2007-07-06

Sensitivity to Distributional Assumptions in Estimation of the ODP Thresholding Function

Wendy Jill Bunn
Brigham Young University - Provo

SENSITIVITY TO DISTRIBUTIONAL ASSUMPTIONS IN ESTIMATION OF THE ODP THRESHOLDING FUNCTION

by

Wendy J. Bunn

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

August 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Wendy J. Bunn

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Natalie J. Blades, Chair
Date	Scott D. Grimshaw
Date	David G. Whiting

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Wendy J. Bunn in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Natalie J. Blades
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

SENSITIVITY TO DISTRIBUTIONAL ASSUMPTIONS IN ESTIMATION OF THE ODP THRESHOLDING FUNCTION

Wendy J. Bunn

Department of Statistics

Master of Science

Recent technological advances in fields like medicine and genomics have produced high-dimensional data sets and a challenge to correctly interpret experimental results. The Optimal Discovery Procedure (ODP) (Storey 2005) builds on the framework of Neyman-Pearson hypothesis testing to optimally test thousands of hypotheses simultaneously. The method relies on the assumption of normally distributed data; however, many applications of this method will violate this assumption. This thesis investigates the sensitivity of this method to detection of significant but nonnormal data. Overall, estimation of the ODP with the method described in this thesis is satisfactory, except when the nonnormal alternative distribution has high variance and expectation only one standard deviation away from the null distribution.

ACKNOWLEDGEMENTS

I would like to express appreciation to those who have offered their support throughout this thesis: my husband, Ben, for showing constant love and devotion; my parents and family, for expressing encouragement and confidence; Dr. Natalie Blades, for giving guidance and expert advice; the BYU Statistics department faculty, for providing me opportunities to stretch my limits; and my Heavenly Father, for blessing me with the ability and strength to succeed.

CONTENTS

TABLES

Table

FIGURES

Figure

1. MULTIPLE COMPARISONS IN GENE EXPRESSION EXPERIMENTS

Advances in modern technology, particularly in high-dimensional biological studies, have yielded vast data repositories. The task of sorting through the volume of resulting data becomes problematic, particularly when the goal is to test multiple hypotheses simultaneously. In response, the theory and methods of hypothesis testing have been refined for a broad range of applications, including magnetic resonance imaging (MRI), proteomics, and gene expression experimentation.

MRI is a widely used technology in medicine which utilizes magnets and radio waves to probe the human body (Harvey et al. 2006). During the imaging procedure, the patient is placed in a strong uniform magnetic field. This causes hydrogen nuclei in the patient's cells to align themselves either parallel or antiparallel to the field. Brief pulses of electromagnetic energy are sent through the field, perpendicular to the direction of the field. Some of the aligned nuclei absorb the pulses of electromagnetic energy and shift out of their alignment with the magnetic field. After the pulse passes, the nuclei emit their additional energy and realign themselves with the field. These energy emissions are recorded as signal output and combined to create the image; computers can rotate this image to construct a three-dimensional map of the body's interior. This imaging process is important because it allows clinicians to view vital body structures in a non-invasive way. Therefore, it is crucial that imaging software is able to sort out the millions of wave measurements into a clear representation of the body.

Improved methods for proteomics and genomics experimentation have recently been developed with the help of computers and automated equipment. In proteomics experimentation, the entire collection of proteins from a cell or organism is separated into smaller subgroups by their polarity, size, and other distinguishing features. From there, the structure and function of these proteins is investigated using tech-

niques such as x-ray crystallography, amino acid sequencing, and mass spectrometry. Implementing each of these methods relies heavily on the proper interpretation of data.

Gene expression experiments use microarrays to determine the function of a specific gene or pinpoint the genes that are expressed in a cell at a certain time. For example, a researcher might be interested in knowing which genes in a human brain cancer tumor cell are being expressed differentially; that is, genes that are expressed at different levels (either higher or lower) when compared to a normal brain cell. Microarray technology makes it easy to assay the expression levels of thousands of genes simultaneously on one array. Due to the expense of these arrays, comparatively few replicates are created (usually fewer than 20). On these precious few arrays, many thousands of genes are compared.

The quest to discover differentially expressed genes among the thousands present on a microarray chip or to detect subtle changes in the density of brain tissue with MRI has introduced concern over how to sort out the valid signal from the data noise. The debate regarding control of errors made in high-dimensional multiple testing situations has been difficult to resolve. The application of statistical analyses to large-scale data sets has motivated many of the recent advancements in multiple testing, particularly the Optimal Discovery Procedure which will be discussed in the next two chapters.

2. MULTIPLE COMPARISON PROBLEM

The problem of multiple comparisons, or multiple testing, has been a focus of statisticians for many decades. Although a procedure for testing a single hypothesis was established in the early 1900s, applying this optimal method to many tests simultaneously continues to challenge statisticians.

2.1 Simple Hypothesis Testing

The modern foundations of hypothesis testing stem primarily from the contributions of R. A. Fisher, Jerzy Neyman, and Egon Pearson. Their ideas can be separated into two different approaches: Fisher's "p-value procedures"(Royall 1997) and Neyman-Pearson likelihood-based inference. The basic hypothesis testing framework begins with the construction of two competing hypotheses, H_0 and H_1, and a parameter of interest, θ; in the simple case, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. After gathering data, the most likely hypothesis should be favored.

Fisher's p-value approach computes a test statistic (a function of the data), such as the t-statistic for a one-sample test, $t = \dfrac{\bar{x}}{s/\sqrt{n}}$ (Fisher 1925), which follows a t-distribution. The probability that a particular extreme statistic was obtained by chance is the p-value, and is computed under the assumptions of the null hypothesis. These p-values are often interpreted as the strength of the evidence against the null hypothesis. A small p-value for a particular test corresponds to a small probability that the data were generated by the null distribution. A large p-value, on the other hand, implies that the observed statistic could reasonably occur if $\theta = \theta_0$. Large p-values suggest that there is insufficient evidence against the null hypothesis, not that the null hypothesis is necessarily true.

After a p-value is computed, a decision must be made: reject the null hypothesis and favor the alternative, or fail to reject the null hypothesis. Using this procedure, very small p-values—less than 0.05 or 0.01—may be declared statistically significant.

This decision results in the subsequent rejection of the null hypothesis in favor of the alternative hypothesis; however, even if the null hypothesis is rejected, there is a chance that a mistake was made. The probability that the null hypothesis was rejected falsely is called the Type I error, or α (there is also a Type II error related to power). A false rejection of the null hypothesis occurs when the randomly sampled data give evidence in favor of the wrong conclusion. In general, this error is serious enough that it needs to be controlled below a small value. In Fisher hypothesis testing, the Type I error (α) can be controlled by fixing it at some acceptable level, such as 0.05 or 0.01. The resulting p-values are then compared to α in order to make a decision. If the p-value is smaller than the chosen α, then the null hypothesis is rejected. Although there is a nonzero probability that a Type I error has been made, it has been controlled to be no more than α.

The hypothesis testing approach proposed by Neyman and Pearson is based on the likelihood ratio. This ratio consists of the likelihood of the null hypothesis given the data, divided by the likelihood of the alternative hypothesis given the data:

$$\lambda(\mathbf{x}; \theta_0, \theta_1) = \frac{L(\theta_0 \mid \mathbf{x})}{L(\theta_1 \mid \mathbf{x})}.$$

If this ratio is greater than 1, then the data support the null hypothesis over the alternative. Conversely, if the ratio is less than one, then the alternative hypothesis is better supported by the data than the null. Slight favoring of one hypothesis over the other is insufficient to make the decision to reject or fail to reject the null hypothesis. A critical region must be constructed such that the null hypothesis should be rejected if the likelihood ratio is less than k. The most powerful method for testing the hypotheses stated previously is given by the Neyman-Pearson lemma (Neyman and Pearson 1928a). This method specifies that a critical region C should be constructed from values of $\mathbf{x}$ so that the likelihood ratio is less than or equal to k

$$C = \{\mathbf{x} \mid \lambda(\mathbf{x}; \theta_0, \theta_1) \leq k\},$$

where k is some constant so that $\mathrm{P}(\mathbf{X} \in C \mid \theta_0) = \alpha$. If the likelihood ratio is less than the specified k, then the null hypothesis will be rejected. Like the p-value procedure, if the null hypothesis is rejected using the likelihood approach, there is still a chance of a Type I error. This error is controlled by adjusting the width of the critical region C. Again, α is set to a reasonable level and then the value of k is chosen according to the formula above, ensuring that the Type I error rate is controlled. This method is optimal because it has the most power to detect true alternative hypotheses. Power is the ability of a test to detect a true alternative hypothesis and is a desirable characteristic of any test. In fact, the critical region defined by C above is the most powerful critical region of size α.

2.2 Multiple Hypothesis Testing

Although the work of Fisher and Neyman and Pearson presented straightforward techniques for testing a single hypothesis, a standardized approach for comparing several quantities in a pair-by-pair fashion still does not exist. The problem of multiple comparisons is a complicated extension of the single hypothesis test scenario. One simple approach involved comparing each possible pair of groups using a two-sided t-test, for a total of $\binom{m}{2}$ tests for m groups, with the resulting p-values all compared to α to determine significance. This reasoning creates a problem with the analysis—the overall probability of a Type I error occurring is actually larger than α. This multiplicity effect (Tukey 1977) can be explained as chance structure that appears in a large number of test statistics (Diaconis 1985, p. 9). This apparent structure shows up when one attempts to indiscriminately use a single-test method in a multiple-comparison problem.

The Neyman-Pearson approach only guarantees its rejection region to be of size α when the test is performed once; there is no assertion that many independent tests performed simultaneously will have an overall error rate equal to α. As an

example, consider 10 independent hypothesis tests. Each of the 10 resulting p-values is compared to $\alpha = 0.05$ to determine significance. The probability that a Type I error has not been made in the first test is $(1 - \alpha) = 0.95$; however, the probability that a Type I error has not been made in any of the 10 tests is $(1 - \alpha)^{10} = 0.599$. Because the probability of a Type I error is inflated, this method is clearly inadequate for arriving at correct decisions about the hypotheses. In order to present more reasonable multiple testing methods, other error measures will be defined that might be desirable to control, depending on the situation. Additionally, methods that have been developed to control these errors will be discussed later in this chapter.

2.3 Error Rates

In the context of multiple testing, there are several error measures that evaluate how well a multiple testing procedure performs. The three most common are the per-comparison error rate (PCER), the family-wise error rate (FWER) and the per-family error rate (PFER). The per-comparison error rate and the family-wise error rate have been used generally over many years, and the per-family error rate is a function of the per-comparison error rate. To illustrate these error rates, consider Table 2.1 below (Benjamini and Hochberg 1995).

Table 2.1: Benjamini and Hochberg table for testing m null hypotheses. Of the m hypotheses tested, m_0 are true and $m - m_0$ are false. There are R rejected hypotheses—V of them are rejected incorrectly and the remaining S are false null hypotheses.

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

The PFER measures how many Type I errors are expected when testing a family of m hypotheses.

$$\text{PFER} = E(V),$$

where V is the number of true null hypotheses rejected (false positive findings). The PCER is a function of the PFER; the expected number of false positives is divided by m, the number of tests, to yield the expected number of false positives per comparison being made.

The family-wise error rate (FWER) measures the probability of at least one Type I error occurring in a set of hypotheses. The FWER is defined as:

$$\text{FWER} = Pr(V \geq 1),$$

where, again, V is the number of false positive findings.

In general, these error rates are related by the following equation:

$$\text{PCER} \leqq \text{FWER} \leqq \text{PFER}.$$

Control of these error rates at a minimal level is the goal of a variety of multiple testing procedures. Hochberg and Tamhane (1987) describe the advantages of controlling either the FWER or the PFER. First, minimizing the FWER for a family with an infinite number of possible inferences is possible; controlling the PFER in this situation is not possible. Also, controlling the FWER produces inferences that are simultaneously correct. Alternatively, minimizing the PFER in a finite family situation also controls the FWER, because the PFER is an upper bound. Controlling the PFER also penalizes the experimenter for testing an excessive number of hypotheses. Besides these error rates, there are other less conservative measures that have been developed to perform well in specific applications, such as microarray gene expression experiments. Standard methods for controlling these rates are the subject of the next section.

2.4 Standard Controlling Methods

> It would be wrong, I argue strongly, to try to find a single multiple comparisons procedure for general use. There can be, very occasionally, "a man for all seasons"; but we do not dare to seek "a single procedure for all experiments."
>
> —From John W. Tukey's preface to "The Collected Works of John W. Tukey"

The vast spectrum of existing multiple comparison procedures began with two basic methods developed by R. A. Fisher (Hochberg and Tamhane 1987): the least significant difference (LSD) and Bonferroni procedures. These two procedures became the foundation for a variety of multiple testing techniques that are widely used today.

The Bonferroni procedure is a single-step method that controls the PFER, thereby controlling the Type I error rate in the strong sense; that is, any combination of true and false null hypotheses will result in a controlled α level. This procedure is based on the Bonferroni inequality (also known as Boole's inequality):

$$P(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i),$$

where $P(E_i)$ is the probability that E_i is obtained for E_i not necessarily disjoint events in the sample space. So, $P(\bigcup_{i=1}^{n} E_i)$ is the probability that any E_i will be obtained. This probability is always less than or equal to the sum of the individual probabilities. In the multiple testing situation, the Bonferroni correction states that if n hypotheses are tested, then α should be set at $\frac{\alpha}{n}$. So, if

$$P(T_i \in \gamma \mid H_0) \leq \frac{\alpha}{n}$$

for all i in n, then using the Bonferroni inequality,

$$P(T_i \in \gamma \mid H_0) \leq \alpha$$

for any i in n. This simple correction becomes problematic as the number of hypotheses increases. For example, if testing 1000 hypotheses individually at $\alpha = 0.00005$,

the null hypothesis would only be rejected for extremely small p-values and this test would not have much power.

Least Signficant Difference (LSD) is a two-step procedure that controls the FWER and thereby controls the Type I error rate in the weak sense; that is, the Type I error is generally controlled when all null hypotheses are true, but under other combinations of hypotheses (i.e., one false hypothesis and the remaining true hypotheses), α will be larger than expected; however, this method is more powerful than the Bonferroni procedure. The LSD procedure works as follows: first, all the comparisons are tested using ANOVA with the null hypothesis of equal group means; next, if this overall test is significant, all pairwise comparisons are tested using a t-test. This method initially seemed to protect against false positives by first testing if all means were equal; however, as mentioned above, this procedure can give a high error rate when only a few means in the set are different from the others.

Both of the procedures discussed above have limitations, and attempts have been made in many cases to remedy them. As a result, many multiple comparison procedures have been developed during the last half-century. Some procedures are intended for use in broad situations, and others are recommended only in specific applications. The existence of a variety of methods echoes the sentiments voiced by John W. Tukey at the beginning of this section and allows the statistician to choose the method that best suits the type of problem. Two commonly used multiple testing methods are Scheffe's S procedure and Tukey's T procedure. These were developed for balanced designs—designs in which an equal number of observations are made for each treatment combination. Note that there are many other procedures, both single-step and multi-step, that will not be included here for the sake of brevity.

Scheffe's S procedure gives corrected simultaneous confidence intervals and is closely related to the ANOVA F-test. Scheffe's procedure adjusts the general formula for confidence intervals by using a t table value instead of an F table value. For a

two-sided confidence interval for one test, the formula is

$$\bar{y_i} - \bar{y_j} \pm \text{table value} \times SE(\bar{y_i} - \bar{y_j}),$$

where the table value is the $100(1 - \frac{\alpha}{2})$ percentile from a t-distribution. The t-distribution resembles the symmetric, unimodal normal distribution but has heavier tails when the sample size is small.

In some multiple testing situations, however, the desired comparisons are linear combinations of the treatment means—they are not just pairwise comparisons. In this case, Scheffe's procedure corrects the table value to that shown in the formula below. Notice that this procedure does not penalize for the number of comparisons being made and does not require contrasts to be specified before data analysis begins.

$$\text{table value} = \sqrt{r-1} \times \sqrt{F(r-1, \text{df}_{\text{error}})},$$

where r is the number of treatment means and n is the total sample size.

Tukey's T procedure gives simultaneous confidence intervals for pairwise comparisons, and is optimal when computing all possible pairwise comparisons. Again, the multiple comparison correction by Tukey lies in the table value for the confidence interval. This correction incorporates the standard range test and distribution into its formula, shown below as confidence intervals for all possible pairwise comparisons.

$$\bar{y_i} - \bar{y_j} \pm \sqrt{MSE} \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \times \frac{q(r, \text{df}_{\text{error}})}{\sqrt{2}}$$

In this equation, $\frac{q(r,\text{df}_{\text{error}})}{\sqrt{2}}$ is the table value from the standard range distribution, r is the number of treatment means, and n is the total sample size. Tukey's T procedure can be modified to apply to unbalanced designs as well, as shown by the methods proposed by Tukey and (independently) Kramer, and those proposed by Miller and Winer.

2.5 FDR Solution

The false discovery rate (FDR) is a Type I error–controlling method formally stated by Benjamini and Hochberg (1995). Developed specifically for multiple comparison situations, this method is attractive when control of the FWER is too strict and rejects too few hypotheses to be useful. Though a small Type I error rate is desirable, in some applications allowing a few falsely rejected hypotheses is worth the resulting gain in power. There is a key distinction, however, between the Type I error rate and the FDR. For example, a Type I error rate of 5% corresponds to 5% of the true null hypotheses being rejected; however, an FDR of 5% means that 5% of hypotheses that are declared significant are in fact true null hypotheses (Storey and Tibshirani 2003). The goal of the false discovery rate is to determine the expected proportion of rejected null hypotheses that have been wrongly declared significant. Benjamini and Hochberg (1995) define the FDR as

$$\text{FDR} = E\left(\frac{V}{R}\middle| R > 0\right) \times P(R > 0).$$

This error rate can be estimated using p-values according to the algorithm from Storey (2002):

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 M \times t}{\#\{p_i \leq t; i = 1, ...M\}}.$$

The quantity $\hat{\pi}_0$ is the estimated proportion of true null hypotheses among the M tested hypotheses, and the denominator calculates the number of p-values, p_i, that are less than the cutoff value t.

Another equivalent estimation method resamples the observations after permuting the treatment group labels. This method, outlined in the supporting appendix of Storey, Xiao, Leek, Tompkins, and Davis (2005b), estimates the FDR as follows for a fixed significance cutoff c:

$$\widehat{\text{FDR}}(c) = \frac{\hat{\pi}_0 \frac{1}{B} \sum_{b=1}^{B} \#\{F_i^{0b} \geq c; i = 1, ..., M\}}{\#\{F_i \geq c; i = 1, ..., M\}},$$

where $\hat{\pi}_0$ is derived from

$$\hat{\pi}_0(c') = \frac{\#\{F_i < c'; i = 1, ..., M\}}{\frac{1}{B}\sum_{b=1}^{B} \#\{F_i^{0b} < c'; i = 1, ..., M\}}.$$

The statistics $F_1^{0b}, F_2^{0b}, ..., F_M^{0b}$ are simulated null statistics generated using the bootstrap of the alternative model residuals added to the null model, and F_i are the observed F test statistics. This method finds the proportion of estimated null statistics that are greater than some cutoff c and divides that proportion by the number of observed F statistics that are greater than c. The resulting value represents the expected proportion of the observed "extreme" (larger than c) F statistics that are actually null. As the algorithm suggests, the estimate for π_0 is first found by resampling, then integrated into the resampling-based FDR estimate analogous to the method in Storey (2002).

Not only estimating, but also controlling the FDR is of particular interest. Benjamini and Hochberg offer a method—referred to by Storey (2002) as the sequential p-value method—to control the FDR at some level q^*. By the algorithm presented by Benjamini and Hochberg, the p-values from the m hypothesis tests are ordered from smallest to largest and a cutoff value $\hat{k}$ is computed as follows:

$$\hat{k} = \max\left\{ i : p_i < \frac{i \times q^*}{m} \right\}.$$

All p-values less than the cutoff $\hat{k}$ are declared significant. A plot showing a set of ordered p-values overlaid with the $\hat{k}$ calculation $\frac{i \times q^*}{m}$ with $q^* = 0.05$ is shown in Figure 2.1.

The cutoff ($\hat{k}$) can be easily identified in Figure 2.1 as the point where the p-values (shown in red) are larger than the computation $\frac{i \times q^*}{m}$ (shown in blue). The value of $\hat{k}$ for these data is close to 100, which is the point just before the red p-values rise above the blue line. In this example, the 101 smallest p-values are declared significant, and the FDR is controlled at $q^* = 0.05$.

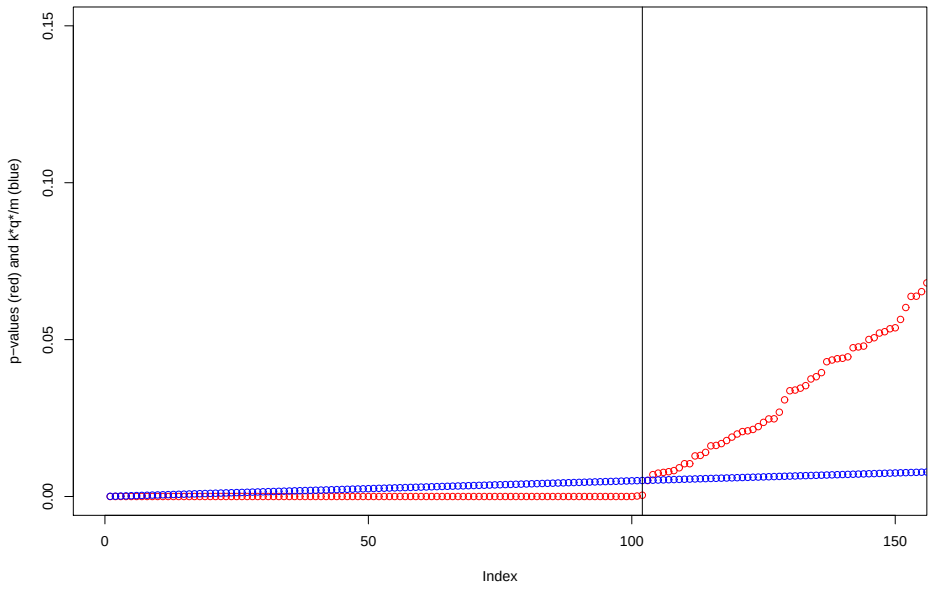


Figure 2.1: Benjamini and Hochberg cutoff $\hat{k}$ for controlling FDR. This figure shows the p-values (in red) plotted against the cutoff $\hat{k}$ computed by $\frac{i \times q^*}{m}$ for each ith p-value (in blue). The cutoff is located at the vertical line, where the p-values exceed the $\frac{i \times q^*}{m}$ line.

In general, if $i = m$, then all null hypotheses are rejected at a level equal to q^*, or α. If $i = 1$, then the hypothesis corresponding to the smallest p-value is rejected and the FDR is controlled by the Bonferroni correction method.

FDR control in a multiple comparison setting can also be considered from a decision theoretic perspective. Müller, Parmigiani, and Rice (2006) discuss several loss functions with associated optimal decision rules that can be applied to a microarray analysis. Let δ_i be an indicator that the ith test is rejected; that is, gene i is determined to be significant. The total number of rejected hypotheses is $D = \sum \delta_i$. The indicator variable r_i is either 0 or 1; it takes on the value 1 if the ith gene is truly differentially expressed and takes on the value 0 otherwise. With this parameterization, the FDR is defined as follows:

$$\text{FDR} = \frac{\sum (1 - r_i)\delta_i}{D}.$$

The ith term in the sum is nonzero when gene i is not differentially expressed, yet it has been called significant (a false discovery). The first loss function from Müller et al. (2006) relates the posterior expectations of false positive and false negative counts ($\overline{FD}$ and $\overline{FN}$, respectively). These counts are defined as

$$\text{FD} = \sum (1 - r_i)\delta_i$$

and

$$\text{FN} = \sum r_i(1 - \delta_i).$$

The loss function uses a linear combination of these quantities as follows:

$$L_N(\delta, z) = c\overline{FD} + \overline{FN}.$$

In this formula, z is a summary statistic, c is a fixed cutoff threshold, and v_i is $\text{P}(r_i = 1 \mid Y)$, the marginal posterior probability of differential expression for gene i given the data Y. The optimal decision indicator, δ^*, recommended by Müller et al. is to declare all genes with marginal probability v_i greater than a threshold t as differentially expressed. This loss function relates false discoveries and false nondiscoveries (both undesirable events), but assumes that all false negatives are equally weighted and that all false positives are equally weighted. Müller et al. also outline loss functions that weight the loss for each gene by how differentially expressed the gene is. In particular, the loss function

$$L_m(m, \delta, z) = -\sum \delta_i m_i + k \sum (1 - \delta_i)m_i + cD$$

uses the parameter m_i to measure differential expression; that is, $m_i = 0$ if the ith gene is not differentially expressed, and $m_i > 0$ if the ith gene is differentially expressed. The first term in this loss function gives a reward for correct discoveries, the second term imposes a penalty on false nondiscoveries (with proportionality constant k), and the last term inhibits the model from finding the expression of all genes significant

(in which case, D would equal n, the number of genes). The optimal decision rule for this loss function is

$$\delta_i^* = I\left\{\bar{m}_i \geq \frac{c}{1+k}\right\}.$$

In other words, all genes with a posterior expectation of differential expression level greater than a fixed cutoff should be called significant. Müller et al. also explain that the sequential p-values method by Benjamini and Hochberg can be approximated by using Bayes' rule and modifications of the following loss function:

$$L_U(\delta, z) \equiv \frac{\overline{FD}}{\alpha D} - g_D = \frac{\overline{FDR}}{\alpha} - g_D.$$

In this function, $g_D = \frac{D}{n}$, so the threshold j should be chosen to have an increment in posterior probability w_j less than $\frac{j\alpha}{n}$. Therefore, the optimal rule is to choose threshold

$$j = \max\left\{i : \Delta w_{B(i),i} \leq \frac{\alpha i}{n}\right\},$$

where $\Delta w_{B(i),i}$ is the increment in posterior probability. This rule looks quite similar to the Benjamini and Hochberg (1995) FDR-controlling method. Müller et al. provide additional details on the formulation of this optimal decision rule.

In an effort to address the limitations of the FDR, a new error measure called the positive false discovery rate (pFDR) was formulated by Storey (2002).

2.6 pFDR Solution

The positive false discovery rate (pFDR) is distinguished from the FDR because it is conditional upon at least one rejected hypothesis of the m tests. Using the terminology of Table 1, R (the number of rejected hypotheses) must be greater than or equal to 1. This seemingly slight adjustment provides substantial advantages in terms of power and error control. Storey (2002) defines the pFDR as

$$\text{pFDR} = E\left(\frac{V}{R}\middle| R > 0\right).$$

The added condition that $R > 0$ introduces a new quantity that is different from the FDR in important ways: it is more liberal and more powerful, meaning that it rejects more hypotheses, but it also controls the Type I error rate so that maximum power can be maintained.

To control the pFDR, the rejection region is fixed and the resulting pFDR is computed. Thus, a cutoff value is determined (i.e., 0.05 or 0.01) and the pFDR for the tests can be estimated, thereby controlling the pFDR. The Storey (2002) method to control the pFDR—as previously mentioned—is fundamentally different from the sequential p-value method of Benjamini and Hochberg. Formerly, the error rate was fixed (α), and then the rejection region was determined. Storey (2002) introduced a reversal by first fixing the rejection region (i.e., reject p-values between 0 and γ, where γ is small) and then estimating the achieved error rate.

In practice, Storey (2002) developed several estimates that can be combined to find the error rate resulting from fixing the rejection region at a certain value γ. A necessary estimate must be made as to the value of π_0, the proportion of null hypotheses that are true out of the m tests performed. Storey (2002) estimated π_0 with

$$\hat{\pi}_0 = \frac{W(\lambda)}{(1-\lambda)m},$$

where $W(\lambda)$ is a function of all "accepted" null hypotheses, and λ is a value between 0 and 1. This estimate can be thought of as the ratio of the observed p-values that are greater than $\lambda, (W(\lambda))$, divided by the section of the range [0,1] greater than λ, which is 1-λ. This ratio should equal the total number of null p-values $\hat{\pi}_0 \times m$ over the range (1). Storey estimated the probability that a given p-value is less than γ as

$$\widehat{P(P \leq \gamma)} = \frac{R(\gamma)}{m},$$

where $R(\gamma)$ is a function of all the rejected null hypotheses. Combining these estimates with a few minor adjustments, Storey created an estimate of the pFDR that is quite

useful under this new controlling method. Given that the rejection region is known (i.e., [0,γ]), the error rate of the m hypothesis tests can be calculated easily using this estimate:

$$\widehat{\text{pFDR}_\lambda}(\gamma) = \frac{W(\lambda) \times \gamma}{(1-\lambda)(R(\gamma) \vee 1)(1-(1-\gamma)^m)}.$$

Storey (2002) also showed that by using this controlling method, the number of hypothesis tests that are rejected increases, while the error rate is still "controlled" at the same level. This gives the pFDR-controlling procedure the advantage of being more liberal, but also more powerful.

As a part of the pFDR-controlling procedure, Storey also introduced a quantity called the q-value, which is conceptually similar to the more familiar p-value. In the context of the pFDR, the q-value is the minimum pFDR that could occur if the p-values smaller than the chosen cutoff γ were rejected. The process for computing the q-value in practice is detailed in Storey (2002) and can be divided into three steps: estimate π_0, choose λ, and compute q-values.

Estimation of π_0 can be performed using a bootstrap method with a range of λ values in the unit interval (Storey, Taylor, and Siegmund 2004), or a smoothing method (Storey and Tibshirani 2003). With the bootstrap method, bootstrapped samples of the p-values are taken and the pFDR is estimated from each sample on the range of λ values. The λ that produces the smallest MSE among the sets of samples is chosen as the optimal λ, which is then used to estimate π_0. With the smoothing method, $\pi_0(\lambda)$ is estimated for each possible λ. Then, the values of $\hat{\pi}_0(\lambda)$ are plotted against the λ values and fitted with a natural cubic spline. The final value of $\hat{\pi_0}$ is the estimated $\hat{\pi}_0(\lambda)$ value of the spline at $\lambda = 1$. Lastly, the q-values are computed. To do this, the pFDR calculation is performed m times on the set of p-values, setting $\gamma = p_{(i)}$ for each pFDR($p_{(i)}$). This step computes the pFDR when the ith ordered p-value and all smaller p-values are declared significant. Next, the q-value for the largest p-value is equal to $\widehat{\text{pFDR}(p_{(m)})}$. The rest of the q-values

are found by choosing the smaller of each pFDR($p_{(i)}$) and the q-value of the next largest p-value; that is, the ith q-value is the minimum of pFDR($p_{(i)}$) and $q_{(i+1)}$ for all $i = m-1, \ldots, 1$. By way of example, Figure 2.2 below shows the p-values of 1000 observations plotted against their corresponding q-values.

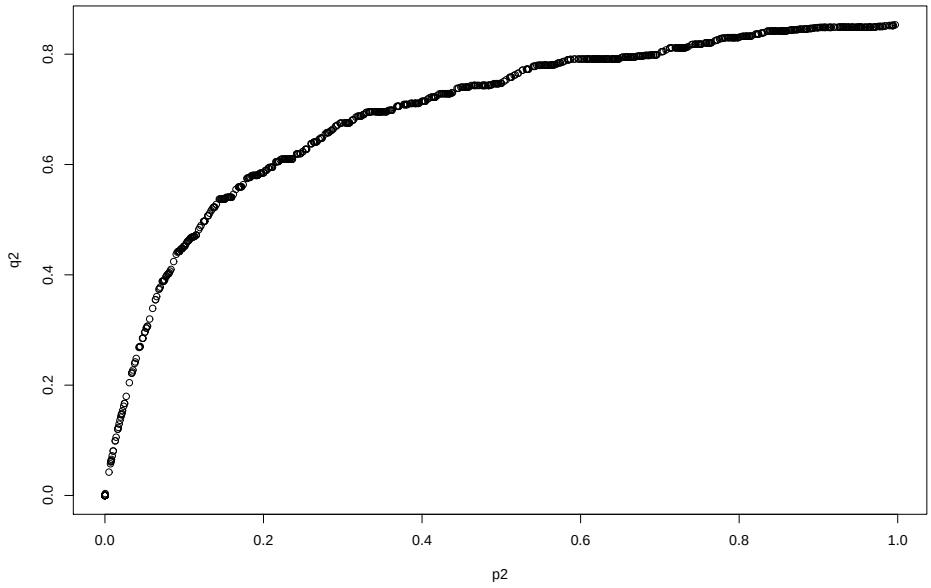


Figure 2.2: Plot of p-values versus their estimated q-values. The p-values in this plot are ordered by size, then the corresponding q-value for each is computed in the three-step process.

In simulations comparing the two procedures, the pFDR procedure performs as well or better than the sequential p-values method in terms of the gains in power, making the pFDR preferable to the FDR as an appropriate error measure.

Storey (2003) also explains that the pFDR can be interpreted from a Bayesian point of view. For a multiple comparison of m tests using the statistics $T_1, ..., T_m$ and a given significance region Γ, the pFDR is

$$\text{pFDR}(\Gamma) = \text{E}\left(\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right),$$

where $V(\Gamma) = \#\{\text{null } T_i : T_i \in \Gamma\}$ and $R(\Gamma) = \#\{T_i : T_i \in \Gamma\}$. The variable $H_i = 0$

if the null hypothesis is true and $H_i = 1$ if the alternative hypothesis is true. The H_i are Bernoulli random variables with *a priori* probability π_0. Applying the Bayesian interpretation, if $T_i \mid H_i \sim (1 - H_i) \times F_0 + H_i \times F_1$ for the null distribution F_0 and the alternative distribution F_1, then

$$\text{pFDR}(\Gamma) = P(H = 0 \mid T \in \Gamma).$$

This statement holds because $P(H_i = 0 \mid T_i \in \Gamma)$ is the same for every $i = 1 \ldots m$. Therefore, the pFDR is simply the posterior probability that the null hypothesis is true, given that the statistic falls in the rejection region. Storey (2003) also gives the q-value a Bayesian interpretation; specifically,

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} P(H = 0 \mid T \in \Gamma_\alpha).$$

Storey shows here that the q-value can actually be thought of as a "posterior Bayesian p-value," or the minimum posterior probability that $H = 0$ for all significance regions that contain the statistic T.

As discussed in this chapter, the majority of research performed in the field of multiple testing has been focused on the problem of creating significance cutoffs for p-values, or modifying the p-values to adjust for multiplicity. Storey (2005) suggests that these methods are all theoretically inadequate because they neglect data structure by testing each hypothesis separately. In the following chapter, the Optimal Discovery Procedure (ODP) developed by Storey (2005) will be presented as a method to optimally test a set of hypotheses as a whole in a similar fashion to Neyman-Pearson hypothesis testing with a single hypothesis.

3. OPTIMAL DISCOVERY PROCEDURE

The majority of research performed in the field of multiple testing has been focused on the problem of creating appropriate significance cutoffs for p-values, or modifying the p-values themselves to adjust for multiplicity. Additionally, many researchers have tried to find the best way to determine the rate at which errors are made, using various error rate measures or controlling methods. Storey (2005) suggests that these methods are all theoretically inadequate because they neglect the inherent structure in the data across hypotheses and test each hypothesis separately. The Optimal Discovery Procedure (ODP) expands the fundamental ideas of Neyman and Pearson—the optimal testing of a single hypothesis—to optimally test a set of hypotheses as a whole (Storey 2005). In many applications, the data are expected to have complex structure; therefore, it makes sense to test these related hypotheses in a joint fashion and assess the significance of each hypothesis relative to the others.

Recall that the Neyman-Pearson lemma uses the likelihood ratio

$$\frac{\text{likelihood under alternative distribution}}{\text{likelihood under null distribution}}$$

to optimally test a single hypothesis. The null hypothesis is rejected if this ratio is greater than the chosen cutoff. The ratio can be thought of as a significance thresholding function of the data (Storey 2005) used to determine the significance of that single hypothesis test. In the same way, the ODP is based on a significance thresholding function of null and alternative hypotheses, but combines information from all hypothesis tests in the set:

$$S_{ODP}(z) = \frac{\text{sum of true alternative densities evaluated at } z}{\text{sum of true null densities evaluated at } z}.$$

The form of this thresholding function is analogous to the Neyman-Pearson likelihood ratio described previously. This function is valuable because it offers a multivariate

approach for analyzing high-dimensional data simultaneously, rather than testing hypothesis i without regard to the information contained in other tests.

The goal of the ODP is to maximize the expected number of true positives for a given number of false positives. That is, a true null hypothesis has a Type I error rate equal to the expected number of false positives from that test. Therefore, if the Type I error rates from all true null hypotheses were added together, the result would be the expected number of false positives, or EFP. In the same way, the sum of the power for each true alternative hypotheses gives the expected number of true positives (ETP). This idea relates to the FDR in the following way:

$$\text{FDR} \approx \frac{\text{EFP}}{\text{EFP+ETP}}.$$

The numerator of this equation looks quite similar to the quantity V and the denominator can be thought of as R in Table 2.1 (Benjamini and Hochberg 1995). As the ETP is maximized for a given EFP, the optimality goal is achieved. Storey (2005) shows that this optimality is equivalent to that reached by the Neyman-Pearson lemma for testing single hypotheses.

3.1 Motivating Example

To illustrate this new procedure, Storey (2005) applies the ODP principles to the following example. This situation tests eight simple hypotheses on normally distributed data with mean μ and unit variance. The hypotheses being tested are $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_i$. Each hypothesis test has a single observed datum z and the true μ values are shown in Table 3.1. The true significance thresholding function S_{ODP} for these data is calculated by

$$S_{ODP}(z) = \frac{\phi(z;-2) + \phi(z;1) + \phi(z;2) + \phi(z;3)}{\phi(z;0) + \phi(z;0) + \phi(z;0) + \phi(z;0)}.$$

The true significance thresholding function for a realized observation z ($S_{ODP}(z)$) is also found in the table.

Table 3.1: Densities and true means for the normal example, as well as true ODP, estimated ODP and Neyman-Pearson rankings of eight hypothesis tests.

Significance test i	1	2	3	4	5	6	7	8
Alternative value of μ_i	-3	-2	-2	-1	1	2	2	3
True value of μ_i	0	-2	0	0	1	2	0	3
Observed datum z_i	1.0	-2.3	-0.02	-0.4	0.5	2.2	-0.1	3.4
ODP rank	4	3	6	8	5	2	7	1
Estimated ODP rank	4	3	6	8	5	2	7	1
UMP unbiased rank	4	2	8	6	5	3	7	1

Inspection of the significance thresholding function reveals that if z_i is from a true alternative density, the value of the numerator will increase because z_i is "probably" close to μ_i. Likewise, if there are other true alternative densities close to z_i, the numerator will increase because z_i is "probably" close to values of other μ_j's; however, if all other densities are true null densities, the sum of true alternative densities will be relatively small, and S_{ODP} will behave like a Neyman-Pearson statistic (Storey 2005).

3.2 Estimation

Though attractively simple, the true ODP significance thresholding function of the data described here and in the previous example cannot be directly evaluated. The primary difficulty with the above formulation is that it requires the true distribution for each test to be known. The true ODP thresholding function separates known true densities into true alternative densities in the numerator and true null densities in the denominator. Although it is not feasible to apply the true ODP thresholding function in practice, there are several methods that can estimate the true thresholding function and overcome this issue.

Because the numeric values of each $S_{ODP}(z_i)$ are important only for finding the ranking of the tests from most to least significant, an equivalent form of the

significance thresholding function is shown in the following formula:

$$\hat{S^*}_{ODP}(z) = \frac{\sum_{i=1}^{m} \phi(z; \hat{\mu}_i)}{\phi(z; 0)}.$$

This estimated form of the true ODP (shown for testing the normal hypotheses in this example) sums the estimated true densities divided by the common hypothesized null density, with $\mu = 0$. The estimated parameters for the true density are calculated using the observed data (in this example, a single data point z) generated by that density. Further, prior identification of the true alternative densities is no longer necessary because the numerator includes estimated densities for all tests, not just those with true alternative hypotheses. Again, because the numeric values of the S_{ODP} statistics are only used to rank the tests, this estimated formula produces the correct ranking even though the statistics have not been scaled.

Storey, Dai, and Leek (2005a) report applicable approaches to estimate the ODP statistic: the "canonical" ODP estimate, weighted estimates, and a nuisance parameter invariance estimate. The canonical ODP estimate generalizes the likelihood ratio test, but is not particularly useful because it requires that the true densities of the null hypotheses be known. This difficulty can be overcome in some cases if a common null density for all null hypotheses ($f(\mathbf{x})$) is known. Thus, the estimated ODP is

$$\hat{S}_{ODP} = \frac{\sum_{i=1}^{m} \hat{g}_i(\mathbf{x})}{f(\mathbf{x})},$$

where $\hat{g}_i(\mathbf{x})$ is the estimated density for the ith observation evaluated at $\mathbf{x}$ and $f(\mathbf{x})$ is the known common null density. A weighted estimate can be computed if weights for the true status of each hypothesis are known. This weighted estimate of the ODP,

$$\hat{S}_{ODP} = \frac{\sum_{i=1}^{m} \hat{g}_i(\mathbf{x})}{\sum_{i=1}^{m} \hat{w}_i \hat{f}_i(\mathbf{x})},$$

changes as the sample size increases. If the null hypothesis is true, the weight $\hat{w}_i$ will go to 1 as n increases, and if the alternative hypothesis is true, $\hat{w}_i$ will go to

0 as n increases. This estimation method is useful when the weights are known. Another estimation method involves the principle of nuisance parameter invariance. The ODP thresholding rule is estimated by imposing the constraint that all of the null distributions f_i are the same, or that $\sum_{i=1}^{m} f_i/m = \sum_{i=1}^{m_0} f_i/m_0$, where m_0 is the number of true null hypotheses. The form of the ODP thresholding rule is then

$$\hat{S}_{ODP} = \frac{\sum_{i=1}^{m} \hat{g}_i(\mathbf{x})}{\sum_{i=1}^{m} \hat{f}_i(\mathbf{x})}.$$

This estimation approach is important to consider because of the effect that nuisance parameters can have on the estimated ODP function (details given in Supplementary Information of Storey et al. (2005a)).

After comparing the ODP with the UMP unbiased test (based on the Neyman-Pearson lemma), Storey's (2005) results show that the ODP performs as well as, or in some situations better than, the UMP test. These findings indicate that an optimal procedure for a single hypothesis case may not be optimal for many simultaneous tests.

For the simulation study described in the following chapter, true and estimated ODP scores will be compared. The true ODP score of observation i is

$$S_{ODP}(z_i) = \frac{\text{sum of true alternative densities evaluated at } z_i}{\text{sum of true null densities evaluated at } z_i}.$$

For these simulations, the estimated score for observation i is

$$\hat{S}_{ODP}(z_i) = \frac{\text{sum of estimated densities evaluated at } z_i}{\text{common null density evaluated at } z_i}.$$

The numerator of this estimated score is a sum of estimated densities—these estimated densities are normal, are centered at the other observation values, and have unit variance. The denominator is a common null density—in this situation it is a standard normal density—used as a scaling factor.

3.3 The ODP Algorithm and Gene Expression Experiments

Recently, the principles of the ODP have been utilized in an important application of multiple testing: the analysis of microarray data with emphasis on gene expression. Discussed briefly in Chapter 1, the motivation of gene expression experiments is to measure the levels of gene expression in different groups of organisms, or in the same organisms over time, and detect differences between groups at the molecular level. These differences can either be up-regulation (an increase in the expression of a gene) or down-regulation (a decrease in the expression of a gene) when compared to a control group. The up- or down-regulated genes are termed "differentially expressed" and are key to discovering the causes and cures for diseases and conditions such as asthma and a variety of cancers. Although identifying individual genes of interest may be the purpose of an experiment, the regulation of gene expression is a complex and intertwined process involving the coordination of many genes at once. For example, a certain gene in a tumor cell may be over-expressed (compared to a normal tissue cell) and in response to this over-expressed gene, a group of related genes may be noticeably under-expressed in that same tumor cell. Therefore, a broader goal of gene expression experimentation may be to understand the ways that genes regulate each other in a dependent fashion.

Expression experiments draw samples of genetic material from test subjects and hybridize them to individual microarrays containing fragments of genes from the subject's genome (the collection of all the genetic information of an organism). In some situations, each sample is applied to only one microarray, while in other situations, the sample may be compared to a reference or control sample within the same array. Often, the key question may be how a particular gene is behaving across all microarray samples.

Applying the ODP to gene expression data, the null hypothesis states that for a particular gene, there is no difference between treatment groups (arrays). This means

that the gene is expressed at the same level, or in relatively the same amounts, in every sample (treatment group). If the alternative hypothesis is true—that is, if there is a significant difference in the expression of the gene depending on which microarray it is located on—then that gene is differentially expressed. Of the observed scores, those with higher scores are more significant because there is a high probability that

(1) there are genes behaving similarly to that gene, or

(2) within the gene, the arrays from different treatments are different from each other (that is, more like their treatment means and less like the overall mean).

3.4 Normality Assumption

To estimate the true ODP, Storey et al. (2005a) use a model containing normal densities. Their model uses these densities in both the denominator and numerator of the thresholding function for a particular gene, i. The use of normal distributions may seem justifiable, especially considering the argument made by Storey et al. (2005a) that gene expression is continuous and approximately normally distributed; however, there are potential consequences if this assumption does not hold true.

The ODP focuses mainly on the estimated thresholding function, not the true thresholding function. In practice, estimating the true ODP is generally required because calculating the true thresholding function implies prior knowledge about the true distribution of every gene and whether or not it is differentially expressed. Estimation of these true densities with normal densities has no effect on the true densities themselves; it can only alter the estimated ODP's ability to identify the differentially expressed genes in an experiment. Storey et al. (2005a) state that because the actual significance of the tests performed is calculated nonparametrically, it is not absolutely necessary to use the correct parametric distribution in estimating the ODP.

On the contrary, there is some evidence suggesting that gene expression data are not always normally distributed. As an example, Slonim (2002) describes the distribution of expression data as being somewhat variable. Slonim indicates that in practice, expression data may display variance heterogeneity between microarrays, and in some cases appear to be generated by a continuous distribution other than a normal. A study by Giles and Kipling (2003) revealed that genes with low expression levels do not correlate strongly with normality, or in other words, they have distinctive nonnormal distributions. Although Storey et al. (2005a) state that the normal density provides a good fit for microarray data, this is not always the case.

There are potential negative consequences of applying an inappropriate approximation to the true ODP. These consequences may vary depending on if the underlying true null density is nonnormal, if the underlying true alternative densities are nonnormal, or both. It is possible that fewer genes that are truly differentially expressed in an experiment may be identified as significant when the estimated ODP is used. Other possible results include an increase in the number of null genes declared significant, or a rearrangement of the estimated significance rankings between differentially expressed genes.

4. SIMULATIONS AND RESULTS

To compare the performance of the estimated ODP function against the true ODP function and the Neyman-Pearson most-powerful procedure in a variety of situations, Storey (2005) performed a simple simulation study. Storey varied three factors in his study: proportion of data from true alternative distributions (.25 or .50), number of observations (48 or 2000), and set of alternative means (-1,1,2,3; 1,2,3; or -2,-1,1,2). All data in Storey's simulation were generated from a normal distribution with mean zero (if true null) or one of the alternative means and unit variance. To explore the effect of nonnormality on ODP estimated scores, this simulation study is extended to include nonnormal observations and test the importance of the normality assumption.

This simulation study investigates how the estimation of the ODP is affected by data that are nonnormally distributed. Four factors are of interest in this simulation: assignment of null and alternative data to normal or nonnormal distributions, proportion of data from the true alternative distribution, nonnormal distribution used, and distance between the null and alternative distribution means (measured in null distribution standard deviations).

4.1 Performance Evaluation

The goal of this simulation study is to identify nonnormal distributions that result in a poorly estimated ODP function. For each simulation, the performance of the estimated ODP compared to the true ODP is measured. Performance will be evaluated using the area under Receiver Operating Characteristic (ROC) curves for both the true and estimated ODP functions, and calculating the area difference between the true and estimated functions. Additionally, the average True Positive Rate (TPR) and False Positive Rate (FPR) for the true and estimated ODP using the 100 highest-ranked observations will be compared.

ROC curves are a graphical representation of the accuracy of a detection method, screening procedure, or test. To construct an ROC curve, the TPR (the proportion of true positives identified out of total true positives) and FPR (proportion of falsely identified positives out of total true negatives) are calculated. The horizontal axis of the graph displays the FPR and the vertical axis displays the TPR. Given these specifications, the curve of an accurate test starts at the origin and rises sharply to a high value of sensitivity, continuing to rise as the FPR increases. Common applications are found in medicine and signal detection theory; examples include screening tests for disease (Jensen et al. 1996), weather forecasting and meteorology (Wilson 2000), and detection of signal intensities on DNA microarray slides (Bilban et al. 2002).

ROC curves are a graphical medium by which two methods or tests can be compared, and the information they provide can be summarized in several ways. Two methods may be compared by considering each method's FPR for a given TPR. Additionally, the area under the ROC curve can be a useful nonparametric summary of overall performance. The total possible area under the curve is 1, and generally the values for area under the curve range from 0.5 (a diagonal line resulting from a random guess) to 1. For some applications, the area under the curve to the right of an FPR of 0.5 is not interesting in terms of evaluating a method's performance; instead, a useful measure of performance is the area under the ROC curve between FPRs of 0 and 0.5 (with possible area values between 0.25 and 0.5). It should be noted, however, that choosing the cutoff 0.5 is a somewhat arbitrary and ad hoc choice.

In the context of ROC curves, there are two quantities that may be used to evaluate performance: area differences and scaled area differences. Area differences are calculated as

Area under ROC curve for true ODP $-$ Area under ROC curve for estimated ODP.

The area differences will fall in the interval [0,1] because the true ODP ROC curve will always outperform the estimated ODP ROC curve. Scaled area differences are

calculated as

$$\frac{\text{Area under ROC curve for true ODP} - \text{Area under ROC curve for estimated ODP}}{\text{Area under ROC curve for true ODP}}.$$

The scaled area differences will also fall in the interval [0,1], but they have an additional interpretation. These values are the proportion of area change from true ODP to estimated ODP, relative to the area under the ROC for the true ODP. The decision to use scaled area differences rather than absolute area differences is rather simple. Storey (2005) states that, theoretically, the true ODP thresholding function is the best possible ranking function in terms of accuracy. That is, there is no way to outperform the true ODP score function; therefore, the true ODP can be used as the baseline to which estimators of that function can be compared. In this way, we aim for an area under the estimated ODP ROC curve that is as close to the area under the true ODP ROC curve as possible, not for any fixed target value. The scaled area differences measure the percent difference between the true ODP (the best-case scenario) and an estimator of that function.

For this simulation set, the scaled area differences and true positives out of the top 100 observations will be averaged over 50 repeated simulations of the same set of conditions, and their standard errors will be reported.

4.2 Preliminary Simulation Results

To begin the series of simulations, a preliminary simulation set is conducted to clarify the factors of interest and investigate potential concerns. In these simulations, comparisons are made between similarly shaped densities from different families. Also, gamma distributions with a common mean and different variances are compared, along with a trial run which varies the proportion of true alternative observations.

4.2.1 Distribution Shape Simulation

The first simulation investigated similarly shaped densities and how the density parameterization might affect the performance of the estimated ODP. In this simulation set, shape and skewness were held relatively constant for three nonnormal densities—a $t(df = 10)$, gamma($\kappa = 9$, $\theta = 0.4$), and lognormal($\mu = 1.5$, $\sigma = 0.0625$)—along with a Normal($\mu = 3.5$, $\sigma = 1$) density for comparison. The expected value of all four densities was approximately 3.5. A plot of the four density functions under consideration is shown in Figure 4.1.

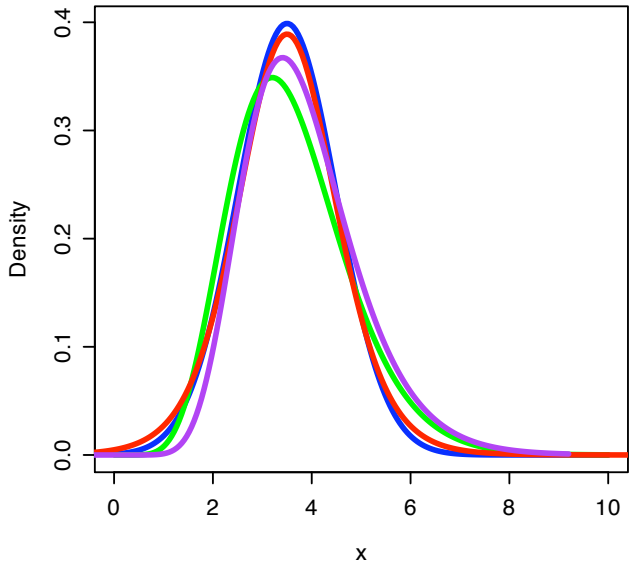


Figure 4.1: Four densities compared in a preliminary simulation: Normal($\mu = 3.5$, $\sigma = 1$) is shown in blue, $t(df = 10)$ is shown in red, gamma($\kappa = 9$, $\theta = 0.4$) is shown in green, and lognormal($\mu = 1.5$, $\sigma = 0.0625$) is shown in purple.

The estimation of the ODP might be affected by characteristics of the density function itself, such as the dependence of mean and variance in the gamma and lognormal distributions. Notice that these four densities are relatively symmetric and have approximately the same shape and spread. In fact, the only important

differences between these four densities are the properties of the individual density families.

The true and estimated ROC curves for each of the densities are compared in Figure 4.2. Notice that visually the estimated ODP curve follows the true ODP curve very closely. In order to evaluate the differences between the estimation of the densities, the scaled area differences are shown in Table 4.1. In all cases, the percent change in area is extremely small (less than half a percent), in some cases less than one quarter of a percent. These results imply that the true ODP is well estimated in these four simulations and that all four distributions are estimated similarly. This second observation supports the conclusion that densities with similar shapes, although they are from different distributional families, produce very similar estimated ODP results.

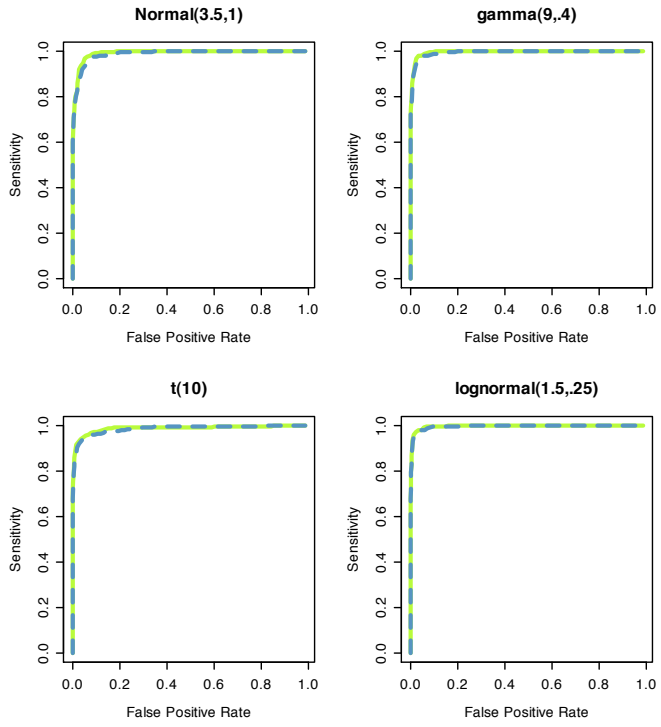


Figure 4.2: ROC curves for each density. The true ODP ROC curve is shown in light green and the estimated ODP ROC curve is shown as a blue dotted line.

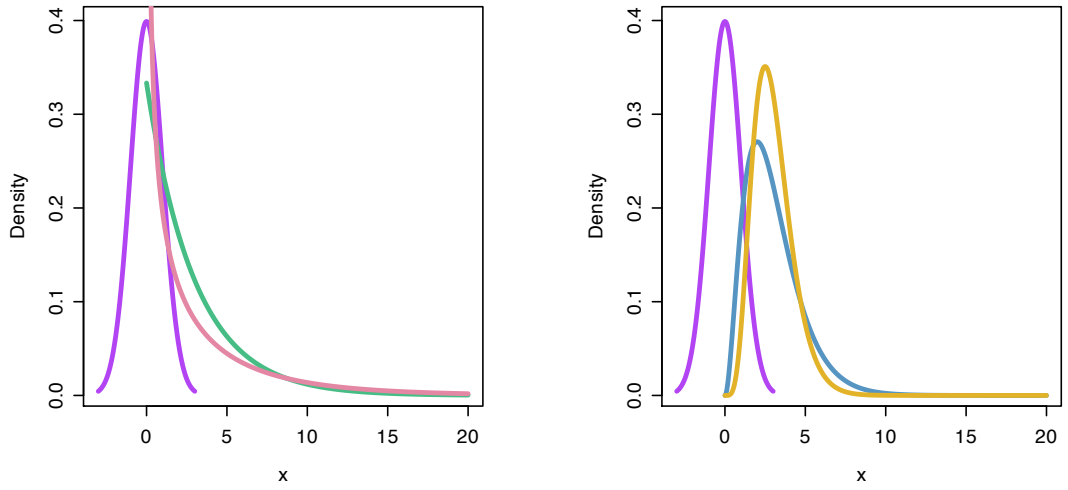
Table 4.1: Summary measures of area differences and scaled area differences to evaluate ODP estimation performance.

Density	Scaled Area Difference (% Change in Area)
Normal	.00194 (.19%)
t	.00445 (.45%)
gamma	.00485 (.49%)
lognormal	.00098 (.10%)

4.2.2 Skewness and Variance Simulation

To investigate the effect of variance on estimating gamma-distributed data, a second set of simulations was conducted. The expected value of each gamma distribution was held constant at 3, but the parameterizations were changed to yield differing variances. Each of four gamma distributions with differing parameterizations but the same expectation was compared to a standard normal null density. The distributions of four extreme examples are shown in Figure 4.3. Notice that in Figure 4.3a, the gamma densities peak at 0, the same location as the peak of the standard normal density. These two gamma densities also have quite long tails. By comparison, the gamma densities in Figure 4.3b peak some distance away from 0, with very low density at 0. They also have smaller variance and, consequently, shorter tails.

When the true and estimated ODP scores were calculated for these four parameterizations, there was a strong connection between estimation performance and density location as determined by variance. For the two distributions with high density close to 0 (the peak for the standard normal null density), the estimation of the ODP was dramatically worse than it was for the two densities that had highest density farther away from 0 (see Figure 4.4). Table 4.2 shows the variations in scaled area differences for the four gamma densities under consideration. As more of the density in the gamma distribution moves away from zero with decreasing variance, the difference between the estimated and true ROC curves decreases. These results indicate



(a) gamma(.5,6) and gamma(1,3)

(b) gamma(3,1) and gamma(6,.5)

Figure 4.3: Panel a: Densities of gamma(.5,6), shown in pink, and gamma(1,3), shown in green, compared to the standard normal density, shown in purple. Panel b: Densities of gamma(3,1), shown in blue, and gamma(6,.5), shown in yellow, compared to the standard normal density, shown in purple.

Table 4.2: Summary measures of scaled area differences to evaluate ODP estimation performance in four different gamma parameterizations under consideration.

Density	Scaled Area Difference (% Change in Area)
gamma(.5,6)	.2603 (26%)
gamma(1,3)	.1102 (11%)
gamma(3,1)	.0211 (2%)
gamma(6,.5)	.0085 (.8%)

that for a given expectation, variability in the density affects how well the true ODP is estimated. Specifically, when regions of high density in the gamma distribution overlap with regions of high density in the standard normal density (characteristic of a high variance gamma), the estimation of the ODP suffers. In the final simulation set, variance is an underlying factor that is incorporated into the design because of its impact on the results in this simulation set.

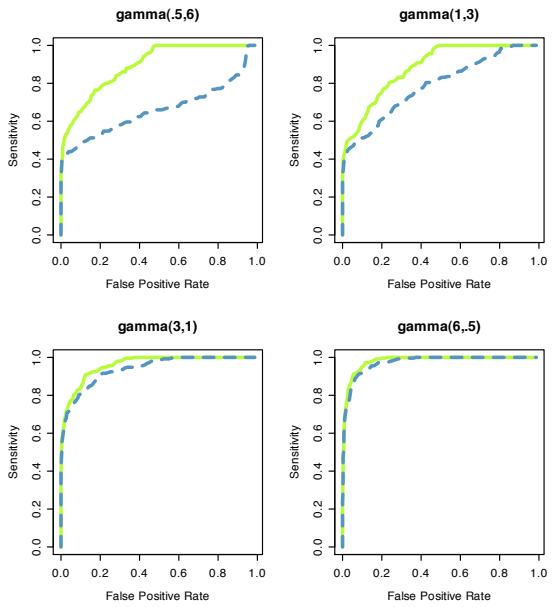
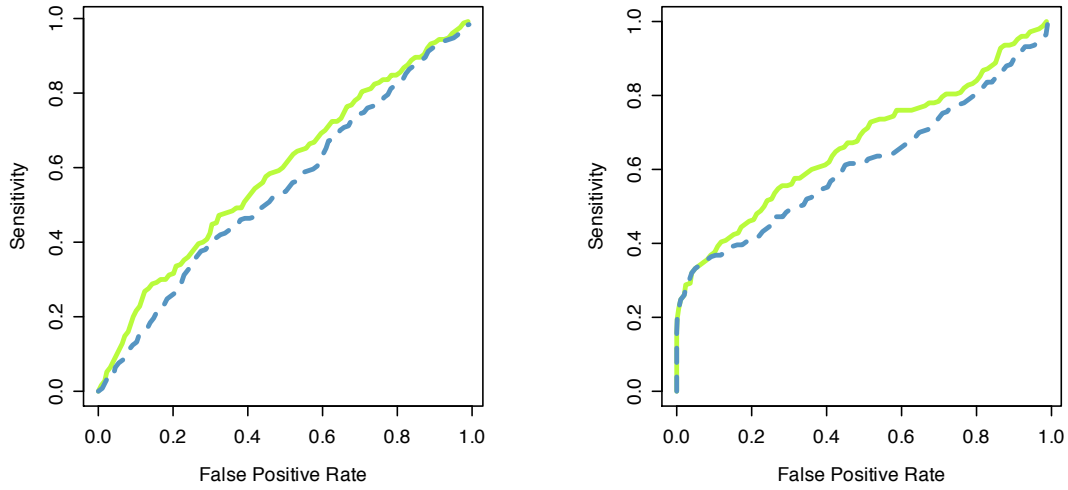


Figure 4.4: ROC curves for each gamma density. The true ODP ROC curve is shown in light green and the estimated ODP ROC curve is shown as a blue dotted line.

4.2.3 Density Location

Two simulations were performed to further investigate the effect of density location relative to the null distribution. First, alternative observations were generated from a normal distribution with mean 0.25 and unit variance. Next, alternative observations were generated from a Cauchy distribution. In both cases, the proportion of alternative observations was constant at 0.25. Both distributions were centered near (or at) zero, but the Cauchy distribution's observations had a wider spread because the distribution has thicker tails and undefined variance. The ROC curves from this simulation are shown in Figure 4.5.

The diagonal line seen in Figure 4.5a indicates that the method is doing little more than randomly guessing whether an observation is null or alternative. The ROC curve for the Cauchy distribution (Figure 4.5b) does not do much better at distinguishing between the two kinds of observations. However, the estimated ODP curve is not far behind the true ODP in both cases.



(a) Normal(.25,1) (b) $t(1)$

Figure 4.5: ROC curves for alternative observations generated from a Normal distribution(.25,1) (panel a) and from a Cauchy distribution (panel b). The true ODP ROC curve is shown in light green and the estimated ODP ROC curve is shown as a blue dotted line.

4.2.4 Proportion of Alternative Observations

In addition to exploring properties of the alternative distributions, the proportion of alternative observations relative to null observations (π_1) is also a factor of interest. The four values for the proportion of alternative observations were chosen as 0.05, 0.10, 0.25, and 0.50. The two smaller values are interesting because they have been suggested as estimated proportions of differentially expressed genes in a sample of microarray gene expression data in many microarray studies (Broberg 2003; Zhang, Yin, and Zhang 2006); the two larger proportions correspond to values used in previous simulations performed by Storey (2005). This simulation set was performed using alternative data from two distributions, the first, gamma(1,3), and the second, gamma(6,.5). Two distributions were used because of a potential interaction between the variance of the distribution and the effect of increasing this proportion. The ROC curves from these two simulations are shown in Figures 4.6 and 4.7.

Notice that the true and estimated curves for the gamma(6,.5) simulation (Figure 4.7) are reasonably close to each other for all values of π_1. Remember that this gamma distribution is also fairly distinct from the null standard normal distribution (recall Figure 4.3b). However, the true and estimated curves for the gamma(1,3) simulation (Figure 4.6) fall closer together; that is, the estimation improves as the proportion of alternative observations increases. The scaled area differences between true and estimated ODP ROC curves are shown in Table 4.3. This higher-variance gamma distribution is close to the null distribution (recall Figure 4.3a), and therefore is more difficult to distinguish. The trend in these results suggests that changing the proportion of alternative observations has a more marked effect upon those alternative distributions that are already difficult to separate from the null distribution. In these cases, as the proportion of alternative observations increases, the estimated ODP performs better (closer to the true ODP curve). However, this conclusion is something of a simplification: although the difference between the simulation with $\pi_1 = 0.25$ and the simulation with $\pi_1 = 0.10$ is essentially impossible to identify by looking at the ROC curve areas, the numerical summary of areas under the curve shows that there is a small lapse in accuracy as the proportion changes. This accuracy lapse is evidenced by a difference between true and estimated ODPs of 0.0702 when the proportion is 0.10 and a difference of 0.0707 when the proportion is 0.25 (see Table 4.3). This is a very small lapse, however, and should not be overly concerning.

Because the results of this simulation set yielded ROC curves with differing shapes, there is some concern over the use of a single-number summary (area under the ROC curve) to compare simulations. Although this nonparametric value is concise, it cannot be used to re-create the original curve. The number of ways that an ROC curve can be constructed and still yield the same area is limitless; perhaps ROC shapes, as well as areas, should be compared. Further investigation into ROC literature yielded no information regarding the characterization of ROC shape. For now, this simulation

set will only use areas under ROC curves and true positives to summarize simulation results.

In general, the preliminary set of simulations provided a useful testing ground on which initial questions about the ODP could be explored. The distribution shape simulations confirmed that the shape of the density, not the density family or parameterization features, is the important characteristic of a nonnormal distribution. The variance and skewness simulations suggested that different distribution variances should be used because higher variance in the alternative distribution was related to comparatively worse estimation of the ODP function. This finding motivated the inclusion of alternative distribution variance as a factor in the next simulation study. Also, the simulations where the density location and π_1 were changed suggested that a very small distance between densities and an increase of π_1 both had negative effects on estimation.

Table 4.3: Summary measures of scaled area differences to evaluate ODP estimation performance with four proportions of alternative observations (alternative observations generated from a gamma(1,3)).

Proportion	AUC_{TRUE}	AUC_{EST}	Scaled Area Difference
0.05	.8868	.7993	.0987
0.10	.9032	.8330	.0777
0.25	.8890	.8183	.0795
0.50	.9001	.8577	.0471

4.3 Simulation Descriptions

The simulations performed for this investigation were subjected to constraints for the sake of simplicity and comparability. The general simulation approach taken for n=1000 observations is as follows. First, $n(1-\pi_1)$ observations were randomly generated from the true null distribution and $n\pi_1$ observations were randomly generated from the true alternative distribution (where π_1 is the proportion of alternative

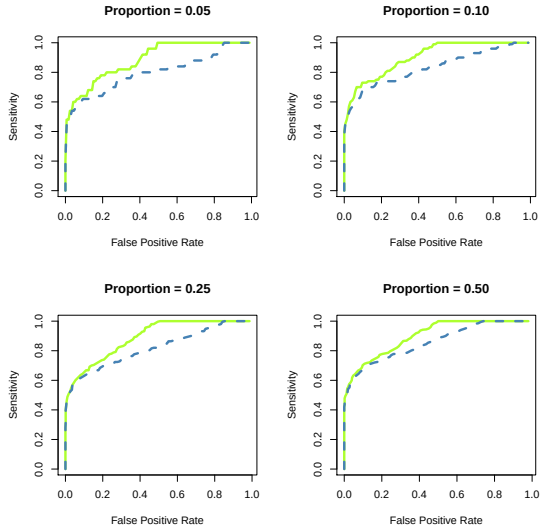


Figure 4.6: ROC curves for alternative observations generated from a gamma(1,3) with proportions of alternative observations of 0.05, 0.10, 0.25, and 0.50, respectively.

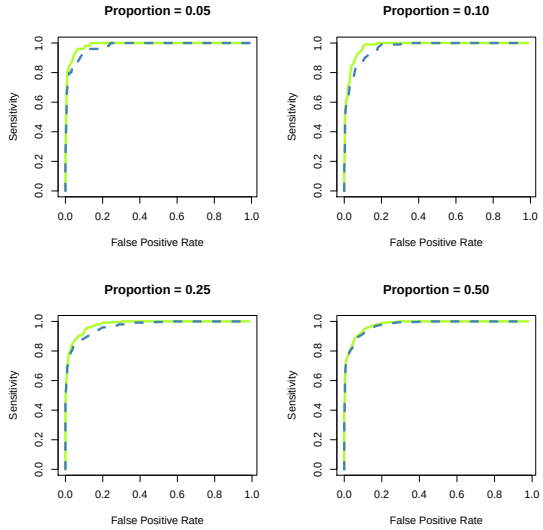


Figure 4.7: ROC curves for alternative observations generated from a gamma(6,.5) with proportions of alternative observations of 0.05, 0.10, 0.25, and 0.50, respectively.

observations). Next, the true ODP scores were calculated using the true alternative densities in the numerator and the true null densities in the denominator. Then, the estimated ODP scores were computed using normal densities (with means determined by the other observations and unit variance) and a standard normal density as the common null distribution in the denominator.

In Storey's motivating example, the common null density was standard normal. Storey et al. (2005a) outline the general form of the estimated ODP as

$$\hat{S}_{ODP}(x) = \frac{\sum_{i=1}^{m} \hat{g}_i(x)}{f(x)},$$

where there are m total significance tests, $\hat{g}_i(x)$ is the estimated density for the ith observation evaluated at x, and $f(x)$ is the common null density. When estimating the ODP function in the context of nonnormal (and sometimes strictly positive) observations, determining what the common null density should be was difficult. Given the lack of guidance afforded by the general formulation for the estimated ODP, it was decided that any choice of null density other than a standard normal would require information about the true null density not usually known in practice. Although using the standard normal as the common null density was not intuitively satisfying, simulated data showed that centering a normal density closer to where the true null density was centered did not influence the rankings among alternative observations produced by the two estimation methods. The decision was made to define the estimated ODP thresholding function as

$$\hat{S}_{ODP}(x) = \frac{\sum_{i=1}^{m} \hat{g}_i(x)}{f(x)},$$

where $\hat{g}_i(x)$ is a normal density evaluated at x using estimated parameters from the ith observation, and $f(x)$ is a standard normal density evaluated at x. This formula was taken to be the "estimated ODP," and was used in all subsequent simulations.

After true and estimated scores were computed, these scores were sorted from largest to smallest, producing a ranking of most to least significant observations. From

these sorted scores, ROC curves for both groups were obtained using knowledge of true null and true alternative observations and areas under those curves were compared. In the process of creating the ROC curves, the true value of π_1 was used, not estimated from the data, although there are methods available to estimate this proportion.

In this set of simulations, there are three possible scenarios, shown in Table 4.4, which are variations of the assignment of null and alternative observations to normal or nonnormal distributions.

Table 4.4: Simulated data distribution scenarios. Scenario 1 consists of normally distributed null data and nonnormally distributed alternative data. Scenario 2 reverses the assignments of normal and nonnormal data. Scenario 3 specifies nonnormal data for both the null and alternative.

Scenario	Null data	Alternative data
1	normal	nonnormal
2	nonnormal	normal
3	nonnormal	nonnormal

As mentioned previously, the proportion of true alternative observations (π_1) will be varied at 0.50, 0.25, 0.10 and 0.05. Varying this proportion is crucial to determine whether the performance of the estimated ODP is affected by the number of true alternatives, specifically at proportion values similar to those seen in gene expression experiments. Also, three nonnormal distributional families will be compared: gamma, lognormal, and t.

For each simulation, observations will be randomly generated from the null and alternative distributions, with proportion of true alternatives equal to π_1. In general, the distance between the expected values of these distributions will be varied at 1, 3, and 5 null standard deviations. Also, the null distribution has unit variance for easier comparability across scenarios and nonnormal distributions, with expected value 0 (for normal or t-distributed observations) or 1 (for nonnormal, strictly positive

observations). The simulations in each of the three scenarios will now be described in detail.

Scenario 1 consists of simulations with normal null observations and nonnormal alternative observations. All null observations are randomly generated from a standard normal distribution. Alternative observations are generated from the three nonnormal distributions with parameterizations corresponding to expected values of 1, 3, and 5 (which are 1, 3, and 5 null standard deviations away from the standard normal null distribution, respectively). However, the gamma, lognormal, and t- distributions are flexible enough to allow for a broad range of possible variances for any given distribution mean. In the preliminary simulation described previously, the variance of the nonnormal distribution had an impact on the estimation performance. To account for this additional complexity, each nonnormal density was evaluated with all possible combinations of three means (1, 3, and 5) and three variances (2, 4, and 8 for the gamma and lognormal; low, medium, and high for the t-distribution), for a total of nine parameterizations with each alternative density. The scenario 1 parameterizations are shown in Tables 4.5 and 4.6.

Table 4.5: Scenario 1 null parameterization

Normal	Mean=0	Variance=1

Scenario 2 consists of simulations with nonnormal null observations and normal alternative observations. Null observations were randomly generated from the three nonnormal distributions with unit variance—the gamma and lognormal distributions had means of 1 and the t-distribution had a mean of 0. Alternative observations were generated from normal distributions with parameterizations corresponding to expected values of 1, 3, and 5 for the t-distribution nulls and 2, 4, and 6 for the lognormal and gamma nulls (which are 1, 3, and 5 null standard deviations away from the null distributions, respectively). To incorporate the concept of different variances

Table 4.6: Scenario 1 alternative parameterizations

		Variance=2 (Low)	Variance=4 (Medium)	Variance=8 (High)
gamma(κ, θ)	Mean=1	$\left(\frac{1}{2}, 2\right)$	($\frac{1}{4}$,4)	($\frac{1}{8}$,8)
	Mean=3	$\left(\frac{9}{2}, \frac{2}{3}\right)$	($\frac{9}{4}$,$\frac{4}{3}$)	($\frac{9}{8}$,$\frac{8}{3}$)
	Mean=5	($\frac{25}{2}$,$\frac{2}{5}$)	($\frac{25}{4}$,$\frac{4}{5}$)	($\frac{25}{8}$,$\frac{8}{5}$)
lognormal(μ, σ)	Mean=1	(-1.099,1.482)	(-.805,1.269)	(-.5493,1.048)
	Mean=3	(.781,.798)	(.915,.606)	(.998,.448)
	Mean=5	(1.471,.527)	(1.535,.385)	(1.571,.277)
t(df)	Mean=1	(1)+1	(5)+1	(10)+1
	Mean=3	(1)+3	(5)+3	(10)+3
	Mean=5	(1)+5	(5)+5	(10)+5

in the alternative observations (featured in scenario 1), the normal distribution was given low, medium, and high variances of 2, 4, and 8, respectively, resulting in nine alternative observation parameterizations. The scenario 2 parameterizations are shown in Tables 4.7 and 4.8.

Table 4.7: Scenario 2 null parameterizations

		Variance=1
gamma(κ, θ)	Mean=1	(1,1)
lognormal(μ, σ)	Mean=1	(-.347,.693)
t(df)	Mean=0	(200)

Scenario 3 consists of simulations with nonnormal null observations and nonnormal alternative observations, and draws on ideas established in the previous two scenarios. Null observations are randomly generated from the three nonnormal distributions with unit variance, as in scenario 2. Alternative observations are generated from nonnormal distributions with parameterizations similar to those in scenario 1 except the gamma and lognormal alternative distributions had to be shifted in order to be the desired number of null standard deviations away from their null counterparts. The scenario 3 parameterizations are shown in Tables 4.9 and 4.10.

Table 4.8: Scenario 2 alternative parameterizations

		Variance=2	Variance=4	Variance=8
For gamma null:	Mean=2	(2,2)	(2,4)	(2,8)
Normal(μ, σ^2)	Mean=4	(4,2)	(4,4)	(4,8)
	Mean=6	(6,2)	(6,4)	(6,8)
For lognormal null	Mean=2	(2,2)	(2,4)	(2,8)
Normal(μ, σ^2)	Mean=4	(4,2)	(4,4)	(4,8)
	Mean=6	(6,2)	(6,4)	(6,8)
For t null:	Mean=1	(1,2)	(1,4)	(1,8)
Normal(μ, σ^2)	Mean=3	(3,2)	(3,4)	(3,8)
	Mean=5	(5,2)	(5,4)	(5,8)

Table 4.9: Scenario 3 null parameterizations

		Variance=1
gamma(κ, θ)	Mean=1	(1,1)
lognormal(μ, σ)	Mean=1	(-.347,.693)
t(df)	Mean=0	(200)

Table 4.10: Scenario 3 alternative parameterizations

		Variance=2 (Low)	Variance=4 (Medium)	Variance=8 (High)
gamma(κ, θ)	Mean=2	(2,1)	(1,2)	$(\frac{1}{2},4)$
	Mean=4	$(8,\frac{1}{2})$	(4,1)	(2,2)
	Mean=6	$(\frac{36}{2},\frac{2}{6})$	$(\frac{36}{4},\frac{4}{6})$	$(\frac{36}{8},\frac{8}{6})$
lognormal(μ, σ)	Mean=2	(.490,.637)	(.347,.833)	(.144,1.048)
	Mean=4	(1.327,.343)	(1.275,.472)	(1.184,.637)
	Mean=6	(1.765,.233)	(1.739,.325)	(1.691,.448)
t(df)	Mean=1	(1)+1	(5)+1	(10)+1
	Mean=3	(1)+3	(5)+3	(10)+3
	Mean=5	(1)+5	(5)+5	(10)+5

4.4 Simulation Results

For the 324 simulation situations discussed in the description section, results and comparisons were obtained and summarized by scenario. Tables of the number of true positives out of the top 100 ranked observations and the scaled area differences

were obtained by averaging over 50 repeated simulations. Graphs of the ROC curves comparing the true and estimated ODP are based on data from a single unreplicated simulation, and hence have more associated uncertainty than the averaged results. The results of these simulations are presented in figures and tables following a more detailed discussion of the outcomes of each scenario. In general, the estimation of the true ODP function improves as the distance between the null and alternative distributions increases, as the variance of the alternative distribution decreases, and as π_1 increases.

In scenario 1, the simulation goal was to use nonnormal alternative observations (from gamma, lognormal, and t) to evaluate the performance of the estimated ODP function. For the t-distribution situations, the estimated ODP function performed only slightly worse than the true ODP function. Although the t-distribution is not normal, it is symmetric and approximately normal (depending on the parameterization), so it is not surprising that the estimated ODP function using normal densities is appropriate in this case. For both the gamma and lognormal distributions the estimation was reasonable, except when the distance between the centers of the two distributions was a single null standard deviation. In the situations where the skewed distributions were close together (see Figures 4.8 and 4.9), there was a large gap between the true and estimated ROC curves. As shown in Tables 4.11 and 4.12, the scaled area differences for the small distance simulations range from 0.26 to 0.71 for the gamma and from 0.14 to 0.53 for the lognormal. Scaled area differences for the t-distribution are shown in Table 4.13. In general, estimation of the ODP suffers when skewed nonnormal distributions that are a single null standard deviation away from the the null distribution are used for the alternative observations.

The objective of the simulations in scenario 2 is to investigate the effect of nonnormal null observations on the estimation of the ODP function. As was observed in the previous scenario, using t-distributed null observations did not drastically affect

the estimation of the ODP function (see Tables 4.16 and 4.25 for the true positives and scaled area differences). As for the gamma and lognormal simulations, those with a small distance between the nonnormal null and normal alternative distributions again produced the worst estimation of the ODP relative to the other simulations; however, none of the scaled area differences exceeded 0.3, as shown in Tables 4.14 and 4.15. In fact, the largest scaled area difference among the skewed distributions occurred with a high variance lognormal, but was only 0.21. The ROC curves for all three distributions with a small distance between the null and alternative distribution means are shown in Figures 4.11, 4.12, and 4.13. It is interesting to note that although many of these simulations have small scaled area differences, in some cases the true ODP function itself does not accurately identify alternative observations. The goal of this estimation method is to obtain the rankings of the true ODP with minimal differences, but clearly there are situations in which the true ODP rankings are not ideal. Overall, the ODP function is reasonably well estimated when nonnormal null observations are used.

In the simulations of scenario 3, the same distributional family was used to generate the null and alternative observations and the performance of the estimated ODP was evaluated for each. As with scenarios 1 and 2, the estimated ODP for the t-distribution performed well, with only minimal scaled area differences, shown in Table 4.19. The gamma and lognormal distributions were both estimated surprisingly accurately. The only non-zero (rounded to four decimal places) scaled area differences in these two sets were for those simulations with high variance alternative distributions that were a small distance from the null distribution, as shown in Tables 4.17 and 4.18. In summary, the estimated ODP is an adequate indicator of significance even when the null and alternative observations are both nonnormally distributed.

A second measure of estimation performance, the number of true positives, was used in this simulation study (see Tables 4.20 to 4.28). Recall that when π_1=0.05,

there are 50 true alternative observations, 100 when π_1=0.10, 250 when π_1=0.25, and 500 when π_1=0.50. Thus, when π_1=0.05, there was a maximum of 50 true positives that could be found in the top 100 observations. Not surprisingly, with a larger proportion of true alternative observations in the sample, the top 100 observations are exclusively true positives when the distance between the centers of the two distributions increases. Inspection of these results reveals conclusions similar to those obtained using the ROC curves and scaled area differences. The skewed nonnormal alternative observations (scenario 1) did not produce a well-estimated ODP when the distance between distribution means was small. Nonnormal null observations, alone or with nonnormal alternative observations, were estimated reasonably; however, the high variance skewed distributions in scenario 3 had noticeable differences in the number of true positives when the distance between distributions was small.

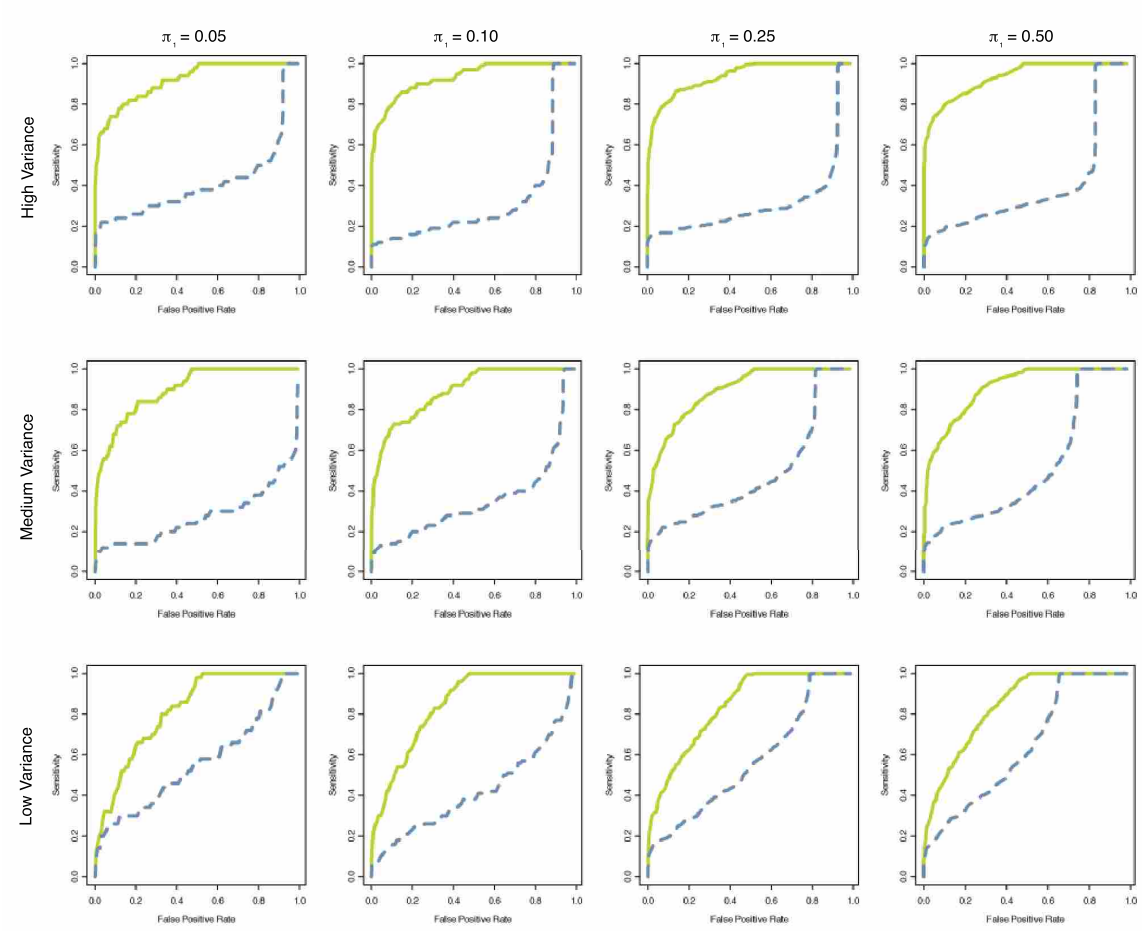


Figure 4.8: Gamma distribution from scenario 1 with mean 1 and high, medium, and low variance. Each row corresponds to an alternative distribution variance that is high (8), medium (4), or low (2). Each column is a different value of π_1, the proportion of true alternative observations, and increases from 0.05 to 0.50, from left to right. All alternative distributions are 1 null standard deviation away from the null distribution.

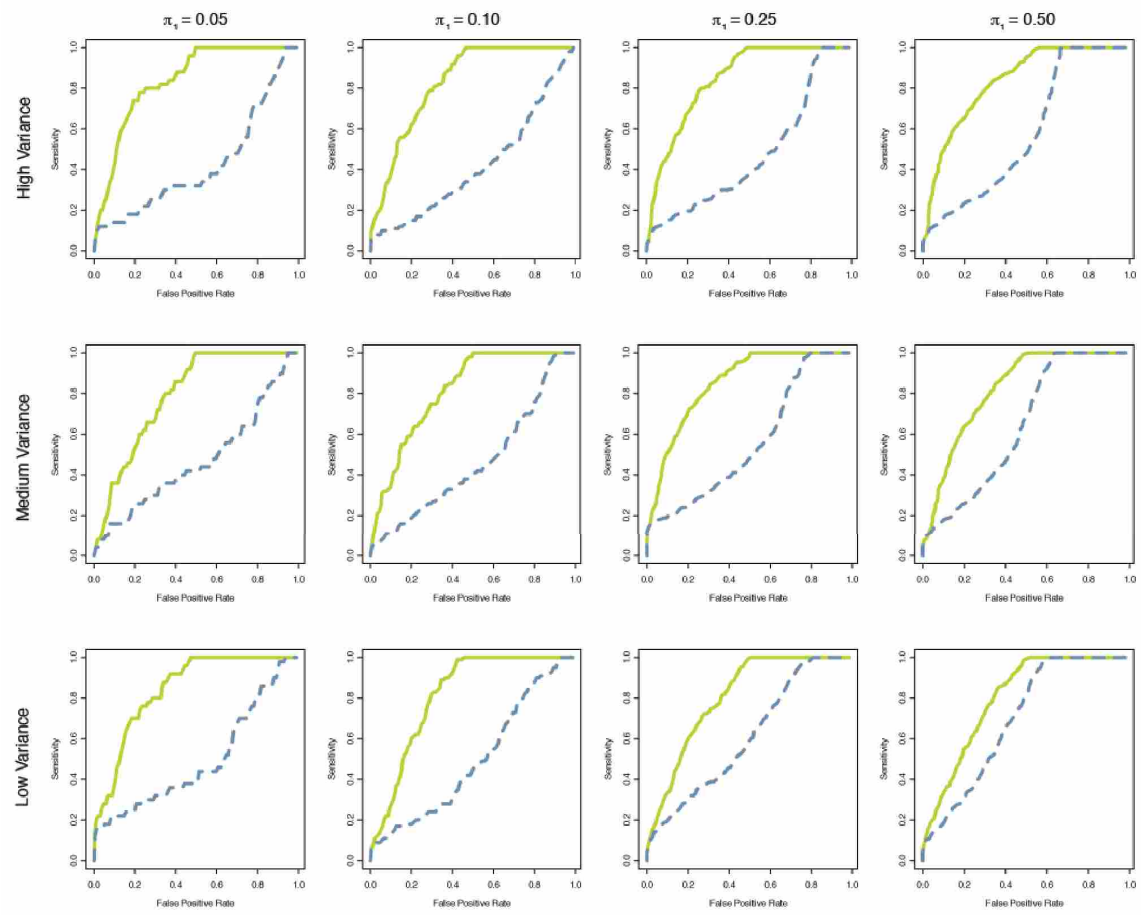


Figure 4.9: Lognormal distribution from scenario 1 with mean 1 and high, medium, and low variance. Each row corresponds to an alternative distribution variance that is high (8), medium (4), or low (2). Each column is a different value of π_1, the proportion of true alternative observations, and increases from 0.05 to 0.50, from left to right. All alternative distributions are 1 null standard deviation away from the null distribution.

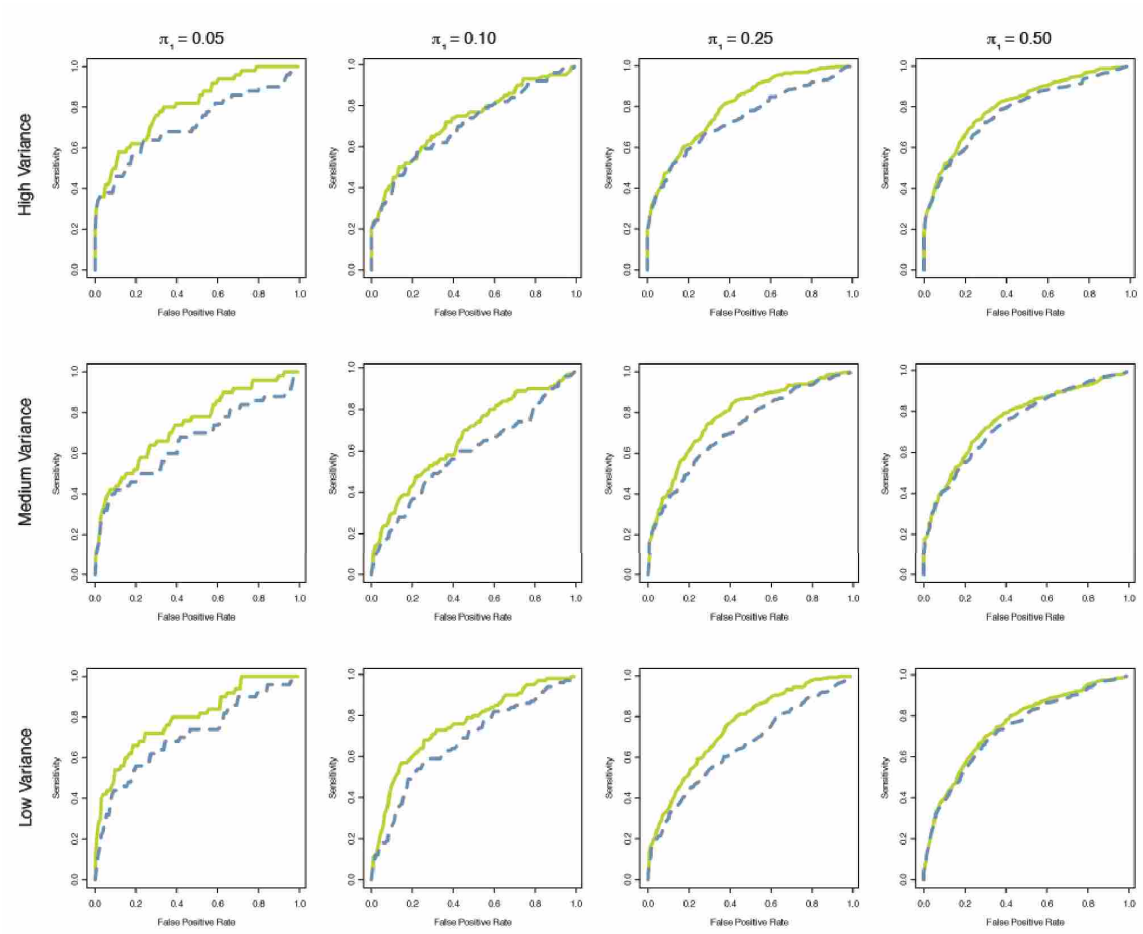


Figure 4.10: Scenario 1 t-distribution with mean 1 and high, medium, and low variance. Each row corresponds to an alternative distribution variance that is high (df=1), medium (df=5), or low (df=10). Each column is a different value of π_1, the proportion of true alternative observations, and increases from 0.05 to 0.50, from left to right. All alternative distributions are 1 null standard deviation away from the null distribution.

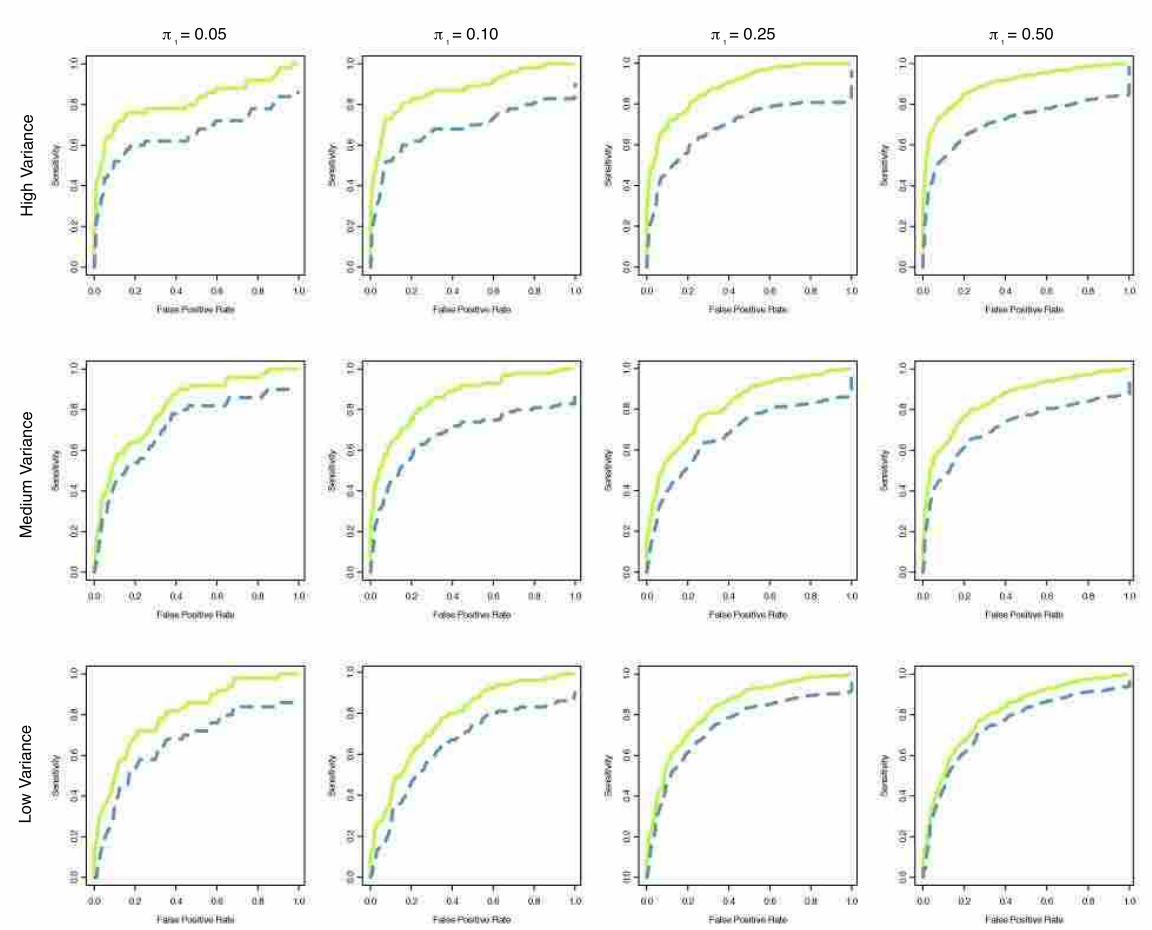


Figure 4.11: Gamma distribution from scenario 2 with mean 1 and normal distribution with mean 2 and high, medium, and low variance. Each row corresponds to an alternative distribution variance that is high (8), medium (4), or low (2). Each column is a different value of π_1, the proportion of true alternative observations, and increases from 0.05 to 0.50, from left to right.

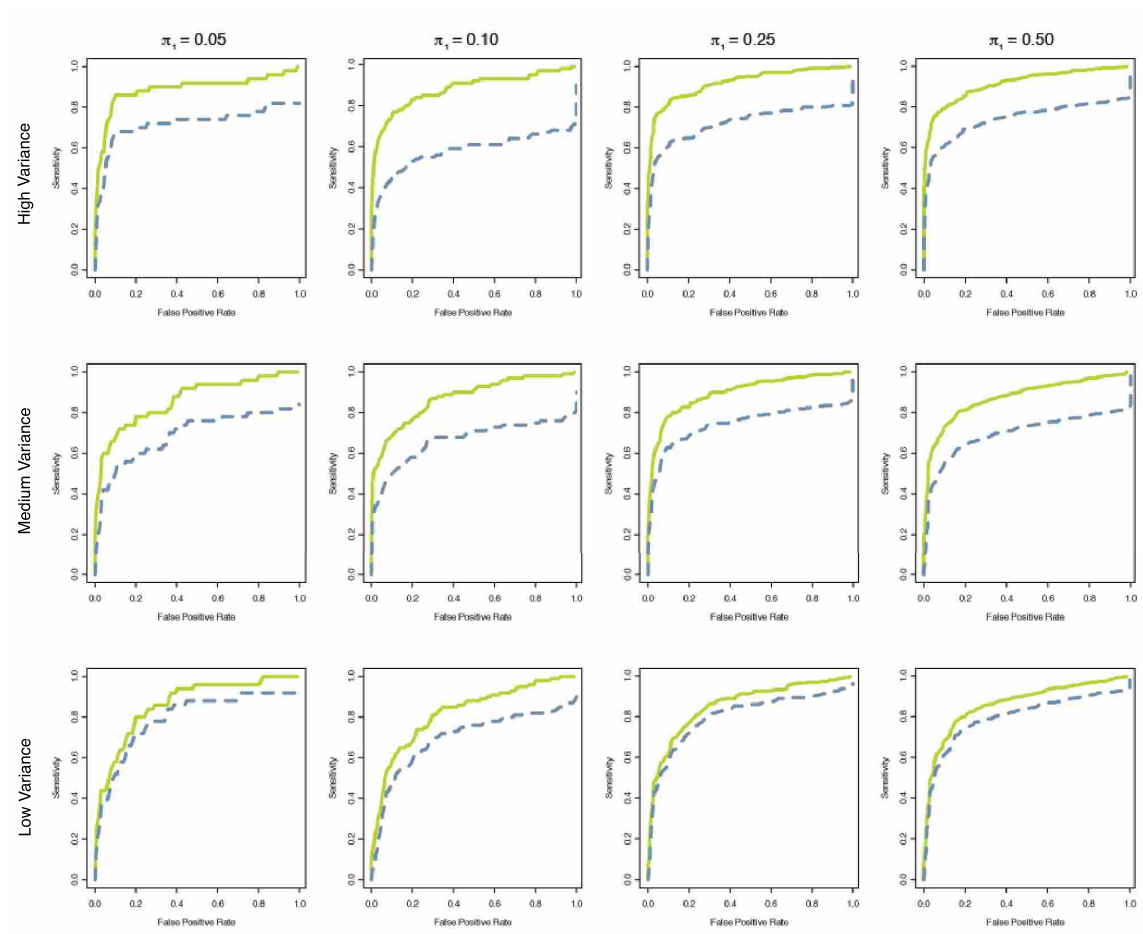


Figure 4.12: Lognormal distribution from scenario 2 with normal mean 2 and high, medium, and low variance. Each row corresponds to an alternative distribution variance that is high (8), medium (4), or low (2). Each column is a different value of π_1, the proportion of true alternative observations, and increases from 0.05 to 0.50, from left to right. All alternative distributions are 1 null standard deviation away from the null distribution.

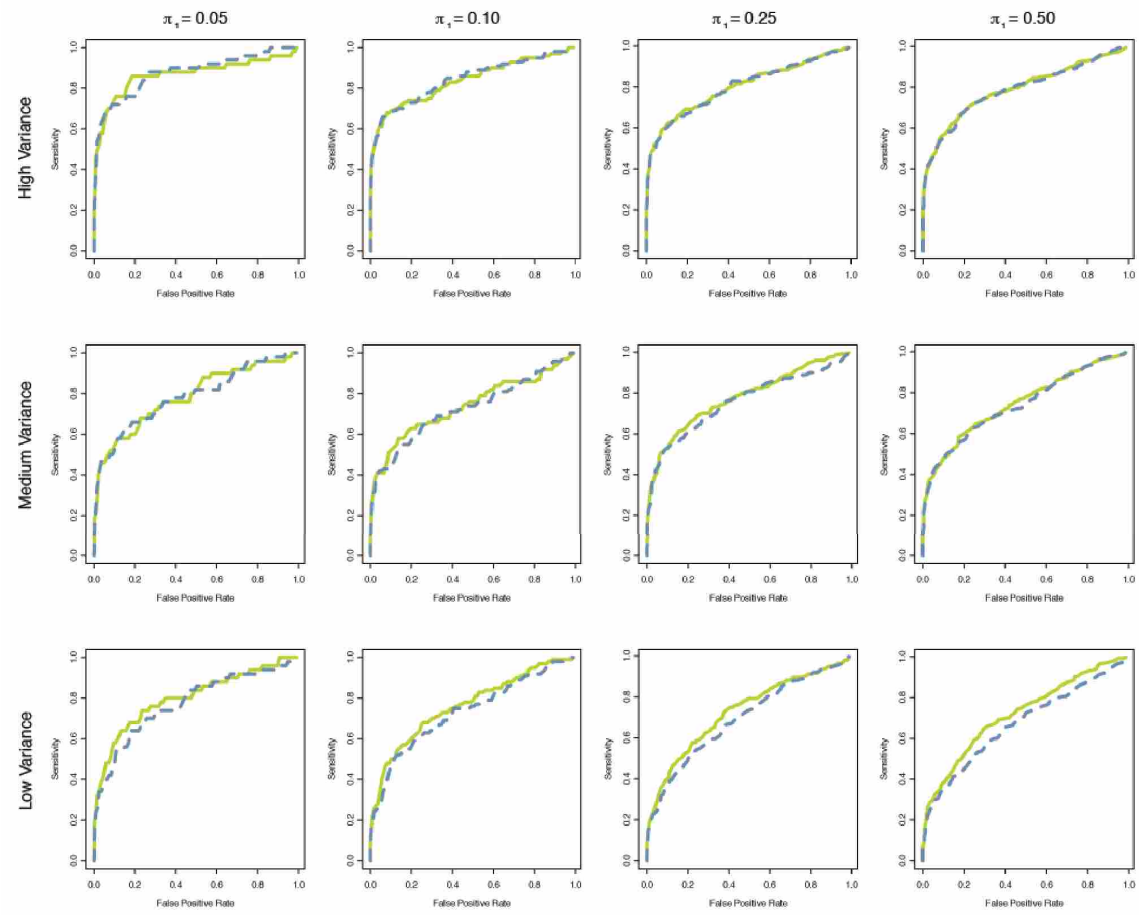


Figure 4.13: Scenario 2 t-distribution compared to a normal with mean 1 and high, medium, and low variance. Each row corresponds to an alternative distribution variance that is high, medium, or low. Each column is a different value of π_1, the proportion of true alternative observations, and increases from 0.05 to 0.50, from left to right. All alternative distributions are 1 null standard deviation away from the null distribution.

Table 4.11: Scaled area differences for scenario 1 using the gamma distribution

		Gamma		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.7125 (0.0087)	0.5994 (0.0104)	0.4454 (0.0106)
	0.10	0.7149 (0.0064)	0.5809 (0.0071)	0.4333 (0.0067)
	0.25	0.6687 (0.0052)	0.53 (0.0053)	0.3663 (0.005)
	0.50	0.5821 (0.0048)	0.4219 (0.0043)	0.2594 (0.0038)
3	0.05	0.1016 (0.0041)	0.0488 (0.0018)	0.0224 (0.0011)
	0.10	0.091 (0.0031)	0.0464 (0.0013)	0.0207 (0.0008)
	0.25	0.0691 (0.0017)	0.0309 (0.0008)	0.0122 (0.0004)
	0.50	0.0357 (0.001)	0.0156 (0.0006)	0.0064 (0.0002)
5	0.05	0.0119 (0.001)	0.0026 (0.0003)	0.0005 (0.0001)
	0.10	0.0112 (0.0008)	0.0022 (0.0002)	0.0003 (0)
	0.25	0.0088 (0.0004)	0.0019 (0.0001)	0.0002 (0)
	0.50	0.0054 (0.0002)	0.0011 (0.0001)	0.0001 (0)

Table 4.12: Scaled area differences for scenario 1 using the lognormal distribution

		Lognormal		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.5323 (0.0098)	0.4561 (0.0103)	0.3687 (0.0094)
	0.10	0.4935 (0.0067)	0.4157 (0.0071)	0.3279 (0.0069)
	0.25	0.4285 (0.0043)	0.3421 (0.0052)	0.2452 (0.0045)
	0.50	0.2998 (0.0037)	0.2282 (0.0044)	0.1438 (0.0034)
3	0.05	0.0657 (0.0024)	0.0346 (0.0018)	0.0166 (0.0008)
	0.10	0.0545 (0.0016)	0.0303 (0.001)	0.0147 (0.0005)
	0.25	0.0396 (0.0012)	0.0191 (0.0007)	0.0086 (0.0004)
	0.50	0.0148 (0.0006)	0.0077 (0.0003)	0.0035 (0.0002)
5	0.05	0.0056 (0.0004)	0.0013 (0.0001)	0.0003 (0)
	0.10	0.0047 (0.0003)	0.001 (0.0001)	0.0002 (0)
	0.25	0.0036 (0.0002)	0.0007 (0.0001)	0.0001 (0)
	0.50	0.002 (0.0001)	0.0004 (0)	0.0001 (0)

Table 4.13: Scaled area differences for scenario 1 using the t-distribution

		t		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.0755 (0.0052)	0.1129 (0.0076)	0.135 (0.0066)
	0.10	0.0617 (0.0037)	0.1042 (0.0053)	0.1179 (0.0044)
	0.25	0.0472 (0.0028)	0.0837 (0.003)	0.0883 (0.003)
	0.50	0.0309 (0.0019)	0.0513 (0.0021)	0.0605 (0.0023)
3	0.05	0.0113 (0.001)	0.0173 (0.0019)	0.0157 (0.0012)
	0.10	0.0081 (0.0012)	0.0164 (0.0011)	0.014 (0.0009)
	0.25	0.0074 (0.0007)	0.0109 (0.0007)	0.0116 (0.0005)
	0.50	0.0022 (0.0004)	0.0062 (0.0004)	0.0068 (0.0003)
5	0.05	0.0012 (0.0008)	0.0016 (0.0004)	0.0005 (0.0002)
	0.10	0.0014 (0.0004)	0.0013 (0.0004)	0.0005 (0.0001)
	0.25	0.0013 (0.0003)	0.0016 (0.0002)	0.0005 (0.0001)
	0.50	0.0013 (0.0003)	0.001 (0.0001)	0.0006 (0.0001)

Table 4.14: Scaled area differences for scenario 2 using the gamma distribution

		Gamma		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.1996 (0.0071)	0.1694 (0.0081)	0.0864 (0.0055)
	0.10	0.2072 (0.0055)	0.1552 (0.0043)	0.0778 (0.0039)
	0.25	0.1768 (0.0035)	0.154 (0.0032)	0.0812 (0.0025)
	0.50	0.153 (0.0021)	0.1424 (0.0019)	0.0779 (0.0018)
3	0.05	0.0692 (0.0056)	0.0208 (0.003)	0.0015 (0.0007)
	0.10	0.0761 (0.0041)	0.0208 (0.0019)	0.0037 (0.001)
	0.25	0.0741 (0.0021)	0.0233 (0.0015)	0.0022 (0.0004)
	0.50	0.0681 (0.0013)	0.0233 (0.0008)	0.0028 (0.0004)
5	0.05	0.0216 (0.0027)	0.002 (0.0009)	0.0001 (0)
	0.10	0.0176 (0.0019)	0.0012 (0.0005)	0 (0)
	0.25	0.0147 (0.0009)	0.0015 (0.0003)	0.0001 (0.0001)
	0.50	0.0151 (0.0007)	0.0012 (0.0002)	0 (0)

Table 4.15: Scaled area differences for scenario 2 using the lognormal distribution

		Lognormal		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.2182 (0.008)	0.1788 (0.009)	0.1007 (0.0059)
	0.10	0.2245 (0.0058)	0.1767 (0.005)	0.0933 (0.004)
	0.25	0.2032 (0.0031)	0.1703 (0.0042)	0.0949 (0.0028)
	0.50	0.1786 (0.0023)	0.1637 (0.0024)	0.098 (0.0022)
3	0.05	0.0818 (0.0053)	0.0258 (0.0032)	0.0029 (0.001)
	0.10	0.075 (0.0034)	0.0269 (0.0029)	0.0018 (0.0006)
	0.25	0.0791 (0.0026)	0.0258 (0.0016)	0.0034 (0.0005)
	0.50	0.079 (0.0016)	0.0275 (0.001)	0.003 (0.0004)
5	0.05	0.0201 (0.0028)	0.0018 (0.0008)	0.0001 (0)
	0.10	0.0223 (0.0018)	0.0018 (0.0005)	0.0001 (0)
	0.25	0.0182 (0.0012)	0.0019 (0.0004)	0.0001 (0.0001)
	0.50	0.0182 (0.001)	0.002 (0.0003)	0.0002 (0.0001)

Table 4.16: Scaled area differences for scenario 2 using the t-distribution

		t		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.0015 (0.0014)	0.0118 (0.0032)	0.065 (0.0071)
	0.10	0.0021 (0.0009)	0.0126 (0.0021)	0.0505 (0.0053)
	0.25	0.0025 (0.0005)	0.0082 (0.0013)	0.0401 (0.004)
	0.50	0.0008 (0.0004)	0.0049 (0.0009)	0.0278 (0.0023)
3	0.05	0.0105 (0.0023)	0.0286 (0.0031)	0.0291 (0.0022)
	0.10	0.0078 (0.0016)	0.0259 (0.002)	0.0255 (0.0013)
	0.25	0.007 (0.0011)	0.0208 (0.0014)	0.0215 (0.0008)
	0.50	0.0044 (0.0007)	0.0127 (0.0009)	0.0131 (0.0005)
5	0.05	0.0101 (0.0018)	0.0079 (0.0009)	0.0017 (0.0003)
	0.10	0.0077 (0.0014)	0.0069 (0.0008)	0.0017 (0.0002)
	0.25	0.0061 (0.0009)	0.0066 (0.0004)	0.0015 (0.0001)
	0.50	0.0057 (0.0007)	0.0053 (0.0003)	0.0012 (0.0001)

Table 4.17: Scaled area differences for scenario 3 using the gamma distribution

		Gamma		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.1559 (0.0097)	0 (0)	0 (0)
	0.10	0.1824 (0.0095)	0 (0)	0 (0)
	0.25	0.1792 (0.0056)	0 (0)	0 (0)
	0.50	0.1825 (0.0046)	0 (0)	0 (0)
3	0.05	0 (0)	0 (0)	0 (0)
	0.10	0 (0)	0 (0)	0 (0)
	0.25	0 (0)	0 (0)	0 (0)
	0.50	0 (0)	0 (0)	0 (0)
5	0.05	0 (0)	0 (0)	0 (0)
	0.10	0 (0)	0 (0)	0 (0)
	0.25	0 (0)	0 (0)	0 (0)
	0.50	0 (0)	0 (0)	0 (0)

Table 4.18: Scaled area differences for scenario 3 using the lognormal distribution

		Lognormal		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.035 (0.0058)	0 (0)	0 (0)
	0.10	0.0392 (0.0053)	0 (0.0001)	0 (0)
	0.25	0.0378 (0.0034)	0 (0.0001)	0 (0)
	0.50	0.0435 (0.0027)	0 (0)	0 (0)
3	0.05	0 (0)	0 (0)	0.0001 (0)
	0.10	0 (0)	0 (0)	0 (0)
	0.25	0 (0)	0 (0)	0 (0)
	0.50	0 (0)	0 (0)	0 (0)
5	0.05	0 (0)	0 (0)	0 (0)
	0.10	0 (0)	0 (0)	0 (0)
	0.25	0 (0)	0 (0)	0 (0)
	0.50	0 (0)	0 (0)	0 (0)

Table 4.19: Scaled area differences for scenario 3 using the t-distribution

		t		
Distance	π_1	High Variance	Medium Variance	Low Variance
1	0.05	0.0691 (0.0067)	0.1104 (0.0064)	0.1308 (0.007)
	0.10	0.0679 (0.0038)	0.1032 (0.0046)	0.1339 (0.0045)
	0.25	0.0527 (0.0024)	0.0823 (0.0026)	0.1016 (0.0036)
	0.50	0.0319 (0.0018)	0.0463 (0.0022)	0.0599 (0.0023)
3	0.05	0.0084 (0.0018)	0.0162 (0.002)	0.0177 (0.0014)
	0.10	0.0107 (0.0012)	0.013 (0.0013)	0.0159 (0.0009)
	0.25	0.0073 (0.0007)	0.0119 (0.0007)	0.0111 (0.0007)
	0.50	0.0038 (0.0004)	0.0064 (0.0004)	0.0068 (0.0005)
5	0.05	0.0026 (0.0007)	0.0014 (0.0007)	0.0012 (0.0003)
	0.10	0.0009 (0.0007)	0.0013 (0.0003)	0.0005 (0.0001)
	0.25	0.0021 (0.0003)	0.0013 (0.0002)	0.0006 (0.0001)
	0.50	0.0013 (0.0002)	0.0012 (0.0002)	0.0005 (0.0001)

Table 4.20: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 1 using the gamma distribution, with their standard errors, based on 50 repeated simulations.

		Gamma					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	37.54 (0.4)	7.22 (0.324)	31.06 (0.447)	9.28 (0.395)	23.36 (0.347)	10.82 (0.428)
	0.10	69.5 (0.502)	14.56 (0.398)	56.92 (0.565)	18.14 (0.561)	41.72 (0.571)	20.24 (0.53)
	0.25	99.7 (0.082)	37.16 (0.773)	94.18 (0.394)	44.7 (0.793)	76.42 (0.633)	48.88 (0.668)
	0.50	100 (0)	70.26 (0.778)	99.84 (0.066)	80.06 (0.824)	95.56 (0.239)	84.1 (0.579)
3	0.05	32 (0.456)	28.76 (0.537)	37.9 (0.362)	34.02 (0.416)	43.3 (0.33)	40.02 (0.418)
	0.10	58.68 (0.669)	54.98 (0.708)	67.14 (0.615)	62.94 (0.635)	76.4 (0.469)	72.44 (0.485)
	0.25	98.18 (0.213)	97.9 (0.222)	99.1 (0.132)	98.92 (0.13)	99.48 (0.1)	99.48 (0.1)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
5	0.05	46.52 (0.264)	44.78 (0.299)	49.28 (0.118)	48.74 (0.156)	49.94 (0.034)	49.84 (0.052)
	0.10	85.86 (0.366)	83.62 (0.41)	93.7 (0.225)	92.14 (0.345)	96.92 (0.187)	96.32 (0.207)
	0.25	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

Table 4.21: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 1 using the lognormal distribution, with their standard errors, based on 50 repeated simulations.

		Lognormal					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	21.76 (0.403)	7.12 (0.336)	18.12 (0.543)	7.26 (0.343)	15.1 (0.461)	7.4 (0.326)
	0.10	40.24 (0.551)	15.2 (0.42)	33.02 (0.523)	16.24 (0.48)	29.08 (0.491)	15.68 (0.506)
	0.25	70.68 (0.559)	36.32 (0.637)	63.44 (0.555)	39.52 (0.689)	57.02 (0.806)	42.56 (0.703)
	0.50	90.02 (0.362)	69.06 (0.715)	86.84 (0.532)	73.68 (0.677)	82.52 (0.599)	74.96 (0.685)
3	0.05	34.36 (0.409)	30.28 (0.454)	39.06 (0.416)	34.82 (0.457)	44.68 (0.283)	40.98 (0.302)
	0.10	61.92 (0.548)	57.5 (0.566)	69.04 (0.405)	64.1 (0.461)	77.44 (0.487)	72.84 (0.546)
	0.25	97.54 (0.277)	97.22 (0.297)	98.94 (0.123)	98.86 (0.121)	98.98 (0.158)	98.92 (0.1710)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	99.96 (0.028)	99.96 (0.028)
5	0.05	48.38 (0.189)	46.8 (0.249)	49.8 (0.064)	49.42 (0.103)	50 (0)	50 (0)
	0.10	89.58 (0.304)	87.14 (0.348)	94.72 (0.239)	93.28 (0.277)	97.5 (0.141)	96.96 (0.156)
	0.25	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

Table 4.22: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 1 using the t-distribution, with their standard errors, based on 50 repeated simulations.

		t					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	21.78 (0.457)	20.2 (0.417)	18.9 (0.456)	15.16 (0.443)	17.14 (0.381)	13.06 (0.349)
	0.10	41.98 (0.643)	39.02 (0.603)	33.86 (0.642)	28.66 (0.603)	33.78 (0.54)	26.98 (0.502)
	0.25	82.84 (0.627)	81.44 (0.647)	65.34 (0.596)	60.7 (0.652)	64.28 (0.559)	58.76 (0.651)
	0.50	99.6 (0.09)	99.58 (0.091)	89.48 (0.47)	88.86 (0.477)	88.86 (0.37)	87.4 (0.408)
3	0.05	42.72 (0.365)	41.74 (0.355)	44.82 (0.257)	42.86 (0.349)	45.3 (0.258)	43.24 (0.3)
	0.10	78.64 (0.464)	76.66 (0.439)	80.64 (0.413)	77.52 (0.44)	81.56 (0.415)	78.18 (0.405)
	0.25	99.78 (0.072)	99.76 (0.073)	99.78 (0.059)	99.72 (0.07)	99.46 (0.096)	99.4 (0.103)
	0.50	100 (0)	100 (0)	99.92 (0.039)	99.92 (0.039)	99.98 (0.02)	99.98 (0.02)
5	0.05	47.48 (0.196)	47.32 (0.226)	49.48 (0.096)	49.34 (0.12)	49.94 (0.034)	49.82 (0.055)
	0.10	92.4 (0.306)	91.88 (0.306)	96.92 (0.169)	96.32 (0.175)	97.38 (0.164)	96.98 (0.168)
	0.25	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

Table 4.23: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 2 using the gamma distribution, with their standard errors, based on 50 repeated simulations.

		Gamma					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	32.36 (0.373)	22.28 (0.388)	27.5 (0.46)	19.76 (0.45)	22.58 (0.518)	18.66 (0.439)
	0.10	60.2 (0.525)	41.74 (0.613)	49.58 (0.631)	36.96 (0.618)	38.88 (0.515)	32.9 (0.511)
	0.25	94.78 (0.322)	76.88 (0.668)	84.86 (0.513)	69.26 (0.697)	70.22 (0.654)	60.42 (0.621)
	0.50	100 (0)	94.46 (0.292)	98.02 (0.217)	89.16 (0.42)	90.38 (0.456)	83.08 (0.581)
3	0.05	37.3 (0.465)	33.98 (0.506)	38.24 (0.364)	37.36 (0.395)	40.58 (0.314)	40.5 (0.315)
	0.10	67.96 (0.459)	62.6 (0.496)	66.28 (0.492)	64.9 (0.52)	68.9 (0.573)	68.62 (0.573)
	0.25	96.8 (0.239)	94.6 (0.359)	94.2 (0.325)	93.3 (0.336)	91.72 (0.447)	91.54 (0.438)
	0.50	99.72 (0.076)	99.14 (0.134)	98.66 (0.155)	98.56 (0.165)	97.32 (0.147)	97.16 (0.141)
5	0.05	44.48 (0.271)	43.5 (0.261)	47.1 (0.214)	47.02 (0.226)	49.28 (0.121)	49.28 (0.121)
	0.10	79.24 (0.426)	77.94 (0.439)	84.06 (0.396)	83.98 (0.397)	88.24 (0.334)	88.24 (0.334)
	0.25	99.2 (0.121)	99.14 (0.128)	98.98 (0.147)	99 (0.146)	99.02 (0.132)	99 (0.14)
	0.50	99.9 (0.052)	99.88 (0.055)	99.62 (0.075)	99.62 (0.075)	99.66 (0.079)	99.66 (0.079)

Table 4.24: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 2 using the lognormal distribution, with their standard errors, based on 50 repeated simulations.

		Lognormal					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	37.32 (0.405)	27.08 (0.43)	33.3 (0.498)	25.44 (0.522)	30.14 (0.446)	26.04 (0.51)
	0.10	68.22 (0.458)	50.24 (0.528)	61.24 (0.526)	49 (0.46)	53.84 (0.498)	47.82 (0.507)
	0.25	98.06 (0.238)	88.46 (0.452)	92.54 (0.416)	82.54 (0.648)	85.14 (0.467)	78.72 (0.521)
	0.50	100 (0)	97.8 (0.202)	99.52 (0.096)	95.68 (0.241)	96.92 (0.226)	93.38 (0.342)
3	0.05	41.2 (0.383)	37.42 (0.39)	43.24 (0.328)	42.16 (0.344)	45.42 (0.289)	45.32 (0.293)
	0.10	75.58 (0.372)	70.32 (0.437)	75.92 (0.446)	74.16 (0.494)	79.46 (0.405)	79.34 (0.4)
	0.25	98.74 (0.136)	97.82 (0.209)	97.98 (0.15)	97.74 (0.159)	97.26 (0.242)	96.94 (0.259)
	0.50	99.82 (0.068)	99.5 (0.096)	99.44 (0.086)	99.2 (0.131)	99.18 (0.13)	99.04 (0.134)
5	0.05	46.86 (0.254)	45.92 (0.271)	48.86 (0.14)	48.78 (0.152)	49.88 (0.046)	49.88 (0.046)
	0.10	86.4 (0.305)	84.8 (0.318)	89.58 (0.309)	89.54 (0.309)	93.4 (0.306)	93.38 (0.305)
	0.25	99.7 (0.087)	99.58 (0.091)	99.18 (0.156)	99.16 (0.155)	99.4 (0.118)	99.44 (0.118)
	0.50	99.9 (0.052)	99.9 (0.052)	99.88 (0.046)	99.78 (0.072)	99.96 (0.028)	99.82 (0.055)

Table 4.25: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 2 using the t-distribution, with their standard errors, based on 50 repeated simulations.

		t					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	27.9 (0.443)	28 (0.455)	23.04 (0.438)	22.34 (0.479)	20.12 (0.484)	16.98 (0.388)
	0.10	52.24 (0.624)	51.86 (0.619)	43.52 (0.49)	42.52 (0.567)	36.64 (0.526)	32.14 (0.569)
	0.25	93.64 (0.439)	93.58 (0.415)	82.38 (0.611)	81.16 (0.528)	71.64 (0.555)	67.24 (0.631)
	0.50	99.96 (0.028)	99.96 (0.028)	98.06 (0.203)	97.9 (0.214)	93.52 (0.353)	92.62 (0.415)
3	0.05	36.18 (0.407)	35.52 (0.45)	38.16 (0.426)	36.42 (0.389)	42.3 (0.351)	39.9 (0.43)
	0.10	67.64 (0.632)	66.14 (0.64)	72.02 (0.517)	69.24 (0.445)	77.14 (0.434)	73.9 (0.455)
	0.25	99.84 (0.052)	99.86 (0.05)	99.76 (0.067)	99.74 (0.069)	99.54 (0.096)	99.52 (0.096)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
5	0.05	44.28 (0.287)	43.48 (0.293)	47.84 (0.165)	47.02 (0.213)	49.58 (0.095)	49.28 (0.134)
	0.10	84.86 (0.394)	83.56 (0.46)	90.98 (0.318)	89.26 (0.383)	95.88 (0.221)	94.82 (0.237)
	0.25	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

Table 4.26: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 3 using the gamma distribution, with their standard errors, based on 50 repeated simulations.

		Gamma					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	16.48 (0.47)	13.28 (0.495)	14.84 (0.396)	14.84 (0.396)	15.06 (0.474)	15.06 (0.474)
	0.10	31.16 (0.534)	26.32 (0.623)	28.32 (0.567)	28.32 (0.567)	27.84 (0.608)	27.84 (0.608)
	0.25	66.12 (0.696)	57.6 (0.71)	59.66 (0.702)	59.66 (0.702)	56.48 (0.583)	56.48 (0.583)
	0.50	90.96 (0.506)	86.62 (0.554)	85.36 (0.612)	85.36 (0.612)	81.02 (0.503)	81.02 (0.503)
3	0.05	30.12 (0.459)	30.12 (0.459)	35.74 (0.421)	35.74 (0.421)	40.34 (0.387)	40.34 (0.387)
	0.10	56.02 (0.609)	56.02 (0.609)	62.82 (0.527)	62.82 (0.527)	66.48 (0.562)	66.48 (0.562)
	0.25	90.68 (0.365)	90.68 (0.365)	91.28 (0.364)	91.28 (0.364)	90.22 (0.398)	90.24 (0.4)
	0.50	98.48 (0.177)	98.48 (0.177)	98.08 (0.223)	98.08 (0.223)	97.2 (0.216)	97.22 (0.216)
5	0.05	44.52 (0.342)	44.52 (0.342)	48.64 (0.151)	48.64 (0.151)	49.86 (0.057)	49.86 (0.057)
	0.10	77.76 (0.392)	77.76 (0.392)	84.04 (0.352)	84.04 (0.352)	89.36 (0.31)	89.36 (0.31)
	0.25	98.32 (0.172)	98.32 (0.172)	98.4 (0.185)	98.4 (0.185)	98.56 (0.167)	98.56 (0.167)
	0.50	99.82 (0.062)	99.82 (0.062)	99.68 (0.083)	99.68 (0.083)	99.46 (0.104)	99.44 (0.104)

Table 4.27: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 3 using the lognormal distribution, with their standard errors, based on 50 repeated simulations.

		Lognormal					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	17.8 (0.529)	17.6 (0.558)	18.8 (0.537)	18.8 (0.537)	20.68 (0.459)	20.68 (0.459)
	0.10	32.26 (0.587)	31.72 (0.586)	34.4 (0.629)	34.4 (0.629)	38.24 (0.622)	38.24 (0.622)
	0.25	65.02 (0.631)	64.06 (0.6)	68.02 (0.771)	68.02 (0.771)	69.94 (0.731)	69.94 (0.731)
	0.50	90.56 (0.359)	90.3 (0.368)	91.08 (0.396)	91.08 (0.396)	89.76 (0.365)	89.76 (0.365)
3	0.05	39.16 (0.372)	39.16 (0.372)	44.54 (0.303)	44.54 (0.303)	47.52 (0.216)	47.52 (0.216)
	0.10	68.12 (0.504)	68.12 (0.504)	73.66 (0.439)	73.66 (0.439)	80.06 (0.448)	80.06 (0.448)
	0.25	94.94 (0.339)	94.94 (0.339)	95.86 (0.262)	95.86 (0.262)	95.82 (0.25)	95.84 (0.25)
	0.50	99.16 (0.126)	99.16 (0.126)	98.88 (0.123)	98.88 (0.123)	98.82 (0.148)	98.86 (0.148)
5	0.05	48.9 (0.155)	48.9 (0.155)	49.88 (0.055)	49.88 (0.055)	50 (0)	50 (0)
	0.10	87.58 (0.413)	87.58 (0.413)	91.68 (0.264)	91.68 (0.264)	94.82 (0.31)	94.82 (0.31)
	0.25	99.24 (0.136)	99.24 (0.136)	99.2 (0.148)	99.18 (0.153)	99.38 (0.11)	99.38 (0.117)
	0.50	99.94 (0.034)	99.94 (0.034)	99.82 (0.068)	99.84 (0.066)	99.88 (0.055)	99.84 (0.052)

Table 4.28: True positives out of the top 100 observations for the true and estimated ODP functions in scenario 3 using the t-distribution, with their standard errors, based on 50 repeated simulations.

		t					
		High Variance		Medium Variance		Low Variance	
Distance	π_1	True	Estimated	True	Estimated	True	Estimated
1	0.05	23.14 (0.474)	20.92 (0.417)	18.16 (0.435)	14.26 (0.445)	17.8 (0.361)	14.06 (0.384)
	0.10	42.14 (0.607)	39.1 (0.595)	35.08 (0.642)	28.94 (0.577)	33.56 (0.525)	26.84 (0.586)
	0.25	80.38 (0.565)	79.46 (0.542)	64.86 (0.703)	60.2 (0.75)	63.2 (0.775)	57.44 (0.726)
	0.50	99.5 (0.115)	99.48 (0.119)	89.12 (0.439)	88.02 (0.386)	86.54 (0.468)	85.68 (0.485)
3	0.05	43.58 (0.283)	42.26 (0.307)	44.52 (0.331)	42.32 (0.373)	45.52 (0.292)	43.04 (0.356)
	0.10	79.14 (0.469)	77.32 (0.52)	80.28 (0.457)	77.42 (0.496)	81.42 (0.424)	78.48 (0.436)
	0.25	99.78 (0.072)	99.78 (0.072)	99.48 (0.091)	99.48 (0.091)	99.3 (0.115)	99.24 (0.123)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	99.98 (0.02)	99.98 (0.02)
5	0.05	47.32 (0.22)	47 (0.219)	49.52 (0.096)	49.28 (0.114)	49.72 (0.076)	49.64 (0.085)
	0.10	92.08 (0.321)	91.54 (0.327)	96.56 (0.214)	95.92 (0.225)	97.28 (0.194)	96.82 (0.193)
	0.25	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	0.50	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

Table 4.29: True Positive Rate and False Positive Rate for scenario 1 using the gamma distribution

			Gamma					
			High Variance		Medium Variance		Low Variance	
Distance	π_1	Value	TPR	FPR	TPR	FPR	TPR	FPR
1	0.05	True	0.751	0.066	0.621	0.073	0.467	0.081
			(0.008)	(0.00042)	(0.009)	(0.00047)	(0.007)	(0.00037)
		Est	0.144	0.098	0.186	0.095	0.216	0.094
			(0.006)	(0.00034)	(0.008)	(0.00042)	(0.009)	(0.00045)
	0.10	True	0.695	0.034	0.569	0.048	0.417	0.065
			(0.005)	(0.00056)	(0.006)	(0.00063)	(0.006)	(0.00063)
		Est	0.146	0.095	0.181	0.091	0.202	0.089
			0.004)	(0.00044)	(0.006)	(0.00062)	(0.005)	(0.00059)
	0.25	True	0.399	0	0.377	0.008	0.306	0.031
			(0)	(0.00011)	(0.002)	(0.00052)	(0.003)	(0.00084)
		Est	0.149	0.08	0.179	0.074	0.196	0.068
			(0.003)	(0.00103)	(0.003)	(0.00106)	(0.003)	(0.00089)
	0.50	True	0.2	0	0.2	0	0.191	0.009
			(0)	(0)	(0)	(0.00013)	(0)	(0.00048)
		Est	0.141	0.059	0.16	0.04	0.168	0.032
			(0.002)	(0.00156)	(0.002)	(0.00165)	(0.001)	(0.00116)
3	0.05	True	0.64	0.072	0.758	0.065	0.866	0.06
			(0.009)	(0.00048)	(0.007)	(0.00038)	(0.007)	(0.00035)
		Est	0.575	0.075	0.68	0.069	0.8	0.063
			(0.011)	(0.00057)	(0.008)	(0.00044)	(0.008)	(0.00044)
	0.10	True	0.587	0.046	0.671	0.037	0.764	0.026
			(0.007)	(0.00074)	(0.006)	(0.00068)	(0.005)	(0.00052)
		Est	0.55	0.05	0.629	0.041	0.724	0.031
			(0.007)	(0.00079)	(0.006)	(0.00071)	(0.005)	(0.00054)
	0.25	True	0.393	0.002	0.396	0.001	0.398	0.001
			(0.001)	(0.00028)	(0.001)	(0.00018)	(0)	(0.00013)
		Est	0.392	0.003	0.396	0.001	0.398	0.001
			(0.001)	(0.0003)	(0.001)	(0.00017)	(0)	(0.00013)
	0.50	True	0.2	0	0.2	0	0.2	0
			(0)	(0)	(0)	(0)	(0)	(0)
		Est	0.2	0	0.2	0	0.2	0
			(0)	(0)	(0)	(0)	(0)	(0)
5	0.05	True	0.93	0.056	0.986	0.053	0.999	0.053
			(0.005)	(0.00028)	(0.002)	(0.00012)	(0.001)	(0.00004)
		Est	0.896	0.058	0.975	0.05	0.997	0.053
			(0.006)	(0.00031)	(0.003)	(0.00016)	(0.001)	(0.00006)
	0.10	True	0.859	0.016	0.937	0.007	0.969	0.003
			(0.004)	(0.00041)	0.002)	(0.00025)	(0.002)	(0.00021)
		Est	0.836	0.018	0.921	0.009	0.963	0.004
			(0.004)	(0.00046)	(0.003)	(0.00038)	(0.002)	(0.00023)
	0.25	True	0.4	0	0.4	0	0.4	0
			(0)	(0)	(0)	(0)	(0)	(0)
		Est	0.4	0	0.4	0	0.4	0
			(0)	(0)	(0)	(0)	(0)	(0)
	0.50	True	0.2	0	0.2	0	0.2	0
			(0)	(0)	(0)	(0)	(0)	(0)
		Est	0.2	0	0.2	0	0.2	0
			(0)	(0)	(0)	(0)	(0)	(0)

5. CONCLUSIONS

Determining the importance of the normality assumption when estimating the ODP with nonnormal observations is crucial to the widespread applicability of the ODP as a method for determining significance. Additionally, situations in which the estimated ODP fails to adequately represent the information in the true ODP should be identified in order to make recommendations about the use of the ODP estimation method. The results of the previous chapter can be summarized with plots of the scaled area differences for the gamma, lognormal, and t-distributions shown in Figures 5.1, 5.2 and 5.3, respectively. There are three general trends that can be observed from these figures and the results in the previous chapter. First, the estimation of the ODP improves as π_1 increases. This observation is plausible because more strength is borrowed between alternative observations as the proportion of alternative observations grows. Second, estimation of the ODP improves as distance between the centers of the null and alternative distributions increases. Because this method aims to identify significant observations, it is easier to accomplish this goal when the underlying distributions are more distinct. Third, the ODP is better estimated when the variance of the alternative distribution is smaller. Alternative observations which fall in a tighter cluster are easier to distinguish from null data.

As shown in this simulation study, differences between the true and estimated ODP are not problematic for all simulations when the null and alternative distributions are at least 3 null standard deviations apart. Alternative distributions with low variance are preferable to high variance, and a larger value of π_1 for a given distance between distributions corresponds to better ODP estimation. Comparing Figures 5.1, 5.2, and 5.3, it is readily apparent that potential problems exist when this estimation method is used for skewed alternative observations which are a small distance away from the null distribution.

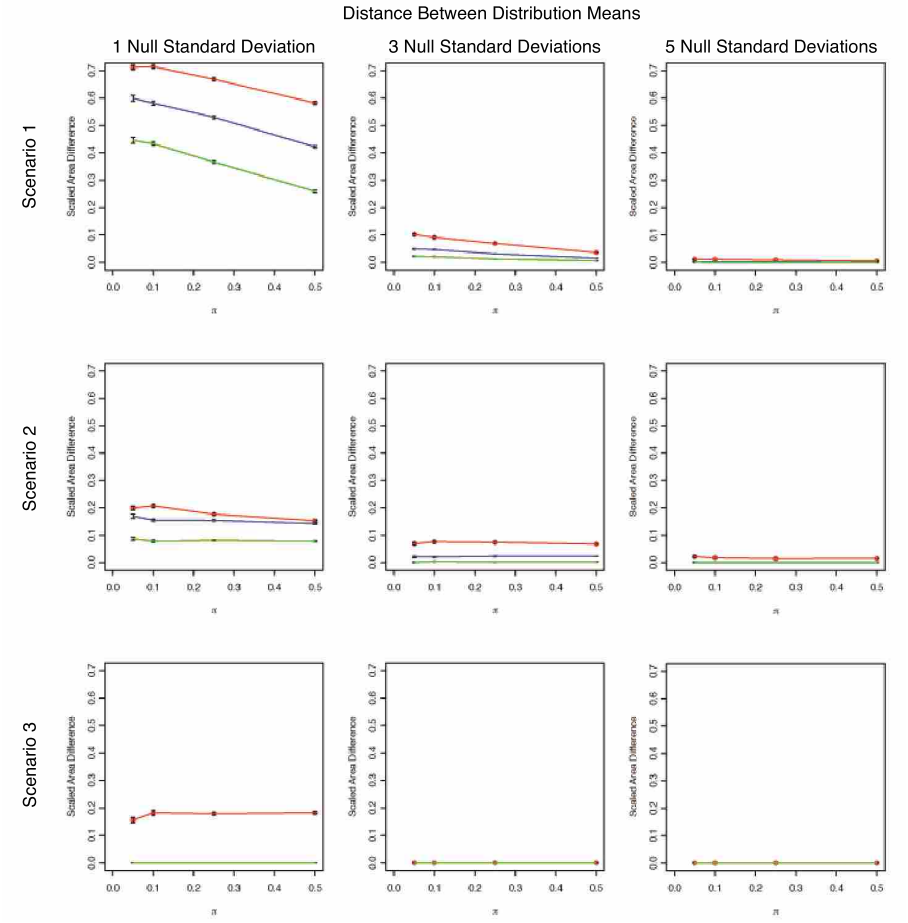


Figure 5.1: Scaled area differences for the gamma distribution. Each plot displays π_1 vs. scaled area difference. The trendlines within the plot represent the three different variance levels for the alternative distribution—high variance, shown in red; medium variance, shown in blue; and low variance, shown in green—with standard error bars for each point.

5.1 Recommendations

Based on the results of the simulation study performed, the proposed ODP estimation method should not be used on data which have skewed alternative observations when the distance between expectations is one null standard deviation. Estimation is considered satisfactory for all observed situations with the t-distribution, and, surprisingly, for most simulations in which both the null and alternative distributions are nonnormal.

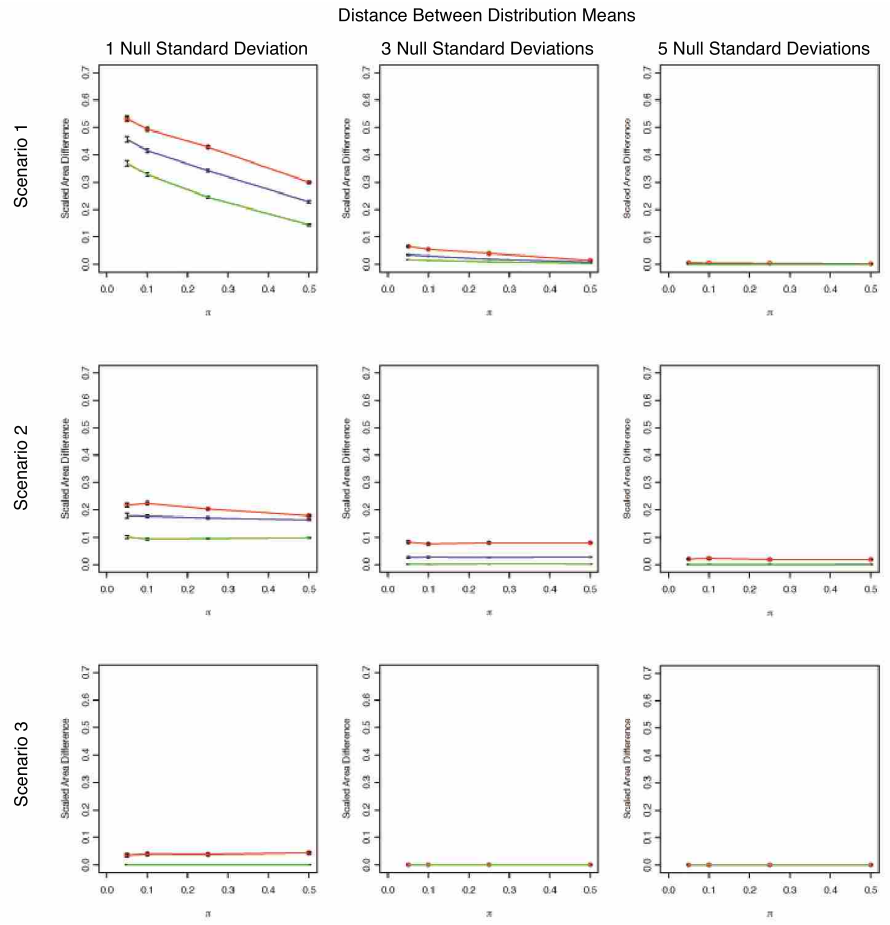


Figure 5.2: Scaled area differences for the lognormal distribution. Each plot displays π_1 vs. scaled area difference. The trendlines within the plot represent the three different variance levels for the alternative distribution—high variance, shown in red; medium variance, shown in blue; and low variance, shown in green—with standard error bars for each point.

5.2 Future Research

Future exploration of ODP methodology could exist on two levels—performing additional simulations and adjusting the current estimation methodology. With regards to simulation extensions, there are three general areas of exploration: factor settings, assumptions, and methods.

Factor settings refers to the levels of factors chosen for this set of simulations. That is, the values of π_1 or the number of null standard deviations separating the

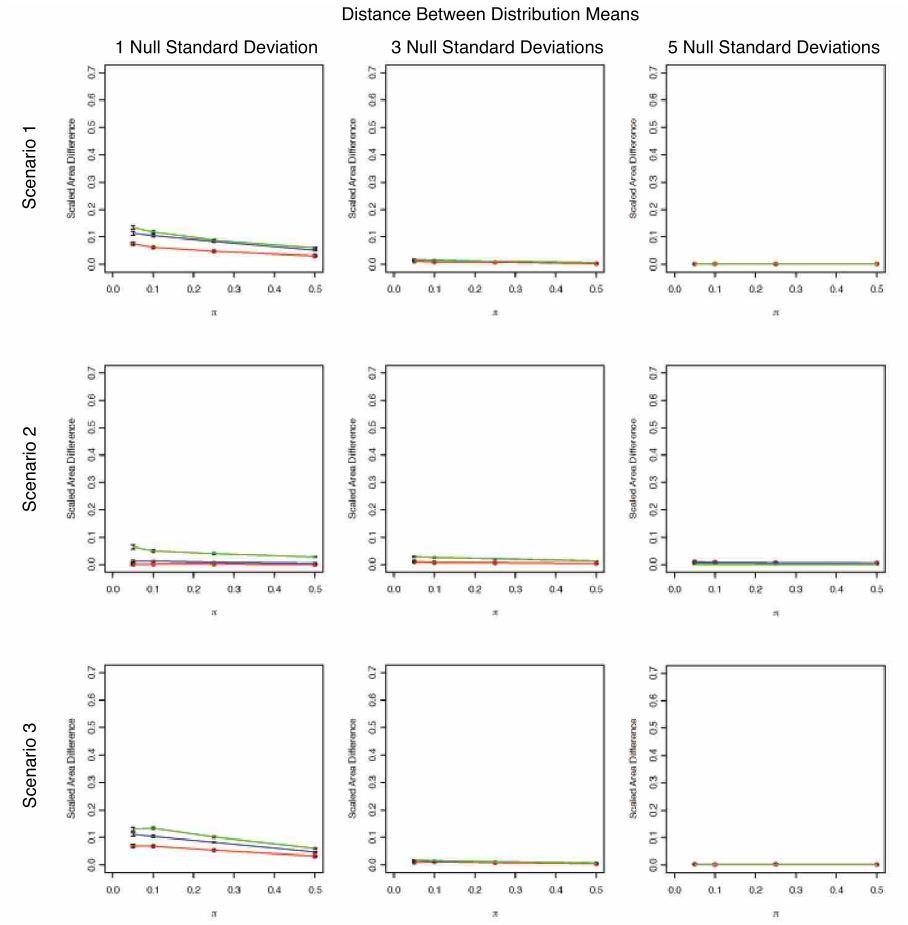


Figure 5.3: Scaled area differences for the t-distribution. Each plot displays π_1 vs. scaled area difference. The trendlines within the plot represent the three different variance levels for the alternative distribution—high variance, shown in red; medium variance, shown in blue; and low variance, shown in green—with standard error bars for each point.

null and alternative distributions could be specified differently from those settings used here. Similarly, the resulting nonnormal parameterizations or even the distributional families themselves could be modified to explore the behavior of other types of nonnormal data in the context of the ODP.

The simulations in this study were run with the requirement that all null distributions have unit variance. This assumption was imposed to establish a degree of comparability across scenarios and nonnormal distributions. If this requirement were

relaxed, or if the null variance were held constant at some other value, a greater range of possibilities could be observed and the performance of the estimated ODP could be evaluated.

The methods used in the construction and estimation of the ODP are consistent with the information in Storey (2005); however, as discussed in the previous chapter, the methodology for incorporating nonnormal null distributions into scenarios 2 and 3 had to be inferred from general theoretical information. In the estimation of the ODP in those scenarios, a standard normal density was placed in the denominator of the score statistic for all simulations. Various other methods exist for estimating the ODP, and this simulation study may be conducted using one of those methods instead of the method used here. Also, the observations were generated by two distributions, one null and one alternative, but could also reasonably consist of mixture distributions instead. Another methodological variation that could be used to evaluate the ODP estimation would be to transform the nonnormal observations (using Box-Cox or other means) to appear more normally distributed before performing the simulations.

Finally, the results of this simulation study may be used to develop a new ODP estimation method. Because the major ODP estimation difficulties occurred with skewed alternative observations, this new estimation method should adapt to accomodate nonnormal data, perhaps by replacing the normal densities with estimated nonnormal densities.

BIBLIOGRAPHY

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Bilban, M., Buehler, L. K., Head, S., Desoye, G., and Quaranta, V. (2002), "Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer," *BMC Genomics*, 3.

Broberg, P. (2003), "Statistical methods for ranking differentially expressed genes," *Genome Biology*, 4, R41.

Diaconis, P. (1985), "Theories of Data Analysis: From Magical Thinking through Classical Statistics," in *Exploring Data Tables, Trends, and Shapes*, eds. Hoaglin, D. C., Mosteller, F., and Tukey, J. W., John Wiley & Sons, pp. 1–36.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Oliver & Boyd.

Giles, P. and Kipling, D. (2003), "Normality of oligonucleotide microarray data and implications for parametric statistical analyses," *Bioinformatics*, 19, 2254–2262.

Harvey, D., Weng, Q., Fletcher, E., DeCarli, C., and Beckett, L. (2006), "A Central Limit Theorem for High-Dimensional Spatially Correlated Processes," in *American Statistical Association Joint Statistical Meetings*, Seattle, WA.

Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, John Wiley & Sons.

Hsu, J. C. (1996), *Multiple Comparisons: Theory and methods*, Chapman & Hall.

Jensen, A. L., Thofner, M. T., and Iverasen, L. (1996), "Application of receiver-operating-characteristic (ROC) curves to veterinary clinical pathology," *Comparative Haematology International*, 6, 176–181.

Ji, Y., Tsui, K.-W., and Kim, K. M. (2006), "A two-stage empirical Bayes method for identifying differentially expressed genes," *Computational Statistics and Data Analysis*, 50, 3592–3604.

Konishi, T. (2004), "Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment," *BMC Bioinformatics*, 5.

Müller, P., Parmigiani, G., and Rice, K. (2006), "FDR and Bayesian Multiple Comparisons Rules," Working Paper 115, Johns Hopkins Department of Biostatistics.

Neyman, J. and Pearson, E. (1928a), "On the Use and Interpretation of Certain Test Criteria for Purpose of Statistical Inference, Part I," *Biometrika*, 20A, 175–240.

— (1928b), "On the Use and Interpretation of Certain Test Criteria for Purpose of Statistical Inference, Part II," *Biometrika*, 20A, 263–294.

Royall, R. (1997), *Statistical Evidence: A likelihood paradigm*, Chapman & Hall.

Shaffer, J. P. (1995), "Multiple Hypothesis Testing," *Annual Review Psychology*, 46, 561–584.

Slonim, D. K. (2002), "From patterns to pathways: gene expression data analysis comes of age," *Nature Genetics Supplement*, 32, 502–508.

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty before 1900*, The Belknap Press of Harvard University Press.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-value," *The Annals of Statistics*, 31, 2013–2035.

— (2005), "The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing," Working Paper 259, University of Washington Department of Biostatistics.

Storey, J. D., Dai, J. Y., and Leek, J. T. (2005a), "The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments," Working Paper 260, University of Washington Department of Biostatistics.

Storey, J. D., Taylor, J., and Siegmund, D. (2004), "Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach," *Journal of the Royal Statistical Society, Series B*, 66, 187–205.

Storey, J. D. and Tibshirani, R. (2003), "Statistical significance for genomewide experiments," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005b), "Significance analysis of time course microarray experiments," *Proceedings of the National Academy of Sciences*, 102, 12837–12842.

Tukey, J. W. (1977), "Some thoughts on clinical trials, especially problems of multiplicity," *Science*, 198, 679–684.

— (1994), *The Collected Works of John W. Tukey*, vol. 8: Multiple Comparisons: 1948-1983, Chapman & Hall.

Wilson, L. J. (2000), "Comments on "Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System"," *Weather and Forecasting*, 15, 361–364.

Zhang, J.-G., Yin, Z.-J., and Zhang, Q. (2006), "A Non-transformation Method for Identifying Differentially Expressed Genes from cDNA Microarrays," *Acta Genetica Sinica*, 33, 80–88.