Hypothesis tests can be categorized as rejection-support tests or acceptance-support tests. Type I error in a rejection-support test has the same effect as type II error in an acceptance-support test. Most hypothesis testing is based on type I error control, but it increases type II error. Similar to how researchers control type I error when performing multiple rejection tests, researchers might want to control type II error in multiple acceptance-support testing. In this thesis, I discuss various methods to control decision errors in multiple testing. As an example, I conduct multiple normality tests with four different normality formulae under four different alternative hypotheses. I discuss estimation of the ratio between the null hypothesis and non-discovery rate, and also balance the trade-off between false discovery rate and non-discovery rate to find more cost-effective critical points. I apply this discussion to real data about breast cancer.

INDEX WORDS: Acceptance support test, Decision error, False discovery rate, Hypothesis test, Multiple testing, Non-discovery rate, Normality test, Rejection support test, Type I error, Type II error
FALSE NEGATIVE CONTROL FOR MULTIPLE ACCEPTANCE-SUPPORT HYPOTHESES TESTING PROBLEM

by

SOYEON JUNG

B.S., Seoul National University, 2007
M.S., Seoul National University, 2010

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015
FALSE NEGATIVE CONTROL FOR MULTIPLE ACCEPTANCE-SUPPORT HYPOTHESES TESTING PROBLEM

by

SOYEON JUNG

Approved:

Major Professors:  Jeongyoun Ahn

Committee:  Jaxk Reeves
            Pengsheng Ji

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015
# Contents

1 Two Types of Hypothesis Tests
   1.1 Rejection-Support Tests ........................................... 2
   1.2 Acceptance-Support Tests ......................................... 3
   1.3 Decision Errors in Single Hypothesis Test ....................... 4

2 Multiple Testing
   2.1 Decision Errors ..................................................... 8
   2.2 Control of Errors .................................................. 12

3 Multiple Normality Testing
   3.1 Normality Tests .................................................... 26
   3.2 Alternative distributions ......................................... 29
   3.3 Trade-off between FDR and NDR when $\pi_0$ is Known ............ 37
   3.4 Estimation of NDR for given FDR when $\pi_0$ is Unknown ........... 46
   3.5 Real Data Example ................................................ 66

4 Conclusions ........................................................................ 77

Bibliography ......................................................................... 79
List of Figures

1.1 Trade-off between type I error and type II error rates in a single hypothesis test when $n=100$, $\sigma=5$, and $\theta=0.5$, 1, and 1.5. The dashed lines are $\alpha=0.01$ and 0.05. ............................................. 6

2.1 FDR and NDR by $\alpha$ for multiple t-testing when $H_0: \mu = 0$, $H_a: \mu > 0$, $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $N(0.3,1)$. ............................................. 15

2.2 FDR, FNR, and NDR by $\alpha$ for multiple t-testing when $H_0: \mu = 0$, $H_a: \mu > 0$, $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $N(0.3,1)$. ............................................. 19

2.3 Boxplot of $\hat{\pi}_0$ by $\lambda$ for multiple t-testing when $H_0: \mu = 0$, $H_a: \mu > 0$, $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $N(0.3,1)$. The horizontal dashed line is true $\pi_0 = 0.5$. ............................................. 22
2.4 The average values of $\hat{NDR}$, true NDR (right panels) by given FDR values in $(0,0.5)$ on a grid of 0.01 for the multiple t-testing when $H_0 : \mu = 0$, $H_a : \mu > 0$, $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $N(0.3,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is q. .......................... 24

3.1 (a) Density plot of $U(0,1)$ and a normal distribution with the same mean and the variance. (b) Density plot of Laplace(0,1) and a normal distribution with the same mean and the variance. (c) Density plot of $\chi^2(1)$, $\chi^2(2)$, $\chi^2(3)$, $\chi^2(4)$, $\chi^2(5)$, and a normal distribution with the same mean and the variance with $\chi^2(5)$. (d) Density plot of $\frac{1}{2} N(0,1) + \frac{1}{2} N(4,1)$, and a normal distribution with the same mean and the variance. .......................... 31

3.2 Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $U(0,1)$. ................................. 32

3.3 Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is Laplace(0,1) ................................. 34

3.4 Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\chi^2(df)$ where df is randomly picked from the set of $\{1,2,3,4,5\}$ ................................. 35

3.5 Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\frac{1}{2} N(0,1) + \frac{1}{2} N(4,1)$. ................................. 37
3.6 NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $U(0,1)$. The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2. .......................... 39

3.7 NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\text{Laplace}(0,1)$. The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2. .......................... 41

3.8 NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\chi^2(\text{df})$ where df is randomly picked from the set of $\{1,2,3,4,5\}$. The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2. .......................... 42

3.9 NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) when the alternative distribution is $\frac{1}{2}N(0,1) + \frac{1}{2}N(4,1)$. The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2. .......................... 44

3.10 Boxplot of $\hat{\pi}_0$ by $\lambda$ in $(0,1)$ on a grid of 0.1 for multiple Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $U(0,1)$. The dashed line is true $\pi_0 = 0.5$. .......................... 47
3.11 Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $U(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels. 

3.12 Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $U(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.

3.13 Boxplot of $\hat{\pi}_0$ by $\lambda$ in $(0,1)$ on a grid of 0.1 for multiple Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is Laplace$(0,1)$. The dashed lines indicate true $\pi_0 = 0.5$.

3.14 Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is Laplace$(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
3.15 Box plots of $\hat{\text{NDR}}$ (left panels), and average values of $\hat{\text{NDR}}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is Laplace$(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.  

3.16 Boxplot of $\hat{\pi}_0$ by $\lambda$ in $(0,1)$ on a grid of 0.1 for multiple Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\chi^2(df)$ where $df$ is randomly picked from the set of $\{1,2,3,4,5\}$. The dashed line is true $\pi_0 = 0.5$.  

3.17 Box plots of $\hat{\text{NDR}}$ (left panels), and average values of $\hat{\text{NDR}}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\chi^2(df)$ where $df$ is randomly picked from the set of $\{1,2,3,4,5\}$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
3.18 Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\chi^2(df)$ where df is randomly picked from the set of $\{1,2,3,4,5\}$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.

3.19 Boxplot of $\hat{\pi}_0$ by $\lambda$ in $(0,1)$ on a grid of 0.1 for multiple Anderson-Darling normality testing when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\frac{1}{2}N(0,1)+\frac{1}{2}N(4,1)$. The dashed line is true $\pi_0 = 0.5$.

3.20 Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\frac{1}{2}N(0,1)+\frac{1}{2}N(4,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.

3.21 Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\frac{1}{2}N(0,1)+\frac{1}{2}N(4,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
3.22 Frequency histograms of 30 variables for 212 malignant images. ........... 67
3.23 Frequency histograms of 30 variables for 357 benign images. ............... 68
3.24 Histograms of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-
    square, and Lilliefors normality tests when \( m = 60, n_{\text{malignant}} = 212, \
    n_{\text{benign}} = 357 \). .............................................................. 69
3.25 Histograms of p-values for average counts in the bin locations by Anderson-
    Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors normality tests
    when \( m = 60, n_{\text{sub-sample}} = 30 \), repetition=100. ....................... 70
3.26 Estimation of \( \pi_0 \) by \( \lambda \) in (0, 1) on a grid of 0.1 for Anderson-Darling,
    Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov)
    normality tests when \( m = 60, n_{\text{malignant}} = 212, n_{\text{benign}} = 357 \). The dashed
    lines indicate \( \hat{\pi}_0 \) when \( \lambda = 0.5 \). ................................. 71
3.27 Estimation of \( \pi_0 \) by \( \lambda \) in (0, 1) on a grid of 0.1 for Anderson-Darling,
    Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov)
    normality tests when \( m = 60, n_{\text{sub-sample}} = 30 \), repetition=100. The
    dotted horizontal lines indicate \( \hat{\pi}_0 \) when \( \lambda = 0.5 \). .................. 72
3.28 Estimation of PON and NDR by given \( q \) with \( \lambda = 0.5 \) for Anderson-
    Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-
    Smirnov) normality tests when \( m = 60, n_{\text{malignant}} = 212, n_{\text{benign}} = 357 \). 74
3.29 Estimation of PON and NDR by given \( q \) in (0,0.38) on a grid of 0.1 with
    \( \lambda = 0.5 \) for Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lil-
    liefors (Kolmogorov-Smirnov) normality tests when \( m = 60, n_{\text{sub-sample}} = \
    30 \), repetition=100. ............................................................... 75
List of Tables

1.1 Four possible outcomes of a hypothesis test ................................................. 4

2.1 Notation of decision outcomes for multiple hypotheses testing .......... 9

2.2 Errors of multiple hypotheses testing with and without the Bonferroni correction when $\alpha = 0.05$ and $m = 2000$ ......................................................... 14

2.3 Errors of multiple hypotheses testing with Bonferroni, BH, BY Corrections 17

2.4 Compare of effectiveness with $\omega_{ratio} = \omega_{slope} = 1$. Point 1 and 3 are chosen by the measure of $e_{ratio}$, and Point 3 is chosen by the measure of $e_{slope}$ ......................................................... 25

3.1 Comparison of performances of the four normality tests for each alternative. Note that we prefer smaller FDR, smaller NDR, and larger area between PON and NDR. The approximation symbol “>” indicates the two tests have similar performance. For the trade-off between FDR and NDR, the greater symbol, “>” indicates more cost-effective ............. 45

3.2 Compare the products of four alternatives for each normality test. Note that we prefer smaller FDR, Smaller NDR, and larger area between PON and NDR. For the trade-off between FDR and NDR, the greater symbol, “>” indicates core cost-effective. “Uf” indicates the uniform, “La” indicates the Laplace, and “Mx” indicates the mixture alternatives ......................................................... 45
3.3 Compare of effectiveness between $q = 0.1$, 0.2, and 0.3 for the SW test with $\omega_{ratio} = \omega_{slope} = 1$ when the alternative is $U(0,1)$. . . . . . . . . . . 50

3.4 Compare of effectiveness between $q = 0.1$, 0.2, and 0.3 for the SW (or AD) test with $\omega_{ratio} = \omega_{slope} = 1$ when the alternative is Laplace. . . . . . . 51

3.5 Compare of effectiveness between $q = 0.05$, 0.1, and 0.2 for the SW (or AD) test with $\omega_{ratio} = \omega_{slope} = 1$ when the alternative is $\chi^2$. . . . . . . . . . . 56

3.6 Compare of effectiveness between $q = 0.05$, 0.1, and 0.2 for the SW (or AD) test with $\omega_{ratio} = \omega_{slope} = 1$ when the alternative is $\chi^2$. . . . . . . . . . . 61

3.7 Comparison of performances of the four normality tests for each alternative about the estimation of $\pi_0$ and NDR, the accuracy and control of FDR at level $q$, and the cost-effectiveness of trade-off between FDR and NDR. . . . 65

3.8 Compare of effectiveness between $q = 0.05$, 0.1, and 0.2 for the SW test with $\omega_{ratio} = \omega_{slope} = 1$ when $m = 60$, $n_{sub-sample} = 30$, repetition=100. . 73
Chapter 1

Two Types of Hypothesis Tests

Hypotheses testing is one of the most important concepts in statistics. It is formulated in two hypotheses: null ($H_0$) and alternative ($H_a$) hypotheses. The first step of a hypothesis testing is to transfer a research question into null and alternative hypotheses. A null hypothesis is typically stated by no difference or no effect, and an alternative hypothesis is expressed by the existence of difference or an effect. In a common hypothesis testing, the alternative hypothesis is what the researcher believes because he/she usually wants to see a difference or an effect. However, a null hypothesis represents the researcher’s theoretical position. Binder (1963) classified hypotheses testing into two groups: rejection-support (RS) testing and acceptance-support (AS) testing. In RS testing, an alternative hypothesis is what the researcher believes, and rejecting the null hypothesis supports the researcher’s theoretical position, whereas in AS testing, a null hypothesis is what the researcher believes. Thus, non-rejecting the null hypothesis supports the researcher’s idea. Thus, the researcher’s theoretical position is the main key that divides hypothesis tests into AS tests and RS tests. In the following sections, we discuss RS and AS testing with examples and also introduce two kinds of decision errors and their meanings in RS and
1.1 Rejection-Support Tests

In RS tests, an alternative hypothesis supports researcher’s theoretical position. We discuss about null and alternative hypotheses, and type I and type II errors in RS testing through an simple example of a t-test.

Suppose we ask Americans what they think about the ideal age to have a first child to test our hypothesis which is that men’s ideal age is higher than women’s. We pick a random sample of 50 men and 50 women to ask their ideas. Then, the mean of ideal age is 27 for men and 25 for women. The standard deviation is 3 for men and 4 for women.

In this case, the null and alternative hypotheses are:

\[ H_0 : \mu_{\text{men}} = \mu_{\text{women}} \]

\[ H_a : \mu_{\text{men}} > \mu_{\text{women}} \]

where \( \mu_{\text{men}} \) is the mean value for men, \( \mu_{\text{women}} \) is the mean value for men

Then, the t-statistic is

\[
T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} = \frac{27 - 25}{\sqrt{3^2/50 + 4^2/50}} = 5.345.
\]

We have to set a significance level to get a critical value. Usually, we set a significance level by the probability of type I error and denote it by \( \alpha \). We discuss type I and type II errors of the decisions in Subsection 1.1.3. A critical value is a cut-off value to determine the boundary between the rejection and non-rejection regions. In this example, if the
mean difference between men and women is higher, then we have a stronger evidence to reject the null hypothesis. If we set $\alpha = 0.05$ as the significance level, then the critical value is $\pm t_{\alpha/2,N_1+N_2-2} = \pm t_{0.025,98} = \pm 1.984$.

Since the test statistic 5.345 is greater than the critical value 1.984, it is in the rejection region, and we can reject the null hypothesis and say “the ideal age between men and women for the first baby are significantly different at the level of 0.05”.

### 1.2 Acceptance-Support Tests

In AS tests, the researcher’s theoretical position is on the null hypothesis. Suppose we want to see whether a random sample comes from a specific population with distribution. Then, we set the hypotheses

\[
    H_0: \text{The data follow a specified distribution},
\]

\[
    H_a: \text{The data do not follow the specified distribution}.
\]

Usually, the researchers want to claim that the data follow the specified distribution, which is the null hypothesis. The test is classified in AS because the researcher’s theoretical position is on the null. In this case, if we reject the null, we learn nothing. Even if we fail to reject the null, it does not necessarily mean that the data follow a specified distribution. Simply speaking, we want to escape the $H_a$ case which is “bad” in the researcher’s view. Note that we usually do not use the expression of “accept $H_0$”. Instead, we conclude that we do not have enough evidence to say the data do not follow the specified distribution.
Statisticians call this kind of test *goodness-of-fit test* because we want to see how well a specified model or distribution fits a data in this type of test. The normality test is a representative example for a goodness-of-fit test, because researchers usually want to avoid the worst situation that the data are not normal. In this case, the test is an AS test. On the other hand, if a researcher wants to prove the data are skewed or follow any other distribution which is not normal, then the test is a RS test. Thus, a goodness-of-fit test is usually an AS test, but we cannot simply classify that all of the goodness-of-test is always an AS test.

### 1.3 Decision Errors in Single Hypothesis Test

A hypothesis test allows only two possible decisions: rejection of the null hypothesis and failure to reject the null at a specified significance level. Since the null is either true or false, a hypothesis test admits four possible outcomes as shown in Table 1.1.

<table>
<thead>
<tr>
<th>$H_0$ is true</th>
<th>Non-Reject $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Negative</td>
<td>False Positive</td>
<td>(Type I error, $\alpha$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$H_a$ is true</th>
<th>False Negative</th>
<th>True Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Type II error, $\beta$)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are two ways to be correct: rejecting the null when it is false and failing to reject it when it is true. The first correct way is referred to as a true positive and the second is referred to as a true negative. There are also two ways to be wrong: rejecting the null when it is true and failing to reject when it is false. The first of these two ways to be wrong is referred to as a false positive or a type I error, and it is noted by $\alpha$. The second
of these two ways to be wrong is referred to as a false negative or a type II error, and it is noted by \( \beta \). The mathematical definitions are

\[
\text{Type I error} = P(\text{reject } H_0 | H_0 \text{ is true}),
\]

\[
\text{Type II error} = P(\text{non-reject } H_0 | H_0 \text{ is false}).
\]

The dependence between \( \alpha \) and \( \beta \) is well studied in the literature (Craiu and Sun, 2008). When a type I error is strictly controlled, a type II error is sacrificed. That is, a type II error is increased with a decrease of type I error. For example, suppose we want to test the mean of a normal population

\[
H_0 : \theta = 0 \\
H_a : \theta > 0
\]

with known variance \( \sigma^2 \) and \( n \) i.i.d. samples. The probability of type II error is calculated by the equation,

\[
\beta(\theta; \alpha) = 1 - \Phi\left(\Phi^{-1}(\alpha) + \frac{\theta}{\sigma\sqrt{n}}\right),
\]

where \( \Phi \) is the cdf of the standard normal distribution.

Assume \( n=100 \) and \( \sigma=5 \). Figure 1.1 shows \( \beta \) is decreasing with increasing of \( \alpha \) for \( \theta = 0.5, 1, \) and 1.5. The \( \alpha \) values are from 0 to 0.3 by 0.01. The dashed lines indicate \( \alpha \) values of 0.01 and 0.05 respectively. The value of type II error is falling when \( \alpha \) is relaxed from 0.01 to 0.05. Usually, in a single RS testing, the \( \alpha \) is pre-decided at a small value and \( \beta \) is used for calculating power to get an appropriate sample size. Actually, power is the rate of correctly rejected hypothesis, so it is a function of type II error. The common
Figure 1.1: Trade-off between type I error and type II error rates in a single hypothesis test when $n=100$, $\sigma=5$, and $\theta=0.5, 1, \text{ and } 1.5$. The dashed lines are $\alpha=0.01$ and 0.05.

Definition of power for single hypothesis test is

$$\text{Power} = P(\text{reject } H_0 | H_0 \text{ is false}) = P(\text{reject } H_0 | H_a \text{ is true}).$$

(1.1)

Thus, power = 1 - $P(\text{fail to reject } H_0 | H_a \text{ is true}) = 1 - \text{type II error}$.

When the mean difference between null and alternative ($\theta$) is bigger, the type II error is dropped dramatically, because the big difference between two means indicates that it is easy to recognize which one is which, and so both types of errors are decreased simultaneously.
If the researcher’s hope is in the alternative as typical testing, he/she wants to prove this is true with strong evidence. Thus, researchers are typically more sensitive to type I error than type II error. In the previous RS test example, the type I and II errors are

I: decide there is a significant difference between men and women when it is not true
II: decide there is no difference between men and women when it is not true.

In this case, they are more likely to claim that the male and female think differently about the ideal age of the first child. Even if type II error occurred, there is nothing to lose because they may not publish the result. However, type I error implies that they may publish that their claim is proved, but it is wrong.

In the same way, when the researcher’s belief is on the null hypothesis, then they must be more sensitive to type II error. For the case of AS test example, the type I and II errors are,

I: decide the data do not follow a specified distribution but it is not true
II: decide the data follow a specified distribution but it is not true

In this case, they are more likely to wish to claim that the data follow the distribution. Type II error implies that they believe the data follow the distribution and move on the next step. Thus, in AS testing, type I and type II errors have the opposite meanings to those in RS testing. (Nickerson, 2000)
Chapter 2

Multiple Testing

Multiple testing refers to the testing of more than one hypothesis simultaneously. When many hypotheses are tested at once, each test has the type I error probability, and there might be a marked increase in the probability of at least one type I errors with the number of hypotheses. We introduce several methods to deal with this multiple testing problem.

2.1 Decision Errors

Broadly speaking, there are two kinds of errors: false positive and false negative. In a single hypothesis test, we call them type I and II errors as we mention in the previous section. In multiple testing, there are several different ways to define false positive and negative. To define errors in multiple testing, we reset the notation. Assume we test $m$ hypotheses, $m_0$ is the number of true null hypotheses, and $R$ is the number of rejected hypotheses. Other values are in Table 2.1. $T_N$ indicates true negative, $F_P$ indicates false positive, $F_N$ indicates false negative, and $T_P$ indicates true positive.
Table 2.1: Notation of decision outcomes for multiple hypotheses testing

<table>
<thead>
<tr>
<th></th>
<th>Non-Reject $H_0$</th>
<th>Reject $H_0$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>$T_N$</td>
<td>$F_P$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>$H_a$ is true</td>
<td>$F_N$</td>
<td>$T_P$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$N$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

2.1.1 False Positives

As noted, researchers care more about false positive in RS testing. We introduce several measures for false positives.

A straightforward extension of the type I error to multiple testing is *per comparison error rate* (PCER). It is the expected value of the number of type I errors over the total number of hypotheses, i.e.,

$$\text{PCER} = E(F_P)/m.$$  

In multiple testing, the PCER is additive such that we are mostly likely to make erroneous conclusions. In other words, the probability of making one or more type I errors among all hypotheses, $1 - (1 - \alpha)^m$, will be expanded. In this sense, the most commonly controlled measure for false positives in multiple testing is the *family-wise error rate* (FWER). It is the probability of at least one false positives,

$$\text{FWER} = P(F_P \geq 1).$$  

When multiple hypotheses tests are performed, the FWER is quickly increased at the rate of $1 - (1 - \alpha)^m$. For example, if we want to conduct multiple testing for 100 independent
hypotheses with a significance level of 0.05, then the FWER is

\[ 1 - (1 - 0.05)^{100} = 0.994. \]

Classical procedures that control FWER are extremely strict and tend to have very small power. Benjamini and Hochberg (1995) argue that control FWER is not quite needed. The control of FWER is important when at least one of the type I error make erroneous conclusions. However, the overall conclusion is not erroneous even if some of the null hypotheses are falsely rejected in many cases. They suggest a new measure for false positive in multiple testing which is False Discovery Rate (FDR). Basically, discovery has the same meaning as rejection. FDR is the expected proportion of false positives among the rejected hypotheses. The most obvious definition of FDR is, \( E(F_P/R) \).

However, there are some problems of this definition. In most cases, the probability that \( R = 0 \) is positive. Another problem is that if all null hypotheses are true, the FDR is equivalent to FWER. If \( F_P = 0 \), then FDR=0, and if \( F_P > 0 \), then FDR=1. To avoid those problems, they defined FDR,

\[
\text{FDR} = E(F_P/R|R > 0)P(R > 0). \tag{2.1}
\]

Storey (2003) argued that some researchers would want the FDR to be 1 when all null hypotheses are true, and that \( P(R > 0) \) is not needed anymore. He named the new FDR to Positive False Discovery Rate (pFDR),

\[
pFDR = E(F_P/R|R > 0).
\]

The term “positive” describes that we have a condition of at least one positive (rejected
hypothesis). Since $P(R > 0)$ is around 1 in most experiments, FDR and pFDR are very similar.

2.1.2 False Negatives

In multiple AS testing, researchers more care about false negatives than false positives like in single AS testing. We introduce three kinds of false negatives of multiple testing. Since the FWER is a measure of false positives, a similar measure in terms of false negatives is the probability of at least one false negatives which we call *Family-Wise type II Error Rate* (FWER.II),

$$\text{FWER.II} = P(F_N \geq 1).$$

Genovese and Wasserman (2002) suggested a new quantity for false negative, *False Non-discovery Rate* (FNR), which is a similar measure as FDR in false positive. It is the expected proportion of false negatives among the non-rejected hypotheses,

$$\text{FNR} = E(F_N/N|N > 0)P(N > 0).$$

However, Craiu and Sun (2008) argued that *Non Discovery Rate* (NDR) is more suitable to trade-off with FDR. It is defined as the expected value of the number of false negatives over the number of alternative hypotheses,

$$\text{NDR} = E(F_N)/m_1. \quad (2.2)$$

Note that the quantity is defined given $m_1 > 0$. Since NDR is the average of type II error rate of single testing, it is a natural extension of the type II error rate for
multiple testing. Also, it is difficult to interpret $1 \text{-FNR}$ as power, because $1 - \text{FNR} = E(T_N/N|N > 0)P(N > 0)$, value depends on the true negatives not true positives. Whereas, $1 - \text{NDR} = E(T_P/m_1)$ is the average of power and depends on true positives. Thus, they choose the trade-off between FDR and NDR instead of between FDR and FNR.

### 2.2 Control of Errors

#### 2.2.1 Bonferroni Correction

Holm (1979) developed the Bonferroni correction procedure to control the probability of at least one false positive in multiple testing. It is considered the most conservative method to control false positives. It is named after Italian mathematician Carlo Emilio Bonferroni for the use of Bonferroni inequality within multiple inference theory.

If we want to control the FWER for the whole family of tests at the level of $\alpha$, the Bonferroni correction tests the individual hypothesis at a significance level of $\alpha/m$ when dependent or independent tests are being performed simultaneously. The corrected type I error for each test is

$$P(\text{reject } H_i|H_0 \text{ is true}) \leq \frac{\alpha}{m}$$

for $1 \leq i \leq m$, then

$$P(\text{reject some } H_i|H_0 \text{ is true}) \leq \alpha$$

which follows from the Bonferroni inequality.

The Bonferroni correction is known as the most conservative correction method for multiple hypotheses testing because it makes the threshold very small. Thus, it is hard
to reject a null hypothesis with the Bonferroni correction. It controls FWER, but the problem is that the Bonferroni correction controls the probability of at least one false positive, so it can increase the probability of false negatives. Moreover in practice, if the tests are not independent, the Bonferroni correction can give rise to a high rate of false negatives. As an illustrative example, we generate $m = 200$ independent data sets for a t-test. The hypotheses are

$$H_0 : \mu = 0,$$

$$H_a : \mu > 0.$$ 

Half of them are from $N(0, 1)$ and others are from $N(0.3, 1)$. We set $\alpha$ at 0.05 and the sample size is 30. Thus, the Bonferroni correction adjusted the threshold at $0.05/200 = 0.00025$, and the corrected FWER is

$$\text{FWER}_{Bof} = P(\text{at least one null is rejected})$$

$$= 1 - P(\text{no one is rejected})$$

$$= 1 - (1 - \alpha/m)^m$$

$$= 1 - (1 - 0.05/200)^{200}$$

$$= 0.049$$

$$< 0.05.$$ 

The FWER is controlled at the level of $\alpha$. We repeat 50 times to get the average values of the errors. Without any kind of correction, FWER is 0.98, FDR is 0.1027, the average of type I error rate is about 0.0544 which is approximately same with the $\alpha = 0.05$, and the average of type II error rate (NDR) is 0.5246 as shown in Table 2.2. With the Bonferroni correction, FWER is 0.04 which means that it is obviously controlled at the level of $\alpha = 0.05$, and also FDR and the average of type I error rate are quite decreased.
However, NDR is 0.9804, which is exceedingly high with the correction. Note that the Bonferroni correction is for controlling FWER originally, but it also affects the other kinds of errors. We change $\alpha$ value from 0 to 0.5 in 0.01 steps to see the trends of FDR and NDR by $\alpha$ as shown in Figure 2.1. The dashed vertical lines correspond to $\alpha = 0.01$ and $\alpha = 0.05$ respectively. There is a steep increase in the average of type II error rate when the threshold is less than 0.05; it implies that NDR is increased incredibly with the Bonferroni method. When the $\alpha$ values are increasing, FDR value is gradually increased while NDR is rapidly decreased.

Table 2.2: Errors of multiple hypotheses testing with and without the Bonferroni correction when $\alpha = 0.05$ and $m = 2000$

<table>
<thead>
<tr>
<th></th>
<th>Ave(Type I error rate)</th>
<th>FWER</th>
<th>FDR</th>
<th>FNR</th>
<th>NDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No correction</td>
<td>0.0544</td>
<td>0.98</td>
<td>0.1027</td>
<td>0.3568</td>
<td>0.5246</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>0.0004</td>
<td>0.04</td>
<td>0.0200</td>
<td>0.4952</td>
<td>0.9804</td>
</tr>
</tbody>
</table>

While many studies focus on reducing false positives with the Bonferroni correction, the increased false negative is often ignored in multiple testing. The problem is that the power is greatly reduced with an increase of false negatives. Thus, many researchers introduced alternative methods to the Bonferroni correction. Among those false positive controlling methods of multiple testing, we discuss several FDR controlling methods in the next subsubsection.

2.2.2 Benjamini-Hochberg and Benjamin-Yekutieli Procedures

Benjamini and Hochberg (1995) presented a procedure to control FDR. Suppose we test $m$ hypotheses: $H_1, H_2, \ldots, H_m$, and want to control FDR at level $q$. The first step of the FDR control procedure is ordering p-values. Denote by $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ the ordered set of p-values corresponding to the tested hypotheses, and by $H_{(i)}$ the
corresponding hypotheses. Let $k$ be the largest $i$ for which

$$p_{(i)} \leq \frac{i}{m} q,$$

then reject all $H_{(i)}$ for $i \leq k$.

The Benjamini and Hochberg (BH) FDR correction is less conservative than the Bonferroni correction because

$$\frac{q}{m} \leq \frac{i}{m} q$$
It means the threshold of the BH method is looser than the threshold of the Bonferroni correction.

Benjamini and Hochberg proved the above procedure satisfies this inequality

$$\text{FDR} = E(F_P/R|P_{m_0+1} = p_1, \ldots, P_m = p + m_1) \leq \frac{m_0}{m} q \leq q,$$

where $F_P$ is the number of false positives, $R$ is the number of rejected hypotheses, $m_0$ is the number of true null hypotheses, $m_1$ is the number of false null hypotheses, and $m$ is the total number of hypotheses. Thus, the FDR is controlled at $q$ for independent test statistics. They present an adaptive procedure because the original FDR procedure is too conservative when some of the hypotheses are in fact false (Benjamini and Hochberg, 2000). Thus, they look for the largest $k$ such that

$$p(k) \leq \frac{k}{m} \frac{q}{\pi_0}, \quad (2.4)$$

and reject all $H(i)$ where $i \leq k$.

The BH procedure requires an assumption: the test statistics are independent. However, dependent test statistics are encountered more often in practice. Benjamini and Yekutieli (2001) prove the theorem,

*If the joint distribution of the test statistics is positive regression dependency structure on the subset of test statistics corresponding to true null hypotheses, the Benjamini Hochberg procedure controls the FDR at level less than or equal to $\frac{m_0}{m} q$."

It implies the BH procedure also controls the false discovery rate when the test statistics have positive regression dependency corresponding to the true null hypotheses which is
this theorem.

In the Benjamini and Yekutieli (BY) method, the critical value is determined by

\[
q \leq \sum_{i=1}^{m} \left( \frac{1}{i} \right)
\]

where \( m \) is the number of hypothesis tests. They proved this procedure always controls the FDR at level less or equal to \( m_0 q / m \) where \( m_0 \) is the number of true null hypotheses, and so the BY method is more conservative to control type I error rate than the BH method. Even though the BH and the BY methods were originally developed to control FDR, they also have an effect to control other false positives. Likewise, the Bonferroni correction also affects other kinds of false positives, not just FWER. Thus, we can compare the errors with the three methods. From the same situation with the previous example in Section 2.2.1, we adjusted the test statistics with the BH and BY methods for a comparison. Table 2.3 shows the FWER and the average of false positive are the most smallest with the B-Y method. The BH method is less conservative than the other two methods. With the BH method, FWER, FDR, and the average of \( F_P \) values are bigger than the ones with other methods. On the other hand, with the BH method, FWER.II, NDR, and the average of \( F_N \) values are smaller than the ones with the other correction methods.

<table>
<thead>
<tr>
<th></th>
<th>Ave(Type I error rate)</th>
<th>FWER</th>
<th>FDR</th>
<th>FNR</th>
<th>NDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No correction</td>
<td>0.0544</td>
<td>0.98</td>
<td>0.1027</td>
<td>0.3568</td>
<td>0.5246</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>0.0004</td>
<td>0.04</td>
<td>0.0200</td>
<td>0.4952</td>
<td>0.9804</td>
</tr>
<tr>
<td>BH</td>
<td>0.0032</td>
<td>0.24</td>
<td>0.0290</td>
<td>0.4725</td>
<td>0.8928</td>
</tr>
<tr>
<td>BY</td>
<td>0.0002</td>
<td>0.02</td>
<td>0.0213</td>
<td>0.4977</td>
<td>0.9908</td>
</tr>
</tbody>
</table>
2.2.3 Control of False Negatives

While most researchers focus on how to control false positives, some researchers (Craiu and Sun, 2008, Genovese and Wasserman, 2002, Sarkar, 2004, 2006, Storey, 2003) discuss false negative quantities in multiple testing. We discuss false negative control with FNR and NDR in this section.

Genovese and Wasserman (2002) introduce FNR as a false negative quantity in multiple testing, and suggest a linear combination of FDR and FNR, $L = FNR + \gamma FDR$, as a measure for comparing multiple testing procedures, where $\gamma$ is a constant.

Also, Sarkar (2004) defines an unbiased FDR-controlling multiple testing procedures by the property:

*An FDR-controlling multiple testing procedure is said to be unbiased at level $\alpha$ if*

$$FDR \leq \alpha \text{ and } FDR + FNR \leq 1.$$

This definition is from the common sense that a good multiple testing should guarantee that the proportion of correct decisions is higher than that of incorrect decisions. Because FDR and FNR are incorrect decisions, $1-FDR$ and $1-FNR$ are correct decisions. It implies

$$FDR + FNR \leq 1 - FDR + 1 - FNR.$$

Craiu and Sun (2008) argue that there is little practical meaning to control FNR in multiple testing, even though FNR is matched with FDR mathematically. Instead, they suggest NDR (non-discovery rate) as a good candidate of false negatives for multiple hypotheses testing, and define as (2.2). Since FNR and NDR are false negatives, both
of them decrease as the increasing $\alpha$ value, while FDR increases as shown in Figure 2.2 with the same setting of Figure 2.1.

![Figure 2.2: FDR, FNR, and NDR by $\alpha$ for multiple t-testing when $H_0 : \mu = 0$, $H_a : \mu > 0$, $m = 200$, $n = 30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0, 1)$, and the alternative distribution is $N(0.3, 1)$. They apply a trade-off between FDR and NDR to find a most cost-effective FDR. They introduce two kinds of cost-effectiveness measures,

$$e_{\text{ratio}} = \omega_{\text{ratio}} \frac{\text{NDR}_2 / \text{NDR}_1}{\text{FDR}_1 / \text{FDR}_2},$$

$$e_{\text{slope}} = \omega_{\text{slope}} \frac{\text{NDR}_2 - \text{NDR}_1}{\text{FDR}_2 - \text{FDR}_1},$$

where $\omega_{\text{ratio}}$ and $\omega_{\text{slope}}$ are weighting factors, and FDR$_1 < $ FDR$_2$. For the first measure,
\( e_{\text{ratio}} > 1 \) implies FDR\(_1\) is more cost-effective than FDR\(_2\). For example, if FDR and NDR are equally interesting, then \( \omega_{\text{ratio}} = 1 \), and \( e_{\text{ratio}} = 1 \) indicates NDR\(_1 \cdot \) FDR\(_1 = \) NDR\(_2 \cdot \) FDR\(_2\). Thus, increasing FDR from 0.01 to 0.1 necessitates a 10-fold decrease in NDR to make the two FDRs equally efficient. The second measure, \( e_{\text{slope}} \) has a more obvious interpretation and is easily seen from the slope on the plot of FDR vs. NDR. When \( e_{\text{slope}} > -1 \), it implies FDR\(_1\) is more cost-effective than FDR\(_2\). If FDR and NDR are equally interested, then the \( \omega_{\text{slope}} = 1 \), and \( e_{\text{slope}} > -1 \) implies FDR\(_2 + \) NDR\(_2 > \) FDR\(_1 + NDR_1\). If the \( e_{\text{slope}} \) is less than negative one, that is if the graph is steep negatively, then a less stringent FDR is more beneficial. When researchers give more weight to the effectiveness of NDR, then the \( \omega_{\text{ratio}} < 1 \). Similarly, if researchers want to give more weight to the effectiveness of FDR, then the \( \omega_{\text{ratio}} > 1 \).

Craiu and Sun (2008) discuss a procedure with four steps to estimate NDR by FDR.

Step1: Estimate \( \pi_0 \).

Step2: Choose FDR=\( q \), e.g., \( q \in (0, \hat{\pi}_0) \), on a grid of 0.01.

Step3: Derive \( R \).

Step4: Estimate NDR.

The first step is to estimate \( \pi_0 \). Schweder and Spjøtvoll (1982) suggested an estimator \( \hat{\pi}_0(\lambda) \) of \( \pi_0 \) which is based on the observed p-values. We consider the following commonly used estimator

\[
\hat{\pi}_0(\lambda) = \frac{\# \{ p_i > \lambda \}}{m(1-\lambda)}, \tag{2.7}
\]

where \( \lambda \) is a constant. Langaas et al. (2005) discuss \( \pi_0 \) estimation based on a p-value threshold. Since p-values that are from the alternative hypothesis are likely to be small,
we expect that most of the p-values within the interval \([\lambda, 1]\) come from the null hypothesis. Thus, we expect that a large majority of p-values in \([\lambda, 1]\) come from \(U(0, 1)\) for a not-too-small \(\lambda\). This indicates the approximation,

\[
E[\#\{p_i > \lambda\}] \approx \pi_0(1 - \lambda).
\]

Therefore, the equation (2.5) is a reasonable estimator of \(\pi_0\). Black (2004) argues that the estimator \(\hat{\pi}_0\) is unbiased when \(\pi_0 = 1\) for any choice of \(\lambda\). When the data are generated from either the null or alternative hypotheses, the uniformity of the p-values dose not hold, and the estimator is biased. For a given \(\lambda\),

\[
E[\hat{\pi}_0(\lambda)] = \frac{E[\#(p_i > \lambda)]}{m(1 - \lambda)} = \pi_0 + \frac{1 - \pi_0}{1 - \lambda} \Pr(p_i > \lambda|H_a),
\]

where \(H_a\) is an alternative hypothesis. The bias of the estimator, \(\Pr(p_i > \lambda|H_a)(1 - \pi_0)/(1 - \lambda)\) is the function of \(\pi_0\). Craiu and Sun (2008) use \(\lambda=0.5\) to estimate \(\pi_0\), then the bias of the estimator is \(2(1 - \pi_0)F_1[F_0^{-1}(0.5)]\). When \(\pi_0\) is small, the bias of \(\hat{\pi}_0\) could be substantial. Also, if the “distance” between the null and alternative populations is small, \(\Pr(p_i > \lambda|H_a)\) could be large, and the bias could be considerable. We use the same example in Figures 2.1 and 2.2 to see the relation between \(\lambda\) and \(\hat{\pi}_0\) empirically. The dashed horizontal line indicates the true \(\pi\) which is 0.05.

When the value of \(\lambda\) increases in the interval \([0, 1]\), \(\hat{\pi}_0\) is decreased generally as shown in Figure 2.3. The figure shows that choosing a large value of \(\lambda\) gives a large variance, which is theoretically supported (Langaas et al., 2005). However, for a small \(\lambda\) value, many p-values from the alternative hypothesis could be included in \(\{p_i > \lambda\}\), and the bias is increased as shown in Figure 2.3. Thus, there is a trade-off between the bias and variance when choosing \(\lambda\). Storey (2002) discusses a method to find an optimal \(\lambda\) which
minimizes the mean-squared error (MSE).

\[ \lambda_{\text{best}} = \arg \min_{\lambda \in [0,1]} (E[\{\hat{\text{FDR}}(\lambda) - \text{FDR}\}^2]). \]

The true MSE is unknown, so he suggests estimating the MSE by using bootstrapping. We discuss more about the estimation of \( \hat{\pi}_0 \) later with simulated data and also real data in Chapters 3 and 4.
To derive $R$, the adaptive BH procedure, equation (2.4) can be performed so that the FDR adjusted p-value corresponding to $p(i)$ is

$$p_{FDR}^{(i)} = \min \left\{ \frac{mp(j)}{j} : j \geq i \right\} = \min \left\{ \frac{mp(i)}{i}, p_{FDR}^{(i+1)} \right\},$$

with $p_{FDR}^{(m)} = p_{(m)}$. Equations (2.4) and (2.8) imply rejecting all hypotheses with $p_{FDR}^{(i)} \leq q/\hat{\pi}_0$, resulting in

$$R = \# \{p_{FDR}^{(i)} \leq q/\hat{\pi}_0 \}. \quad (2.9)$$

To estimate NDR, for a chosen FDR level $q$ such that $1 \leq q \leq \pi_0$,

$$\text{NDR} = E[F_N] = 1 - E[T_P] = 1 - \frac{E[R - F_P]}{m_1} = 1 - \frac{(1 - q/P(\text{R} > 0))E[R]}{(1 - \pi_0)m}, \quad (2.10)$$

An estimate for NDR is obtained by replacing $E[R]$ with an observed value $R$, and $\pi_0$ with $\hat{\pi}_0$,

$$\hat{\text{NDR}} = \left\{1 - \frac{(1 - q)\bar{R}}{(1 - \hat{\pi}_0)m}\right\}I(\hat{\pi}_0 < 1). \quad (2.11)$$

To illustrate the procedure of estimating NDR, we use the same example of Figure 2.1 and 2.2. That is, we conduct a multiple t-testing with a total number of 200 hypotheses, $\pi_0=0.5$, each sample size is 30, and 50 repetitions to get average values. However, the $\pi_0$ is unknown in practice, so we suppose $\pi_0$ is unknown and estimate it with $\lambda=0.5$. We choose FDR in $(0,0.5)$, on a grid of 0.01. With equation (2.11), we can estimate NDR by FDR with the same setting as the previous example. In this case, we can calculate the true NDR values since we know the true $\pi_0$. The Figure 2.4 shows that the estimates NDR values have a quite similar pattern with the true NDR values. We pick three points to compare the effectiveness. Also, the true FDR values are almost same with the
given FDR values. The given FDR values of three points are 0.05, 0.1, and 0.2, and the matched NDR values are 0.7448, 0.4962, and 0.1788 as shown in Figure 2.4.

![Figure 2.4: The average values of \( \hat{NDR} \), true NDR (right panels) by given FDR values in (0,0.5) on a grid of 0.01 for the multiple t-testing when \( H_0 : \mu = 0 \), \( H_a : \mu > 0 \), \( m=200 \), \( n=30 \), repetition=50, \( \pi_0 = 0.5 \), \( \lambda = 0.5 \), the null distribution is \( N(0,1) \), and the alternative distribution is \( N(0.3,1) \). The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is q. For simplicity, we assume \( \omega_{ratio} = \omega_{slope} = 1 \). The two kinds of cost-effectiveness between the first and second points are

\[
\epsilon_{ratio}(1 \ vs. \ 2) = \frac{NDR_2/NDR_1}{FDR_1/FDR_2} = 1.3323 > 1,
\]

\[
\epsilon_{slope}(1 \ vs. \ 2) = \frac{NDR_2 - NDR_1}{FDR_2 - FDR_1} = -4.9732 < -1.
\]
Since $e_{\text{ratio}}(1 \ vs. \ 2)$ is greater than 1, FDR$_1$ is more cost-effective than FDR$_2$. However, the $e_{\text{slope}}(1 \ vs. \ 2)$ is smaller than -1, and so FDR$_2$ is more cost-effective than FDR$_1$ for the value of $e_{\text{slope}}(1 \ vs. \ 2)$. The conclusions by $e_{\text{ratio}}$ and $e_{\text{slope}}$ could be different.

$$e_{\text{ratio}}(2 \ vs. \ 3) = \frac{\text{NDR}_3/\text{NDR}_2}{\text{FDR}_2/\text{FDR}_3} = 0.7207 < 1,$$

$$e_{\text{slope}}(2 \ vs. \ 3) = \frac{\text{NDR}_3 - \text{NDR}_2}{\text{FDR}_3 - \text{FDR}_2} = -3.1738 < -1.$$

For the case of comparison between the second and third points, two kinds of cost-effectiveness indicate the same conclusion which is that FDR$_3$ is more cost-effective than FDR$_2$.

Table 2.4: Compare of effectiveness with $\omega_{\text{ratio}} = \omega_{\text{slope}} = 1$. Point 1 and 3 are chosen by the measure of $e_{\text{ratio}}$, and Point 3 is chosen by the measure of $e_{\text{slope}}$.

<table>
<thead>
<tr>
<th></th>
<th>1 vs. 2</th>
<th>2 vs. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{\text{ratio}}$</td>
<td>1.3323</td>
<td>0.7207</td>
</tr>
<tr>
<td>$e_{\text{slope}}$</td>
<td>-4.9732</td>
<td>-3.1738</td>
</tr>
</tbody>
</table>

As shown in Table 2.4, the measure of $e_{\text{ratio}}$ implies that point 1 is more cost-effective than point 2, and point 3 is also more cost-effective than point 2, whereas $e_{\text{slope}}$ implies that point 3 is the most cost-effective among the three points. Thus, we need to also compare point 1 and point 3.

The $e_{\text{ratio}}(1 \ vs. \ 3)$ is equal to 0.9602, which means point 3 is more cost-effective than point 1, and $e_{\text{slope}}(1 \ vs. \ 3)$ is equal to -3.7736 which implies point 3 is more cost-effective than point 1. Over all, the point 3 is most cost-effective among the three points. If the weighted values are changed, the decisions could be changed.
Chapter 3

Multiple Normality Testing

We discuss multiple normality tests in this chapter as an example of multiple acceptance-support tests. Normality tests are used to determine if a random sample of $n$ comes from a normal population. The null and alternative hypotheses are:

$$H_0: \text{The distribution is normal.}$$

$$H_a: \text{The distribution is not normal.}$$

Generally, the researcher’s theoretical position is on the null hypothesis in a normality test, which means it is an acceptance support test.

3.1 Normality Tests

In this section, we discuss five popular normality tests, each of which has some advantages and some disadvantages or restrictions.

The Chi-Square goodness of fit test is one of the most common normality tests. We
compare the observed and the expected frequencies. The expected are from a specified model or a distribution. In general, the chi-square test statistic is of the form

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where $k$ is a number of classes, $O_i$ is the observed frequency, and $E_i$ is the expected frequency. This statistic follows a $\chi^2$ distribution with $k - 1$ degrees of freedom, and the p-value can be calculated. If the test statistic is large, it means that the observed values do not fit the model or the assumed distribution. This test must be used with sufficient sample data. Usually, the required expected frequency is at least five, and if some of the counts are less than five, you may need to combine some classes in the tails.

The Kolmogorov-Smirnov (KS) test (Massey Jr, 1951) is based on the empirical distribution function. The test statistic is defined as

$$D_n = \sup_x |F_n(x) - F(x)|,$$

where $F_n(x)$ is the empirical distribution function and $F(x)$ is a specified distribution function. An advantage is that the critical value does not depend on the underlying cumulative function that is being tested. Lilliefors (1967) discuss that the KS test has at least two advantages over the chi-square test. The SW test can be used with small sample sizes, and it is often more powerful than the chi-square test for any sample size. However, there are some important disadvantages. This test can be applied only to continuous distributions and it is more sensitive near center than at the tails. The most important disadvantage is that the parameters are estimated from the data, then the critical value is not valid anymore.
The *Anderson-Darling* (AD) test (Anderson and Darling, 1952, 1954) is a modification of the KS test and gives more weight to the tails than the KS test. Thus, the AD test is more sensitive to the variance in the tails of the distribution than the KS test. The test statistic is defined as

\[
A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x),
\]

(3.3)

to test the hypothesis that a random sample \(X_1, \ldots, X_n\), with empirical distribution \(F_n(x)\). As shown in the equation (3.3), the weight function \(w(x) = \left[ F(x)(1 - F(x)) \right]^{-1} \) places more weights when \(F(x)\) is near 0 or 1, that is, at the tails. However, the critical values for the AD test are dependent on the specific distribution \(F(x)\), thus the critical value is different for each different distribution which is being tested.

The *Lilliefors* (LF) test (Lilliefors, 1967) is another modification of the KS test, but the parameters are estimated based on the sample. Thus, if the parameters are unknown, we prefer the LF test. The LF statistic is defined by

\[
L = \max_x |S_n(x) - F(x)|,
\]

(3.4)

where \(S_n(x)\) is the sample cumulative function and \(F(x)\) is the cumulative normal distribution function with \(\mu = \bar{X}\) and \(\sigma^2 = \frac{s^2}{n-1}\).

Shapiro and Wilk (1965) introduce a new statistical procedure for testing of normality. Suppose we want to test whether a sample \(x_1, \ldots, x_n\) come from a normally distributed
population. The *Shapiro-Wilk* (SW) test statistic is defined by

\[
W = \frac{\left( \sum_{i=1}^{n} a_i x(i) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},
\]

where \( x_i \) are random sample, \( x(i) \) is the ordered sample values, \( a_i \) are constants given by \( (a_1,...,a_n) = m^T V^{-1} (m^T V^{-1} m)^{-1/2} \),

\( m = (m_1,...,m_n)^T \),

\( m_1,...,m_n \) are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and \( V = (v_{ij}) \) be the \( n \times n \) covariance matrix of those order statistics.

The value of \( W \) in the equation (3.5) is between zero and one. If \( W \) is a small value, then we can reject the normality hypothesis. Whereas, if \( W = 1 \), it indicates the normality of the data. The study of Keskin (2006) showed that the SW test has the highest power for the normality test. Also, Razali and Wah (2011) showed SW test is the most powerful normality test for any type of alternative distribution and sample sizes whereas the KS test is the least powerful.

### 3.2 Alternative distributions

Now we consider alternative hypotheses for a normality test. If we set a normal distribution as a null hypothesis, then all other distributions can be alternative. When the alternative is not specified, the problem is that the type II error \( (\beta) \) cannot be calculated. Then, the power cannot be calculated because it is a function of \( \beta \). If we want to design a powerful normality test for each problem, we have to calculate the power. Thus,
researchers have to choose the alternative according to the context. For example, Razali and Wah (2011) studied power comparison of four different normality tests. They pick the alternative from 14 different non-normal distributions to cover various standardized skewness and kurtosis. They consider seven symmetric distributions: U(0,1), Beta(2,2), t(300), t(10), t(7), t(5), and Laplace, and seven asymmetric distributions: Beta(6,2), Beta(2,1), Beta(3,2), χ²(20), χ²(4), Gamma(4,5), and Gamma(1,5). We discuss four kinds of alternative distributions in this thesis: U(0,1), Laplace(0,1), χ²(df) where df is randomly picked from {1,2,3,4,5}, and a mixture of normal distribution. The density plots of four picked alternative distributions are shown in Figure 3.1.

3.2.1 Uniform Distribution

If a researcher infers that data follow a distribution which has a constant probability, a uniform distribution is one possibility for the alternative distribution of the data. We consider U(0,1) as an alternative for a normality test. Figure 3.1 (a) shows a density plot of U(0,1) and the normal distribution with the same mean and variance with U(0,1). As an example, we generate m = 200 independent data sets. Half of them are from U(0,1) and others are from N(0,1). Thus, we know the actual π₀ = 0.5 in this case. The sample size for each test is 30. We use four different normality tests: Anderson-Darling (AD) normality test, Shapiro-Wilk (SW) normality test, Pearson chi-square (PC) normality test, and Lilliefors (LF) normality tests. Figure 3.2 shows histograms of p-values by four different multiple normality tests with 20 breaks. The bars with diagonal stripes indicate that the data are from the normal distribution which is the null distribution. We expect that under the null distribution, the p-values follow a uniform distribution over the interval [0, 1]. We expect that the number of observed p-values in a rejection region, (0, α), under the alternative distribution is higher than under the null distribution of the
Figure 3.1: (a) Density plot of $U(0, 1)$ and a normal distribution with the same mean and the variance. (b) Density plot of Laplace(0,1) and a normal distribution with the same mean and the variance. (c) Density plot of $\chi^2(1), \chi^2(2), \chi^2(3), \chi^2(4), \chi^2(5)$, and a normal distribution with the same mean and the variance with $\chi^2(5)$. (d) Density plot of $\frac{1}{2}N(0, 1) + \frac{1}{2}N(4, 1)$, and a normal distribution with the same mean and the variance.
test. As an example, suppose $\alpha=0.05$, each of the first left bar indicates the number of rejected hypotheses. For the AD and SW tests, the proportion of the striped area is small. However, the proportion of the striped area is about 0.5 in the first bar for the PC test. By the equation (1.1), the power of each test is represented by the white section of the first bar on the left in each histogram at the level of $\alpha = 0.05$. With our histograms in Figure 3.2, the AD and SW tests have more power than the PC or LF test.

### 3.2.2 Laplace Distribution

If a researcher estimates that data follow a distribution which has a higher kurtosis than a normal distribution, one possibility is that the data follow a Laplace distribution. Laplace distribution can be a candidate. We consider Laplace ($\mu = 0, b = 1$) distribution as an alternative for a normality test where $\mu$ is a location parameter and $b$ is a scale parameter. Figure 3.1 (b) shows a density plot of Laplace(0,1) and the normal distribution with the same mean and variance with the Laplace distribution. We generate $m = 200$ independent data sets, and half of them are from Laplace(0,1). The rest of the setting are the same as the example of the uniform distribution. Figure 3.3 shows histograms of p-values by four different multiple normality tests with 20 breaks. With the histograms, AD, SW, and LF tests have more power than the PC test at the level of $\alpha = 0.05$. The remarkable thing is that the LF test is as powerful as the AS or SW test for the Laplace alternative. Thus, we recommend the AD, SW, and LF tests when the expected alternative is a Laplace distribution.
Figure 3.2: Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $U(0, 1)$.
Figure 3.3: Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is Laplace(0,1)
3.2.3 $\chi^2$ Distribution

If a researcher infers that the data follow an asymmetric distribution, one possibility is that the data follow a $\chi^2$ distribution. We consider $\chi^2(\text{df})$ distribution where df are randomly picked from the set of \{1,2,3,4,5\} as an alternative for a normality test. Figure 3.1 (c) shows density plots of $\chi^2(1)$, $\chi^2(2)$, $\chi^2(3)$, $\chi^2(4)$, and $\chi^2(5)$ distributions and the normal distribution with the same mean and variance with $\chi^2(5)$ distribution.

Similar to with the uniform and Laplace alternative distributions, we generate $m = 200$ independent data sets, and half of them are from $\chi^2(\text{df})$ distribution where df are randomly picked from the set of \{1,2,3,4,5\}. The rest of the setting are the same as the example of uniform and Laplace distributions. Figure 3.4 shows histograms of p-values by four different multiple normality tests when the alternative is the $\chi^2$ distribution. The SW test is remarkably the most powerful at the level of $\alpha = 0.05$.

3.2.4 Mixture of Normal Distributions

A mixture of normal distributions can be another candidate as an alternative if a researcher estimates that the alternative is a bimodal or multi-modal distribution. We consider a mixture distribution of two different normal distributions $N(0,1)$ and $N(4,1)$, and combine them by a Bernoulli random choosing with a half and half probabilities. Figure 3.1 (d) shows a density plot of the mixture distribution and the normal distribution as the same mean and the variance with the mixture distribution.

In a similar way to the normality testing with other alternative distributions, we generate $m = 200$ independent data sets from $\frac{1}{2}N(0,1) + \frac{1}{2}N(4,1)$. The rest of the settings are
Figure 3.4: Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\chi^2(df)$ where df is randomly picked from the set of \{1,2,3,4,5\}
the same as other distributions. Figure 3.5 shows histograms of p-values by four different multiple normality tests when the alternative is bimodal. Is the histograms, all of the normality tests are powerful at the level of $\alpha = 0.05$.

Overall, the AD and SW tests are more powerful than the PC or LF test for any alternative distribution of the data. The asymmetric or bimodal alternatives are more detectable for all of those normality tests. Especially, the mixture normal distribution is the most detectable as a non-normal distribution.

3.3 Trade-off between FDR and NDR when $\pi_0$ is Known

In this section, we discuss the trade-off between FDR and NDR with the following alternatives: uniform, Laplace, $\chi^2$, and a mixture distribution. We use the same simulation examples as in Section 3.2. We assume $\pi_0$ is known to be 0.5, and FDR and NDR can be calculated using equations (2.1) and (2.2). In practice, the real $\pi_0$ might be unknown, and FDR and NDR values cannot be calculated, but should be estimated. Thus, we assume $\pi_0$ is unknown in Section 3.3, and estimate the $\pi_0$, the number of rejections, and NDR with $\lambda = 0.5$ using equations (2.7), (2.9), and (2.11) respectively. We are also interested in $E[N]/m$ which is the proportion of non-rejection among all $m$ tests. We call this quantity Proportion Of Non-rejection (PON). We conduct the AD, SW, PC, and LF normality tests for comparison.
Figure 3.5: Histogram of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\frac{1}{2}N(0, 1) + \frac{1}{2}N(4, 1)$.
Figure 3.6: NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is U(0,1). The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2.
The calculated FDR, NDR and PON are shown in Figure 3.6 when the alternative is U(0,1). For all of the normality tests, the FDR is more slowly increasing compared to the rate of decrease of NDR when $\alpha$ increase from 0 to 0.5 on a grid of 0.01. The FDR increases by about 0.05 and the NDR decreases by about 0.2 when $\alpha$ is increased from 0.05 to 0.1 with the AD and SW tests. When $\alpha = 0.2$, the NDR is around 0.4 with the AD test, and 0.2 with the SW test, which indicates the NDR is more quickly dropped for the SW than AD test. With the PC and LF tests, the FDR value is not changed a lot as increasing $\alpha$ values. For example, FDR value is about 0.3 when $\alpha=0.05$, and it is around 0.4 with the PC or LF tests. For the PC test, the FDR and NDR values have stepped lines because the PC test uses finite classes to calculate a $\chi^2$ statistic in equation (3.1). Thus, even though the data sets are different, the p-values could be exactly same if the number of observed frequencies are same in each class. In the generated data, many p-values are exactly same for the PC test, thus the FDR and NDR values are not changed with a small change of $\alpha$. The distance between PON and NDR is $(F_N + T_N)/m - F_N/m = (T_N - F_N)/200$. Thus, larger area between PON and NDR indicates that the test is better at declaring normal as normal and non-normal as non-normal. The AD and SW tests have larger area between PON and NDR than others as shown in the Figure 3.6. For the PC and LF tests, the FDR value are remarkably bigger than AD or SW tests, and the NDR values are bigger and the area between PON and NDR is smaller than AD or SW tests. The AD and SW tests have similar pattern, but the SW test has a little smaller FDR and NDR values, and larger area between PON and NDR for any $\alpha$ value than the AD test. Thus, we recommend the SW or AD tests when the alternative is a uniform distribution.

When the alternative is Laplace(0,1), the AD, SW, and LF have similar patterns for
Figure 3.7: NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is Laplace(0,1). The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2.
FDR and NDR and PON values as shown in Figure 3.7. When $\alpha$ is smaller than 0.3, the PC test has bigger FDR and NDR values than other tests. Even though the SW test has a little bigger NDR values and smaller area between PON and NDR than the AD test, and the AD, SW, and LF tests show similar performance for the Laplace alternative.

When the alternative is the $\chi^2$ distribution, the result of all of the normality tests are similar, especially, the AD and SW have similar patterns to each other, and the PC and LF tests have similar patterns of the FDR and PON as shown in Figure 3.8. We consider $\frac{1}{5}\chi^2(1) + \frac{1}{5}\chi^2(2) + \frac{1}{5}\chi^2(3) + \frac{1}{5}\chi^2(4) + \frac{1}{5}\chi^2(5)$ as the $\chi^2$ alternative. The NDR is noticeably dropped with a small increase of $\alpha$. For example, when $\alpha=0.05$, the NDR is around 0.2 with the AD and SW tests, and 0.4 with the PC and LF tests. Furthermore, when $\alpha=0.1$, the NDR is around 0.1 with the AD and SW tests, and 0.3 with the PC and LF tests. For all the normality tests, the areas between PON and NDR are larger for the $\chi^2$ alternative than the $U(0,1)$ or Laplace alternatives. Among the four tests, the AD and SW tests have smaller NDR values than the PC and LF tests. Thus, the AD and SW tests have larger areas between PON and NDR values than the PC and LF tests. Thus, we recommend the AD or SW normality tests for the $\chi^2$ alternative.

When the alternative is a mixture of two or more distributions, the FDR and NDR values are dependent of the distance between distributions which are used for the mixture. We consider $\frac{1}{2}N(0,1)+\frac{1}{2}N(4,1)$, so the mean different is 4, and the NDR values are dropped quickly as increasing $\alpha$ value more than $U(0,1)$ or Laplace, but less than $\chi^2$ alternative as shown in Figure 3.9. Among the four normality tests, the AD and SW tests have a smaller NDR values than the PC or LF tests. Because the PON values are similar for all tests, the AD and SW tests have larger area between PON and NDR than the PC or LF
Figure 3.8: NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when the alternative distribution is $\chi^2(\text{df})$ where df is randomly picked from the set of $\{1,2,3,4,5\}$. The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2.
Figure 3.9: NDR, FDR, and PON by $\alpha$ by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) when the alternative distribution is $\frac{1}{2} N(0, 1) + \frac{1}{2} N(4, 1)$. The dashed lines indicates that $\alpha$ is 0.05, 0.1, and 0.2.
tests. Whereas, the FDR values have a similar increasing pattern when the alternative
is a mixture distribution for all of the four tests. Also, the FDR patterns are similar to
the FDR values of the $\chi^2$ alternative. Thus, we recommend the SW or AD tests for the
mixture distribution.

Table 3.1: Comparison of performances of the four normality tests for each alternative.
Note that we prefer smaller FDR, smaller NDR, and larger area between PON and NDR.
The approximation symbol “$>$” indicates the two tests have similar performance. For the
trade-off between FDR and NDR, the greater symbol, “$>$” indicates more cost-effective.

<table>
<thead>
<tr>
<th>$H_a$</th>
<th>FDR</th>
<th>NDR</th>
<th>PON-NDR</th>
<th>Trade-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>SW &lt; AD &lt; LF &lt; PC</td>
<td>SW &lt; AD &lt; LF ≈ PC</td>
<td>SW &gt; AD &gt; LF ≈ PC</td>
<td>SW ≈ AD &gt; PC ≈ LF</td>
</tr>
<tr>
<td>Laplace</td>
<td>AD ≈ SW ≈ LF &lt; PC</td>
<td>AD &lt; SW ≈ LF &lt; PC</td>
<td>AD &gt; SW ≈ LF &gt; PC</td>
<td>AD &gt; SW ≈ LF ≈ PC</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>SW ≈ AD ≈ LF &lt; PC</td>
<td>SW ≈ AD &lt; LF ≈ PC</td>
<td>SW ≈ AD &lt; LF ≈ PC</td>
<td>SW ≈ AD &gt; LF ≈ PC</td>
</tr>
<tr>
<td>Mixture</td>
<td>AD ≈ SW ≈ LF &lt; PC</td>
<td>AD &lt; SW &lt; LF &lt; PC</td>
<td>AD &gt; SW &lt; LF &gt; PC</td>
<td>AD ≈ SW &gt; LF ≈ PC</td>
</tr>
</tbody>
</table>

Table 3.2: Compare the products of four alternatives for each normality test. Note that
we prefer smaller FDR, Smaller NDR, and larger area between PON and NDR. For the
trade-off between FDR and NDR, the greater symbol, “$>$” indicates core cost-effective.
“Uf” indicates the uniform, “La” indicates the Laplace, and “Mx” indicates the mixture
alternatives.

<table>
<thead>
<tr>
<th>Test</th>
<th>FDR</th>
<th>NDR</th>
<th>PON-NDR</th>
<th>Trade-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>Mx ≈ $\chi^2$ &lt; La &lt; Uf</td>
<td>$\chi^2$ &lt; Mx &lt; La &lt; Uf</td>
<td>$\chi^2$ &gt; Mx &gt; Uf ≈ La</td>
<td>$\chi^2$ &gt; Mx &gt; Uf &gt; La</td>
</tr>
<tr>
<td>SW</td>
<td>Mx ≈ $\chi^2$ &lt; La ≈ Uf</td>
<td>$\chi^2$ &lt; Mx &lt; La &lt; Uf</td>
<td>$\chi^2$ &gt; Mx &gt; Uf &gt; La</td>
<td>$\chi^2$ &gt; Mx &gt; Uf &gt; La</td>
</tr>
<tr>
<td>PC</td>
<td>Mx ≈ $\chi^2$ ≈ La &lt; Uf</td>
<td>$\chi^2$ &lt; Mx &lt; La &lt; Uf</td>
<td>$\chi^2$ ≈ Mx &gt; La &gt; Uf</td>
<td>$\chi^2$ &gt; Mx &gt; Uf &gt; La</td>
</tr>
<tr>
<td>LF</td>
<td>Mx ≈ $\chi^2$ &lt; La &lt; Uf</td>
<td>$\chi^2$ &lt; Mx &lt; La &lt; Uf</td>
<td>$\chi^2$ ≈ Mx &gt; La &gt; Uf</td>
<td>$\chi^2$ &gt; Mx &gt; Uf &gt; La</td>
</tr>
</tbody>
</table>

Overall, the SW and AD tests have smaller NDR values than the PC or LF test. The
summary of the discussion in this section is in Tables 3.1 and 3.2. When the alternative
is $\chi^2$ or the mixture distributions, the LF tests also fine in terms of NDR controlling.
However, if the alternative is the uniform, the PC and LF tests have higher error for
both NDR and FDR than the AD or SW tests. The NDR values are quickly dropped
and the FDR values are slowly increased when the alternative is the $\chi^2$ or the mixture distributions as increase of $\alpha$. For the view of the trade-off between FDR and NDR, we can make the NDR values small with a little increase of the FDR for the asymmetric and the mixture distributions. However, when the alternative is the uniform or the Laplace, the NDR values are slowly decreasing as $\alpha$ increases, the NDR value is around 0.6 and FDR value is 0.4 when $\alpha=0.2$ with the PC test when the alternative is the uniform. Thus, it is hard to make the NDR values smaller for a trade-off with a decreasing FDR when the alternative is the uniform or the Laplace distributions.

3.4 Estimation of NDR for given FDR when $\pi_0$ is Unknown

In this section, we discuss an estimation of NDR for given FDR values. We use the same examples in Section 3.2 and 3.3, but assume that the $\pi_0$ is unknown. We follow a procedure of Craiu and Sun (2008). The first step is estimating $\pi_0$ using the equation (2.7) with $\lambda = 0.5$. The next step is choosing an interval for $q \in (0, \hat{\pi}_0)$. The last step is that we estimate the number of rejections ($R$) and NDR values by the equations (2.9) and (2.11) respectively. We also calculate the true FDR and NDR values to see how accurate the estimations are.

3.4.1 The Uniform Alternative

When the alternative is the uniform distribution, the values of $\hat{\pi}_0$ are very different as the tests and $\lambda$ values. The larger $\lambda$ values give larger variances generally, but the bias is increased with small $\lambda$ values. With $\lambda = 0.5$, the SW test estimates $\pi_0$ very well as
shown in Figure 3.10. For the AD test, \( \pi_0 = 0.5 \) is in the range of the \( \hat{\pi}_0 \) when \( \lambda = 0.5 \), but choosing \( \lambda \geq 0.6 \) might be better for the \( \pi_0 \) estimation. For the PC test, \( \pi_0 = 0.5 \) is out of the range of the \( \hat{\pi}_0 \) when \( \lambda = 0.5 \), but if \( \lambda \) is bigger or equal to 0.6, \( \hat{\pi}_0 \) can be estimated pretty well. For the LF test, the estimated values of \( \pi_0 \) is far from the true \( \pi_0 \) which is 0.5. We choose \( q \) values in \((0,0.5)\) for the four normality tests. The grid is 0.05 for the boxplots of \( \hat{\text{NDR}} \) and 0.01 for the average values of \( \hat{\text{NDR}} \).

For the AD and SW tests, the \( \hat{\text{NDR}} \) values are very similar with the true NDR values, and the true FDR values are also very similar with the given \( q \) values as shown in Figure 3.11. Whereas, for the PC and LF tests, NDR values are overestimated and the true FDR values are also biased compared to the given \( q \) values as shown in Figure 3.12. As shown in the left panels in Figures 3.11 and 3.12, the PC and LF tests have the variances of \( \hat{\text{NDR}} \) that are much bigger than the AD or SW tests. For the view of the trade-off between NDR and FDR, the FDR with the SW test is the most cost-effective. For example, when \( q = 0.2 \), the \( \hat{\text{NDR}} \) values are 0.483, 0.213, 0.928, 0.939 for the AD, SW, PC, and LF test respectively. We calculate cost-effectiveness using the equations (2.5) and (2.6) for the SW test as shown in Table 3.3. Among the three points, \( q = 0.3 \) has the most cost-effectiveness for the SW test.

For accuracy of estimating NDR, small variances of \( \hat{\text{NDR}} \), and the cost-effectiveness of the trade-off, we recommend the SW or AD tests when the alternative is the uniform, and this recommendation is in line with the result of Section 3.3.
Figure 3.10: Boxplot of $\hat{\pi}_0$ by $\lambda$ in $(0,1)$ on a grid of 0.1 for multiple Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0, 1)$, and the alternative distribution is $U(0, 1)$. The dashed line is true $\pi_0 = 0.5$.

Table 3.3: Compare of effectiveness between $q = 0.1$, 0.2, and 0.3 for the SW test with $\omega_{ratio} = \omega_{slope} = 1$ when the alternative is $U(0, 1)$.

<table>
<thead>
<tr>
<th></th>
<th>0.1 vs. 0.2</th>
<th>0.2 vs. 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{ratio}$</td>
<td>0.591</td>
<td>0.261</td>
</tr>
<tr>
<td>$e_{slope}$</td>
<td>-5.077</td>
<td>-1.757</td>
</tr>
</tbody>
</table>
Figure 3.11: Box plots of $\hat{N}_{DR}$ (left panels), and average values of $\tilde{N}_{DR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $U(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
Figure 3.12: Box plots of $\hat{N}_{\text{DR}}$ (left panels), and average values of $\hat{N}_{\text{DR}}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $U(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
3.4.2 The Laplace Alternative

When the alternative is the Laplace distribution, the estimation of $\pi_0$ is not accurate as shown in Figure 3.13 with $\lambda = 0.5$. However, the median values of the $\hat{\pi}_0$ are exactly 0.5 for the PC test when $\lambda \geq 0.7$. The average values of $\hat{\pi}_0$ with $\lambda = 0.5$ are 0.693, 0.722, 0.691, and 0.747 for the AD, SW, PC, and LF tests respectively. We choose $q$ values in (0, 0.5) for all four normality tests.

The AD and SW tests have exactly same average values of $\hat{\text{NDR}}$, NDR, and FDR when $q \leq 0.47$. For the AD and SW tests, the $\hat{\text{NDR}}$ values are underestimated as shown in Figure 3.14, and the true FDR values are very similar with the give $q$ values. For the PC test, the estimation of NDR is the most accurate among the four tests, but FDR values are not controlled at the level of $q$ as shown in Figure 3.15. For the LF test, the $\hat{\text{NDR}}$ values are also underestimated, and FDR values are controlled at the level of $q$ as shown in Figure 3.15. For the view of the trade-off between NDR and FDR, the PC test is the least cost-effective. For example, when $q = 0.2$, the $\hat{\text{NDR}}$ values are 0.325, 0.325, 0.744, 0.441 for the AD, SW, PC, and LF tests respectively. We calculate cost-effectiveness for the SW test as shown in Table 3.4, note that the AD test have the same result when the alternative is the Laplace. Among the three points, $q = 0.1$ has the most cost-effectiveness with the measure of the ratio, whereas $q = 0.3$ has the most cost-effectiveness with the measure of slope.

Even though the AD and SW tests show less accuracy of estimating NDR than the PC or LF tests, they have smaller variances of $\hat{\text{NDR}}$, and more cost-effectiveness of the trade-off between the FDR and NDR. Thus, we recommend the SW or AD tests when the alter-
Figure 3.13: Boxplot of $\hat{\pi}_0$ by $\lambda$ in (0,1) on a grid of 0.1 for multiple Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is Laplace$(0,1)$. The dashed lines indicate true $\pi_0 = 0.5$.

Table 3.4: Compare of effectiveness between $q = 0.1$, 0.2, and 0.3 for the SW (or AD) test with $\omega_{\text{ratio}} = \omega_{\text{slope}} = 1$ when the alternative is Laplace.

<table>
<thead>
<tr>
<th>$q$</th>
<th>0.1 vs. 0.2</th>
<th>0.2 vs. 0.3</th>
<th>0.1 vs. 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{\text{ratio}}$</td>
<td>1.305</td>
<td>0.910</td>
<td>1.188</td>
</tr>
<tr>
<td>$e_{\text{slope}}$</td>
<td>-1.729</td>
<td>-1.277</td>
<td>-1.503</td>
</tr>
</tbody>
</table>
Figure 3.14: Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is Laplace(0,1). The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
Figure 3.15: Box plots of $\hat{N}\text{DR}$ (left panels), and average values of $\hat{N}\text{DR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is Laplace$(0,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
native is the Laplace in terms of the trade-off between the FDR and NDR.

3.4.3 The $\chi^2$ Alternative

Another alternative for a normality test is the $\chi^2$ distribution. The estimation of $\pi_0$ is much more improved than the uniform and Laplace alternatives as shown in Figure 3.16. For the AD and SW test, the $\hat{\pi}_0$ values are stable around the true $\pi_0 = 0.5$ for all range of $\lambda$. For the PC and LF tests, the $\hat{\pi}_0$ is overestimated than the true $\pi_0$ with $\lambda = 0.5$. Since the average values of $\hat{\pi}_0$ with $\lambda = 0.5$ are 0.518, 0.508, 0.542, and 0.569 for the AD, SW, PC, and LF tests respectively, and we choose $q$ values in $(0,0.5)$ for the all four normality tests.

When the alternative is the $\chi^2$, both of the FDR and NDR are much smaller than the uniform or Laplace alternative as shown in Figures 3.17 and 3.18. For the AD and SW tests, the $\chi^2$ alternative has smaller variances of $\hat{\text{NDR}}$ than the uniform or Laplace alternatives as shown in the left panels of Figures 3.11, 3.14, and 3.17. Also, the $\hat{\text{NDR}}$ values are similar with the true NDR values, and the FDR values are well controlled at the level of $q$. Especially, the $\hat{\text{NDR}}$ values are very similar with the NDR values, and FDR values are also very similar with the $q$ values for the SW test. For the PC test, the variances of the $\hat{\text{NDR}}$ are remarkably bigger than the AD or SW tests as shown in the upper left panel of Figure 3.18. The $\hat{\text{NDR}}$ values have a similar pattern with the true NDR values, but $\hat{\text{NDR}}$ values are increasing when $q \geq 0.34$, because the number of rejections ($R$) is very slowly increased, whereas $q$ values are increased regularly in the equation (2.11) as shown in the upper right panel in Figure 3.18. Also, the FDR values
are not controlled at the level of \( q \) for the PC test. For the LF test, the \( \hat{NDR} \) values are well estimated compared to the true NDR values, and FDR is controlled at \( q \) as shown in the bottom right panel of Figure 3.18. However, the variances of \( \hat{NDR} \) are similarly large with the PC test as shown in the bottom left panel of Figure 3.18. For the view of the trade-off between NDR and FDR, the SW test is the most cost-effective, and the AD test is less but similar with the SW test, whereas the PC test is the least. For example, when \( q = 0.2 \), the \( \hat{NDR} \) values are 0.052, 0.039, 0, 0.133 for the AD, SW, PC, and LF tests respectively. For all four normality tests, the \( \chi^2 \) alternative is more cost-effective than the uniform of Laplace alternative for the view of the trade-off between the FDR and NDR. We calculate cost-effectiveness for the SW test with \( q = 0.05, 0.1, \) and 0.2 as shown in Table 3.5. Among the three points, \( q = 0.05 \) has the most cost-effectiveness with the measure of the ratio, whereas \( q = 0.2 \) has the most cost-effectiveness with the measure of slope.

Similar to the uniform alternative, we recommend the SW or AD tests when the alternative is the \( \chi^2 \) distribution since the estimation of NDR is more accurate, \( \hat{NDR} \) values have smaller variances, and the FDR and NDR values are more cost-effective than the PC or LF tests. This recommendation is in line with the result of in Section 3.3.

Table 3.5: Compare of effectiveness between \( q = 0.05, 0.1, \) and 0.2 for the SW (or AD) test with \( \omega_{ratio} = \omega_{slope} = 1 \) when the alternative is \( \chi^2 \).

<table>
<thead>
<tr>
<th>( q )</th>
<th>0.05 vs. 0.1</th>
<th>0.1 vs. 0.2</th>
<th>0.05 vs. 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e_{ratio} )</td>
<td>1.147</td>
<td>0.851</td>
<td>0.976</td>
</tr>
<tr>
<td>( e_{slope} )</td>
<td>-1.371</td>
<td>-0.530</td>
<td>-0.811</td>
</tr>
</tbody>
</table>
Figure 3.16: Boxplot of $\hat{\pi}_0$ by $\lambda$ in (0,1) on a grid of 0.1 for multiple Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\chi^2(df)$ where df is randomly picked from the set of \{1,2,3,4,5\}. The dashed line is true $\pi_0 = 0.5$. 
Figure 3.17: Box plots of $\widehat{\text{NDR}}$ (left panels), and average values of $\widehat{\text{NDR}}$, true NDR (right panels) by given $q$ values in (0,0.5) on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\chi^2(\text{df})$ where df is randomly picked from the set of \{1,2,3,4,5\}. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
Figure 3.18: Box plots of $\hat{\text{NDR}}$ (left panels), and average values of $\hat{\text{NDR}}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\chi^2(\text{df})$ where df is randomly picked from the set of $\{1,2,3,4,5\}$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
3.4.4 The Normal Mixture Alternative

The last alternative for the normality test is the mixture distribution of $\frac{1}{2}N(0,1) + \frac{1}{2}N(4,1)$. The $\pi_0$ is well estimated with $\lambda = 0.5$ for the AD, SW, and PC tests as shown in Figure 3.19. For the AD and SW tests, the median values of $\hat{\pi}_0$ are very stable for all given $\lambda$ values. For the PC test, the true $\pi_0$ is the same with the median value of $\hat{\pi}_0$ with $\lambda = 0.5$. The average values of $\hat{\pi}_0$ with $\lambda = 0.5$ are 0.508, 0.519, 0.504, and 0.552 for the AD, SW, PC, and LF tests respectively. Since the values are all greater than 0.5, we choose $q$ values in (0,0.5) for the all four normality tests.

When the alternative is the mixture normal distribution, the NDR is well estimated which means that the $\hat{\text{NDR}}$ values are similar with the true NDR values for all four tests, and the FDR is also well controlled at the level of $q$ as shown in Figures 3.20 and 3.21. For the AD test, the variances of the $\hat{\text{NDR}}$ are the smallest compared to other tests, and also compared to other alternatives as shown in the top left panel of Figure 3.20. The variances of the $\hat{\text{NDR}}$ are also small for the SW test when $q \geq 0.1$ as shown in the bottom left panel of Figure 3.20. For the PC test, the NDR is a little underestimated when $q < 0.46$, but it is overestimated when $q \geq 0.46$ as shown in the top right panel in Figure 3.21. The variances of the $\hat{\text{NDR}}$ for the mixture normal alternative are small compared to other alternatives for the PC test. For the LF test, the $\hat{\text{NDR}}$ has similar pattern with the PC test, but the variances of the $\hat{\text{NDR}}$ are larger than the PC test for the most given $q$ values.

For the view of the trade-off between NDR and FDR, the AD and SW tests are the more cost-effective than the PC or LF tests. For all four normality tests, the mixture alternative is more cost-effective than the uniform of Laplace alternatives, but less than
the $\chi^2$ alternative. As an example, we calculate cost-effectiveness for the SW test with $q = 0.05, 0.1,$ and $0.2$ as shown in Table 3.6. Among the three points, $q = 0.2$ has the most cost-effectiveness for both of the measures.

Table 3.6: Compare of effectiveness between $q = 0.05, 0.1,$ and $0.2$ for the SW (or AD) test with $\omega_{ratio} = \omega_{slope} = 1$ when the alternative is $\chi^2.$

<table>
<thead>
<tr>
<th>$q$</th>
<th>0.05 vs. 0.1</th>
<th>0.1 vs. 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_{ratio}$</td>
<td>0.822</td>
<td>0.424</td>
</tr>
<tr>
<td>$\epsilon_{slope}$</td>
<td>-6.857</td>
<td>-1.884</td>
</tr>
</tbody>
</table>

Similar to the uniform or $\chi^2$ alternatives, we recommend the AD or SW tests, since the estimation of NDR is more accurate than the PC test, the AD and SW tests have smaller variances of $\hat{NDR}$ than the LF tests, and they have more cost-effectiveness of the trade-off between the FDR and NDR than the PC or LF tests. This recommendation is in line with the result of in Section 3.3.

Table 3.7 shows the summary of the discussion in this section. The chosen $\lambda = 0.5$ works well to estimate $\pi_0$ for the SW test with the uniform alternative, for the AD and SW tests with the $\chi^2$ alternative, and for the AD, SW, and PC tests with the mixture alternative. When the alternative is the Laplace, the median value of $\hat{\pi}_0$ is stable for the AD, SW, and LF tests, even though the estimation is not accurate. Also, the median value of $\hat{\pi}_0$ is stable near the true $\pi_0$ for the AD and SW tests with the $\chi^2$ and mixture alternatives. For the estimation of NDR, the AD and SW tests are more accuracy than the PC or LF test except for the Laplace alternative. The PC test is less accuracy than other tests generally for the estimation of NDR, but the test is most accurate with the Laplace alternative. For the all four alternatives, the AD and SW tests have smaller
Figure 3.19: Boxplot of $\hat{\pi}_0$ by $\lambda$ in $(0,1)$ on a grid of 0.1 for multiple Anderson-Darling normality testing when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, the null distribution is $N(0, 1)$, and the alternative distribution is $\frac{1}{2}N(0, 1) + \frac{1}{2}N(4, 1)$. The dashed line is true $\pi_0 = 0.5$. 
Figure 3.20: Box plots of $\hat{\text{NDR}}$ (left panels), and average values of $\hat{\text{NDR}}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Anderson-Darling and Shapiro-Wilk normality test (top-right panel) when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\frac{1}{2}N(0,1)+\frac{1}{2}N(4,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
Figure 3.21: Box plots of $\hat{NDR}$ (left panels), and average values of $\hat{NDR}$, true NDR (right panels) by given $q$ values in $(0,0.5)$ on a grid of 0.01 by Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m=200$, $n=30$, repetition=50, $\pi_0 = 0.5$, $\lambda = 0.5$, the null distribution is $N(0,1)$, and the alternative distribution is $\frac{1}{2}N(0,1)+\frac{1}{2}N(4,1)$. The dashed vertical and horizontal lines are FDR and NDR values of three picked points, and the dotted increasing line is $q$ on the right panels.
variance of $\hat{NDR}$ than the PC or LF tests. For the FDR, the AD and SW tests are more accurate which means that the FDR values are more similar with the given $q$ values than the PC or LF test. When the alternative is $\chi^2$ or mixture, the FDR values are similar with the given $q$ values for the all four tests. The LF test controls well the FDR at the level of given $q$. For the trade-off between the FDR and NDR, the SW and AD tests are more cost-effective than the PC or LF tests. Overall, the $\chi^2$ and the mixture normal alternatives are better for the estimation of $\hat{\pi}_0$ and $\hat{NDR}$, and they are more cost-effective for the trade-off between FDR and NDR than the uniform or Laplace alternatives.

Table 3.7: Comparison of performances of the four normality tests for each alternative about the estimation of $\hat{\pi}_0$ and NDR, the accuracy and control of FDR at level $q$, and the cost-effectiveness of trade-off between FDR and NDR.

<table>
<thead>
<tr>
<th>$H_a$</th>
<th>$\hat{\pi}_0$</th>
<th>$\hat{NDR}$</th>
<th>FDR</th>
<th>Trade-off</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda = 0.5$</td>
<td>Stability</td>
<td>Accuracy</td>
<td>Var</td>
</tr>
<tr>
<td>Uniform</td>
<td>SW</td>
<td>None</td>
<td>AD,SW</td>
<td>SW</td>
</tr>
<tr>
<td>Laplace</td>
<td>None AD,SW,LF</td>
<td>PC AD,SW</td>
<td>AD,SW</td>
<td>AD,SW,LF</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>AD,SW AD,SW</td>
<td>AD,SW,LF</td>
<td>SW</td>
<td>AD,SW,LF</td>
</tr>
<tr>
<td>Mixture</td>
<td>AD,SW,PC AD,SW</td>
<td>AD,SW,LF</td>
<td>AD</td>
<td>All</td>
</tr>
</tbody>
</table>

65
3.5 Real Data Example

The data set, Wisconsin diagnostic breast cancer (Wolberg et al., 1993), is taken from the UCI machine learning repository [http://archive.ics.uci.edu/ml]. Total 569 tissue images were analyzed, and 357 were classified as benign and 212 were classified as malignant. The computer vision diagnostic system extracts ten real-valued features for each cell nucleus.

a) radius (mean of distances from center to points on the perimeter)
b) perimeter
c) area
d) compactness (perimeter$^2$/area )
e) smoothness (local variation in radius lengths)
f) concavity (severity of concave portions of the contour)
g) concave points (number of concave portions of the contour)
h) symmetry
i) fractal dimension (coastline approximation)
j) texture (standard deviation of gray-scale values)

The mean value, standard error, and worst or largest (average of the three worst or largest values) of each features were computed for each image. Thus, we have total 30 features for each image. Many papers use this data set with a normality assumption (Wolberg et al., 1993, 1995). We conduct multiple normality tests for each variable and estimate NDR values. The frequency histograms of each variable for malignant and benign samples are shown in Figures 3.22 and 3.23 respectively.
Figure 3.22: Frequency histograms of 30 variables for 212 malignant images.
Figure 3.23: Frequency histograms of 30 variables for 357 benign images.
3.5.1 Estimation of $\pi_0$

The data set has 30 real-valued variables, but we had to test for benign and malignant separately for each variable. Thus, the total number of tests was $m = 60$. The sample size for the data is 357 for benign and 212 for malignant samples. We conduct the AD, SW, PC, and LF normality tests for each variable. We expect approximately half of them are decided to follow a normal distribution based on the histograms in Figures 3.22 and 3.23. However, most of the p-values are under 0.05, so the histograms of p-values are right-skewed as shown in Figure 3.24 which means most of the variables seem to be non-normal distributions. This result is probably due to the large sample size. Thus, we randomly pick sub-samples ($n = 30$) from the data set without replacement, repeat 100 times, and calculate the average frequency for each bin of the histogram as shown in Figure 3.25. The p-values are more spread to the right side than the p-values in Figure 3.24. The AD and SW tests have a similar pattern. For the LF test, the frequency graph seems to be less right skewed than the AD or SW tests. The PC test has a quite different distribution. Next, we estimate the $\pi_0$ with the whole data as shown in Figure 3.26, and with the repeated sub-sample data as shown in Figure 3.27. For the whole data, the $\hat{\pi}_0$ values are 0.067, 0.167, 0.125, and 0.167 for the AD, SW, PC, and LF tests respectively with $\lambda = 0.5$. For the sub-sample data with 100 repetitions, the $\hat{\pi}_0$ values are 0.394, 0.369, 0.508, and 0.479 for the AD, SW, PC, and LF tests respectively with $\lambda = 0.5$. Thus, with the sub-sample data, the $\hat{\pi}_0$ is much larger than the value with the whole data set for all four tests. For the sub-sample data, the $\hat{\pi}_0$ values are decreased by increasing $\lambda$ values for all four tests, and average values of $\hat{\pi}_0$ are more stable for the SW test than other tests as shown in Figure 3.27.
Figure 3.24: Histograms of p-values by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors normality tests when $m = 60$, $n_{\text{malignant}} = 212$, $n_{\text{benign}} = 357$. 
Figure 3.25: Histograms of p-values for average counts in the bin locations by Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors normality tests when $m=60$, $n_{sub-sample}=30$, repetition=100.
Figure 3.26: Estimation of $\pi_0$ by $\lambda$ in (0,1) on a grid of 0.1 for Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m = 60$, $n_{\text{malignant}} = 212$, $n_{\text{benign}} = 357$. The dashed lines indicate $\hat{\pi}_0$ when $\lambda = 0.5$. 
Figure 3.27: Estimation of $\pi_0$ by $\lambda$ in (0,1) on a grid of 0.1 for Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m = 60$, $n_{sub-sample} = 30$, repetition=100. The dotted horizontal lines indicate $\hat{\pi}_0$ when $\lambda = 0.5$. 
3.5.2 Estimation of NDR for given FDR

We estimate the NDR with four different \( \lambda \) values for each test with the whole data and the repeated sub-sample data as shown in Figures 3.28 and 3.29. With the whole data, the \( \hat{NDR} \) values are much less than the \( \hat{NDR} \) values with the sub-sample data. Especially, \( \hat{NDR} \) for the SW test is almost zero for all given \( q \) values. The big sample size might make the NDR extremely small. Thus, we estimate NDR, and discuss the trade-off between FDR and NDR with the sub-sample data. We tried \( \lambda = 0.3, 0.4, 0.5, \) and 0.6 to estimate NDR, and for all \( \lambda \) values which we tried, the SW test has the smallest \( \hat{NDR} \) values, the AD test is the next, the LF test is the next, and the PC test has the largest values. The \( \hat{NDR} \) values seem to not be very different for the four different \( \lambda \) values. We keep \( \lambda = 0.5 \) as Chapters 2 and 3. As similar with the \( \hat{NDR} \) pattern, the SW has the smallest \( \hat{PON} \), the AD test is the next, the LF is the next, and the PC has the largest \( \hat{PON} \) for all given \( q \) values as shown in Figure 3.29. The areas between the \( \hat{PON} \) and \( \hat{NDR} \) are similar for all four tests. Among the four alternative distributions in Chapter 3, the mixture normal distribution has the most similar \( \hat{NDR} \) patterns with this data set. For the view of the trade-off between FDR and NDR, the SW test is the most cost-effective, and this is in line with the result in Table 3.7. We calculate the two measures (2.5) and (2.6) of cost-effectiveness for the SW test as shown in Table 3.8. Among the three chosen points, \( q = 0.2 \) is the most cost-effective with the measure of the ratio, whereas \( q = 0.1 \) is the most cost-effective with the measure of slope.

Table 3.8: Compare of effectiveness between \( q = 0.05, 0.1, \) and 0.2 for the SW test with \( \omega_{ratio} = \omega_{slope} = 1 \) when \( m = 60, n_{sub-sample} = 30, \) repetition=100.

<table>
<thead>
<tr>
<th>( q )</th>
<th>0.05 vs. 0.1</th>
<th>0.1 vs. 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \epsilon_{ratio} )</td>
<td>0.999</td>
<td>0.398</td>
</tr>
<tr>
<td>( \epsilon_{slope} )</td>
<td>-2.285</td>
<td>-0.914</td>
</tr>
</tbody>
</table>

74
Figure 3.28: Estimation of PON and NDR by given $q$ with $\lambda = 0.5$ for Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m = 60$, $n_{malignant} = 212$, $n_{benign} = 357$. 
Figure 3.29: Estimation of PON and NDR by given $q$ in (0,0.38) on a grid of 0.1 with $\lambda = 0.5$ for Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors (Kolmogorov-Smirnov) normality tests when $m = 60$, $n_{sub-sample} = 30$, repetition=100.
Chapter 4

Conclusions

A hypothesis test is classified as an acceptance-support test or a rejection-support test. The first purpose of this thesis is to discuss the classification in terms of multiple testing. We argue the decision errors in each type of test in Chapter 1. Type I errors in RS testing have similar meaning to type II errors in AS testing. The second purpose is the error control in RS and AS multiple testing. The Bonferroni correction, and FDR control methods are to control false positives in RS multiple testing, but false negative or type II error are sharply increased with the false positive control methods as shown in Chapter 2. Thus, we discuss the trade-off between FDR and NDR to control false negative for AS multiple testing. As an example of multiple AS testing, we conduct multiple normality testing in Chapters 3 and 4. We choose four popular normality tests: Anderson-Darling, Shapiro-Wilk, Pearson chi-square, and Lilliefors, and four alternative distributions: uniform, Laplace, $\chi^2$, and mixture normal. We discuss the estimation of $\pi_0$ and NDR, and the trade-off between FDR and NDR in Chapter 3. The cost-effectiveness of the trade-off is different for each normality test and each alternative. The AD and SW tests are more cost-effective than the PC or LF test for all alternatives, and the $\chi^2$ and mixture
normal alternatives have less NDR values than the uniform or Laplace alternative for all normality tests. In Chapter 4, we conduct the same four normality tests with a breast cancer data set to apply our discussion to a real data.

The contribution of this thesis is that we illuminate the classification the hypothesis testing to two types, and clarify the role of decision errors in each type of testing. Moreover, we extend the methods of estimation of $\pi_0$ and NDR, and the trade-off between FDR and NDR to multiple AS testing. Also, we apply the normality tests and estimate NDR to the real data.

We estimate $\pi_0$ with a fixed $\lambda = 0.5$ for all tests and all alternatives. However, we do not discuss the optimization of $\lambda$ for our testing. The $\hat{\pi}_0$ is critical to the accuracy of the estimation of the number of rejections and NDR. Black (2004), Langaas et al. (2005), Schweder and Spjøtvoll (1982), Storey (2002) study how to estimate $\pi_0$. With their method, we could estimate $\pi_0$ more accurately and differently for each test and alternative, and estimate NDR more accurately. Furthermore, we choose an interval for $q \in (0, \hat{\pi}_0)$, and estimate the number of rejections and NDR values in this thesis. A more appropriate way for multiple AS testing is to fix NDR values and estimating FDR values. However, the number of rejections is dependent for the FDR value, and it is hard estimate FDR by a given NDR for a normality test. It might be a future work. Lastly, the sample size is critical for the p-value. For example, if a sample size is too big, most of the the null hypothesis might be rejected. We keep $n_{\text{sub-sample}} = 30$ for the simulated data and pick the sub-sample $n = 30$ from the entire real data set. If the sample size is different, the result might be changed. Sample size might be more critical for some tests than others. Thus, sample size effect is another issue for the future.
Bibliography


