RANDOMIZATION TEST AND CORRELATION EFFECTS IN HIGH DIMENSIONAL DATA

by

XIAOFEI WANG

M.S., Chinese Academy of Agricultural Sciences, 2008

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Gary Gadbury

Copyright

Abstract

High-dimensional data (HDD) have been encountered in many fields and are characterized by a “large p, small n” paradigm that arises in genomic, lipidomic, and proteomic studies. This report used a simulation study that employed basic block diagonal covariance matrices to generate correlated HDD. Quantities of interests in such data are, among others, the number of ‘significant’ discoveries. This number can be highly variable when data are correlated. This project compared randomization tests versus usual t-tests for testing of significant effects across two treatment conditions. Of interest was whether the variance of the number of discoveries is better controlled in a randomization setting versus a t-test. The results showed that the randomization tests produced results similar to that of t-tests.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank all faculty and staff members in Department of Statistics at Kansas State University. Thanks a lot for the opportunity given to me to pursue a degree in Statistics. Thanks a lot for their guidance and supervision on my study and daily basis.

I would like to extend my gratitude to my committee members, Dr. Gary Gadbury, Dr. Sunghun Park, and Dr. Weixin Yao. Special thanks to my advisor, Dr. Gary Gadbury, for his conuntless help for my research work. Without his endless patience, invaluable advice, and intensive care, this work would not be a possibility.

Finally, a hearty appreciation is given to my family and friends.

Chapter 1 - Introduction

With the development of modern technologies, high-dimensional data (HDD) have been encountered in many fields, especially in areas of biological and medical research. A common characteristic of HDD is that the number of features or variables is very large in the dataset, while the sample size is relatively small. It refers to a "large p, small n" paradigm (West, 2003) that arises in genomic, lipidomic, proteomic, and microarray studies. In particular, microarray technology enables researchers to monitor the expression levels of tens of thousands of genes simultaneously. The context of the microarray experiments can be found in Allison et al. (2005) and Göhlmann et al. (2009). This report will focus on microarray data, but the technical discussions are applicable to studies involving HDD.

A common interest in microarray studies is to detect the genes that are differentially expressed because of a treatment effect. In light of this goal, we need to estimate the difference in gene expression and test whether the observed differences are induced by treatment or by chance. In this process, two important issues emerge. One is the choice of an appropriate reference distribution for computing valid P-values. The other is the correlation structure among thousands of genes, which can cause correlation among test statistics from multiple tests. See Hu et al. (2010) to get more details about invalid P-values due to an incorrect reference distribution.

Results given by P-values can be misleading if required assumptions are violated. Among the various techniques to quantify the difference across levels of treatment for a single gene, the student's t-test is a common choice since it is relatively robust to moderate violations of the normality assumption. In addition, the sample means of moderately large samples are approximately normal distribution based on the central limit theorem even though the individual data values may not be normally distributed (Boneau, 1960; Edgell et al., 1984).

However, the sample sizes were often small in early microarray experiments because of the cost of arrays, and the gene expression may not be normally distributed (Lee et al., 1999). So, resampling-based procedures (RBPs) or nonparametric tests became an alternative for researchers. RBPs have certain advantages, such as eliminating some parametric assumptions and being robust and flexible to accommodate different test statistics (Gadbury et al., 2003). See Mehta (2006) to get more details about two RBPs – randomization and the bootstrap.

Recently, microarray technologies have become more affordable. For example, a sample size of 173 was used for the real case study in this report. Thus, sample size is a lesser limitation for the application of a student's t-test to HDD. However, the correlation structure is still a challenge for high-dimensional investigations since many genes are co-regulated in an organism, and estimation of the dependence structure is unrealistic because of the very large number of features relative to the available samples. In addition, the simulation of realistic data is very complicated in practice. Paranagama (2011) discussed the challenges of generating realistic data. Most studies used a rigid dependence structure for simulation studies (Gadbury et al., 2003; Hu et al., 2011). Gadbury et al. (2008) proposed a plasmode method for generating data which were closer to the structure of real data. Paranagama (2011) extended this method and suggested a new plasmode method to simulate data with more original structure preserved in the datasets.

In this report, we employed a simple rigid dependence structure, i.e., basic block diagonal covariance matrices (equicorrelaiton matrices and normal distribution theory) to generate correlated data for simulations, so that we can use a correct test statistic to compute the P-value for a single gene. Then, the P-values will be valid for all genes and the distribution of P-values will be expected to be uniform under the global null hypothesis and no correlation effects. A global null hypothesis is defined as the null hypothesis is true for every test. However, the distribution of P-values may vary widely from experiment to experiment when the HDD are correlated , even under the global null hypothesis. In some cases, there are fewer P-values clustering near 0 than expected under the global null hypothesis. In other cases, the distribution of P-values is more likely to cluster near 0. In addition, the distribution of P-values is expected to be clustering near 0 when the alternative hypothesis is true. So, the distribution of P-values may give misleading results when HDD are correlated.

Next, we will focus on comparison of randomization tests with usual t-tests for testing of significant effects across two treatment conditions in correlated HDD. Correlated tests can give misleading results, in part because the variance of the number of discoveries is inflated using t-tests (Hu et al., 2010). So, of interest was whether the variance of the number of discoveries is better controlled in a randomization setting versus using a t-test. In fact, a randomization test, where entire high-dimensional vectors are permuted across treatment conditions, helps preserve the dependence structure in HDD, but it does not accommodate the correlation effects in high-dimensional testing situations (Efron, 2007; Efron 2010) .

In the context of this research, the results show that the randomization tests produce results similar to results from t-tests. When the data are uncorrelated, P-values are valid and the distribution of P-values is meaningful to illustrate biological interests. However, the distribution of P-values may give misleading results if the data are correlated. Thus, randomization tests are no direct assistance with the correlation effects in HDD.

The remainder of this report is organized as follows. Firstly, some key references will be reviewed which are closely related to this field. Then, the simulation studies will be described in Chapter 3. In Chapter 4, we will present a real case study involving a microarray experiment. Finally, the conclusion and further study will be given in Chapter 5.

Chapter 2 - Literature Review

The investigation of high-dimensional data (HDD) has been an active topic in recent research. This chapter will briefly review some literature applicable to this report.

2.1 Introduction of high-dimensional data

High-dimensional data are defined as sets of data in which the number of features compared to the sample size is very large. They refer to a "large p, small n" paradigm arising in "omics" studies. Among these studies, microarray data are a good example of HDD. For microarray studies, two treatment conditions are very common (e.g. Control vs Treatment). The following table shows the structure of HDD for two treatment conditions. A measure of genetic expression level for the ith gene of the jth experimental unit is denoted by Y_{ij}^{c} for control group and Y_{ij}^{t} for treatment group.

Table 2.1 HDD structure for two treatment conditions

	Control			Treatment		
EU / i	1	…	n_1	1	…	n_2
1	Y_{11}^{c}	…	$Y_{1n_1}^{c}$	Y_{11}^{t}	…	$Y_{1n_2}^{t}$
…	Y_{i1}^{c}	…	$Y_{in_1}^{c}$	Y_{i1}^{t}	…	$Y_{in_2}^{t}$
K	Y_{k1}^{c}	…	$Y_{kn_1}^{c}$	Y_{k1}^{t}	…	$Y_{kn_2}^{t}$

In microarray data analysis, of interest is to identify the genes that are differentially expressed between two treatment groups. So, multiple tests are performed for K features simultaneously. This inflates the number of type I errors because K is very large and the probability is $1-(1-\alpha)^k$ for at least one is falsely rejected null hypothesis. There are several conventional methods employed to deal with this problem by controlling FWER (i.e., the

probability of one or more Type I errors), such as Bonferroni, Dunett, Tukey, ect. But they are too conservative when the number of comparisons is very large. Take the Bonferroni method for example, it sets the significance cut-off at α/K which is a very small value when K is large and it is known to be too conservative for HDD analysis. In our real case study, the significance cut-off is at 0.05/3790 =1.32x10^{-5} if we want to control FWER at 0.05. It is too conservative and may fail to detect the genes that are differentially expressed between two conditions.

In 1995, Benjamini and Hochberg proposed their landmark work by introducing the false discovery rate (FDR). FDR is the expected proportion of false findings among all significant discoveries. Let K be the total number of tests, R be the significant discoveries at some threshold, and P be the number of non-significant results. In a real dataset, only these three measures are known. Other accuracy measures are given in Table 2.2. Then, FDR is defined by

$$FDR = \begin{cases} E(V/R) & R > 0 \\ 0 & R = 0 \end{cases}$$

. In addition, Benjamini and Yekutieli (2001) proved that the same procedure with a slight modification controlled the false discovery rate under a particular positive dependency structure. Pawitan et al. (2006) developed a procedure for estimating FDR using a latent variable approach under general dependence.

Table 2.2 Notations for accuracy measures

	True Ho	True Ha	Total
Declared significant	V	S	R
Declared non-significant	U	T	P
Total	L	M	K

L is the number of true null hypotheses

M is the number of true alternative hypotheses

V is the number of false positives (Type I error)

S is the number of true positives

T is the number of false negatives (Type II error)

U is the number of true negatives

Storey (2002) defined the FDR as $FDR = E\left(\frac{V}{R} \mid R > 0\right) P(R > 0)$ and "positive" false discovery rate (pFDR) as $pFDR = E\left(\frac{V}{R} \mid R > 0\right)$. The q value is a measure of significance in terms of FDR. It is the minimum pFDR at which the individual test may be significant. The q-value is calculated as follows (Storey, 2002):

For the K hypothesis tests, compute the P-values $P_1, ..., P_K$.

Order the P-values $P_{(1)} \le ... \le P_{(K)}$.

Set $\hat{q}(P_{(K)}) = p\hat{FDR}(P_{(K)})$.

Set $\hat{q}(P_{(i)}) = \min\left\{p\hat{FDR}(P_{(i)}), \hat{q}(p_{(i+1)})\right\}$ for i=K-1, K-2, …, 1.

2.2 Application of t-test and randomization test in HDD study

For comparison of two treatment conditions, a two-sample t-statistic is often computed for testing of the significance. Allison et al. (2002) employed this parametric approach to produce a P-value and then used a mixture model (uniform distribution and beta distribution) on a distribution of P-values to analyze the gene expression data. Pawitan (2005) illustrated the necessity of FDR for analysis of microarray data by applying the usual t-test. Cao (2011) suggested a method to produce the simultaneous critical values for rejection regions by using t-statistics.

As we mentioned before, the sample sizes are small in the early microarray studies and the assumptions of the parametric approach may not always be met. Thus, randomization tests have some advantages to be considered as an alternative. Gadbury et al. (2003) relied on 'nonparametric' randomization tests to produce an exact P-value for each gene under an additive model when the small sample size is small. They also showed that the distribution of P-values on all genes can yield valuable information to answer the biological question – whether the gene expression levels are significantly different between control and treatment group. Barry et al. (2005) described a permutation-based framework, significance analysis of function and expression (SAFE) to analyze data from gene expression studies. Subramanian et al. (2005) proposed a method called gene set enrichment analysis (GSEA), and then did permutations to compute the statistics for assessing the significant gene-sets. Efron and Tibshirani (2007)

suggested two improvements to GSEA. Dudoit (2002) used a permutation procedure to detect the differentially expressed genes by estimating the adjusted P-values. Hall and Tajvide (2002) suggested a permutation approach to test the equality of distributions in an HDD setting. In addition, some researches are involved in the comparison of t-tests and permutation tests. Tsai et al. (2003) compared the type I error and power of the t-tests and permutation test for detecting differentially expressed genes between two microarray sample sets.

2.3 Correlation effects in HDD analysis

Since many genes are co-regulated in an organism, correlation exists among gene expression levels (Qui et al., 2005). It is a challenge for investigating HDD because it is unrealistic to estimate the dependence structure, and difficult to simulate realistic data (Hu et al., 2010; Paranagama, 2011).

In 2010, Hu et al. illustrated the distribution of P-values in HDD analysis. They showed that a histogram of valid P-values is expected to be approximately uniform between 0 and 1 under the global null hypothesis. But when the valid P-values are correlated, the distribution of P-values under the global null hypothesis may vary widely from experiment to experiment and may give us misleading results. Also, the research showed that the P-values are no longer expected to be uniformly distributed under the global null hypothesis if they are computed from incorrect statistical tests.

In addition, Hu et al. (2010) discussed the issues about correlated P-values under the global null hypothesis. They adopted the idea of Schweder and Spjøtvoll (1982) and simulated the test statistics using a restrictive dependence structure, which means that all pairwise correlations of two test statistics are equal within a block while the test statistics are uncorrelated in different blocks. Let p_i be a P-value for the ith feature, α be a threshold, m be the number of blocks, and N_b be the number of tests declared not significant for one block with size b.

D_i is defined as follows, $D_i = \begin{cases} 0 & p_i \le \alpha \\ & \\ 1 & p_i > \alpha \end{cases}$, which is a Bernoulli variable. Then, N_b is the summation of D_i. The expected value of N_b is $E_{H_0}(N_b) = E_{H_0}\left(\sum_{i=1}^{b} D_i\right) = b(1-\alpha)$ and the

variance of N_b is $V_{H_0}(N_b) = V_{H_0}\left(\sum_{i=1}^{b} D_i\right) = b\alpha(1-\alpha) + b(b-1)Cov(D_1, D_2)$. The covariance of D_1 and D_2 is given as,

$$\begin{aligned} Cov(D_1, D_2) &= P(D_1 = 1, D_2 = 1) - (1-\alpha)^2 \\ &= P(|T_1| \le z_{\alpha/2}, |T_2| \le z_{\alpha/2}) - (1-\alpha)^2 \\ &= \Phi(z_{\alpha/2}, z_{\alpha/2}) - \Phi(-z_{\alpha/2}, z_{\alpha/2}) - \Phi(z_{\alpha/2}, -z_{\alpha/2}) \\ &+ \Phi(-z_{\alpha/2}, -z_{\alpha/2}) - (1-\alpha)^2, \end{aligned}$$

where T1 and T2 represent two test statistics with correlation equal to ρ, Φ(z1, z2) is a bivariate normal CDF of (T1, T2) evaluated at (z1, z2). Let N_0 be the number of genes declared significant out of total K genes under the global null hypothesis. Then, the expected value of N_0 is $E_{H_0}(N_0) = K\alpha$ and the variance of N_0 is $V_{H_0}(N_0) = mVar(N_b)$.

2.4 Relationship between dependence structure and correlation of test statistics

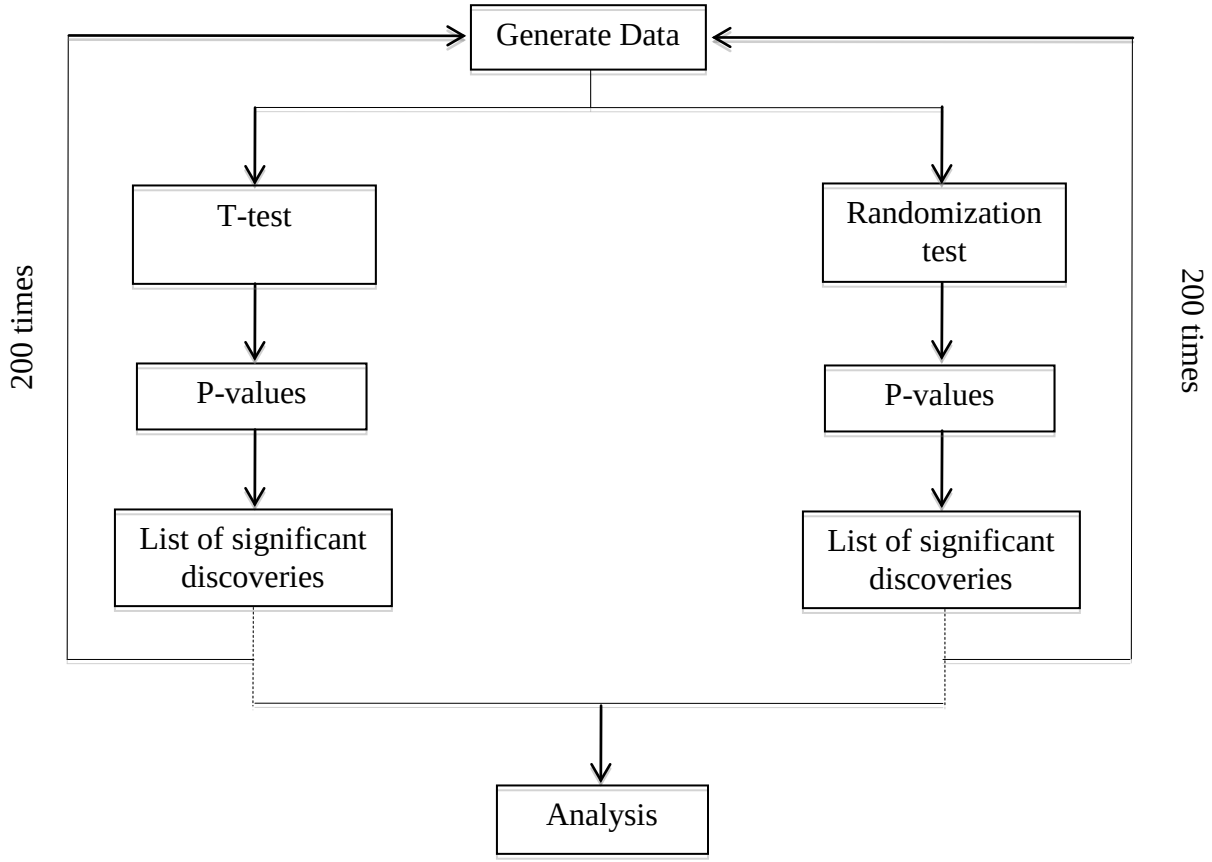
Paranagama (2011) discussed the conditional density of test statistics and illustrated the relationship between dependence structure of the data and correlation of two test statistics. The equation is given by $\rho_{z_{ij}} = \frac{\rho_x \sigma_i \sigma_j + \rho_y \tau_i \tau_j}{\sqrt{(\sigma_i^2 + \tau_i^2)(\sigma_j^2 + \tau_j^2)}}$, where it is assumed that two features within a group have a bivariate normal distribution. For this equation, ρ_x represents the correlation between the ith and jth features in group 1, ρ_y denotes the correlation between these two features in group 2, and σ and τ are the variances for these two features in two groups, respectively.

For the rigid dependence structure, it is assumed that the correlations are the same between two features in two groups and the variances are also the same. Thus, this formula could be simplified by $\rho_{z_{ij}} = \frac{2\rho\sigma^2}{\sqrt{2\sigma^2 * 2\sigma^2}} = \rho$, which means that the correlation between two test statistics is the same as two features in the dataset.

Chapter 3 - Simulation Study

In this section, the results are reported from simulation studies which were designed to compare the performance of randomization tests with usual t-tests. Figure 3.1 depicts the procedure for the simulation study. First, the HDD need to be generated. Then, both t-test and randomization test are employed to analyze the created datasets and a P-value from a test for each gene is calculated using these two tests. Based on P-values, we can get a list of significant discoveries by counting the number of statistically significant tests, one of the quantities of interests for this research. This simulation is repeated 200 times to compute the mean and standard deviation of the number of discoveries. Finally, the results of randomization tests are compared with that of t-tests. For a randomization setting, it is column-wise permutation meaning that the entire high-dimensional samples are permuted across the two treatment conditions.

Figure 3.1 Procedure for Simulation Study



3.1 Simulating high-dimensional datasets

For simulating the data, basic block diagonal covariance matrices in a multivariate normal distribution are employed to create the correlated HDD, where the correlation structure is easily parameterized. The datasets are generated for a total of 1000 genes for two treatment conditions (K=1000, G=2). A 200 × 200 equicorrelation matrix ($\sum_{200}$) is used to incorporate the correlation into the simulations, which has 1 on the diagonal entries and ρ on the off-diagonal entries. Thus, the correlation between all pairs is ρ within each block, and the features are uncorrelated in different blocks. This simulation study used only non-negative values for ρ since the equicorrelation block diagonal matrix was not positive definite when $\rho < 0$. In addition, we assumed that the dependence structure among genes is the same in the two treatment groups and that the expression levels of the same gene are independent in different samples (arrays).

In this study, there were a total of 18 cases for the simulation study based on different means, sample sizes, and correlations. We generated expression levels on 1000 genes and n samples for each group, where n=10 and 20. For the genes chosen to be differentially expressed, data were generated using a mean corresponding to an effect size on the standard normal scale. On average, 25% of the features were simulated as differentially expressed genes between the two groups. So, 250 genes had a higher average expression in the treatment group, while others were not differentially expressed in the two groups. For the different correlations, we changed ρ over 3 values of 0 (independence), 0.3 (moderate dependence), and 0.6 (strong dependence). In addition, variances were set equal to 1. The details of parameters are displayed in table 3.1.

3.2 Quantities of interests

One of the quantities of interests is the P-values. Since the simulation is repeated 200 times, there are 200 vectors of P-values with a length of 1000. We count the number of significant discoveries based on P-values that are below a threshold, and denote this number as *No*. In this study, 3 thresholds were used. For each threshold, there were 200 numbers for the counts of significance. Based on these numbers, the empirical values were computed for the mean and standard deviation of *No*. In addition, the theoretical values also can be calculated for some cases based on the theories mentioned in the literature review.

Table 3.1 Cases for the Simulation Study

Cases	Sample Size (n)	Correlation (ρ)	Standardized Effect Size	$\mu_{Control}$	$\mu_{Treatment}$
1	20	0	0	0	0
2	20	0.3	0	0	0
3	20	0.6	0	0	0
4	10	0	0	0	0
5	10	0.3	0	0	0
6	10	0.6	0	0	0
7	20	0	1.58	0	0.5
8	20	0.3	1.58	0	0.5
9	20	0.6	1.58	0	0.5
10	10	0	1.58	0	0.7
11	10	0.3	1.58	0	0.7
12	10	0.6	1.58	0	0.7
13	20	0	3.16	0	1
14	20	0.3	3.16	0	1
15	20	0.6	3.16	0	1
16	10	0	3.16	0	1.4
17	10	0.3	3.16	0	1.4
18	10	0.6	3.16	0	1.4

3.3 Results of the simulation studies

The focus of the simulation was to assess the performance of randomization tests compared to t-tests. Of interest was whether the variance of the number of discoveries is better controlled in a randomization setting for correlated HDD.

3.3.1 Correlated P-values on the number of significant discoveries

The empirical values were compared between t-tests and randomization tests for mean and standard deviation of the number of significant discoveries. Also, the theoretical values were

compared with empirical values under the global null hypothesis. The outcomes are displayed in table 3.2-3.4 showing that the randomization tests produce similar results to the t-tests. SS represents the sample size and ES represents effect size for each group on a standardized scale. In cases 1-6, the effect size is 0, meaning that no genes are differentially expressed across the two treatment groups (that is, the global null hypothesis is true). In cases 7-12, the effect size is 1.58, meaning that some genes are differentially expressed across the two treatment groups. In cases 13-18, it is also for the simulations under the alternative hypothesis; however, the effect size is twice the cases of 7-12. E(*No*) is the expected number of significant discoveries out of 1000 genes. Sd(*No*) is the standard deviation of the number of significant discoveries.

As mentioned in the literature review, the expected number of differentially expressed genes is $K\tau$ under the global null hypothesis. It only depends on the threshold τ for significance. If the global null hypothesis is true, it is expected that there will be $K\tau$ statistically significant tests (genes). These are then type I errors if the global null hypothesis is true. If the number of rejected null hypothesis substantially exceeds the expected value under the global null hypothesis, it could be interpreted as evidence that some genes are differentially expressed across two groups.

From table 3.2, we can conclude that the theoretical values for the expected number of significant genes are the same for different sample sizes and correlations at the same threshold. For the different thresholds, the expected values increase as the threshold increases. For the standard deviation, it increases as the correlation increases within a threshold. For the empirical values from t-tests and randomization tests, they are very close to the theoretical values for both expected number and standard deviation of significant discoveries. For example, if the threshold τ is 0.01 and the sample size is 20, the theoretical expected number of significant genes is 10, and the empirical values from the t-test and randomization test are 9.83 and 9.78 for the uncorrelated structure, 9.54 and 9.59 for the moderate correlation, and 10.51 and 10.51 for the strong correlation, respectively. For a threshold of 0.01 and n = 20, the empirical values for standard deviation are 3.01, 7.77, and 14.53 for t-tests and 3.01, 7.69, and 14.35 for randomization tests as the correlation increases. These values are close to the theoretical values. In addition, we can see that the empirical values for the standard deviation increase within the same correlation structure as the threshold increases. When the correlation ρ is 0 and the sample size is 20, the theoretical values for standard deviation are 3, 7, and 9, respectively, for different

thresholds. The corresponding empirical values are 3.01, 6.51, and 9.67 for the t-tests and 3.01, 6.37, and 9.46 for the randomization tests. Also, we can conclude that the randomization tests produced results similar to that of t-tests based on the empirical values of the expected number and standard deviation for the number of significant discoveries.

In cases 7-12, 25% of the genes were simulated differentially expressed across the two treatments. For the genes chosen to be differentially expressed, data were generated using a mean corresponding to the effect size of 1.58 on the standard normal scale. When the alternative hypothesis was true, the empirical values were compared for the expected number and standard deviation of significant discoveries from both t-tests and randomization tests. We did not have a formula for the expected number or standard deviation of number of discoveries when the alternative hypothesis is true for some genes.

From table 3.3, we can see that the empirical values for the expected number of significant results increase as the threshold increases within a sample size and correlation structure. For example, when the sample size is 20 and correlation is 0, the expected values from the t-tests are 43.88, 122.50, and 190.96 with the increasing thresholds, and they are 43.71, 122.37, and 191.06 for the randomization tests. Also, the empirical values for the expected number are close for different sample sizes and correlations at the same threshold. For instance, at the threshold 0.01, the expected values from the t-tests are 43.88 for the uncorrelated structure, 43.15 for the moderate correlation, and 38.61 for the strong correlation, and they are 43.71, 43.05, and 38.54, respectively, from the randomization tests. In addition, the empirical values exceed the theoretical values for expected number of significant results under the global null hypothesis, which indicates that there are some genes differentially expressed between two treatment groups. For the standard deviation, it has the same pattern as the cases under the global null hypothesis. The standard deviation increases as the correlation structure becomes stronger. For example, when the sample size is 20 and the threshold is 0.01, the empirical values for standard deviation are 5.65, 17.22, and 26.62 from the t-tests and 5.60, 17.06, and 26.59 from the randomization tests as the correlation increases. Also, the standard deviation increases within the same correlation structure as the threshold increases. For instance, when the sample size is 20 with the uncorrelated structure, the empirical values for standard deviation are 5.65, 8.16, and 11.95 from the t-tests and 5.60, 8.30, and 11.77 from the randomization tests with the increasing

thresholds. In addition, from the table we can also conclude that the randomization tests produce similar results to t-tests, just like cases 1 – 6 under the global null hypothesis.

Table 3.2 Simulation Results for Cases 1-6

For E(*No*) and Sd(*No*), the numbers of the first column are the theoretical values for different thresholds. The last two columns are the empirical values from t-tests and randomization tests respectively.

Cases	SS	ES	ρ	τ	E(*No*)			Sd(*No*)		
					theor.	t-test	rand.test	theor.	t-test	rand.test
1	20	0	0	0.1	100	98.98	98.81	9	9.67	9.46
				0.05	50	49.53	49.52	7	6.51	6.37
				0.01	10	9.83	9.78	3	3.01	3.01
2	20	0	0.3	0.1	100	96.99	97.07	33	32.92	32.85
				0.05	50	48.35	48.43	23	21.69	21.79
				0.01	10	9.54	9.59	8	7.77	7.69
3	20	0	0.6	0.1	100	103.37	103.38	65	61.91	61.65
				0.05	50	52.06	52.12	45	41.59	41.35
				0.01	10	10.51	10.51	17	14.53	14.35
4	10	0	0	0.1	100	100.90	100.79	9	9.63	9.59
				0.05	50	50.43	50.37	7	7.08	7.11
				0.01	10	10.20	10.23	3	3.35	3.24
5	10	0	0.3	0.1	100	99.38	99.28	33	27.15	27.24
				0.05	50	49.12	49.07	23	17.28	17.29
				0.01	10	9.33	9.39	8	5.70	5.58
6	10	0	0.6	0.1	100	98.64	98.32	65	68.30	68.18
				0.05	50	49.52	49.46	45	48.05	48.02
				0.01	10	10.52	10.51	17	19.52	19.76

Table 3.3 Simulation Results for Cases 7-12

For E(*No*) and Sd(*No*), the first column is the empirical values from t-tests and the second is the values from randomization tests.

Cases	SS	ES	ρ	τ	E(*No*)		Sd(*No*)	
					t-test	rand.test	t-test	rand.test
7	20	1.58	0	0.1	190.96	191.06	11.95	11.77
				0.05	122.50	122.37	8.16	8.30
				0.01	43.88	43.71	5.65	5.60
8	20	1.58	0.3	0.1	189.51	189.70	36.78	36.51
				0.05	120.50	120.28	30.36	30.21
				0.01	43.15	43.05	17.22	17.06
9	20	1.58	0.6	0.1	183.42	183.36	59.77	59.93
				0.05	114.66	114.37	49.15	49.06
				0.01	38.61	38.54	26.62	26.59
10	10	1.58	0	0.1	185.63	185.53	11.60	11.68
				0.05	117.15	117.25	9.88	9.88
				0.01	38.31	38.00	5.73	5.65
11	10	1.58	0.3	0.1	185.39	185.39	38.17	38.14
				0.05	115.72	115.84	32.13	32.22
				0.01	38.63	38.29	18.24	18.11
12	10	1.58	0.6	0.1	182.83	182.78	61.24	61.21
				0.05	113.56	113.63	48.95	49.15
				0.01	36.41	36.09	26.10	25.97

In cases 13-18, 25% of the genes were simulated differentially expressed using a mean corresponding to an effect size of 3.16 on the standard normal scale. From table 3.4, we can conclude that the pattern of empirical values for the expected number is similar to the cases with an effect size of 1.58. It increases as the threshold increases within a sample size and correlation structure, and the values are very similar for different sample sizes and correlations at the same threshold. However, the magnitude of the expected values is greater than the cases with an effect size of 1.58 with the same parameters. For example, when the sample size is 20 and the correlation is 0, the expected values are 175.65 and 175.56 respectively for a t-test and randomization test at a threshold of 0.01, while they are only 43.88 and 43.71 when the effect size is 1.58. For the standard deviation, it increases as the correlation increases within the same threshold. However, as the threshold increases within the same correlation, it is not always increasing like the cases with an effect size of 1.58. Take the case with a sample size of 20 and correlation of 0.3 for an example, the values of the standard deviation are 21.77, 17.45, and 21.28 for the t-tests and 21.93, 17.56, and 21.44 for the randomization tests with increasing thresholds. This discrepancy may be due to simulation error in that only 200 simulations per case were carried out. Increasing this number may help to resolve this issue, and it will be left for future work. In addition, based on the empirical values from the t-tests and randomization tests, we can conclude that the randomization tests produce similar results to that of the t-tests like the cases described above.

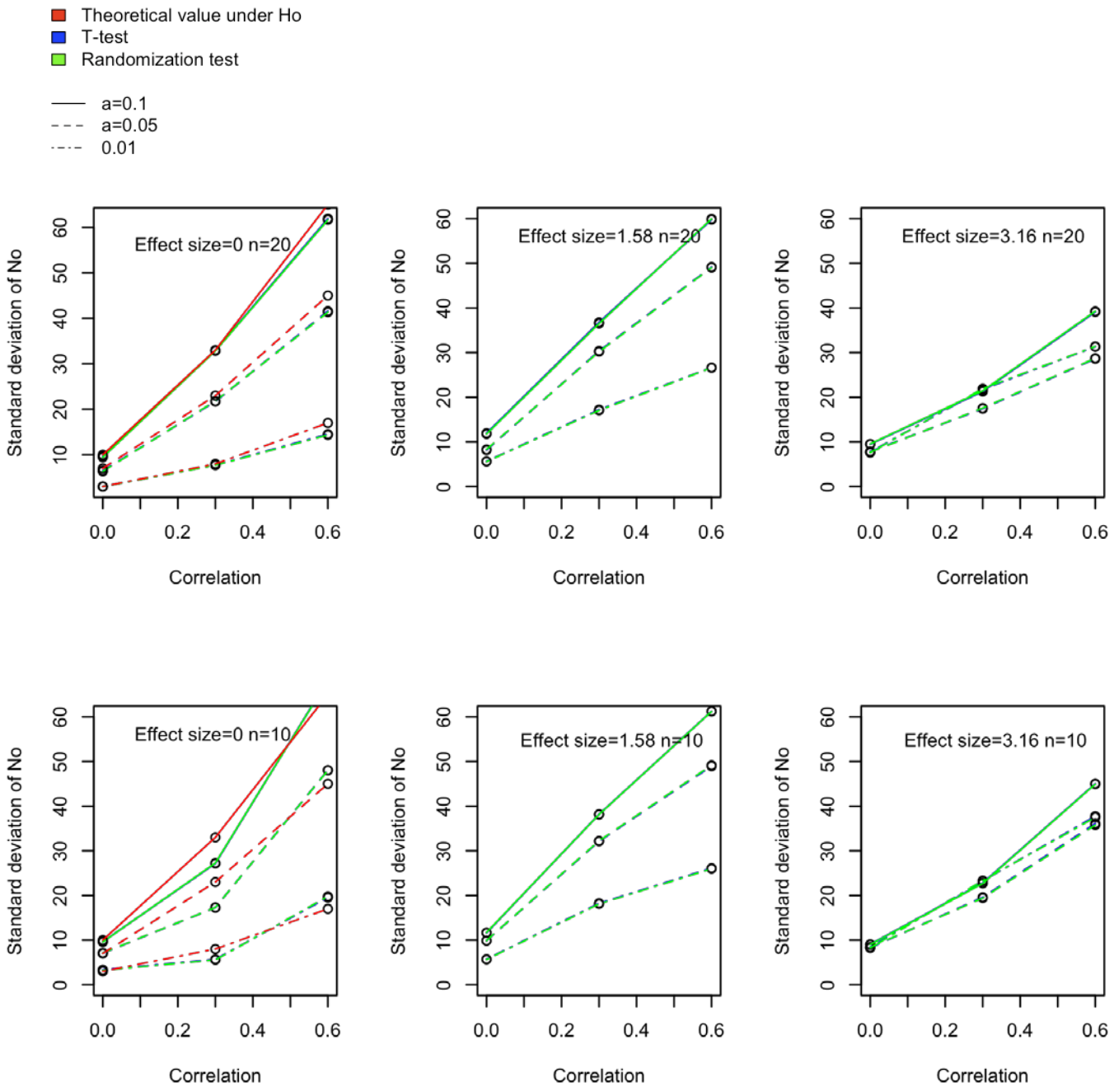
Figure 3.2 illustrates the changing pattern for the standard deviation of the number of significant results visually. The red line represents the theoretical values under the global null hypothesis. The blue represents the empirical values from t-tests and the green is the empirical values from randomization tests. From the figures, we can see that the standard deviation increases as the correlation increases. Also, it becomes greater when the threshold increases within the same correlation except cases with an effect size of 3.16. For the empirical values from the two tests, they are so similar that the blue lines and green lines overlap and exist almost as the same line.

Table 3.4 Simulation Results for Cases 13-18

For E(*No*) and Sd(*No*), the first column is the empirical values from t-tests and the second is the values from randomization tests.

Cases	SS	ES	ρ	τ	E(*No*)		Sd(*No*)	
					t-test	rand.test	t-test	rand.test
13	20	3.16	0	0.1	306.84	306.69	9.46	9.52
				0.05	255.34	255.30	7.85	7.73
				0.01	175.65	175.56	7.72	7.60
14	20	3.16	0.3	0.1	303.14	303.25	21.28	21.44
				0.05	252.73	252.71	17.45	17.56
				0.01	175.85	175.57	21.77	21.93
15	20	3.16	0.6	0.1	303.10	303.07	39.11	39.26
				0.05	253.50	253.54	28.62	28.71
				0.01	176.08	176.01	31.31	31.35
16	10	3.16	0	0.1	303.55	303.52	9.06	9.01
				0.05	247.81	247.96	8.28	8.37
				0.01	159.40	158.64	8.43	8.23
17	10	3.16	0.3	0.1	303.37	303.28	22.90	22.69
				0.05	247.59	247.50	19.59	19.41
				0.01	159.04	157.78	23.25	23.31
18	10	3.16	0.6	0.1	308.03	308.12	45.02	44.94
				0.05	249.29	249.11	36.10	35.78
				0.01	157.23	156.32	37.76	37.50

Figure 3.2 Comparison of Standard Deviation of the Number of Discoveries

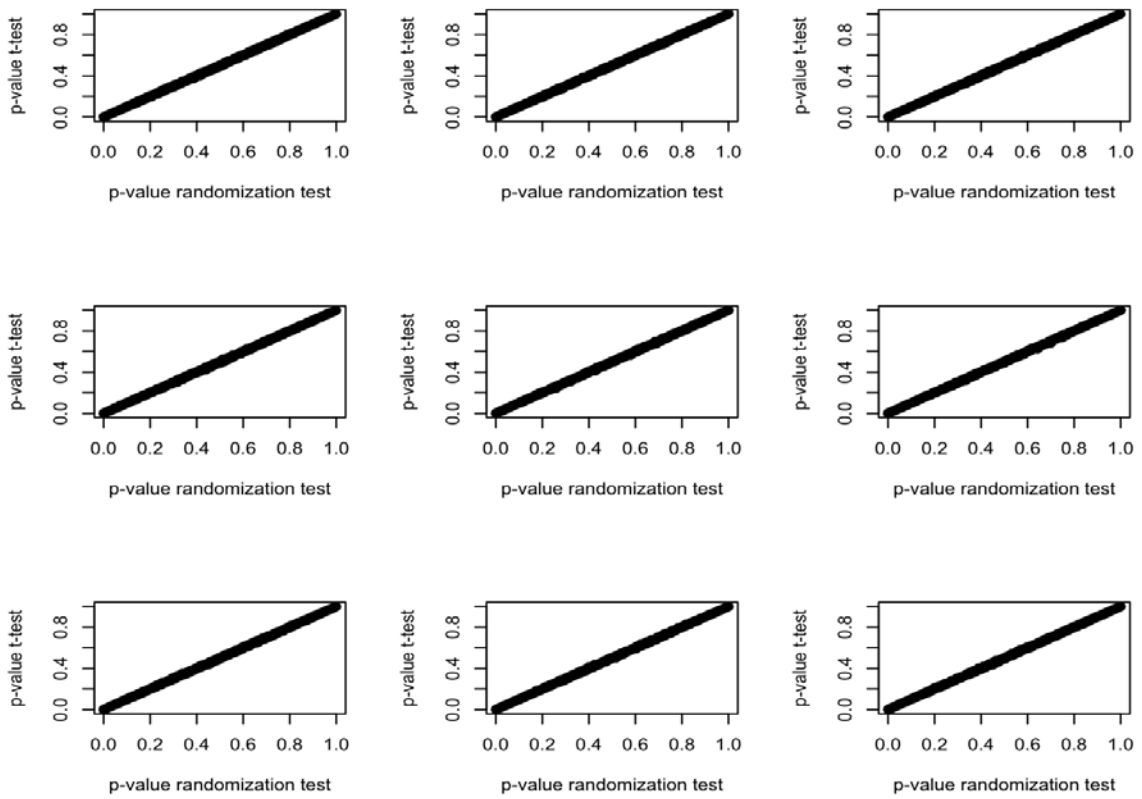


3.3.2 Comparison of P-values between t-tests and randomization tests

To study the relationship of P-values between the two tests, we compare the P-values from both t-tests and randomization tests. Figures 3.3 to 3.6 illustrate the relationship of P-values between these two tests. The horizontal axis represents the P-values from the randomization tests and the vertical represents those from the t-tests.

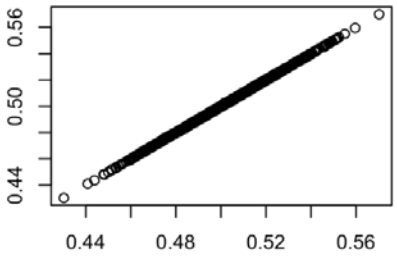
Figure 3.3 is plotted for the P-values of data sets individually in case 1 (SS=20, ES=0, ρ=0). There are 9 vectors of P-values which are randomly selected from a total of 200 vectors. From these figures, we can conclude that the P-values from t-tests and randomization tests are highly correlated. One thing is important here, we have to make sure that the P-values from a t-test and randomization test are calculated for the same data set. In figure 3.3, all points are almost on the diagonal line meaning that the P-values are very similar from the two tests. For other cases, the plots of P-values have the same pattern as that of case 1 so we do not need to describe all of the figures. Thus, the P-values from the t-tests and randomization tests are highly correlated, regardless of sample size and/or correlation.

Figure 3.3 Comparison of P-values for 9 Data Sets in cases 1

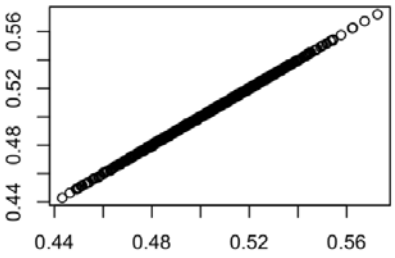


Figures 3.4 to 3.6 are plotted as the mean of 200 vectors of P-values for the cases with a sample size of 20. From these figures, we can see that the results of the t-tests and randomization tests are highly correlated. For the cases 1-3, which are under the global null hypothesis, the range of the mean of P-values is narrower than that of the individual P-values since the variance of a mean is smaller than that of an individual P-value. For the cases under the alternative hypothesis, there is a gap in the mean of P-values between differentially and non-differentially expressed genes because of the shrinkage of variance. For plots of other cases with a sample size of 10, the pattern is the same as the cases with a sample size of 20 so we do not display all of the figures.

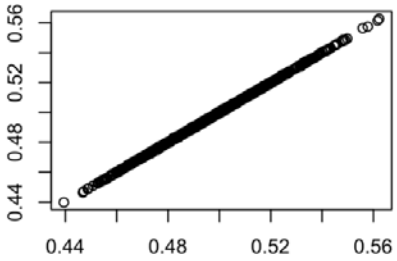
Figure 3.4 Comparison the Mean of P-values for Cases 1-3



Case 1 (ρ =0)

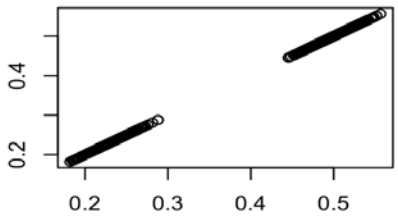


Case 2 (ρ =0.3)

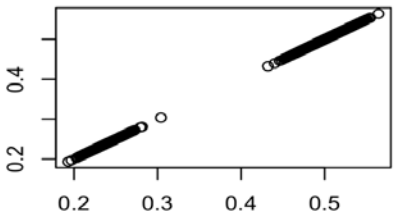


Case 3 (ρ =0.6)

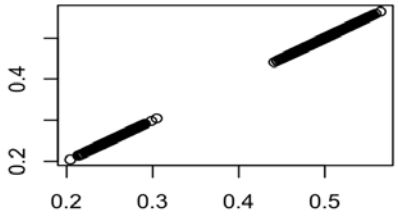
Figure 3.5 Comparison the Mean of P-values for Cases 7-9



Case 7 (ρ =0)

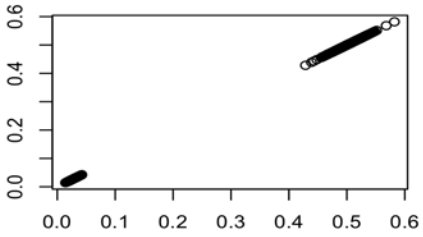


Case 8 (ρ =0.3)

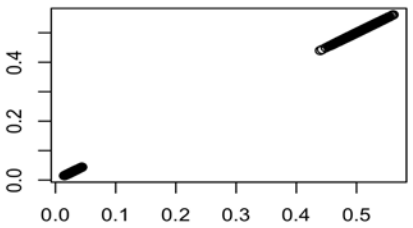


Case 9 (ρ =0.6)

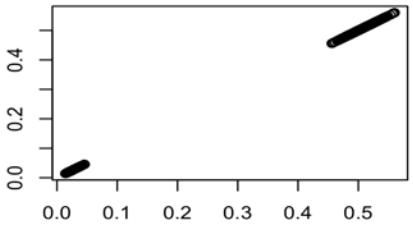
Figure 3.6 Comparison the Mean of P-values for Cases 13-15



Case 13 (ρ =0)



Case 14 (ρ =0.3)



Case 15 (ρ =0.6)

3.3.3 Distribution of P-values

As mentioned in the literature review, when the global null hypothesis is true, the valid P-values from multiple tests were expected to have a uniform distribution for the uncorrelated HDD. In our study, cases 1 and 4 were simulated for the uncorrelated structure ($\rho = 0$) and under the global null hypothesis. The two figures on the first row in figure 3.7 and 3.8 are plotted for cases 1 and 4, respectively. One is plotted for the P-values from the t-tests and the other is for the randomization tests. Shown are 4 histograms of P-values for each case which were randomly selected from a total of 200 vectors. Also, the P-values from the two tests were calculated from the same data set. From these figures, we can see that the histogram of P-values is approximately uniform on interval 0 and 1 for both cases. However, when the HDD are correlated, the distribution of P-values may vary widely from experiment to experiment. The other figures describe this pattern visually for the cases which were simulated as correlated HDD and under the global null hypothesis. They are not a uniform distribution. In some occasions, there are fewer P-values clustering near 0 than the expected under the global null hypothesis, which is difficult to interpret. In other occasions, there are many more P-values clustering near 0, which is a sign that some genes may be differentially expressed, even though the global null hypothesis is true. So, the distribution of P-values from multiple tests can give misleading results when HDD are correlated. However, the distribution of P-values is plausible for moderate correlation structure. In addition, we can conclude that the pattern of histogram of P-values is very similar between the t-tests and randomization tests for all cases under the global null hypothesis.

When there are some genes differentially expressed across two treatments, the distribution of P-values shows a clustering near 0. Figures 3.9 to 3.12 are plotted for the cases under the alternative hypothesis. Figures 3.9 and 3.10 display the distribution of P-values for the cases with an effect size of 1.58, and figures 3.11 and 3.12 are plotted for the cases with an effect size of 3.16. From these figures, we can see that the P-values are more likely to cluster near 0 for the uncorrelated data when the alternative is true. For the correlated HDD, the P-values also show a pattern of clustering near 0. However, the signal is stronger for the effect size of 3.16 compared to 1.58. In addition, we can see that the distribution of P-values is very similar between the sample size of 10 and 20 with the same parameters. What is more important, it can be concluded that the histograms of P-values are also similar between the t-tests and randomization tests for the cases under the alternative hypothesis.

Figure 3.7 Histograms of P-values for Cases 1-3 (SS=20 ES=0)

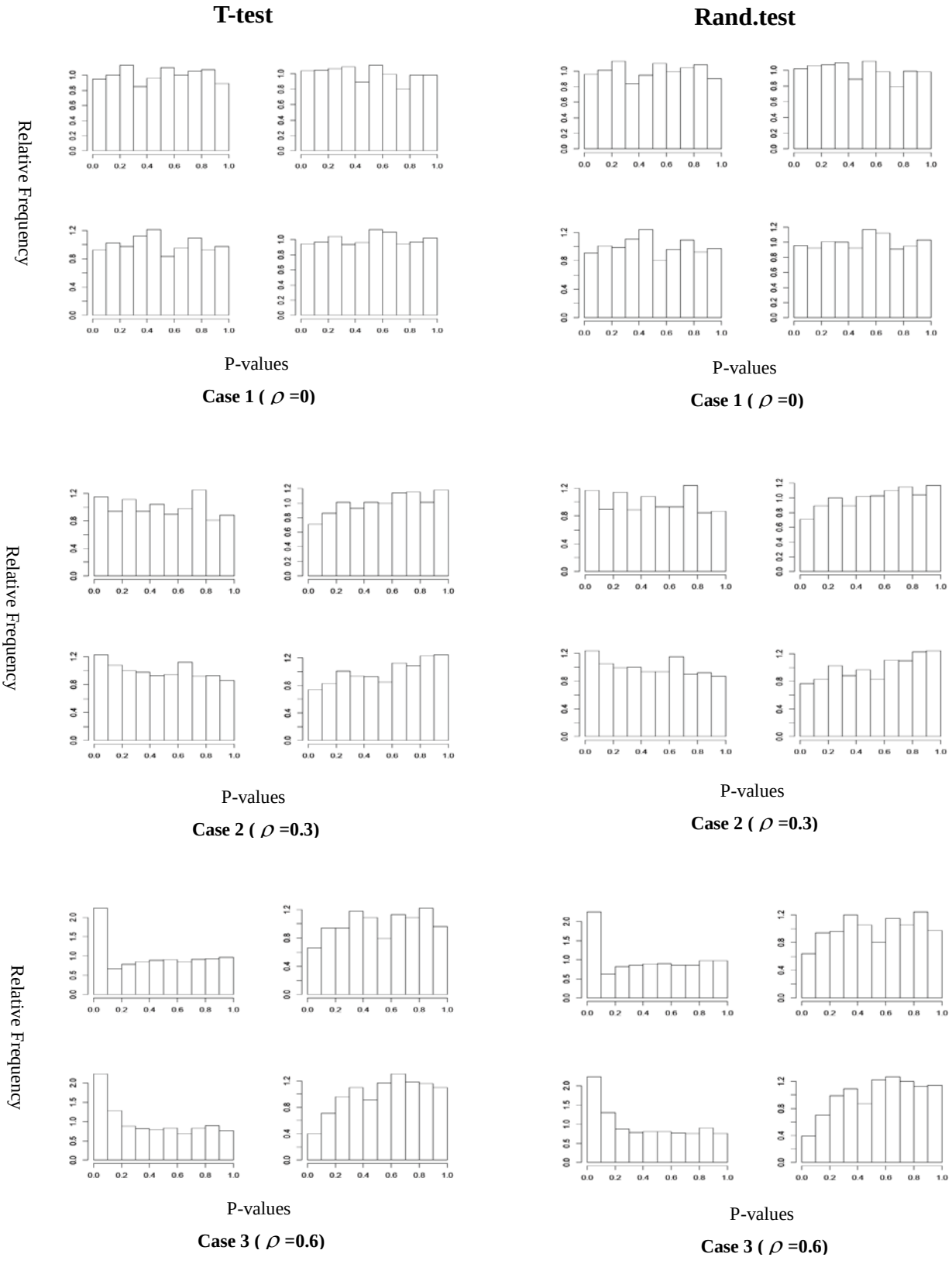


Figure 3.8 Histograms of P-values for Cases 4-6 (SS=10 ES=0)

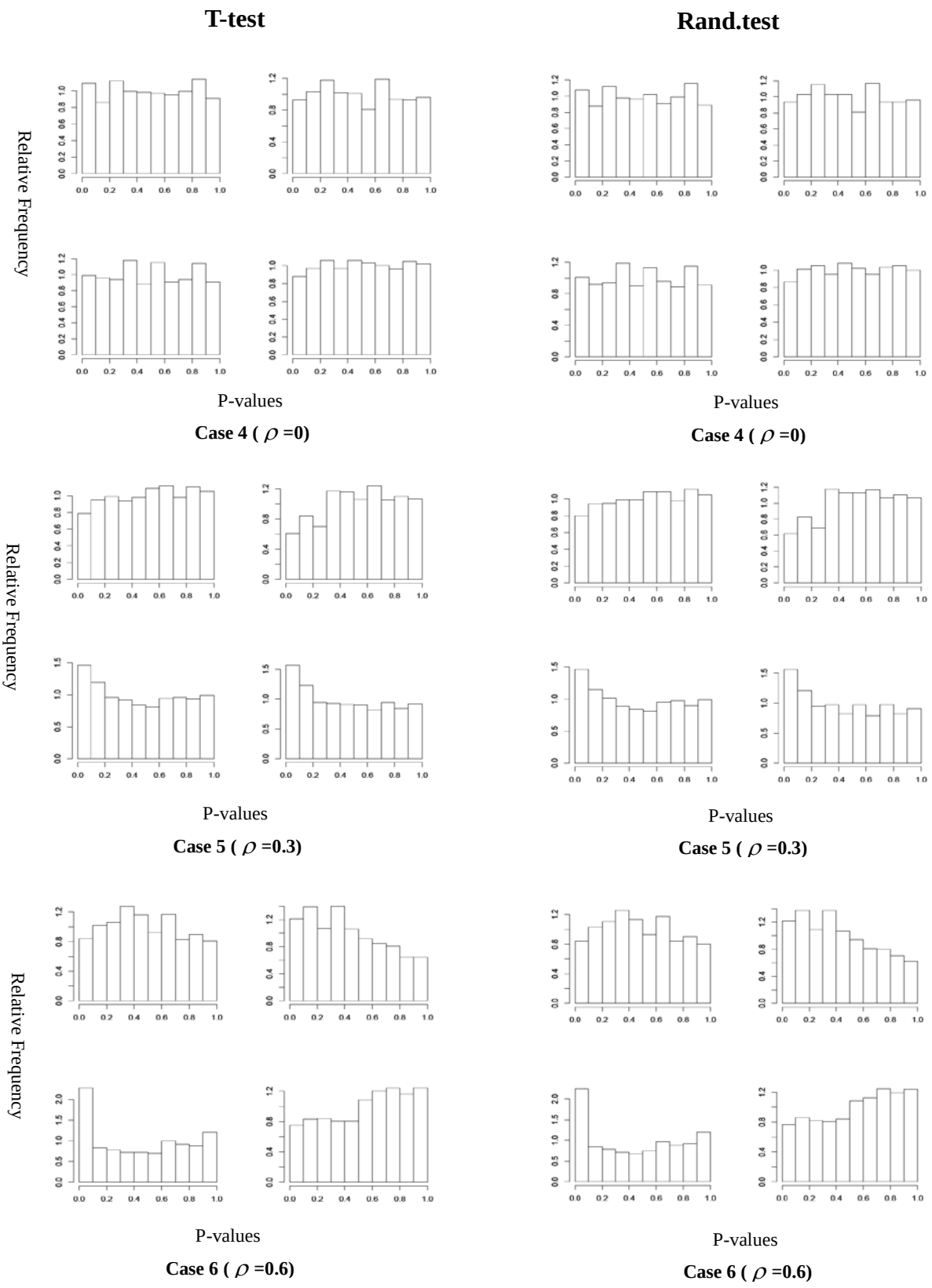


Figure 3.9 Histograms of P-values for Cases 7-9 (SS=20 ES=1.58)

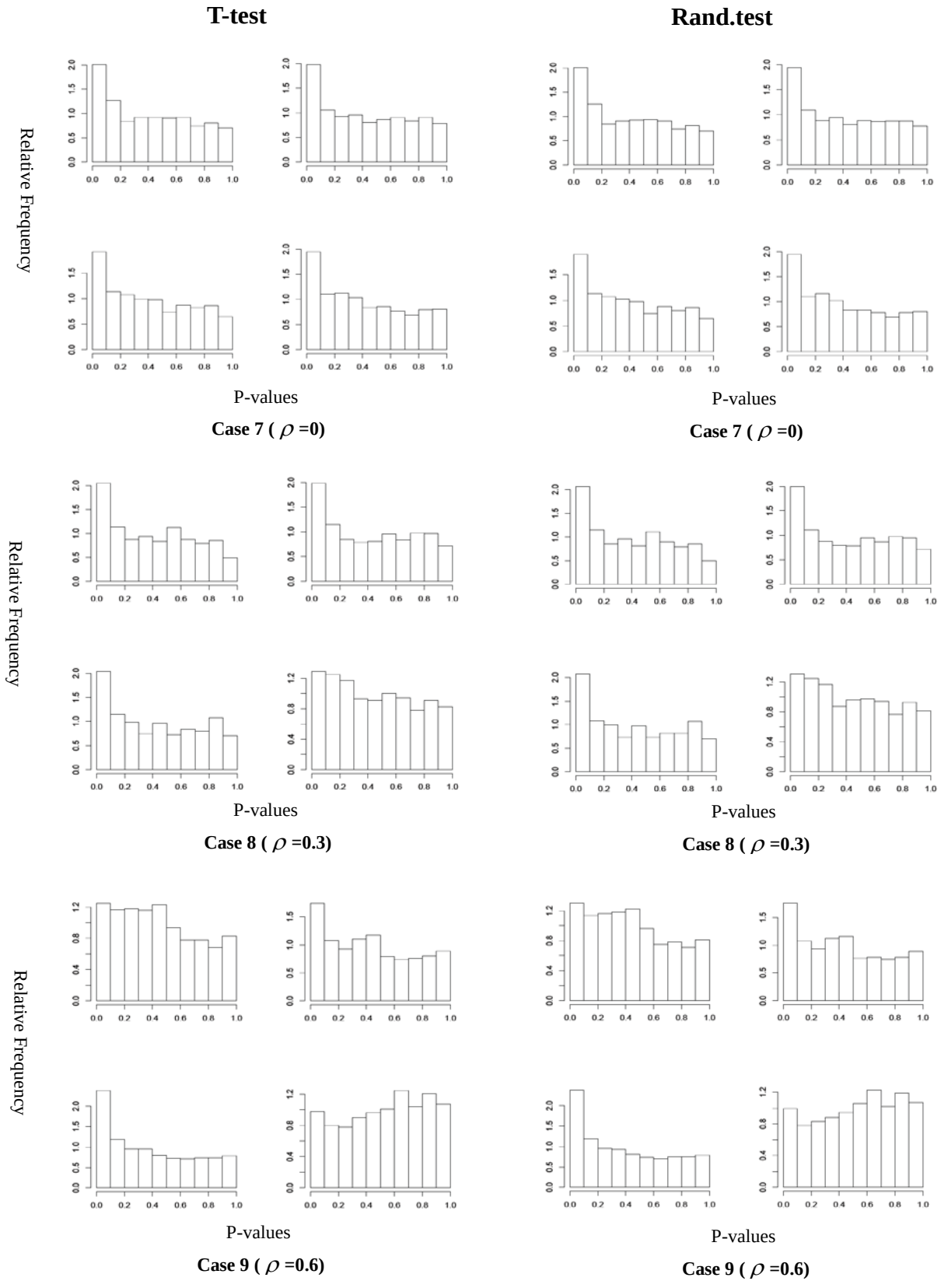


Figure 3.10 Histograms of P-values for Cases 10-12 (SS=10 ES=1.58)

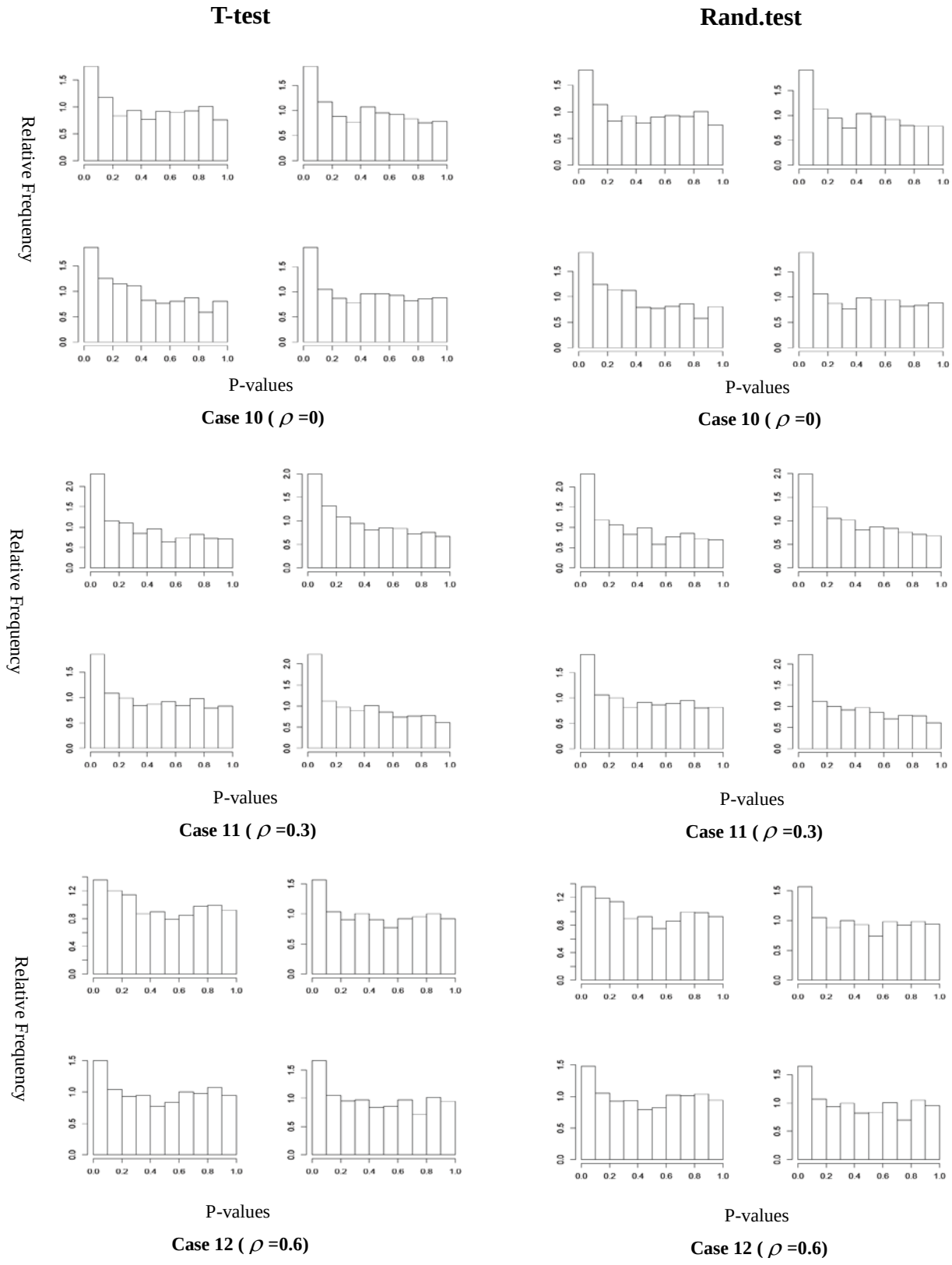


Figure 3.11 Histograms of P-values for Cases 13-15 (SS=20 ES=3.16)

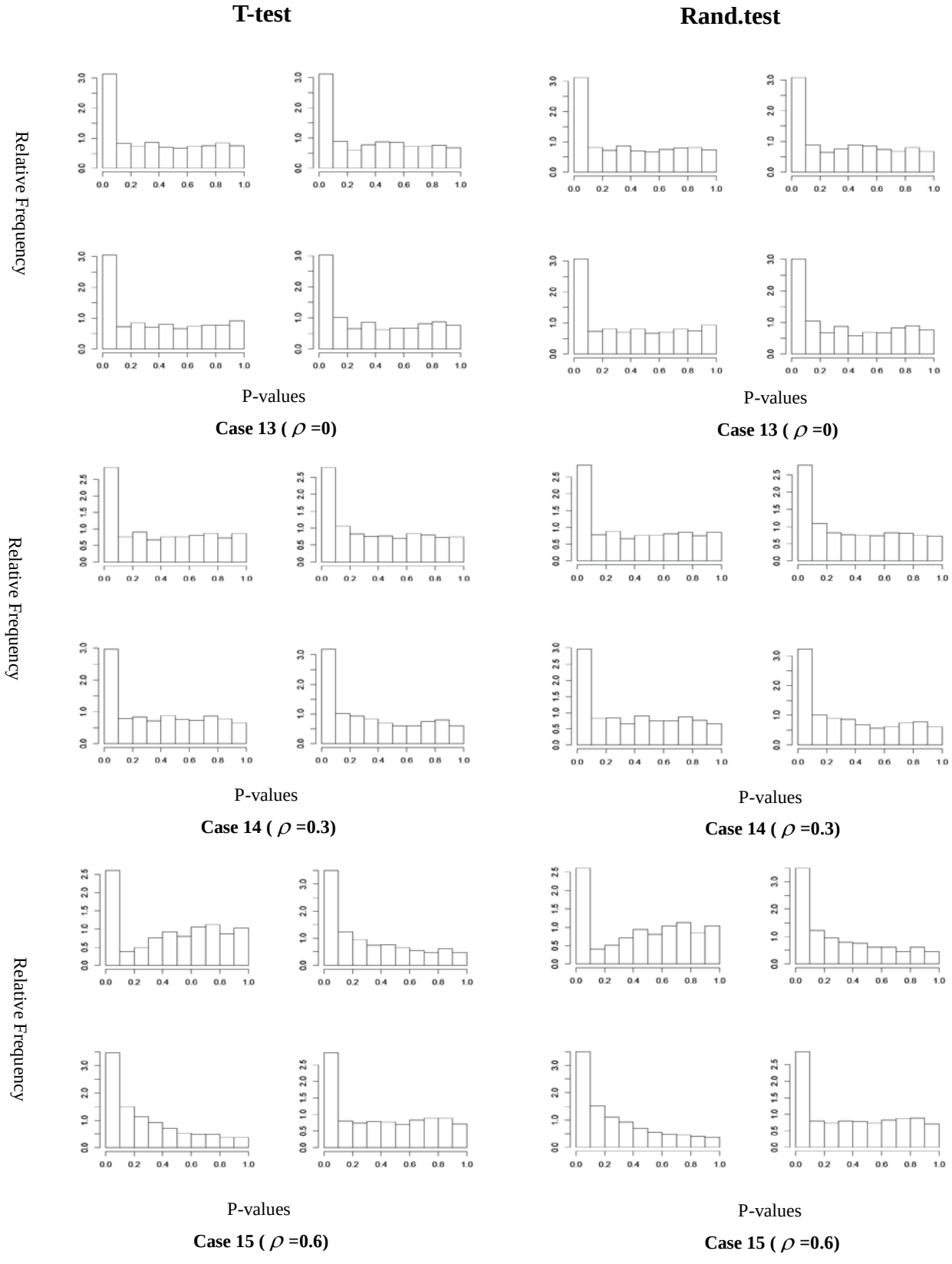
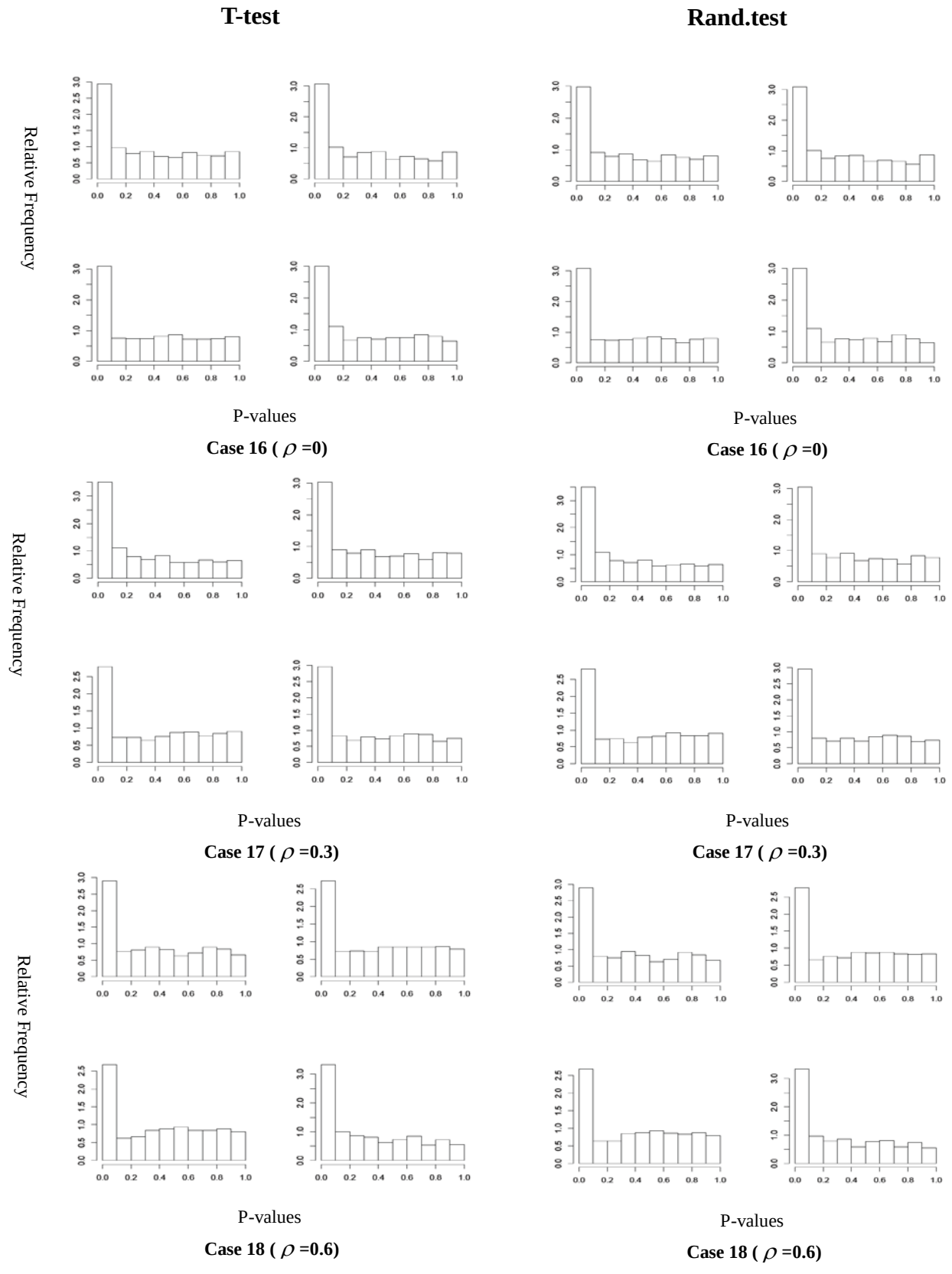


Figure 3.12 Histograms of P-values for Cases 16-18 (SS=10 ES=3.16)



Chapter 4 - A Real Case Study

In this report, a real microarray dataset was used to illustrate the results comparing the randomization tests with t-tests.

4.1 Multiple Myeloma and Bone Lesions Dataset

A microarray dataset in multiple myeloma patients with and without bone lesions was introduced in Tian et al (2003). There were 173 individuals in total. 137 subjects were with bone lytic lesions and 36 of them were without bone lytic lesions. To obtain the data at the gene expression level, U95Av2 microarrays were used for hybridization , and MAS software, version 5.01, was used to quantify the intensity values (referred to as signals). After initial filtration, the number of probe sets was reduced to 3970.

4.2 Results for Multiple Myeloma Dataset

In this microarray study, of interest was to detect the genes differentially expressed across the two treatment conditions. The number of false positive results among the rejected null hypotheses need to be controlled. As mentioned in the literature review, FDR controls the expected proportion of false discoveries among the rejected null hypotheses.

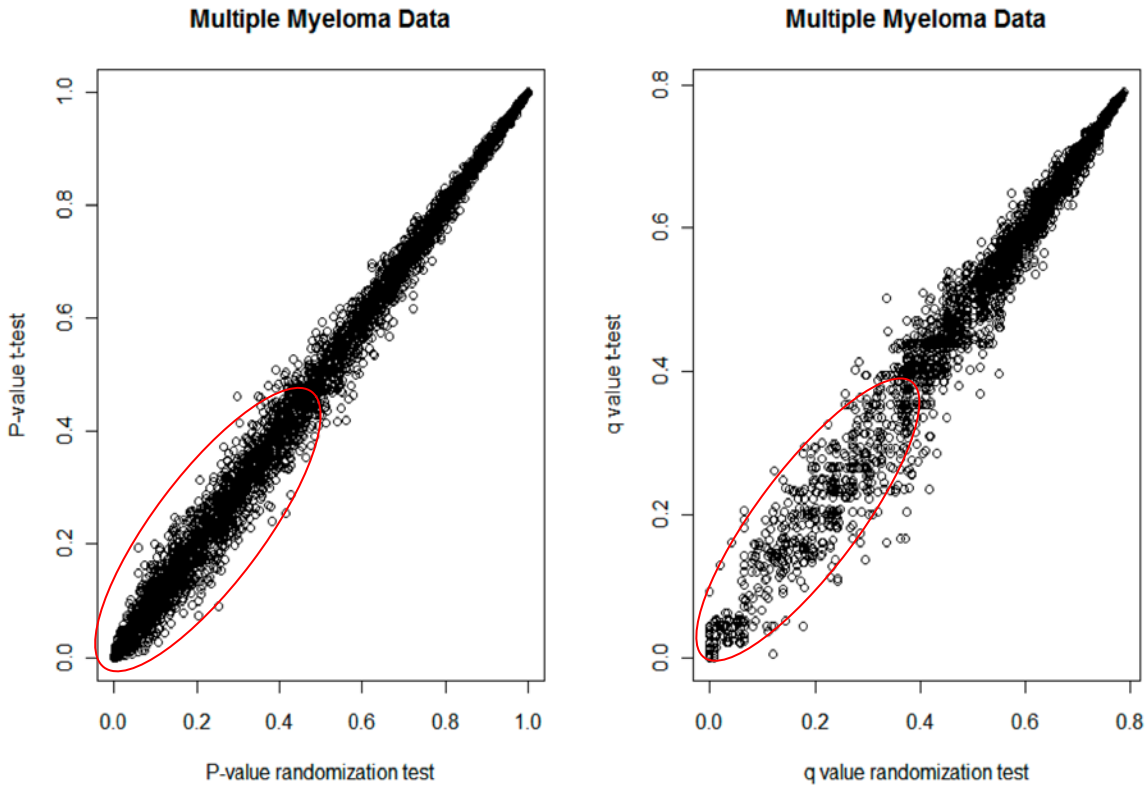
For the real case study, the quantities of interests and analysis are similar to that of the simulation study. First, the P-values were calculated for each gene by a randomization test and t-test, respectively. Then, based on the P-values, we computed the q-values. Finally, we assessed the performance of a randomization test compared with a t-test.

4.2.1 Comparison of P-values and q-values between the t-tests and randomization tests

To study the relationship of P-values and q-values between the t-tests and randomization tests, the results from these two tests were compared. Figure 4.1 displays this comparison for the P-values and q-values, respectively. The left figure is plotted for the P-values and the right is for the q-values. The horizontal axis represents the results of randomization tests and the vertical represents those of t-tests. From these figures, we can conclude that the results are highly correlated between a randomization test and t-test. The correlation is 0.995 for P-values and 0.894 for q-values, respectively, between these two tests. In addition, we can see that the number

of points in the circled area is less for the plot of q-values than that of P-values since a q-value corresponds to a P-value that has been adjusted for FDR control. For example, 5% FDR means that among the significant discoveries, 5% are true nulls.

Figure 4.1 Comparisons for P-values and q-values



4.2.2 Distribution of P-values and q-values

Figures 4.2 and 4.3 are plotted for the histograms of P-values and q-values from both t-test and randomization test. From these figures, we can conclude that the histogram patterns are very similar between these two tests for this case study. Also, based on the cumulative number of significant calls (table 4.1) derived from the two tests, we can see that the values are close at the different thresholds. In addition, from the histograms of P-values, we can see that the P-values are more likely to cluster near 0, which is a sign that some genes may be differentially expressed between two treatments. Based on the study of correlation densities in Paranagama (2011), the departure of the correlation distribution from independence is subtle for this data set. Thus, we

can make a conclusion that there are some genes differentially expressed across two disease state sets.

Table 4.1 Cumulative Number of Significant Calls

	P-values			q-values		
	<0.01	<0.05	<0.1	<0.01	<0.05	<0.1
T-test	203	514	773	10	79	119
Randomization test	201	497	770	35	72	131

Figure 4.2 Histograms of P-values

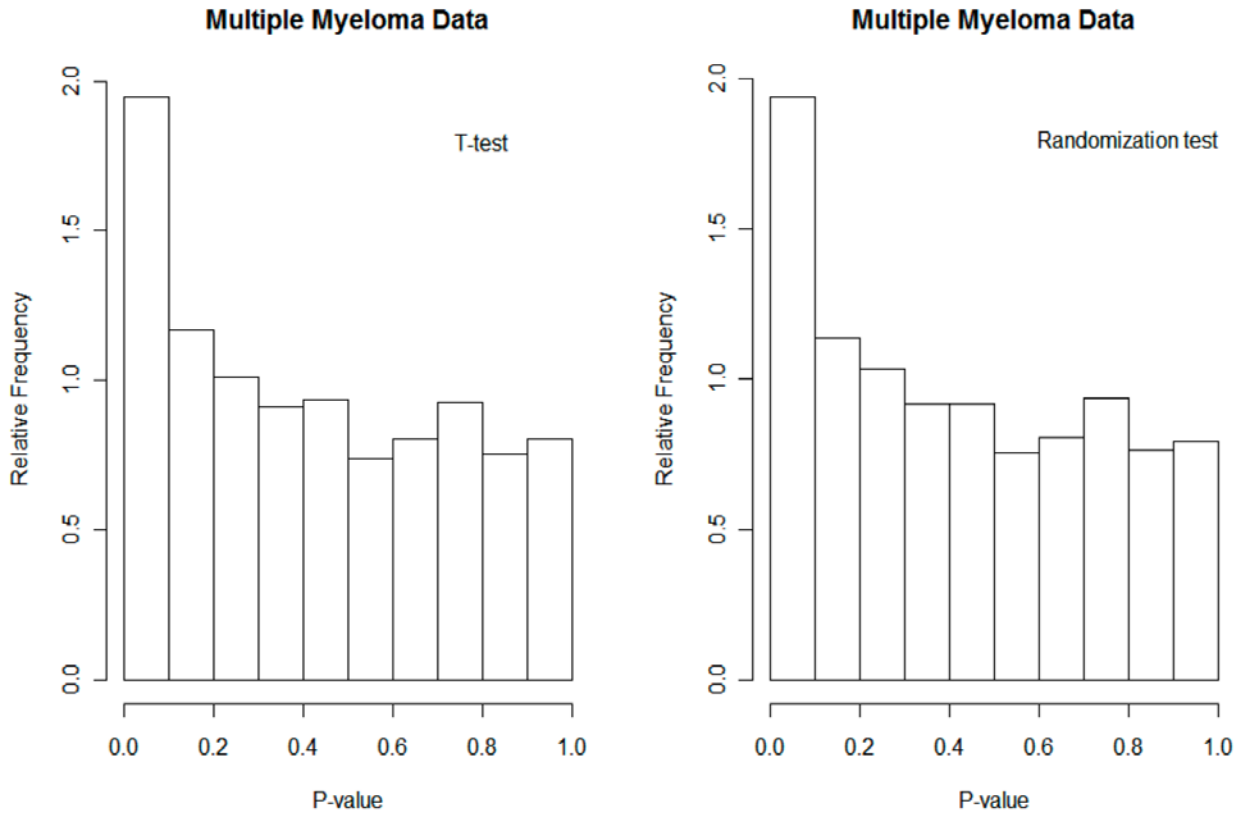
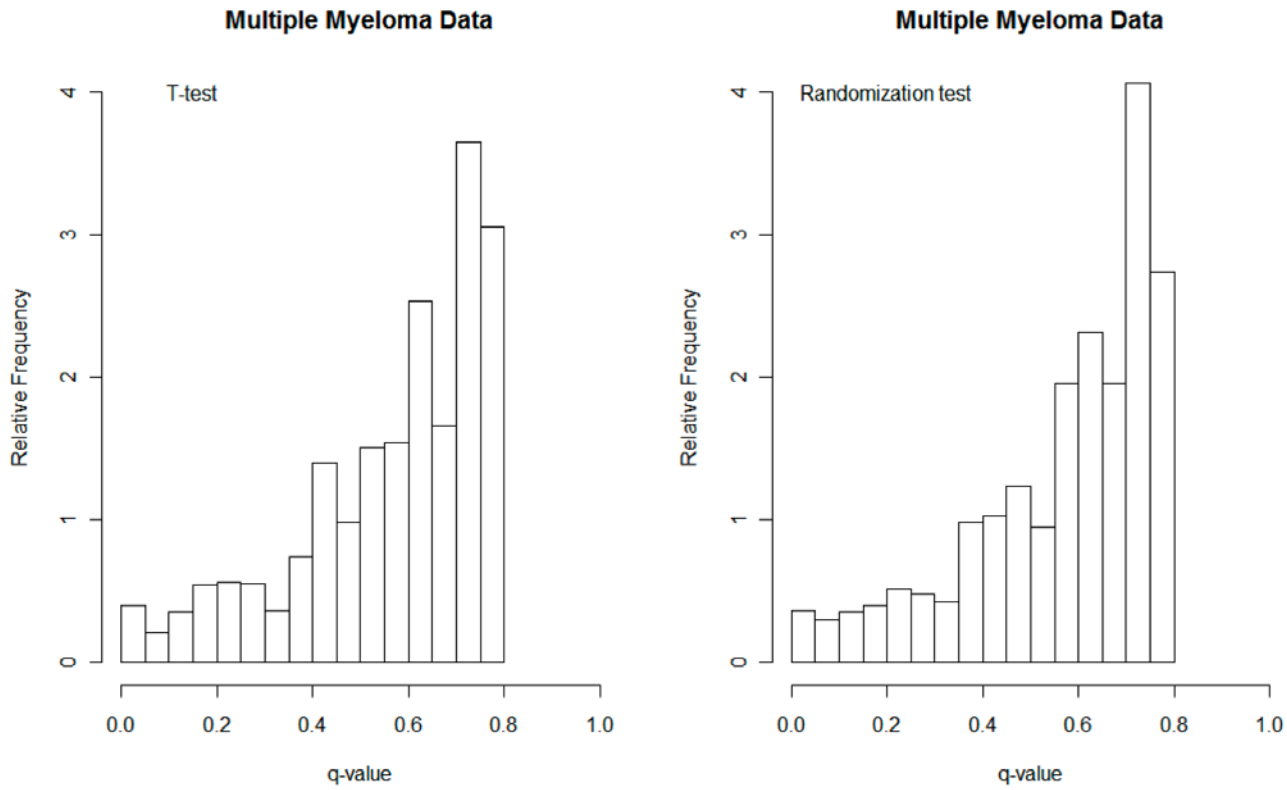


Figure 4.3 Histograms of q-values



Chapter 5 - Concluding Remarks and Future Work

In this report, we focused on the comparison of randomization tests with usual t-tests to assess the performance of randomization tests in the presence of non-estimable dependence structure. The dependence structure cannot be estimated since the number of features is very large relative to the available samples. For a randomization setting, it is column-wise permutation across treatment conditions which helps preserve the correlation structure between genes. In addition, the results from simulations and a real case study showed that the randomization tests produced similar results to that of t-tests. For uncorrelated data, the P-values were valid since the correct reference distribution was used for the test statistics, and the distribution of them was meaningful to analyze the HDD. However, when HDD are not independent, the results may be misleading when illustrating biological results using the distribution of the P-values. So, we can make a conclusion that randomization tests are no direct assistance with the correlation effects in HDD.

For a correlation study, it is unrealistic to estimate a (K×K)-dimensional dependence structure from observed data on available sample units. However, we can estimate the correlation between any pair of genes to determine the presence of apparent structure and the extent of a departure from an independence structure. So, for future work, we can study the empirical pairwise-correlations between genes to evaluate the performance of randomization tests in the presence of a dependence structure.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., & Prolla, T. A., et al. (2002). A mixture model approach for the analysis of microarray gene expression data. Computational Statistics & Data Analysis, 39, 1-20.

Allison, D. B. (2006). DNA microarrays and related genomics techniques : designs, analysis, and interpretation of experiments. Chapman & Hall/CRC, Boca Raton.

Barry, W., Nobel, A., &Wright, F. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics, 21, 1943-1949.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the Flase Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological, 57, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 29, 1165-1188.

Bonear, C. (1960). The effects of violations of assumptions underlying the t-test. Psychological Bulletin, 57, 49-64

Cao, H., & Kosorok, M. (2011) Simultaneous critical values for t-tests in very high dimensions. Bernoulli, 17, 347-394.

Dudoit, S., Yang, Y., Callow, M., & Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12, 111-139.

Edgell, S., & Noon, S. (1984). Effect of violation of normality on the t-test of the correlation-coefficient. Psychological Bulletin, 95, 576-583.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association, 102, 93-103.

Efron, B. (2007). Large-scale inference : empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, Cambridge.

Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. Annals of Applied Statistics, 1, 107-129.

Gadbury, G. L., Page, G. P., Heo, M., Mountz, J. D., & Allison, D. B. (2003). Randomization tests for small samples: an application for genetic expression data. Journal of the Royal Statistical Society Series C-Applied Statistics, 52, 365-376.

Gadbury, G. L., Xiang, Q., Yang, L., Barnes, S., Page, G. P., & Allison, D. B. (2008). Evaluating Statistical Methods Using Plasmode Data Sets in the Age of Massive Public Databases: An Illustration Using False Discovery Rates. Plos Genetics.

Göhlmann, H., & Talloen, W. (2009). Gene expression studies using affymetrix microarrays. In: Mathematical and computational biology series. Taylor & Francis.

Hall, P., & Tajvidi, N. (2002) Permutation tests for equality of distributions in high-dimensional settings. Biometrika, 89, 359-374.

Hu, X., Gadbury, G. L., Xiang, Q., & Allison, D. B. (2010). Illustrations on using the distribution of a p-value in high dimensional data analysis. Advances and Applications in Statistical Sciences, 191-213.

Lee, C., Klopp, R., Weindruch, R., & Prolla, T. (1999). Gene expression profile of aging and its retardation by caloric restriction. Science, 285,1390-1393.

Mehta, T., Zakharkin, S., Gadbury, G., & Allison, D. (2006). Epistemological issues in omics and high-dimensional biology: give the people what they want. Physiological Genomics, 28, 24-32.

Paranagama, D. C. (2011). Correlation and variance stabilization in the two group comparison case in high dimensional data under dependencies. PhD. Dissertation, Kansas State University.

Pawitan, Y., Calza, S., & Ploner, A. (2006). Estimation of false discovery proportion under general dependence. Bioinformatics, 22, 3025-3031.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics, 21, 3017-3024.

Qiu, X., Klebanov, L., & Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. Statistical Applications in Genetics and Molecular Biology.

Schweder, T., & Spjøtvoll , E. (1982). Plots of P-values to evaluate many tests simultaneously. Biometrica, 69, 493-502.

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B-Statistical Methodology. 64, 479-498.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., et al. (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. New England Journal of Medicine, 349, 2483-2494.

Tsai, C., Chen, Y., & Chen, J. (2003). Testing for differentially expressed genes with microarray data. Nucleic Acids Research 31(9), e52.

West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. Bayesian Statistics 7 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford Univ. Press. MR2003537, 733–742.

Appendix A - R Programs

```
##############################################################################
##
##    Function to generate correlated HDD employing basic block diagonal covariance matrices.
##
##    Input:
##       m -- the number of rows (genes), n -- the number of columns (sample size),
##       rol -- the correlation coefficient, b -- the block size, u -- the value to form the mean vector.
##
##     Output:
##       a matrix (m×n) with a correlation of rol within the block.
##
##############################################################################

 normdata<-function(m,n,rol,b,u){

 R1<-matrix(rol,b,b)
 R2<-diag(1-rol,b)
 R<-R1+R2
 cov<-R
 X<-rep(0,n)

 for(i in 1:(m/b)){
 x<-rmvnorm(n=n,mean=c(rep(u,50),rep(0,b-50)),sigma=cov)
 X<-cbind(X,x)
 }

 X<-X[,-1]
 y<-t(X)
 y<<-y

}
```

```
##############################################################################
##
##   This code computes P-values for an input data set using t-tests.
##
##   Input:
##       data -- a data matrix with k rows and 2n columns.
##       k is the number of genes. It is assumed there are
##       two groups with n cases in each group.
##
##   Output: pval-- a permanent object.
##
##############################################################################

Pval.code<-function(data){

  n<-dim(data)[1]
  num<-dim(data)[2]/2
  sum1<-0
  sum2<-0
  sum1.2<-0
  sum2.2<-0

  for(i in 1:num){
  sum1<-data[,i]+sum1
  sum2<-data[,(num+i)]+sum2
  sum1.2<-sum1.2+data[,i]^2
  sum2.2<-sum2.2+data[,(num+i)]^2
  }

  mu1<-sum1/num
  mu2<-sum2/num
  var1<-(sum1.2-num*mu1^2)/(num-1)
  var2<-(sum2.2-num*mu2^2)/(num-1)
  sd.cr<-sqrt((var1+var2)/num)
  #sd.cr1<<-sd.cr
  tstat<-(abs(mu1-mu2)/sd.cr)
  #print(mean(tstat))
  #using pooled degrees of freedom here
  pval<-(1-pt(tstat,(2*num-2)))*2
  pval<<-pval

  }
```

```
##############################################################################
##
##   This code does an approximate randomization test for two sample data.
##
##   Input:
##        data -- a data matrix with k rows (genes) and 2n columns (samples). The first n=N/2
##        columns are one treatment group, and the second n=N/2 columnes is the
##        second treatment group.
##        its -- the number of iterations desired.
##
##    Output:
##         pval-- a permanent object.
##
##############################################################################

Pval.rand.approx<-function(data,its){
pval<-0
k<-dim(data)[1]
N<-dim(data)[2]
n<-N/2
R.Test<-rep(0,k)

sum1<-0
sum2<-0

for(i in 1:n){
sum1<-data[,i]+sum1
sum2<-data[,(n+i)]+sum2
}

mu1<-sum1/n
mu2<-sum2/n
d.obs<-mu1-mu2
d.obs<-round(d.obs,4)
d.obs<<-d.obs

for(j in 1:its){
tt<-sample(c(rep(1,n),rep(0,n)))
x<-data[,tt==1]
y<-data[,tt==0]
sum1a<-0
sum2a<-0

for(i in 1:n){
sum1a<-x[,i]+sum1a
sum2a<-y[,i]+sum2a
```

```
}

mu1a<-sum1a/n
mu2a<-sum2a/n
d.rand<-mu1a-mu2a
d.rand<-round(d.rand,4)
test<-rep(0,k)
test[abs(d.rand)>=abs(d.obs)]<-1
R.Test<-R.Test+test
}

pval<-R.Test/its
pval<<-pval

}
```

```
##############################################################################
##
##   Function computing the theoretical values for the standard deviation of the
##   number of significant discoveries.
##
##   Input:
##        r -- correlation coefficient within a block.
##        a -- thresholds.
##        k -- block size.
##        m -- number of blocks.
##
##   Output:
##         matrix with columns the threshold, the variance for the number of
##         significant tests in one block of size k, and the standard deviation
##         for the number of significant discoveries out of k*m genes.
##
##############################################################################

library(mvtnorm)

testvar=function(r,a=c(.2,.15,.1,.05,.01,.001,.0001),k,m){

 n=length(a)
 res=rep(0,n)

 for(i in 1:n){
    rho=cbind(c(1,r),c(r,1))
    ai=a[i]
    t=qnorm(1-ai/2)
    print(c(ai,t))
    #compute covariance assuming equal correlations among all pairs
    p=pmvnorm(upper=c(t,t),corr=rho)-pmvnorm(upper=c(t,-t),corr=rho)-pmvnorm(upper=c(-
t,t),corr=rho)+pmvnorm(upper=c(-t,-t),corr=rho)
    v1=k*ai*(1-ai)
    v2=k*(k-1)*(p[[1]]-(1-ai)^2) #variance of number of rejection
    v=v1+v2
    res[i]=v
 }

 sd=sqrt(m*res)
 return(cbind(a,res,sd))
}
```

```
##############################################################################
##
##   This function is to plot the p-values (t-test vs rand.test) for 1000
##   genes in individual data set. There are 9 vectors of P-values which are randomly
##   selected from a total of 200 simulations.
##
##   Input:
##        p.t -- p-values from t-tests.
##        p.rand -- p-values from randomization tests for the same data set as t-tests.
##
##   Output:
##         produce figure 3.2.
##
##############################################################################

image.plot<-function(p.t,p.rand){

 par(mfrow=c(3,3))

 for (i in 1:9){
 j=sample(1:200,1)
   x=p.rand[,j]
 y=p.t[,j]
 plot(x,y,xlim=c(0,1),ylim=c(0,1),xlab="p-value randomization test",ylab="p-value t-test")
 }

}
```

```
##############################################################################
##
##    This function is to plot the histogram of p-values for 1000
##    genes in individual data set. There are 4 vectors of P-values which are randomly
##    selected from a total of 200 simulations.
##
##    Input:
##        p.vector -- a vector of P-values for 1000 genes.
##
##    Output:
##        a histogram of P-values.
##
##############################################################################

hist.plot<-function(p.vector){

 par(mfrow=c(2,2))

 for (i in 1:4){
    j=sample(1:200,1)
    x=p.matrix[,j]
    hist(x,freq=FALSE,main="",xlim=c(0,1),xlab="",ylab="")
 }

}
```

```
#############################################################################
##
##    Function for performing row-wise t-tests for a given matrix
##    (works for unequal sample size). Grouping variable must be provided
##    separately and is used to group columns in the data matrix.
##
##    Input:
##         data1 -- subjects in columns and variables in rows.
##         groups -- grouping variable used to group subjects.
##
##    Output:
##         matrix with columns t-statistic values, p-values and degrees of freedom.
##
#############################################################################

rowttest<-function(data1, groups) {

dottest<-function(d, g) {
x<-d[g==unique(g)[1]]
y<-d[g==unique(g)[2]]
t<-t.test(x, y, alternative="two.sided")
c(t.stat=t$statistic, p.val=t$p.value, df=t$parameter)
}

t(apply(X=data1, MARGIN=1, FUN=dottest, g=groups))
}
```

```
#############################################################################
##
##   Function for computing an approximate p-value for a randomization test
##   for two treatments.
##
##   Input:
##        X -- a vector of responses of units to a control treatment.
##        Y -- a vector of responses of units to a test treatment.
##        its -- the number of iterations desired.
##
##   Output:
##        a list object, say r,  with D and P.
##        d is the observed treatment different.
##        p is  the approximated p-value for randomization test.
##
#############################################################################

random.approx<-function(X,Y,its){

 XY<-c(X,Y)
 N<-length(XY)
 n1=length(X)
 d<-mean(Y)-mean(X)

 D.null<-0

 for (i in 1:its){
    treat<-sample(1:N,n1)
    d.null<-mean(XY[treat])-mean(XY[-treat])
    D.null[i]<-d.null
 }

 d<-d
 D.null<-D.null
 pval<-sum(abs(D.null)>=abs(d))/its

r=list(D=d,P=pval)
r

}
```

```
##############################################################################
##
##    Function for performing row-wise randomization tests for a given matrix
##    (works for unequal sample size). Grouping variable must be provided separately
##    and is used to group columns in the data matrix.
##
##    Input:
##         data1 -- subjects in columns and variables in rows.
##         groups -- grouping variable used to group subjects.
##
##    Output:
##          matrix with columns observed treatment difference and p-values.
##
##############################################################################


row.rand.test<-function(data1,groups){

 do.rand.test<-function(d,g){
    x<-d[g==unique(g)[1]]
    y<-d[g==unique(g)[2]]
    rand<-random.approx(x,y,10000)
    c(obs.d=rand$D, p.val=rand$P)
 }

 t(apply(data1,MARGIN=1, FUN=do.rand.test, g=groups))

}
```