3-24-2015

Evaluation of Some Statistical Methods for the Identification of Differentially Expressed Genes

Andrew L. Haddon
Florida International University, ahadd003@fiu.edu

Recommended Citation
Haddon, Andrew L., "Evaluation of Some Statistical Methods for the Identification of Differentially Expressed Genes" (2015). *FIU Electronic Theses and Dissertations.* Paper 1913.
http://digitalcommons.fiu.edu/etd/1913

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

EVALUATION OF SOME STATISTICAL METHODS FOR THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Andrew Haddon

2015

To: Dean Michael R. Heithaus
College of Arts and Sciences

This thesis, written by Andrew Haddon, and entitled Evaluation of Some Statistical Methods for the Identification of Differentially Expressed Genes, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

Wensong Wu

BM Golam Kibria, Co-Major Professor

Florence George, Co-Major Professor

Date of Defense: March 24, 2015

The thesis of Andrew Haddon is approved.

Dean Michael R. Heithaus
College of Arts and Sciences

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2015

ACKNOWLEDGMENTS

I wish to thank all of the members of my committee for their support and guidance through every stage of this thesis project. Their direction and availability was always just a door-knock away. I would like to thank my major professor, Dr. Florence George for her continuous support as a professor and a mentor beginning from when I was an undergraduate in a completely different department. Her immense knowledge and advice has been an endless help throughout the time of research and writing of this thesis as well as the time as a student under her supervision. I would also like to thank my co-major professor, Dr. BM Golam Kibria for his encouragement, time, and insightful ideas. His instructive wisdom, guidance, and humor were a prime contribution to this thesis. I wish to give thanks to Dr. Wensong Wu as well for her constructive comments and teachings.

Lastly, my sincere thanks also go out to my parents, siblings, and Paula for their love, patients, and constant support throughout my studies.

ABSTRACT OF THE THESIS

EVALUATION OF SOME STATISTICAL METHODS FOR THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

by

Andrew Haddon

Florida International University, 2015

Miami, Florida

Professor Florence George, Co-Major Professor

Professor B.M. Golam Kibria, Co-Major Professor

Microarray platforms have been around for many years and while there is a rise of new technologies in laboratories, microarrays are still prevalent. When it comes to the analysis of microarray data to identify differentially expressed (DE) genes, many methods have been proposed and modified for improvement. However, the most popular methods such as Significance Analysis of Microarrays (SAM), samroc, fold change, and rank product are far from perfect. When it comes down to choosing which method is most powerful, it comes down to the characteristics of the sample and distribution of the gene expressions. The most practiced method is usually SAM or samroc but when the data tends to be skewed, the power of these methods decreases. With the concept that the median becomes a better measure of central tendency than the mean when the data is skewed, the tests statistics of the SAM and fold change methods are modified in this thesis. This study shows that the median modified fold change method improves the power for many cases when identifying DE genes if the data follows a lognormal distribution.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER I INTRODUCTION

Analysis of DNA microarrays has become a popular topic in the past years. Microarray technology has allowed researchers to observe thousands of gene expressions all at once. Gene expression in cells is of relevance because it allows a way to pinpoint disease markers that are related to medical treatments (Troyanskaya *et al*., 2002). A job that many researchers may want to perform would be to identify which genes in a cell are differentially expressed. For example, a researcher may need to conduct an experiment to discover differentially expressed genes between two experimental conditions. For explanation purposes this could be between healthy patients and patients who have a condition of interest such as cancer. Microarray analysis will allow the researcher to find which genes are expressed differently between these two groups of patients. The researchers will then be able to develop a treatment that targets these specific genes and create a more effective type of therapy. Further information on microarray technology can be found in Majtán *et al.* (2004).

Over the years many methods have been studied to perform the analysis of microarray data. These methods can be categorized into two types, parametric methods and nonparametric methods. Examples of parametric methods are the *t*-test, Bayes *t*-test (Baldi and Long, 2001), an analysis of variance approach, and the B-statistic method (Smyth, 2004). Nonparametric methods, on the other hand, have become very attractive in this field of research because of the previous costs of microarray experiments and the availability of replicated data has made it difficult to obtain large samples (Zhang, 2007). Nonparametric methods include Significance Analysis of Microarrays (SAM) proposed by Tusher *et al*. (2001), samroc, which uses a very similar test statistic to SAM's in addition to the use of a receiver operating characteristic (ROC) curve (Broberg, 2003), the mixture model method (MMM) (Pan, 2003), nonparametric empirical Bayes method (Efron *et al.,* 2001), and the Zhao-Pan method (Zhao and Pan, 2003).

A variety of comparisons between methods have been performed in the past to find which method is most reliable in discovering true differentially expressed genes. The main purpose in these comparisons is to find the method that correctly identifies the highest proportion of the true differentially expressed (DE) genes as DE while maintaining a small proportion of equivalently expressed (EE) genes being falsely identified as DE.

One of the most widely used methods for microarray analysis is the previously mentioned SAM (Zhang, 2007). However, SAM is not a completely robust method and some shortcomings arise. Many researchers have attempted to modify the method in order to make it more reliable. When the number of significant genes is fairly large in a data set, the estimated number of significant genes by SAM is affected and the test is less powerful. As a solution, Pan *et al*. (2003) suggested the use of MMM to estimate the distribution of the null and test statistic. The MMM allows for identifications of a rejection region for any type 1 error rate. In another attempt to fix this bias, Van de Wiel (2004) proposes a method using rank scores within SAM. Just by replacing the data with rank scores, the tendency of SAM to produce a biased estimate of DE genes is eliminated. The results are only valid though when the number of samples, *N,* is not "too small". On the basis of the test statistic used in SAM, Broberg (2003) created the samroc method. Broberg found that when the number of DE genes is large, then the samroc method is likely to work better than SAM. However, in most of the tests performed, the two methods worked just as well as each other when samroc did not outperform SAM.

Breitling *et al.* (2004) adopted another approach to identify differentially expressed genes called rank product in an attempt to exceed SAM. The results showed that, while being a simpler method than SAM, rank product outperformed SAM in identifying DE genes, even with very small data sets. It is also seen that the rank product method performed very similarly to fold change. Fold change (FC) is a popular method often used because of its simplicity and easy understanding (Tarca, 2008). There are some concerns with the fold change method that will be mentioned later in Chapter 2.

Comparisons across methods are interesting because each method usually results in outcomes without much agreement. In Jeffery *et al.* (2006) it is found that only 8 to 21% of the genes are commonly identified between the ten different methods being compared including SAM, samroc, fold change, and rank product. The study shows that many factors such as number of genes and number of samples influences which method will obtain the best result. It is concluded that rank product works well under settings with low number of samples and the ROC curve performed well under data sets with large sample sizes. The conclusion by Kim *et al.* (2006) is similar to that of Jeffery *et al.* (2006), noting that the sample size, distribution, and equal variance assumptions of each test greatly impact which test performs better. Our study shows that samroc performed best under the normal distribution and equal variance setting, as well as slightly exceeding SAM in both large and small sample cases. However, SAM outperformed samroc when the data follows a lognormal distribution.

Despite the advancement of next generation sequencing (NGS) as an alternative to microarrays, research in analysis of microarrays is still very relevant. Researchers in labs are more comfortable and confident with using microarrays as the technology has been around for a long time and it is less complicated than NGS (Baker, 2013). Figuring out the most efficient method to identify differentially expressed genes under particular data settings can help master the data analysis step in microarray research.

The focus of the present study is a comparison of the top performing and popular methods SAM, samroc, rank product, and fold change along with modified versions of the SAM method and the fold change rule. As it is evident in Kim *et al.* (2006) and Jeffery *et al.* (2006), sample size and distributional assumption of the data largely impacts the decision of which is the superior method to choose when identifying differentially expressed genes. The aim of this thesis was found after evaluating previous research and understanding the biggest drawbacks in this area. Several settings of normally distributed data, lognormal cases, and various sample sizes will be tested under each of the methods. For the first time, a modification that uses median in place of the mean in the test statistics of SAM and the fold change rule will be made in this thesis. The modifications follow from the concept that the median is a better measure of central tendency than the mean when describing skewed data. The expectation is that using the median will better represent the average gene expressions when the microarray data follows a skewed distribution. The modification to fold change will be shown to improve results in identifying differentially expressed genes under skewed data settings. A table of cutoff values for fold change and its modified version is also included in the present study.

This thesis is organized as follows. In Chapter 2, the statistical techniques are given. A simulation study under the different settings of distribution and sample size is performed on each of the methods in Chapter 3. Chapter 4 will include the application and analysis of the methods to the widely reviewed leukemia dataset from Golub *et al.* (1999). Finally, conclusions will be made along with a statement of some concerns and future possible research in Chapter 5.

CHAPTER II STATISTICAL METHODS

This section will consist of a review of several favored statistical methods for identifying differentially expressed genes in microarray datasets. The performance of the methods on data that follow a normal distribution and a lognormal distribution are of interest. Let the i^{th} gene expression level of the j^{th} sample under condition 1 be represented by X_{ij} and the i^{th} gene expression level of the k^{th} sample under condition 2 be represented by Y_{ik}, where j=1,…,J, k=1,…,K, which represents replicates under condition 1 and 2 respectively. The gene number is represented by i, where i=1,…,n. For this study n=5000 genes. The number of genes, n, was chosen to be 5000 based on the work of Schwender *et al*. (2003) and Zhang's (2007) research.

SAM

The test statistic in SAM is very similar to the test statistic from the simple t-test. The difference lies on the introduction of a small constant, s_0, in the denominator. The test statistic for SAM is as follows:

$$d(i) = \frac{\overline{X}_i - \overline{Y}_i}{s(i) + s_0}, \tag{2.1}$$

where X_i is the expression of the i^{th} gene under experimental condition 1 and Y_i is the expression of the i^{th} gene under experimental condition 2 (i =1,…,n). Further, $\overline{X}_i$and $\overline{Y}_i$ are the mean expression levels under conditions 1 and 2 respectively for gene i.

The "gene-specific scatter" or standard deviation $s(i)$ is defined:

$$s(i)=\sqrt{\frac{1/J+1/K}{J+K-2}\bullet\left\{\sum_{j=1}^{J}(X_{ij}-\overline{X}_i)^2+\sum_{k=1}^{K}(Y_{ik}-\overline{Y}_i)^2\right\}}, \tag{2.2}$$

where J is the number of replicates in experimental condition 1 and K is the number of replicates in experimental condition 2 (Zhang, 2007).

The constant, s_0, is added in order to correct the issue that the traditional *t*-test faces. The problem with the *t*-test occurs when genes have low expression levels and yield a small sample variance. The combination of those two factors lead to producing a large test statistic making it very likely that the gene will be identified as DE. The value of s_0 represents a percentile of the standard deviation values of all the genes. The method to compute this value can be found on Page 30 of the SAM user guide (Chu *et al.*, 2002).

In order to find which genes are DE, SAM calls an algorithm to create the null scores by pooling the data together across the two treatments per gene *B* times, where *B* is the total number of permutations. For each permutation, SAM finds the null statistic by using the same formula as the original test statistic, resulting in a total of *B* null statistics for each gene. The mean of the null statistic is then found for each gene and plotted against the ordered test statistic. The absolute differences between the two values are then found and compared against a cutoff value to determine whether or not there is a significant difference (Tusher *et al.,* 2001). The cutoff value can be obtained by following the method explained on Page 29 of the SAM user guide (Chu *et al.*, 2002).

Samroc

Broberg's (2003) approach to identifying lists of significant genes while minimizing the rate of false positives and false negatives consists of ranking genes in order of likelihood of being differentially expressed. The test statistic is similar to that of SAM, however the constant s_0, is chosen in a different manner (Kim *et al.* 2006). The test statistic looks like such:

$$d(i)_{samroc} = \frac{\overline{X}_i - \overline{Y}_i}{s(i) + s_0} \, . \tag{2.3}$$

Plotting the number of false negatives against the number of false positives as a proportion of the total number of genes for various cutoff values creates the ROC curve. This can be seen in Figure 1. By using every combination of s_0 and significance level α to obtain the false positive and false negative proportions, the final value of s_0 is chosen from the combination that produced the shortest distance, c, to the origin, where there would be no false negatives or false positives in Figure 1.

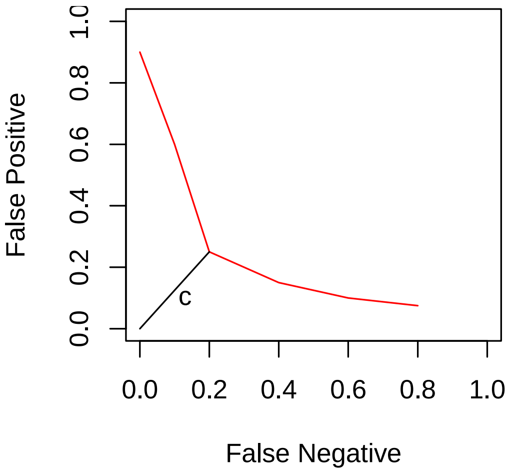


Figure 1. Example of an ROC curve. Graph obtained from Broberg (2003).

The choice of the particular combination allows the selection of s_0 to minimize (2.4) where the sum of FN and FP are the proportion of incorrectly identified genes (Broberg 2003).

With the data arranged with the rows representing each gene and the columns representing different samples, samroc uses repeated permutations of the columns in order to simulate the null distribution such as in SAM. The test statistic is calculated for each arrangement and compared to the original observed test statistic to find the p-value, the probability of obtaining a value as or more extreme (Broberg, 2003).

$$p_i = \frac{\#\left\{ d(j)^{*b} : \left| d(j)^{*b} \right| \geq \left| d(i) \right| \right\}}{B \bullet M}, \tag{2.4}$$

where *d(i)* is the observed test statistic for the i^{th} gene, *B* is the number of permutations, *M* is the number of genes, and $d(j)^{*b}$ is the value of the null statistic for the j^{th} gene and b^{th} permutation. Values of p_i that exceed the selected significance level, α, are considered differentially expressed.

Fold Change

According to McCarthy and Smyth (2009), the earliest publications in analyzing microarray data to identify differentially expressed genes used the fold change rule. The fold change rule is defined as follows (Kim *et al.,* 2006):

$$FC_i = \frac{\max(\overline{X}_i, \overline{Y}_i)}{\min(\overline{X}_i, \overline{Y}_i)}, \tag{2.5}$$

where $\overline{X}_i$ and $\overline{Y}_i$ are the mean expression levels under conditions 1 and 2 respectively for gene *i*. The typical accepted cutoff value for the fold change rule is $FC_i > 2$ (McCarthy and Smyth, 2009). McCarthy and Smyth also mention that a disadvantage of the fold change rule is that it does not take variability into consideration. Since it does not account for variability, it makes it difficult to make sense of a set cutoff value. The shortfalls of the fold change rule led to the development of more sophisticated tests such as SAM, however they also have their flaws and do not have the intuitive appeal which the fold change rule has (Breitling *et al.,* 2004).

Rank Product

The rank product method was created with overcoming the problems of fold change in mind, while being statistically rigorous and simple at the same time (Breitling *et al.,* 2004). After the rank product method gained popularity as a method to detect differentially expressed genes in microarray data, Koziol (2010) extended the process to a two sample setting. Koziol defines the test statistic as follows:

$$RP_i = \left(\prod_{j=1}^{j} R_{ij} \right)^{1/J} \div \left(\prod_{k=1}^{K} R_{ik} \right)^{1/K}, \tag{2.6}$$

where *J* is the number of replicates in experimental condition 1, *K* is the number of replicates in experimental condition 2, and the rank is taken among the expressions in a single sample, across the *n* genes, for each sample. R_{ij} represents those ranks assigned to the i^{th} gene under condition 1 and R_{ik} will be those ranks assigned to the i^{th} gene under condition 2. Further, the monotone log transformation is taken on the test statistic to obtain a better approximation of the null distribution and the resulting statistic is:

$$\log(RP_i) = (1/J)\sum_{j=1}^{J}\log(R_{ij}) - (1/K)\sum_{k=1}^{K}\log(R_{ik}), \tag{2.7}$$

According to Koziol (2010), “the exact distribution of *log(RP*$_i$*)* can be tedious” so a normal approximation of the distribution should be adequate, especially for large samples. If there is skewness in the data, then this approximation may not be adequate.

SAM Using Median

It has been shown that microarray data is consistent and well approximated by the lognormal distribution (Hoyle *et al.,* 2002). The lognormal distribution is known to be a skewed distribution and the best measure of central tendency for this type of distribution is the median (Hozo *et al.,* 2005, Manikandan, 2011). Behind this reasoning, the following modifications to improve the accuracy of SAM to correctly identify differentially expressed genes are proposed in this project.

The first modification will consist of a test statistic as such:

$$d_{M1}(i) = \frac{\tilde{X}_i - \tilde{Y}_i}{\tilde{s}(i) + s_0}, \tag{2.8}$$

Instead of using the average expression levels of the ith gene under condition 1 and 2, $\overline{X}_i$ and $\overline{Y}_i$, when calculating the test statistic, the median expression levels for the ith gene, $\tilde{X}_i$ and $\tilde{Y}_i$ under each condition is used.

$$\tilde{X}_i = median(X_{i1}, X_{i2}, ..., X_{iJ}) \tag{2.9}$$

$$\tilde{Y}_i = median(Y_{i1}, Y_{i2}, ..., Y_{iK}) \tag{2.10}$$

The same substitution of the median for the mean will be done when calculating the standard deviation.

$$\tilde{s}(i) = \sqrt{\frac{1/J + 1/K}{J + K - 2} \bullet \left\{ \sum_{j=1}^{J} (X_{ij} - \tilde{X}_i)^2 + \sum_{k=1}^{K} (Y_{ik} - \tilde{Y}_i)^2 \right\}}. \tag{2.11}$$

The second modified test statistic will be as follows:

$$d_{M2}(i) = \frac{\overline{X}_i - \overline{Y}_i}{\tilde{s}(i) + s_0}, \tag{2.12}$$

where the numerator stays as the difference in mean expression levels, however the denominator is using the modified standard deviation with the median (2.11).

Median Fold Change

With the prevailing use among biologists as seen in Sikora-Wohlfeld *et al.* (2013) because of its attractive nature and simplicity, the following modification is made to the fold change rule:

$$FC_{M\,i} = \frac{\max(\tilde{X}_i, \tilde{Y}_i)}{\min(\tilde{X}_i, \tilde{Y}_i)}. \tag{2.13}$$

The mean expression level, $\overline{X}_i$, under condition 1 and the mean expression level, $\overline{Y}_i$, under condition 2 from the fold change formula FC_i (2.5) is changed to the median expression level under each condition respectively for the i^{th} gene. Use the definitions (2.9) and (2.10) for $\tilde{X}_i$ and $\tilde{Y}_i$

The modification of replacing the mean with the median expression level of the i^{th} gene will better identify differentially expressed genes when the microarray data is following a skewed distribution such as lognormal as seen in the results in the following chapter.

The cutoff values for both, the original fold change rule in Chapter 2.3 and the modified fold change seen here, can also be found in the Appendix and Chapter 3.2 respectively. The cutoff values are selected in order to obtain a probability of rejecting the null hypothesis when it is true, type 1 error rate, of 0.05. For the purpose of this study, the type 1 error rate represents the probability of declaring a gene DE when it is truly not.

CHAPTER III SIMULATION STUDY

Since a theoretical comparison among the test statistics is not possible, a simulation study has been conducted to compare the performance of the test statistics in this chapter. In this section, the performance of SAM, samroc, fold change, rank product and the proposed modifications of SAM and fold change using median are compared by applying the methods to simulated gene expression data sets. The methods are compared under two cases: the data is simulated to follow a normal distribution and the data is simulated to follow a lognormal distribution. For both cases, simulations of several combinations of sample sizes have been done. For the lognormal case, the data was simulated to have three different levels of skewness, slight, moderate, and high.

Simulation Techniques

The simulation is performed by generating 5000 genes where 500 of them are knowingly differentially expressed. A matrix, *W*, is generated of size (5000 x (*J*+*K*)), *J* is the number of samples from condition 1 and *K* represents the number of samples from condition 2. As stated earlier, each data point in the matrix represents a gene expression, X_{ij} and Y_{ik}. The i^{th} gene expression level under condition 1 is represented by X_{ij} and the i^{th} gene expression level under condition 2 is represented by Y_{ik}. Matrix *W* will be designed as such:

$$W = \begin{bmatrix} X_{11} & \dots & X_{1,J-1} & X_{1J} & Y_{11} & & Y_{1,K-1} & Y_{1K} \\ X_{21} & \dots & X_{2,J-1} & X_{2J} & Y_{21} & & Y_{2,K-1} & Y_{2K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ X_{4999,1} & \dots & X_{4999,J-1} & X_{4999,J} & Y_{4999,1} & & Y_{4999,K-1} & Y_{4999,K} \\ X_{5000,1} & \dots & X_{5000,J-1} & X_{5000,J} & Y_{5000,1} & & Y_{5000,K-1} & Y_{5000,K} \end{bmatrix}. \tag{3.1}$$

The comparison between SAM, samroc, fold change, rank product, and the proposed modifications using median are performed under cases of randomly generated data from the normal distribution and lognormal distribution. For the cases under the normal distribution, the data follows the model:

$$X_{ij} = z_{ij} + \begin{cases} \delta_{ij} & if \quad 1 \le i \le 250 \\ \theta_{ij} & if \quad 251 \le i \le 500, \qquad for \quad j = 1,...,J \\ 0 & otherwise \end{cases} \tag{3.2}$$

$$Y_{ik} \sim N(0,1) \quad for \quad k = 1,...,K \tag{3.3}$$

where $z_{ij} \sim N(0,1)$, $\delta_{ij} \sim N(1.5,1)$, and $\theta_{ij} \sim N(-1.5,1)$.

For the cases under the lognormal distribution different levels of skewness are considered: slightly, moderately, and highly skewed. The levels of skewness will be implemented by setting $\sigma = 1,\ 1.2,\ 1.5$ respectively.

$$X_{ij} = \begin{cases} \eta_{ij} & if \quad 1 \le i \le 250 \\ \phi_{ij} & if \quad 251 \le i \le 500, \qquad for \quad j = 1,...,J \\ \varsigma_{ij} & otherwise \end{cases} \tag{3.4}$$

$$Y_{ik} \sim \ln N(0,1) \quad for \quad k = 1,...,K \tag{3.5}$$

where $\eta_{ij} \sim \ln N(1.5,\sigma)$, $\phi_{ij} \sim \ln N(-1.5,\sigma)$, and $\varsigma_{ij} \sim \ln N(0,1)$.

The choice of the sample sizes under condition 1 and 2, values of *J* and *K*, were chosen in order to cover a variety of situations that an experimenter may face when using real data and to be consistent with previous studies on microarray data. Sample sizes of (4,4) and

(10,26) were chosen as in Kim *et al.* (2006) and Zhang's (2007) study where the latter is also the sample size of the Leukemia data from Baldi and Long (2001). The sample size (8,8) was also chosen since it is of same size as the apolipoprotein AI (Apo AI) dataset from Callow *et al.* (2000) that has been analyzed in Chapter 4. For a thorough analysis covering more possibilities, sample sizes on a scale of 5 from 10 to 25 were also chosen for *J* and *K*. All of the sample sizes can be seen in Table 2. For the purpose of this study, the process of simulating a data set and running the methods under each setting was 500 times, while the previously mentioned studies of Zhang (2007) and Schwender *et al*. (2003) used 100 simulations for such comparisons.

Results and Discussion

An advantage of simulating microarray data is that the exact genes that are differentially expressed are known. After each method is performed on the simulated data sets, the total number of genes that were correctly identified as DE, true positives (TP), and the total number of genes that were incorrectly identified as DE, false positives (FP), were recorded. With the number of TP and FP known, then the type 1 error rate and the power were calculated to perform the comparison of methods. The null hypothesis for microarray analysis is that the i^{th} gene under condition 1 is the same as under condition 2 i.e., it is not DE, versus the alternative where the i^{th} gene under condition 1 is significantly different from the i^{th} gene under condition 2 i.e., the i^{th} gene is DE. The hypotheses are important to note in order to find the type 1 error rate, the probability of rejecting the null hypothesis given that it is in fact true, and the power, the probability of correctly rejecting a false null hypothesis. In terms of the microarray analysis done here the type 1 error rate reduces to the number of genes incorrectly identified as differentially expressed, FP, divided by the total number of equivalently expressed genes, 4500, and power reduces to the number of correctly identified differentially expressed genes, TP, divided by the total number of actual differentially expressed genes, 500.

$$\begin{aligned} P(type\ 1\ Error) &= \frac{P(reject\ null \cap null\ is\ true)}{P(null\ is\ true)} \\ &= \frac{FP/5000}{4500/5000} = \frac{FP}{4500} \end{aligned} \tag{3.6}$$

$$\begin{aligned} Power &= \frac{P(reject\ null \cap null\ is\ false)}{P(null\ is\ false)} \\ &= \frac{TP/5000}{500/5000} = \frac{TP}{500} \end{aligned} \tag{3.7}$$

To compare across each of the methods properly, it is important that the type 1 error rate is approximately 0.05 or less. A type 1 error rate of more than 0.05 would not be desirable to most investigators.

A smaller set of 20 simulations was done to begin to examine the power and type 1 error rate for the two modified methods of SAM and the modified fold change. The preliminary results were also used to gauge an estimate of the proper cutoff values for fold change in order to obtain the desired type 1 error rate. The preliminary results revealed that both modified SAM methods using median did not improve the current SAM method. Table 1 shows the power and estimated probability of type 1 error for small sample size (4,4) and large sample size (10,26) under a lognormal distribution. The original SAM method maintained a higher power than both modified versions over all the tested sample sizes. The results for only the simulations under the lognormal distribution are shown since the modification using median is intended to improve results when the microarray data is skewed. However, the difference in performance under data that is normally distributed is similar, with SAM outperforming the modified versions as expected.

Table 1. Power and P(type 1 error) for SAM and modified SAM. For simulated data under lognormal distribution and standard deviation of 1

Sample Size	(4,4)		(10,26)	
	Power	P(type 1)	Power	P(type 1)
SAM	0.0642	0.0020	0.5674	0.0174
SAM Modification 1	0.0438	0.0015	0.4606	0.0431
SAM Modification 2	0.0494	0.0018	0.5122	0.0139

After obtaining the preliminary results and obtaining a point of reference for proper cutoff values for the fold change methods, the full simulations, as previously explained, were performed without continuing further with the modified SAM methods. Shown in Figure 2, the simulations carried out with the data following a lognormal distribution displayed that the modified fold change with median, original fold change, and rank product performed better than the SAM and samroc methods across all sample sizes.

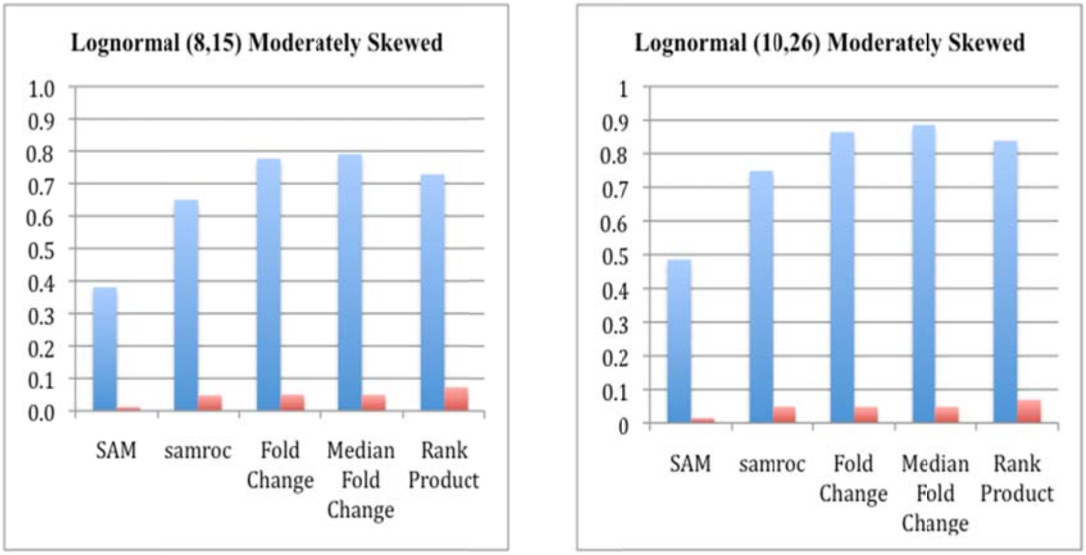


Figure 2. Power and P(type 1 error) under lognormal distribution. The graphs from left to right are from the simulations of sample size (8,15) and (10,26) with moderate skew. The blue columns correspond to the power and the red columns correspond to the P(type 1 error).

The simulations carried out under the lognormal distribution revealed settings where the SAM method turns out to be the weakest of the methods. SAM worked rather poorly for all sample size combinations where at least one of the conditions had sample size less than 15. For settings where both conditions had 15 or more samples, SAM worked decently with a power most of the time above 0.70 except for few situations where the data was moderately skewed and in all cases that were highly skewed. In highly skewed settings, SAM was rather poor. The samroc method followed similar trends as SAM, however, samroc was more robust in respect to sample size. The performance of samroc was much better than SAM under settings where both conditions had sample sizes of 10 or higher. The values of power and type 1 error rate for each setting under a lognormal distribution are given in Table 2. Levels of skewness are indicated by L=slightly skewed, M=moderately skewed, and H=highly skewed.

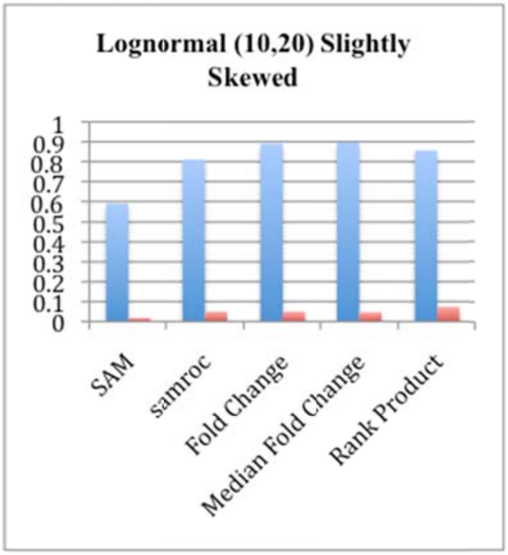
Table 2. Power and P(type 1 error) for simulations under lognormal distribution

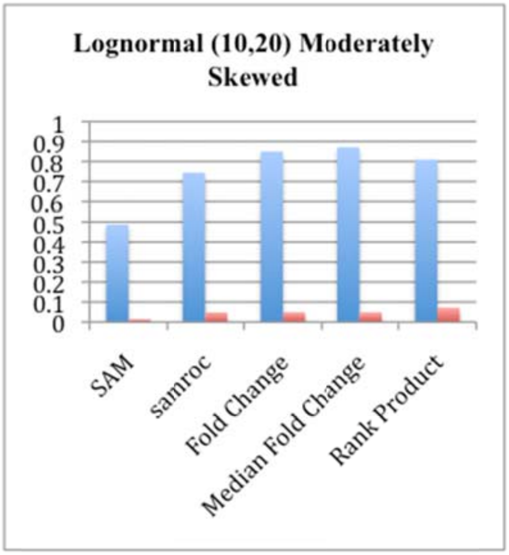
(*J*,*K*) and Skew	Power					P(type 1 error)				
	SAM	sam-roc	FC	Med. FC	Rank Prod.	SAM	sam-roc	FC	Med. FC	Rank Prod.
(4,4) L	0.0613	0.3986	0.4395	0.4445	0.4677	0.0022	0.0448	0.0427	0.0409	0.0809
M	0.0338	0.3790	0.4538	0.4742	0.4528	0.0014	0.0433	0.0440	0.0478	0.0791
H	0.018	0.3599	0.4638	0.4790	0.4318	0.0010	0.0413	0.0438	0.0474	0.0768
(8,8) L	0.3594	0.6436	0.7109	0.7235	0.6725	0.0116	0.0473	0.0472	0.0470	0.0807
M	0.2415	0.5840	0.6854	0.7072	0.6415	0.0080	0.0452	0.0481	0.0480	0.0790
H	0.1224	0.5245	0.6468	0.6818	0.5965	0.0041	0.0417	0.0476	0.0480	0.0772
(8,15) L	0.4194	0.7142	0.8144	0.8182	0.7677	0.0132	0.0492	0.0496	0.0492	0.0745
M	0.3806	0.6500	0.7767	0.7907	0.7280	0.0118	0.0477	0.0500	0.0494	0.0728
H	0.3229	0.5843	0.7154	0.7521	0.6649	0.0098	0.0444	0.0497	0.0490	0.0703
(8,20) L	0.4305	0.7243	0.8409	0.8579	0.8091	0.0134	0.0499	0.0472	0.0481	0.0716
M	0.4137	0.6596	0.8005	0.8225	0.7616	0.0125	0.0483	0.0475	0.0482	0.0691
H	0.3864	0.5935	0.7327	0.7820	0.6932	0.0114	0.0462	0.0473	0.0485	0.0667
(8,25) L	0.4369	0.7208	0.8638	0.8730	0.8383	0.0135	0.0498	0.0496	0.0489	0.0687
M	0.4341	0.6653	0.8248	0.8408	0.7874	0.0131	0.0489	0.0496	0.0490	0.0663
H	0.4236	0.5999	0.7543	0.7961	0.7116	0.0126	0.0471	0.0498	0.0491	0.0637

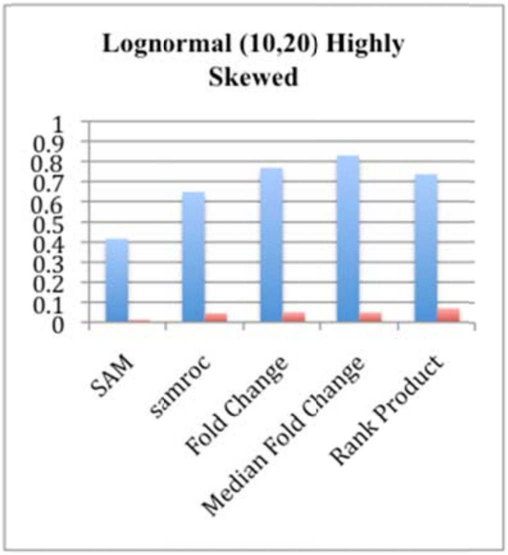
(12,8) L	0.5427	0.7418	0.7979	0.8108	0.7452	0.0174	0.0481	0.0490	0.0493	0.0813
M	0.3970	0.6644	0.7409	0.7845	0.7080	0.0126	0.0456	0.0410	0.0485	0.0795
H	0.2371	0.5836	0.6944	0.7531	0.6531	0.0074	0.0427	0.0475	0.0484	0.0774
(10,15) L	0.5984	0.7959	0.8588	0.8601	0.8132	0.0187	0.0488	0.0490	0.0480	0.0773
M	0.4731	0.7204	0.8154	0.8284	0.7703	0.0144	0.0473	0.0489	0.0469	0.0754
H	0.3608	0.6320	0.7427	0.7928	0.7028	0.0106	0.0443	0.0488	0.0470	0.0732
(10,20) L	0.5895	0.8107	0.8905	0.8973	0.8559	0.0181	0.0496	0.0499	0.0468	0.0744
M	0.4880	0.7440	0.8498	0.8713	0.8099	0.0146	0.0482	0.0497	0.0491	0.0722
H	0.4168	0.6474	0.7674	0.8290	0.7354	0.0121	0.0454	0.0498	0.0487	0.0699
(10,26) L	0.5567	0.8149	0.9068	0.9149	0.8860	0.0168	0.0499	0.0486	0.0486	0.0722
M	0.4855	0.7487	0.8648	0.8861	0.8378	0.0145	0.0491	0.0484	0.0485	0.0699
H	0.4476	0.6574	0.7806	0.8432	0.7560	0.0132	0.0470	0.0488	0.0489	0.0673
(15,15) L	0.8165	0.8931	0.9176	0.9106	0.8735	0.0261	0.0489	0.0485	0.0482	0.0821
M	0.6801	0.8087	0.8747	0.8889	0.8318	0.0210	0.0470	0.0485	0.0483	0.0805
H	0.4591	0.6829	0.7765	0.8544	0.7588	0.0136	0.0434	0.0483	0.0480	0.0789
(15,20) L	0.8581	0.9160	0.9461	0.9471	0.9169	0.0267	0.0485	0.0496	0.0496	0.0794
M	0.7353	0.8465	0.9063	0.9261	0.8758	0.0226	0.0482	0.0497	0.0496	0.0778
H	0.5245	0.7177	0.8052	0.8913	0.7973	0.0153	0.0449	0.0495	0.0496	0.0762
(15,25) L	0.8775	0.9242	0.9596	0.9585	0.9433	0.0272	0.0492	0.0481	0.0485	0.0777

M	0.7625	0.8622	0.9219	0.9390	0.9039	0.0230	0.0481	0.0482	0.0485	0.0755
H	0.5577	0.7374	0.8211	0.9036	0.8243	0.0162	0.0462	0.0485	0.0487	0.0738
(20,20) L	0.9223	0.9537	0.9692	0.9735	0.9423	0.0289	0.0485	0.0495	0.0473	0.0829
M	0.8192	0.8876	0.9327	0.9592	0.9073	0.0250	0.0475	0.0495	0.0473	0.0815
H	0.5906	0.7447	0.8247	0.9350	0.8312	0.0172	0.0445	0.0497	0.0471	0.0800
(20,25) L	0.9416	0.9634	0.9796	0.9827	0.9656	0.0295	0.0483	0.0500	0.0493	0.0812
M	0.8542	0.9098	0.9491	0.9715	0.9346	0.0260	0.0478	0.0499	0.0495	0.0797
H	0.6364	0.7702	0.8422	0.9488	0.8602	0.0184	0.0457	0.0497	0.0495	0.0780
(25,25) L	0.9657	0.9784	0.9882	0.9890	0.9770	0.0303	0.0486	0.0500	0.0488	0.0844
M	0.8903	0.9319	0.9624	0.9810	0.9507	0.0268	0.0476	0.0500	0.0488	0.0829
H	0.6616	0.7857	0.8550	0.9644	0.8841	0.0187	0.0453	0.0499	0.0487	0.0815

As Table 2 shows, the fold change method and the modified fold change method using median were consistently the top two methods across all sample sizes and all skewness settings for the lognormal data. The modified version of fold change with median worked better than the original fold change for all of the simulated sample sizes, obtaining higher levels of power while maintaining a type 1 error rate of 0.05 or smaller. It can also be seen in Table 2 that as the level of skewness rises, the modified version of the fold change method with median further improves over the original fold change. For each sample size simulated, as skewness increases, the difference in power between the original fold change and median fold change increases, with the latter having the higher power. This relationship is illustrated in Figure 3. The improvement in the fold change method was anticipated because the modified version replaced the mean with the median and for the lognormal data, which is a skewed distribution, the median is a more accurate measurement of the central tendency as Manikandan (2011) stated.







Sample Size (10,20)	Difference in Power (Med. FC – Orig. FC)
Slightly Skewed	.0068
Moderately Skewed	.0215
Highly Skewed	.0616

Figure 3. Power and P(type 1 error) under lognormal distribution for sample size (10,20) for different levels of skewness. The blue columns correspond to the power and the red columns correspond to the P(type 1 error**)**.

Even though the median fold change method constantly had the better power as the sample size increased, it is evident that when there are at least 15 samples of each condition and the skewness is not too heavy, all the methods work very similarly, producing about the same power and type 1 error rate. The similar performance between methods toward the higher number of sample sizes leaves the decision of which method to use for analysis of microarray data to the researcher depending on which assumptions best match the data and the method of choice. SAM, samroc, and fold change all have the assumption that the genes share equal variance while rank product assumption is more relaxed allowing the variance to be about equal (Kim *et al.,* 2006, Breitling *et al.,* 2004).

Using the median fold change method is appealing because of its ease and performance compared to the other methods when the data are assumed to follow a lognormal distribution however the choice to use it should be determined by what the researcher knows about the data. The cutoff values for this test were chosen in order to obtain a type 1 error rate of no more than 0.05 and are given in Table 3. The researcher will have to determine whether or not the cutoff values represent a meaningful difference in the genes that are being analyzed.

Table 3. Cutoff values for fold change and median fold change with at most 0.05 P(type 1 error) under lognormal distribution

Sample Size	Fold Change	Median Fold Change
(4,4)	5.00	4.75
(8,8)	3.23	3.18
(8,15)	2.81	2.79
(8,20)	2.72	2.65
(8,25)	2.62	2.58
(10,10)	2.88	2.93
(10,15)	2.65	2.65
(10,20)	2.52	2.49
(10,26)	2.46	2.42
(15,15)	2.42	2.44
(15,20)	2.29	2.28
(15,25)	2.22	2.22
(20,20)	2.16	2.14
(20,25)	2.08	2.06
(25,25)	2.00	2.00

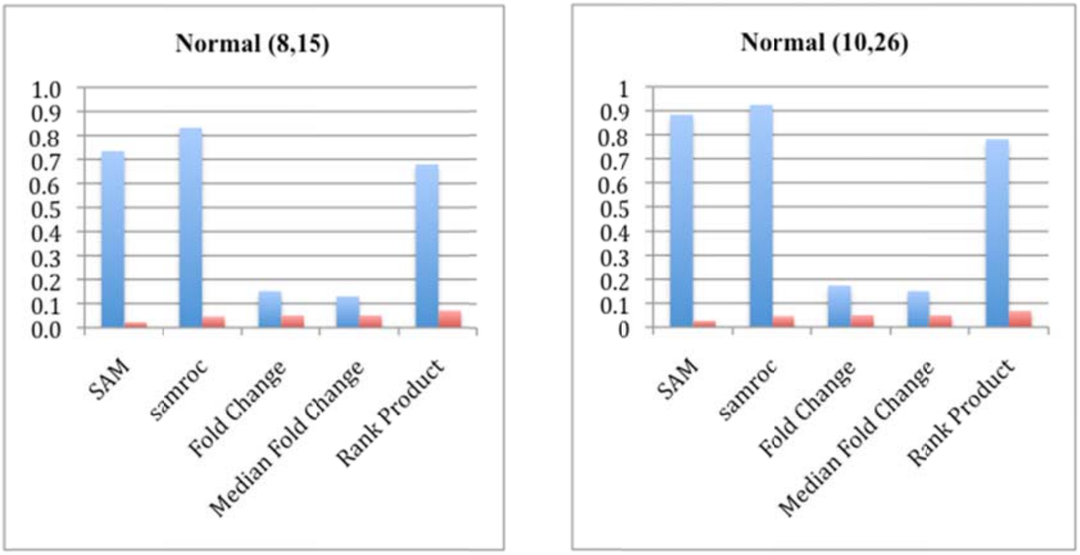


Figure 4. Power and P(type 1 error) under normal distribution. The graphs from left to right are from the simulations of sample size (8,15) and (10,26). The blue columns correspond to the power and the red columns correspond to the P(type 1 error).

The simulations carried out with the data following a normal distribution displayed that SAM, samroc, and rank product substantially performed better than fold change and the modified fold change with median across all sample sizes. Visuals of this outcome for two of the sample sizes simulated, (8,15) and (10,26), can be seen in Figure 4.

For the case of small and equal sample size (4,4) as seen in Figure 5, samroc and rank product perform similar to each other. The similar performance is understandable because the assumptions for the null distribution approximation for rank product that is being used are that the sample sizes are similar in size and that the distribution is close to normal.

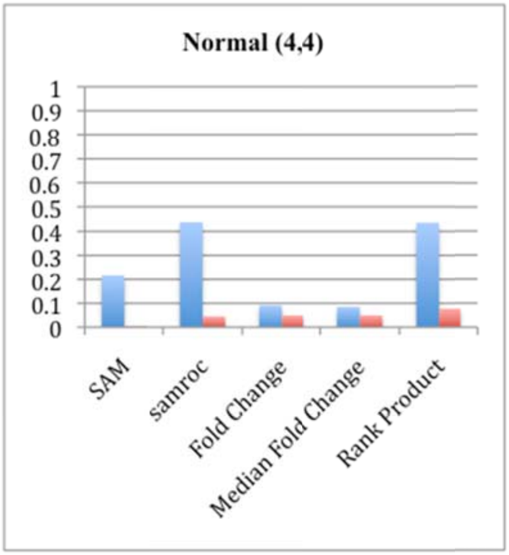


Figure 5. Power and P(type 1 error) under normal distribution for sample size (4,4). The blue columns correspond to the power and the red columns correspond to the P(type 1 error).

As the sample sizes increase when the data is following a normal distribution, samroc and SAM become increasingly better, increasing in power, with samroc being better than the latter. Once both sample sizes are at least 15, the two methods become almost the same making the choice between the two irrelevant. Over all the different sample sizes, from small to large, equal and unequal, when the variances are assumed to be the same and the distribution is following normal, samroc is the better performing method. The increasing similarity in samroc and SAM as sample size gets larger can be seen in Table 4. Table 4 shows the power and type 1 error rate for each method for each sample size used for data following a normal distribution.

Table 4. Power and P(type 1 error) for simulations under normal distribution

(J,K)	Power					P(type 1 error)				
	SAM	samroc	FC	Med. FC	Rank Prod.	SAM	samroc	FC	Med. FC	Rank Prod.
(4,4)	0.2168	0.4381	0.0910	0.0849	0.4360	0.0047	0.0452	0.0493	0.0495	0.0774
(8,8)	0.5297	0.7246	0.1210	0.1068	0.6084	0.0157	0.0452	0.0494	0.0497	0.0775
(8,15)	0.7341	0.8319	0.1512	0.1294	0.6788	0.0222	0.0461	0.0504	0.0499	0.0704
(8,20)	0.7930	0.8648	0.1609	0.1392	0.7101	0.0240	0.0466	0.0498	0.0497	0.0672
(8,25)	0.8262	0.8821	0.1640	0.1435	0.7317	0.0251	0.0467	0.0497	0.0494	0.0643
(10,10)	0.6702	0.8115	0.1329	0.1155	0.6683	0.0201	0.0454	0.0494	0.0494	0.0781
(10,15)	0.7996	0.8762	0.1540	0.1304	0.7220	0.0242	0.0460	0.0497	0.0495	0.0739
(10,20)	0.8538	0.9061	0.1662	0.1441	0.7558	0.0260	0.0463	0.0499	0.0499	0.0706
(10,26)	0.8835	0.9239	0.1739	0.1505	0.7802	0.0268	0.0464	0.0500	0.0496	0.0680
(15,15)	0.8862	0.9337	0.1566	0.1323	0.7802	0.0261	0.0452	0.0492	0.0494	0.0790
(15,20)	0.9307	0.9583	0.1704	0.1488	0.8214	0.0278	0.0455	0.0494	0.0499	0.0765

(15,25)	0.9507	0.9694	0.1837	0.1565	0.8479	0.0287	0.0460	0.0501	0.0500	0.0741
(20,20)	0.9625	0.9789	0.1732	0.1517	0.8558	0.0286	0.0456	0.0497	0.0498	0.0806
(20,25)	0.9772	0.9865	0.1868	0.1608	0.8844	0.0295	0.0460	0.0498	0.0497	0.0787
(25,25)	0.9885	0.9931	0.1890	0.1630	0.9055	0.0296	0.0462	0.0500	0.0499	0.0819

Sample size (10,26) replaced (10,25) since it is the sample size used in the real data set analysis and is close enough to the size of which it replaces. A table of the cutoff values used to obtain a type 1 error rate of at most 0.05 for the fold change method and its modification can be found in the Appendix.

CHAPTER IV APPLICATION

To illustrate the findings of this thesis, two real data sets, the leukemia data set from Golub *et al.* (1999) and the Apo AI data from Callow *et al.* (2000) are analyzed in this chapter.

Leukemia Data

The leukemia data consists of 7129 genes and a total of 38 samples, 11 of the samples are from acute myeloid leukemia (AML) patients and 27 are from acute lymphoblastic leukemia (ALL) patients. For this study, two cases are analyzed for all methods, randomly selecting 10 AML samples and 26 ALL samples as in Kim *et al.* (2006) and randomly selecting 4 samples from AML and ALL as in Broberg (2003). As seen in Chapter 3 when the data set contained a large number of samples under both conditions the choice of method was not so vital. A larger difference in power of the methods was expressed when the data had fewer sample sizes.

To preprocess the data, as in Kim *et al.* (2006), the median was subtracted from each gene expression and then divided by the interquartile range (IQR) per each sample.

$$preprocessed\ X_{ij} = \frac{X_{ij} - median(X_j)}{IQR_j}, \tag{4.1}$$

where j=1,…,J and IQR=upper quartile-lower quartile. The same formula (4.1) is used for k=1,…,K as well to find X_{ik}.

To compare each of the five methods, the same 50 reference genes that were deemed significant in Broberg (2003) were used. According to Broberg (2003), biological evidence and a statistical analysis on the full data set led to the belief that these 50 reference genes are differentially expressed under the comparison of AML and ALL patients. All the genes were ranked by the absolute value of their test statistic and the largest was given rank=1, second largest was given rank=2, and so on. The average ranks of the 50 reference genes were used to compare across the methods and are given in Table 5 for both sample size settings.

Table 5. Average ranks of the reference genes in the leukemia dataset

(J,K)	SAM	samroc	Fold Change	Median Fold Change	Rank Product
(4,4)	688.06	614.98	1573.42	1552.54	2926.36
(10,26)	1056.3	142.22	1323.9	1328.52	2002.28

The lowest average rankings for both sample size settings are given by the samroc and SAM methods as shown in Table 5. The samroc method worked extremely well compared to the other methods for the large sample setting. The outcome could be explained by the results of an F-test to check equal variance, where only 23.7% of the genes were found to satisfy this assumption as found by Kim *et al.* (2006). The violation of the equal variance assumption can affect the results of the fold change methods as well as SAM and samroc since each assumes equal variance. A Shapiro test for normality of the genes was also performed and found that 58.76% of the genes satisfied the normality assumption. When considering multiple test error, there may be more genes that were rejected as normal and as seen in Chapter 3, the proposed median fold change method does not perform as well when the data is normal. Contrary to what was expected, the difference in the performance of SAM and samroc was much larger for the large sample size setting than under the small sample size setting. The larger margin of performance when the sample size increased could be explained by the random sampling when choosing the samples for the (10,26) setting. The simulations from Chapter 3 showed that as the sample sizes grew, the methods performed more alike. Fold change, median fold change, and SAM however did not show to have such a large difference in performance under the large sample size setting. The modified median fold change method produced a smaller average rank than the original fold change method under the small sample setting. This shows an improvement with the modification made in this project. Further real data sets should be tested in order to check this finding but these results are very promising.

Apo AI Data

The Apo AI dataset consists of 5548 genes and 16 samples. Out of the 16 samples, 8 were from control mice and the other 8 samples were from mice with the Apo AI gene knocked out. The 8 mice that had the Apo AI gene knocked out will have a very low high-density lipoprotein cholesterol level and the delivery of the cholesterol to the liver will be affected (Callow *et al.,* 2000). The data were preprocessed in the similar way as the leukemia data (4.1), as was done by Kim *et al.* 2006. The difficulty when attempting to analyze this dataset is that there has not been reference genes adopted as biologically significant from previous studies as there was with the leukemia data.

To compare each of the methods, it was intended to select our own reference genes by finding the top 5% significant genes identified by each method and then finally selecting the common significant genes between the five methods. The same strategy was done by Kim *et al.* (2006) to select reference genes. The average ranks of the reference genes should have been taken and compared as was done with the leukemia data however, no common genes were found significant between all five methods. The idea expressed in Jeffery *et al.* (2006) that only a very low percentage of genes will be found significant between multiple methods is supported by these results. Table 6 shows the number of genes that were commonly found between each pair of the five methods.

Table 6. Number of common identified significant genes in the Apo AI dataset.

Methods	SAM				
samroc	42	samroc			
Fold Change	0	25	Fold Change		
Median Fold Change	0	35	39	Median Fold Change	
Rank Product	33	182	1	3	Rank Product

In addition to Table 6, there are only a few three-way combinations of the methods that share common identified differentially expressed genes. Together, SAM, samroc, and rank product found 33 common significant genes, samroc, fold change and median fold change found 13 in common, and samroc, median fold change, and rank product found only 1 gene in common significant. The conflicting result between methods is one of the drawbacks of microarray analysis. There is a large inconsistency between the different methods to identify which genes are identified as significantly different between two groups.

A Shapiro test was performed to test the normality assumption on the Apo AI data set and found that 4450 of the 5548 genes, 80.21% are normally distributed. In reference to the simulations performed in Chapter 3, samroc performed the best when the data followed normal, so the number of significant genes in common between the fold change methods and samroc can be compared. Even though the modification to the fold change method was intended to improve the identification of significant genes when the microarray data followed a skewed distribution, Table 6 shows that median fold change has 10 more significant genes in common with samroc than the original fold change method does with samroc. These results show that the proposed modified version of fold change with the median can be an improvement over the original in cases when the data may be approximately normal. However, since there were no reference genes truly known to significantly have a biological difference between the knockout Apo AI mice and the control mice, from this dataset analysis, it is best to note the challenge of identifying truly differentially expressed genes when analyzing real data sets. As we can see the choice of the method for analysis can make a large difference in which genes are called differentially expressed.

CHAPTER V CONCLUSION

A comparison of the performance of popular testing procedures for identifying differentially expressed genes from microarray data such as SAM, samroc, fold change and rank product was conducted. On the basis of the assumption that microarray data are related to the lognormal distribution from Hoyle *et al.* (2002) and the familiar idea that the median is a better measurement of central tendency than the mean when describing skewed data as expressed in Manikandan (2011), modifications were attempted on two methods. The test statistics of SAM and fold change were modified, replacing the mean gene expression values with the median.

The six procedures were applied to simulated datasets under various settings of sample sizes to represent real situations when dealing with microarray data and different levels of skewness. The test statistic modifications to SAM with median did not result in a higher power than the original SAM in the analysis of lognormal data early in this research so the testing was continued without the two SAM modifications. The lack of improvement could have been because of the choice of the constant s_0. The value s_0 may need to be adjusted in order to correctly minimize the coefficient of variation of the test statistic.

In the analysis of the simulated lognormal distribution, different levels of skewness were considered. Fold change and the modified median fold change were consistently the top performing methods for all levels of skewness of the lognormal data. For small sample sizes the results had shown that the popular SAM method performed very poorly while the proposed median fold change method out performed all other methods throughout the tested sample sizes and levels of skewness. The SAM method was the worst performing method for the simulated lognormal data with its power reducing as the data became more and more skewed, on the contrary, the modified fold change method improved its performance as skewness grew.

In the analysis of the simulated normally distributed data, samroc was the most powerful method consistently across all sample sizes. The difference in performance between samroc and SAM however was not much different after both sample sizes were 15 or larger. The fold change method and its modified version with median performed rather poorly in the analysis of the simulated normal data.

An analysis on a real microarray datasets was also performed to evaluate how the methods and the proposed modification would perform in a real situation. The leukemia dataset from Golub *et al.* (1999) was analyzed with the five remaining methods, SAM, samroc, rank product, fold change, and the median fold change method. The samroc method performed the best for the large and small sample setting when evaluating the average ranks of the 50 reference genes that were declared as biologically significant in Broberg (2003). Though the median fold change method did not perform the best, it did outperform the original fold change method under the (4,4) sample size and performed almost the same for the (10,26) sample size. Kim *et al.* (2006) found that only a small number of the genes in the leukemia data set satisfied the normality assumption, 31.5%, by a Kolmogorov-Smirnov test, so it is assumed that this data is not normally distributed and only 23.7% of the genes satisfy the equal variance assumption. The violation of the assumptions could have affected the performance of the two fold change methods since they do assume that variances are equal. The same assumption of equal variance applies to SAM and samroc while the rank product method is more robust and only requires about equal variances. Therefore maybe further analysis of other real data sets could give us a better idea of the performance of the different methods.

While the analysis on the Apo AI dataset showed that the median fold change method was an improvement to the original fold change, it also gave a nice visualization of how the different methods are inconsistent with each other when identifying differentially expressed genes. Not many genes were found in common when comparing across multiple methods, displaying the difficulty of finding truly differentially expressed genes when analyzing microarray data.

A table with the suggested cutoff values to control for a type 1 error rate of 0.05 or less for the fold change and median fold change methods was also provided at the end of Chapter 3. These cutoff values can be suggested only as guidelines and need to be evaluated by the researcher to determine if they represent an actual biological difference in the particular genes being analyzed. The fold change method is appealing because it is very simple to use but the fact that the cutoff value does not take into account the variability in the data, there is apparent weakness to the method.

The procedures presented here considered various sample sizes and simulated samples modeled after normal distributed data and lognormal distributed data at different levels of skewness. Under all the simulations equal variance was simulated between both conditions. When dealing with real microarray data there will be many cases where the genes for the two conditions being analyzed will not have equal variance. The equal variance assumptions of the models presented here will have a meaningful effect on how they perform and should be taken into account when choosing which method is best to use. Simulations with samples of unequal variances should be carried out to further evaluate how the median fold change method performs. Throughout all the simulations and the leukemia data analyses, samroc performed quite well and at many times was the most powerful. However in the analysis of the lognormal data, rank product was slightly better in addition to the fold change methods. Modifying the test statistic of samroc, which is closely related to the SAM test statistic, with the median in place of the mean may have a positive effect on the performance of samroc when the data is skewed.

One last thing to consider is the relevance of microarrays and the analysis of microarrays today when newer technology is becoming popular. With the advancements and reduced costs in next generation sequencing one may ask if there is any reason to use microarrays. Generally microarrays are less complicated and easier to work with and prepare than NGS (Baker 2013). A point Baker (2013) brings up is that since microarrays have been around for a while now, many researchers have become very comfortable with the use of microarrays in practice and have gotten used to interpreting their results. From the research done here, it can be seen that each method varies widely under different situations so the search for the most consistent and best performing method is still sought after.

LIST OF REFERENCES

Baker SC. Next-generation sequencing vs. microarrays [Online]. *Genetic Engineering and Biotechnology News.* [http://www.genengnews.com/gen-articles/next-generation-sequencing-vs-microarrays/4689/]. 14 Jan. 2013.

Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 2001; 17: 509-19.

Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* 2004; 573(1-3): 83-92.

Broberg P. Ranking genes with respect to differential expression. *Genome Biology* 2003; 4:41.

Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* 2000; 10: 2022-29.

Chu G, Narasimhan B, Tibshirani R, Tusher V. SAM Significance Analysis of Microarrays-User guide and technical document [Online]. [http://www-stat.stanford.edu/~tibs/SAM/sam.pdf]. 2002.

Efron B, Tibshirani R, Storey JD, Tusher V. Emperical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; 96: 1151-1160.

Golub TR, Slonim DK, Tamajo P, Huard C, Gaosenbeek M, Mesirov JP, Coller H, Loh ML, Dowing JK, Claigiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531-7.

Hoyle DC, Rattray M, Jupp R, Brass A. Making sense of microarray data distributions. *Bioinformatics* 2002; 18(4): 576-584.

Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Medical Research Methodology* 2005; 5(1): 3-20.

Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006; 7: 359.

Kim SY, Lee, JW, Sohn IS. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. *Statistical Methods Medical Research* 2006; 15: 3.

Koziol JA. The rank product method with two samples. *FEBS Letters* 2010; 548(21): 4481-4484.

Majtán T, Bukovská G, Timko J. DNA microarrays--techniques and applications in microbial systems. *Folia Microbial (Praha).* 2004; 49(6): 635-64.

Manikandan S. Measures of central tendency: median and mode. *Journal of Pharmacology and Pharmacotherapeutics* 2011; 2(3): 214-215.

McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 2009; 25(6): 765-771.

Pan W. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* 2003; 19(11): 1333-1340.

Pan W, Lin J, CT Le. A mixture model approach to detecting differentially expressed genes with microarray data. *Funct. Integr. Genomics* 2003; 3: 117-124.

Schwender H, Krause A, Ickstadt K. Comparison of the emperical bayes and the significance analysis of microarrays. *Technical Report*. SFD 475: Dortmund, Germany: University of Dortmund; 2003.

Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLOS Computational Biology* 2013; 9(11): e1003342.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 2004; 3: Article 3.

Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics Gynecology* 2006; 195(2): 373-388.

Troyanskaya OG, Garber ME, Brown PO, Botstein D, and Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *BMC Bioinformatics* 2002; 18:11: 1454-1461.

Tusher VG., Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; 98: 5116-21.

Van de Weil MA. Significance analysis of microarrays using rank scores. *Kwantitatieve Methoden* 2004; 71: 25-37.

Zhang S. A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics* 2007; 8:230.

Zhao Y, Pan W. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* 2003; 19: 1046-54.

APPENDICES

Appendix 1 The following graphs are results from simulations of lognormal distributed data. The blue columns correspond to the power and the red columns correspond to the P(type 1 error).

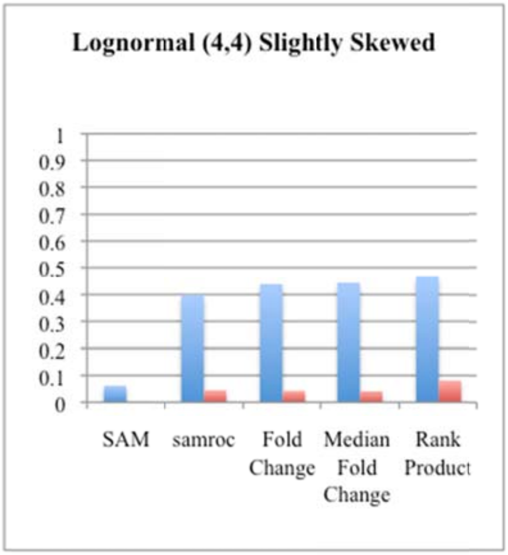


Figure 1.1

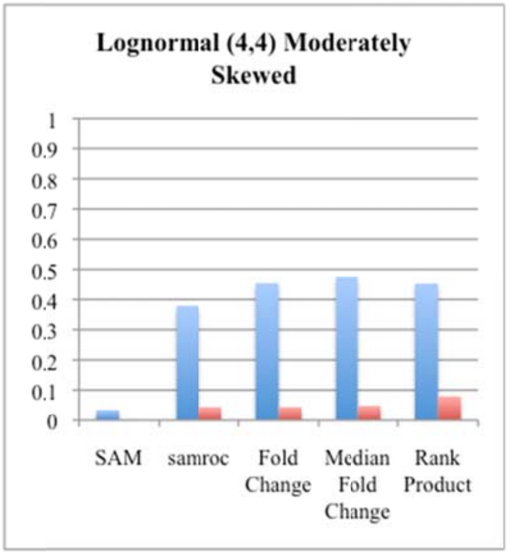


Figure 1.2

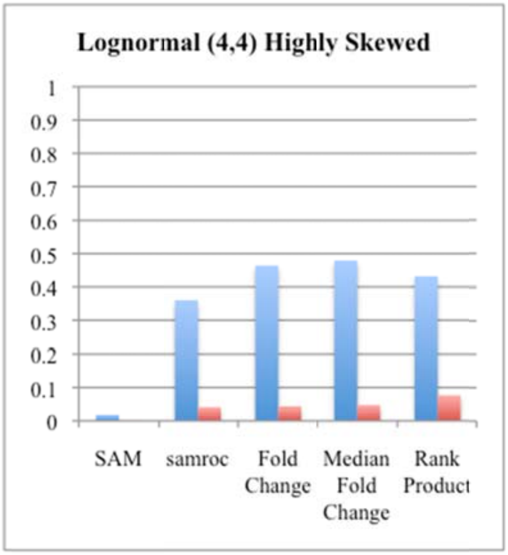


Figure 1.3

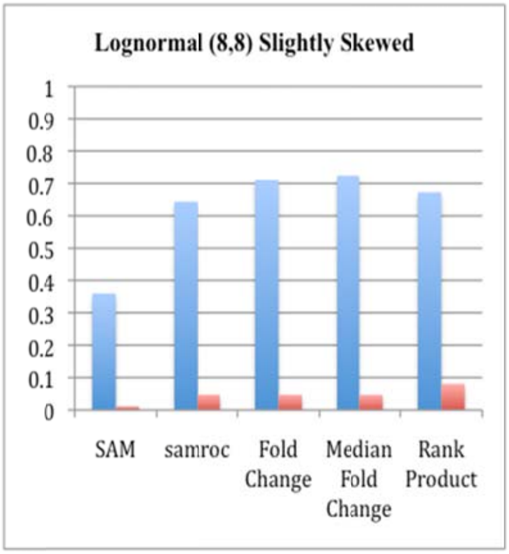


Figure 1.4

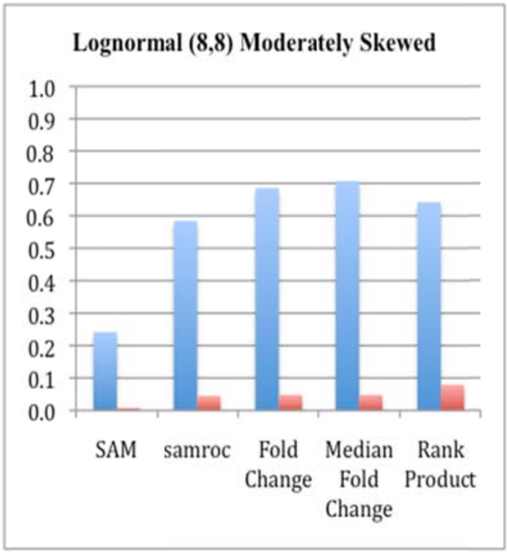


Figure 1.5

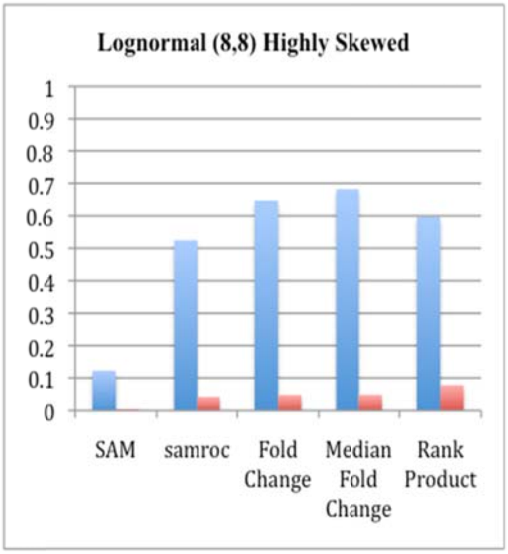


Figure 1.6

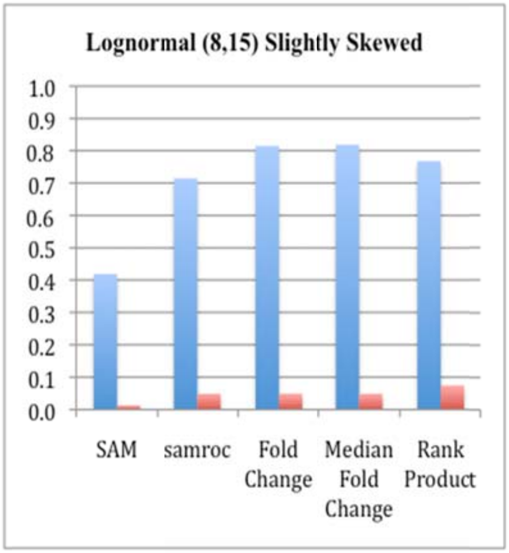


Figure 1.7

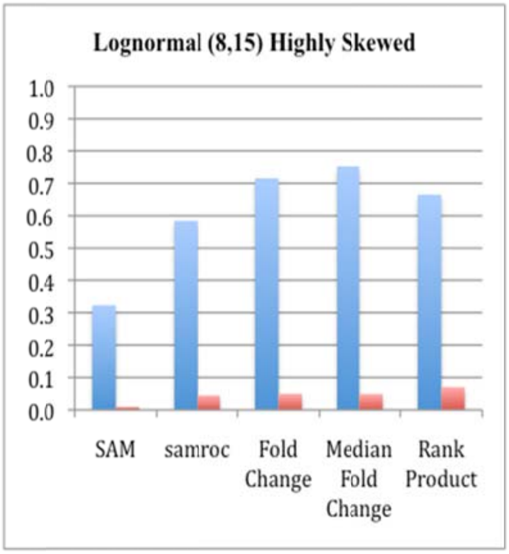


Figure 1.8

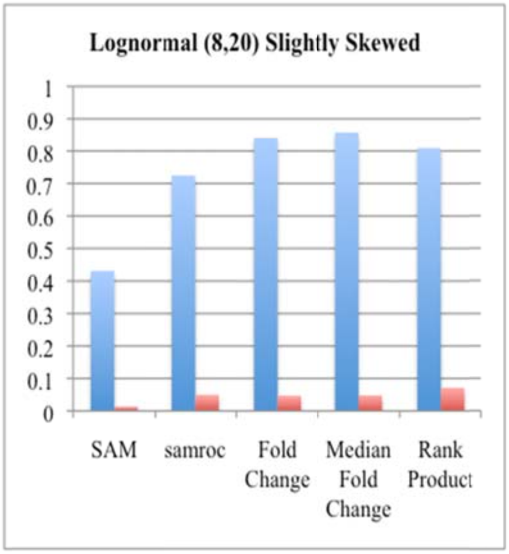


Figure 1.9

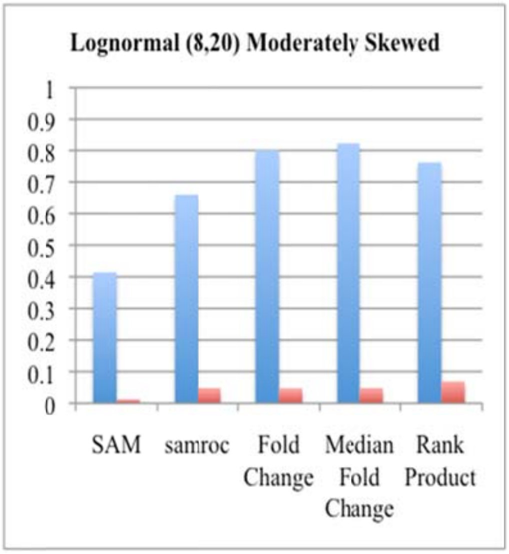


Figure 1.10

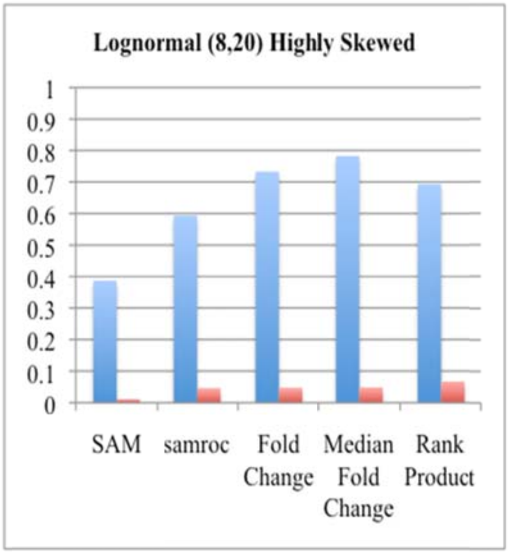


Figure 1.11

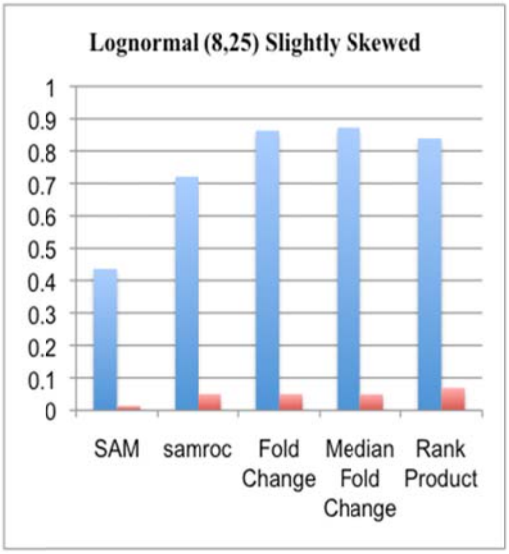


Figure 1.12

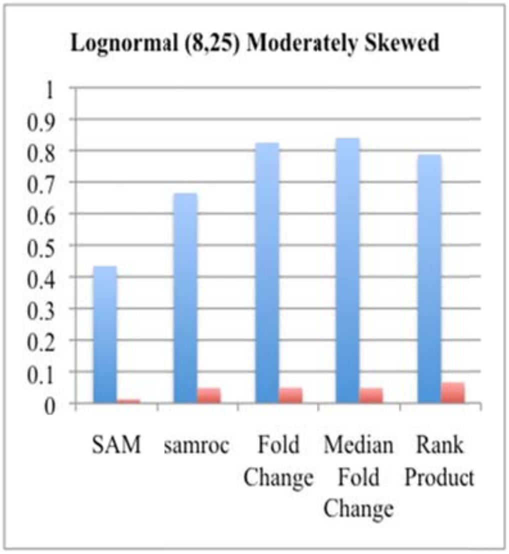


Figure 1.13

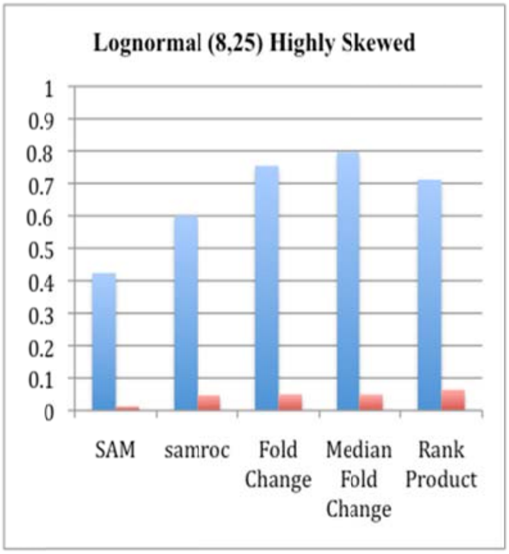


Figure 1.14

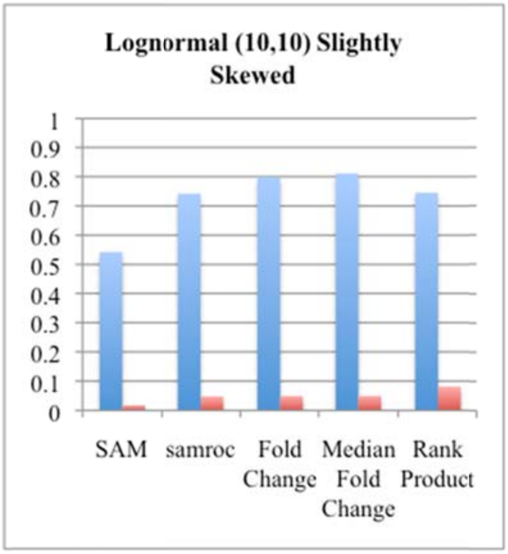


Figure 1.15

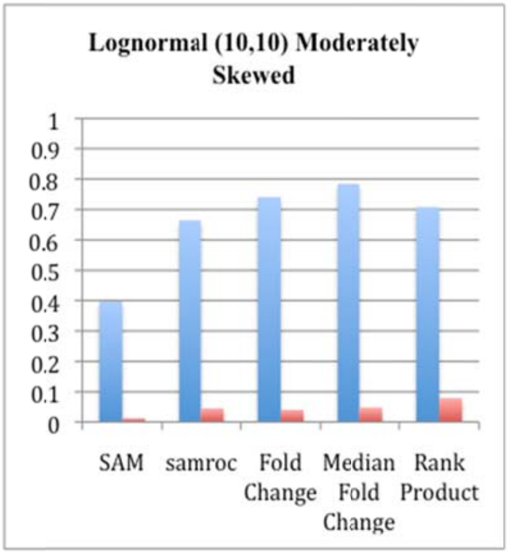


Figure 1.16

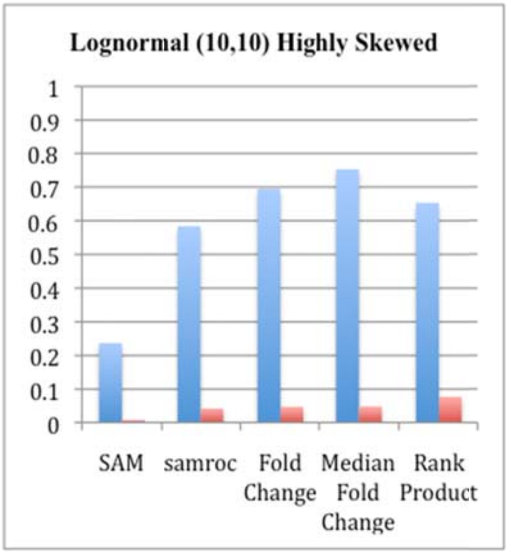


Figure 1.17

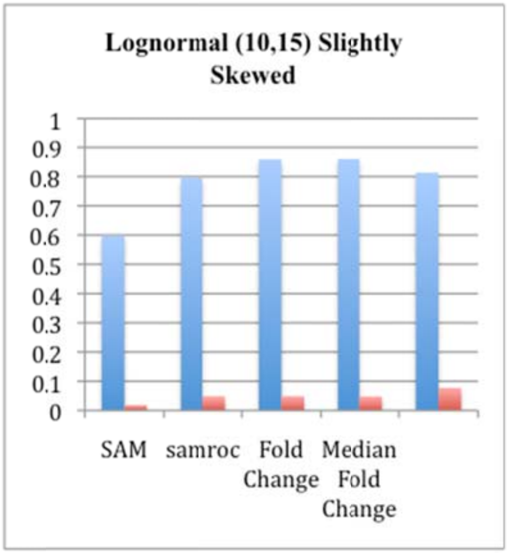


Figure 1.18

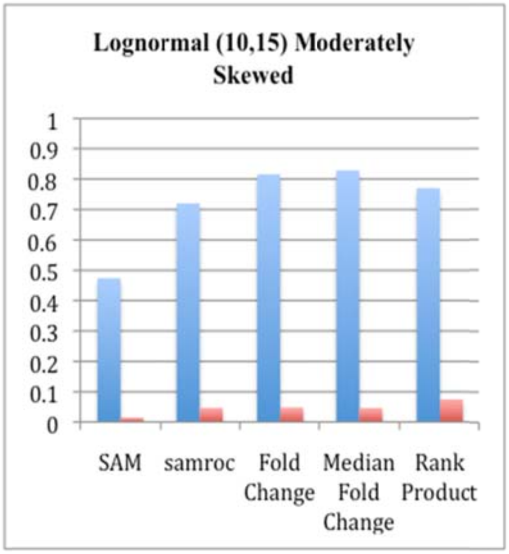


Figure 1.19

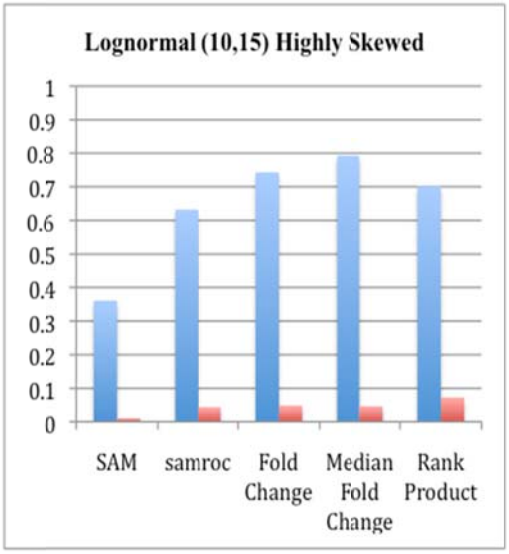


Figure 1.20

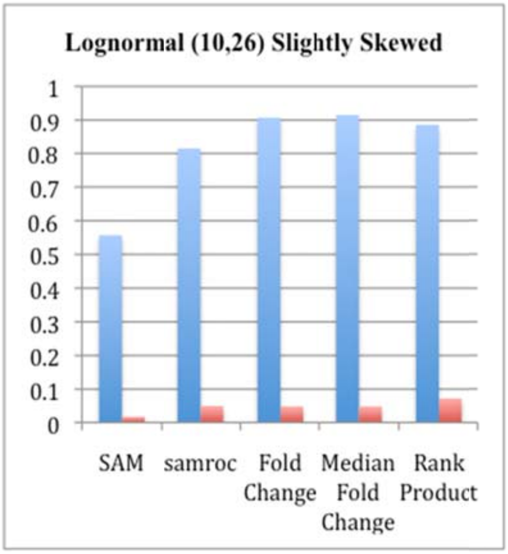


Figure 1.21

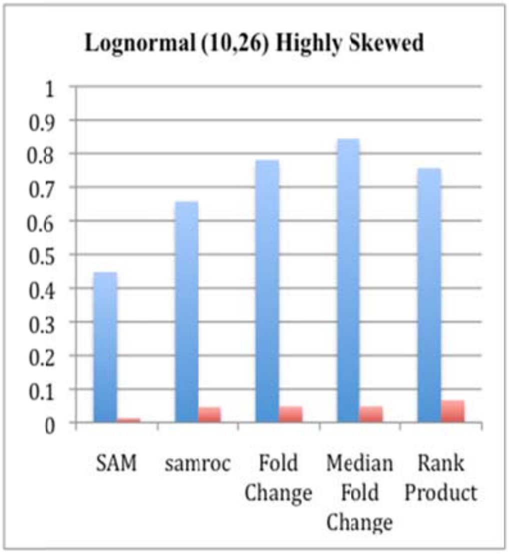


Figure 1.22

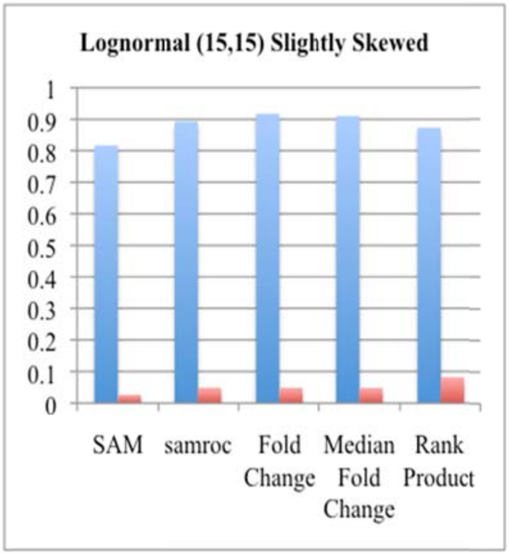


Figure 1.23

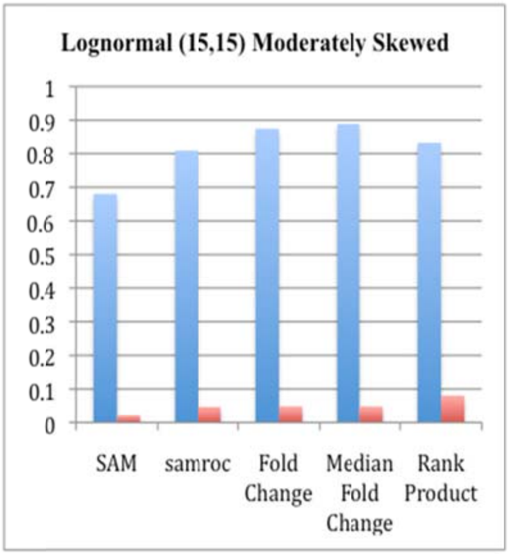


Figure 1.24

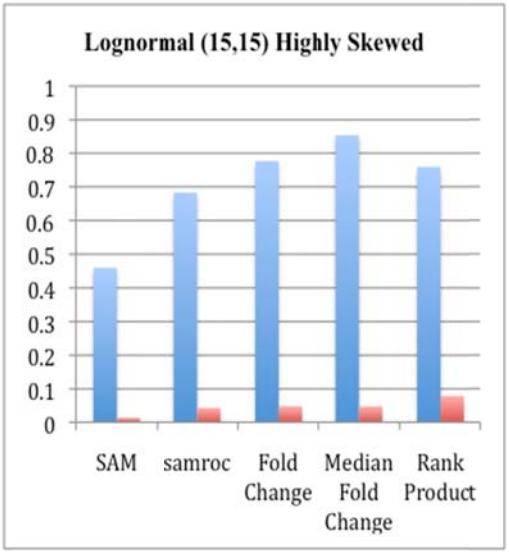


Figure 1.25

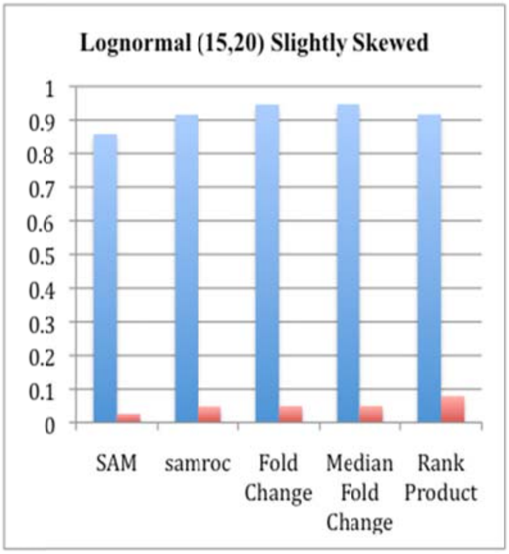


Figure 1.26

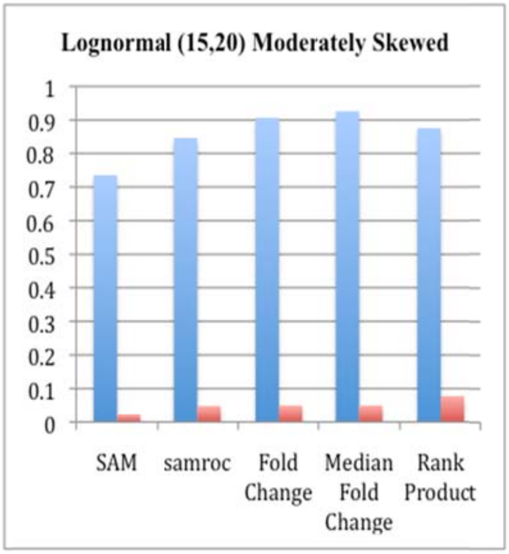


Figure 1.27

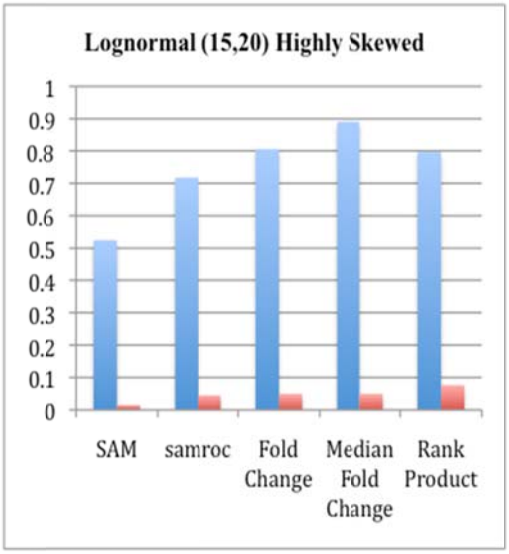


Figure 1.28

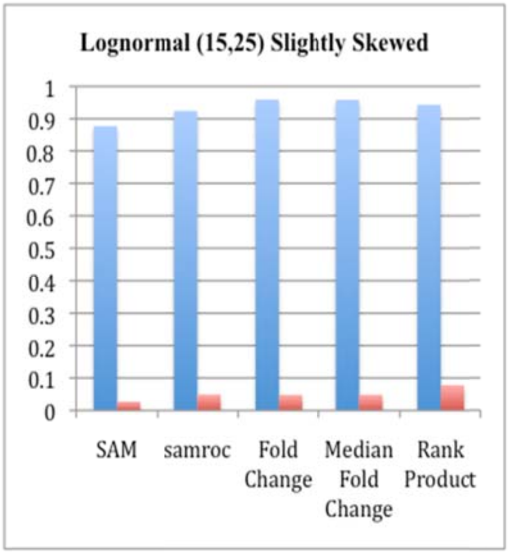


Figure 1.29

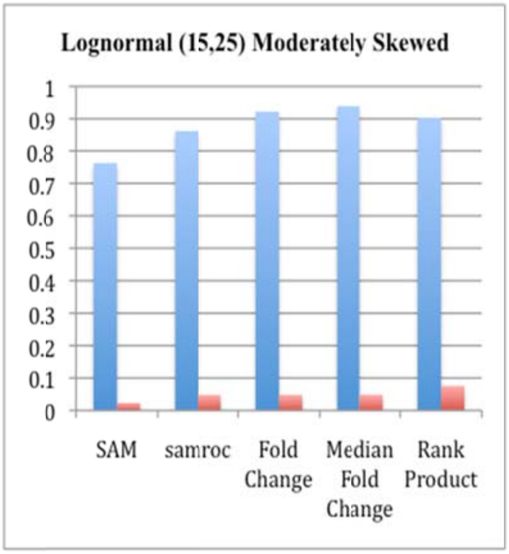


Figure 1.30

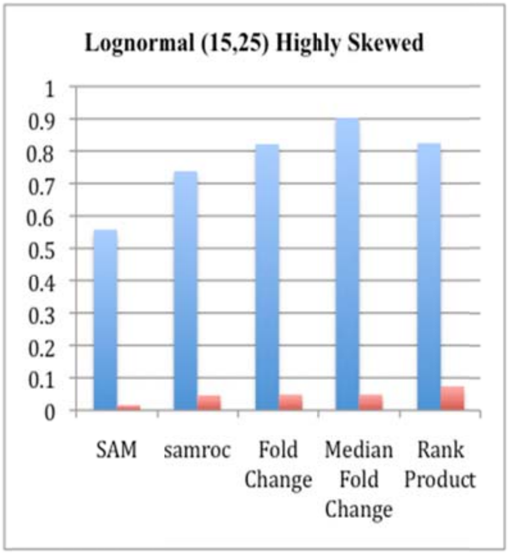


Figure 1.31

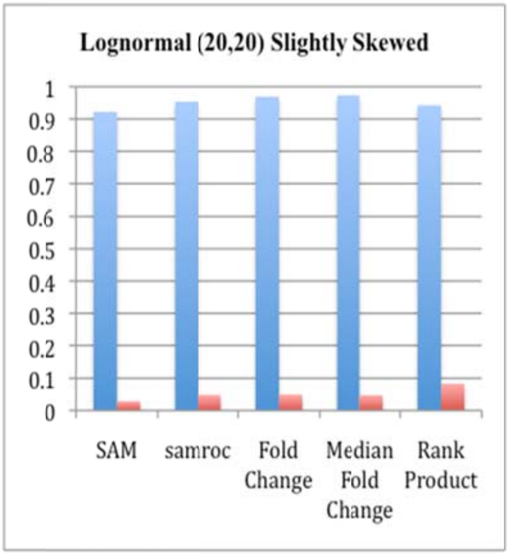


Figure 1.32

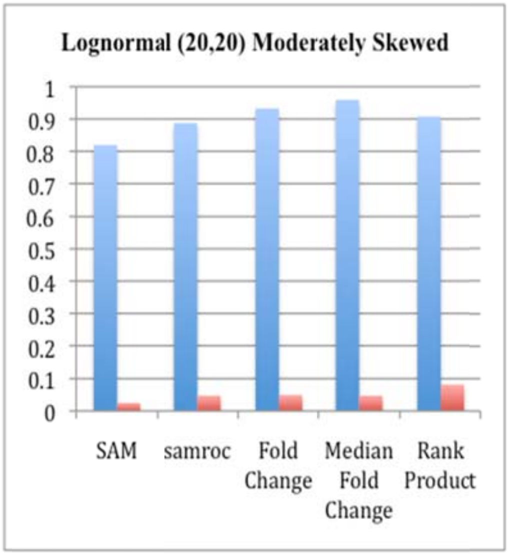


Figure 1.33

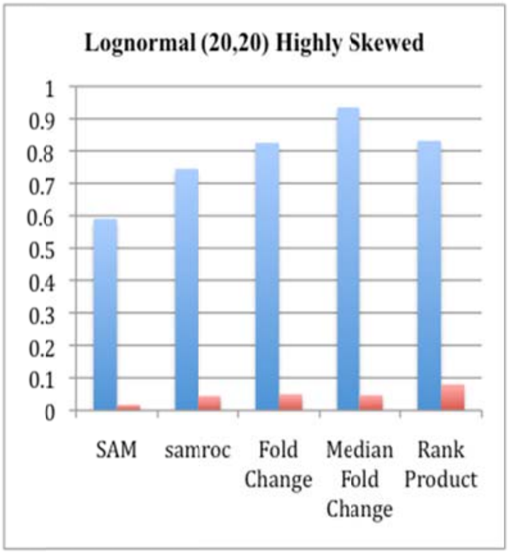


Figure 1.34

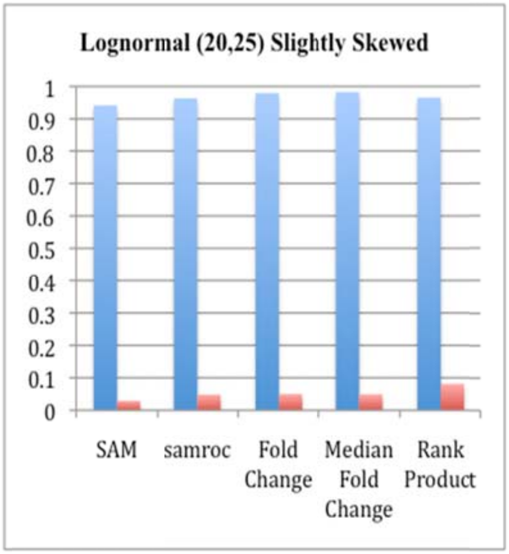


Figure 1.35

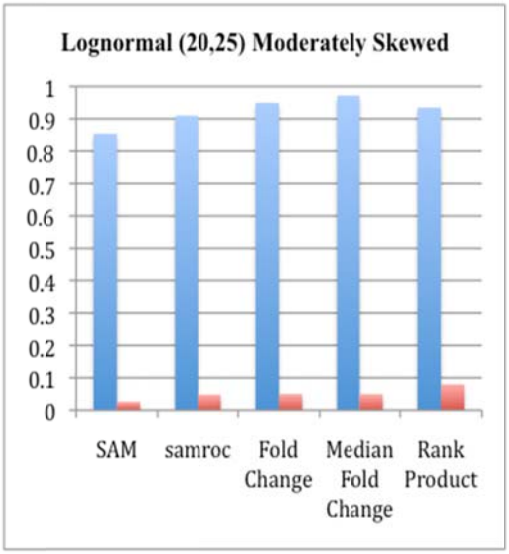


Figure 1.36

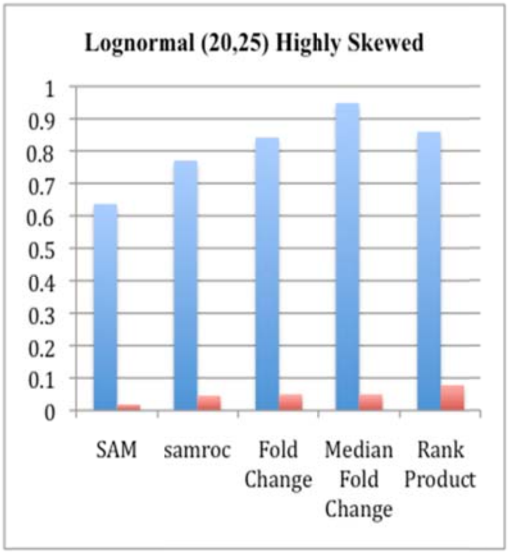


Figure 1.37

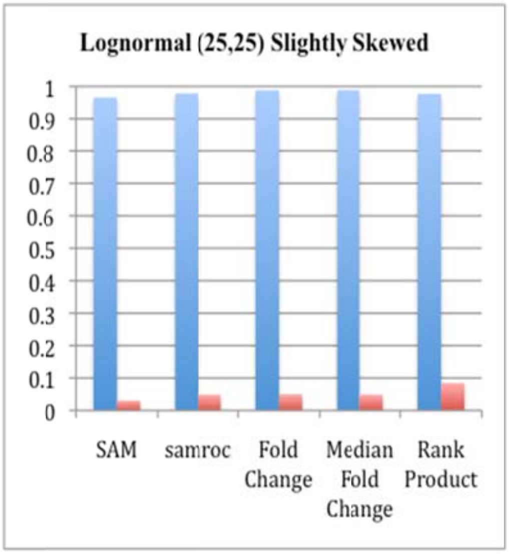


Figure 1.38

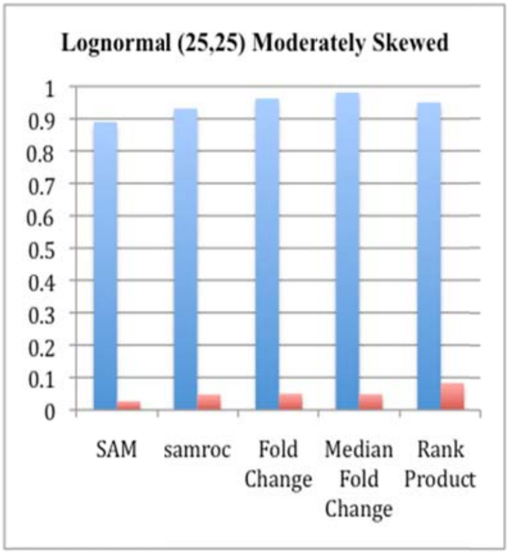


Figure 1.39

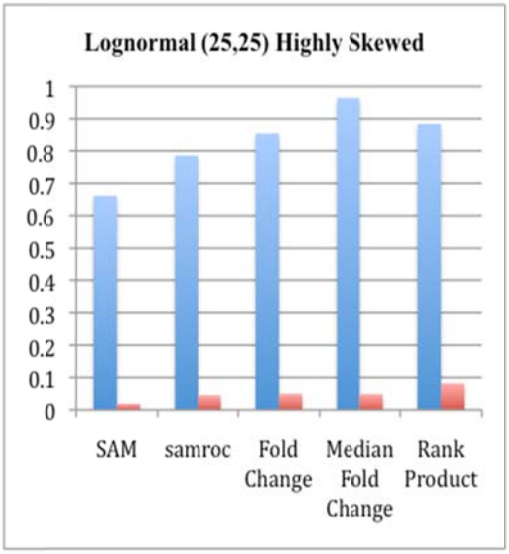


Figure 1.40

Appendix 2 The following graphs and tables are results from simulations of normally distributed data. The blue columns correspond to the power and the red columns correspond to the P(type 1 error).

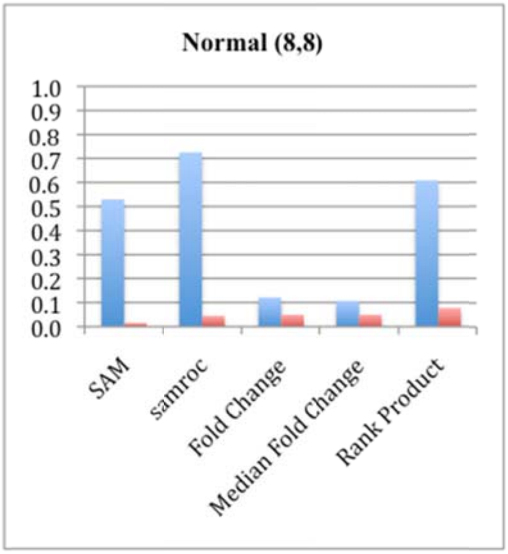


Figure 2.1

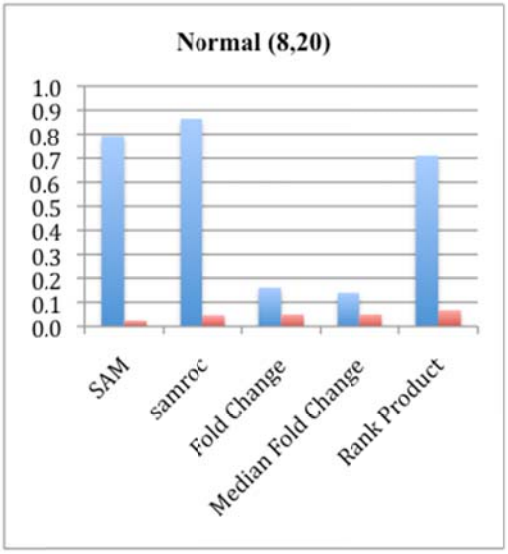


Figure 2.2

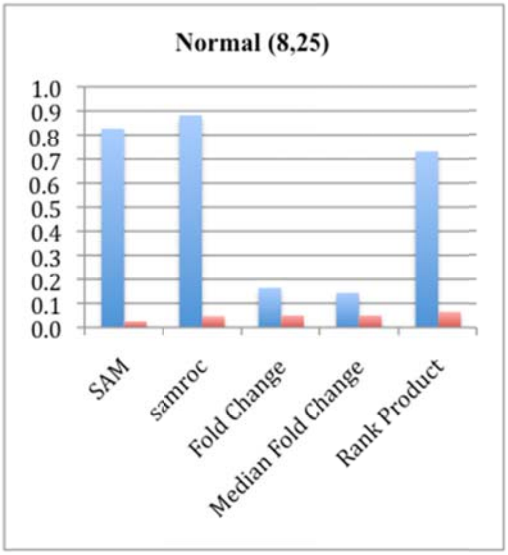


Figure 2.3

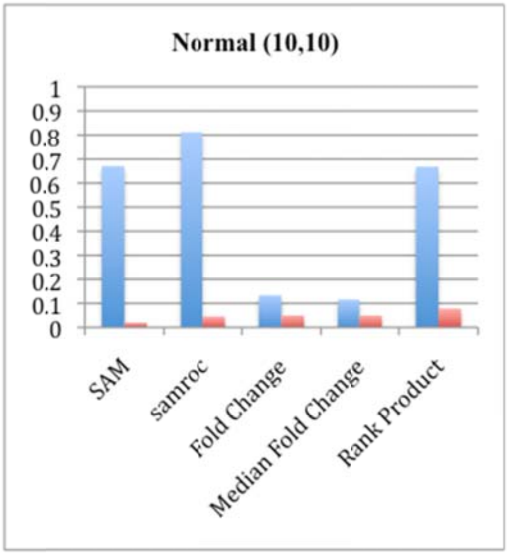


Figure 2.4

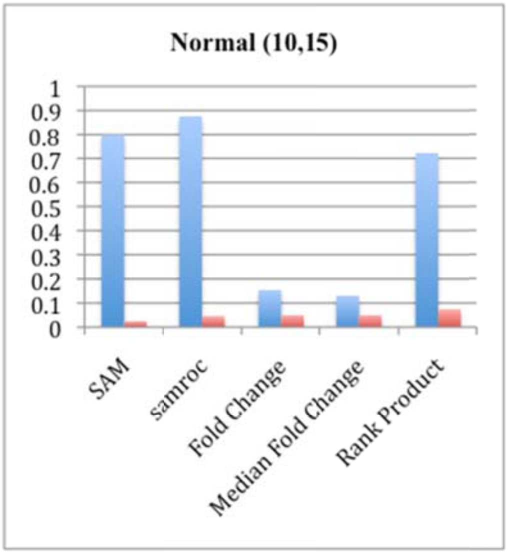


Figure 2.5

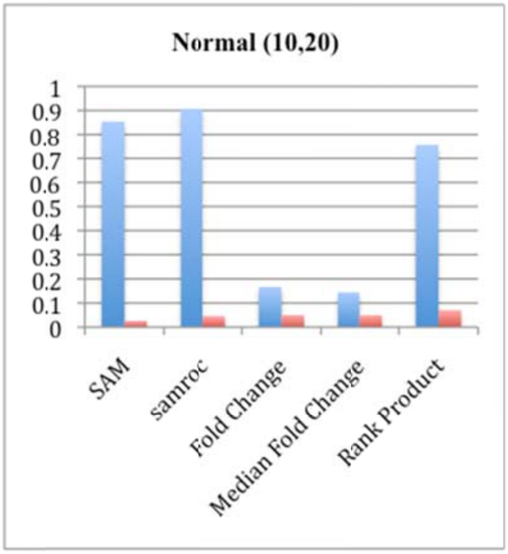


Figure 2.6

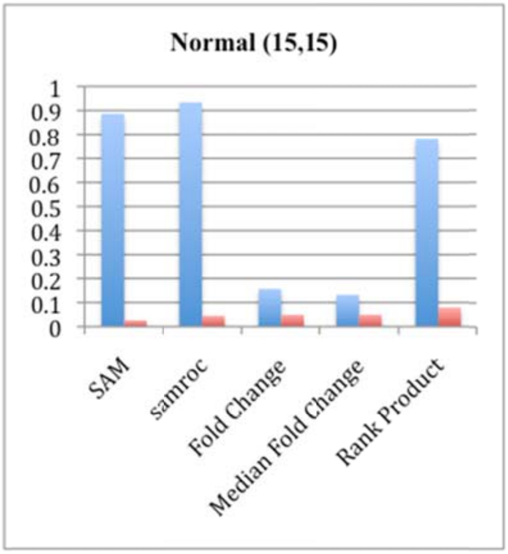


Figure 2.7

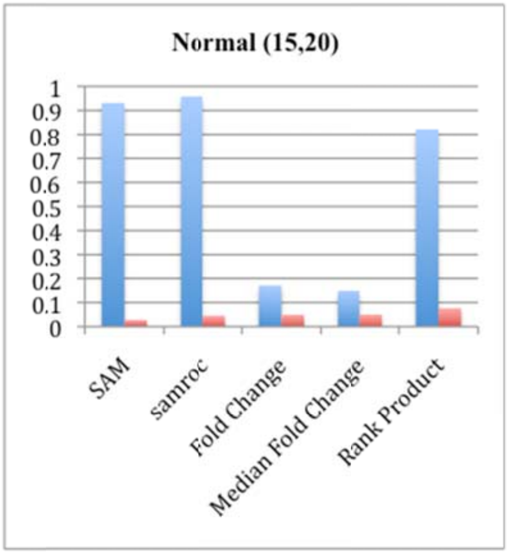


Figure 2.8

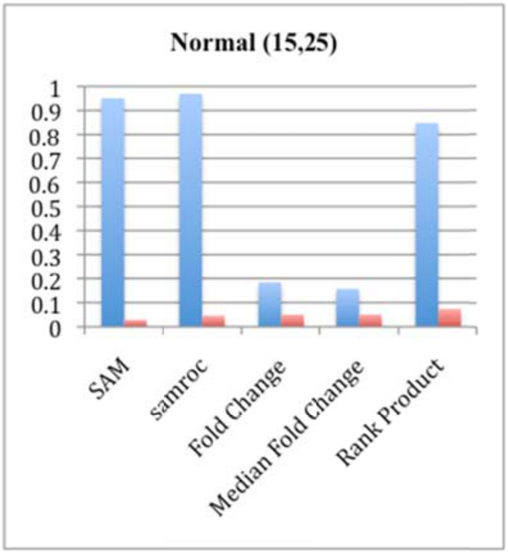


Figure 2.9

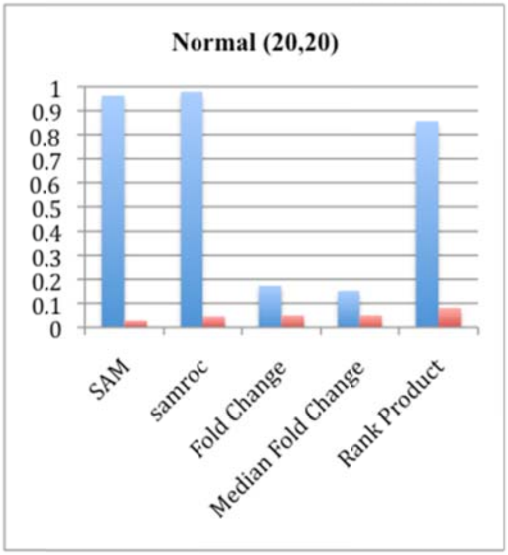


Figure 2.10

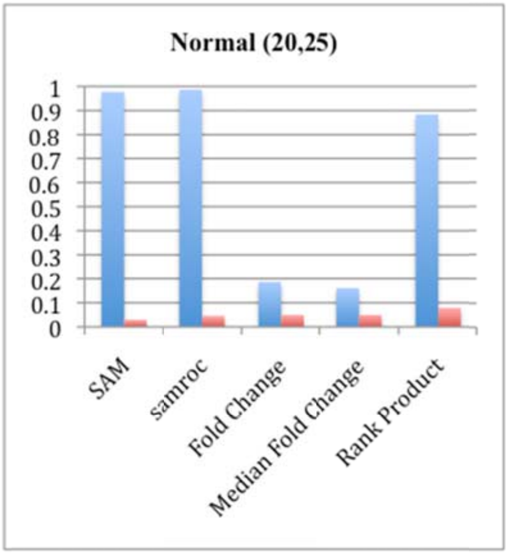


Figure 2.11

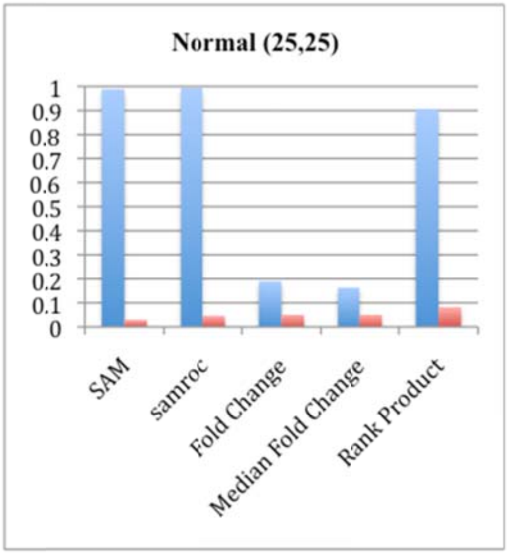


Figure 2.12

Table 2.1 Cutoff values for fold change and median fold change with at most 0.05 P(type 1 error) under normal distribution

Sample Size	Fold Change	Median Fold Change
(4,4)	6.38	6.38
(8,8)	6.38	6.38
(8,15)	6.57	6.53
(8,20)	6.95	6.90
(8,25)	7.40	7.20
(10,10)	6.40	6.40
(10,15)	6.45	6.45
(10,20)	6.70	6.65
(10,26)	6.95	6.90
(15,15)	6.40	6.40
(15,20)	6.45	6.45
(15,25)	6.49	6.49
(20,20)	6.35	6.35
(20,25)	6.37	6.37
(25,25)	6.32	6.32