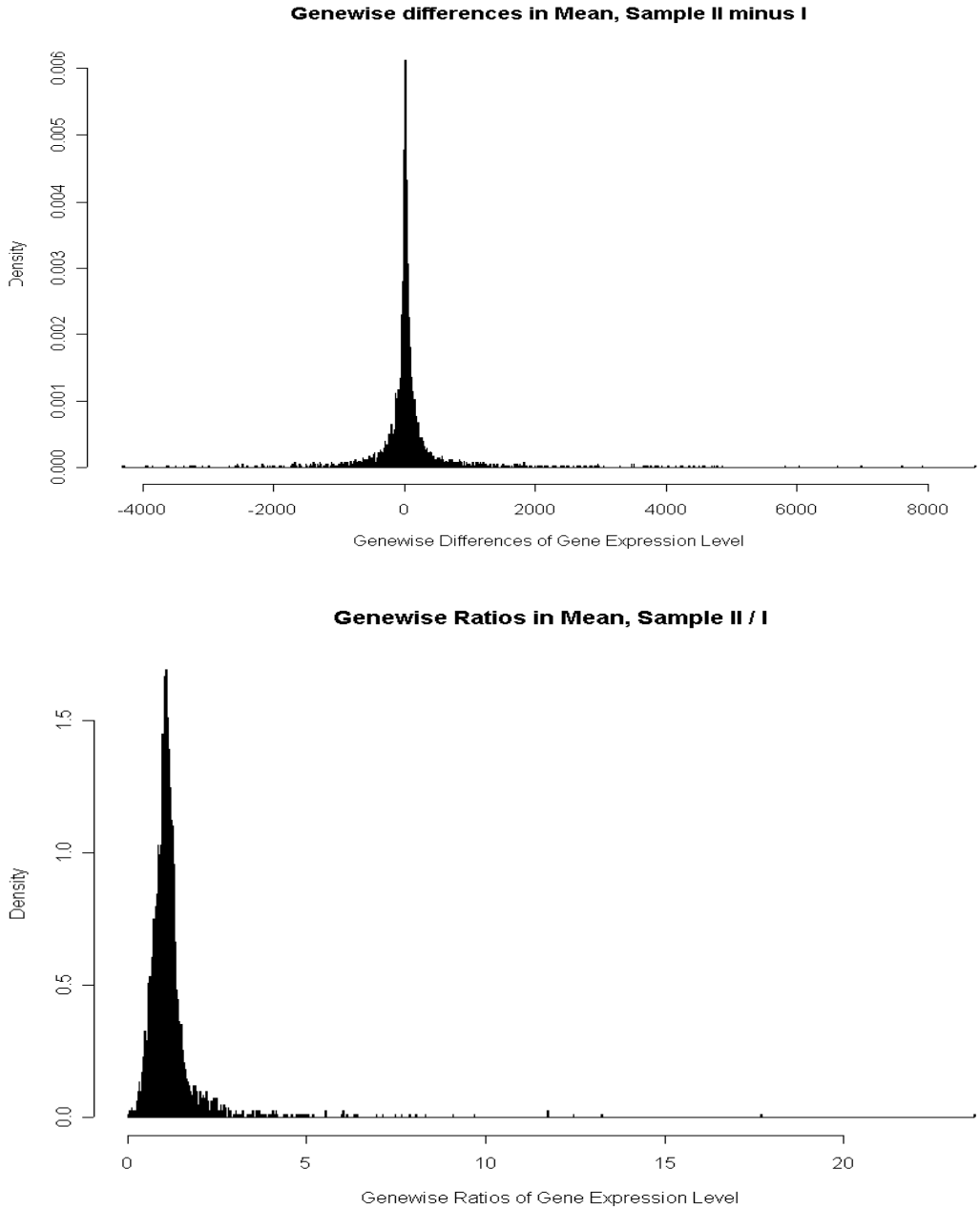

Figure 2 Genewise mean differences (left) and ratios (right) of tumor sample minus normal samples and tumor sample divided by normal sample.

The majority of the genes do not expressed differently, so the genewise differences centered at 0 as the genewise ratios centered at 1 correspondingly. Because of the different magnitude of the gene expression levels between different experiments, comparison of the absolute differences is nonsense. The fold change is meaningful in this case, and the majority of the fold changes among all the datasets are between 0 and 5, whereas 0 means very big difference. There are also relative differences in variations of the data, which can be seen from Figure 3:

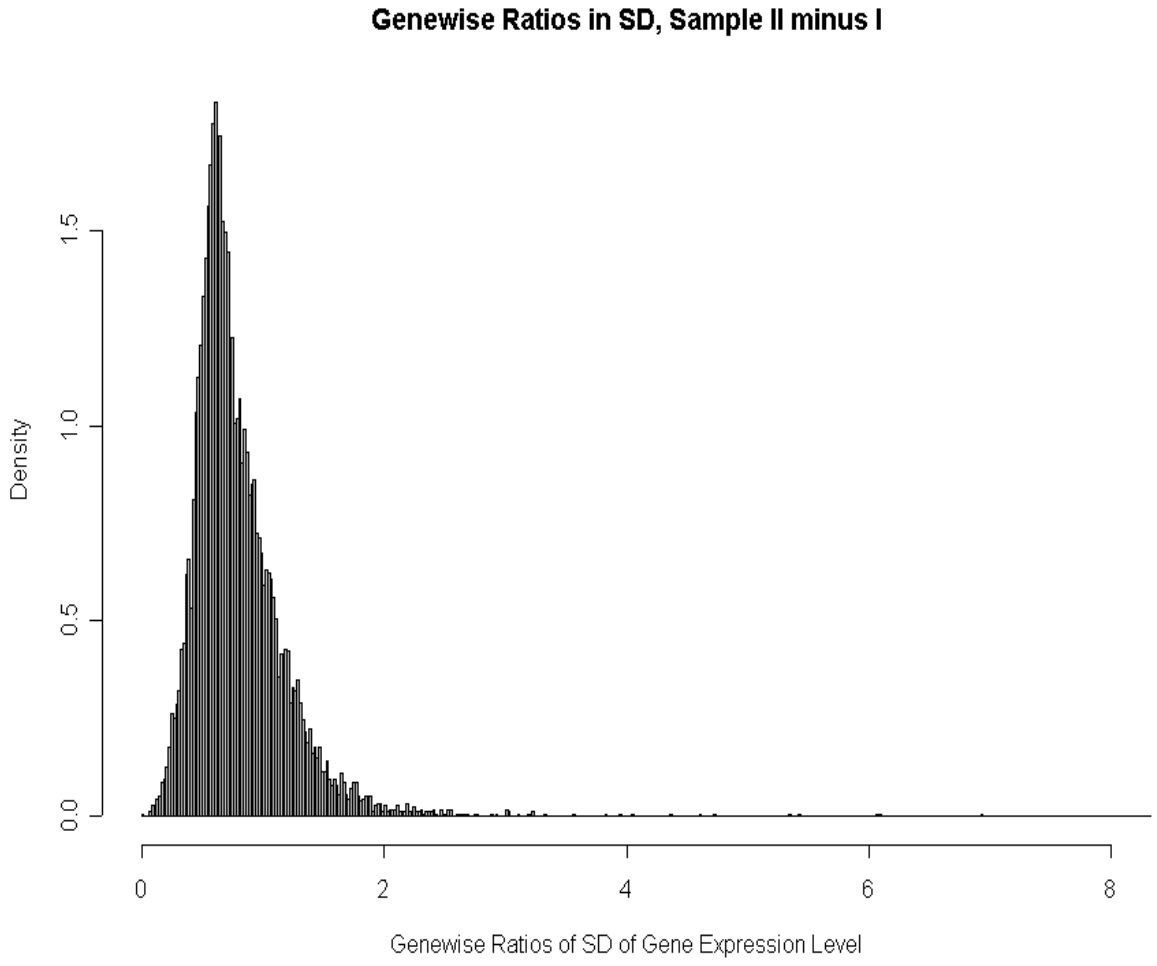

Figure 3 Genewise ratios of Standard Deviation (SD) of tumor sample divided by normal sample.

It is shown in Figure 3 that majority genes expressions from two samples are variance homogeneous (ratios around 1). But there are also part of genes have serious variance heterogeneity between two samples, which can be seen in the histogram (ratios near to 0 or near to 8). To describe the distributions more precisely, the third and forth moments (Skewness and Kurtosis) for the data are also calculated and shown in Figure 4:

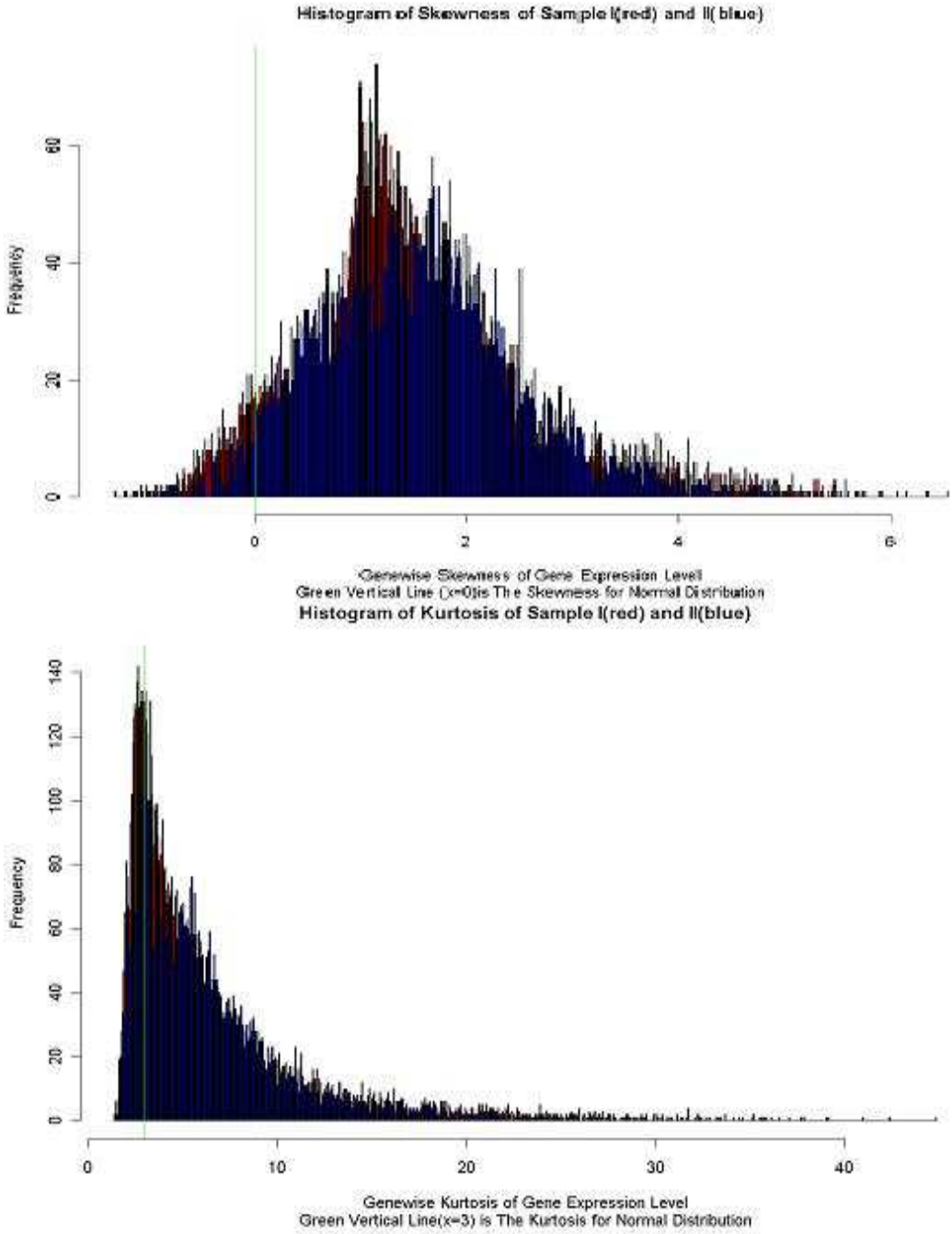

Figure 4 The histograms of Skewness (left) and Kurtosis (right), green lines show the corresponding value for the normality.

The histograms show that the Skewness and Kurtosis differ more or less between two samples. The genes expressions data in all the samples deviate from normality (see the green lines for skewness and kurtosis for normality). Since the skewness and kurtosis are correlated, the joint distributions of skewness and kurtosis are shown in Figure 5:

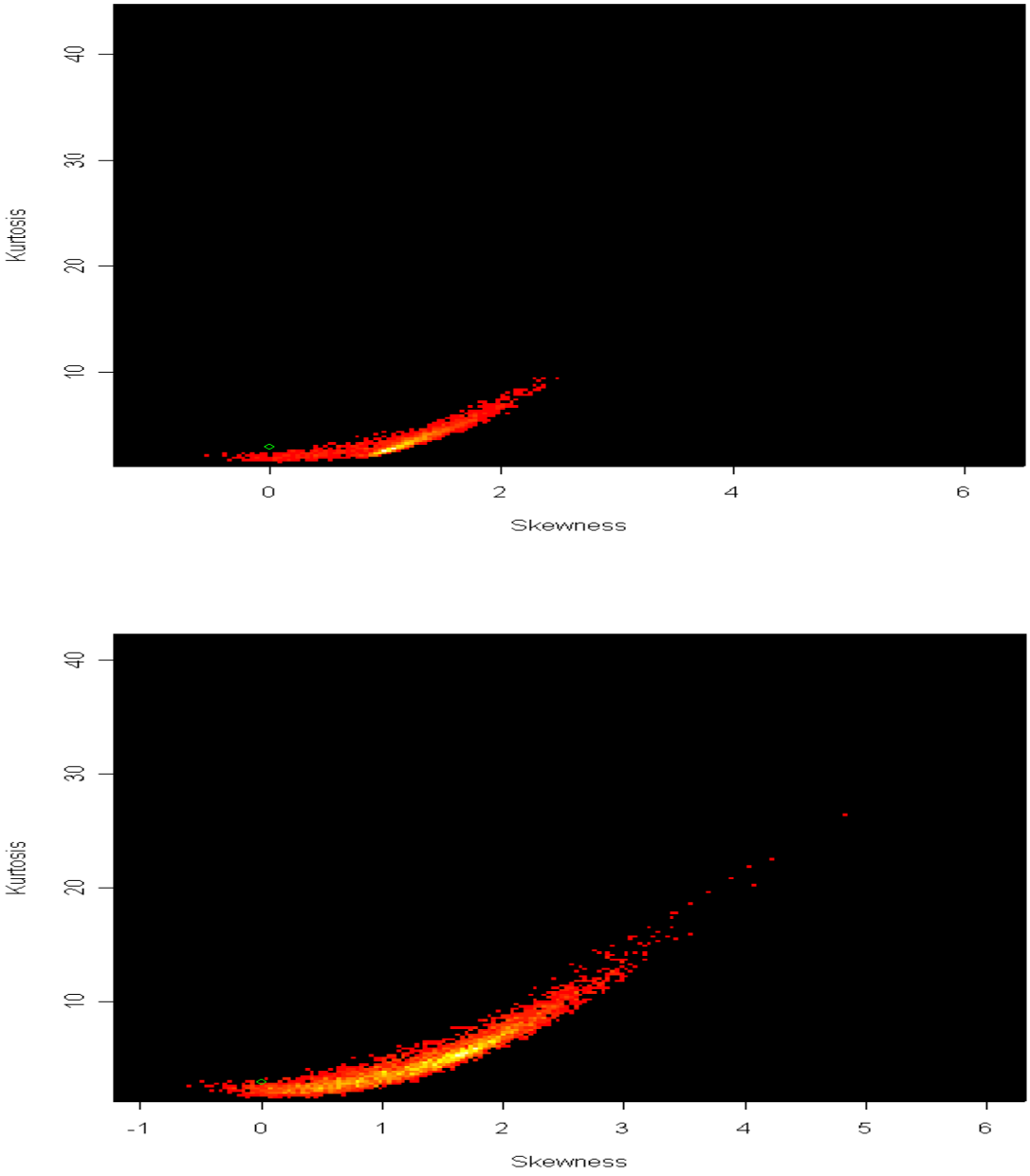

Figure 5 Joint distributions of skew and kurtosis of Normal sample (left) and tumor sample (right). The green points indicate the value (0,3) for normality.

From the joint distributions, it is easier to see that most of the gene expression data deviate from normality, and some of them are even highly skewed and with sharp distribution curves. The skewness locate between –1 to 2 most frequently, there are also highly skewed data, which is very likely the consequences of extreme values. After excluding the extreme values (usually the highest and lowest 0.5% of the array will be deleted empirically, thus the central

part of the distribution curve is better described by skewness and kurtosis), the skewness and kurtosis can be shown in Figure 6:

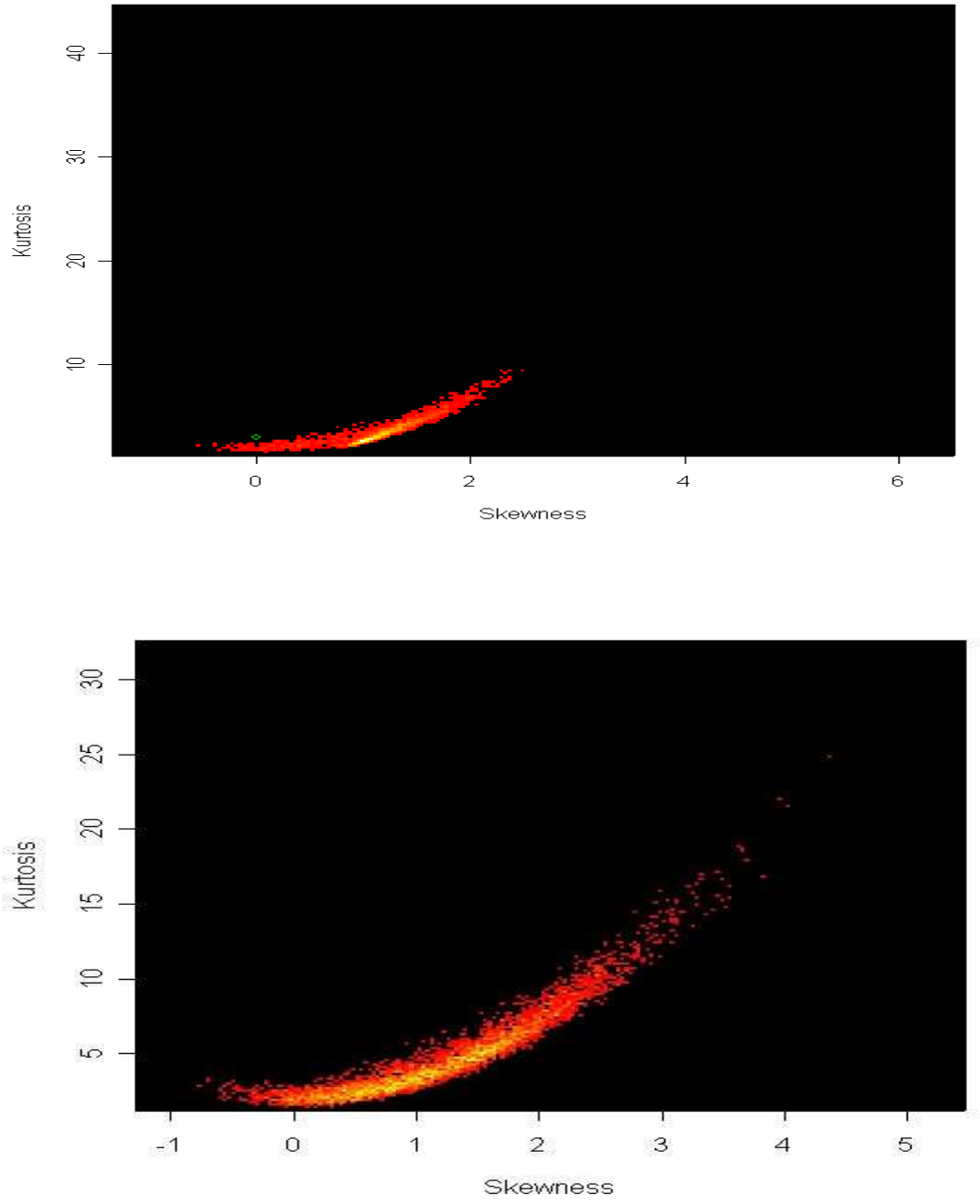

Figure 6 Joint distributions of skew and kurtosis of Normal sample (left) and tumor sample (right) after excluding outliers.

Actually the very extreme skew and kurtosis are stilled caused by extreme values in the dataset; this phenomenon can be modeled as mixture distribution, which will be discussed later in this section.

For detecting and visualizing the differentially expressed genes, Volcano Plot was proposed (Wolfinger *et al.*, 2001). The properties of Volcano Plot and comparison with other graphical presentations will be discussed in detail in section 5. The Volcano Plot for the example data is shown in Figure 7:

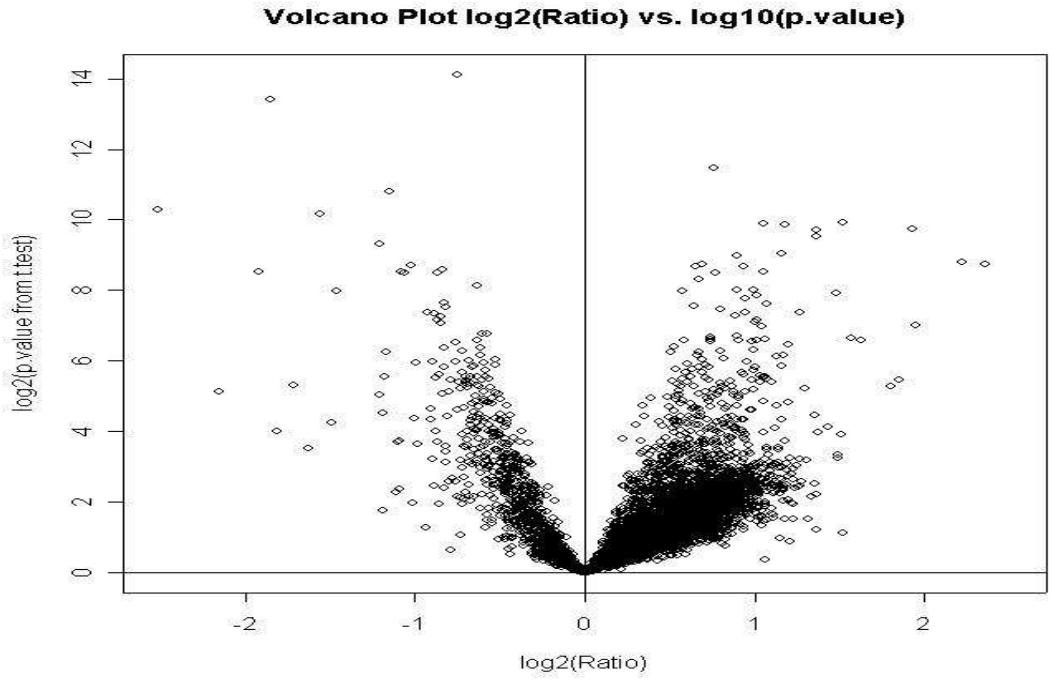

Figure 7 The Volcano Plot of Singh data with base-2 logarithm of fold change as abscissa versus minus base-10 logarithm of p values from t test.

The Volcano Plot also shows that there are majority of genes did not express differentially (i.e. small fold changes and large *p* values from the *t* test). But the usage of *t* test is doubted because the nonnormality nature of the data. P values from other nonparametric tests seem more appropriate for this purpose, which will be shown also in next section.

In horticulture science, the experimental data have also such problems. For example, the experiment conducted by Schneider and Tatilioglu (Schneider; Tatilioglu, 1996) is designed to study the protein band samples of two different Genotypes of chives. The protein band is generated through gel. Here the influence on the band samples under rising dosages (0ppm, 10ppm, 15ppm, 20ppm) of Tetracycline was shown for tetracycline sensitive and tetracycline insensitive plants. The endpoint is the integrals of the 18-kilo Dalton Mitochondria Protein band. The process of the integral values under rising dosage of Tetracycline is examined as a function of the Genotype. Per Genotype and dosage three repetitions are accomplished. The descriptive statistics of the dataset are shown in Table 2.

Table 2 Data from Schneider and Tatilioglu's experiments

Dose of Tetra.	0ppm	10ppm	15ppm	20ppm
Mean	665.00	587.17	408.08	225.38
SD	255.57	262.12	217.87	112.11

Tetra is the abbreviation of Tetracycline, SD means Standard Deviation.

With the increase of dose of Tetracycline, the mean and the standard deviation of the endpoints becomes much smaller. For the analysis of such data, the test should have the robustness against the variance heterogeneity.

Candidate Tests and Simulation Study

Let $X_{11}, X_{12}, \ldots X_{1n}$ be i.i.d. with distribution function $F_1(x)$ and $X_{21}, X_{22}, \ldots X_{2n}$ be i.i.d. with distribution function $F_2(x)$.

Parametric two-sample test

t test

When the data are normally distributed, further we assume that the variances are homogeneous, that is $\sigma^2{}_1 = \sigma^2{}_2 = \sigma^2$

Test statistics:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $S^2{}_{pool} = \frac{(n_1 - 1)S^2{}_1 + (n_2 - 1)S^2{}_2}{n_1 - n_2 - 2}$

$S^2{}_1$ and $S^2{}_2$ are unbiased estimators of σ^2 calculated from sample 1 and sample 2, respectively. Degree of freedom: $df = n_1 - n_2 - 2$. The statistic T has t distribution with *df* degrees of freedom, and we use this statistic for inferential purpose about the two population means under the above assumptions.

Welch *t* test

We assume here, $X_1 \sim \mathrm{N}(\mu_1, \sigma_1{}^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2{}^2)$, $\sigma_1{}^2 \neq \sigma_2{}^2$

Test statistics:

$T_{welch} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(df_{welch})$, where the approximate number of Degrees of freedom is

$$\mathrm{df}_{\mathrm{welch}} = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1-1) + \left(S_2^2/n_2\right)^2/(n_2-1)}$$

Modifications of the t test

Student *t* test, Welch *t* test (or use both *t* tests after logarithm of the raw data) are commonly used. The t test is a simple, statistically based method for detecting differentially expressed genes. In replicated experiments, the error variance can be estimated for each gene from the log ratios, and a standard t test can be conducted for each gene (Callow, 2000); the resulting t statistic or p values can be used to determine which genes are significantly differentially expressed. This gene-specific t test is not affected by heterogeneity in variance across genes because it only uses information from one gene at a time. It may, however, have low power because the sample size, i.e. the number of RNA samples measured for each condition is small. In addition, the variances estimated from each gene are not stable: for example, if the estimated variance for one gene is small, by chance, the t value can be large even when the corresponding fold change is small. It is possible to compute a global t test, using an estimate of error variance that is pooled across all genes, if it is assumed that the variance is homogeneous between different genes (Arfin, 2000). This is effectively a fold-change test because the global t test ranks genes in an order that is the same as fold change; that is, it does not adjust for individual gene variability. It may therefore suffer from the same biases as a fold-change test if the error variance is not truly constant for all genes.

As noted above, the error variance (the square root of which gives the denominator of the t tests) is hard to estimate and subject to erratic fluctuations when sample sizes are small. More stable estimates can be obtained by combining data across all genes, but these are subject to bias when the assumption of homogeneous variance is violated. Modified versions of the t test find a middle ground that is both powerful and less subject to bias.

In the ‘significance analysis of microarrays’ (SAM) version of the t test (known as the S test) (Tusher, 2001), a small positive constant is added to the denominator of the gene-specific t test. With this modification, genes with small fold changes will not be selected as significant; this removes the problem of stability mentioned above. The regularized t test (Baldi, 2001) combines information from gene-specific and global average variance estimates by using a

weighted average of the two as the denominator for a gene specific t test. The B statistic proposed by Lonnstedt and Speed (2002) is a log posterior odds ratio of differential expression versus non-differential expression; it allows for gene-specific variances but it also combines information across many genes and thus should be more stable than the t statistic. The t and B tests based on log ratios can be found in the Statistics for Microarray Analysis (SMA) package; the S test is available in the SAM software package; and the regularized t test is in the Cyber T package. In addition, the Bioconductor has a collection of various analysis tools for microarray experiments. Additional modifications of the t test are discussed by Pan (2002).

Nonparametric two-sample test

Asymptotic Wilcoxon test

Let r_{1i} be the rank of X_{1i} in the combined sample, that is X_{1i} is the r_{1i} th smallest in the combined sample, then $W = \sum_{i=1}^{n} r_{1i}$, the sum of ranks of the X_1's, is defined to be the Wilcoxon statistics (Wilcoxon, 1945). The null distribution of the statistic approximates normal when sample size is large.

Exact Wilcoxon test

Using the same Wilcoxon test statistic defined above, the null distribution of the statistic is replaced by exact distribution rather than normal approximation. It is used when sample size is small.

Maximally Selected Rank test

The maximally selected rank test is proposed for the classification problem, the idea is to use an optimal cut point to distinguish two samples. Let r_{1i} be the rank of X_{1i} in the combined sample, and $a_n(1)...a_n(n)$ denote some scores, μ is a pre specified cut point. A simple linear rank statistics (Hajek and Sidak, 1967, p.61) is defined as $S_{n\mu} = \sum_{i=1}^{n} c_\mu(X_i) a_n(r_{1n}) = \sum_{X_i \le \mu} a_n(r_{1n})$, where $c_\mu(X_i) = I_{X_i \le \mu}$ are regressors depending on μ. Furthermore, when the scores are set equal to the ranks, i.e. $a_n(i) = i$, $S_{n\mu}$ equals to the Wilcoxon statistic in section 3.3. The exact distribution of maximally selected rank statistics is derived by Hothorn and Lausen (2002).

Conover-Salsburg test

Conover and Salsburg (1988) investigate two kinds of Lehmann alternatives (Lehmann, 1953), which only a subset of treated sample will show an improvement. They proposed to use scores $a_n(i) = [i/(N+1)]^4$ for the statistic defined by Hajek and Sidak (1967) $S = \sum_{i=1}^{n} a_n(i)$.

Mood's Median Test

Mood's median test (Mood, 1950) is a nonparametric test, which is alternative to Wilcoxon test when variances of two samples are heterogeneous. The test statistic is M = the number of X_2 values that exceed the median of the combined samples (middle observation if $n_1 + n_2$ is odd, and average of the middles ones if $n_1 + n_2$ is even). The distribution of M is

$$P(M_{n1,n2} = m) = \frac{\binom{n_1}{\frac{n_1+n_2}{2} - m}\binom{n_2}{m}}{\binom{n_1+n_2}{\frac{n_1+n_2}{2}}},\ m = 0,1,...n_2.$$ A similar formula holds if $n_1 + n_2$ is odd.

Cramer-von Mises test

The Cramer-von Mises statistic, $W^2 = \int_{-\infty}^{\infty} [F_2(x) - F_1(x)]^2 \, dF_1(x)$, where $F_i(x)$ is the empirical CDF based on sample *i*,. The statistic was suggested independently by Cramer (1928) and von Mises (1931).

Kolmogorov Smirnov test

The Kolmogorov Smirnov test (Kolmogorov, 1973) based on the statistic $W = \max(F_2(x) - F_1(x))$ where $F_i(x)$ is the empirical CDF based on sample *i*,.

Proposed Maximum Test

In many applications, the precise form of the model underlying the data is not known; however, several scientifically plausible ones are available. Often optimal tests for each of them exist. Unfortunately, use of any one optimal test may lead to a loss of power under another model.

There are different solutions for such problem, such as adaptive inference when the models are very far apart (Bickel, 1982), Maximin efficiency robust test (Gastwirth, 1966, 1985) and Maximum test (Tarone, 1981; Fleming & Harrington, 1991)

A nonparametric test can be performed using a linear rank statistic

$$T = \sum_{i=1}^{N} g(i) V_i ,$$

where $g(i)$ are real valued scores, and $V_i = 1$ when the ith smallest of the $N = n + m$ observations is from the first sample and $V_i = 0$ otherwise. Two-sample tests based on T are distriubtion-free. Under H_0, we have

$$E(T) = \frac{n}{N} \sum_{i=1}^{N} g(i)$$

$$Var(T) = \frac{mn}{N^2 (N-1)} \left[N \sum_{i=1}^{N} g^2(i) - \left(\sum_{i=1}^{N} g(i) \right)^2 \right]$$

and the standardized statistic

$$\frac{T - E(T)}{\sqrt{Var(T)}}$$

follows asymptotically a standard normal distribution (Büning & Trenkler, 1994, pp. 127-130; Hajek et al., 1999, pp. 57-63).

3 optimal tests to use:

1. Gastwirth test G (short tails)

$$g(i) = \begin{cases} i - \dfrac{N+1}{4} & for \quad i \le \dfrac{N+1}{4} \\ 0 & for \quad \dfrac{N+1}{4} < i < \dfrac{3(N+1)}{4} \\ i - \dfrac{3(N+1)}{4} & for \quad i \ge \dfrac{3(N+1)}{4} \end{cases}$$

2. LT test (long tails)

$$g(i) = \begin{cases} -\left(\left[\dfrac{N}{4} \right] + 1 \right) & for \quad i < \left[\dfrac{N}{4} \right] + 1 \\ i - \dfrac{N+1}{2} & for \quad \left[\dfrac{N}{4} \right] + 1 \le i \le \left[\dfrac{3(N+1)}{4} \right] \\ \left[\dfrac{N}{4} \right] + 1 & for \quad i > \left[\dfrac{3(N+1)}{4} \right] \end{cases}$$

   Here, $[x]$ denotes the highest integer less than or equal to x.

3. Brunner (Brunner and Munzel, 2000) test (Nonparametric Behrens-Fisher Problem)
   To formulate a nonparametric Behrens-Fisher problem, we consider the relative treatment effect $p = P(X_{11} < X_{21}) + \frac{1}{2} P(X_{11} = X_{21})$.

The random variable X_{11} is called to *tend to smaller (larger)* values than the random variable X_{21} if $p > \frac{1}{2}$ $\left(p < \frac{1}{2}\right)$ and the two random variables are called tendentiously equal if $p = \frac{1}{2}$.

To estimate the relative treatment effect p and to derive its asymptotic distribution it is more convinient to express p in terms of the disbribution functions. To this end, we use the so-called *normalized version* $F_i(x) = \frac{1}{2}\left[F_i^-(x) + F_i^+(x)\right]$ of the distribution function (Ruymgaart, 1980) where $F_i^-(x) = P(X_{il} < x)$ is the left-continuous version and $F_i^+(x) = P(X_{il} \le x)$ is the right-continuous version of the distribution function. Then, the relative treatment effect p can be written as $p = \int F_1 dF_2$ and the hypothesis of no treatment effect is written as $H_0^p : p = \int F_1 dF_2 = \frac{1}{2}$. We note that $H_0^F : F_1 = F_2 = F$ implies $H_0^p : p = \frac{1}{2}$ because $\int F dF = \frac{1}{2}$, which follows from integration by parts.

To estimate the relative treatment effect p, the distribution functions F_1 and F_2 are replaced by their empirical counterparts $\hat{F}_i(x) = \frac{1}{2}\left[\hat{F}_i^-(x) + \hat{F}_i^+(x)\right]$.

$$\hat{F}^-{}_i(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c^-(x - X_{ik})$$

$$\hat{F}^+{}_i(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c^+(x - X_{ik})$$

$$\hat{F}_i(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c(x - X_{ik})$$

where

$$c^-(x) = \begin{cases} 0, & x \le 0 \\ 1, & x > 0 \end{cases} \text{ called left-continuous}$$

$$c^+(x) = \begin{cases} 0, & x < 0 \\ 1, & x \ge 0 \end{cases} \text{ called right-continuous}$$

$$c(x) = \frac{1}{2}\left[c^+(x) + c^-(x)\right] \text{ called normalized}$$

version of indicator function.

Let $H(X) = \sum_{i=1}^{2} \frac{n_i}{N} F_i(x)$ denote the combined distribution function and let $\hat{H}(X) = \sum_{i=1}^{2} \frac{n_i}{N} \hat{F}_i(x)$ denote the normalized version of the combined empirical distribution function. Note that $R_{ij} = N \cdot \hat{H}(X_{ik}) + \frac{1}{2}$ is the rank of X_{ik} among all

N observations. Let $\overline{R}_{i\cdot} = n^{-1}{}_i \sum_{k=1}^{n_i} R_{ik}, i = 1,2$, denote the mean of the ranks R_{ik} in the ith sample.

Then, it follows that

$$\hat{p} = \int \hat{F}_1 d\hat{F}_2 = \frac{1}{n_1}\left(\overline{R}_{2\bullet} - \frac{n_2 + 1}{2}\right)$$

is an unbiased and consistent estimator for the relative treatment effect p.

The statistic $\sqrt{N}\left(\hat{p} - \frac{1}{2}\right) \Big/ \sigma_N = \left(\overline{R}_{2\bullet} - \overline{R}_{1\bullet}\right) \Big/ \sqrt{N\sigma_N^2}$ has, asymptotically, a standard normal distribution under $H_0^p : p = \frac{1}{2}$, where

$$\sigma_N{}^2 = N\left[\sigma_1^2 / n_1 + \sigma_2^2 / n_2\right].$$

The variances σ_1^2 and σ_2^2 are unknown and must be estimated from the data.

$$\hat{\sigma}_i^2 = S_i^2 / \left(N - n_i\right)^2,$$

where

$$S_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(R_{ik} - R_{ik}^{(i)} - \overline{R}_{i\bullet} + \frac{n_i + 1}{2}\right)^2$$

is the empirical variance of $R_{ik} - R_{ik}^{(i)}$, $R_{ik}^{(i)} = N \cdot \hat{H}\left(X_{ik}\right) + \frac{1}{2}$ denotes the (within) rank of X_{ik} among n_i observations within the ith sample.

Thus, under H_0^p, the statistic

$$W_N^{BF} = \frac{\sqrt{N}\left(\hat{p} - \frac{1}{2}\right)}{\hat{\sigma}_N} = \frac{1}{\sqrt{N}} \cdot \frac{\overline{R}_{2\bullet} - \overline{R}_{1\bullet}}{\hat{\sigma}_N}$$

has, asymptotically , a standard normal distribution under hypothesis

The asymptotical distribution of the Maximum of the above three standardized statistics is possible to derive. Assume that the standardized optimal statistics are asymptotically jointly multivariate normal with correlation matrix $\{\rho_{ij}\}$. And the joint density function of the three standardized statistics is $f(x_1, x_2, x_3)$. Define the maximum of X_1, X_2, X_3 are X_m, and when $X_m \le x$, it means each of $X_i \le x, i = 1,2,3$. Then the Asymptotic distribution of X_m is just the integral $\int_{X_i < x_m} f(x_1, x_2, x_3) dx_1 dx_2 dx_3$. But the asymptotic distribution of a maximum statistic may be not available, or, if available, the asymptotic approximation can be poor (Freidlin & Korn, 2002). Therefore, this test can, for large sample-sizes be performed simulation-based only, i.e. Bootstrap or permutation-based. The Bootstrap or permutation distribution of the

maximum statistic can be obtained with resampling with or without replacement respectively. For example, in the permutation case, the data from two samples are pooled together, and in each resampling, generate a permutation of the pooled data, and take the first n_1 numbers as the first group and the next n_2 as second group. Perform the test and save the statistics. Repeat this process for N, say 4,000, times. We can use the 4,000 stored statistics to generate the empirical distribution of the statistic, and then the decision of the test can be made from the observed statistic depending on the empirical distribution. The only difference between a permutation resampling and Bootstrap resampling is that Bootstrap resampling sample from the raw data with replacement.

Distribution systems

To model the nonnormality of the real data, distribution systems are needed. The suggestion to use "typical" nonnormality, such as lognormal, Beta, Gamma, Weibull, student t distributions etc., has been made by Pearson and Please (1975). In the literature, there appear other methods for generating nonnormality, such as adding outliers, using extreme nonnormality (chi square, rectangular, lognormal, exponential, t, Cauchy distributions), transformation to unknown nonnormality and Tabular.

But the above method is either hard for the researcher to manipulate the distribution parameters (mean, variance, skewness, kurtosis, etc.) or hard to implement in Monte Carlo simulation. Thus, some other distribution systems are adapted for generating nonnormality in the simulation study, which fulfill the following requirements: they should have a priori known parameter, enable the researcher to change distributions with the least amount of difficulty, be realistic simulations of empirical distributions, capable of generating widely different distributions, and should operate as efficiently as possible.

The first used system is Fleishman system (Fleishman, 1978). The idea behind is a polynomial transformation, and will be called the power method. The transformation is of the form $Y = a + bX + cX^2 + dX^3$, where X is a random variate distributed normally with zero mean and unit variance, $N(0,1)$, Y will have a distribution dependent upon the constants. With the restriction of mean, variance, skew and kurtosis, the four coefficients (*a*, *b*, *c* and *d*) can be found (for details see Fleishman, 1978). Some of the coefficients for certain combination of skew and kurtosis are tabulated, which are used in this report. One thing to

notice is the limitation of Fleishman system, the possible space of the skew and kurtosis can be described by a parabola: $skew^2 < 0.0629576 \times kurtosis + 0.0717247$.

Histograms of some nonnormal distributions generated by Fleishman system are shown in Figure 8:

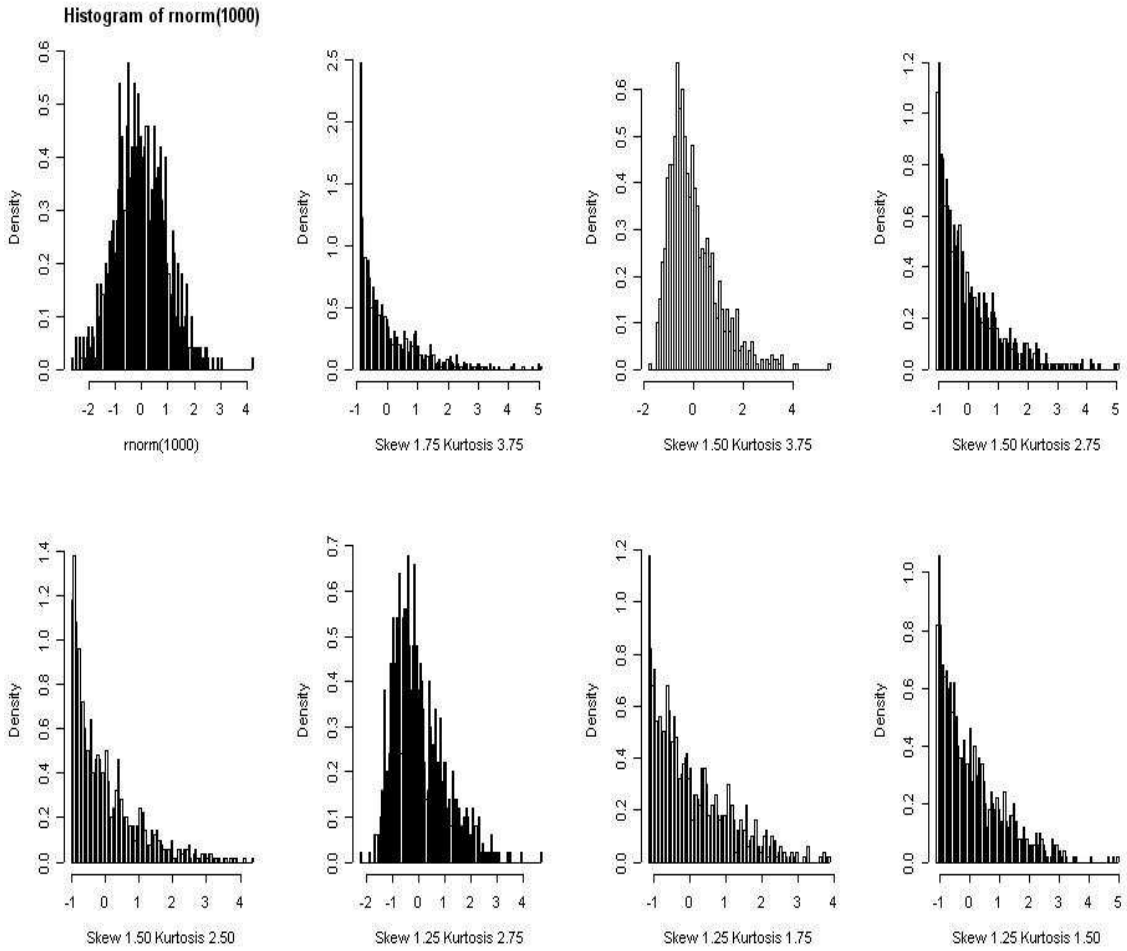

Figure 8 Some distributions generated by Fleishman system

Due to the limitation of Fleishman system, other distribution systems are used; the second used system is Johnson distribution. Starting from a continuous random variable Z whose distribution is unknown and is to be approximated and subsequently sampled, Johnson (1949) proposed a set of four normalizing translations. These translations have the general form

$$X = \gamma + \delta \cdot g\left(\frac{Z-\xi}{\lambda}\right)$$

Where X is a standard normal random variate (that is $X \sim N(0,1)$), γ and δ are shape parameters, λ is a scale parameter, ξ is a location parameter, and $g(\cdot)$ is a function whose form defines the four distribution families in the Johnson translation system,

$$g(y)=\begin{cases}\ln(y) & for\ S_L\ \ (\log normal)\ \ Family\\ \ln\left[y+\sqrt{y^2+1}\right] & for\ \ S_U(unbounded)\ \ Family\\ \ln(y/(1-y)) & for\ \ \ S_B(bounded)\ \ Family\\ y & for\ \ \ S_N(normal)\ \ Family\end{cases}$$

In the support of all these four functions, it can be proven to be monotonic increasing.

The random variable Z has the following characteristics:

Cumulative distribution:

$$F_Z(z)=\Phi\left(\gamma+\delta\cdot g\left(\frac{z-\xi}{\lambda}\right)\right)$$

Density function:

$$f_Z(z)=\delta\cdot g'\left(\frac{z-\xi}{\lambda}\right)\phi\left(\gamma+\delta\cdot g\left(\frac{z-\xi}{\lambda}\right)\right)$$

According to Hill *et al.* (1976), the Johnson curves can be fitted using Moment Matching Estimate, which is implemented by in Software R, package *SuppDists*.

As shown before, the extreme skew and kurtosis are caused by the extreme values exists in the dataset. In fact that not only unimodal distribution exists in real data set, such extreme values can also be understood as another components of the distribution, consequently the distribution system should be able to generate bimodal or multimodal distributions. Mixture distribution is used for this purpose, though Johnson system has the same functionality. In this study, only two normal components are used. Suppose $X_1 \sim N(\mu_1,\sigma_1^{\,2})$ has the (Probability Density Function) PDF $f_1(x_1)$, and $X_2 \sim N(\mu_2,\sigma_2^{\,2})$ has the PDF $f_2(x_2)$. Since the support of X_1 and X_2 are $(-\infty,\infty)$, if X is a random variable which may come from the above two populations, the support of X is also $(-\infty,\infty)$. Random variable X has the PDF $f(x)=af_1(x)+(1-a)f_2(x)$. (a is called mixing probability or proportion, a of the times from the first component, *1-a* of the times from the second component.)

Since:

$$\int_{-\infty}^{\infty} f(x)dx=\int_{-\infty}^{\infty}(af_1(x)+(1-a)f_2(x))dx=\int_{-\infty}^{\infty}af_1(x)dx+\int_{-\infty}^{\infty}(1-a)f_2(x)dx=a+(1-a))=1$$

The mean of X is $E(X)=a\mu_1+(1-a)\mu_2)$

The variance of X is $V(X)=a(\sigma_1^{\,2}+\mu_1^{\,2})+(1-a)\left(\sigma_2^{\,2}+\mu_2^{\,2}\right)-(a\mu_1+(1-a)\mu_2)^2$

Two mixture distributions are shown in Figure 9:

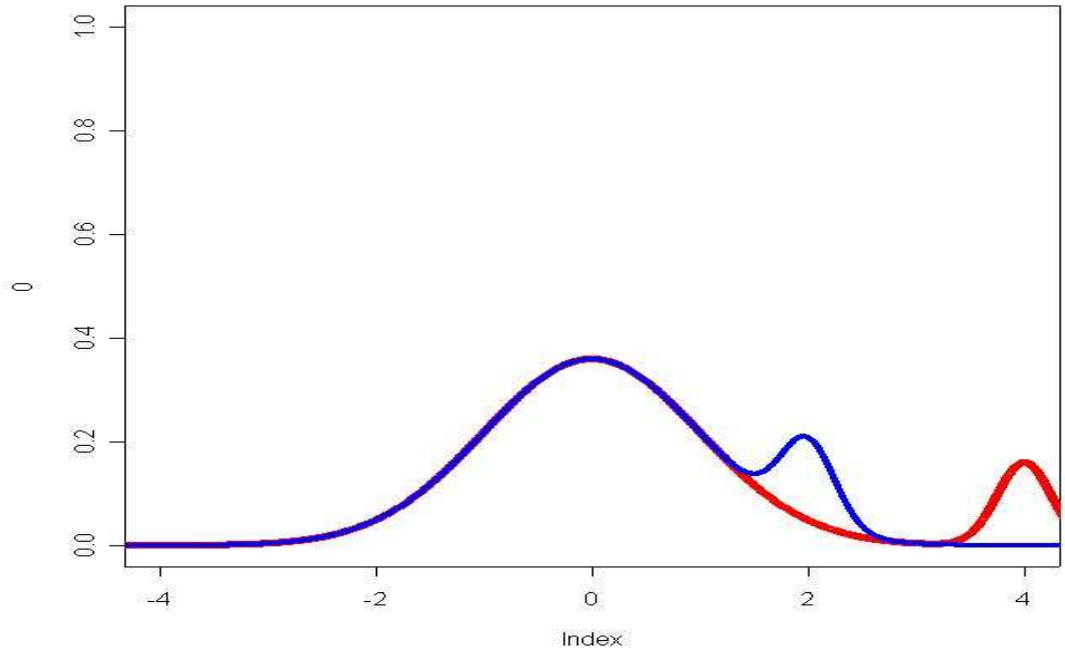

Figure 9 Two mixture distributions. Blue one with 0.9*N*(0,1) +0.1*N*(2,0.25), and red one with 0.9*N*(0,1)+0.1*N*(4,0.25).

The above introduced distribution system are flexible to model and simulate the real data, but the mechanism of generating the real data might be more complex, the simple transformation and mixture might not be sufficient for investigation.

Monte-Carlo simulation

Outline of Monte-Carlo Simulation

The different statistical tests can be compared via Monte-Carlo simulation. In the simulation, the random variate will be generated under different specific conditions. For example, when the type I error of *t* test is under investigation. The random data should be generated under the null hypothesis condition. The underlying distribution of the data is set to be normal distribution with equal variances. Use software R for instance, the normally distributed random number can be generated using command rnorm with other necessary arguments. The *t* test is also a function in software R with command name t.test. Conduct *t* test for the generated random variate *N*, say 10,000, times and count the number of significant results. The number of significant results divided by the simulation number 10,000 is the estimation of the type I error rate of t test under the null hypothesis. If one of the two samples is shifted

from the other, the power of the sample can be also estimated using the same process. The confidence intervals of the estimated type I error rate and the power can be also calculated using the inference for proportions. The estimated type I error and the power are denoted by *p*, the confidence interval is $p \pm z\sqrt{\frac{p(1-p)}{N}}$, where z depends on the level of confidence desired.

Generating of Random Number

There are different ways for the random number generation. Here only a general idea of generating random variables of different types is given. This is done by simply determining values of the uniform variable. Consider the following fact:

Let random variable *Y* be distributed normally over the unit interval $0 < y < 1$. Suppose that $F(x)$ is a distribution function of the continuous type, which is strictly increasing when $0 < F(x) < 1$. If the relationship $Y = F(X)$ is defined, the inequalities $X \le x$ and $F(X) \le F(x)$ are equivalent. Thus, with $0 < F(x) < 1$, the distribution of X is

$$\Pr(X \le x) = \Pr[F(X) \le F(x)] = \Pr[Y \le F(x)]$$

because $Y = F(X)$. However, $\Pr(Y \le y) = G(y)$, so we have

$$\Pr(X \le x) = G[F(x)] = F(x),\ 0 < F(x) < 1.$$

That is, the distribution function of X is $F(x)$. This result permits us to simulate random variables of different types. This is done by simply determining values of the uniform variable *Y*, usually with a computer.

But in the simulation of random variables using uniform random variables, it is frequently difficult to solve $y = F(x)$ for *x*. Thus other methods are necessary. For instance, consider the important normal case in which we desire to determine *X* so that it is $N(0,1)$. Of course, once *X* is determined, other normal variables can then be obtained through *X* by the transformation $Z = \sigma X + \mu$. To simulate normal variables, Box and Muller (Hogg, Craig, 1995) suggested the following transformations, let Y_1, Y_2 be a random sample from the uniform distribution over $0 < y < 1$. Define X_1 and X_2 by

$$X_1 = (-2\ln Y_1)^{1/2}\cos(2\pi Y_2)$$

$$X_2 = (-2\ln Y_1)^{1/2}\cos(2\pi Y_2)$$

The random variables X_1 and X_2 are proved to be independent standard normal random variables. For the proof see for example Hogg and Craig.

Simulation Results and Discussion

Condition I: Normality with variance heterogeneity

From the investigation of Microarray datasets, there are roughly 25% genes fulfil the assumption of normality (via Kolmogorov-Smirnov testing). But the variances are heterogenous for some of the genes. Thus the first condition considered here is Normality data with variance heterogeneity.

The simulation is designed with sample size 10 only; both samples are normally distributed with mean difference 0,0.5,1,1.5,2,2.5,3, standard deviance ratio 1, 1.5, 2, 3, 5, 7,10. Three candidate tests, Welch *t* test, Exact Wilcoxon test and maximally selected test are used. 10,000 replications and 0.05 significant level are used.

The power matrices for three tests are shown in table 3-5:

Table 3 Power matrix for Welch *t* test. SDR means standard deviance ratio, MD means mean difference.

SDR / MD	1	1.5	2	3	5	7	10
0	0.050	0.056	0.044	0.063	0.050	0.045	0.049
0.5	0.172	0.120	0.119	0.075	0.057	0.058	0.054
1	0.555	0.359	0.254	0.138	0.087	0.066	0.071
1.5	0.868	0.707	0.478	0.279	0.127	0.079	0.088
2	0.985	0.898	0.739	0.440	0.203	0.130	0.076
2.5	1.000	0.983	0.920	0.628	0.284	0.189	0.111
3	1.000	0.998	0.981	0.780	0.378	0.238	0.143

Table 4 Power matrix for maximally selected rank test.

SDR / MD	1	1.5	2	3	5	7	10
0	0.059	0.072	0.111	0.213	0.336	0.451	0.490
0.5	0.161	0.125	0.170	0.232	0.375	0.435	0.530
1	0.471	0.338	0.289	0.312	0.377	0.468	0.537
1.5	0.773	0.633	0.517	0.405	0.470	0.451	0.555
2	0.937	0.773	0.718	0.612	0.538	0.532	0.546
2.5	0.934	0.940	0.834	0.762	0.579	0.578	0.575
3	0.958	0.883	0.940	0.851	0.683	0.601	0.624

Table 5 power matrix for Exact Wilcoxon test.

SDR / MD	1	1.5	2	3	5	7	10
0	0.045	0.053	0.049	0.075	0.088	0.077	0.094
0.5	0.155	0.115	0.106	0.083	0.071	0.079	0.093
1	0.511	0.327	0.240	0.139	0.101	0.123	0.105
1.5	0.833	0.660	0.446	0.273	0.165	0.113	0.123
2	0.979	0.869	0.698	0.438	0.226	0.159	0.120
2.5	1.000	0.980	0.884	0.616	0.294	0.240	0.167
3	1.000	0.995	0.965	0.751	0.392	0.278	0.190

The simulation results show that when the variances are homogeneous, all the tests control the alpha. But even with no mean difference but variance heterogeneity, the Welch *t* test is the best candidate because the alpha is always controlled regardless of the ratio of standard deviation. The Exact Wilcoxon and Maximally Selected test behaves too liberal in this case.

Condition II: Nonnormality with Fleishman System

The distribution is skewed from the investigation of real microarray data, after deleting the extreme values; the skew and kurtosis are limited to a domain that Fleishman can be used to describe the shapes of the underlying density curves. In this scenario, five tests (Welch t test, Asymptotic Wilcoxon test, Exact Wilcoxon test, Maximally selected test and Mood's median test) are compared under normal (N(0,1) vs. N(0.75,1), sample size 25 with expected power 75%), Fleishman with Skew 1.5 Kurtosis 3.75 and Fleishman with Skew 2 and Kurtosis 7 distributions. Sample sizes (25,25) (20,20) (15,15) (10,10) are used.

The simulation results under three distributions can be seen in Table 6-8:

Table 6 Power matrix for five tests under normality.

Under the null					
Sample size	T	MS	EW	AW	Mood
25:	0.0510	0.0530	0.0534	0.0560	0.0568
20:	0.0488	0.0542	0.0482	0.0482	0.0146
15:	0.0516	0.0602	0.0486	0.0486	0.0164
10:	0.0464	0.0536	0.0416	0.0416	0.0104
Under the alternative					
25:	0.7442	0.6046	0.7098	0.7162	0.6510
20:	0.6374	0.5250	0.6186	0.6186	0.3578
15:	0.5022	0.4274	0.4708	0.4708	0.2764
10:	0.3398	0.2982	0.3070	0.3070	0.1684

Table 7 Power matrix for five tests under Fleishman with skew 1.5, kurtosis 3.75

Under the null					
Sample size	T	MS	EW	AW	Mood
25:	0.0500	0.0522	0.0484	0.0502	0.0570
20:	0.0524	0.0534	0.0524	0.0524	0.0150
15:	0.0470	0.0624	0.0494	0.0494	0.0188
10:	0.0404	0.0540	0.0430	0.0430	0.0104
Under the alternative					
25:	0.7492	0.8832	0.8570	0.8602	0.7602
20:	0.6472	0.7854	0.7658	0.7658	0.4692
15:	0.5410	0.6788	0.6216	0.6216	0.3732
10:	0.3962	0.4396	0.4364	0.4364	0.2320

Table 8 Power matrix for five tests under Fleishman with skew 2, kurtosis 7

Under the null					
Sample size	T	MS	EW	AW	Mood
25:	0.041	0.050	0.040	0.043	0.049
20:	0.045	0.052	0.052	0.053	0.015
15:	0.046	0.049	0.038	0.038	0.019
10:	0.038	0.055	0.047	0.047	0.014
Under the alternative					
25:	0.735	0.951	0.914	0.918	0.839
20:	0.6472	0.7854	0.7658	0.7658	0.4692
15:	0.560	0.800	0.709	0.709	0.465
10:	0.391	0.526	0.494	0.494	0.289

The simulation results above show that under the null, regardless of the distributions all the candidate tests controls alpha. When the data is normally distributed, the t test is proven to be the best test for all sample sizes. But when the data is skewed, Maximally selected test behaves the best, for example as shown in Table 7, the power of MS test under all sample sizes are much higher than t test and also slightly higher than Wilcoxon test. Mood's median test works well in moderate sample sizes (25-15), but the power decreases dramatically when the sample size falls to be 10, which is not surprising since more than 12 observations is recommended in the literature.

Condition III: Mixture of two Normals

Conover-Salsburg test is a good candidate test when only a subset of treated sample will show an improvement. The simulation for this test is designed as Sample Size 25 balanced, with 5000 simulation replicates.

The simulation results of this test show that it controls alpha, 0.0476, under the null. And when two samples are $N(0,1)$ and. $N(0.75,1)$, the power of Conover-Salsburg test is 0.6216, whereas the expected power for t test is 0.75. When the two samples are designed to $N(0,1)$ and $0.7N(0,1)+0.3N(5,1)$, The power of CS test is 0.7566 whereas *t* test has the power of 0.2423.

Condition IV: Simulation for Maximum Test

Since the characteristics of microarray data include not only one problem at a time, the most powerful test in each case cannot work well for all the genes in microarray data. The Maximum of some standardized statistics is proposed. The simulation study is conducted to check the alpha robustness and the power of this test compared with t test and Wilcoxon test for sample size 25. The simulation number are set to be 1,000 since the Bootstrap procedure are very computation intensive. The Bootstrap replication is 400 for the Maximum test.

The 3 components of Maximum Test are simulated separately under normality to check the alpha and power.

The results are shown in table 9:

Table 9 Simulation study for the 3 components of Maximum Test

Test / Distribution	Gastwirth Test	Long Tail Test	Brunner Test	Expected alpha and Power of *t* test
N(0,1) vs. N(0,1)	0.0476	0.049	0.0457	0.05
N(0,1) vs.(0.75,1)	0.659	0.651	0.749	0.75

The simulation results shows that each test controls alpha in normal case, but the *t* test is the most powerful test. But Brunner Test is a very good test, which controls the type I error and gives almost the same power like *t* test.

Since the Gastwirth test is powerful when the data have short tail (small kurtosis), the LongTail test is powerful when the data have long tail (larger kurtosis) and Brunner Test is powerful for the non-parametric Behrens-Fisher problem. The combination of these three tests is hoped to have better performance than other candidate tests.

Under the normality with homogeneous variance the Maximum Test is proved to control alpha and has the power smaller than *t* test, which can be seen in table 10.

Table 10 Type I error and power of Maximum test under normality and variance homogeneity

Test / Distribution	Maximum Test	Expected alpha and power of Welch *t* test
N(0,1) vs. N(0,1)	0.0525	0.050
N(0,1) vs. N(0.75,1)	0.6848	0.750

Again the results show that the *t* test is more powerful than Maximum Test when data is normally distributed, and the reason why the Maximum Test is less powerful than *t* test is that some price must be paid for the Gastwirth Test and the TongTail Test, which are not as powerful as the *t* test or the Brunner Test.

Under the variance heterogeneity, the type I error and power of Maximum Test can be shown in table 11:

Table 11 Type I error and power of Maximum test under normality and variance heterogeneity.

Test / Distribution	Maximum Test	Welch *t* test	Exact Wilcoxon Test
N(0,1) vs. N(0,2)	0.057	0.046	0.074
N(0,1) vs. N(0.75,2)	0.324	0.371	0.349

The simulation result shows that under normality and variance heterogeneity, exact Wilcoxon test does not control alpha, whereas Maximum Test and Welch *t* test has the type I error near to the nominal 0.05. The power of Welch *t* test is higher than Maximum Test in this case.

When the data is nonnormal, for example Fleishman distribution with skew 2 and kurtosis 7, the simulation results of Maximum test can be shown in table 12:

Table 12 Type I error and power of Maximum under nonnormality and variance heterogeneity.

Test / Distribution	Maximum Test	Welch *t* test	exact Wilcoxon
Fleishman(0,1,2,7) vs. Fleishman(0,1,2,7)	0.033	0.037	0.042
Fleishman(0,1,2,7) vs. Fleishman(0,2,2,7)	0.158	0.056	0.172
Fleishman(0,1,2,7) vs. Fleishman(0.75,1,2,7)	0.936	0.745	0.912
Fleishman(0,1,2,7) vs. Fleishman(0.75, 2,2,7)	0.270	0.356	0.267

The simulation results show that under nonnormality (right skewed) and variance homogeneity, the Maximum Test, Welch t test and exact Wilcoxon test controls alpha, but the Maximum Test and Welch t test is conservative. Under variance heterogeneity, only Welch t test controls alpha, both the Maximum Test and exact Wilcoxon test are anticonservative. Under variance homogeneity, the Maximum Test has the highest power among three tests. But under variance homogeneity, Welch *t* test performs best which guarantee the type I error and have higher power than other two tests.

This result implies that in the situation of nonnormality and variance heterogeneity, Welch *t* test is the best candidate test though the assumption of normality is violated. The Maximum Test is conducted with more realistic assumption, but it does not control the type I error under this condition.

Graphical Presentation of the Test Results

The test results, such as *p* values or the confidence interval of some parameters can be presented graphically. On is the so-called Volcano Plot.

Volcano plot

The test results can be presented efficiently using the so-called Volcano Plot, Volcano Plot is nothing but a scatter plot of the base 2 logarithm of the ratio of means versus the base 10

logarithm of the p value from the statistical test. One disadvantage of the Volcano Plot is that the distance of two points is hard to interpret since the x- and y-axis are in two different scales. The microarray data will be used to illustrate this kind of plot.

The identification of the important differentially expressed genes from massive amounts of microarray data is an interesting and current biostatistics problem. Different approaches have been proposed, such as testing procedures, e.g. significance analysis of microarrays (Tusher et al. 2001) and graphical tools, e.g. the MAplot (shows log fold-change as a function of mean log expression level) (Cope et al., 2004). The volcano plot (Wolfinger et al., 2001), a combination of testing and graphical approaches, is a simple scatter plot, where p is the two-sided p-value of the common two-sample t-test. In Figure 1 such a plot is given for the Affymetrix-type oligonucleotide arrays by Shipp et al. (2002) and the interpretation is as follows: the abscissa <0 indicates under expression, >0 over expression, the ordinate indicates non-significant finding and significant findings, where the nominal or multiplicity-adjusted false positive error rates can be used. Therefore, genes can be identified which are both significant (low p-values) and relevant (high log-ratios). This plot implicitly assumes the approximate validity of the Gaussian distribution. Although probe (expression) level data were pre-processed several-fold (normalized, log-transformed), doubts on the Gaussian distribution and variance homogeneity assumption exist. One example of this plot can be seen in figure 10.

Modified volcano plot

The reason for graphing statistical significance (p-value) versus biological relevance (ratio) needs an explanation, particularly because the p-value alone is currently the gold standard for reporting statistical comparisons between treatments and controls in bio-medical publications. The p-value is a single probability [0, 1] estimated from the effect difference, variance, sample size, and based on the fulfillment of the underlying test assumptions. Why is the simultaneous consideration of significance and relevance particularly important for microarray data? The objective is the identification of highly over- or under-expressed genes. Although the same design is used for all genes, different sample sizes and different variances occur at evaluation, e.g. the plotted lymphoma data (Shipp et al., 2002) use 19 chips each, but the sample sizes and variances for gene S62696_s_at (unigene-ID) are $n_1 = 5$, $n_2 = 11$ and $\sigma_1^2 = 142237, \sigma_2^2 = 22904$. Also, the distribution between different genes in the same experiment may be different. Therefore, using the t-test based on its p-value alone (even after

log-transformation) can lead to serious misclassifications. A further question that arises is if the presentation of significance versus relevance is appropriate when the significance is obtained from a t-test of the difference ($\overline{x}_{Treatment} - \overline{x}_{Control}$) and the relevance is represented by the ratio $\frac{\overline{x}_{Treatment}}{\overline{x}_{Control}}$. The consideration of the ratio seems to be biologically appropriate even if for log-transformed data the original multiplicative model is transferred into an additive model. However, if the ratio is an appropriate measure for relevance, then the use of the p-value from a parametric test for ratios is consequent. This test is according to Sasabuchi (1988):

$$t_i = \frac{\overline{X}_{Treatment} - \theta \overline{X}_{Control}}{MSE \sqrt{\frac{1}{n_{Treatment}} + \frac{\theta^2}{n_{Control}}}} \propto t_{df,1-\alpha/2}$$

where MSE denotes the common mean square error estimator, *df* is the degrees of freedom $df = n_{Treatment} + n_{Control} - 2$ and $t_{df,1-\alpha/2}$ the quantile of the t distribution. In comparison with the common t-test, this test inherently needs an a-priori definition of a threshold θ, e.g. θ =2, the so-called 2-fold rule. Furthermore, the question of whether a xy-graph of p-value vs. ratio estimate is appropriate arises. The ratio represents a percentage of k-fold change, and the p-value a probability. We propose the presentation of statistical significance by the upper/lower limit of the confidence interval instead of the p-value, because confidence intervals offer information about the distance from the null-hypothesis (distance to 1), the direction of the effect (larger/smaller than 1), and the variability (width) simultaneously. Although confidence intervals for the difference are frequently used in biomedical research, the ratio-to-control confidence intervals can be directly medically interpreted for some problems (Feuerstein et al. 1997). Sometimes the ratio problem is transformed via log-transformation into a difference problem which assumes log-normal distributed endpoints. A two-sided parametric confidence interval for a ratio according to Fieller (1954) is:

$$\theta_{upper} = \frac{\overline{x}_{Control}\overline{x}_{Treatment} + \sqrt{a\overline{x}^2_{Treatment} + b\overline{x}^2_{Control} - ab}}{\overline{x}^2_{Control} - a}$$

$$\theta_{lower} = \frac{\overline{x}_{Control}\overline{x}_{Treatment} - \sqrt{a\overline{x}^2_{Treatment} + b\overline{x}^2_{Control} - ab}}{\overline{x}^2_{Control} - a}$$

$$a = \frac{MSE}{n_{Control}} t^2_{df=n_{Control}+n_{Treatment}-2,1-\alpha/2} \text{ and } b = \frac{MSE}{n_{Treatment}} t^2_{df,1-\alpha/2}$$

The side condition $\bar{x}^2_{Control} > a$ is simply a one-sided test for control mean values larger than zero, i.e., this approach is limited to non-zero control effects.

Sometimes skewed data, multimodal distributed data or data with outliers requiring a non-parametric confidence interval can be observed in microarray data. According to Hothorn and Munzel (2002), the two-sided non-parametric confidence interval for the ratio $\delta = \frac{med(x_{i,Treatment})}{med(x_{i,Control})}$ is $\left[\delta_{(w)}; \delta_{(n_{Treatment} n_{Control} - w + 1)}\right]$ where $w_{n_{Treatment}, n_{Control}, 1-\alpha/2}$ denotes the lower quantile of the Wilcoxon test. The asymptotic or exact confidence interval can be estimated using Hodges-Lehman (1963) confidence intervals with the R-package *exactRankTest* after log-transformation of the raw data. One example can be see in the following figure:

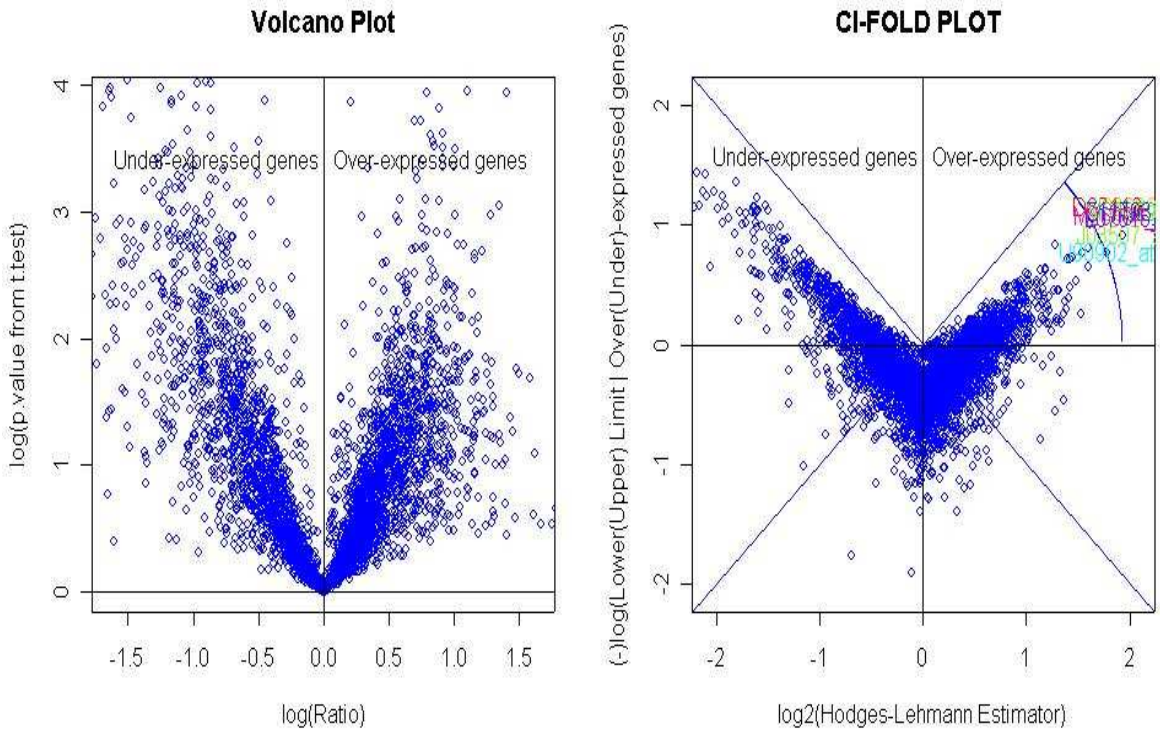

Figure 10 Original (a) and modified volcano plot (b) for lymphoma data

Other modifications of the volcano plot are available, e.g. the p-value based on a small sample test using local pooled errors (Jain et al. (2003). Multiple volcano plots were proposed, e.g. for interspecies comparisons and different organ tissues.

2. CHAPTER B

Introduction to Inference for Proportions

In clinical trails, frequently we need to compare some new drugs or treatments with the classical ones. Therefore we divided all the patients randomly to two groups, one group treated with new drugs, the other treated with old one or placebo. After a period of time, we want to see how many patients in each group recovered from the disease, and whether this new drugs or treatment are really more efficient compared with the classical ones, In these cases the goal of the user would be to find out whether the new developed treatment shows any (statistically) significant response at all. The natural way is to use the 2×2 contingency tables. Then the problems are inverted to the comparison of two proportions.

For instance, the data from table 13:

Table 13 Cross-Classification of Smoking By Lung Cancer (Doll and Hill, 1950)

Smoker	YES	NO	Total
Cases	688	21	709
Controls	650	59	709

the 2×2 contingency table above gives us the data from a survey, the table can be transformed to be the following table 14:

Table 14 Estimated Conditional Distributions

Smoker	YES	NO	Total
Cases	0.96	0.04	1.0
Controls	0.92	0.08	1.0

Then the two-sample test is to compare the “yes” proportions between two groups. The two 2×2contingency tables can be formalized to be:

Table 15 The Observed 2×2 contingency table, x.

Response	Success	Failure	Row_Total
Population1	X_{11}	X_{12}	N_1
Population2	X_{21}	X_{22}	N_2
Col_Total	M_1	M_2	N

The estimated conditional distribution table is given by:

Table 16 Estimated Conditional Distributions

Response	Success	Failure	Row_Total
Population1	π_{11}	π_{12}	1.0
Population2	π_{21}	π_{22}	1.0

Here $\pi_i = X_{i1}/n_1$, π_{i2} =1-π_{i1}. The **difference of proportions** of successes, π_{11}-π_{21}, is a basic comparison of proportions. The hypothesis can be formalized as:

H_0: $\pi_1 \le \pi_2$

H_A: $\pi_1 \ge \pi_2$

The difference of proportions falls between –1.0 and +1.0. But a value π_{11}-π_{21} of fixed size may have greater importance when both p_i are close to 0 or 1 than when they are not. For instance, the difference between 0.010 and 0.001 may be more noteworthy than the difference between 0.510 and 0.501, even though both are 0.009. In such cases, we need some other kind of statistics to show the difference.

An alternative is **Odds Ratio**, the odds is defined:

$$\Omega = \pi / 1-\pi$$

Ω is nonnegative, with Ω >1.0 when a success is more likely than a failure. The Odds Ratio is given by:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The odds ratio can equal any nonnegative number. The condition $\Omega_1 = \Omega_2$, θ=1, corresponds to independence of X and Y. When $1 < \theta < \infty$, subjects in row 1 are more likely to have a success than are subjects in row 2; that is $\pi_1 < \pi_2$. For the proportions just given, the odds ratios are $\theta_1 = \frac{0.010/(1-0.010)}{0.001/(1-0.001)} \approx 10$ and $\theta_2 = \frac{0.510/(1-0.510)}{0.501/(1-0.501)} \approx 1$. The odds of success in row 1 are 10 times the odds in row 2. This does not mean that the probability $\pi_1 = 10\pi_2$.

If we need more direct interpretation of our comparison, we need the **risk ratio** (or relative risk), which is defined below:

$$\rho = \frac{\pi_1}{\pi_2}$$

It can be any nonnegative number. A relative risk of 1.0 corresponds to independence. For the proportions just given, the relative risk are $0.010/\ 0.001 \approx 10.0$ and $0.410/0.401 \approx 1.02$. By this means we can see the difference of the two proportions in relative point of view.

There are some relationships between Odds Ratio and Relative Risk. From definition

odds ratio = relative risk$\left(\frac{1-\pi_2}{1-\pi_1}\right)$.

Their magnitudes are similar whenever the probability π_i of the outcome of interest is close to zero for both groups. In our example, the risk ratio and the odds ratio are almost the same with value 10 because both 0.010 and 0.001 are all close to zero. Because of this similarity, when each π_i is small, the odds ratio provides a rough estimate of the relative risk.

The sample relative risk is $r = \hat{\pi}_1/\hat{\pi}_2$. Like the odds ratio, it converges to normality faster on the log scale. The asymptotic standard error of log r is

$$\sigma(\log r) = \left(\frac{1-\pi_1}{\pi_1 n_1} + \frac{1-\pi_2}{\pi_2 n_2}\right)^{1/2}.$$

The Wald interval exponentiates endpoints of $\log r \pm z_{\alpha/2}\hat{\sigma}(\log r)$. It works well but can be somewhat conservative. There is an alternative method, score method (Koopman 1984,). The fact that the score intervals are computationally more complex than Wald intervals the principle behind them is simple. However, currently they are not available in standard software.

The following are three score test statistics for Binomial Ratio.

Suppose π_1 is the response rate of an experimental treatment and π_2 is the response rate of an active control treatment. Define the ratio of binomial proportions as (1.3)

$$\rho = \frac{\pi_1}{\pi_2}$$

In a non-inferiority clinical trial the objective is not to demonstrate that the experimental treatment is superior to the control but rather to demonstrate that the experimental treatment is not significantly inferior. Accordingly a non-inferiority margin, ñ0, is specified a priori and we test the null hypothesis of inferiority

$H_0 : \rho \geq \rho_0$ versus the one sided alternative hypothesis of non-inferiority $H_1 : \rho < \rho_0$

The test is carried out under the assumption that ñ is at its threshold null value $\rho = \rho_0$. Let $y \in \Omega$ denote any generic 2×2 table of the form of Table 1.3 that might be observed if we generated n1 independent Bernoulli trials each with probability π_1 and n2 independent Bernoulli trials each with probability π_2. The probability of observing any $y \in \Omega$. under H0 is

$f_{\pi_1,\rho_0}(y) = \binom{n_1}{X_{11}}\binom{n_2}{X_{21}}\pi_1^{X_{11}}(1-\pi_1)^{X_{12}}(\rho_0\pi_1)^{X_{21}}(1-\rho_0\pi_1)^{X_{22}}$ The test statistic (see Miettinen and Nurminen, 1985) is defined as

$$D(y) = \frac{\hat{\pi}_2 - \rho_0\hat{\pi}_1}{\sqrt{\frac{(\tilde{\pi}_2)(1-\tilde{\pi}_2)}{n_2} + \frac{\rho_0{}^2(\tilde{\pi}_1)(1-\tilde{\pi}_1)}{n_1}}}$$

where $\hat{\pi}_j = \frac{x_{j1}}{n_j}$ for j = 1, 2, and $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are the maximum likelihood estimates of π_1 and π_2, respectively, restricted under the null hypothesis to satisfy the requirement that $\tilde{\pi}_2/\tilde{\pi}_1 = \rho_0$. Miettinen and Nurminen (1985) have shown that one may obtain these restricted maximum likelihood estimates by solving a quadratic likelihood equation.

Thus

$\tilde{\pi}_1 = \frac{-B - \sqrt{B^2 - 4AC}}{2A}$ and $\tilde{\pi}_2 = \rho_0\tilde{\pi}_1$,

where $A = \rho_0 N$, $B = -(\rho_0 n_2 + X_{21} + n_1 + \rho_0 X_{11})$, $C = X_{11} + X_{21}$

Under H_0 this test statistic has mean 0 and variance 1.

For the construction of the unconditional exact confidence interval for the risk ratio, we have three choices of test statistics for test based interval estimation. Suppose we take n_1 independent Bernoulli samples from treatment 1 and n_2 independent Bernoulli samples from treatment 2. Let $y \in \Omega$ denote any generic 2 × 2 table that might be observed, and let x be the 2x2 table that was actually observed. Define

$$\hat{\pi}_j = \frac{x_{j1}}{n_j}$$

for j = 1, 2. The unstandardized test statistic

$D(y)=\frac{(X_{21}+0.5)(n_1+0.5)}{(y_{11}+0.5)(n_2+0.5)}$ is to compute an exact confidence interval for ρ. This statistic was proposed initially by Gart and Nam (1988). The 0.5 terms were necessary to ensure that the statistic and its reciprocal are defined whenever y11 or y12 are zero. We have observed that the unstandardized statistic is extremely conservative, leading to wider confidence intervals and larger p-values than could be obtained by other exact methods such as Agresti and Min (2001). Therefore we also use a test based exact confidence interval using the standardized statistic

$$D(y)=\frac{\hat{\pi}_2-\rho_0\hat{\pi}_1}{\sqrt{\frac{(\tilde{\pi}_2)(1-\tilde{\pi}_2)}{n_2}+\frac{(\tilde{\pi}_1)(1-\tilde{\pi}_1)}{n_1}}}$$

for j = 1, 2, and $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are the maximum likelihood estimates of π_1 and π_2 ,respectively, restricted under the null hypothesis to satisfy the requirement that $\tilde{\pi}_2/\tilde{\pi}_1=\rho_0$. The use of this test statistic has been proposed by Miettinen and Nurminen (1985) for asymptotic confidence intervals and by Chan and Zhang (1999) for exact confidence intervals. Confidence intervals derived by the above standardized statistic are shorter than corresponding intervals derived by the unstandardized statistic.

Miettinen and Nurminen confidence Interval

The test statistic is adopted and assumed to have a standard normal distribution. The asymptotic $100\times(1-\alpha)\%$ confidence interval $(\tilde{\rho}_*,\tilde{\rho}^*)$ is obtained by inverting the corresponding one-sided hypothesis tests. Thus $\tilde{\rho}_*$ satisfies the equality

$$1-\Phi\left\{\frac{(X_{21}/n_2)-\rho_*(X_{11}/n_1)}{\sqrt{\frac{(\tilde{\pi}_2)(1-\tilde{\pi}_2)}{n_2}+\frac{(\tilde{\pi}_1)(1-\tilde{\pi}_1)}{n_1}}}\right\}=\alpha/2$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are the maximum likelihood estimates of π_1 and π_2 ,respectively, under the restriction that $\tilde{\pi}_2/\tilde{\pi}_1=\rho_*$. Similarly ρ^* satisfies the equality

$$\Phi\left\{\frac{(X_{21}/n_2)-\rho^*(X_{11}/n_1)}{\sqrt{\frac{(\tilde{\pi}_2)(1-\tilde{\pi}_2)}{n_2}+\frac{(\tilde{\pi}_1)(1-\tilde{\pi}_1)}{n_1}}}\right\}=\alpha/2$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are the maximum likelihood estimates of π_1 and π_2 ,respectively, under the restriction that $\tilde{\pi}_2/\tilde{\pi}_1 = \rho^*$.

Katz, Baptista *et. al.* method

This method was first proposed by Katz, Baptista, Azen and Pike (1978), and subsequently modified by Gart and Nam (1988). It assumes that, in large samples, under the alternative hypothesis $\rho = \rho_0$ the statistic

$$\log D(y) = \log \frac{X_{12} + 0.5}{n_2 + 0.5} - \log \frac{X_{11} + 0.5}{n_1 + 0.5}$$

is approximately normal with mean $\log \rho_0$ and variance

$$\hat{s}^2 = \frac{1}{X_{21} + 0.5} - \frac{1}{n_2 + 0.5} + \frac{1}{X_{11} + 0.5} - \frac{1}{n_1 + 0.5}$$

therefore the asymptotic $100 \times (1-\alpha)\%$ confidence interval for $\log \rho$ is

$\log D(x) \pm z_{\alpha/2} \hat{s}$

An asymptotic two-sided p-value based on the above log statistic is

$p_2 = 2(1 - \Phi(|D(x)|))$

Koopman method

Koopman's (1984) method is based on inverting a chi-square test under the alternative hypothesis $\rho = \rho_0$. Under this hypothesis the test statistic

$$U_{\rho_0}(D(y)) = \frac{(X_{11} - n_1\hat{\pi}_1)^2}{n_1\hat{\pi}_1(1-\hat{\pi}_1)} + \frac{(X_{21} - n_2\hat{\pi}_2)^2}{n_2\hat{\pi}_2(1-\hat{\pi}_2)}$$

is distributed asymptotically as chi-square with 1 df. $\hat{\pi}_1$ and $\hat{\pi}_2$ are the maximum likelihood estimates of π_1 and π_2, under the restriction $\rho = \rho_0$. Koopman (1984) has provided the following expressions for

$$\hat{\pi}_1 = \frac{\rho_0(n_2 + X_{11}) + X_{21} + n_1 - [\{\rho_0(n_2 + X_{11}) + X_{21} + n_1\} - 4\rho_0 N(X_{11} + X_{21})]^{1/2}}{2N}$$

and $\hat{\pi}_2 = \hat{\pi}_1 \rho_0$. At the observed value, x, an approximate $100 \times (1-\alpha)\%$ two-sided confidence region for ρ is thus given by $\{\rho : U(D(x)) < \chi^2_{1,1-\alpha}\}$ where $\chi^2_{1,1-\alpha}$ is the

$1-\alpha$ fractile of the chi-square distribution with 1 *df*. One can establish that U_p is a convex function of ρ. Therefore the above confidence region is an interval of the form (ρ_*, ρ^*) where

$U_{\rho_*}(D(x)) = U_{\rho^*}(D(x)) = \chi^2_{1,1-\alpha}$

An asymptotic two-sided p-value based on the Koopman statistic is

$p_2 = \Pr(\chi_1^2 \geq U_\rho(x))$ evaluated at $\rho = 1$,where χ_1^2 is a random variable distributed as chi-square with 1 *df*.

Suppose X is from a binomial distribution bin(n, p). Our goal is to construct a (1-α)% confidence interval for the parameter p. The most widely used or known is based on an asymptotic normal approximation to the distribution of $\hat{\pi} = X/n$

Wald: $\hat{\pi} \pm z_{\alpha/2}\sqrt{\sigma^2(\hat{\pi}, n)},$

Where $z_{\alpha/2}$ is the 1-α/2 quantile of the standard normal distribution, and

$\sigma^2(\hat{\pi}, n) = \hat{\pi}(1-\hat{\pi})/n$

is the variance of $\hat{\pi}$. The above so-called standard interval is known to perform poorly. Wald-tests do not control α(type I error). I.E. Wald interval has a bad performance for some n's and p's. (e.g. Agresti and Caffo,2000).A much better alternative is to use the score interval: These references showed that a much better confidence interval for a single proportion is based on inverting the test with standard error evaluated at the null hypothesis, which is the score test approach. This confidence interval, due to Wilson (1927), is the set of p0 values for which $|\hat{\pi} - \pi_0|/\sqrt{\pi_0(1-\pi_0)/n} < z_{\alpha/2}$ which is

$$\hat{\pi}\left(\frac{n}{n+z^2_{\alpha/2}}\right) + \frac{1}{2}\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{n+z^2_{\alpha/2}}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z^2_{\alpha/2}}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right)\right]}$$

The mid point of this interval is

$$\hat{\pi}\left(\frac{n}{n+z^2_{\alpha/2}}\right) + \frac{1}{2}\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right)$$

it can be written as

$$\frac{\hat{\pi}n + \frac{z^2\alpha/2}{2}}{n + z^2\alpha/2}$$

recall $z^2\alpha/2 \approx 4$, $\hat{\pi}n = X$, i.e. number of success can be written as $\frac{X+2}{n+4}$.

Add-4 Method for the Difference of Two Proportions

Agresti and his Co-workers proposed a simple approach for constructing a confidence interval for a binomial proportion. They noticed that, and as a simplification proposed adding 4 pseudo-observations with one-half as successes and the other half as failures to obtain a modified estimator of $\pi, \tilde{\pi} = (X+2)/(n+4).$ Then their Adding-4 confidence interval is obtained by using $\tilde{p}$ in the Wald interval:

Adding-4:

$$\tilde{\pi} \pm z_{\alpha/2}\sqrt{\sigma^2(\tilde{\pi}, n+4)}$$

It performs surprisingly well. Under the same idea, the difference of two proportions is showed below: We observe two independent binomial variables:

$X_1 \sim bin(n_1, \pi_1)$ and $X_2 \sim bin(n_2, \pi_2)$

The goal is to construct a (1-α)% confidence interval for $\pi_2 - \pi_1$. The Wald interval is

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\sigma^2(\hat{\pi}_1, n_1) + \sigma^2(\hat{\pi}_2, n_2)},$$

where $\hat{\pi}_1 = X_1/n_1$ and $\hat{\pi}_2 = X_2/n_2$. Its performance is not satisfactory, as for one binomial proportion. The score interval can be extended, but it lacks a close form. Agresti and Caffo (2000) generalize the Adding-4 method as

Adding-4:

$$\tilde{\pi}_2 - \tilde{\pi}_1 \pm z_{\alpha/2}\sqrt{\sigma^2(\tilde{\pi}_1, n_1+2) + \sigma^2(\tilde{\pi}_2, n_2+2)},$$

Where $\tilde{\pi}_i = (X_i + 1)/(n_i + 2)$ for $i = 1,2$. The add-4 approach works quite good for the test of the difference of proportions.

Proposed Confidence Interval for the Ratio of Two Proportions

Let random variate (*r.v.*) $X \sim Bin(n,\pi)$. Observations from *n* Bernoulli trails x, LM estimate of success probability π is $\hat{\pi} = \frac{x}{n}$. From CLT, $\hat{\pi}$ is normally distributed with mean π and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Let π_1, π_2 denote the success probabilities of two samples, the **relative risk** is defined as $\rho = \pi_1/\pi_2$. **Sample relative risk** is $r = \hat{\pi}_1/\hat{\pi}_2$, log transform is $\log(r)$

Since $\hat{\pi}_i$ is asymptotically normally distributed, $\log(\hat{\pi}_i)$ is differentiable for $0 < \pi_i \leq \hat{1}$, then under mild condition using delta method, it is easy to show that $\log(\hat{\pi}_i)$ is distributed normally with expectation $\log(\pi)$, and standard error $\left(\frac{1-\hat{\pi}_i}{n_i\hat{\pi}_i}\right)^{1/2}$.

Consequently, $\log(r)$ is a linear combination of two normally distributed variables, and then it is again a normally distributed random variable with expectation $\log(\rho)$ and standard error $\sigma(\log(r)) = \left(\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}\right)^{1/2}$.

Add-4 method for the Ratio of two Proportions

We define new statistics (add 4 statistic) $\tilde{\pi}_1, \tilde{\pi}_2$ as $\tilde{\pi}_i = \frac{X_i + 2}{n_i + 4} \quad i = 1,2$

Because $\tilde{\pi}_i$ is only a linear transformation of X_i, which is asymptotically normally distributed random variable, $\tilde{\pi}_i$ is easily proven to be normally distributed with the following expectation and standard error.

Expectation of $\tilde{\pi}$ $E(\tilde{\pi}) = \frac{n\pi + 2}{n+4}$

Standard error of $\tilde{\pi}$ $\sigma(\tilde{\pi}) = \frac{\sqrt{n\pi(1-\pi)}}{n+4}$

bias of $\tilde{\pi}$ is $bias_\pi(\tilde{\pi}) = E_\pi(\tilde{\pi}) - \pi = \frac{2-4\pi}{n+4}$

Sample add-4 relative risk is $\tilde{r} = \tilde{\pi}_1/\tilde{\pi}_2$

Log-transformation of $\tilde{\pi}$ risk is under the mild condition, and then according to the delta method, the expectation and the standard error of $\tilde{\pi}$ can be derived as:

Expectation: $E(\log(\tilde{\pi})) = \log\left(\frac{n\pi+2}{n+4}\right) \xrightarrow{d} \log(\pi)$

Standard Error: $\sigma(\log(\tilde{\pi})) = \frac{\sqrt{n(1-\tilde{\pi})}}{(n+4)\sqrt{\tilde{\pi}}} \xrightarrow{d} \sqrt{\frac{1-\pi}{n\pi}}$

Log-transformation of the sample add-4 relative risk is a linear combination of two normally distributed *r.v.* , then the expectation and the standard error of $\log(\tilde{r})$ are:

Expectation: $E(\log(\tilde{r})) = \log\left(\frac{n_1\pi_1}{n_1+4} \times \frac{n_2+4}{n_2\pi_2}\right) \xrightarrow{d} \log\left(\frac{\pi_1}{\pi_2}\right)$

Standard Error: $\sigma(\log(\tilde{r})) = \sqrt{\frac{n_1(1-\tilde{\pi}_1)}{(n_1+4)^2\tilde{\pi}_1} + \frac{n_2(1-\tilde{\pi}_2)}{(n_2+4)^2\tilde{\pi}_2}} \xrightarrow{d} \sqrt{\frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}}$

Consequently, from the large sample inference theory, Wald confidence interval is constructed as:

$$\left[\log\left(\frac{\tilde{\pi}_1}{\tilde{\pi}_2}\right) - \sqrt{\frac{n_1(1-\tilde{\pi}_1)}{(n_1+4)^2\tilde{\pi}_1} + \frac{n_2(1-\tilde{\pi}_2)}{(n_2+4)^2\tilde{\pi}_2}}, \log\left(\frac{\tilde{\pi}_1}{\tilde{\pi}_2}\right) + \sqrt{\frac{n_1(1-\tilde{\pi}_1)}{(n_1+4)^2\tilde{\pi}_1} + \frac{n_2(1-\tilde{\pi}_2)}{(n_2+4)^2\tilde{\pi}_2}}\right]$$

Simulation Results and Discussion

Sample size 50, balanced case, success probability for sample 1: 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 1,00, Risk ratio: 1 (to check the control for alpha!) and success probability for sample 2: p_1/r . Significant level: 0.05. Coverage probability when nominal probability 0.05 is shown in Table 8:

Table 8: Coverage probability of p1 from 0.01 to 1.00 when true ratio = 1

p1	0.01	0.05	0.10	0.25	0.50	0.60	0.75	0.95	1.00
Cov. Prob.	1	0.9963	0.9837	0.9614	0.951	0.9428	0.9609	0.9941	1.00

When p_1 is zero, the coverage probability is 1.00.It can be shown in Figure 11:

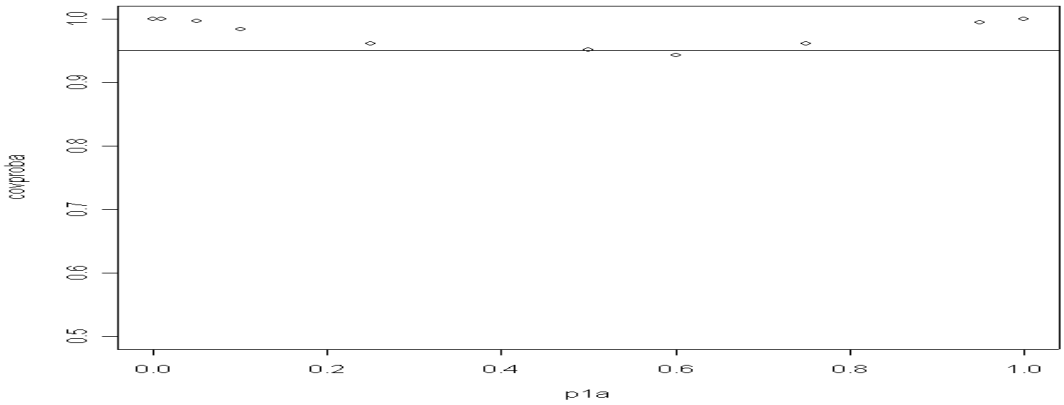

Figure 11 Coverage probability of add-4 confidence interval.

More closely to the range of p1 from 0.35 to 0.65, the results can be shown in Table 13:

Table 17 Coverage probability of p1 from 0.25 to 0.65 when true ratio = 1

p1	0.35	0.40	0.45	0.50	0.55	0.60	0.65
Cov. Prob.	0.9523	0.9511	0.9555	0.9542	0.9488	0.9534	0.9537

It can be shown in Figure 11:

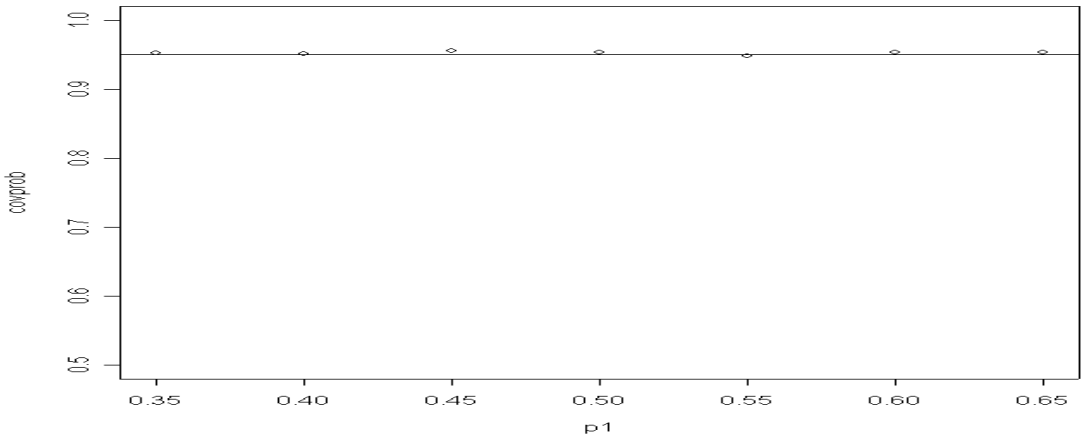

Figure 12 Coverage probability of add-4 confidence interval. p_1 from 0.35 to 0.65.

3. GENERAL DISCUSSION

When there is no assumption on the underlying distribution of the data, there is no uniformly most powerful test. As shown in the simulation study in Chapter A, in each specific condition, there is locally most powerful test. For instance, when the data is normally distributed with variance homogeneity, the sample *t* test is the most powerful test, whereas with variance heterogeneity, the Welch *t* test is the most powerful test. Under nonnormality, such as lognormal or certain Fleishman distribution, the Wilcoxon test is the most powerful test when there is only a location shift between the two groups. And in the mixture case, the Conover Salsburg Test is the most powerful test to detect the subset of response. There is no universe winner for all the conditions.

A simple test, which is appropriate for all the condition, seems impossible, but there are methods to combine different tests. Since the mixture distribution case is rare, in this thesis the test is proposed to solve the problem with a priori unknown unimodal distribution, which can be distributions with short tail, long tail, skew and heteroscedasticity. The Maximum test is proposed for this purpose.

But the simulation result shows that although the Maximum test begins with the more realistic assumption (only the unimodal distribution of data), it does not behave as efficient as the Welch *t* test when the underlying distribution of the data is nonnormal and variance heterogeneity. This also implies that when the data are skewed and the variances of two samples are not equal, the Welch *t* test is still a good candidate while the Welch *t* test is robust against nonnormality. Thus, in the application of Microarray data, the commonly used Welch *t* test is reasonable.

The new add-4 confidence interval for the ratio of two proportions is proved to maintain the nominal level of confidence (0.95) when the true ratio is 1. Especially when the first proportion is from 0.35 to 0.65, the actual level of coverage is very close to 0.95. Further study can be done when the true ratio is other numbers than 1 and sample size is smaller than 50.

REFERENCE:

"Student" (1908). *Biometrika*, 6, 1-25

Agresti A, (2001) Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures *AM STAT* 55 (3): 261-261

Agresti A, (2002) Categorical Data Analysis 2nd Edition , A JOHN WILEY & SONS, INC.,PUBLICATION

Agresti A, Min YY, (2001) On small-sample confidence intervals for parameters in discrete distributions, *Biometrics* 57 (3): 963-971 SEP

Agresti A., and Caffo, B. (2000), "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures," *The American Statistician*, 54, 280-288

Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, Hatfield GW: (2000) Global gene expression profiling in Escherichia coli K12. The effects of integration host factor. *J Biol Chem*, 275:29672-29684.

Armstrong,S.A., Staunton,J.E., Silverman,L.B., Pieters,R., den Boer,M.L., Minden,M.D., Sallan,S.E., Lander,E.S., Golub,T.R. and Korsmeyer,S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet*., 30, 41–47.

Baldi P, Long AD, (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509-519.

Balkin SD, Mallows CL, (2001) An adjusted, asymmetric two-sample t test *AM STAT* 55 (3): 203-206

Bartlett MS. (1953) Approximate confidence intervals, II. More than one unknown parameter. *Biometrika* 40:306–317.

Baumgartner W, Weiß P, (1998) Schindler H: A nonparametric test for the general two-sample problem. *Biometrics*; 54:1129-1135.

Bickel, P. J. (1982) On adaptive estimation. *Ann. Statist*. 10: 647–671.

Bioconductor [http://www.bioconductor.org]

Brownie C, Boos DD, Hughesoliver J., (1990) Modifying the t and ANOVA F Tests when Treatment is Expected to Increase Variability Relative to Controls *Biometrics* 46 (1): 259-266

Brunner, E. & Munzel, Ullrich, (2000) The Nonparametric Behrens-Fisher Problem: Asympotic Theory and a Small-Sample Approximation. *Biometrical Journal* 42: 17-25

Büning, H. & Trenkler, G. (1994) Nichtparametrische statistische Methoden, 2nd edn (Berlin: De Gruyter).

Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res*, 10:2022-2029.

Chan ISF, Zhang Z (1999). Test based exact confidence intervals for the difference of two binomial proportions. *Biometrics* ,55:1201-1209.

Conover, W.J. and Salsburg, D.S. (1988) Locally most powerful tests for detecting treatment effects when only a subset of patients can be expressed to “respond” to treatment. *Biometrics*, 44, 189-196

Cope, L.M., Irizarry, R.A., Jaffee, H.A., Wu, Z., and Speed, T.P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323-331.

Cramer, H. (1928) *Skand.Aktura*., 11, 141-180

Cyber T [http://www.igb.uci.edu/servers/cybert/]

Feuerstein, T.J., Rossner, R. and Schumacher, M. (1997) How to express an effect mean as percentage of a control mean? *J. Pharmacol. Toxicol. Methods*, **37**, 187-190.

Fieller, E. (1954) Some problems in interval estimation. *J. Royal Statist. Soc*. **B16**, 175-185.

Fleming, T. R. and Harrington, D. P. (1991). Counting Processes and Survival Analysis. Wiley,
New York.

Fleishman, A.I. (1978), A method for simulating non-normal distributions. *Psychometrika,* 43, 521-532

Freidlin, B. & Korn, E.L. (2002) A testing procedure for survival data with few responders, *Statistics in Medicine*, 21, pp. 65-78.

Garber,M.E., Troyanskaya,O.G., Schluens,K., Petersen,S., Thaesler,Z., Pacyna-Gengelbach,M., van de Rijn,M., Rosen,G.D., Perou,C.M., Whyte,R.I. et al. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, 98, 13784–13789.

Gart JJ. (1985) Approximate tests and interval estimation of the common relative risk in the combination of 2×2 tables. *Biometrika* 72(3):673–677.

Gart JJ, Nam J. (1988) Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics* 44:323–338.

Gastwirth, J.L., 1966. On robust procedures. *J. Amer. Statist. Assoc*. 61 (316), 929–948.

Gart JJ, Nam J, (1988) Approximate Interval Estimation Of The Ratio Of Binomial Parameters S - A Review And Corrections For Skewness *Biometrics* 44 (2):323-338

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular

classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.

Gruvberger,S., Ringner,M., Chen,Y., Panavally,S., Saal,L.H., Borg,A., Ferno,M., Peterson,C. and Meltzer,P.S. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*., 61, 5979–5984.

Hajek, J. and Sidak, Z. (1967). Theory of Rank Tests. New York: Academic Press.

Hajek, J., Sidak, Z. & Sen, P. K. (1999) Theory of Rank Tests (San Diego: Academic Press).

Hill, I.D., Hill, R., and Holder, R.L. (1976) Fitting Johnson curves by moments. *Appl. Statist*., 25, 180-189.

Hodges, J.L. and Lehmann, E.L. (1963) Estimates of location based on rank tests. *Ann. Math. Statist*. **34**, 598-611.

Hogg, R.V. and Craig, A.T. (1995) Introduction to Mathematical Statistics (Fifth Edition), Prentice Hall.

Hothorn, L.A. and Ma, D. 2004, VocanoCI: a modified volcano plot using confidence intervals for simple presentation of microarray data, Technical Report [http://www.bioinf.uni-hannover.de]

Hothorn, T. and Munzel, U. (2002): Non-parametric confidence interval for the ratio. Report University of Erlangen, Department Medical Statistics (*www.imbe.uni-erlangen.de*)

Hothorn,T. Lausen, B. 2002, On the exact distribution of maximally selected rank statistics. Computational Statistics & Data Analysis, 43, 121-137.

Huang,Y., Prasad,M., Lemon,W.J., Hampel,H., Wright,F.A., Kornacker,K., LiVolsi,V., Frankel,W., Kloos,R.T., Eng,C., Pellegata,N.S. and de la Chapelle,A. (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc. Natl Acad. Sci. USA*, 98, 15044–15049.

Huber, W., von Heydebreck, A., Stültmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18, S96-S104

Hwang, D., Schmitt, W. and Stephanopoulos, G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics,* 18, 1184-1193

Jain, N., Thatte1, J., Braciale, T., and Ley K. (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics,* **19**, 1945-1951.

Johnson, N. L. (1949). *Biometrika***,** 36, 149-176.

Katz D, Baptista J, Azen SP, et al. (1978) Obtaining Confidence-Intervals For Risk Ratio Incohort Studies *Biometrics* 34 (3): 469-474

Kepler,T., Crosby,L. and Morgan,K. (2002) Normalization and analysis of DNA microarray data by self consistency and local regression. *Genome Biol.*, *3*, research 0037.1–0037.12.

Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kolmogorov, A.N. (1973). *G. Inst. Ital. Attuari*., 4, 83-91 (in Italian)

Koopman PAR.(1984) Confidence intervals for the ratio of two binomial proportions. *Biometrics* 40:513–517.

Kutalik, Z. Inwald, J. Gordon, S.V., Hewinson, R.G., Butcher, P. Hinds, J., Cho, K.H. and Wolkenhauer, O. (2004) Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions in Mycobacterium bovis. Bioinformatics, 20, 357-363

Lachenbruch P.A. (2003) Proper metrics for clinical trials: transformations and other procedures to remove non-normality effects. Statistics in Medicine, 22,3823-3842

Lehmann, E.L. (1953) Ann. Math. Statist., 24, 23-43.

Lepage, Y.(1977): A class of nonparametric tests for location and scale parameters. *Communications in Statistics-Theory and Methods* 24,649-659.

Lonnstedt I, Speed T (2002) Replicated microarray data. *Statistica Sinica*, 12-31.

Ludbrook J. and Dudley, (1998) H. Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*.; 52(2):127-132.

Miettinen O, Nurminen M. (1985) Comparative analysis of two rates. *Statistics in Medicine* 4:213–226.

Mood, A.M. (1950). *Introduction to the Theory of Statistics.* McGraw-Hill, New York

Nam J. (1995) Confidence limits for the ratio of two binomial proportions based on likelihood scores: non-iterative method. *Biometrical Journal* 37:375– 379.

Neuhäuser, M., Liu, P.-Y. & Hothorn, L.A. (1998): Nonparametric tests for trend: Jonckheere's test, a modification and a maximum test. *Biometrical Journal* 40, 899-909.

Nthangeni M, Algina J, (2001) Type I error rate and power of some alternative methods to the independent samples t test *EDUC PSYCHOL MEAS* 61 (6): 937-957

Nurminen M, Miettinen, (1990) Confidence Intervals for the ratio of the parameters of 2 Independent Binomials. *Biometrics* 46 (1): 269-271

Pan W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546-554.

Pearson, E.E., Please, N.W., (1975) Relation between the shape of population distribution of four simple test statistics. *Biometrika*, 62, 223-241.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, 32(suppl.), 496–501.

R. Doll and A.B. Hill, (1950) British Mel. J., 30, 739-748.

R: [http://www.r-project.org]

Razzaghi M, (2002) The use of distribution mixtures for dose-response modeling in toxicological experiments, *Environmentrics* 13 (5-6): 657-667
Robert R. SOKAL, F. James ROHLF (1995):Biometry 3rd Edition W.H. Freeman and Company, U.S.A.

Ruymgaart, F. H., (1980) A unified approach to the asymptotic distribution theory of certain midrank statistics. In: Statistique non Parametrique Asymptotique, 1±18, J. P. Raoult (Ed.), Lecture Notes on Mathematics, No. 821, Springer, Berlin.

SAM: Significance Analysis of Microarrays [http://www.stat.stanford.edu/%7Etibs/SAM]

Sasabuchi, S. (1988) A multivariate one-sided test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor, *Memoirs of the Faculty of Science, Kyushu University,* **A42,** 9-19.

Schneider,R.,Tatlioglu,T.(1996) Molecular investigations on tetracycline and temperature sensitivity of cytoplasmic mal sterility in *Allium schoenoprasum* L. Beitr. *Züchtungsforschung* 2, 202-205.

Shipp,M.A., Ross,K.N., Tamayo,P., Weng,A.P., Kutok,J.L., Aguiar,R.C., Gaasenbeek,M., Angelo,M., Reich,M., Pinkus,G.S. et al. (2002)Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, 8, 68–74.

Shipp,M.A., Ross,K.N., Tamayo,P., Weng,A.P., Kutok,J.L., Aguiar,R.C., Gaasenbeek,M., Angelo,M., Reich,M., Pinkus,G.S. et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C.,Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P. et al. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203–209.

StatXact 4 User Manual, Cytel Software Corporation.

Tarone, R. E. (1981) . On the distribution of the maximum of the log-rank statistic and the modified Wilcoxon statistic. *Biometrics* 37: 79-85.

Torsten Hothorn, Berthold Lausen, (2002) On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43, 121-137

Troendle JF, Blair RC, Rumsey D, et al., (1997) Parametric and non-parametric tests for the overall comparison of several treatments to a control when treatment is expected to increase variability, *STAT MED* 16 (23): 2729-2739

Tusher VG, Tibshirani R, Chu G: (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*, 98,5116-5121.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

von Mises, N. (1931). Wahrscheinlichkeitsrechnung. Deuticke, Leipzig.

Welch, B.L. (1947). *Biometrika*, 34, 28-35

Wilcoxon, F. (1945) *Biometrics*, 1, 80-83

Wilson, E.B. (1927) Probable inference, the law of succession, and statistical inference. J. Amer. Statist. Assoc. 22:209-212.

Wolfinger R.D., Gibson G., Wolfinger E.D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput. Biol*., **8**, 625-637.

Yang,Y., Dudoit,S., Lin,D., Peng,V., Ngai,J. and Speed,T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., 30, e15.

Yeoh,E.J., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A. et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1, 133–143.

APPENDIX

R codes for the candidate tests which is not available in R or any R package.

R codes for the descriptive statistics of Microarray data:

```
############## A function to change all the zero values in matrix to NaN
####################################
############################## 29.01.2004 by Donghui Ma
#######################################################
zerotoNaN <- function(M){
        diM     <- dim(M)
        M1      <- matrix(as.numeric(M>0),diM[1],diM[2])
        M2      <- M2 <-( M + abs(M))/2
        Mf      <- M2/M1
        Mf
                                }
#############################################################################################
##################
Outlier.test        <- function(x,lower.b,upper.b)      {
        x           <- x[complete.cases(x)]
        lower.outlier       <- sum(as.numeric(x< lower.b))
        upper.outlier       <- sum(as.numeric(x> upper.b))
        outlier             <- c(lower.outlier,upper.outlier)
        outlier
                                                                        }

#############################################################################################
##################
################################### To calculate SKewness and Kurtosis
###################################
pianfengdu <- function(x) {
        n  <- length(x)
        mx <- mean(x)
        sx  <- sd(x)
        Skew <- (sum((x-mx)^3))/((n-1)*sx^3)
        Kurtosis <- (sum((x-mx)^4))/((n-1)*sx^4)
        SK <- c(Skew,Kurtosis)
        SK
                                    }
```

```
##########################################################################################
##################
pool            <- read.table("C:/Temp/singh.txt",header=T)                    # read the txt file to a data
frame           #
no              <- length(pool[1,])-1                                         # get the total number of the
observation #
pool.matrix <- as.matrix(pool[,2:no])                                         # read the data frame to a
matrix          #
pool.matrix     <- zerotoNaN(pool.matrix)                                             # get rid of negative
and give NaN            #
ng              <- length(pool.matrix[,1])                                    # get the number of genes
                #
g1.matrix       <- as.matrix(pool[,grep("N", names(pool))])       # read the frist group to a matrix
        #
n1              <- length(g1.matrix[1,])                                      # sample size 1
                        #
g2.matrix       <- as.matrix(pool[,grep("T", names(pool))])       # read the second group to a matrix #
n2              <- length(g2.matrix[1,])                                      # sample size 2
                        #
g1.matrix       <- zerotoNaN(g1.matrix)                                               # get rid of the
negative to NaN         #
g2.matrix       <- zerotoNaN(g2.matrix)                                               # get rid of the
negative to Nan         #
##########################################################################################
##################
###################################### define Information
Matrics##########################################
#       Mean            SE      CV      Skewness        Kurtosis Normality      Lognormality
lower.outlier upper.outlier #
GeneInfo.pg <- GeneInfo.rmoutlier.pg <- matrix(NA,no,9) # pooled two group after pre-testing
        #
GeneInfo.g1 <- GeneInfo.rmoutlier.g1 <- matrix(NA,ng,9) # for the first group
        #
GeneInfo.g2 <- GeneInfo.rmoutlier.g2 <- matrix(NA,ng,9) # for the second group
                #
################################## calculation the summary statistics
#######################################
# pool all the genes for each sample, to get the 0.25 and 0.975 quantile, they are lower.bound & upper.bound#
# for detecting outliers respectively
                                        #
```

```
pool.g1 <- pool.g2 <- c()
for(g in 1:ng)      {
        # for sample I #
        complete.g1         <- as.numeric(g1.matrix[g,])
        complete.g1         <- complete.g1[complete.cases(complete.g1)]
        nober.g1            <- length(complete.g1)
        # pool all the genes for sample I #
        if (nober.g1 > 2 ) {
        SD.com.g1           <- (complete.g1-mean(complete.g1,na.rm=T))/sd(complete.g1)
        pool.g1     <- c(pool.g1,SD.com.g1)
                                        } # end of pooling #
        # for sample II #
        complete.g2         <- as.numeric(g2.matrix[g,])
        complete.g2         <- complete.g2[complete.cases(complete.g2)]
        nober.g2            <- length(complete.g2)
        # pool all the genes for sample II #
        if (nober.g2 > 2)  {
        SD.com.g2           <- (complete.g2-mean(complete.g2,na.rm=T))/sd(complete.g2)
        pool.g2       <- c(pool.g2,SD.com.g2)
                                    } # end of pooling #
                            } # end of for g in 1:ng #
lower.b.g1      <- quantile(pool.g1,prob=0.025,na.rm=T)
upper.b.g1      <- quantile(pool.g1,prob=0.975,na.rm=T)
lower.b.g2      <- quantile(pool.g1,prob=0.025,na.rm=T)
upper.b.g2      <- quantile(pool.g1,prob=0.975,na.rm=T)
# calculation for the information matrices #
for(g in 1:ng)      {
        # for sample I#
        complete.g1         <- as.numeric(g1.matrix[g,])
        complete.g1         <- complete.g1[complete.cases(complete.g1)]
        nober.g1            <- length(complete.g1)
        if (nober.g1 >= 20)         {
        GeneInfo.g1[g,1] <- mean(complete.g1)
        GeneInfo.g1[g,2] <- sd(complete.g1)
        GeneInfo.g1[g,3] <- GeneInfo.g1[g,2]/GeneInfo.g1[g,1]
        complete.g1     <- (complete.g1-mean(complete.g1,na.rm=T))/sd(complete.g1)
        SkewKurtosis.g1  <- pianfengdu(complete.g1)
        GeneInfo.g1[g,4] <- SkewKurtosis.g1[1]
        GeneInfo.g1[g,5] <- SkewKurtosis.g1[2]
        GeneInfo.g1[g,6] <- as.numeric(shapiro.test(g1.matrix[g,])$p.value > 0.05)
```

```
GeneInfo.g1[g,7] <- as.numeric(shapiro.test(log(g1.matrix[g,]))$p.value > 0.05)
Outliers.g1          <- Outlier.test(complete.g1,lower.b.g1,upper.b.g1)
GeneInfo.g1[g,8] <- Outliers.g1[1]
GeneInfo.g1[g,9] <- Outliers.g1[2]
                        } # end for if #
# for sample II #
complete.g2          <- as.numeric(g2.matrix[g,])
complete.g2          <- complete.g2[complete.cases(complete.g2)]
nober.g2             <- length(complete.g2)
if (nober.g2 >= 20)      {
GeneInfo.g2[g,1] <- mean(complete.g2)
GeneInfo.g2[g,2] <- sd(complete.g2)
GeneInfo.g2[g,3] <- GeneInfo.g2[g,2]/GeneInfo.g2[g,1]
complete.g2   <- (complete.g2-mean(complete.g2,na.rm=T))/sd(complete.g2)
SkewKurtosis.g2  <- pianfengdu(complete.g2)
GeneInfo.g2[g,4] <- SkewKurtosis.g2[1]
GeneInfo.g2[g,5] <- SkewKurtosis.g2[2]
GeneInfo.g2[g,6] <- as.numeric(shapiro.test(g2.matrix[g,])$p.value > 0.05)
GeneInfo.g2[g,7] <- as.numeric(shapiro.test(log(g2.matrix[g,]))$p.value > 0.05)
Outliers.g2          <- Outlier.test(complete.g2,lower.b.g2,upper.b.g2)
GeneInfo.g2[g,8] <- Outliers.g2[1]
GeneInfo.g2[g,9] <- Outliers.g2[2]

                        } # end for if #
        } # end of for g #


# Graphical presentation #
# for the mean #
GI1.g1 <- hist(GeneInfo.g1[,1],freq=F,breaks=1000)
GI1.g2 <- hist(GeneInfo.g2[,1],freq=F,breaks=1000)
plot(GI4.g1,col="red",main="Histogram of Means of Sample I(red) and II(blue)",xlab="Genewise Mean of
Gene Expression Level")
lines(GI4.g2,col="blue")
# for the deviation #
GI2.g1 <- hist(GeneInfo.g1[,2],freq=F,breaks=1000,xlim=c(0,500))
GI2.g2 <- hist(GeneInfo.g2[,2],freq=F,breaks=1000,xlim=c(0,500))
plot(GI2.g1,col="red",main="Histogram of SDs of Sample I(red) and II(blue)",xlab="Genewise SD of Gene
Expression Level",,xlim=c(0,500))
lines(GI2.g2,col="blue",xlim=c(0,500))
```

```
# for the CV #
GI3.g1 <- hist(GeneInfo.g1[,3],freq=F,breaks=1000)
GI3.g2 <- hist(GeneInfo.g2[,3],freq=F,breaks=1000)
plot(GI3.g2,col="blue",main="Histogram of CVs of Sample I(red) and II(blue)",xlab="Genewise CV of Gene Expression Level",xlim=c(0,3))
lines(GI3.g1,col="red")
# for the Skewness #
GI4.g1 <- hist(GeneInfo.g1[,4],freq=F,breaks=1000)
GI4.g2 <- hist(GeneInfo.g2[,4],freq=F,breaks=1000)
plot(GI4.g1,col="red",main="Histogram of Skewness of Sample I(red) and II(blue)",xlab="Genewise Skewness of Gene Expression Level",sub="Green Vertical Line (x=0)is The Skewness for Normal Distribution")
lines(GI4.g2,col="blue")
abline(v=0,col="green")

# for the Kurtosis #
GI5.g1 <- hist(GeneInfo.g1[,5],freq=F,breaks=1000)
GI5.g2 <- hist(GeneInfo.g2[,5],freq=F,breaks=1000)
plot(GI5.g1,col="red",main="Histogram of Kurtosis of Sample I(red) and II(blue)",xlab="Genewise Kurtosis of Gene Expression Level",sub="Green Vertical Line(x=3) is The Kurtosis for Normal Distribution")
lines(GI5.g2,col="blue")
abline(v=3,col="green")

# Joint distribution of Skewness and Kurtosis#
GI45.g1 <- hist2d(GeneInfo.g1[,4],GeneInfo.g1[,5],na.rm=T,nbins=200,xlab="Skewness",ylab="Kurtosis")
GI45.g2 <- hist2d(GeneInfo.g2[,4],GeneInfo.g2[,5],na.rm=T,nbins=200,xlab="Skewness",ylab="Kurtosis")
persp(GI45.g1$x,GI45.g1$y,GI45.g1$counts,col="red",main="2D Histogram of Skewness and Kurtosis of Sample I(red) and II(blue)",xlab="Skewness",ylab="Kurtosis",zlab="Frequency",xlim=c(-1.4,7),ylim=c(0,45))
#points(GI45.g1$x,GI45.g1$y,GI45.g1$counts,col="blue")


par(new=TRUE)
persp(GI45.g2$x,GI45.g2$y,GI45.g2$counts,col="blue",main="2D Histogram of Skewness and Kurtosis of Sample I(red) and II(blue)",xlab="Skewness",ylab="Kurtosis",zlab="Frequency",xlim=c(-1.4,7),ylim=c(0,45))
# add a line for normal !!!!!!!!!!!!!!!!!!#

# Genewise comparision #
hist((GeneInfo.g2[,1]-GeneInfo.g1[,1]),freq=F,breaks=1000,col="grey",xlim=c(-300,300),main="Genewise differences in Mean, Sample II minus I",xlab="Genewise Differences of Gene Expression Level")
# Mean Ratio #
```

```
hist((GeneInfo.g2[,1]/GeneInfo.g1[,1]),freq=F,breaks=1000,col="grey",main="Genewise Ratios in Mean,
Sample II / I",xlab="Genewise Ratios of Gene Expression Level")

hist((GeneInfo.g2[,2]-GeneInfo.g1[,2]),freq=F,breaks=1000,col="grey",main="Genewise differences in SD,
Sample II minus I",xlab="Genewise Differences of SD of Gene Expression Level",xlim=c(-100,100))
# SD Ratios#
hist((GeneInfo.g2[,2]/GeneInfo.g1[,2]),freq=F,breaks=1000,col="grey",main="Genewise Ratios in SD, Sample
II minus I",xlab="Genewise Ratios of SD of Gene Expression Level",xlim=c(0,8))

hist((GeneInfo.g2[,3]-GeneInfo.g1[,3]),freq=F,breaks=1000,col="grey",main="Genewise differences in CV,
Sample II minus I",xlab="Genewise Differences of CV of Gene Expression Level")
hist((GeneInfo.g2[,4]-GeneInfo.g1[,4]),freq=F,breaks=1000,col="grey",main="Genewise differences in
Skewness, Sample II minus I",xlab="Genewise Differences of Skewness of Gene Expression Level")
hist((GeneInfo.g2[,5]-GeneInfo.g1[,5]),freq=F,breaks=1000,col="grey",main="Genewise differences in
Kurtosis, Sample II minus I",xlab="Genewise Differences of Kurtosis of Gene Expression Level")
GI45.g2m1 <- hist2d((GeneInfo.g2[,4]-GeneInfo.g1[,4]),(GeneInfo.g2[,5]-
GeneInfo.g1[,5]),na.rm=T,,xlab="Skewness",ylab="Kurtosis")
persp(GI45.g2m1$x,GI45.g2m1$y,GI45.g2m1$counts)
```

R codes for VolcanoCI Plot:

```
#####################The input file should be in such Format####################
# GeneID   Control1 ... Controln          Treatment1 ... Treatmentn
# GeneName1        12.3          23.3          89.0          90.0          #
# ...              ..            ..            ..            ..            #
# geneNameN        ..            ..            ..            ..            #
################################################################################
########################Read the Data file to R using the follwing code        #
pool        <- read.table("C:/Temp/shipp.txt",header=T)
no                  <- length(pool[1,])-1  # get the total number of the observation#
pool.matrix <- as.matrix(pool[,2:no])  # read the data frame to a matrix       #
ng                  <- length(pool.matrix[,1])      # get the number of genes     #
g1.matrix   <- as.matrix(pool[,grep("DLBC", names(pool))])      #the 1st group  #
g2.matrix   <- as.matrix(pool[,grep("FSCC", names(pool))])  #the 2nd group     #
n1                  <- length(g1.matrix[1,])        # sample size 1               #
n2                  <- length(g2.matrix[1,])        # sample size 2               #
# If data file contains negative obseravtions which should be casewise deleted  #
# run the following code to substitute the negative number with NA          #
pool.matrix[pool.matrix <= 0] = NA
g1.matrix[g1.matrix <= 0] = NA
g2.matrix[g2.matrix <= 0] = NA
```

```
################################################################################
################################R-codes for CIs ##################################
library(exactRankTests)
RM.CI <- function(x1,x2,alpha=0.05,min.n1n2 = 2)        {
        x1          <- x1[complete.cases(x1)]
        x2          <- x2[complete.cases(x2)]
        n1 <- length(x1)
        n2 <- length(x2)
        LB          <- NA
        EST         <- NA
        UB          <- NA

        if (n1 >=min.n1n2 & n2 >= min.n1n2)          {
                n <- n1 + n2
                x <- c(x1,x2)
                # a factor coding the two groups #
                cell <- factor(rep(c("Normal","Tumor"),c(n1,n2)))
                mydata <- data.frame(expression = x,cell,row.names=NULL)
                pos <- wilcox.exact(I(log(expression)) ~ cell, data = mydata, alternative =
"two.sided",conf.int=TRUE,conf.level=(1-alpha))
                LB <- exp(pos$conf.int)[1]
                UB <- exp(pos$conf.int)[2]
                EST<- exp(pos$estimate[[1]])
                                        } # end of if n1 n2 >=2  #
        CI <- c(LB,EST,UB)
        CI
                                                }
# two-sided Fieller CI #
fieller.CI <- function(x1,x2,alpha=0.05,min.n1n2=2)      {
        x1          <- x1[complete.cases(x1)]
        x2          <- x2[complete.cases(x2)]
        n1          <- length(x1)
        n2          <- length(x2)
        LB          <- UB <- est <- NA
        if (n1 >=min.n1n2  & n2 >= min.n1n2 )        {
                mx1         <- mean(x1)
                mx2         <- mean(x2)
                v           <- n1+n2-2
                sp          <- (var(x1)*(n1-1)+var(x2)*(n2-1))/v
                CT          <- qt(alpha/2,v)
```

```
                g           <- ((sp)*(CT^2))/(n1*(mx1^2))
                if (g < 1)  { # g < 1 just direct calculation #
                            est         <- mx2/mx1
                            LB          <- (est - sqrt(g*(est^2+(1-g)*(n1/n2))))/(1-g)
                            UB          <- (est + sqrt(g*(est^2+(1-g)*(n1/n2))))/(1-g)
                            }
                if (g >=1)  {
                            est.inv     <- mx1/mx2
                            est         <- 1/est.inv
                            g.inv       <- ((sp)*(CT^2))/(n2*(mx2^2))
                            LB          <- 1/((est.inv + sqrt(g.inv*(est.inv^2+(1-g.inv)*(n2/n1))))/(1-
g.inv))
                            test        <- (est.inv - sqrt(g.inv*(est.inv^2+(1-g.inv)*(n2/n1))))/(1-g.inv)
                            UB          <- ifelse(test>0,1/test,Inf)
                            }
                            } # end of if n1 & n2 #
        CI <- c(LB,est,UB)
        CI
                                                }


################################PLotting Part#######################################
par(mfrow=c(1,2)) ############# Plot Volcano and CI-FOLD together ###############

################################### Volcano Plot ###################################
############ only plot and give the coordinates for the plotting  ##############
Volcano.LR<- function(min.n1n2)             {
log2.gene.EST <- gene.EST <- gene.VTP <- rep(NA,ng)
for (i in 1:ng)         {
        x1          <- g1.matrix[i,]
        x2          <- g2.matrix[i,]
        x1          <- x1[complete.cases(x1)]
        x2          <- x2[complete.cases(x2)]
        n1 <- length(x1)
        n2 <- length(x2)
        if (n1>=min.n1n2 & n2 >=min.n1n2)           {
        gene.temp  <- t.test(x1,x2)
        mx1         <- gene.temp$estimate[[1]]
        mx2         <- gene.temp$estimate[[2]]
        if (mx1 !=0 & mx2 != 0) {log2.gene.EST[i]<- log2(mx2/mx1)}
```

```
            gene.VTP[i]             <- -log10(gene.temp$p.value)
                                                } # end of if n1 n2 #
                        }  # end of for ng #
x.min <- quantile(log2.gene.EST,probs=0.01,na.rm=T)
x.max <- quantile(log2.gene.EST,probs=0.99,na.rm=T)
x.lim <- max(abs(x.min),abs(x.max))



y.max <- quantile(gene.VTP, probs=0.99,na.rm=T)
y.lim <-abs(y.max)

# plotting part #
par(bg="white")
plot(log2.gene.EST,gene.VTP,col="blue",xlim=c(-x.lim,x.lim),ylim=c(0,y.lim),main="Volcano
Plot",xlab="log(Ratio)",ylab="log(p.value from t.test)")
abline(v=0,h=0) # for the x,y aixes #
text(-x.lim/2,y.lim-0.5,labels="Under-expressed genes")
text(x.lim/2,y.lim-0.5,labels="Over-expressed genes")
return(list(log2.gene.EST,gene.VTP))
                                                }# end of Vocalno.LR#
Vol.LR <- Volcano.LR(15)  # at least 15 observations in each group #
##############################################################################################

############### a function for CI-FOLD Plot,log2.gene.EST & gene.VTP ###############
CI.FOLD.plot<- function(method="nonpar", # nonpar= Ratio of two media; fieller = fieller's method
                                    alpha = 0.05,  # alpha for CI
                                    col="blue",
                                    pch = 16,
                                    xlab = "log2(Hodges-Lehmann Estimator)",
                                    ylab = "(-)log(Lower(Upper) Limit | Over(Under)-expressed genes)",
                                    main = "CI-FOLD PLOT ",
                                    aid.lines = TRUE, # abline(c(0,1)) and abline(c(0,-1))#
                                    fold.arc = 2,
                                    VIG.names = TRUE,global = TRUE)          {
gene.EST <- log2.gene.EST <- gene.VTP <- rep(NA,ng)
if (method == "nonpar")             { # for the nonpar CI method #
for (i in 1:ng)             {
            gene.temp  <- RM.CI(g2.matrix[i,],g1.matrix[i,],alpha=alpha,min.n1n2=15)
            gene.EST[i]             <- gene.temp[2]
   if (is.na(gene.EST[i])==FALSE)             {
```

```
            log2.gene.EST[i]<- log2(gene.EST[i])
            if (gene.EST[i] >=1)    {
                        gene.VTP[i] <- log2(gene.temp[1])  # over  expression then Lower bound#
                                    } # end of if gene.EST #
            if (gene.EST[i] < 1)    {
                        gene.VTP[i] <- -log2(gene.temp[3]) # under expression then upper bound#
                                    }
                                                }
                        }  # end of for ng #
                                    } # end of the nonpar CI method#
if (method == "fieller")            { # for the fieller CI method #
for (i in 1:ng)             { # only 2 samples all > 2 obs calculated Fieller CI #
            gene.temp  <- fieller.CI(g2.matrix[i,],g1.matrix[i,],min.n1n2=15)
            gene.EST[i]             <- gene.temp[2]
   if (is.na(gene.EST[i])==FALSE)           {
            log2.gene.EST[i]<- log2(gene.EST[i])
            if (gene.EST[i] >=1)    {
                        gene.VTP[i] <- log2(gene.temp[1])  # over  expression then Lower bound#
                                    } # end of if gene.EST #
            if (gene.EST[i] < 1)    {
                        gene.VTP[i] <- -log2(gene.temp[3]) # under expression then upper bound#
                                    }
                                                } # for if is.na #
                        } # end of for i #
                                    } # end of fieller CI method #

x.min <- quantile(log2.gene.EST,probs=0.01,na.rm=T)
x.max <- quantile(log2.gene.EST,probs=0.9999,na.rm=T)
x.lim <- max(abs(x.min),abs(x.max))
y.max <- quantile(gene.VTP, probs=0.9999,na.rm=T)
y.lim <-abs(y.max)
ss <- max(x.lim,y.lim)

# plotting part #
if (global==TRUE) { plot(log2.gene.EST,gene.VTP,col=col,main=main,xlab=xlab,ylab=ylab,xlim=c(-
ss,ss),ylim=c(-ss,ss)) }
else {plot(log2.gene.EST,gene.VTP,col=col,main=main,xlab=xlab,ylab=ylab,xlim=c(-ss,ss),ylim=c(0,ss))}
abline(v=0,h=0) # for the x,y aixes #
text(-ss/2,ss-0.5,labels="Under-expressed genes")
text(ss/2,ss-0.5,labels="Over-expressed genes")
```

```
if (aid.lines == TRUE) { abline(c(0,1),type="p",col=col) ; abline(c(0,-1),col=col,type="p") }
# for the fold.arc #
lfc <- log2(fold.arc)
abp <- lfc/sqrt(2)
dots.over<- seq(abp,lfc,by=0.001)
dots.under<- seq(-lfc,-abp,by=0.001)
vrtd.over <- sqrt(lfc^2-dots.over^2)
vrtd.under<- sqrt(lfc^2-dots.under^2)
lines(dots.over,vrtd.over,col=col)
lines(dots.under,vrtd.under,col=col)
if (VIG.names==TRUE)            {
        CanMarker.over      <- which(log2.gene.EST > 0 & gene.VTP > 0 &
log2.gene.EST^2+gene.VTP^2>lfc^2)
        CanMarker.under <- which(log2.gene.EST < 0 & gene.VTP > 0 &
log2.gene.EST^2+gene.VTP^2>lfc^2)
        text(log2.gene.EST[CanMarker.over],gene.VTP[CanMarker.over],labels=as.character(pool[CanMark
er.over,1]),col=rainbow(length(as.character(pool[CanMarker,1]))))
        text(log2.gene.EST[CanMarker.under],gene.VTP[CanMarker.under],labels=as.character(pool[CanM
arker.under,1]),col=rainbow(length(as.character(pool[CanMarker,1]))))
                                }
return(list(log2.gene.EST,gene.VTP))
                                        }# end of CIFOLD #
#CI.FOLD.plot(method="nonpar",VIG.names=T,global=T)



a<-CI.FOLD.plot(method="nonpar",
                                aid.lines = TRUE,
                                fold.arc = 4.5,
                                VIG.names = T,global = T)
```

R codes for the proposed Maximum Test:

```
MAX.test <- function(x,...) UseMethod("MAX.test")

MAX.test.default <-
function(x,y,alternative = c("two.sided","less","greater"),
          method = c("Statistic","p.value"),
          approx = c("normal","t"), # only useful when use method p.value #
          resampling = c("Bootstrap","Permutation"),
          resam.num = 400 ,
          conf.level = 0.95, ...)
{
alternative = match.arg(alternative)
method <- match.arg(method)
approx <- match.arg(approx)
resampling <- match.arg(resampling)
DNAME <- paste(deparse(substitute(x)), "and", deparse(substitute(y)))

if(!is.numeric(x)) stop("`x' must be numeric")
if(!is.numeric(y)) stop("`y' must be numeric")
x <- x[complete.cases(x)]
y <- y[complete.cases(y)]
m <- length(x)
if(m < 1)
 stop("not enough x observations")
n <- length(y)
if(n < 1)
 stop("not enough y observations")
N <- m + n
z <- c(x,y)

if (method == "Statistic")
{
 if (resampling == "Bootstrap")
 {
  MAX.obs <- MAX.MIN(x,y,alternative = alternative)$MaxStat[[1]]
  MAX.boot <- rep(NA,resam.num)
  for (i in 1:resam.num )
  {
   z.b <- sample(z,replace=T) # with replacement#
```

```
  x.b <- z.b[1:m]
  y.b <- z.b[(m+1):(m+n)]
  MAX.boot[i] <- MAX.MIN(x.b,y.b,alternative = alternative)$MaxStat[[1]]
 }
}# end of if resampling bootstrap #
if (resampling == "Permutation")
{
 MAX.obs <- MAX.MIN(x,y,alternative = alternative)$MaxStat[[1]]
 MAX.boot <- rep(NA,resam.num)
 for (i in 1:resam.num )
 {
  z.b <- sample(z) # without replacement #
  x.b <- z.b[1:m]
  y.b <- z.b[(m+1):(m+n)]
  MAX.boot[i] <- MAX.MIN(x.b,y.b,alternative = alternative)$MaxStat[[1]]
 }
}
PVAL.boot <- switch(alternative,
                    two.sided = ((sum(MAX.boot >= abs(MAX.obs))+sum(MAX.boot <= -
abs(MAX.obs)))+1)/(resam.num+1),
             less = (sum(MAX.boot <= MAX.obs)+1)/(resam.num+1),
                    greater =  (sum(MAX.boot >= MAX.obs)+1)/(resam.num+1))

RVAL <- list(DataName = DNAME,
             Method = method,
             MAX.observed = MAX.obs,
             p.value = PVAL.boot)
return(RVAL)
} # end of if method "statistic" #

if (method == "p.value")
{
MIN.obs <- MAX.MIN(x,y,alternative = alternative,approx = approx)$MinPval[[1]]
MIN.boot <- rep(NA,resam.num)
 if (resampling == "Bootstrap")
 {
  for (i in 1:resam.num )
   {
  z.b <- sample(z,replace=T)
  x.b <- z.b[1:m]
```

```
  y.b <- z.b[(m+1):(m+n)]
  MIN.boot[i] <- MAX.MIN(x.b,y.b,alternative = alternative,approx = approx)$MinPval[[1]]
  }
 }# end of if resampling Bootstrap #
 if (resampling == "Permutation")
 {
  for (i in 1:resam.num )
  {
  z.b <- sample(z)
  x.b <- z.b[1:m]
  y.b <- z.b[(m+1):(m+n)]
  MIN.boot[i] <- MAX.MIN(x.b,y.b,alternative = alternative,approx = approx)$MinPval[[1]]
  }
 }# end of if resampling Permutation #

PVAL.boot <- (sum(MIN.obs <= MIN.boot)+1)/(resam.num+1)
RVAL <- list(DataName = DNAME,
             Method = method,
             MIN.observed = MIN.obs,
             p.value = PVAL.boot)
return(RVAL)
} # end of if method "p.value"#

} # end of Max.test.default #

# Max of Statistics and Min of P value #

MAX.MIN <- function(x,y,alternative = c("two.sided","less","greater"),approx = c("normal","t"))
{
alternative <- match.arg(alternative)
approx <- match.arg(approx)
sp.g  <- GLR.test(x,y,alternative = alternative,type = "G")
sp.l  <- GLR.test(x,y,alternative = alternative,type = "L")
sp.h  <- GLR.test(x,y,alternative = alternative,type = "H")
sp.b  <- c(Brunner.test(x,y)$Statistic ,Brunner.test(x,y,approx = approx)$p.value)

Max.Stat <- max(sp.g[[1]],sp.l[[1]],sp.h[[1]], sp.b[[1]])
Min.Pval <- min(sp.g[[2]],sp.l[[2]],sp.h[[2]], sp.b[[2]])
RVAL <- list(MaxStat = Max.Stat,
             MinPval = Min.Pval)
```

```
return(RVAL)
}

#############################################################################
################## generalized Linear Rank Tests ######################
#########by Donghui Ma##### depends on package exactRankTests #########
#############################################################################
cscores.more <- function(y, type=c("Gastwirth", "LongTail",
                    "HoggFisherRandles"), int = FALSE,
                    maxs=length(y), ... ) {
type <- match.arg(type)
if (!(all.equal(floor(maxs),maxs))  || maxs < 1)
stop("maxs is not an positiv integer")
N <- length(y)
RET <- switch(type,"Gastwirth" = {
                        r <- rank(y)
                        r[r <= (N+1)/4] <- r[r <= (N+1)/4]-(N+1)/4
                        r[r > (N+1)/4 & r < 3*(N+1)/4] <- 0
                        r[r >= 3*(N+1)/4] <- r[r >= 3*(N+1)/4]-3*(N+1)/4
                        r},
                        "LongTail" = {
                        r <- rank(y)
                        r[r < floor(N/4)+1] <- -(floor(N/4)+1)
                        r[r >= floor(N/4)+1 & r <= floor(3*(N+1)/4)] <-
                r[r >= floor(N/4)+1 & r <= floor(3*(N+1)/4)]-(N+1)/2
                        r[r > floor(3*(N+1)/4)] <- (floor(N/4)+1)
                        r},
                        "HoggFisherRandles"={
                        r <- rank(y)
                        r[r <= (N+1)/2] <- r[r <= (N+1)/2]-(N+1)/2
                        r[r >  (N+1)/2] <- 0
                        r}
                ) # end of switch #
attr(RET, "scores") <- type
RET

                                                }# end of function cscores.more #
GLR.test <- function(x,...) UseMethod("GLR.test")

GLR.test.default <-
```

```
function(x,y,alternative = c("two.sided","less","greater"),
            type = c("Gastwirth", "LongTail", "HoggFisherRandles"),
            exact = NULL, conf.int = FALSE, conf.level = 0.95, ...)
{
   alternative <- match.arg(alternative)
   type <- match.arg(type)
   if(conf.int) {
      if(!((length(conf.level) == 1)
          && is.finite(conf.level)
          && (conf.level > 0)
          && (conf.level < 1)))
         stop("conf.level must be a single number between 0 and 1")
   } # end of if(conf.int)#
   DNAME <- paste(deparse(substitute(x)), "and", deparse(substitute(y)))
   if(!is.numeric(x)) stop("`x' must be numeric")
   if(!is.numeric(y)) stop("`y' must be numeric")

   x <- x[complete.cases(x)]
   y <- y[complete.cases(y)]
   m <- length(x)
   if(m < 1)
      stop("not enough x observations")
   n <- length(y)
   if(n < 1)
      stop("not enough y observations")
   N <- m + n
   r <- cscores.more(c(x, y), type = type)
   T <- sum(r[seq(along = x)])
   ET     <- (m/N)*sum(r)
   VT     <- ((m*n)/((N^2)*(N-1)))*(N*sum(r^2)-(sum(r)^2))
   STATISTIC <- (T-ET)/sqrt(VT)
   PVAL <- switch(alternative, two.sided = 2*(1-pnorm(abs(STATISTIC))),less = pnorm(STATISTIC),
            greater = 1-pnorm(STATISTIC)   )
   RVAL <- list(Statistic = STATISTIC,
                  p.value = PVAL)
   return(RVAL)

} # end of GLR.test.default #

##############################################################################
```

```
#################### Brunner Test Biom. J 2000 42, 17-25  ################
############################### by Donghui Ma ##############################
Brunner.test <- function(x,y,alternative = c("two.sided","less","greater"),
                    conf.level = 0.95,approx = c("normal","t"),...)   {
alternative <- match.arg(alternative)
approx <- match.arg(approx)
DNAME <- paste(deparse(substitute(x)), "and", deparse(substitute(y)))
if(!is.numeric(x)) stop("`x' must be numeric")
if(!is.numeric(y)) stop("`y' must be numeric")
x <- x[complete.cases(x)]
y <- y[complete.cases(y)]
m <- length(x)
if(m < 1)
 stop("not enough x observations")
n <- length(y)
if(n < 1)
 stop("not enough y observations")
N <- m + n
dim(x) <- m
dim(y) <- n
# normalized empirical distribution function of data#
F <- function(data,xx){(0.5*(sum((xx-data)>0)+sum((xx-data)>=0)))/length(data)}
# normalized combined empirical distribution function #
H <- function(xx) {length(x)*F(x,xx)/N + length(y)*F(y,xx)/N}
# Rank of X.ij among all N observations #
R.N <- function(xx) { N*H(xx)+0.5 }
# mean of ranks R.ij in the ith sample #
R.M.I <- function(data) {
 RMI <- sum(apply(data,1,R.N))/length(data)
 RMI }
# empirical variance of R.ij -R.ij(i) (within rank of X.ij among ni obs in ith sample)#
S2 <- function(data)          {
 # Rank of X.j among all n observations, within ith sample Rank #
 R.n <- function(xx) { length(data)*F(data,xx)+0.5}
 mn <- length(data)
 S22 <- (1/(mn-1))*sum((apply(data,1,R.N)-apply(data,1,R.n)-R.M.I(data)+(mn+1)/2)^2)
 S22                              } # end of S2#
Var.n <- function(data) {S2(data)/(N-length(data))^2}
Var.N <- N*(Var.n(x)/m+Var.n(y)/n)
# Calculation of Statistics for Brunner Test #
```

```
STATISTIC <- (R.M.I(y)-R.M.I(x))/sqrt(N*Var.N)
names(STATISTIC) <- "Brunner"
# p value calculation #
if (approx == "normal")
{
PVAL <- switch(alternative, two.sided = 2*(1-pnorm(abs(STATISTIC))),less = pnorm(STATISTIC),
            greater = 1-pnorm(STATISTIC)   )

RVAL <- list(Statistic = STATISTIC,
                p.value = PVAL,
                Null.Hypothesis = "relative effect = 1/2",
                Alternative.Hypothesis = paste("Relative effect ",
                  deparse(alternative), "1/2"))
return(RVAL)
}
if (approx == "t")
{
df <- t.test(x,y,alternative = alternative,conf.level = conf.level)$parameter
PVAL <- switch(alternative, two.sided = 2*(1-pt(abs(STATISTIC),df)),less = pt(STATISTIC,df),
            greater = 1-pt(STATISTIC,df)   )
RVAL <- list(Statistic = STATISTIC,
                p.value = PVAL,
                Null.Hypothesis = "relative effect = 1/2",
                Alternative.Hypothesis = paste("Relative effect ",
                  deparse(alternative), "1/2"))
return(RVAL)
}
                                                } # End of Brunner Test #
```

R codes for Mixture of Normals:

```
##############################################################################################
# function to generate random numbers under mixture of at most three distributions#
# By Donghui Ma                                                    #
# 21.08.03                                                  #
##############################################################################################
#all the arguments should be given, even the third component is empty, 0,0,0 given#

rmix <- function(n,m1,s1,p1,m2,s2,p2,m3,s3,p3) {

mv  <- c(m1,m2,m3)
mpv <- c(p1,p2,p3)
```

```
dn  <- c(1,2,3)
x   <- rep(NA,n)
rm  <- matrix(data=0,nrow=n,ncol=3)
P1  <- sample(dn,n,replace = TRUE, prob = mpv)
for (i in 1:n)  {
        rm[i,P1[i]] <- 1
        x[i] <-
rm[i,1]*rnorm(1,mean=0,sd=1)+rm[i,2]*rnorm(1,mean=5,sd=1)+rm[i,3]*rnorm(1,mean=10,sd=1)

                } # end of for loop #
x



                                                } # end of the function#
```

R codes for the Fleishman System

The test.txt file can be found in the disk attached. ma <- read.table("C:/Temp/test.txt",header=T)
ma.matrix <- as.matrix(ma[,2:5])

```
rFleishman <- function(n,m=0,std=1,SK.ABCD,...)    {
        x1 <- rnorm(n)
        x2 <- SK.ABCD[[1]]+x1*SK.ABCD[[2]]+(x1^2)*SK.ABCD[[3]]+(x1^3)*SK.ABCD[[4]]
        x  <- m + std*x2
        x
                                                }
```

R codes for Mood Median Test:

```
# Large Sample min(n1,n2) > 12 mood median test #
median.atest <- function(x1,x2)      {
        ZC <- qnorm(0.975, mean=0, sd=1, lower.tail = TRUE)
        sx1<- sort(x1)
        sx2<- sort(x2)
        n1 <- length(x1)
        n2 <- length(x2)
        N  <- n1 + n2
        s  <- ceiling(N/2)
        vs <- (n1/2)-0.5-ZC*sqrt((n1*n2)/(4*N))
        v3 <- round(vs)
        v4 <- round(n1-vs)+1
        L  <- sx2[s-n1+v3+1]- sx1[n1-v3]
        U  <- sx2[s-n1+v4]  - sx1[n1-v4+1]
        CI <- c(L,U)
        CI
                        } # end of the function #


# calculation of factorial #
jiecheng <- function(x)     {
        jc <- 1
        if (x == 0) jc <- 1
        else       {
                for (i in 1:x)       {
                        jc <- jc*i
                                }
                }
        jc
                        }
# calculation of combination#
zuhe <- function(n,k)       {
        zuhe <- jiecheng(n)/jiecheng(k)/jiecheng(n-k)
        zuhe
                        }

# exact mood median test#
median.etest <- function(x1,x2)      {
        p1  <- c()
```

```
p1[1] <- 0
p2  <- c()
p2[1] <- 0
i    <- 1
n1 <- length(x1)
n2 <- length(x2)
N  <- n1 + n2
s  <- ceiling(N/2)
v  <- max(n1-s,0):min(n1,N-s)
lv <- length(v)
j    <- 1
if (sum(p1)<=0.025){
        p1[i] <- (zuhe(n1,v[i])*zuhe(n2,N-s-v[i]))/zuhe(N,N-s)
        i <- i + 1
        p2[i] <- 0
                }
if (sum(p2)<=0.025){
        p2[j] <- (zuhe(n1,v[lv+1-j])*zuhe(n2,N-s-v[lv+1-j]))/zuhe(N,N-s)
        j <- j + 1
        p2[j] <- 0
                }
CI <- c(v[i-1],v[lv-j])
CI


                        }
```

ACKNOWLEDGMENT

First of all, I am greatly indebted to all my families for their continuous love and care. Without their support, it would have been impossible to finish the study successfully.

I am very grateful to my supervisor Professor Dr. Ludwig Hothorn for his guidance starting the inception of the research topic to the completion of this thesis.

I am grateful to Mr. Dilba, Mr. Chen and my wife Sun Jin, for the intensive discussion of mathematical statistics.

Many thanks to all staff members of the Bioinformatics Unit, with their kind help, the study and the life here become much easier.

I would like to thank the University of Hannover for the scholarship ward.