

# MULTIPLE SIGNIFICANCE TESTS AND THEIR RELATION TO $P$ -VALUES

A Thesis Submitted to the College of  
Graduate Studies and Research  
In Partial Fulfillment of the Requirements  
For the Degree of Master of Science  
In the Department of Mathematics and Statistics  
University of Saskatchewan  
Saskatoon

By

Xiao Bo (Alice) Li

©Copyright Xiao Bo (Alice) Li, August 2008. All rights reserved.

## PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics  
McLean Hall  
106 Wiggins Road  
University of Saskatchewan  
Saskatoon Saskatchewan  
S7N 5E6

## ABSTRACT

This thesis is about multiple hypothesis testing and its relation to the  $P$ -value. In Chapter 1, the methodologies of hypothesis testing among the three inference schools are reviewed. Jeffreys, Fisher, and Neyman advocated three different approaches for testing by using the posterior probabilities,  $P$ -value, and Type I error and Type II error probabilities respectively. In Berger's words "Each was quite critical of the other approaches." Berger [8] proposed a potential methodological unified conditional frequentist approach for testing. His idea is to follow Fisher in using the  $P$ -value to define the strength of evidence in data and to follow Fisher's method of conditioning on strength of evidence; then follow Neyman by computing Type I and Type II error probabilities conditioning on strength of evidence in the data, which equal the objective posterior probabilities of the hypothesis advocated by Jeffreys [26].

Bickis [3] proposed another estimate on calibrating the null and alternative components of the distribution by modeling the set of  $P$ -values as a sample from a mixed population composed of a uniform distribution for the null cases and an unknown distribution for the alternatives. For tackling multiplicity, exploiting the empirical distribution of  $P$ -values is applied. A variety of density estimators for calibrating posterior probabilities of the null hypothesis given  $P$ -values are implemented. Finally, a noninterpolatory and shape-preserving estimator based on B-splines as smoothing functions is proposed and implemented.

## ACKNOWLEDGEMENTS

I am most indebted to my supervisor, Professor Mikelis G. Bickis, for accepting me as a graduate student in the Department of Mathematics and Statistics. This occurred at a crucial time for me, at a point when my dreams had been broken. I appreciate his invaluable guidance, his tireless encouragement, his patient understanding, and his magnificent support during my Master of Science program, as well as the critical review and comments on this thesis. Despite my initial apprehensions, each minute of my weekly meetings with him enriched my pleasant memories in the Department of Mathematics and Statistics. Many of the ideas were incubated in the discussion with him. Thank you Prof. Bickis for all your contributions.

My advisory committee members, Professor Chris Soteros, Professor John Martin, and Professor Raj Srinivasan, provided an excellent study environment and made my program here fruitful. I am most appreciative! I would like to extend my thankful thoughts to Professor Kelly to be on my thesis committee. I am happy to have met and worked with them over the past years.

I would like to extend my sincere gratefulness to Professors Longhai Li and Xulin Guo for their help beyond my area of study.

I want to thank my friends and colleagues, in particular Yaling Yin, Mahshid Atapour, and Zhidong Zhang, for their friendship and help that made my study life easy and interesting.

Special thanks to all of you who have been helpful and kind to me from the first day of my transfer here.

I would also like to acknowledge the financial support of the Department of Mathematics and Statistics, the Centennial Equity Merit Scholarship of University of Saskatchewan, and the Scholarship of the Canadian Federation of University Women Saskatoon Canada.

Last but by no means least, I wish to thank my little boy, Richard Jia Li, for his patient understanding and loving support. As a mother, I owe a lot to him.

# Contents

<b>1</b>	<b>THE TESTING PROCEDURES FROM THREE INFERENCE SCHOOLS</b>	<b>3</b>
1.1	Hypothesis testing . . . . .	3
1.1.1	Bayesian methodology for hypothesis testing . . . . .	4
1.1.2	Fisher's $P$ -value for hypothesis testing . . . . .	7
1.1.3	Neyman-Pearson frequentist methodology for hypothesis testing . . . . .	8
1.1.4	Connecting $P$ -value and Neyman-Pearson testing and randomizing a test . . . . .	9
1.2	Example . . . . .	10
1.3	The unified conditional frequentist testing . . . . .	12
1.3.1	Statistic and an unknown parameter . . . . .	12
1.3.2	Introduction to conditioning statistic and test . . . . .	12
1.3.3	The unified conditional statistic and test recommended by Berger . . . . .	15
1.3.4	The properties of conditional frequentist test and potential agreements . . . . .	18
1.3.5	Calibration of $P$ -values for testing simple hypothesis . . . . .	22
<b>2</b>	<b>DETERMINING POSTERIOR <math>P</math>-VALUES FROM EMPIRICAL DISTRIBUTIONS</b>	<b>29</b>
2.1	Multiplicity of hypothesis testing and the false discovery rate . . . . .	29
2.2	Exploiting the empirical distribution of $P$ -values . . . . .	31
2.3	Modeling $P$ -values as a sample from a mixed population . . . . .	34
2.4	Density estimation based on kernel methods . . . . .	36
2.5	Posterior probabilities of the null hypothesis given the $P$ -value and implementation . . . . .	39
<b>3</b>	<b>THE EMPIRICAL <math>P</math>-VALUE CALIBRATION</b>	<b>42</b>
3.1	Smoothing and density estimation . . . . .	42
3.2	B-splines as smoothing and estimating functions . . . . .	44
3.2.1	B-splines . . . . .	45
3.2.2	Smoothing function and B-spline resulting from the $k^{th}$ order centred difference of a truncated power function . . . . .	46
3.2.3	Computing and evaluating B-splines . . . . .	49
3.3	The estimates of $S'(q)$ based on the smoothing B-splines . . . . .	53

# List of Figures

1.1	Posterior density distribution(used with permission of Dr. Mikelis G. Bickis) . . . .	5
1.2	Probability density function of $P$ -values under $H_0$ and $H_1$ . . . . .	17
1.3	Proportion of tests having true nulls given 2 sample standard deviations from the null mean . . . . .	27
1.4	Proportion of tests having true nulls given half of the sample standard deviation from the null mean . . . . .	28
2.1	CDF and ECDF of 100% $N(0, 1)$ null cases . . . . .	32
2.2	CDF of $Q$ -values from four normal mixed populations . . . . .	33
2.3	ECDF of $Q$ -values from four normal mixed populations . . . . .	33
2.4	Cutoff set up by Bonferroni method and FDR . . . . .	34
2.5	Kernel density estimation . . . . .	38
2.6	Posterior probability of the null hypothesis given $Q$ -value equals odds ratio of the slopes. . . . .	40
2.7	Posterior probability of the null hypothesis given $Q$ -value equals odds ratio of the slopes . . . . .	40
3.1	Fidelity to the data and smoothness compared between different orders of B-splines) 43	
3.2	Base of different orders of B-splines . . . . .	50
3.3	Histogram and smoothing B-splines of Order $0^{th}$ and $1^{st}$ . . . . .	55
3.4	Noninterpolatory, shape-preserving, and slope of Order 2 B-spline estimation . . . .	57
3.5	Estimated cumulative distribution using Order 2 and Order 3 Bsplines . . . . .	62
3.6	Order 2 Bspline estimators for the posterior probability of null hypothesis . . . . .	63
3.7	Histosplines . . . . .	64
3.8	Estimated PDF of $Q$ -value, $S'(q)$ = derivative of estimated CDF Order 3, is the $2^{nd}$ -order B-splines as shown in Lemma (3.6). . . . .	64
3.9	Estimated PDF and the differences at the right end of the distribution before and after the transformation . . . . .	65
3.10	Improving the precision of estimators by increasing the number of knots . . . . .	66

# List of Tables

1.1	Calibrating $P$ -values . . . . .	23
2.1	Outcomes of $n$ hypotheses . . . . .	30

# INTRODUCTION

In this thesis, we will be examining and implementing hypothesis testing approaches. By modeling the set of  $P$ -values as a sample from a mixed population, we will calibrate posterior probabilities of the null hypothesis, and hence will consider how posterior probabilities of the null hypothesis and  $P$ -values from the significance tests are related.

As an overview, the thesis comprises three chapters. In Chapter 1, we introduce the Bayesian and frequentist procedures, where the procedures reviewed can be found in [30] and [25]; then move on to describe a unified conditional frequentist testing methodology proposed by Berger [8].

In Chapter 2, various ideas for handling the multiplicity of tests such as false discovery rate, developed by Benjamini and Hochberg, and Bonferroni method are discussed. Then we consider calibrating the null and alternative components of the distribution by modeling the set of  $P$ -values as a sample from a mixed population proposed by Bickis [3]. The posterior probabilities of the null hypothesis given  $P$ -values from the empirical  $P$ -value distribution are computed based on kernel probability density estimation. This methodology is aimed to distinguish the sub-population of nulls, to calibrate  $P$ -value by computing the posterior probability of the null hypothesis in the light of deviations from uniformity of the empirical distribution of  $P$ -values, and to make inferences about estimates of the probability of the null hypothesis being true.

In the final chapter, we are concerned with the probability density estimation for  $P$ -values resulting from the significance tests. After defining B-splines based on centred differences as smoothing functions, we develop a noninterpolatory and shape-preserving density estimator. The results related to the properties of our noninterpolatory and shape-preserving density estimator are proved. In general, B-spline is defined with the aid of divided differences or recurrence relation (cf. DeBoor [15] Page.131). However, our reformulation of B-splines based on centred differences results in more accurate and stable density estimates compared with the kernel probability density estimators and other interpolatory spline density estimators. Although spline smoothing approach to non-parametric regression curve fitting is widely applicable (cf. Silverman [42]), our noninterpolatory and shape-preserving density estimator resulting from B-splines as smoothing functions based on centred differences has not yet been found in the literature in the desired form. Finally, these techniques are illustrated using simulated data. MATLAB coding with some further details

and comments for carrying out these simulations is presented in the Appendix.

Chapter 1 is basically a literature review of hypothesis testing procedures. The contents in Chapter 2 are original work of Dr. Mikelis G. Bickis, my supervisor. In Chapter 3, we establish the relation between B-splines smoothing functions and centred differences, and explicitly express and convert the B-splines smoothing function from the centred difference to truncated polynomials instead of the recurrence relation. Most of the results and proofs, the definition of smoothing B-splines resulting from centred differences, and the practical description and implementation of the probability density estimation based on the smoothing B-splines are our own.

# Chapter 1

## THE TESTING PROCEDURES FROM THREE INFERENCE SCHOOLS

### 1.1 Hypothesis testing

Generally hypothesis testing is a decision making problem with a number of possible outcomes. In particular, hypothesis testing provides an objective framework for making decisions using probabilistic methods, rather than relying on subjective impressions whose conformity with the data are needed to be tested. As well, hypothesis testing provides a decision making criterion that is consistent for all people even though people can form different opinions by looking at data. The hypothesis testing considered can be formulated in terms of the null hypotheses, denoted by  $H_0$  and the alternative hypotheses, denoted by  $H_1$ . Quite often the inferential process can be summarized in the verification of some statements about an unknown quantity  $\theta$ , belonging to a parameter space  $\Theta$ .

Consider the two disjoint subsets  $\Theta_0$  and  $\Theta_1$  of  $\Theta$ . The hypotheses constituted can be parameterized as follows:

Under  $H_0 : \theta \in \Theta_0$

Under  $H_1 : \theta \in \Theta_1$

If the subset of the parameter space defining a hypothesis contains a single element, the hypothesis is said to be simple. In another words, let  $X$  be a random variable with a probability density function  $f(x|\theta)$ , that is,  $X \sim f(x, \theta)$ . If a statistical hypothesis is a statement about the distribu-

tion of  $X$ , the hypothesis completely specifies  $f(x, \theta)$ . Otherwise, it is said to be composite, and under a composite hypothesis it is only specified that the observational distribution belongs to a family. When a hypothesis is simple,  $\Theta_0 = \{\theta_0\}$  and/or  $\Theta_1 = \{\theta_1\}$ .

In general, any decision about the truth or falsity of the hypothesis based on experimental evidence is subject to error, which is not only random error that results from experimental measurements but also occasional decision errors such as Type I and Type II errors. We will discuss Type I and Type II errors later. Suppose that the only possible decisions are whether  $H_0$  is true or  $H_1$  is true. All outcomes in hypothesis testing typically refer to the null hypothesis. Hence if one decides  $H_0$  is true, then  $H_0$  is accepted; if one decides  $H_1$  is true, then  $H_0$  is rejected.

There are three different approaches for testing, advocated by Jeffreys [26], Fisher [18], and Neyman [30] by using the posterior probabilities,  $P$ -value, and Type I error and Type II error probabilities respectively, but, as quoted from Berger [8], “Each was quite critical of the other approaches.” If one makes the wrong decision, one suffers a loss. From a Bayesian perspective, one would try to minimize the expected loss. Also one may have many alternative hypotheses  $H_1, \dots, H_k$  that can be compared through  $P(H_i|x), i = 1, \dots, k$ . Under the frequentist perspective, however, it is important to have the only two hypotheses  $H_0$  and  $H_1$ .

### 1.1.1 Bayesian methodology for hypothesis testing

A Bayesian’s approach for testing is to find an optimal procedure that minimizes some risk function, which is especially useful for such decision making. Bayesians view probability as degree of belief about unknown parameters and combine the prior belief with the information provided by the data in a study to produce a posterior distribution. Similar to a classical sampling distribution that is centred around a parameter estimate and used to calculate confidence intervals from the frequentist perspective, the posterior distribution can be employed to construct a credibility region for the unknown parameters from the Bayesian perspective.

For example, consider a random variable  $X$ , the number of success, having a binomial distribution,

$$X \sim \text{Bin}(n, \theta)$$

where  $n$  is the number of trials and  $\theta$  is the success probability.

Consider a situation where the prior belief is Beta( $\alpha, \beta$ ) distributed:

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \cdot \theta^{(\alpha-1)} \cdot (1 - \theta)^{(\beta-1)},$$

where  $\alpha$  and  $\beta$  are called hyperparameters.

Using Bayes' Theorem, the posterior density is:

$$p(\theta|x) \propto p(x|\theta) \cdot p(\theta) = \frac{1}{B(\alpha + x, \beta + n - x)} \cdot \theta^{(\alpha+x-1)} \cdot (1 - \theta)^{(\beta+n-x-1)},$$

$$\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

The posterior is  $\text{Beta}(\alpha + x, \beta + n - x)$  distributed with hyperparameters  $(\alpha + x, \beta + n - x)$  updated given that the prior belief is  $\text{Beta}(\alpha, \beta)$  distributed with hyperparameters  $(\alpha, \beta)$ . The posterior

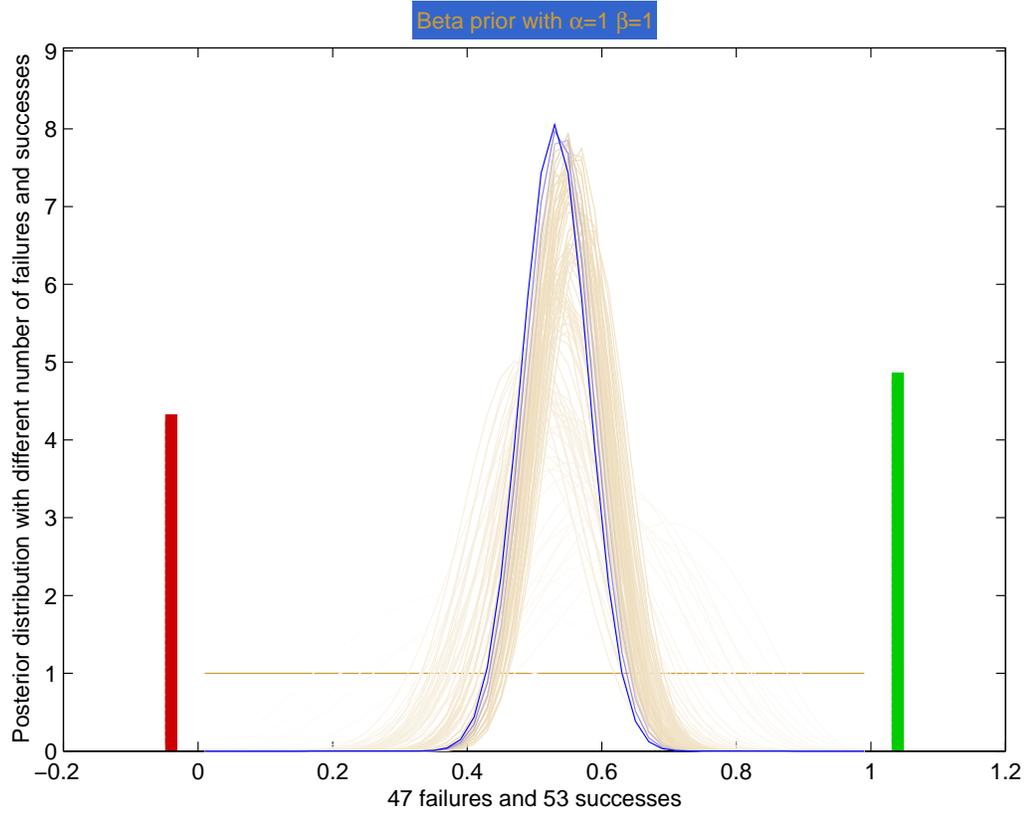


Figure 1.1: Posterior density distribution(used with permission of Dr. Mikelis G. Bickis)

probability density function with the data  $X \sim \text{Bin}(n, 0.5)$  is shown as Figure (1.1), where the prior hyperparameters  $\alpha = 1$  and  $\beta = 1$ , the number of trials  $n$  is increased from 1 to 100, and the number of success  $X = 53$  when  $n = 100$ . After 100 trials, with 53 successes (i.e., correct decisions) and 47 failures, we wish to choose between two hypotheses:  $H_0$  (i.e.,  $\theta = \theta_0$ ) and  $H_1$  (i.e.,  $\theta \neq \theta_0$ ), where  $\theta$  and  $\theta_0$  are two-dimensional and the two components of  $\theta$  are  $\alpha$  and  $\beta$ .

A Bayesian hypothesis test (Jeffreys [26]) proceeds by contrasting two quantities: the probability of the observed data  $x$  given  $H_0$  (i.e.,  $\theta = \theta_0$ ) and the probability of the observed data  $x$  given  $H_1$  (i.e.,  $\theta \neq \theta_0$ ). Also as quoted from Migon [31], “it suffices to examine the posterior prob-

abilities  $p(H_0|x)$  and  $p(H_1|x)$ . If the posterior probability  $p(H_0|x) > p(H_1|x)$ , then  $H_0$  should be accepted as the most plausible hypothesis for  $\theta$ . In this case, it can be said that  $H_0$  is preferable to  $H_1$ . Otherwise,  $H_1$  is preferable to  $H_0$ .”

Once again using Bayes’ Theorem:

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)} \propto p(x|\theta) \cdot p(\theta)$$

$$p(H_0|x) \propto p(x|H_0) \cdot p(H_0)$$

$$p(H_1|x) \propto p(x|H_1) \cdot p(H_1)$$

$$\frac{p(H_0|x)}{p(H_1|x)} = \frac{p(H_0)}{p(H_1)} \cdot \frac{p(x|H_0)}{p(x|H_1)}$$

The ratio  $\frac{p(x|H_0)}{p(x|H_1)}$  is the Bayes factor, denoted by  $B(x)$ , and it quantifies the evidence that the data provide for  $H_0$  against  $H_1$ . In accordance with Berger [8], assuming equal prior plausibility for the testing (prior indifference of  $\theta$ ), the posterior probability for  $H_0$  is given as follows.

$$p(H_0|x) = \frac{B(x)}{1 + B(x)}. \quad (1.1)$$

Under unequal prior plausibility for the testing, let us assign a lump prior probability  $\pi_0$  to a simple hypothesis  $H_0$ , that is,  $p(H_0) = \pi_0$  even though one will have that  $p(H_j) = p(H_j|x) = 0$  when  $H_j$  is a simple hypothesis,  $j = 0, 1$ , and the prior distribution of  $\theta$  is continuous. So, if  $H_1$  is the complement of a simple hypothesis, then  $p(H_1) = \pi_1 = 1 - \pi_0$  and this probability is distributed over the different values of  $\theta$  under  $H_1$ .

Let the prior density of  $\theta$  under  $H_1$  be  $\theta|H_1 \sim f(\theta)$ . As  $H_0$  is a simple hypothesis, it follows that  $p(x|H_0) = p(x|\theta_0)$ .

The marginal likelihood of  $H_1$  based on  $X$  is found by integrating over all possible values of  $\theta$ :

$$p(x|H_1) = \int_{\theta - \{\theta_0\}} p(x|\theta, H_1) \cdot p(\theta|H_1) d\theta = \int_{\theta} p(x|\theta) \cdot f(\theta) d\theta.$$

Also, the marginal prior for  $\theta$  has continuous and discrete parts of which the cumulative distribution function of  $\theta$  under  $H_1$  is  $F(\theta)$  and therefore the marginal density of  $X$  is:

$$p(x) = \int p(x|\theta) dF(\theta) = \pi_0 \cdot p(x|\theta_0) + \pi_1 \cdot \int p(x|\theta) \cdot f(\theta) d\theta = \pi_0 \cdot p(x|\theta_0) + \pi_1 \cdot p(x|H_1).$$

The Bayes factor corresponding to  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$  is:

$$B(x) = \frac{p(x|\theta_0)}{\int_{\theta} p(x|\theta) \cdot f(\theta) d\theta}.$$

As noted above, the relative odds between  $H_0$  and  $H_1$ ,  $B(x)$ , does not take into account the prior odds  $\pi_0/(1 - \pi_0)$ , which is a Bayesian measure of the goodness of fit of a given model to the data set. Thus  $B(x) > 1$  indicates that  $H_0$  fits the data better than  $H_1$ .

Based on assigning equal prior probabilities of  $1/2$  to the two hypotheses and applying the Bayes theorem, Jeffreys [26] approach for testing proceeded by:

- Compute the Bayes factor  $B(x) = \frac{p(x|H_0)}{p(x|H_1)}$ .
- Reject  $H_0$  as  $B(x) \leq 1$ ; otherwise, accept  $H_0$ .
- Report the posterior probability of the hypothesis as in equation (1.1) or equation (1.2).

$$p(H_1|x) = \frac{1}{1 + B(x)} \quad (1.2)$$

Note that Bayesian hypothesis testing depends on prior distributions.

### 1.1.2 Fisher's $P$ -value for hypothesis testing

Fisher [18] believed that there must exist a logic of inductive inference that would yield a correct answer to any statistical problem. By using such an inductive logic, the statistician would be freed from prior assumptions of the Bayesian school. Fisher's significance testing is based on the  $P$ -value. The concept of  $P$ -value is defined as follows.

**Definition 1.1 (Definition of  $P$ -value)** *The  $P$ -value is the probability of getting something at least extreme as the observed result assuming the null hypothesis is true, that is,*

$$P\text{-value} = \Pr(t(X) \geq t(x)|H_0).$$

So the  $P$ -value is referred to as the maximum probability of the most extreme event that actually happened. More generally, consider a family of extreme events and let  $\{A_t : t \in T\}$  be a nested collection of extreme events and  $T$  is totally ordered set. We need that  $A_t \subset A_s$  if  $t > s$ , and have a null hypothesis which is a statistic assigning a family of probabilities  $P_\theta(A_t)$  to the extreme events. Let  $t^* = \sup\{t : X \in A_t\}$ . Henceforth, the  $P$ -value is formally defined as:

$$P\text{-value} = \sup_{\theta \in \Theta_0} P_\theta(A_{t^*}).$$

As defined above, the  $P$ -value is computed assuming the null hypothesis is true. The  $P$ -value is, however, not the probability of the null hypothesis  $H_0$ . A large  $P$ -value (close to 1) is not an evidence in favor of  $H_0$ .

Let us proceed with a simple hypothesis  $H_0 : \theta = \theta_0$  by Fisher's significance testing. Suppose one observes data  $X = x$ , where  $X \sim f(x|\theta)$ , and the test is as follows.

- Choose a test statistic  $T = t(X)$ , where large values of  $T$  reflect evidence against  $H_0$ .
- Compute the probability  $p = \Pr(t(X) \geq t(x)|H_0)$ , the  $P$ -value, where  $x$  is the specific observed value and  $X$  is the random variable.
- Reject  $H_0$  if  $p$  is small enough since small  $p$  indicates an unlikely event and, hence, an unlikely hypothesis  $H_0$ .

### 1.1.3 Neyman-Pearson frequentist methodology for hypothesis testing

In a frequentist hypothesis testing procedure, first one needs to specify a null hypothesis, say  $H_0 : \theta = \theta_0$  and a alternative hypothesis, say  $H_1 : \theta = \theta_1$ . The testing then can be proceeded by:

- Construct a test statistic  $T = t(X)$ , where large values of  $T$  reflect evidence against  $H_0$ .
- Reject  $H_0$  if  $T \geq c$ , where  $c$  is a critical value resulting from the pre-chosen significance level  $\alpha$ ; or specify a rejection region  $\Gamma_\alpha$  (critical region), and then reject or accept the null hypothesis  $H_0$  depending on whether or not the observed value of the test statistic is within the critical region.
- Compute Type I and Type II error probabilities,  $\alpha = \Pr(\text{Reject } H_0|H_0 \text{ true})$  and  $\beta = \Pr(\text{Accept } H_0|H_1 \text{ true})$

**Definition 1.2 (Definition of the power function of a test)** *The power function,  $\pi(\theta)$ , of a test of  $H_0$  is the probability of rejecting  $H_0$  when the true value of the parameter is  $\theta$*

For simple hypotheses  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta = \theta_1$ , we have  $\pi(\theta_0) = \alpha$ , and  $\pi(\theta_1) = 1 - \beta$ , where  $\alpha$  and  $\beta$  are Type I and Type II error probabilities.

The distribution of  $T$  under the null hypothesis is known, but it is not necessarily known under the alternative hypothesis. The distribution of  $T$  under the alternative is, however, needed to compute the power as defined in Definition (1.2). As noted above when proceeding by using Neyman-Pearson hypothesis testing, one might reject  $H_0$  when  $H_0$  is true; or might fail to reject  $H_0$  when  $H_0$  is false. Neyman-Pearson frequentist approach for testing just report unconditional Type I and Type II error probabilities, based on the predetermined significance level regardless of the actual scales of evidence in the data.

As discussed previously, Neyman-Pearson hypothesis testing is to reject  $H_0$  if  $T \in \Gamma_\alpha$  with taking risk to Type I error, and fail to reject otherwise with commitment to Type II error. It is desirable to minimize  $\alpha$  and  $\beta$  simultaneously, but the two probabilities can not be controlled at the same time. A traditional way for a simple null hypothesis  $H_0$  for a given  $T = t(X)$  is to assign an upper bound, which is referred to as the significance level or the size of test for a composite hypothesis  $H_0$ , to the Type I error rate  $\Pr(T \in \Gamma|H_0)$ , and to attempt to minimize Type II error

rate  $\Pr(T \notin \Gamma|H_1)$ , or equivalently to maximize power  $\Pr(T \in \Gamma|H_1)$ . This is so-called controlling the Type I error rate.

### 1.1.4 Connecting $P$ -value and Neyman-Pearson testing and randomizing a test

More generally, suppose we have a test where  $H_0$  is rejected when test statistic  $T \in \Gamma_\alpha$ , the rejection region. The rejection region corresponding to the level  $\alpha$  is denoted by  $\Gamma_\alpha$ , satisfying  $\Pr(T \in \Gamma_\alpha|H_0) \leq \alpha$ .

Let  $t$  be the observed value of  $T$ . Then evaluation of  $\Pr(T > t|H_0)$  gives an idea of how extreme the observed value is under  $H_0$ . The notion of the  $P$ -value is useful for us to determine the size  $\alpha$  at which we would reject  $H_0$  based on the information actually obtained.

$$H_0 \text{ is rejected} \iff P\text{-value} < \alpha$$

On the other hand, for a composite null hypothesis  $H_0$ , the size  $\eta$  of a test (or size of critical region) is the maximum probability of rejecting  $H_0$  when  $H_0$  is true (maximized over the values of the parameter under  $H_0$ ), that is,

$$\eta = \sup_{\theta \in \Theta_0} \Pr(\text{Rejection of } H_0 | \theta \in \Theta_0) \quad (1.3)$$

However, it is not always possible to obtain tests of any pre-specified level exactly, that is,  $\eta \leq \alpha$  because of the discreteness of the random variable.

**Definition 1.3 (Definition of a conservative test)** *A test is said to be conservative if  $\eta < \alpha$ .*

The conservative test may result in a loss of the power as defined in Definition (1.2). In order to get the most powerful test, we should increase the rejection region as large as possible to increase the power as discussed in Section (1.1.3).

There is an alternative approach that allows us to obtain tests of an exact level even for discrete distribution. The alternative is known as randomized tests where any pre-determined level is obtained after realization of an additional independent Bernoulli experiment with success probability conveniently chosen to complete the difference between  $\alpha$  and  $\eta$  in equation (1.3).

**Definition 1.4 (Definition of a randomized test)**

A randomized test says that if

$$\frac{P_{\theta_1}}{P_{\theta_0}} = \frac{p(x|\theta_1)}{p(x|\theta_0)} > k$$

then reject  $H_0$

$$\text{if } \frac{P_{\theta_1}}{P_{\theta_0}} = \frac{p(x|\theta_1)}{p(x|\theta_0)} < k$$

then accept  $H_0$

If  $\frac{P_{\theta_1}}{P_{\theta_0}} = k$ , then observe a binary random variable  $U$ , which is independent of the data, and reject if  $u = 1$ , where  $P(U = 1)$  is defined by

$$P_{\theta_0} \left( \frac{P_{\theta_1}}{P_{\theta_0}} > k \right) + P(U = 1) \cdot P_{\theta_0} \left( \frac{P_{\theta_1}}{P_{\theta_0}} = k \right) = \alpha.$$

## 1.2 Example

Berger [8] puts it in this way: “Jeffreys, Fisher, and Neyman not only disagreed as to statistical foundations, but also reported considerably different practical conclusions.”

Let’s consider Example (1.1) to see how they reported the results.

### Example 1.1 (Taken from Berger [8])

Suppose that the data,  $X_1, \dots, X_n$ , are i.i.d. from the normal distribution with the unknown mean  $\theta$  and  $\sigma^2$  known, that is,

$$X_1, \dots, X_n \sim N(\theta, \sigma^2),$$

and  $n = 10$ .

Berger [8] considered two different possible observed data,  $z = \frac{\sqrt{n} \cdot \bar{x}}{\sigma} = 2.3$ , or  $z = 2.9$ , where  $\bar{x}$  is the sample mean.

### Jeffreys’ methodology

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

We have the equal prior probabilities, that is,  $p(H_0) = p(H_1) = 1/2$ , and the prior of  $\theta$  under  $H_1$  is Cauchy distributed with parameters equal to 0 and  $\sigma$ , denoted by  $\theta|H_1 \sim \text{Cauchy}(0, \sigma)$ . So, the prior density of  $\theta$  under  $H_1$  is

$$f(\theta|H_1) = \frac{\sigma}{\pi \cdot (\theta^2 + \sigma^2)}$$

where  $\sigma = 1$  in accordance with Berger [8].

Therefore, the marginal likelihood of  $H_1$  based on  $X_1, \dots, X_n$  or  $z = 2.3$  respectively, ( $z = 2.9$ ) is:

$$p(z|H_1) = \int_{\theta} p(z|\theta) \cdot f(\theta|H_1) d\theta = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot (z-\theta)^2} \cdot \frac{\sigma}{\pi \cdot (\theta^2 + \sigma^2)} d\theta$$

As discussed in Section (1.1.1), the posterior probabilities of  $H_0$ , corresponding to  $z = 2.3$  (or  $z = 2.9$ ), are as follows.

$$\Pr(H_0|x_1, \dots, x_n) = \Pr(\theta = 0|z) = \frac{B(z)}{1 + B(z)}$$

where the Bayes factor is:

$$B(z) = \frac{p(z|H_0)}{p(z|H_1)} = \frac{\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot z^2}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot (z-\theta)^2} \cdot \frac{\sigma}{\pi \cdot (\theta^2 + \sigma^2)} d\theta},$$

so  $B(2.3) = 0.3891316812$  and  $B(2.9) = 0.1276643975$ . Henceforth,

$$\Pr(\theta = 0|z = 2.3) = 0.28,$$

$$\Pr(\theta = 0|z = 2.9) = 0.11.$$

Since  $B(z) < 1$ ,  $H_0$  is then rejected.

**Fisher's methodology:**

$$H_0 : \theta = 0$$

If  $z = 2.3$ ,

then  $P$ -value =  $\Pr(z \geq 2.3) + \Pr(z \leq -2.3) = 0.021$

If  $z = 2.9$ ,

then  $P$ -value =  $\Pr(z \geq 2.9) + \Pr(z \leq -2.9) = 0.0037$

Therefore, Fisher would report the  $P$ -values:  $p = 0.021$  or  $p = 0.0037$ . Since  $P$ -value  $< 0.05$ ,  $H_0$  is then rejected.

**Neyman-Pearson frequentist methodology:**

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

The test statistic is  $T(X) = \frac{\sqrt{n} \cdot \bar{X}}{\sigma}$ . Hence,  $T(x) = 2.3$  or  $2.9$ .

Given the pre-chosen significance level  $\alpha = 0.05$ , the rejection region would be:

$$\Gamma_\alpha = \{T : T > 1.96\} \cup \{T : T < -1.96\}$$

Neyman-Pearson frequentist approach for testing would also reject  $H_0$  since  $T \in \Gamma_\alpha$ , and one would then just report  $\alpha = 0.05$  in either case, unconditional Type I error probability based on the predetermined significance level.

From this example, one can see they reported differently. The three approaches to testing can lead to quite different practical conclusions both in statistics and science. In particular, as more complex statistical analysis is increasingly used in science areas, the perceived disagreement between Bayesian and frequentist approaches seems to loom larger than it does in the statistical

community. Berger [8] proposed a potential unified conditional frequentist approach to testing.

### 1.3 The unified conditional frequentist testing

Methods to reconcile the different approaches to testing that are proposed by Fisher, Jeffreys and Neyman using  $P$ -values, posterior probability, and Type I and Type II error probabilities, respectively, are needed. From both frequentist and Bayesian perspectives as discussed previously, the most promising route to a compromise is to derive Bayesian inference procedures that can also be justified by their behavior in repeated sampling from the model. Sellke et al. [40] outline such a route in the context of testing a null hypothesis. Berger [8] proposed a potential methodological unified conditional frequentist approach to testing. The idea is to follow Fisher in using the  $P$ -value to define the strength of evidence in data and to follow Fisher's method of conditioning on strength of evidence; then follow Neyman by computing Type I and Type II error probabilities conditioning on strength of evidence in the data, which equal the objective posterior probabilities of the hypothesis advocated by Jeffreys, assuming the hypotheses have equal prior probabilities of 0.5.

#### 1.3.1 Statistic and an unknown parameter

**Definition 1.5 (Definition of sufficient statistic)** *Let  $X$  be a random variable with a probability density function  $p(x|\theta)$ . Then the statistic  $T = T(x)$  is sufficient for the unknown parameter  $\theta$  if the conditional probability density function of  $X$  given  $T = T(x)$  does not depend on  $\theta$ , that is,*

$$p(x|t, \theta) = p(x|t).$$

**Definition 1.6 (Definition of ancillary statistic)** *Let  $X$  be a random variable with a probability density function  $p(x|\theta)$ . Then the  $S = S(x)$  is an ancillary statistic for the unknown parameter  $\theta$  if*

$$p(s|\theta) = p(s),$$

*that is, a statistic that has a distribution that does not depend on  $\theta$ .*

#### 1.3.2 Introduction to conditioning statistic and test

Conditional inference is one of the most important concepts in both statistical theory and statistical methodology. In the Bayesian paradigm, conditioning is automatic, for instance,

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

while in the frequentist paradigm, there is no general theory as to how to condition even though frequentists do condition in various circumstances. However, the use of conditioning in the pure frequentist school was comparatively sporadic since according to Berger [8] Neyman rarely addressed the issue. Fisher would use conditional variables to eliminate nuisance parameters, as in the Fisher exact test where he chose  $S$  to be the marginal totals in a contingency table and then computed  $p$ -values conditioning on these marginal totals ([17] and [18]), which is also described in Example (1.2). In addition, Fisher recommended that statisticians routinely condition on an ancillary statistic. It is actually possible to eliminate unknown nuisance parameters and obtain exact size  $\alpha$  tests (the maximum probability of rejecting null hypothesis  $H_0$  over the values of the parameters under  $H_0$ ) by conditional tests based on conditional arguments. For instance, if a sufficient statistic  $S$  exists for an unknown nuisance parameter  $\theta$ , then the distribution of  $X|S$  will not depend on  $\theta$ . This technique will be illustrated for the two-sample binomial test as follows.

**Example 1.2 (Conditional test)**

Let  $X$  and  $Y$  be independent distributed as  $X \sim \text{Bin}(n_1, p_1)$  and  $Y \sim \text{Bin}(n_2, p_2)$ . We wish a size  $\alpha$  test as follows.

$$H_0 : p_1 = p_2 \equiv p \quad \text{vs.} \quad H_1 : p_1 < p_2$$

where  $p$  is unknown. Let  $q = 1 - p$ .

Since  $X$  and  $Y$  are independent, under  $H_0$  the joint density of  $X$  and  $Y$  is:

$$f(x, y) = \binom{n_1}{x} \binom{n_2}{y} p^{x+y} \cdot q^{n_1+n_2-(x+y)}. \tag{1.4}$$

Let the conditioning statistic be  $S = X + Y$ . We now prove  $S$  is a sufficient statistic for the common unknown  $p$ , assuming  $H_0$  is true.

Since the moment generating functions of  $X$  and  $Y$  are

$$M_x(t) = (pe^t + q)^{n_1} \text{ and } M_y(t) = (pe^t + q)^{n_2},$$

and  $X$  and  $Y$  are independent, the moment generating function of  $X + Y$  is

$$M_{x+y}(t) = M_x(t) \cdot M_y(t) = (pe^t + q)^{n_1+n_2} \tag{1.5}$$

Therefore,  $S = X + Y \sim \text{Bin}(n_1 + n_2, p)$ , and

$$f_S(s|p) = \binom{n_1 + n_2}{s} p^s \cdot q^{n_1+n_2-s} \tag{1.6}$$

Consider the statistic  $T(s) = S = X + Y$ . Equation (1.6) can be reformulated as follows.

$$\begin{aligned} f_S(s|p) &= \binom{n_1 + n_2}{s} p^s \cdot q^{n_1+n_2-s} \\ &= g(s) \cdot f(t, p) \end{aligned}$$

where  $f(t, p) = p^t \cdot q^{n_1+n_2-t}$  and  $g(s) = \binom{n_1 + n_2}{s}$ , and both are non-negative functions. Based on Neyman's factorization criterion, that is, the statistic  $T$  is sufficient for the unknown parameter  $\theta$  if and only if

$$p(x|\theta) = f(t, \theta) \cdot g(x)$$

where  $f$  and  $g$  are non-negative functions, hence  $S = X + Y$  is sufficient for the unknown  $p$  if  $H_0$  is true.

This suggests considering a test based on the conditional distribution of  $(X, Y)$  given  $S = s$ . Because  $Y = S - X$ , it suffices to base the test on the conditional distribution of  $Y$  given  $S = s$ . Under  $H_0$ , from equation (1.6)  $S \sim \text{Bin}(n_1 + n_2, p)$  and thus

$$\begin{aligned} f_{Y|s}(y) &= \frac{f_{S,Y}(s, y)}{f_S(s)} \\ &= \frac{f_{X,Y}(s-y, y)}{f_S(s)} \\ &= \frac{\binom{n_2}{y} \binom{n_1}{s-y} p^s \cdot q^{n_1+n_2-s}}{\binom{n_1+n_2}{s} p^s \cdot q^{n_1+n_2-s}} \\ &= \frac{\binom{n_2}{y} \binom{n_1}{s-y}}{\binom{n_1+n_2}{s}} \\ & \quad y = 0, \dots, s; \quad s = 0, \dots, n_1 + n_2. \end{aligned}$$

So  $Y|s \sim \text{Hypergeometric}(s, n_2, n_1 + n_2)$ , namely  $Y|s$  has a hypergeometric distribution. This distribution does not involve  $p$ , and an exact size  $\alpha$  critical region (or by randomizing the test for the discrete distribution as discussed in Section (1.1.4) can be determined under  $H_0$  for any given observed value of  $s$ . For  $H_1 : p_1 < p_2$ , the best critical region would be for large  $y$ . Thus, reject  $H_0$  for a size  $\alpha$  test if

$$\sum_{i=y}^s \frac{\binom{n_2}{i} \binom{n_1}{s-i}}{\binom{n_1+n_2}{s}} \leq \alpha, \quad (1.7)$$

or equivalently reject  $H_0$  if  $y \geq k(s)$  where  $k(s)$  is the critical value and depends on the observed value of  $s$ , i.e.  $k(s)$  is the smallest integer such that equation (1.7) holds. Tests for other

alternatives can be obtained in a similar manner.

**Lemma 1.1 (Lemma on the size of conditional test)** *A conditional size  $\alpha$  test also is a size  $\alpha$  test unconditionally.*

**Proof:** Let  $T$  be test statistic,  $S$  be conditional statistic, and  $f(s)$  be the probability density function of  $S$ .

For a conditional size  $\alpha$  test,  $\Pr(T \geq k(s)|s) = \alpha$ , then we have

$$\begin{aligned} \Pr(T \geq k(S)) &= \int_s \Pr(T \geq k(s)|s) f(s) ds \\ &= E_S\{\Pr(T \geq k(S)|S)\} \\ &= E_S(\alpha) \\ &= \alpha, \end{aligned}$$

where  $\Pr(T \geq k(S)|S)$  is the conditional probability of the event  $T \geq k(S)$ , given the event  $S = s$ , if  $\Pr(S = s) \neq 0$ .

### 1.3.3 The unified conditional statistic and test recommended by Berger

To be precise as to the type of conditioning statistic and conditional test recommended by Berger, first we discuss the definitions of conditional frequentist error probabilities, as quoted from Berger [8], and then Berger's unified conditional frequentist test. Further we consider an example of conditional frequentist testing to illustrate how to find a conditioning statistic that measures the amount of evidence in the data for or against the null hypothesis and then to report the frequentist error probabilities conditioning on this statistic.

In the case of testing simple hypothesis

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1,$$

one determines a conditioning statistic  $S(x)$ .

**Definition 1.7 (Definitions of conditional frequentist error probabilities (CEP))**

$$\begin{aligned} \alpha(s) &= \Pr(\text{reject } H_0 | S(x) = s) = \Pr(\text{Type I error} | S(x) = s) \\ \beta(s) &= \Pr(\text{accept } H_0 | S(x) = s) = \Pr(\text{Type II error} | S(x) = s) \end{aligned}$$

Consider simple hypothesis  $H_0$  and  $H_1$  with absolutely continuous densities. In Fisherian statistics, the most commonly used measure of evidence is the  $P$ -value, so it is natural to con-

sider choosing  $P$ -value as the conditioning statistic for the conditional testing. Based on Definitions (1.1) of  $P$ -value,  $P$ -values under  $H_0$  and  $H_1$  for a conditioning statistic are defined as follows.

**Definition 1.8 (Definitions of  $P$ -values for a conditioning statistic)** *For simple hypotheses  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta = \theta_1$ , let  $p_0$  be the  $P$ -value under  $H_0$ , and  $p_1$  be the  $P$ -value under  $H_1$ , that is,*

$$p_0 = \Pr(t(X) \geq t(x)|H_0);$$

$$p_1 = \Pr(t(X) \leq t(x)|H_1).$$

So a conditioning statistic based on the  $P$ -values above is defined as follows.

**Definition 1.9 (Definition of a conditioning statistic  $S$  based on the  $P$ -values)**

$$S = \max\{p_0, p_1\}$$

Conditioning on  $P$ -values from Definition (1.8), Berger's conditional frequentist test proceeds by rejecting  $H_0$  when  $p_0 \leq p_1$  and accepting otherwise, and then computing the Type I and Type II conditional error probabilities (CEP) as defined in Definition (1.7). The resulting test,  $T^C$ , is defined by:

$$T^C = \begin{cases} \text{if } p_0 \leq p_1 \\ \quad \text{reject } H_0 \text{ and report Type I CEP, } \alpha(x) = \frac{B(x)}{1+B(x)} \\ \text{if } p_0 > p_1 \\ \quad \text{accept } H_0 \text{ and report Type II CEP, } \beta(x) = \frac{1}{1+B(x)} \end{cases} \quad (1.8)$$

where  $B(x)$  is the Bayes factor.

A direct application of Bayes' Theorem as in Section (1.1.1) shows that  $\alpha(x)$  and  $\beta(x)$  are precisely the Bayesian posterior probabilities, as defined in equation (1.1) and equation (1.2), assuming the hypotheses have equal prior probabilities of 0.5. Berger [9] shows that this equivalence holds generally when testing simple hypotheses. Let us work out the following example to demonstrate how to set up the conditional frequentist testing.

**Example 1.3 (Testing of simple hypotheses taken from Sellke et al. [40])**

It is desired to test:

$$H_0 : X \sim \text{Uniform}(0, 1) \quad \text{vs.} \quad H_1 : X \sim \text{Beta}(1/2, 1).$$

The probability density functions under  $H_0$  and  $H_1$  are illustrated in Figure (1.2)

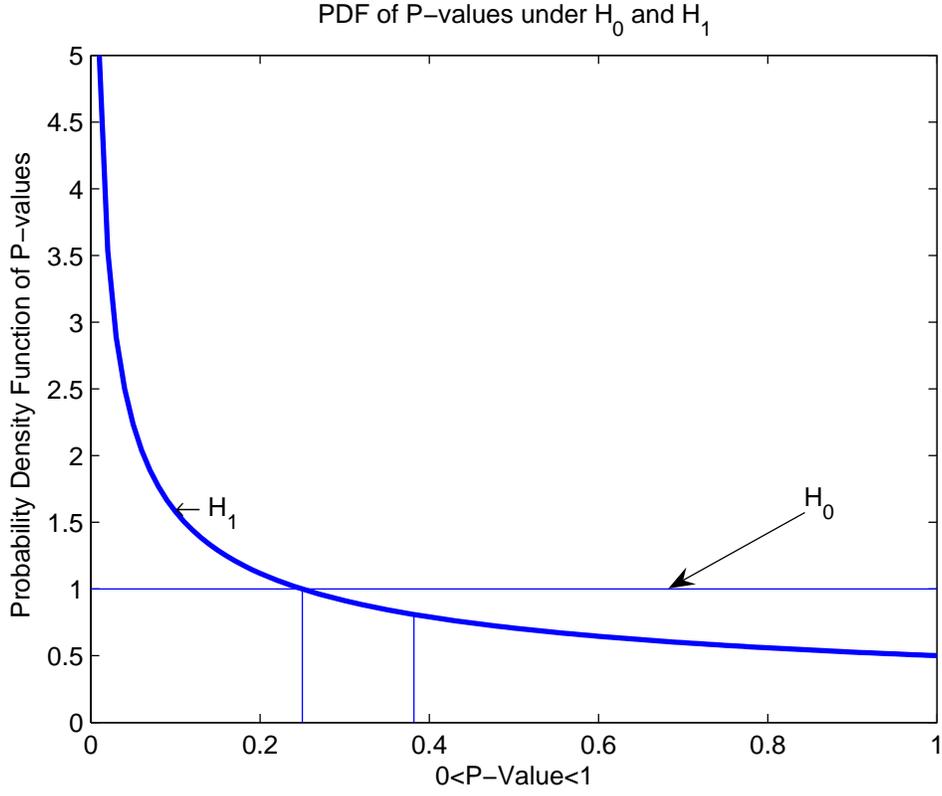


Figure 1.2: Probability density function of  $P$ -values under  $H_0$  and  $H_1$ .

$$f_0(x|H_0) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1.9)$$

$$f_1(x|H_1) = \begin{cases} (2\sqrt{x})^{-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1.10)$$

The Bayes factor is then

$$B(x) = \frac{f_0(x)}{f_1(x)} = 2\sqrt{x}$$

Now we compute the  $P$ -values  $p_0$  and  $p_1$  under  $H_0$  and  $H_1$ .

$$p_0 = \Pr(X \leq x|H_0) = x$$

$$p_1 = \Pr(X \geq x|H_1) = 1 - \sqrt{x}$$

Thus the conditioning statistic as defined in Definition (1.9) is

$$S = \max\{x, 1 - \sqrt{x}\}.$$

Therefore the resulting conditional frequentist test is

$$T^C = \begin{cases} \text{if } x \leq 0.382 \\ \text{reject } H_0 \text{ and report Type I CEP, } \alpha(x) = \frac{B(x)}{1+B(x)} = (1 + 1/2x^{-1/2})^{-1} \\ \text{if } x > 0.382 \\ \text{accept } H_0 \text{ and report Type II CEP, } \beta(x) = \frac{1}{1+B(x)} = (1 + 2\sqrt{x})^{-1} \end{cases} \quad (1.11)$$

where the Bayes factor is  $B(x) = 2\sqrt{x}$ . Note that Type I CEP and Type II CEP  $\alpha(x)$  and  $\beta(x)$  in equation (1.11) vary with the strength of evidence in the data and do not exhibit unnatural behavior for either small or large values of the observation  $X$ . However, there is a possible oddity for middle values of  $X$ . For example, when  $x = 0.36 < 0.382$ , then the conclusion of testing based on equation (1.11) is to reject  $H_0$  and report Type I CEP  $\alpha(0.36) = 0.55$ , which means one might make a decision with an error probability larger than 0.5. While  $H_0$  has formally been rejected, the fact that the reported conditional error probability is so high conveys the clear message that this is a very uncertain conclusion. For those uncomfortable with this mode of operation, we have this quote from Berger [8]: “note that it is possible to, instead, specify an ordinary rejection region (say, at the unconditional  $\alpha = 0.05$  level), find the ‘matching’ acceptance region (which would essentially be the 0.05 level rejection region if  $H_1$  were the null hypothesis), and name the region in the middle the no-decision region. The conditional test would be the same as before, except that one would now state ‘no decision’ when the data are in the middle region. The CEPs would not be affected by this change, so that it is primarily a matter of preferred style of presentation (whether to give a decision with a high CEP or simply state no decision in that case).” (See Berger [8] for more details on no decision region.)

### 1.3.4 The properties of conditional frequentist test and potential agreements

As noted above, the rejection region of the conditional frequentist test need not be specified in advance; it is determined as  $\{x : p_0(x) \leq p_1(x)\}$ . Classically, one is used to controlling Type I error probability through choice of the rejection region. For the conditional frequentist test, however, the unconditional Type I and Type II error probabilities  $\alpha$  and  $\beta$  are not used as the reported error probabilities. The conditional Type I and Type II error probabilities  $\alpha(x)$  and  $\beta(x)$  computed as in equation (1.8) are used instead, which more closely aligns  $P$ -values with posterior probabilities.

Berger’s conditional frequentist test should have been attractive to Neyman because it is fully compatible with Neyman-Pearson theory, which relies on comparing null and alternative densities. For example, the conditioning statistic as defined in Definition (1.9) is computed using null and

alternative densities. Moreover, the frequentist test results in error probabilities fully varying with the data that eliminates the major criticism of Neyman-Pearson frequentist approach. However, Neyman rarely addressed conditioning in spite of the criticisms from Fisher, so Berger [8] puts in this way: “it is difficult to speculate as to his reaction to the conditional frequentist test and to use of the conditioning statistic as defined in Definition (1.9). Another feature of  $T^C$  in equation (1.8) that Neyman might have taken issue with is that conditioning does affect optimality properties such as power if being used to alter the decision rule. As well, Neyman could well have been critical of the specification of rejection region of the conditional frequentist test as defined in equation (1.8).”

Several properties of the conditional frequentist test  $T^C$  in equation (1.8) would have certainly appealed to Fisher. First, the conditional frequentist test  $T^C$  is employing  $P$ -values to measure strength of evidence in data as Fisher recommended, and then conditioning upon strength of evidence is utilized. Second, the resulting test yields Type I and Type II error probabilities  $\alpha(x)$  and  $\beta(x)$ , computed as in equation (1.8), which fully vary with the strength of evidence in the data, an essential property that caused Fisher to be critical of Neyman-Pearson testing. Regarding the conditional statistic as noted above, however, Fisher would have questioned the use of  $S = \max\{p_0, p_1\}$  that is obviously neither an ancillary statistic nor a sufficient statistic as in Example (1.2) and also the use of the bigger one between  $p_0$  and  $p_1$  instead of the smaller one as a conditioning statistic. Another feature of  $T^C$  in equation (1.8) that Fisher might have been critical of is that an alternative hypothesis is necessarily needed to define the conditional frequentist test  $T^C$ .

As discussed previously, one can think of  $T^C$  as converting  $P$ -values into the conditional frequentist error probabilities while retaining the features of Type I and Type II error probabilities. Moreover, the conditional frequentist error probabilities, resulting from the conditional frequentist test defined in equation (1.8) and fully varying with the data, is precisely equal the objective posterior probability defined in equation (1.1) and equation (1.2). Therefore, the conditional frequentist and objective Bayesian end up reporting the same error probabilities.

Since an objective Bayesian would typically use, as the rejection region, the set of potential data for which  $\Pr(H_0|x) \leq 1/2$ , Jeffreys might have disagreed with the specified rejection region predetermined as in equation (1.8).

In the regard of the rejection region based on the conditioning statistic  $S = \max\{p_0, p_1\}$ , as noted above, Berger [8] recommends no decision region as an alternative rejection region if the reported conditional frequentist error probabilities is high when the null hypothesis is rejected. Let us reconsider Example (1.3). When  $x = 0.25$ , one rejects  $H_0$  and reports Type I conditional

frequentist error, as computed based on equation (1.8),

$$\alpha(0.25) = 0.5.$$

For these inconclusive data that provides no real evidence for or against the null hypothesis  $H_0$ , one can specify an ordinary rejection region by using the unconditional significance level, find the corresponding acceptance region, and refer the region in the middle as the no-decision region. As advocated by Berger [8], the conditional test would be the same as before and the conditional frequentist error probabilities, resulting from the conditional frequentist test defined in equation (1.8) would not be changed although one could simply state “no decision” in that case rather than giving a decision with a high conditional frequentist error probability.

In addition to the conditioning statistic  $S = \max\{p_0, p_1\}$ , Sellke et al. [40] consider other conditioning statistic such as an ancillary conditioning statistic, a conditioning statistic resulting from “intrinsic significance” based on a type of conditioning defined through likelihood concepts, “equal probability continuum” conditioning statistic, and the conditioning variable  $S = \min\{p_0, p_1\}$ , in the context of Example (1.3). The resulting conditional frequentist tests from the above conditioning statistics based on the same calculations as those in Example (1.3) are listed in the following using Example (1.3), which has the Bayes factor,  $B(x) = 2\sqrt{x}$ .

**Ancillary conditioning statistic:**

**Definition 1.10 (Definition of an ancillary conditioning statistic  $S$ )**

$$S = \max\{B(x), 2 - B(x)\}$$

A basic calculation shows that the statistic as defined in Definition (1.10) is an ancillary statistic, having the same distribution under  $H_0$  as under  $H_1$ . Computing the resulting Type I and Type II conditional error probabilities (CEP) as defined in Definition (1.7) yields the following test for Example (1.3).

$$T^A = \begin{cases} \text{if } x \leq 1/4, & \text{reject } H_0 \text{ and report Type I CEP } \alpha(x) = \sqrt{x}; \\ \text{if } x > 1/4, & \text{accept } H_0 \text{ and report Type II CEP } \beta(x) = 1/2. \end{cases}$$

The Type II conditional error probability in  $T^A$  is not satisfactory because  $\beta(x)$  remains constant although  $B(x) = 2\sqrt{x}$  varies as  $x$  varies from  $1/4$  to  $1$ . In particular, this constant is  $1/2$ , which suggests that one is doing no better than random choice of a hypothesis from the perspective of Type II error. This also violates the desire for error probabilities that vary with the strength of evidence in the data.

**Intrinsic significance:**

**Definition 1.11 (Definition of  $S$  resulting from intrinsic significance level)**

$$S = \max\{B(x), 1/B(x)\}$$

The conditional Type I error resulting from the test conditioning on  $S$  as defined in Definition (1.11) is referred to as the intrinsic significance level. Computing the Type I and Type II conditional error probabilities (CEP) as defined in Definition (1.7) yields the following test.

$$T^I = \begin{cases} \text{if } x \leq 1/4, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(x) = \begin{cases} 1 & \text{if } 0 < x < 1/16 \\ (1 + (16x^2)^{-1})^{-1} & \text{if } 1/16 < x < 1/4 \end{cases} \\ \text{if } x > 1/4, & \text{accept } H_0 \text{ and report Type II CEP} \\ & \beta(x) = (1 + 4x)^{-1}. \end{cases}$$

However, the Type I conditional error probability in  $T^I$ ,  $\alpha(x)$ , exhibits unnatural behavior. It is obviously unable to report  $\alpha(x) = 1$  when  $x < 1/16$ .

**Equal probability continuum conditioning statistic:**

**Definition 1.12 (Definition of equal probability continuum conditioning statistic  $S$ )**  $S(x)$  is chosen so that

$$\alpha(x) = \beta(x), \text{ for some } x.$$

The resulting test is:

$$T^E = \begin{cases} \text{if } x \leq 0.397, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(x) = (1 + (x^{-3/4} - 1)^{1/3})^{-1}; \\ \text{if } x > 0.397, & \text{accept } H_0 \text{ and report Type II CEP} \\ & \beta(x) = (1 + (x^{-3/4} - 1)^{-1/3})^{-1}. \end{cases}$$

Regarding the Type II conditional error probability in  $T^E$ , as noted above, the equal probability continuum conditioning statistic  $S$  results in the test that has the undesirable property that the Type II conditional error probability  $\beta(x) \rightarrow 0$  as  $x \rightarrow 1$ . This is unnatural because  $B(x) = 2$ , which hardly suggests that the decision to accept  $H_0$  would be “error-free.”

**Conditioning on a statistic against  $H_0$  and  $H_1$ :**

**Definition 1.13 (Definition of statistic based on the  $P$ -values against  $H_0$  and  $H_1$ )**

$$S = \min\{p_0, p_1\}$$

This conditioning statistic  $S = \min\{p_0, p_1\}$  yields quite different answers with those  $S = \max\{p_0, p_1\}$  does. Indeed, the resulting conditional error probabilities from the test conditioning on  $S$  as defined in Definition (1.13) are such that  $\alpha(x) \rightarrow 1/3$  as  $B(x) \rightarrow 0$ , while  $\beta(x) \rightarrow 0$  as  $B(x) \rightarrow 2$ , neither of which is sensible. Hence, this conditioning statistic should not be acceptable.

Regarding Type I or Type II conditional error probability of the above conditional frequentist tests, however, these conditioning arguments do not lead to fruitful conditional frequentist testing in general (See Sellke et al. [40] for more details on  $\alpha(x)$  or  $\beta(x)$  unnatural behavior with small or large values of the observation  $X$ ).

As noted above, an alternative hypothesis is necessary to define conditional frequentist testing  $T^C$  in equation (1.8). Berger [8] proposed a general method on how a conditional frequentist test can be done when there is no specified alternative by creating a generic nonparametric alternative. However, developing specific alternatives for important null hypotheses can be very difficult as shown in Berger [8], so he proposed calibrating  $P$ -values to test a null hypothesis when there is no alternative hypothesis.

### 1.3.5 Calibration of $P$ -values for testing simple hypothesis

It is well known that  $P$ -values under the null hypothesis are uniform distributed if the test statistic is continuous and is of exact size. This can be shown as follows.

**Lemma 1.2 (Lemma on  $P$ -values distribution under the null hypothesis)** *Suppose that  $n$  independent tests about the same hypothesis  $H_0 : \theta = \theta_0$  have been performed using different data sets and were based on independent statistics  $T_1, \dots, T_n$  with continuous distributions under  $H_0$ . Let  $P(T_1), \dots, P(T_n)$  be their respective  $P$ -values, then  $P(T_1), \dots, P(T_n)$  form a random sample from the uniform distribution on  $(0,1)$ .*

**Proof:** The  $P$ -value is the statistic  $P(T) = \Pr(T \geq t|H_0) = 1 - \Pr(T \leq t|H_0) = 1 - F_T(t)$ , where  $F_T(t)$  is the distribution function of the test statistic  $T$ .

By Probability Integral Transformation (cf. Bain [7], Page 201), it follows that  $F_T(T) \sim \text{Uniform}(0, 1)$ , and also obviously  $P(T) = 1 - F_T(T) \sim \text{Uniform}(0, 1)$ . Since  $P(T_i), i = 1 \dots n$ , are functions of independent statistics and all with the same distribution,  $P(T_i), i = 1 \dots n$ , constitute a random sample from the uniform distribution on  $(0,1)$ .

Therefore, one can reduce the original hypothesis to the generic null hypothesis as follows.

$$H_0 : P \sim \text{Uniform}(0, 1)$$

where  $P$  denotes the  $P$ -value.

The Bayes factor corresponding to  $H_0 : P \sim \text{Uniform}(0, 1)$  vs.  $H_1 : P \sim f(p|\theta)$ , where  $\theta$  is unknown with the prior density  $\pi(\theta)$ , is:

$$B(p) = \frac{1}{\int_{\theta} f(p|\theta) \cdot \pi(\theta) d\theta}$$

For these  $P$ -values, Sellke et al. [40] developed a lower bound on the Bayes factor  $B(p)$ .

$$B(p) \geq -e \cdot p \cdot \log(p) \text{ if } p < 1/e, \tag{1.12}$$

where  $p$  denotes the  $P$ -value. Sellke et al. [40] also justify the lower bound on the Bayes factor of  $H_0$  to  $H_1$  in equation (1.12), assuming the density  $f_1(y)$  of  $Y = -\log(p)$  under  $H_1$  has a decreasing failure rate.

Following from equation (1.12), we have the lower bound on the Type I conditional error probability of  $T^C$  as defined in equation (1.8),  $\alpha_{\theta}(p)$ , for  $p < 1/e$

$$\inf_{\theta} \alpha_{\theta}(p) = \left( 1 + \frac{1}{\inf_{\theta} B(p)} \right)^{-1} = \left( 1 + \frac{1}{-e \cdot p \cdot \log(p)} \right)^{-1}$$

The lower bound on the Type I conditional error probability (or the posterior probability of  $H_0$ ) is as follows.

$$\alpha(p) \geq (1 + (-e \cdot p \cdot \log(p))^{-1})^{-1} \text{ if } p < 1/e, \tag{1.13}$$

Various  $P$ -values and their Bayesian calibrations as shown in equation (1.12) are presented in Table (1.1). In terms of the frequentist Type I conditional error probability in rejecting  $H_0$ , the calibrations as shown in equation (1.13) are also presented in Table (1.1).

Calibrations of $P$ -values as Bayes Factor and Conditional Error Probability						
$P$ -values	0.2	0.1	0.05	0.01	0.005	0.001
$B(p) \geq -ep \log(p)$	0.870	0.625	0.407	0.125	0.072	0.0188
$\alpha(p) \geq (1 + (-ep \log(p))^{-1})^{-1}$	0.465	0.385	0.289	0.111	0.067	0.0184

Table 1.1: Calibrating  $P$ -values

As noted from Table (1.1),  $p = 0.05$  translates into odds  $B(0.05) = 0.407$  of  $H_0$  to  $H_1$ , and frequentist error probability  $\alpha(0.05) = 0.289$  in rejecting  $H_0$ . It is pretty clear that  $p = 0.05$  does not indicate particularly strong evidence against  $H_0$  (roughly 1 to 2.5). Even  $p = 0.01$  corresponds to only 8 to 1 odds against  $H_0$ .

We simulate the calibration on Bayes factors provided by  $P$ -values for  $H_0$  to  $H_1$  from a variety of different distributions as follows. Here, it is noted that we have nonparametric alternatives

and will simply collect all the  $P$ -values from a number of tests, composed of null hypotheses and alternative hypothesis, and will record how often the null hypothesis is true for  $P$ -values at various levels. Throughout the simulations, the proportion of these tests having true null hypotheses given the initial proportion of true nulls is to be illustrated.

Suppose that each test  $j$  is based on normal data (known variance  $\sigma_j$ ) with mean  $\theta_j$ , so that our hypothesis testing is as follows.

$$H_0 : \theta_j = 0 \quad \text{vs.} \quad H_1 : \theta_j \neq 0$$

We must choose  $\pi_0$ , the initial proportion of true null hypotheses, and also the values of  $\theta_j$  under the alternative hypotheses. For each hypothesis, one then generates normal data with mean  $\theta_j$ , and computes the corresponding  $P$ -value, defined for the usual test statistic,

$$T(X) = \frac{\sqrt{n_j} \cdot \overline{X_j}}{\sigma_j},$$

$$p = 2 [1 - \Phi(T(X))] \tag{1.14}$$

where  $n_j$ ,  $\sigma_j$ , and  $\overline{X_j}$  are the sample size, standard deviation, and sample mean corresponding to the  $j^{\text{th}}$  hypothesis test;  $\Phi$  is the standard normal cumulative distribution function.

After generating a large series of tests, one looks at the subset of  $p$ -values which are near a specified value, such as 0.05 and 0.01. For instance, we can look at those tests for which  $0.0455 \leq p \leq 0.05$  and  $0.009 \leq p \leq 0.01$ . One then simply notes the proportion of such tests for which  $H_0$  is true. A MATLAB code for carrying out this simulation is given in the Appendix, which also discusses some further details, such as choice of the alternatives  $\theta_j$  from different distributions.

We create a histogram that indicates where the  $p$ -values, defined in equation (1.14), fall that are generated from the null hypotheses, and also a histogram of the  $p$ -values generated under the alternative hypotheses. We illustrate the histograms corresponding to  $0.0455 \leq p \leq 0.05$  and  $0.009 \leq p \leq 0.01$ ; the histograms that would result from such  $p$ -values under the null hypotheses are represented in Figure (1.3) and (1.4) by the white columns and under the alternative hypotheses by the black columns.

Under the alternative hypotheses, we must choose  $n_j$ ,  $\sigma_j$ , and  $\theta_j$ . A variety of possible specifications of  $\theta_j$ , the means of the alternatives, are implemented. In our simulation, the number  $l$  of the usual normal test statistic,  $T(X) = \frac{\sqrt{n_j} \cdot \overline{X_j}}{\sigma_j}$ , are generated with the known standard deviation,  $\sigma$ , and sample size,  $n$ . We choose  $\sigma_j = 1$  and  $n = 20$  below. Based on Sellke et al. [40], “The specific choices of  $n_j$ ,  $\sigma_j$ , and  $\theta_j$  are irrelevant and could vary from test to test; all that really matters is the choice of the  $\eta_j = \frac{\sqrt{n_j} \cdot \theta_j}{\sigma_j}$  .” As follows,  $\eta_j$  is featured by the value of  $a$ . Actually

the value of  $a$  measures the separation of the means under the nulls and the alternatives, which is demonstrated in Figure (1.3) and (1.4). As quoted from Sellke et al. [40], “ Finding the value of  $a$  that minimizes the proportion of true nulls is an interesting exercise.” In Figure (1.3), we choose  $a$  equals 2 sample standard deviations from the null mean, that is,  $a = \frac{\sqrt{5}}{5}$ ; and half of the sample standard deviation from the null mean, that is,  $a = \frac{\sqrt{5}}{20}$  in Figure (1.4). Another feature that must be specified are  $\pi_0$ , the initial proportion of true nulls, and  $\theta_1$ , the means under the alternatives. The simulation could be conducted with any desired sequence of alternative means, but the simulation below accommodates six cases. In our simulation, however, we consider  $\theta_1$  not only under symmetric distributions but also under nonsymmetric distributions as in (e) and (f) of Figure (1.3) and (1.4).

The distribution for null hypotheses is  $X_0 \sim N(0, \frac{\sigma^2}{n})$ .

The distribution for alternative hypotheses is  $X_1 \sim N(\theta_j, \frac{\sigma^2}{n})$ .

We consider the six cases for alternatives means  $\theta_j$ . The black columns in Figure (1.3) and (1.4) give the corresponding numbers of true alternatives of  $p$ -values over the ranges  $0.0455 \leq p \leq 0.05$  and  $0.009 \leq p \leq 0.01$  with the means under alternative hypotheses as follows.

- all alternative means  $\theta_1$  are fixed at the value  $a$ .
- all alternative means  $\theta_1$  are randomly generated from a normal distribution with mean 0 and standard deviation  $a$ , that is,  $\theta_1 \sim N(0, a)$ .
- all alternative means  $\theta_1$  are randomly generated from the corresponding positive half normal distribution above, that is,  $\theta_1 \sim N(0, a)$  and  $X_1 \sim N(|\theta_1|, \frac{\sigma^2}{n})$ .
- all alternative means  $\theta_1$  are randomly generated from a uniform distribution on the interval  $(-a, a)$ , that is,  $\theta_1 \sim \text{Uniform}(-a, a)$ .
- 50% alternative means  $\theta_1$  and 50% negative alternative means  $-\theta_1$  are randomly generated from a exponential distribution with mean  $a$ , that is,  $|\theta_1| \sim \text{Exp}(a)$ .
- all alternative means  $\theta_1$  are randomly generated from a shifted exponential distribution with mean  $a$ , that is, PDF of  $\theta \sim \text{Exp}(a)$  is shifted to the left with  $a$  units.

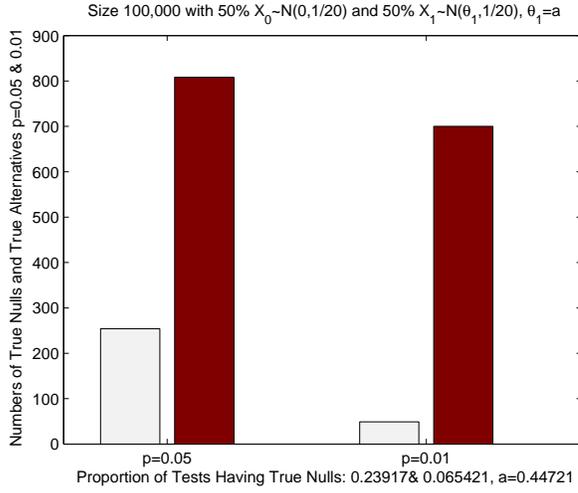
We calibrate the proportion of  $T$ -values in  $(1.96, 2]$  (that is, with  $0.0455 \leq p \leq 0.05$ ), and in  $(2.576, 2.616]$  (that is, with  $0.009 \leq p \leq 0.01$ ) for which the null hypothesis is true (See the Appendix for the implementation).

When  $p \approx 0.05$  in Figure (1.3) and (1.4), for the six cases considered in our simulation and if the initial percentage of true nulls is 50%, the corresponding minimum percentages is 23%. This corresponding minimum percentages is the same as Sellke et al. [40].

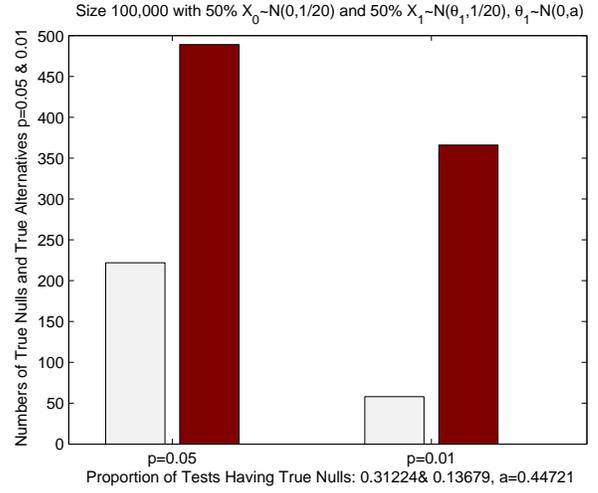
As demonstrated in Figure (1.3) and (1.4), smaller values of  $p$  are more likely under the alternatives than under the nulls, but the degree to which this is so is rather modest for  $p$ -values

in common regions. For instance, a  $p$ -value in the interval  $(0.04, 0.05)$  is essentially equally likely to occur under the nulls as under the alternatives when  $a = \frac{\sqrt{5}}{20}$  in Figure (1.4); and is more likely to occur under the alternatives when  $a = \frac{\sqrt{5}}{5}$  in Figure (1.3). Thus observing, say,  $p = 0.046$  provides no evidence in favor of the null or the alternative.

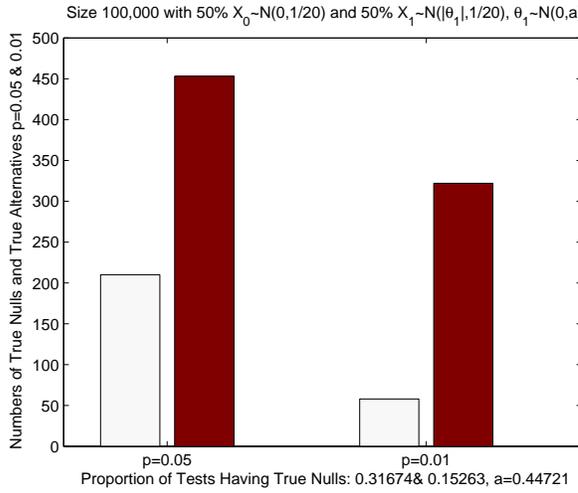
The natural question to ask is whether the qualitative nature of the phenomenon observed in Figure (1.3) and (1.4) is due to the particular choice we made for the alternatives. Sellke et al. [40] put in this way: “It can be shown that, no matter how one chooses the sample size, standard deviation, and sample mean corresponding to each test under the alternatives, at most 3.7% of the  $p$ -values will fall in the interval  $(0.04, 0.05)$ , so that a  $p$ -value near 0.05 provides at most 3.7 to 1 odds in favor of the alternative hypothesis test.” This is actually just a restatement of the earlier observation that, if 50% of the nulls are initially true, then at least 23% of those with a  $p$ -value near 0.05 will be true. The clear message is that knowing that the data are rare under the nulls is of little use unless one determines whether or not they are also rare under the alternatives.



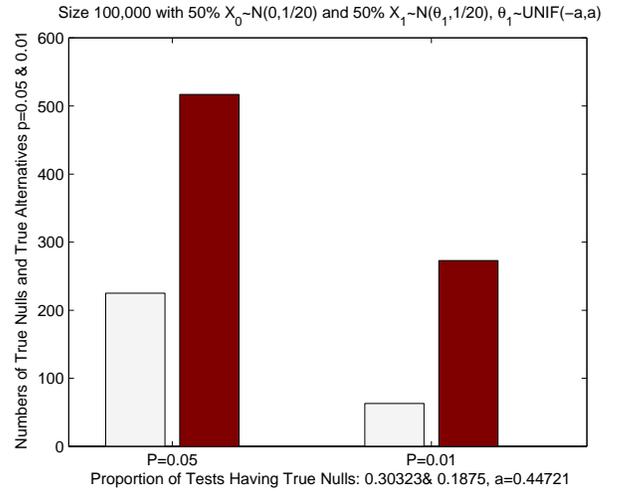
(a)  $\theta_1$  fixed at  $a$



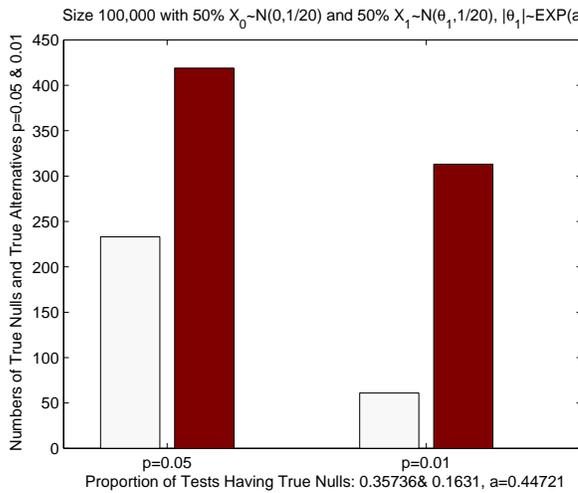
(b)  $\theta_1$  generated from normal distribution



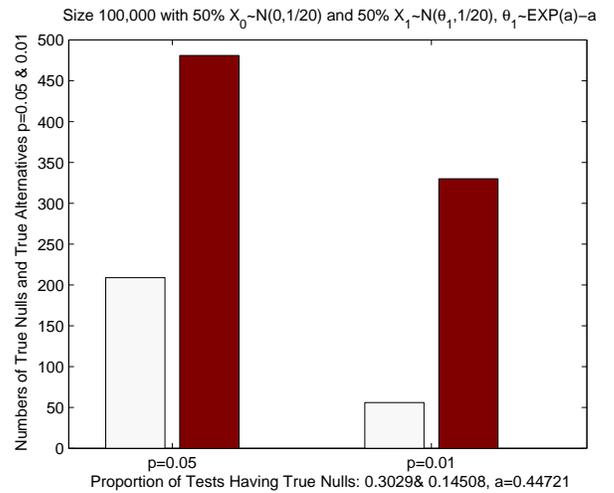
(c)  $\theta_1$  from the positive half normal distribution



(d)  $\theta_1$  generated from uniform distribution

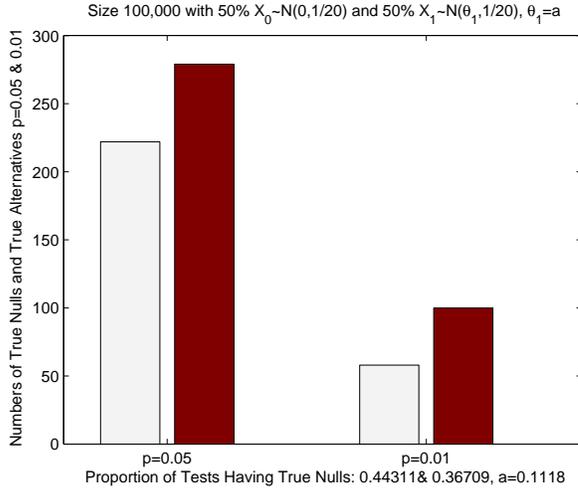


(e) 50% $\theta_1$  and 50%  $-\theta_1$  from exponential

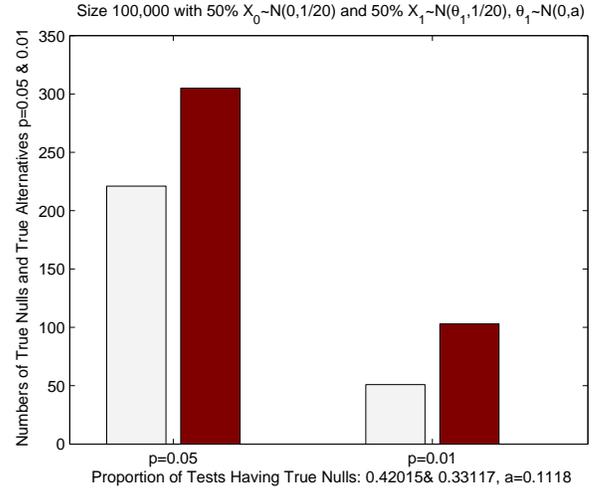


(f)  $\theta_1$  from exponential with mean  $a$  shifted  $a$  units to left

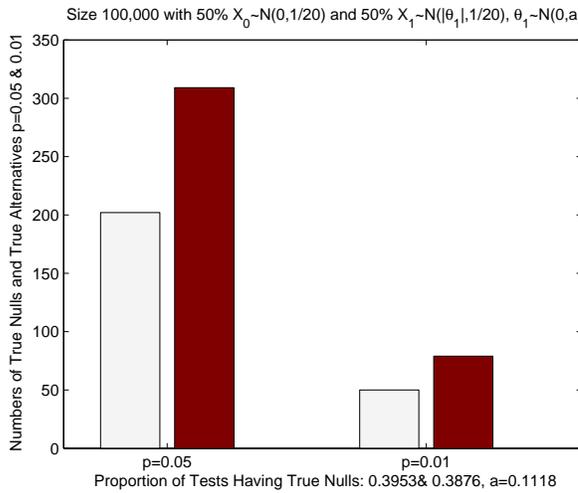
Figure 1.3: True  $H_0$  (White) and true  $H_1$  (Black) over the ranges  $0.0455 \leq p \leq 0.05$  and  $0.009 \leq p \leq 0.01$  with  $a = \frac{\sqrt{5}}{5}$  equal to 2 sample standard deviations from the null mean.



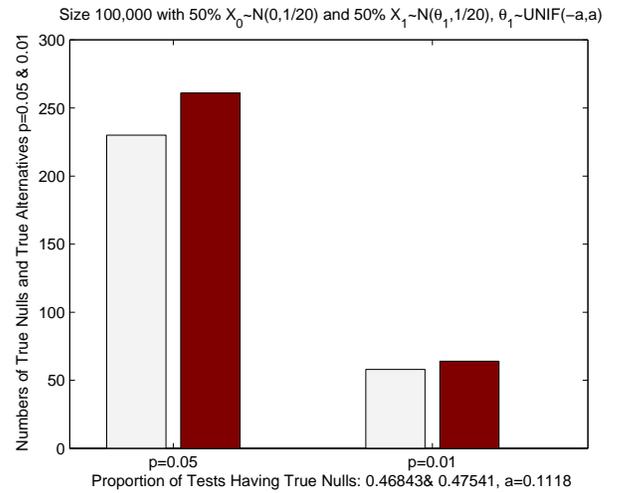
(a)  $\theta_1$  fixed at  $a$



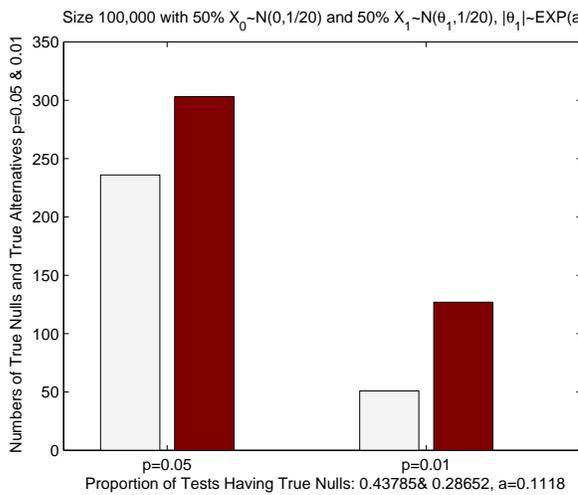
(b)  $\theta_1$  generated from normal distribution



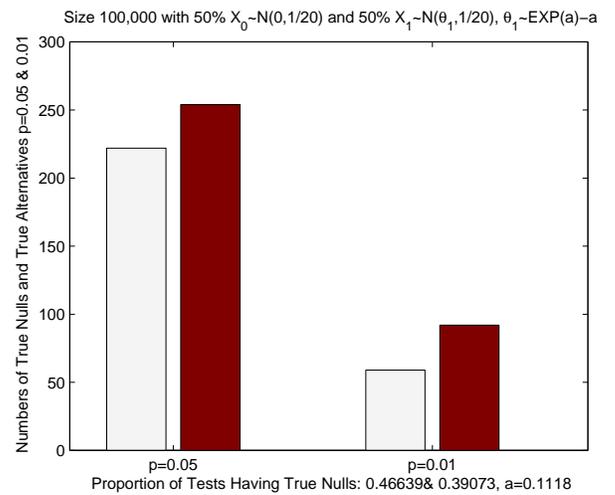
(c)  $\theta_1$  from the positive half normal distribution



(d)  $\theta_1$  generated from uniform distribution



(e) 50% $\theta_1$  and 50%  $-\theta_1$  from exponential



(f)  $\theta_1$  from exponential with mean  $a$  shifted  $a$  units to left

Figure 1.4: True  $H_0$  (White) and true  $H_1$  (Black) over the ranges  $0.0455 \leq p \leq 0.05$  and  $0.009 \leq p \leq 0.01$  with  $a = \frac{\sqrt{5}}{20}$  equal to half of the sample standard deviation from the null mean.

## Chapter 2

# DETERMINING POSTERIOR *P*-VALUES FROM EMPIRICAL DISTRIBUTIONS

### 2.1 Multiplicity of hypothesis testing and the false discovery rate

Multiple hypothesis testing refers to the testing of more than one hypothesis simultaneously. A number of approaches for multiple hypothesis testing are reviewed by Shaffer [35]. As discussed in Chapter 1, in testing any simple hypothesis, generally conclusions based on some test statistic are uncertain. We typically choose an acceptable maximum probability of rejecting the true null hypothesis (significance level), thus committing Type I error, and base the conclusion on the value of this test statistic meeting the specification. However, when many hypotheses are tested and each test has a specified Type I error probability, the probability that some Type I errors are committed potentially increases with the number of hypotheses. In another words, if  $n$  independent hypothesis tests are performed, the experiment-wide significance level  $\alpha$  is given by

$$\alpha = 1 - (1 - \alpha_{\text{per test}})^{\text{number of tests}}.$$

Numerous methods have been proposed for dealing with this multiple testing problem. For instance, Bonferroni method is aimed to retain the same overall Type I error rate (rather than a higher rate) in multiple testing by reducing the size of the allowable error  $\alpha_{\text{per test}}$  by the number of tests. The resulting overall  $\alpha$  does not exceed the desired limit without requiring any

independence assumption.

However, it can be demonstrated that simple techniques such as the Bonferroni method are conservative as defined in Section (1.1.4), so there has been a great deal of attention paid to developing better techniques such that the overall rate of false positives (Type I error) can be maintained without inflating the rate of false negatives (Type II error) unnecessarily as discussed in Section (1.1.3). Henceforth, when performing multiple hypothesis testing simultaneously in the analysis of large data sets, one should pay attention not only to false discovery rate, developed by Benjamini and Hochberg [10], but also to the false negative rate since traditional concepts of size and power are unable to handle the multiplicity of tests. The definition of the false discovery rate is presented shortly.

Regarding the compound error measure of multiple hypothesis testing, as noted above, it is necessary to measure the overall error rate when we handle a number of tests simultaneously. Benjamini and Hochberg [10] proposed a compound error measure based on the false rejections for multiple testing.

Consider  $n$  null hypotheses  $H_1, \dots, H_n$  simultaneously, of which  $n_0$  are true nulls. Let  $H_i = 0$  when the  $i^{th}$  null hypothesis is true and  $H_i = 1$  otherwise. The outcome of the  $n$  tests above are categorized in Table (2.1) Based on Table (2.1), a false discovery rate FDR is defined as follows.

	Accept null	Reject null	Total
Null true	U	V	$n_0$
Null false	T	S	$n_1 = n - n_0$
	W	R	$n$

Table 2.1: Outcomes of  $n$  hypotheses

**Definition 2.1 (Definition of false discovery rate)**

$$\text{FDR} = \text{E} \left[ \frac{V}{\max(R, 1)} \right] = \text{E} \left[ \frac{V}{R} \mid R > 0 \right] \cdot \Pr(R > 0) \tag{2.1}$$

Estimates of false negative rates based on the empirical distribution of  $P$ -values from many significance tests were proposed by Bickis, Bleuer, and Krewski [4]. Bickis [3] proposed another estimate on calibrating the null and alternative components of the distribution by modeling the set of  $P$ -values as a sample from a mixed population composed of a uniform distribution for the null cases and an unknown distribution for the alternatives. The mixture distribution will be discussed further from a Bayesian perspective. Conditioning on the actual mixture of nulls and alternatives in the data set instead of an arbitrarily prespecified false discovery rate, such methodology will allow one to set a threshold of significance, to measure the separation between the nulls and positives, and to filter out the null distribution by evaluating the posterior probabilities of the

null hypothesis given  $P$ -values. As well, in his paper, the techniques are illustrated using both real and simulated data.

$P$ -value is defined as Definition (1.1). In order to simplify the following mathematical discussion,  $Q$ -value is introduced.

**Definition 2.2 (Definition of  $Q$ -value)**

$$Q\text{-value} = 1 - P\text{-value}, \quad 0 \leq Q\text{-value} \leq 1, \quad (2.2)$$

which means that large  $Q$ -values, close to one, represent strong evidence against the null hypothesis. Note that the  $Q$ -value, as defined in equation (2.2), has no relationship with  $Q$ -value of Storey [44].

## 2.2 Exploiting the empirical distribution of $P$ -values

As discussed previously, the distribution of  $P$ -values are exploited to measure evidence against the null hypothesis rather than a binary decision-making on the basis of some arbitrarily predetermined significance level, which ignores the variation with the strength of evidence in the data sets.

**Definition 2.3 (Definition of empirical cumulative distribution function (ECDF))** *Let  $x_1, \dots, x_n$  be a set of data and let  $y_1 < y_2 < y_3, \dots, < y_n$  be the ordered values of the data set. Then the empirical cumulative distribution function based on this data set can be represented as follows:*

$$F_n(x) = \begin{cases} 0 & x < y_1 \\ \frac{j}{n} & y_j \leq x < y_{j+1} \\ 1 & y_n \leq x \end{cases} \quad (2.3)$$

Provided that the global null hypothesis is true, the cumulative distribution function (CDF) and empirical cumulative distribution function (ECDF) of  $Q$ -values are illustrated in Figure (2.1).

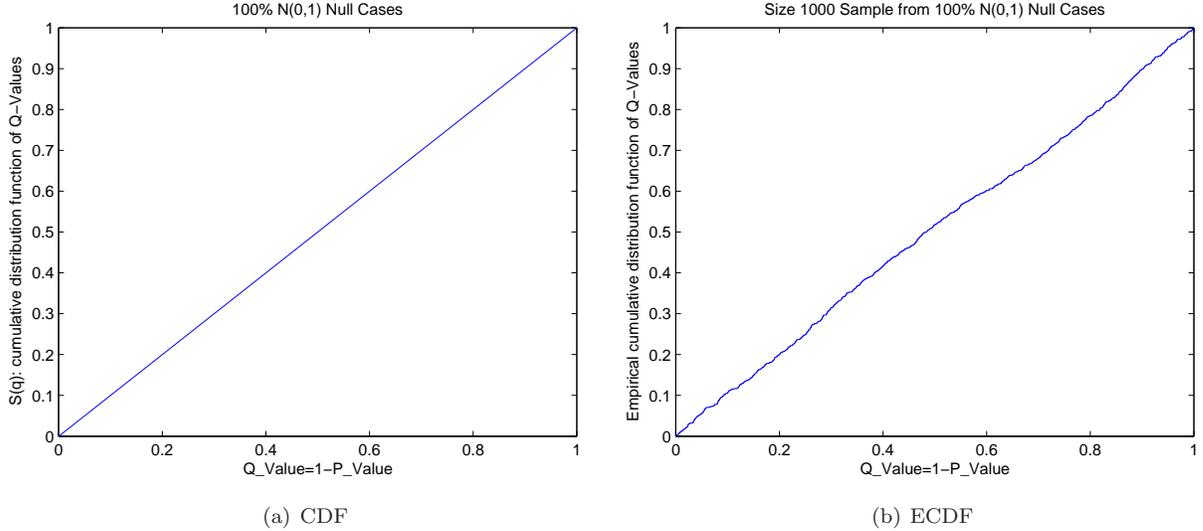


Figure 2.1: Cumulative distribution function and empirical cumulative distribution function of  $Q$ -values with size 1000 sample from 100%  $N(0, 1)$  null cases.

In Figure (2.2) and Figure (2.3), cumulative distribution functions and the empirical cumulative distribution functions of a collection of observed  $P$ -values from a normal mixed population composed of a uniform distribution for the null cases and the other distribution for the alternatives are shown as follows. From both the cumulative distribution functions in Figure (2.2) and the empirical cumulative distribution functions in Figure (2.3), a deviation from uniformity is indicated. In the following discussions, the deviation from uniformity is calibrated and then the posterior  $P$ -value from empirical distributions is determined. In order to clarify what follows and simplify mathematical discussions, one-tailed  $P$ -values are used and only cases in which the alternative is truly one-sided are considered since we are investigating any deviations from the null. In Figure (2.2), the normal mixed populations consist of the proportion of  $\pi_0$  of  $N(0, 1)$  for the null cases and the proportion of  $1 - \pi_0$  of  $N(1, 1)$  or  $N(2, 1)$  for the alternative cases. Here, the proportion of  $\pi_0$  for the null cases is fixed. Figure (2.3) shows the empirical distributions from 1000  $Q$ -values. As well, Figure (2.3) does show that there are more  $Q$ -values close to one in the left two panels than those in the right two panels, indicating mixing of different components and giving evidence that there exist indeed positive cases and some relationship between  $P$ -values and the posterior probability of the null hypothesis. Moreover, such a conclusion can be reached without having to commit to any particular prespecified significance level in order to set an arbitrary cut-off value for a binary conclusion. Figure (2.4) is set up using conventional hypothesis testing approach. This cut-off value totally ignores the variation with the strength of evidence in the data sets and the shape of the empirical distribution. For handling the multiplicity of tests whenever the number of tests increases, the cutoff is shifted to the right. However, the shifted cutoff using the Bonferroni method makes it very difficult to achieve significance with thousands

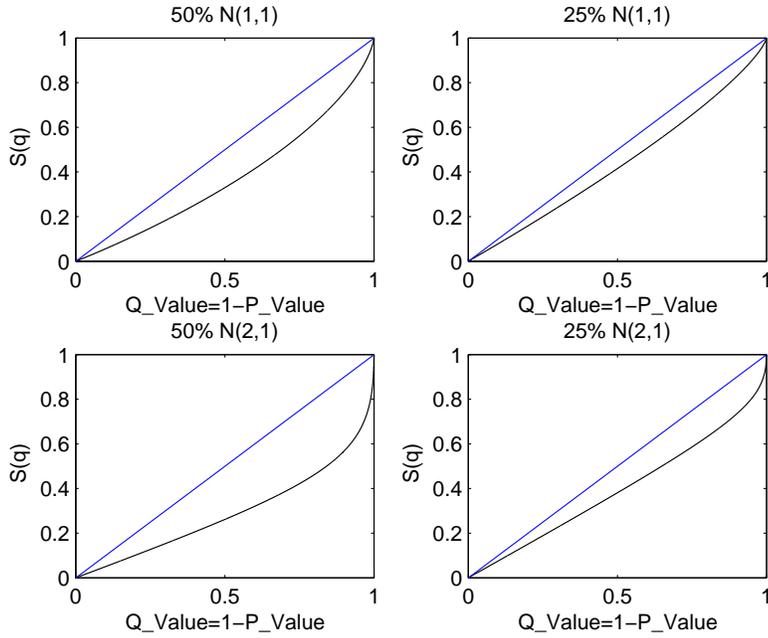


Figure 2.2:  $S(q)$ , cumulative distribution functions of  $Q$ -values of the four different normal mixed populations compared with  $100\%N(0, 1)$  null cases corresponding to the diagonal lines.

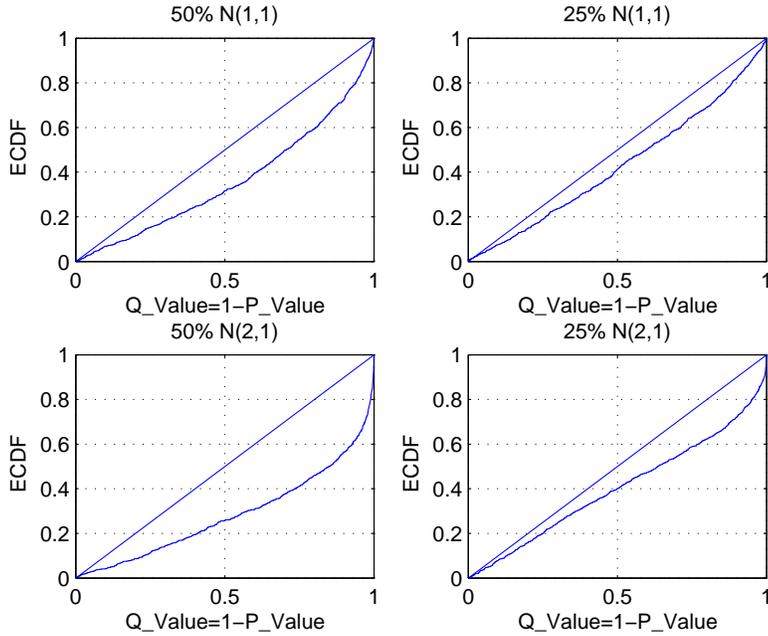


Figure 2.3: Empirical cumulative distribution functions of  $Q$ -values with size 1000 sample from the four different mixed populations compared with  $100\%N(0, 1)$  null cases corresponding to the diagonal lines.

of tests typically arising in bioinformatics. One popular multiple hypothesis testing approach is based on the false discovery rate of Benjamini and Hochberg [10]. As illustrated in Figure (2.4),

this method is equivalent to drawing a line with reciprocal slope equal to the prespecified false discovery rate from the upper right-hand corner. Hence, the rejection region is composed of points to the right of the first intersection point at which the empirical distribution crosses that line with the desired slope. If the corresponding  $Q$ -value is within the rejection region, the null hypothesis test is rejected.

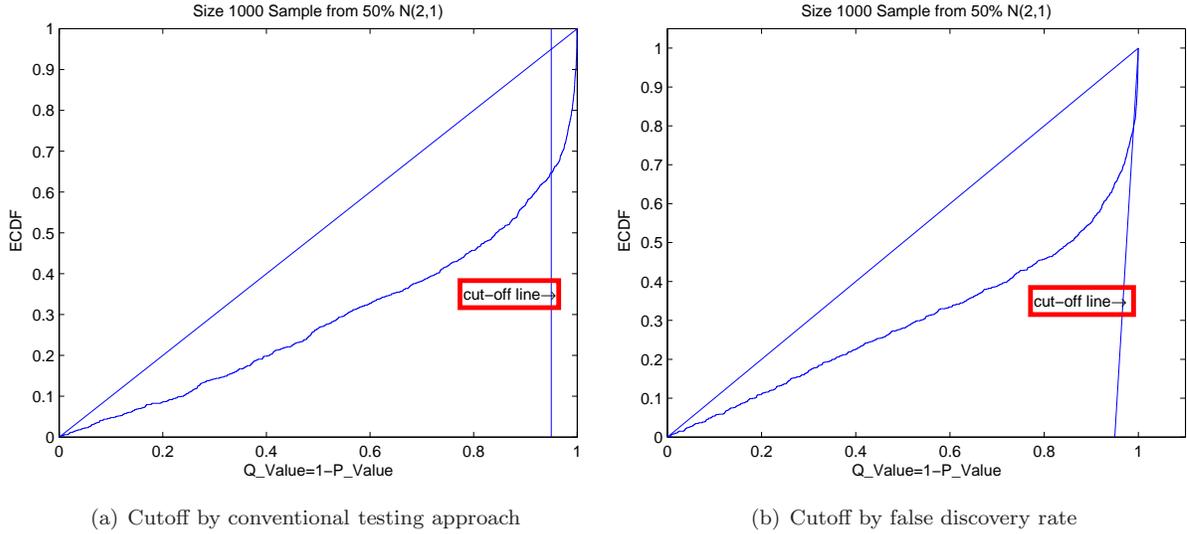


Figure 2.4: Cutoff set up by conventional approach such as Bonferroni method in the left panel; and in the right panel cutoff based on the false discovery rate proposed by Benjamini and Hochberg, with size 1000 sample from 50% $N(2, 1)$  null cases.

### 2.3 Modeling $P$ -values as a sample from a mixed population

As noted above, the hypothesis testing approach is aimed to control the false discovery rate. From a Bayesian perspective, however, the following approach, proposed by Bickis [3], is to view this as inference problem in which one recognizes the existence of a sub-population of nulls and a sub-population of alternatives instead of a problem of binary decision. In particular, the approach is aimed to distinguish the sub-population of nulls as well as possible, to calibrate  $P$ -value by computing the posterior probability of the null hypothesis, given a  $P$ -value in the light of deviations from uniformity of the empirical distribution of  $P$ -values, and to make inferences about estimates of the probability of the null hypothesis being true.  $Q$ -value (or  $P$ -value, equivalently a test statistics  $T$ -value) is therefore employed to calibrate the posterior probabilities of the null hypothesis given  $P$ -values, which are also referred to as the posterior probabilities of being in error by accepting the directional conclusion (one-tailed) suggested by the data sets.

Suppose that our null hypothesis is denoted by  $H_0$ , the alternative hypotheses is denoted by  $H_1$  which could be composite alternatives, a mixture of the various alternatives, and our test

statistics  $T$  is continuous, where the null hypothesis is to be rejected for large values of  $T$  i.e. one-tailed.

Let

$$\text{Under } H_0 : F_0(t) = \Pr(T \leq t | H_0 \text{ true})$$

$$\text{Under } H_1 : F_1(t) = \Pr(T \leq t | H_1 \text{ true})$$

and let their probability density functions, corresponding to  $F_0(t)$  and  $F_1(t)$ , be  $f_0(t)$  and  $f_1(t)$  respectively.

Therefore, the  $P$ -value is :

$$P = \Pr(T \geq t | H_0) = 1 - \Pr(T \leq t | H_0) = 1 - F_0(t) \quad (2.4)$$

and for any  $0 < v < 1$ ,

$$\Pr(P \geq v | H_1) = \Pr(T \leq F_0^{-1}(1 - v) | H_1) = F_1(F_0^{-1}(1 - v)) \quad (2.5)$$

From equation (2.2)  $Q = 1 - P$  and from equation (2.4), equation (2.5) can be simplified as follows.

If  $F_0(t)$  and  $F_1(t)$  are continuous,

$$\text{Under } H_0 : S(q | H_0) \equiv \Pr(Q \leq q) = q, \quad \text{and}$$

$$\text{Under } H_1 : S(q | H_1) \equiv \Pr(Q \leq q) = F_1(F_0^{-1}(q))$$

Let a proportion  $\pi_0$  of the cases be null and  $S(q)$  be the cumulative distribution function of  $Q$ -values from the mixed population:

$$S(q) = \Pr(Q \leq q) \quad (2.6)$$

$$= \Pr((Q \leq q) \cap (H_0 \cup H_1)) \quad (2.7)$$

$$= \Pr((Q \leq q) \cap H_0) + \Pr((Q \leq q) \cap H_1) \quad (2.8)$$

$$= \Pr((Q \leq q) | H_0) \cdot \Pr(H_0) + \Pr((Q \leq q) | H_1) \cdot \Pr(H_1) \quad (2.9)$$

$$= \sum_i S(q | H_i) \cdot \Pr(H_i) \quad (2.10)$$

$$= S(q | H_0) \cdot \Pr(H_0) + S(q | H_1) \cdot \Pr(H_1) \quad (2.11)$$

$$= \pi_0 \cdot q + (1 - \pi_0) \cdot F_1(F_0^{-1}(q)) \quad (2.12)$$

Hence,

$$S'(q) = \pi_0 + (1 - \pi_0) \cdot \lambda(t) \quad (2.13)$$

where

$$\begin{aligned}
\lambda(t) &= \frac{d}{dq} F_1(F_0^{-1}(q)) \\
&= F_1'(F_0^{-1}(q)) \cdot \frac{d}{dq} F_0^{-1}(q) \\
&= F_1'(F_0^{-1})/F_0'(F_0^{-1}) \\
&= F_1'(t)/F_0'(t) \\
&= f_1(t)/f_0(t)
\end{aligned}$$

Note that in order to simplify the mathematical discussion, here,  $\lambda(t) = f_1(t)/f_0(t)$  is the reciprocal of the Bayes factor,  $B(x) = \frac{p(x|H_0)}{p(x|H_1)}$ , as discussed in Chapter 1.

Since we are handling one-tailed  $P$ -values, the tidiest situation is one in which

$$\lim_{t \rightarrow -\infty} \lambda(t) = 0 \quad \text{and} \quad \lim_{t \rightarrow \infty} \lambda(t) = \infty. \quad (2.14)$$

equation (2.14) means that extreme values will give overwhelming evidence in favor of either the null or alternative hypotheses. Assuming equation (2.14) and plugging in equation (2.13), we can see that

$$S'(0) = \pi_0. \quad (2.15)$$

Recall that the goal of the approach discussed above is to separate out the uniform component by viewing the  $Q$ -values as a sample from a mixed distribution containing a uniform component. In order to distinguish the uniform component from the mixture of components, it is appealing to seek a density estimate of  $Q$ -values that indicates mixing of components. We describe the  $Q$ -value sample by estimating its density in a nonparametric way. A kernel density estimator on the  $Q$ -values, for example, could be employed to produce a family of density estimators based on a number of smoothing parameters, which is discussed as follows.

## 2.4 Density estimation based on kernel methods

Silverman [43] provides a practical description of density estimation based on kernel methods.

**Definition 2.4** *Let  $X_1, \dots, X_n$  denote a sample of size  $n$  from a random variable with density  $f$  unknown and observed values  $x_1, \dots, x_n$ . The kernel density estimate of  $f$  at the point  $x$  is given by*

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K((x - x_i)/h),$$

where the function  $K(t)$  is said to be the kernel of the estimator. In general, the kernel  $K(t)$

satisfies the conditions

$$\int_{-\infty}^{\infty} K(t)dt = 1.$$

It is well known that the performance of kernel density estimators depends crucially on the value of the smoothing parameter, which is commonly referred to as the bandwidth of the estimator, which controls the degree of smoothness that the resulting function exhibits.

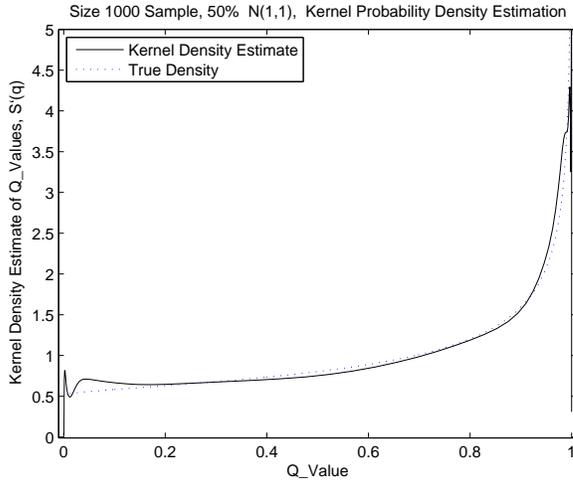
In SAS, PROC KDE produces kernel density estimates based on the usual Gaussian kernel (i.e., the Gaussian density with mean 0 and standard deviation 1), whereas S-PLUS has a function density which produces kernel density estimates with a default kernel, the Gaussian density with mean 0 and standard deviation 1/4. The program R also has a function density which produces kernel density estimates with a default kernel, the Gaussian density with mean 0 and standard deviation 1.

The MATLAB ksdensity function (Reference: Bowman et al. [12]) does this by using a kernel smoothing function and an associated bandwidth to estimate the density as follows.

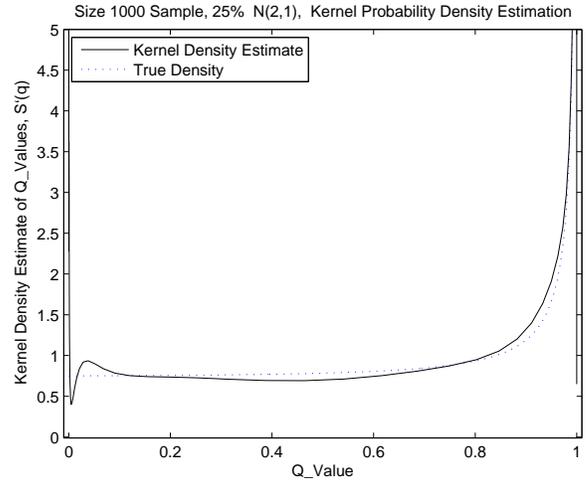
$[F, XI] = \text{KSDENSITY}(X)$  computes a probability density estimate of the sample in the vector  $X$ . KSDENSITY evaluates the density estimate at 100 points covering the range of the data.  $F$  is the vector of density values and  $XI$  is the set of 100 points. The estimate is based on a normal kernel function, using a bandwidth that is a function of the number of points in  $X$ .

Since the probability density estimation, KSDENSITY, does not handle the compact support  $[L, U]$  well, KSDENSITY transforms  $X$  using a logit function, estimates the density of the transformed values, and transforms back to the original scale. For instance, the support of our  $Q$ -values is  $[0, 1]$ , so KSDENSITY uses the transformation  $\log \frac{q}{1-q}$ , and then transforms back to produce the  $Q$ -value density estimator.

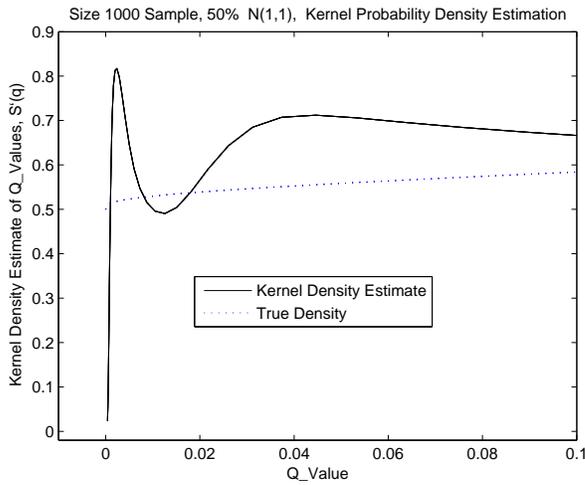
Figure (2.5) is produced with (default) normal kernel functions with default bandwidth chosen automatically according to data sets.



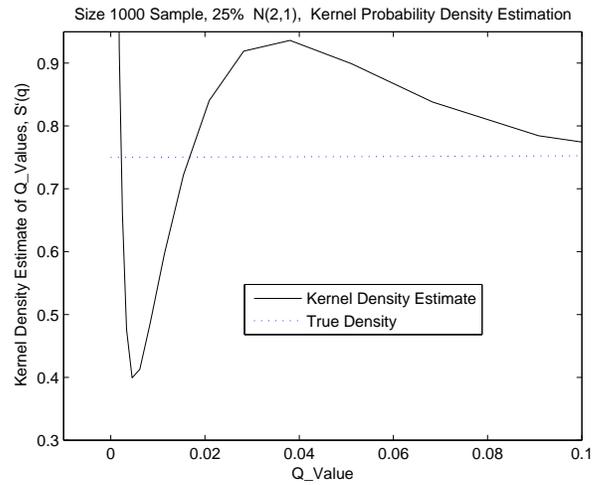
(a) Complete PDF estimate



(b) Complete PDF estimate



(c) Magnification of lower end above



(d) Magnification of lower end above

Figure 2.5: Using kernel density estimation, estimated PDF plots of  $Q$ -values sampled from the mixed population with the sample size  $n = 1000$ , of which in the left panel 500 are simulated from  $N(0, 1)$  and 500 are simulated from  $N(1, 1)$ ; in the right panel 750 are from  $N(0, 1)$  and 250 are from  $N(2, 1)$ . The two graphs in the upper panel show complete estimated PDF (solid lines) compared with the true PDF (dotted lines) of  $Q$ -values. The two graphs in the lower panel magnify respectively lower ends of the estimated PDF plots above to show that estimated PDF using the kernel density estimation do not fit to the data at the left end of the distribution.

The density estimators by the kernel approximation at the two endpoints  $Q$ -value = 0 and  $Q$ -value = 1 are so wiggly as shown in Figure (2.5).

## 2.5 Posterior probabilities of the null hypothesis given the $P$ -value and implementation

Here from a Bayesian paradigm, the interpretation of the posterior probability of null hypothesis given  $P$ -values is proposed. Suppose that  $\pi_0$  is known and let  $\omega = \frac{1 - \pi_0}{\pi_0}$  be the prior odds against the null hypothesis. Then, given a  $Q$ -value (or equivalently a test statistic  $T$ -value), the posterior probability of the null hypothesis is

$$\begin{aligned} \Pr(H_0|T = t) &= \frac{\pi_0 \cdot F'_0(t)}{\pi_0 \cdot F'_0(t) + (1 - \pi_0) \cdot F'_1(t)} \\ &= \frac{1}{1 + \frac{1 - \pi_0}{\pi_0} \cdot \frac{F'_1(t)}{F'_0(t)}} \\ &= \frac{1}{1 + \omega \cdot \lambda(t)} \end{aligned}$$

Since

$$\begin{aligned} S'(q) &= \pi_0 + (1 - \pi_0) \cdot \lambda(t) \\ &= \pi_0(1 + \omega \cdot \lambda(t)) \end{aligned}$$

From equation (2.15), we have that

$$\Pr(H_0|Q = q) = \frac{S'(0)}{S'(q)} \tag{2.16}$$

Equation (2.16) is interpreted geometrically as the following Figure (2.6). From equation (2.16) and Figure (2.6), we thus see that the posterior probability of the null hypothesis given a  $P$ -value (or equivalently a  $Q$ -value ) is given by a ratio of slopes.

Based on numerical differentiation, one can estimate this ratio of slopes by a ratio of slopes of secants from the empirical cumulative distribution (ECDF) as shown in Figure (2.7).

However, the secant estimator is somehow unstable since the empirical distribution (ECDF) is rather wiggly.

Since  $S'(q)$  can be estimated by numerical differentiation, we estimate  $1/S'(Q(i))$  by

$$\frac{1}{S'(Q(i))} = \frac{Q(i+k) - Q(i-k)}{S(Q(i+k)) - S(Q(i-k))} = \frac{Q(i+k) - Q(i-k)}{2k/n}, \tag{2.17}$$

where we can choose  $k$  to be any integer and  $Q$ -values' frequencies are evenly spaced.

However, these procedures discussed above give rather rough estimates of the posterior prob-

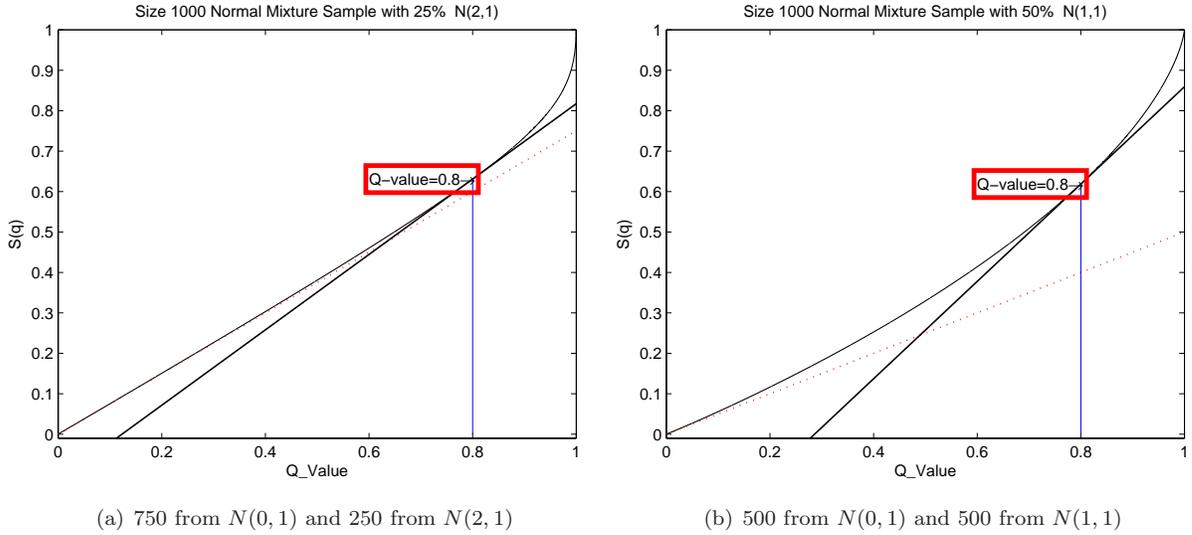


Figure 2.6: Posterior probability of the null hypothesis given  $Q$ -value equals odds ratio of the slopes.

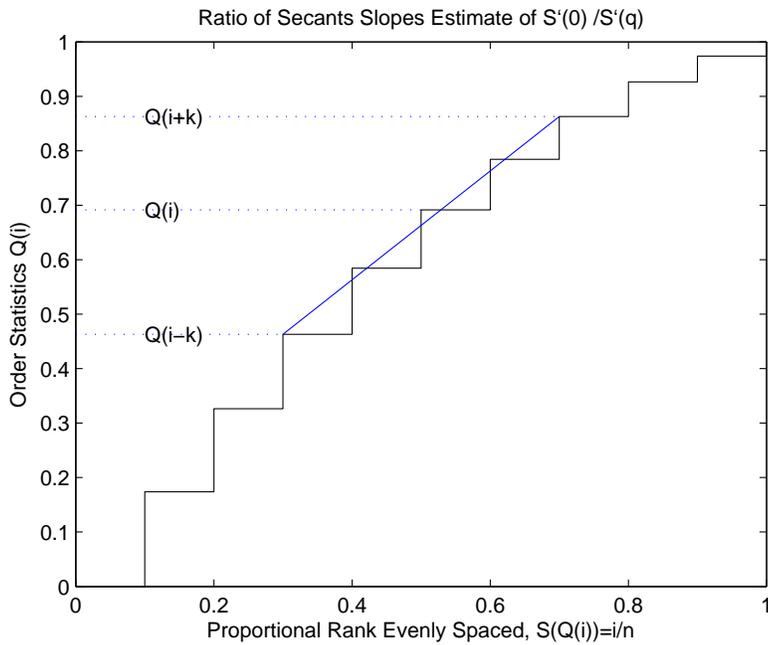


Figure 2.7: Posterior probability of the null hypothesis given  $Q$ -value equals odds ratio of the slopes

ability given the  $Q$ -value even though these estimates could be smoothed, for example, using the S-PLUS function `loess-smooth`. The function `loess-smooth` returns a list of values at which the loess curve is evaluated by running `loess-smooth(x, y, span, degree, family)`, where  $(x, y)$  like  $((x_i, y_i), i = 1, 2, \dots, n)$  are data points; the smoothing parameter, `span`, is related to the bandwidth of approximation; the smoothing parameter, `degree`, is overall degree of locally fitted

polynomial; the smoothing parameter, family, is about the smoothing function which is gaussian or symmetric.

Hence it is appealing to seek a smoothing density estimate of  $Q$ -values that approximates mixing of components.

To take advantage of the particular nature of the density, more appropriate density estimators such as smoothing B-spline estimates are presented in Chapter 3.

Based on B-splines as smoothing functions and taking into account properties of the density, such as support on  $[0, 1]$ , known uniform component, and convex empirical cumulative distribution function, our smoothing density estimate is developed and properties of various estimates of  $S'(q)$  are investigated in Chapter 3.

## Chapter 3

# THE EMPIRICAL $P$ -VALUE CALIBRATION

### 3.1 Smoothing and density estimation

Density estimation is closely linked computationally to smoothing. If one has a random sample from some distribution with observed values  $x_1, \dots, x_n$ , the natural estimate of the distribution function is the empirical CDF (ECDF), which is an average of  $n$  point masses at the observations. While the true CDF is often continuous and differentiable, the ECDF is neither, so it is natural to seek methods for approximating, smoothing, and even discovering patterns in data.

Density estimation methods can be thought of as applying smoothing techniques to a basic estimator, such as a histogram, where there is a tradeoff between fidelity to the data and smoothness. A curve forced to pass through all of the data points is rarely smooth. At the other extreme, the curve like  $f(x) = \bar{X}$  ( $\bar{X} = \sum_{i=1}^n x_i/n$ , the sample mean from some distribution with observed values  $x_1, \dots, x_n$ ), while very smooth, rarely captures all of the important features of the data. For instance, compared with the  $2^{nd}$  order B-spline in Figure (3.1), the smoothness of the  $0^{th}$  order (the histogram) and the  $1^{st}$  order B-splines is tradeoff against fidelity to the data. The degree of smoothness and fidelity to data points  $((x_i, y_i), i = 1, 2, \dots, n)$  can be measured by the penalized residual sum of squares [43]:

$$SS(\lambda) = \sum_{i=1}^n w_i (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx \quad (3.1)$$

where  $\lambda \geq 0$  is the smoothing parameter, and  $w_i (i = 1, 2, \dots, n)$  is the weighting parameter to adjust the residual of sum of squares. The first term on the right side of equation (3.1),  $\sum_{i=1}^n w_i (y_i - f(x_i))^2$ , is the weighted residual of sum of squares; the second term on the right

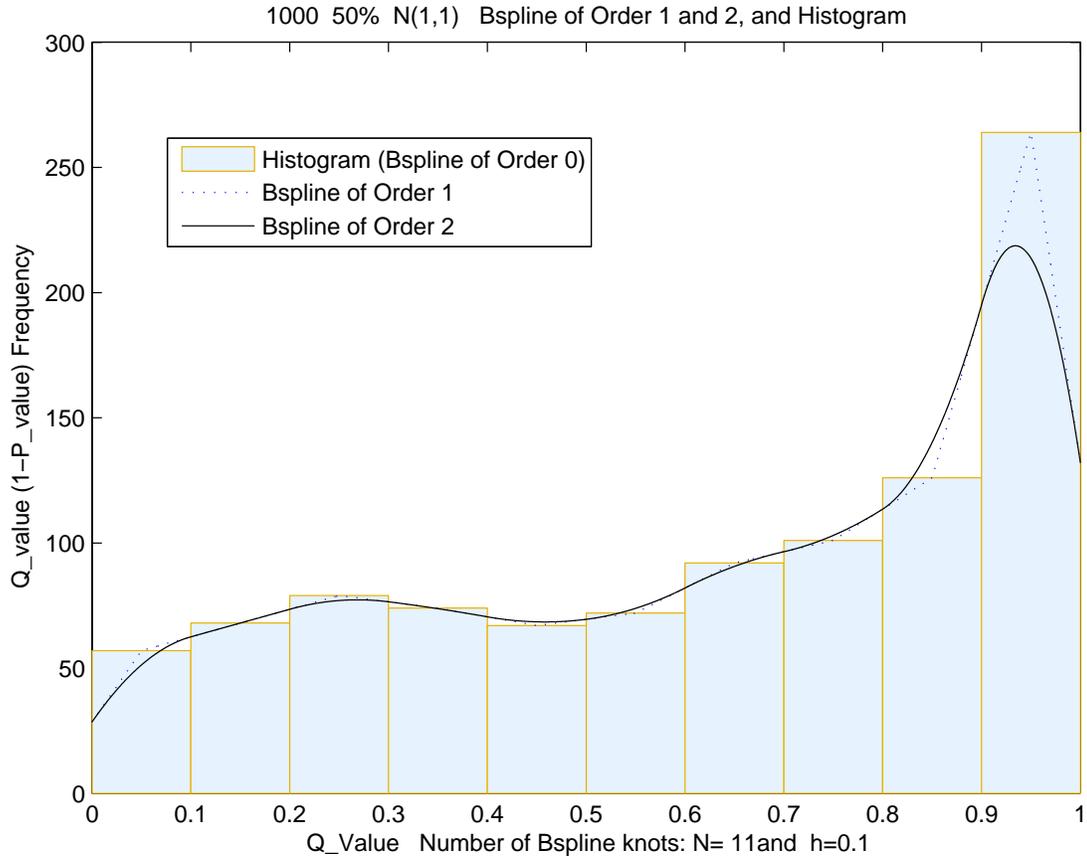


Figure 3.1: Fidelity to the data and smoothness compared between different orders of B-splines)

side of equation (3.1),  $\int_a^b (f''(x))^2 dx$ , is the roughness penalty. One then fits a curve  $f(x)$  to the data minimizing (3.1) over all twice continuously differentiable functions. If  $f(x)$  is linear, then it contributes nothing to the second term, the roughness penalty, so the roughness penalty approach is an extension of the ordinary least square estimation method. The smoothing parameter  $\lambda$  is the tradeoff between goodness of fit and smoothness. The larger the smoothing parameter  $\lambda$ , the more weight on the smoothness. If  $\lambda = 0$ , that is, a fitted function is allowed that is arbitrarily wiggly, then  $f(x)$  simply interpolates every data point with flexible slopes that might produce extreme roughness, provided that there are no duplicates among the  $x'_i$ s. On the other hand, if  $\lambda \rightarrow \infty$ , then  $f(x)$  should be chosen so that  $f''(x) = 0$  everywhere, which is a least square linear regression line. However, there has been considerable theoretical work on the problem of choosing the smoothing parameter  $\lambda$  less subjectively, which controls the degree of tradeoff between low variability (smoothness) and low bias (closeness to the data) [43]. De Boor [15] constructs his cubic smoothing splines by specifying the maximum allowable residual sum of squares, the first term on the right side of equation (3.1), instead of  $\lambda$ . As well, one can employ cross-validation

(CV) to choose  $\lambda$ . The idea is to hold out one observation at a time, say the  $i^{\text{th}}$  observation, then to construct a curve fit function  $f_{-i,\lambda}(x)$  under which the missing data point is best predicted by the remaining  $n - 1$  observations. The cross-validation function is defined as follows [46]:

$$CV(\lambda) = \sum_{i=1}^n n^{-1} (y_i - f_{-i,\lambda}(x_i))^2. \quad (3.2)$$

The smoothing parameter  $\lambda$  is chosen to minimize  $CV(\lambda)$  in equation (3.2). Actually this ordinary cross validation leaving out one point at a time is designed to obtain an unbiased estimate of the predictive mean squared error (PMSE). At a single point  $x$ , the PMSE is simply

$$\text{PMSE}(x, \lambda) = E [f_\lambda(x) - E(Y|x)]^2,$$

where  $f_\lambda(x)$  is a curve fit function predicted by all the observed data points. For a given value of  $\lambda$ , the cross-validation function  $CV(\lambda)$  simply measures the average square error when each point is predicted using only the remaining points in the sample, so simple cross-validation chooses the value of  $\lambda$  which minimizes  $CV(\lambda)$ . However, the method of generalized cross-validation (GCV) finds  $\lambda$  so as to minimize an estimate of

$$GCV(\lambda) = \sum_{i=1}^n n^{-1} \cdot \text{PMSE}(x_i, \lambda). \quad (3.3)$$

More details on the difference between the two criteria can be found in Green, P.J. and Silverman, B.W. [20].

As quoted from Wegman et al. [46], “Interpolating splines are predicated on nonnoisy data. As such they have limited use in a statistical setting, although in several circumstances they do make an appearance. More to the point, it is desirable in a statistical framework to create a type of smoothing spline that could pass near, in some sense, to the data but not be constrained to interpolate exactly.”

## 3.2 B-splines as smoothing and estimating functions

If we let  $\delta(x)$  denote the Dirac delta function at  $x$ , which is a generalized function that is zero everywhere except at  $x$ , where it has an infinite point mass integrating to unity, then we can represent the “natural” density estimate as  $\sum_{i=1}^n \delta(x_i)/n$ . Although this formulation is not very good for such purposes as estimating the value of the density at a point, what we generally do believe is that regions in which we have observations are likely to have greater density than ones in which no observations are seen. Then it is natural to “smear out” the observed point masses over larger regions which include them, and to base estimates of the density on local

averaging probability mass. Therefore, an approach to density estimation by applying B-splines and smoothing techniques to a basic estimator such as the histogram is developed. Silverman [42] gives an overview of spline smoothing approach to non-parametric regression curve fitting.

In this section, we define the  $k^{th}$  order B-splines as appropriately scaled Order  $(k+1)$  centred differences of the truncated power function. Since B-spline provides bases for certain spline spaces whose spline functions can be obtained by forming linear combinations, this gives rise to the B-spline representation(s) for a smoothing function characterized with shape-preserving or variation diminishing.

### 3.2.1 B-splines

Let  $X = \{x_1, \dots, x_n\}$  be an ordered set of real numbers, hereafter called “knots” equally spaced. Thereby,

$$x_1 < x_2 < x_3, \dots, < x_n \text{ and } h = x_i - x_{i-1} (i = 1, 2, \dots, n).$$

**Definition 3.1** The  $k^{th}$  order truncated power function  $x_+^k$  is denoted by  $x_+^k = \max\{0, x\}^k \equiv (0 \vee x)^k$ , where  $0^0$  is interpreted as 0.

**Definition 3.2** We say  $\delta_h f$  is the centred difference of a function  $f$  at the point  $x \in [x_{i-1}, x_i]$  if  $\delta_h f(x) = f(x + h/2) - f(x - h/2)$ .

Let  $h = 1$ . The  $1^{st}$  degree centred difference of the  $0^{th}$  order truncated power function is referred to as the  $0^{th}$  order B-spline, which is modified at the point  $x = 1/2$ , though.

**Definition 3.3 (Definition of the  $0^{th}$  order B-spline)**

$$B^0(x) = \begin{cases} \delta x_+^0 = (x + 1/2)_+^0 - (x - 1/2)_+^0 & = \begin{cases} 0 & |x| > 1/2 \\ 1 & |x| < 1/2 \\ 1/2 & |x| = 1/2 \end{cases} \end{cases}$$

If  $h \neq 1$ , generally, linear transformation is performed.

$$B^0(x) = \begin{cases} \delta_h \left(\frac{x-t}{h}\right)_+^0 = \left(\frac{x-t}{h} + 1/2\right)_+^0 - \left(\frac{x-t}{h} - 1/2\right)_+^0 & = \begin{cases} 0 & |x-t| > h/2 \\ 1 & |x-t| < h/2 \\ 1/2 & |x-t| = h/2 \end{cases} \end{cases} \quad (3.4)$$

**Definition 3.4 (Definition of the  $k^{th}$  order centred difference)** The  $k^{th}$  order centred difference of a function  $f$  is denoted by  $\delta_h^k f(x) = \delta_h(\delta_h^{k-1} f(x))$ .

**Definition 3.5 (Definition of the  $1^{st}$  order smoothing function)**  $f_1(x)$  is referred to as the  $1^{st}$  order smoothing function of  $f(x)$  if  $f_1(x) = \frac{\delta_h}{h} \int_0^x f(t) dt = \frac{1}{h} \int_{x-h/2}^{x+h/2} f(t) dt$

As noted above,  $f_1(x)$  is the integral of  $f(x)$ , so the 1<sup>st</sup> order smoothing function of  $f(x)$ ,  $f_1(x)$ , is smoother than  $f(x)$ , that is, more differentiable.

Actually as Taylor expansion, if  $f(x)$  is third-order differentiable,

$$\frac{\delta_h f(x)}{h} = \frac{f(x+h/2) - f(x-h/2)}{h} = f'(x) + f'''(\xi)h^2/24$$

$$x - h/2 \leq \xi \leq x + h/2,$$

and henceforth,

$$f_1(x) = \int_0^x \frac{\delta_h}{h} f(t) dt = \int_0^x (f'(t) + f'''(\xi)h^2/24) = f(x) + O(h^2),$$

Therefore, the 1<sup>st</sup> order smoothing function of  $f(x)$  is also approximate to  $f(x)$ .

**Definition 3.6 (Definition of Order  $k$  smoothing function)**  $f_k(x)$  is said to be the  $k^{\text{th}}$  order smoothing function of  $f(x)$  if

$$f_k(x) = \frac{\delta_h}{h} \int_0^x f_{k-1}(t) dt = \frac{1}{h} \int_{x-h/2}^{x+h/2} f_{k-1}(t) dt.$$

Without the loss of the generalization because of the linear transformation  $\frac{x-t}{h}$ , let  $h = 1$ . We will investigate properties of B-splines based on Definition (3.6) of smoothing function and develop estimates of  $S'(q)$  based on smoothing B-splines characterized with shape-preserving or variation diminishing, which is considered in Section (3.3).

### 3.2.2 Smoothing function and B-spline resulting from the $k^{\text{th}}$ order centred difference of a truncated power function

**Definition 3.7 (Definition of Order  $k$  smoothing B-spline)** For any non-negative integer  $k$ , the  $k^{\text{th}}$  order smoothing function of  $B^0(x)$  is referred to as the  $k^{\text{th}}$  order smoothing B-spline, denoted by  $B^k(x)$ .

**Lemma 3.1 (Lemma on order  $K$  smoothing function)** For the  $k^{\text{th}}$  order smoothing B-spline,

$$B^k(x) = \delta^{k+1} \{x_+^k/k!\}.$$

**Proof:** We proceed by induction on  $k$  to establish this.

For  $k = 0$ ,  $B^0(x) = \delta^1 \{x_+^0\}$  follows from Definition (3.3). Consider then the case  $k = 1$ . Based on Definition (3.5) of Order 1 smoothing function,

$$B^1(x) = \int_{x-1/2}^{x+1/2} B^0(t) dt$$

$$\begin{aligned}
&= \int_{x-1/2}^{x+1/2} \{(t+1/2)_+^0 - (t-1/2)_+^0\} dt \\
&= \{(t+1/2)_+ - (t-1/2)_+\} \Big|_{x-1/2}^{x+1/2} \\
&= (x+1)_+ - 2x_+ + (x-1)_+ \\
&= \delta^2 x_+
\end{aligned}$$

For  $k = 1$ , the lemma has been proved. Assume that the lemma has been proved for the  $(k-1)^{th}$  order smoothing function  $B^{k-1}(x)$ :

$$B^{k-1}(x) = \delta^k \{x_+^{k-1}/(k-1)!\}$$

On that hypothesis, we shall prove the lemma for the index  $k$ .

$$\begin{aligned}
B^k(x) &= \int_{x-1/2}^{x+1/2} B^{k-1}(t) dt \\
&= \int_{x-1/2}^{x+1/2} \delta^k \{t_+^{k-1}/(k-1)!\} dt \\
&= \delta^k \left\{ \int_{x-1/2}^{x+1/2} t_+^{k-1}/(k-1)! dt \right\} \\
&= \delta^k \left\{ t_+^k/k! \Big|_{x-1/2}^{x+1/2} \right\} \\
&= \delta^k \{ \delta x_+^k/k! \} = \delta^{k+1} \{x_+^k/k!\}
\end{aligned}$$

Although Lemma (3.1), based on centred differences, allows us to compute  $B^k(x)$ , the following lemma simplify the computation of  $B^k(x)$ . Moreover, most salient properties of our noninterpolatory estimation approach are based on Lemma (3.3).

In order to prove Lemma (3.3), we need to introduce the concepts of forward difference and shift operator.

The definition of forward difference is due to Abramowitz et al. [1].

**Definition 3.8** We say  $\Delta_h f$  is the forward difference of a function  $f$  at the point  $x \in [x_{i-1}, x_i]$  if  $\Delta_h f(x) = f(x+h) - f(x)$ .

**Definition 3.9** We say  $\Xi^h f$  is the shift operator of a function  $f$  at the point  $x \in [x_{i-1}, x_i]$  if  $\Xi^h f(x) = f(x+h)$ , that is,

$$f(x) \xrightarrow{\Xi^h} f(x+h).$$

Henceforth, our centred difference based on Definition (3.2) can be reformulated in terms of the

forward difference and the shift operator as follows.

$$\delta = \Delta \Xi^{-1/2} \quad (3.5)$$

Likewise, we have the same result Lemma (3.2) as equation (3.5) for the  $k^{\text{th}}$  order centred difference based on Definition (3.4).

**Lemma 3.2 (Relation between centred and forward difference)**

$$\delta^{k+1} = \Delta^{k+1} \Xi^{-(k+1)/2}$$

**Proof:** We proceed by induction on  $k$  to establish this.

For  $k = 0$ , the lemma follows from equation (3.5).

Consider then the case  $k = 1$ . Based on Definition (3.4),

$$\begin{aligned} \delta^2 f(x) &= \delta(\delta f(x)) \\ &= \delta f(x + 1/2) - \delta f(x - 1/2) \\ &= \Delta \Xi^{-1/2} f(x + 1/2) - \Delta \Xi^{-1/2} f(x - 1/2) \\ &= \Delta f(x) - \Delta f(x - 1) \\ &= f(x + 1) - 2 \cdot f(x) + f(x - 1) \\ &= \Delta f(x) - \Delta f(x - 1) \\ &= \Delta^2 f(x - 1) \\ &= \Delta^2 \Xi^{-1} f(x) \end{aligned}$$

For  $k = 1$ , the lemma has been proved. Assume that the lemma has been proved for the  $k^{\text{th}}$  order centred difference  $\delta^k$ :

$$\delta^k = \Delta^k \Xi^{-k/2}.$$

On that hypothesis, we shall prove the lemma for the index  $k + 1$ .

$$\begin{aligned} \delta^{k+1} f(x) &= \delta(\delta^k f(x)) \\ &= \delta^k f(x + 1/2) - \delta^k f(x - 1/2) \\ &= \Delta^k \Xi^{-k/2} f(x + 1/2) - \Delta^k \Xi^{-k/2} f(x - 1/2) \\ &= \Delta^k f(x + 1/2 - k/2) - \Delta^k f(x - 1/2 - k/2) \\ &= \Delta^{k+1} f(x - (k + 1)/2) \\ &= \Delta^{k+1} \Xi^{-(k+1)/2} f(x) \end{aligned}$$

The following result for the  $k^{th}$  order forward difference is referred to Abramowitz et al. [1] (Page 877).

$$\Delta^{k+1}f(x) = \sum_{j=0}^{k+1} (-1)^j \binom{k+1}{j} f(x+k+1-j) \quad (3.6)$$

**Lemma 3.3 (Lemma on computing  $B^k(x)$ )**

$$B^k(x) = \sum_{j=0}^{k+1} (-1)^j \binom{k+1}{j} \left(x + \frac{k+1}{2} - j\right)_+^k / k!.$$

**Proof:** Based on Lemma (3.1) and equation (3.6), we have

$$\begin{aligned} B^k(x) &= \delta^{k+1} \{x_+^k / k!\} \\ &= \Delta^{k+1} \Xi^{-(k+1)/2} \{x_+^k / k!\} \\ &= \Delta^{k+1} \left\{ [x - (k+1)/2]_+^k / k! \right\} \\ &= \sum_{j=0}^{k+1} (-1)^j \binom{k+1}{j} \left(x + \frac{k+1}{2} - j\right)_+^k / k! \end{aligned}$$

We know from Lemma (3.3),  $B^k(x)$  is a piecewise degree  $k$  polynomial that belongs to continuity class  $C^{k-1}(\mathfrak{R})$ , the set of functions  $f$  for which the  $(k-1)^{th}$  order derivative of  $f$ ,  $f^{k-1}(x)$ , exists and is continuous throughout the real line. The discontinuities of the  $k^{th}$  order derivative of  $B^k(x)$  are  $x_j = j - \frac{k+1}{2}$ ,  $j = 0, 1, \dots, k+1$ , which are knots of Order  $k$  B-spline,  $B^k(x)$ .

We also infer from Lemma (3.3) that  $B^k(x) = 0$  if  $x \notin \left[-\frac{k+1}{2}, \frac{k+1}{2}\right]$ ; and  $B^k(x) > 0$  if  $x \in \left(-\frac{k+1}{2}, \frac{k+1}{2}\right)$ .

The following lemmas are obvious (cf. Lemma (3.3) and Definition (3.7)).

**Lemma 3.4 (Lemma on support of B-spline  $B^k(x)$ )** For any non-negative integer  $k$ , if  $x \notin \left[-\frac{k+1}{2}, \frac{k+1}{2}\right]$ , then  $B^k(x) = 0$ .

**Lemma 3.5 (Lemma on positivity of B-spline  $B^k(x)$ )** For any non-negative integer  $k$ , if  $x \in \left(-\frac{k+1}{2}, \frac{k+1}{2}\right)$ , then  $B^k(x) > 0$ .

### 3.2.3 Computing and evaluating B-splines

To illustrate our noninterpolatory probability density estimation methodology, we turn to evaluating B-splines introduced in Section (3.2.1) and (3.2.2).

Likewise, as discussed in DeBoor [15], a basis for the linear space of piecewise polynomial functions of degree  $k$  that are globally of class  $C^{k-1}(\mathfrak{R})$  consists of such basis splines or B-splines. This gives rise to our B-splines representation for functions from some class. The graphs of  $B^j(x)$ , for

$j = 0, 1, 2, 3$ , are shown in Figure (3.2).

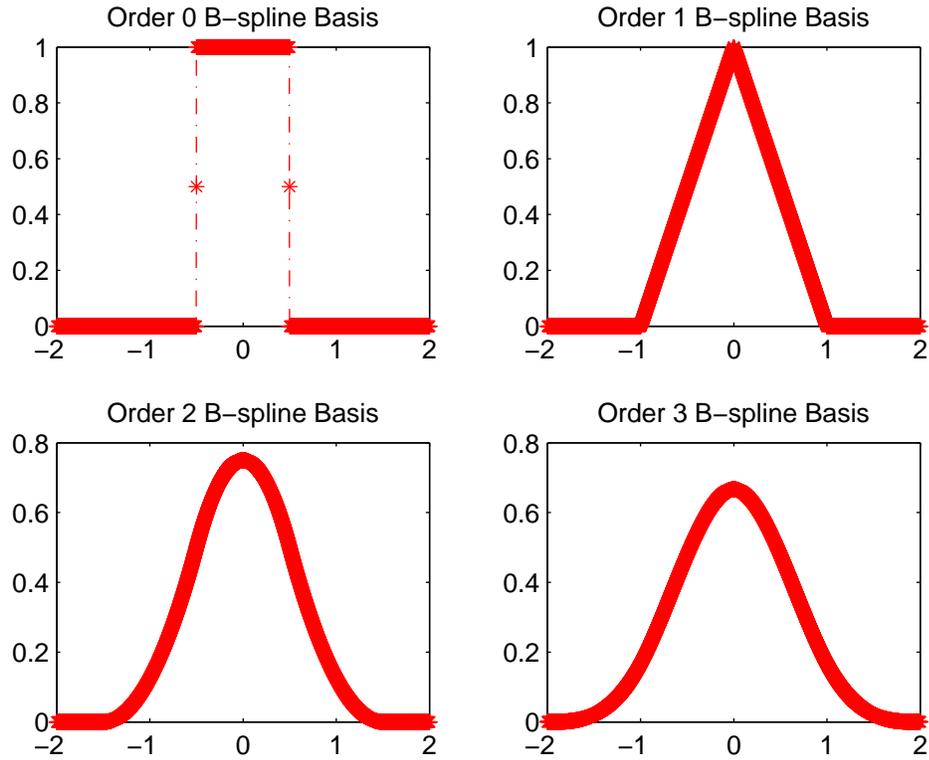


Figure 3.2: Base of different orders of B-splines

$$B^0(x) = \delta x_+^0 = (x + 1/2)_+^0 - (x - 1/2)_+^0 = \begin{cases} 0 & |x| > 1/2 \\ 1 & |x| < 1/2 \\ 1/2 & |x| = 1/2 \end{cases}$$

$$B^1(x) = \delta^2 x_+ = \begin{cases} 0 & |x| \geq 1 \\ 1 - |x| & |x| < 1 \end{cases}$$

$$B^2(x) = \delta^3 \{x_+^2/2!\} = \begin{cases} 0 & |x| \geq 3/2 \\ -x^2 + \frac{3}{4} & |x| < 1/2 \\ \frac{1}{2}x^2 - \frac{3}{2}|x| + \frac{9}{8} & 1/2 \leq |x| \leq 3/2 \end{cases}$$

$$B^3(x) = \delta^4 \{x_+^3/3!\} = \begin{cases} 0 & |x| \geq 2 \\ \frac{1}{2}|x|^3 - x^2 + \frac{3}{2} & |x| \leq 1 \\ \frac{1}{6}|x|^3 + x^2 - 2|x| + \frac{4}{3} & 1 \leq |x| \leq 2 \end{cases}$$

Now, in a sequence of lemmas, we shall develop the important properties of the B-spline family

$\{B^k\}$ .

**Lemma 3.6 (Lemma on derivative of B-splines)** *The derivatives of B-spline functions are calculated as follows, for  $k \geq 1$ :  $(B^k(x))' = B^{k-1}(x + 1/2) - B^{k-1}(x - 1/2)$ .*

Based on Definition (3.6) of Order  $k$  smoothing function and Definition of Order  $k$  B-spline, that is,

$$B^k(x) = \int_{x-1/2}^{x+1/2} B^{k-1}(t) dt$$

the lemma is obviously correct.

**Lemma 3.7 (Lemma on B-splines as smoothing functions)** *Let  $f_k(x)$  be Order  $k$  smoothing function of  $f(x)$ . Therefore,*

$$f_k(x) = (1/h) \cdot \int_{-\infty}^{\infty} B^{k-1}((x-t)/h) f(t) dt$$

**Proof:** We proceed by induction on  $k$  to establish this.

Consider the case  $k = 1$ . From equation (3.4):

$$B^0\left(\frac{x-t}{h}\right) = \begin{cases} 0 & |x-t| > h/2 \\ 1 & |x-t| < h/2 \\ 1/2 & |x-t| = h/2 \end{cases}$$

Based on Definition (3.5) of order 1 smoothing function,

$$\begin{aligned} f_1(x) &= (1/h) \cdot \int_{x-h/2}^{x+h/2} f(t) dt \\ &= (1/h) \cdot \int_{x-h/2}^{x+h/2} B^0\left(\frac{x-t}{h}\right) f(t) dt \\ &= (1/h) \cdot \int_{-\infty}^{\infty} B^0\left(\frac{x-t}{h}\right) f(t) dt \end{aligned}$$

The lemma is obviously correct for  $k = 1$ . Assume that the lemma has been proved for the  $(k-1)^{th}$  order smoothing function  $f_{k-1}(x)$ , that is,

$$f_{k-1}(x) = (1/h) \cdot \int_{-\infty}^{\infty} B^{k-2}\left(\frac{x-t}{h}\right) f(t) dt$$

On the basis of this assumption, we shall prove the lemma for the index  $k$ .

$$f_k(x) = (1/h) \cdot \int_{-\infty}^{\infty} B^0\left(\frac{x-\tau}{h}\right) f_{k-1}(\tau) d\tau$$

$$\begin{aligned}
&= (1/h^2) \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} B^0\left(\frac{x-\tau}{h}\right) B^{k-2}\left(\frac{\tau-t}{h}\right) f(t) dt d\tau \\
&= (1/h^2) \cdot \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} B^0\left(\frac{x-\tau}{h}\right) B^{k-2}\left(\frac{\tau-t}{h}\right) d\tau \right\} f(t) dt
\end{aligned}$$

Let  $\frac{x-\tau}{h} = \xi$ .

$$\begin{aligned}
&(1/h) \cdot \int_{-\infty}^{\infty} B^0\left(\frac{x-\tau}{h}\right) B^{k-2}\left(\frac{\tau-t}{h}\right) d\tau \\
&= \int_{-\infty}^{\infty} B^0(\xi) B^{k-2}\left(\frac{x-t}{h} - \xi\right) d\xi \\
&= \int_{-1/2}^{1/2} B^{k-2}\left(\frac{x-t}{h} - \xi\right) d\xi \\
&= \int_{\frac{x-t}{h} - 1/2}^{\frac{x-t}{h} + 1/2} B^{k-2}(t) dt \\
&= B^{k-1}\left(\frac{x-t}{h}\right)
\end{aligned}$$

Hence,

$$f_k(x) = (1/h) \cdot \int_{-\infty}^{\infty} B^{k-1}\left(\frac{x-t}{h}\right) f(t) dt.$$

Because Order  $k$  smoothing function of  $f(x) \equiv 1$  ( $-\infty < x < \infty$ ) is always 1, Lemma (3.8) on integral of B-splines follows based on Lemma (3.7).

**Lemma 3.8 (Lemma on integral of B-spline)**

$$(1/h) \cdot \int_{-\infty}^{\infty} B^k((x-t)/h) dt = 1$$

If  $h = 1$ , equivalently,

$$\int_{-\infty}^{\infty} B^k(x) dx = \int_{-\frac{k+1}{2}}^{\frac{k+1}{2}} B^k(x) dx = 1$$

$$k = 0, 1, 2, \dots$$

Lemma (3.8) means that the area between the curve  $B^k(x)$  and Axis X is 1.

Moreover,  $B^k(x)$  is a probability density function.

**Lemma 3.9 (Lemma on partition of unity for B-splines)** For all  $k$ , we have

$$\sum_{i=-\frac{k+1}{2}}^{\frac{k+1}{2}} B^k((x-x_i)/h) = 1$$

This proof is referred to DeBoor [15] (Page 110).

From a probabilistic points of view,  $B^k((x-x_i)/h)$  is a conditional probability density func-

tion given  $x$ .

### 3.3 The estimates of $S'(q)$ based on the smoothing B-splines

Based on B-spline elegant properties previously discussed, we introduce a noninterpolatory estimation methodology. Given a function  $f(x)$  and a set of sampling data points  $(x_i, y_i = f(x_i))$ ,  $h_i = x_{i+1} - x_i$ ,  $i = 1, 2, \dots, n$ , we define a spline function by the equation:

$$S_f(x) = \sum_{i=1}^n w_i B_i^k((x - x_i)/h_i) \quad (3.7)$$

where a sequence of parameters  $\{w_i\}(i = 1, 2, \dots, n)$  are weighting parameters which are considered shortly, and each  $B_i^k(x)$  is a  $k^{\text{th}}$ -order B-spline ( $k = 0, 1, 2, \dots$ ). The set  $\{B_i^k\}$  depends only on the ordered set of knots  $X = \{x_1, \dots, x_n\}$  equally spaced and the order  $k$ . Each  $B_i^k$  has limited support from Lemma (3.3); in particular,  $B_i^k(x) \neq 0$  only if  $x \in (-\frac{k+1}{2}, \frac{k+1}{2})$ , corresponding the interval  $(x_i, x_{i+k})$  by linear transformation  $\frac{x - x_i}{h_i}$ . As a consequence, the value of  $S_f(x)$  at a point depends on at most  $k + 1$  nonzero  $B_i^k$ , which are also nonnegative and sum to one at each  $x$ . Therefore, we can consider  $S_f(x)$  as an expectation. Silverman [42] chooses parameters  $\{w_i\}(i = 1, 2, \dots, n)$  from a finite-dimensional Bayesian formulation of the curve estimation problem.

If the errors  $\{y_i - f(x_i)\}(i = 1, 2, \dots, n)$  are normally distributed and the data are identically independently normally distributed with mean  $f(x_i)$  and equal variance, then the first term on the right side of equation (3.1),  $\sum_{i=1}^n w_i (y_i - f(x_i))^2$ , is the negative log-likelihood function for the parameter  $f$ , that is,

$$l(f) = -\frac{1}{2\sigma^2} \sum_{i=1}^n w_i (y_i - f(x_i))^2.$$

The second term on the right side of equation (3.1),  $\int_a^b (f''(x))^2 dx$ , then can be viewed as the negative logarithm of a prior density for  $f$ . Therefore the penalized likelihood is equal to

$$l_{\text{post}}(f) = -\frac{1}{2\sigma^2} \sum_{i=1}^n w_i (y_i - f(x_i))^2 - \frac{1}{2} \lambda \int_a^b (f''(x))^2 dx,$$

so that the minimizer of equation (3.1) can be thought of as a posterior mean value for the regression function. We interpret this prior entirely on the space of spline curves with knots on the data points. Hence, each possible  $S_f(x)$  can be written as a linear combination of B-splines,

$$S_f(x) = \sum_{i=1}^n w_i B_i^k((x - x_i)/h_i).$$

Silverman [42] shows that how this prior distribution can be thought of as a prior multivariate

normal distribution on the coefficients  $w_i$ . Any interested user should examine Silverman [42], which distills a great amount of theoretical knowledge and practical advice.

However, our methodology concentrates these parameters entirely on the space of spline curves with knots on the data points. Suppose we are estimating a probability density function  $f(x)$  on a finite interval  $(a, b)$ , given independent observations  $x_1, \dots, x_n$  from  $f(x)$  and then we choose  $f(x_i)$   $i = 1, 2, \dots, n$ , the frequencies of the independent observations  $x_1, \dots, x_n$  as a sequence of weighted parameters.

$$S_f(x) = \sum_{i=1}^n f(x_i) B_i^k((x - x_i)/h_i) = \sum_{i=1}^n y_i B_i^k((x - x_i)/h_i) \quad (3.8)$$

The  $0^{th}$ -order smoothing function of  $f(x)$  is a step function as follows, which passes through all of the data points:

$$f_0(x) = \sum_{i=1}^n y_i B^0((x - x_i)/h_i)$$

Based on Definition (3.5) of Order 1 smoothing function and Lemma (3.1) on Order  $k$  smoothing function,

$$B^1((x - x_i)/h_i) = (1/h_i) \cdot \int_{x-h_i/2}^{x+h_i/2} B^0((t - x_i)/h_i) dt$$

we can infer that the  $1^{st}$ -order smoothing function of  $f(x)$  is

$$f_1(x) = \sum_{i=1}^n y_i B^1((x - x_i)/h_i)$$

which is also the interpolation scheme, that is, this piecewise linear function  $f_1(x)$  passes through all of the data points. Therefore, the  $k^{th}$ -order smoothing function of  $f(x)$  is just equation (3.8) as follows:

$$f_k(x) = \sum_{i=1}^n y_i B^k((x - x_i)/h_i) \quad (3.9)$$

As well, we can conclude that the  $k^{th}$ -order smoothing function of  $f(x)$  is the same as the  $(k-1)^{th}$ -order smoothing function of  $f_1(x)$

Moreover, the salient properties of our noninterpolatory estimation approach are as follows.

When  $k = 0$  or  $k = 1$ , equation (3.8) is the interpolation scheme. For  $k > 1$ , however, the spline function  $S_f(x)$  defined by equation (3.8) does not interpolate any prescribed set of nodes. The properties of this estimation scheme are these:

- If  $f(x)$  is a linear function, then  $S_f(x) = f(x)$ .
- For any linear function  $l(x)$ ,  $S_f - l$  has no more variations in sign than  $f - l$ . In another words, the spline estimation  $S_f$  crosses any particular straight line at most as many times as does  $f$  itself. This proof would take us far afield.

- If  $f \geq 0$ , then  $S_f \geq 0$ .
- If  $|f| \leq M$ , then  $|S_f| \leq M$ .
- $S_f$  is a linear operator, that is,  $S(\alpha f + \beta g) = \alpha S(f) + \beta S(g)$ .

It is easy to prove that the histogram is the linear combination of the  $0^{th}$ -order B-splines with the knots that are the bins' edge points of the histogram, provided that there are no duplicates among the edge points. The coefficients of the linear combination are the frequencies among the bins of the histogram. This is also shown in Figure (3.3). From Figure (3.3), we can see that the  $0^{th}$ -

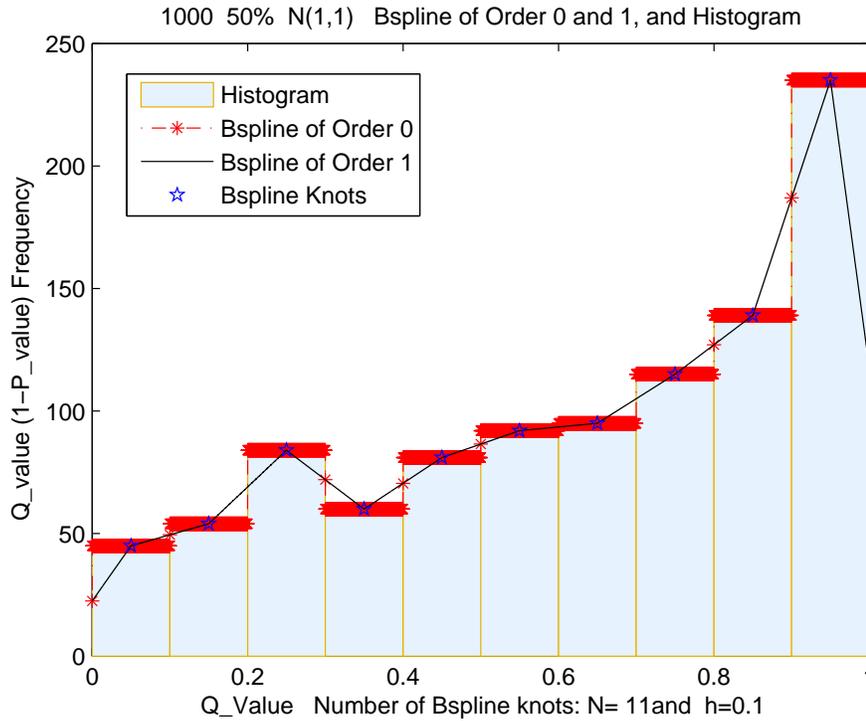


Figure 3.3: Histogram and smoothing B-splines of Order  $0^{th}$  and  $1^{st}$

order smoothing B-spline, the estimated histogram, is just the histogram, which is a step-function and discontinuous at the spline knots  $i \cdot h$  ( $i = 0, 1, 2, \dots, n - 1$ ;  $h = \frac{x_n - x_1}{n - 1}$ , bandwidth); the  $1^{st}$ -order smoothing B-spline belongs to continuity class  $C(\mathfrak{R})$  that means the smoothing function itself is continuous but not everywhere differentiable. The first degree derivative of the  $1^{st}$ -order smoothing B-spline, step-function, is discontinuous at the spline knots  $(i + 1/2) \cdot h$  ( $i = 0, 1, 2, \dots, n - 1$ ;  $h = \frac{x_n - x_1}{n - 1}$ , bandwidth). However, both the  $0^{th}$ -order and the  $1^{st}$ -order smoothing B-splines are not smooth enough for us to estimate the  $P$ -value probability density function  $S'(1 - p)$ , so we will introduce higher order B-splines as estimates. In particular, the properties of the  $2^{nd}$ -order and  $3^{rd}$ -order smoothing B-splines are investigated.

### Theorem 3.1 (Approximation, shape-preserving, and slope of Order 2 B-splines)

Consider an unknown function  $f(x)$  and data points  $(x_i, y_i)$   $i = 1, 2, \dots, n$  from  $f(x)$  with equally spaced nodes  $\{x_i\}$   $i = 1, 2, \dots, n$ . Let  $x_{i+1/2} = (x_i + x_{i+1})/2$ . For the  $2^{nd}$ -order smoothing function  $f_2(x)$ ,

$$f_2(x_i) - y_i = (1/8) \cdot (y_{i+1} - 2y_i + y_{i-1}) \text{ (approximation accuracy),} \quad (3.10)$$

$$f_2''(x_i) = (y_{i+1} - 2y_i + y_{i-1})/h^2 \text{ (shape-preserving),} \quad (3.11)$$

$$f_2'(x_{i+1/2}) = (y_{i+1} - y_i)/h \text{ (slope),} \quad (3.12)$$

$$f_2(x_{i+1/2}) = (1/2) \cdot (y_i + y_{i+1}), \quad \text{for } i = 1, 2, \dots, n. \quad (3.13)$$

**Proof:** Recall that  $f_2(x) = \sum_{i=1}^n y_i B^2((x - x_i)/h_i)$  from equation (3.9), where  $h_i = x_{i+1} - x_i$ . Then compute  $f_2(x)$  at the points  $x_i$  and  $x_{i+1/2}$ :

$$\begin{aligned} f_2(x_i) &= \sum_{j=1}^n y_j B^2((x_i - x_j)/h) \\ &= \sum_{j=1}^n y_j B^2(i - j), \text{ by Lemma (3.4)} \\ &= y_{i-1} B^2(1) + y_i B^2(0) + y_{i+1} B^2(-1) \\ &= y_i + (1/8) \cdot (y_{i+1} - 2y_i + y_{i-1}) \end{aligned}$$

$$\begin{aligned} f_2(x_{i+1/2}) &= \sum_{j=1}^n y_j B^2(i - j + 1/2), \text{ by Lemma (3.4)} \\ &= y_i B^2(1/2) + y_{i+1} B^2(-1/2) \\ &= 1/2(y_i + y_{i+1}) \end{aligned}$$

$$\begin{aligned} f_2'(x_{i+1/2}) &= (y_i B'^2(1/2) + y_{i+1} B'^2(-1/2))/h \\ &= (y_{i+1} - y_i)/h \\ f_2''(x_i) &= y_{i-1} B''^2(1) + y_i B''^2(0) + y_{i+1} B''^2(-1) \\ &= (y_{i+1} - 2y_i + y_{i-1})/h^2 \end{aligned}$$

In term of the geometric interpretation of Theorem (3.1), the shape of the estimated probability density function is dominated by a set of frequencies of sampling data points  $(x_i, y_i = f(x_i))$   $i = 1, 2, \dots, n$ . An estimated function  $f_2(x)$  is convex if  $f_2''(x_i) = (y_{i+1} - 2y_i + y_{i-1})/h^2 > 0$ , concave if  $f_2''(x_i) = (y_{i+1} - 2y_i + y_{i-1})/h^2 < 0$ , and  $f_2(x)$  is linear if  $f_2''(x_i) = (y_{i+1} - 2y_i + y_{i-1})/h^2 = 0$  when  $x \in (x_{i-1}, x_{i+1})$ . This is also shown in Figure (3.4).

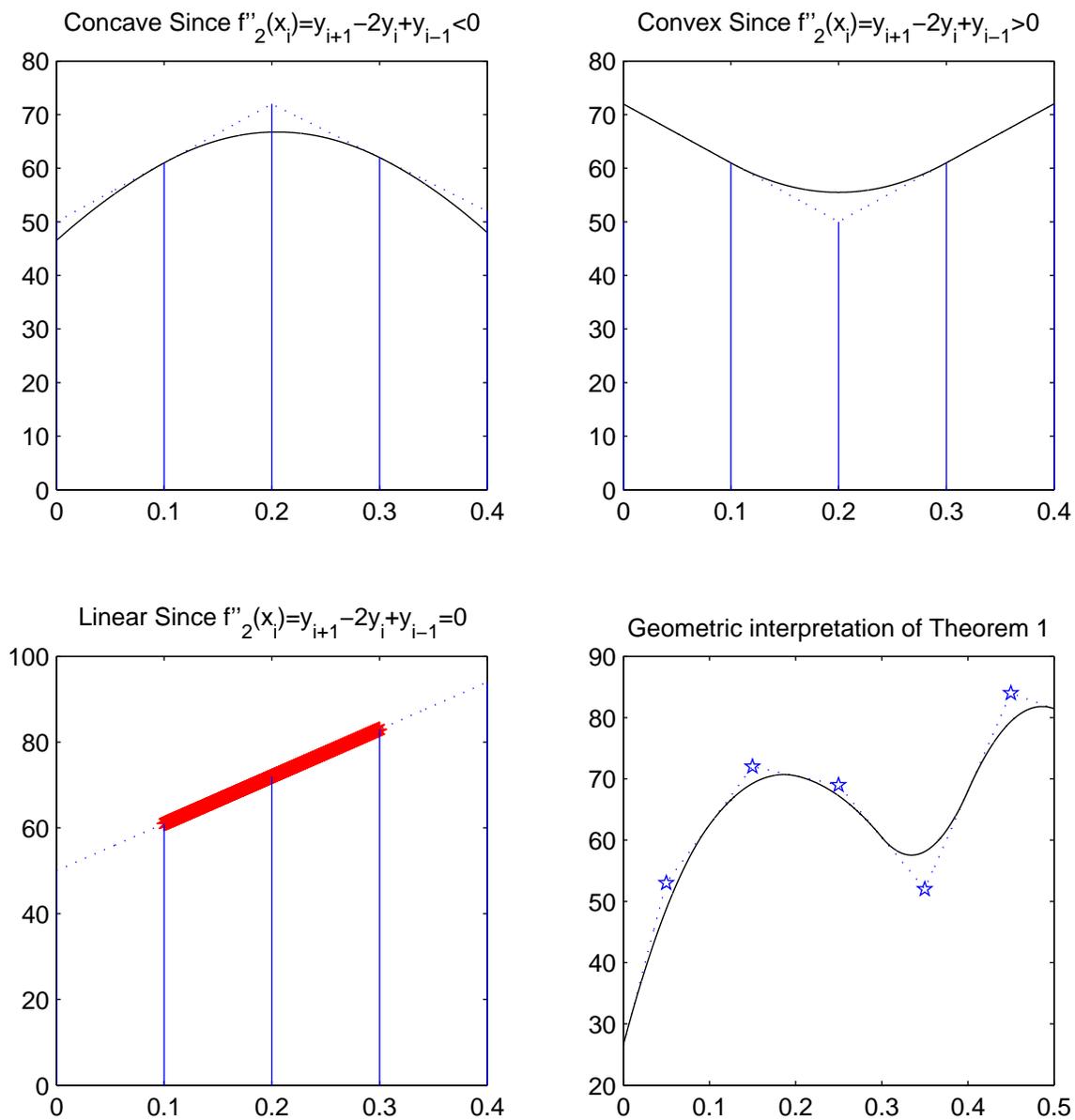


Figure 3.4: Noninterpolatory, shape-preserving, and slope of Order 2 B-spline estimation

**Reducing the approximation error by sharpening data points:** Actually if  $f(x) \in C^2(\mathfrak{R})$ , then using the Taylor expansion

$$f_2(x_i) = y_i + (1/8) \cdot (y_{i+1} - 2y_i + y_{i-1}) = y_i + h^2/8 f''(\xi),$$

$$x_{i-1} < \xi < x_{i+1}.$$

Hence, the useful estimate for the error in the approximation is established. The error can be reduced in density estimation by sharpening the data points as the following equation:

$$\hat{y}_i = y_i - (1/8) \cdot (y_{i+1} - 2y_i + y_{i-1}), \quad i = 1, 2, \dots, n. \quad (3.14)$$

When the data points  $(x_i, y_i = f(x_i))$   $i = 1, 2, \dots, n$  are sharpened from equation (3.14), we probably use additional knots such as  $(x_0, y_0)$  and  $(x_{n+1}, y_{n+1})$  for sharpening  $(x_1, y_1)$  and  $(x_n, y_n)$ . Based on Theorem 3.1 (3.10), if we need our density estimator to pass through the two end-points  $(x_1, y_1)$  and  $(x_n, y_n)$ , the additional knots can be extrapolated by the following equation, equation (3.15).

$$y_0 = 2y_1 - y_2, \quad \text{and} \quad y_{n+1} = 2y_n - y_{n-1}. \quad (3.15)$$

Applying the sharpened data points to the new  $2^{nd}$ -order smoothing function  $\hat{f}_2(x)$ , we have from Theorem (3.1)

$$\hat{f}_2(x_i) = \hat{y}_i + (1/8) \cdot (\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}).$$

By plugging in equation (3.14), we have

$$\hat{f}_2(x_i) = y_i - (h^4/64) \cdot f^{(4)}(\xi) = y_i + O(h^4),$$

$$x_{i-2} < \xi < x_{i+2}, i = 1, 2, \dots, n$$

Therefore, the error in the approximation with the sharpened data points is reduced from  $O(h^2)$  to  $O(h^4)$ .

Similarly, the properties of the  $3^{rd}$ -order smoothing B-splines can be inferred. For the  $3^{rd}$ -order smoothing function,  $f_3(x) = \sum_{i=1}^n y_i B^3((x - x_i)/h_i)$  from equation (3.9).

### **Theorem 3.2 (Approximation, shape-preserving, and slope of Order 3 B-splines)**

Consider an unknown function  $f(x)$  and data points  $(x_i, y_i)$   $i = 1, 2, \dots, n$  from  $f(x)$  with equally spaced nodes  $\{x_i\}$   $i = 1, 2, \dots, n$ . Let  $x_{i+1/2} = (x_i + x_{i+1})/2$ . For the  $3^{rd}$ -order smoothing

function  $f_3(x)$ ,

$$f_3(x_i) - y_i = (1/6) \cdot (y_{i+1} - 2y_i + y_{i-1}) \text{ (approximation accuracy),} \quad (3.16)$$

$$f_3''(x_i) = (y_{i+1} - 2y_i + y_{i-1})/h^2 \text{ (shape-preserving),} \quad (3.17)$$

$$f_3'(x_{i+1/2}) = (y_{i+1} - y_i)/h + (1/(8h)) \cdot (y_{i+2} - 3y_{i+1} + 3y_i - y_{i-1}) \text{ (slope),} \quad (3.18)$$

$$f_3(x_{i+1/2}) = (1/2) \cdot (y_i + y_{i+1}) + (1/48) \cdot (y_{i-1} + y_{i+2} - y_i - y_{i+1}), \text{ for } i = 1, 2, \dots, n. \quad (3.19)$$

In order to fit the new  $3^{rd}$ -order smoothing function, we sharpen the data points by equation (3.20):

$$\hat{y}_i = y_i - (1/6) \cdot (y_{i+1} - 2y_i + y_{i-1}), \quad i = 1, 2, \dots, n. \quad (3.20)$$

As well, applying the sharpened data points to the new  $3^{rd}$ -order smoothing function  $\hat{f}_3(x)$ , we have from Theorem (3.2)

$$\hat{f}_3(x) = \hat{y}_i + (1/6) \cdot (\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}).$$

By plugging in equation (3.20), we have

$$\hat{f}_3(x) = y_i + O(h^4), i = 1, 2, \dots, n.$$

Therefore, the error in the approximation with the sharpened data points is also reduced from  $O(h^2)$  to  $O(h^4)$ . There are many other ways to sharpen data (cf. Sheather [36])

From Theorem (3.1), Theorem (3.2), and the Weierstrass Approximation Theorem, we investigate whether continuous functions can be approximated to any desired precision by increasing the number of knots while the order  $k$  is being fixed.

From equation (3.9),  $f_k(x)$ , the  $k^{th}$ -order smoothing function of  $f(x)$ , reproduces constant  $C$ , that is,  $f_k(x) = f(x)$  in case the function  $f(x) = C$  for all  $x$  in  $[x_1, x_n]$ . This is so because, by Lemma (3.9), B-splines sum up to one. This property of B-splines, together with the fact that B-splines are non-negative and have small support, by Lemma (3.5) and Lemma (3.4), makes it easy to establish an estimate for the error in the noninterpolatory approximation scheme equation (3.9) as follows.

The following definition is due to DeBoor [15], which is about the rate at which

$$\max_{x_1 \leq x \leq x_n} |f(x) - f_k(x)| \rightarrow 0^+ \text{ as } h \rightarrow 0^+.$$

**As usual, a set of knots is prescribed:**

$$a = x_1 < x_2 < x_3 < \dots < x_n = b$$

**Definition 3.10 (Definition of modulus of continuity of function  $f$ )** Whether  $f$  is continuous or not on  $[a, b]$ , the modulus of continuity of  $f$  at  $h$  is denoted by  $\omega(f; h)$  and is defined by

$$\omega(f; h) \equiv \max_{|s-t| \leq h} |f(s) - f(t)|$$

It is obvious that  $\omega(f; h)$  is monotone in  $h$  and subadditive in  $h$ , that is,

$$\omega(f; h_1) \leq \omega(f; h_1 + h_2) \leq \omega(f; h_1) + \omega(f; h_2)$$

for nonnegative  $h_1$  and  $h_2$ . As well, conclude that  $\omega(f; ch) \leq c\omega(f; h)$  for nonnegative  $c \leq n - 1$ .

The following theorem is based on Kincaid's theorem on spline function approximation [28]. The constant covered by the approximation error bound ( $k\omega(f; h)$ ) on Kincaid's spline function is  $k$ , the order of B-spline, with the estimating function

$$f_k(x) = \sum_{i=1}^n f(x_{i+2}) \cdot B_i^k((x - x_i)/h_i).$$

However, the constant covered by the approximation error bound ( $\frac{k+1}{2}\omega(f; h)$ ) on our theorem as follows is  $\frac{k+1}{2}$  with the estimating function as equation (3.9), that is,

$$f_k(x) = \sum_{i=1}^n f(x_i) \cdot B_i^k((x - x_i)/h_i).$$

**Theorem 3.3 (Theorem on spline function approximation)** If  $f(x)$  is a function on  $[a, b]$ , then  $f_k(x)$  satisfies

$$\max_{a \leq x \leq b} |f(x) - f_k(x)| \leq \frac{k+1}{2}\omega(f; h),$$

where  $h = \max_{-[\frac{k+1}{2}]+1 \leq i \leq n+[\frac{k+1}{2}]} |x_i - x_{i-1}| = \max_{-[\frac{k+1}{2}]+1 \leq i \leq n+[\frac{k+1}{2}]} h_i$ , and  $[\frac{k+1}{2}]$  gives the integer part of  $\frac{k+1}{2}$ .

**Proof:** Take a point  $\hat{x}$  in some interval  $[x_j, x_{j+1}] \subseteq [a, b]$ . Then from equation (3.9),

$$f_k(\hat{x}) = \sum_{i=j-[\frac{k+1}{2}]+1}^{j+[\frac{k+1}{2}]} f(x_i) \cdot B^k((\hat{x} - x_i)/h),$$

while based on Lemma (3.9),

$$f(\hat{x}) = f(\hat{x}) \cdot \sum_{i=j-[\frac{k+1}{2}]+1}^{j+[\frac{k+1}{2}]} B^k((\hat{x} - x_i)/h) = \sum_{i=j-[\frac{k+1}{2}]+1}^{j+[\frac{k+1}{2}]} f(\hat{x}) \cdot B^k((\hat{x} - x_i)/h).$$

Therefore,

$$f(\hat{x}) - f_k(\hat{x}) = \sum_{i=j-\lfloor \frac{k+1}{2} \rfloor + 1}^{j+\lfloor \frac{k+1}{2} \rfloor} \{f(\hat{x}) - f(x_i)\} \cdot B^k((\hat{x} - x_i)/h).$$

Taking absolute values on both sides, and using Lemma (3.5) on positivity and Lemma (3.9) on partition of unity of B-spline, we have

$$|f(\hat{x}) - f_k(\hat{x})| \leq \sum_{i=j-\lfloor \frac{k+1}{2} \rfloor + 1}^{j+\lfloor \frac{k+1}{2} \rfloor} |f(\hat{x}) - f(x_i)| \cdot B^k((\hat{x} - x_i)/h)$$

$$\leq \max_{j-\lfloor \frac{k+1}{2} \rfloor + 1 \leq i \leq j+\lfloor \frac{k+1}{2} \rfloor + 1} |f(\hat{x}) - f(x_i)|.$$

For  $i$  in the range  $j - \lfloor \frac{k+1}{2} \rfloor + 1 \leq i \leq j + \lfloor \frac{k+1}{2} \rfloor$ , we have

$$|\hat{x} - x_i| \leq \frac{k+1}{2}h.$$

By the definition of modulus of continuity  $\omega(f; h)$  and the monotonicity and subadditivity of  $\omega(f; h)$ , we conclude that

$$\max_{a \leq x \leq b} |f(x) - f_k(x)| \leq \frac{k+1}{2} \omega(f; h).$$

If  $f(x)$  is continuous, then

$$\lim_{\delta \downarrow 0} \max_{|s-t| \leq \delta} |f(s) - f(t)| = 0.$$

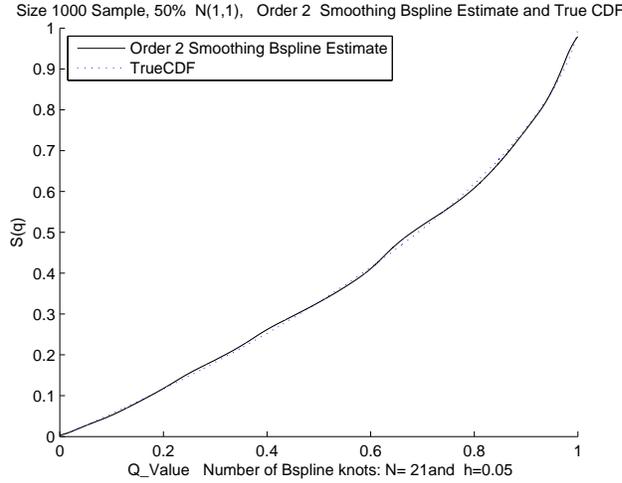
Therefore,

$$\lim_{\delta \downarrow 0} \max_{x_1 \leq x \leq x_n} |f(x) - f_k(x)| = 0$$

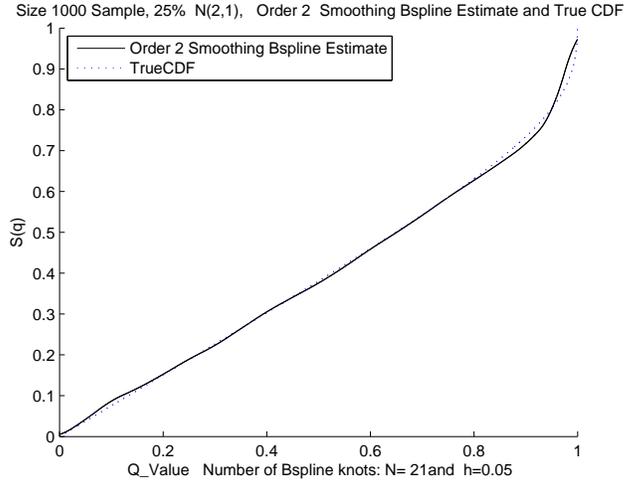
Hence, as the density of the knots is increased, continuous functions can be approximated to arbitrary precision by the  $k^{th}$ -order smoothing function  $f_k(x)$ .

In summary, our noninterpolatory estimation  $S_f$ , that is,  $f_k$  from equation (3.9) maps probability density functions to probability density functions and convex or concave functions to convex or concave functions. Also,  $f_k$  provides a local estimation to a probability density function  $f$ . The function  $f_k(x)$  on the interval  $[x_i, x_{i+1}]$  depends only on the values of a set of frequencies of sampling data points  $(x_i, y_i = \int_{x_{i-1/2}}^{x_{i+1/2}} f(t)dt = S(x_{i+1/2}) - S(x_{i-1/2}))$ ,  $i = 1, 2, \dots, n$ , from an unknown population with a probability density function  $f(x)$  at the  $k+1$  “nearby” data points  $x_{i-k}, \dots, x_i$ , where  $S(x)$  is an empirical cumulative distribution function,  $x_{i-1/2} = x_i - h/2$  and  $x_{i+1/2} = x_i + h/2$ . In particular, if these data points  $(x_{i-k+j}, y_{i-k+j})$ ,  $j = 0, 1, \dots, k$ , lies on a straight line, then  $f_k$  on the interval  $[x_i, x_{i+1}]$  coincides with the same straight line.

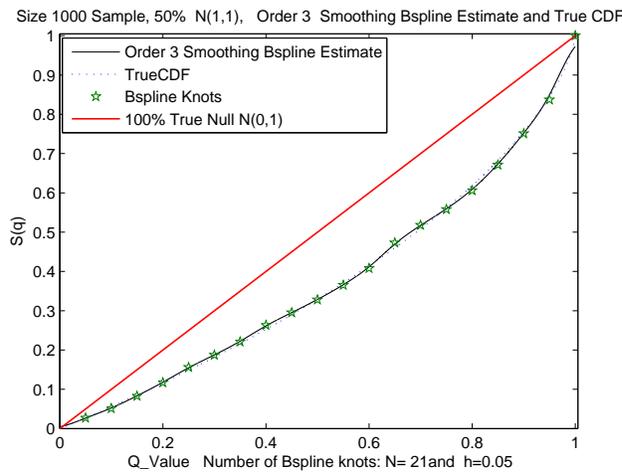
The various estimations of  $P$ -value cumulative distribution function  $S(1-P)$  and probability density function  $S'(1-P)$ , and then the posterior probability density function of null hypothesis,  $S'(0)/S'(Q)$ ,  $Q = 1 - P$ , based on our noninterpolatory approximation are illustrated in the following figures from Figure (3.5) to Figure (3.10).



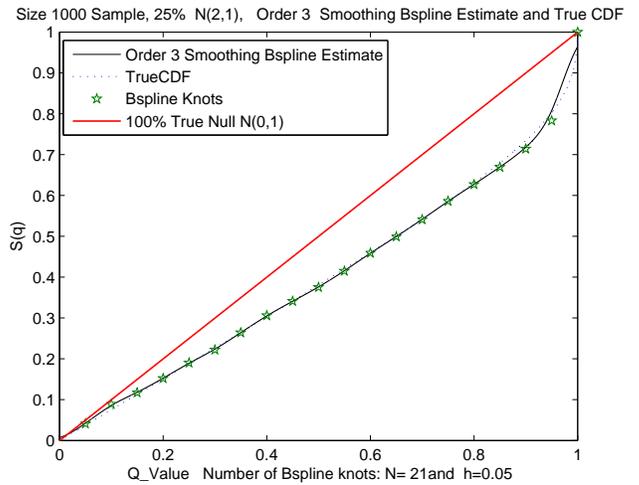
(a) 50% $N(1, 1)$  Order 2 B spline estimator



(b) 25% $N(2, 1)$  Order 2 B spline estimator

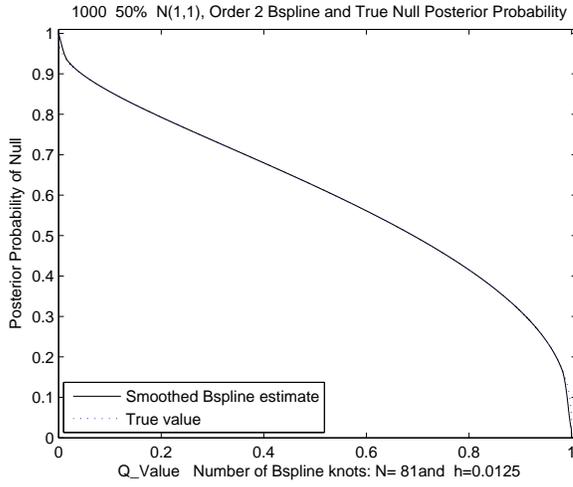


(c) 50% $N(1, 1)$  Order 3 B spline estimator

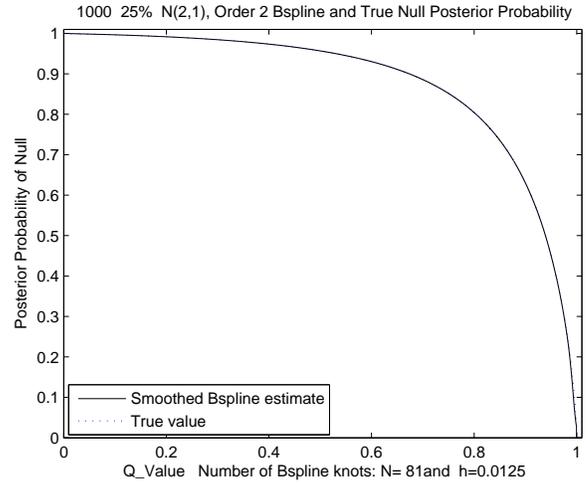


(d) 25% $N(2, 1)$  Order 3 B spline estimator

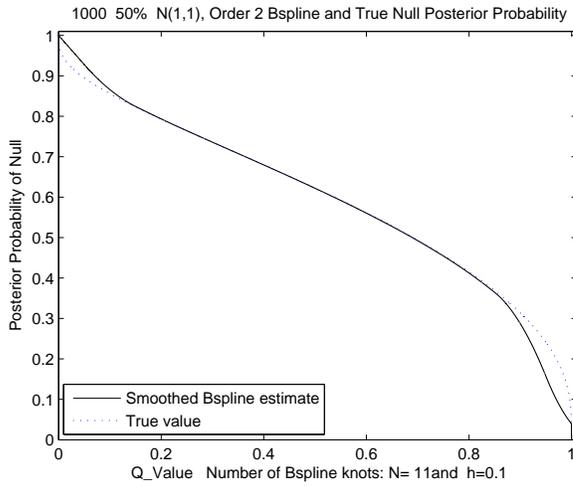
Figure 3.5: Estimated cumulative distribution (CDF) fit to the ECDF for  $Q = 1 - P$  sampled from the mixed population with the sample size  $n = 1000$ , of which in the left panel 500 are simulated from  $N(0, 1)$  and 500 are simulated from  $N(1, 1)$ ; in the right panel 750 are from  $N(0, 1)$  and 250 are from  $N(2, 1)$ . The two graphs in the upper panel show the estimated CDF (solid lines) compared with the true CDF (dotted lines) of  $Q$ -values using Order 2 Bsplines; and the two graphs in the lower panel show the estimated CDF (solid lines) compared with the true CDF (dotted lines) of  $Q$ -values using Order 3 Bsplines with scatterplots as B spline knots, respectively.



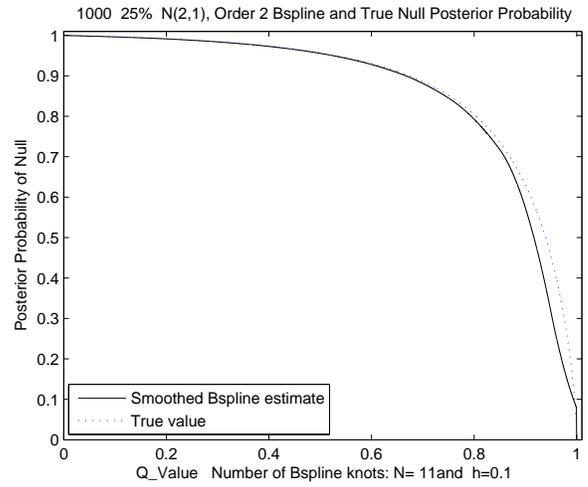
(a) 50% $N(1, 1)$  Bspline knots  $N = 81$



(b) 25% $N(2, 1)$  Bspline knots  $N = 81$



(c) 50% $N(1, 1)$  Bspline knots  $N = 11$



(d) 25% $N(2, 1)$  Bspline knots  $N = 11$

Figure 3.6: Order 2 Bspline estimators for the posterior probability of null hypothesis given  $Q = 1 - P$  sampled from the mixed population with the sample size  $n = 1000$ , of which in the left panel 500 are simulated from  $N(0, 1)$  and 500 are simulated from  $N(1, 1)$ ; in the right panel 750 are from  $N(0, 1)$  and 250 are from  $N(2, 1)$ . The two graphs in the upper panel show the estimated posterior probability of null hypothesis with the number of Bspline knots, 81 (solid lines) compared with the true values (dotted lines); and the two graphs in the lower panel show the estimated posterior probability of null hypothesis with the number of Bspline knots, 11, respectively. Compared with the true values, the estimators with more knots in the upper panel are more accurate than those in the lower panel. Hence, the precision of the estimators can be improved by increasing the number of knots as described in Theorem (3.3).

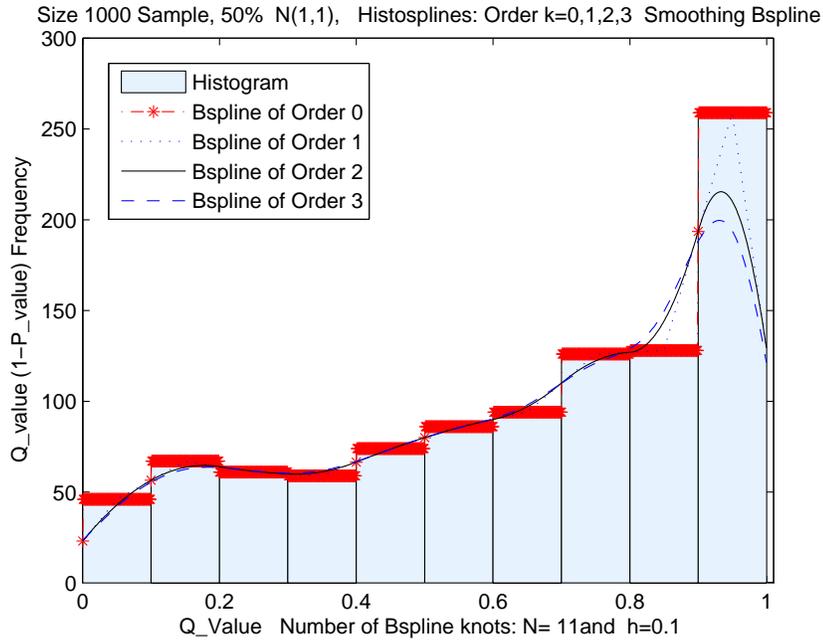


Figure 3.7: Histospines, the estimated histograms of  $Q = 1 - P$  sampled from the mixed population with the sample size  $n = 1000$ , of which 500 are simulated from  $N(0, 1)$  and 500 are simulated from  $N(1, 1)$ , using Order  $k=0,1,2,3$  smoothing B-splines based on equation (3.8). As demonstrated, the histogram is the linear combination of the  $0^{th}$ -order B-splines with the knots that are the bins' edge points of the histogram, and the coefficients of the linear combination are a sequence of frequencies among the bins of the histogram. The dotted line is the  $1^{st}$ -order B-spline estimator, which is continuous but not everywhere differentiable as illustrated. The solid line and the dashed line are the  $2^{nd}$ -order and  $3^{rd}$ -order B-splines estimators whose properties are shown in Theorem (3.1) and (3.2).

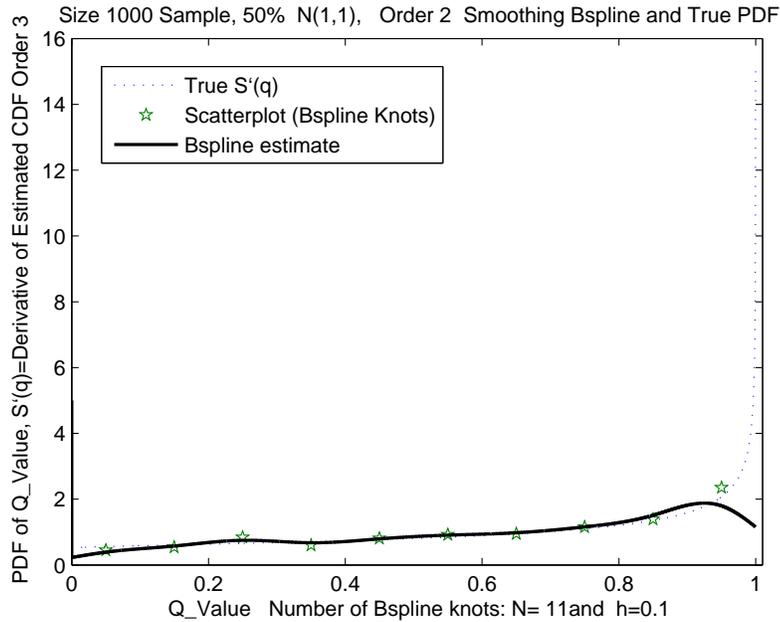
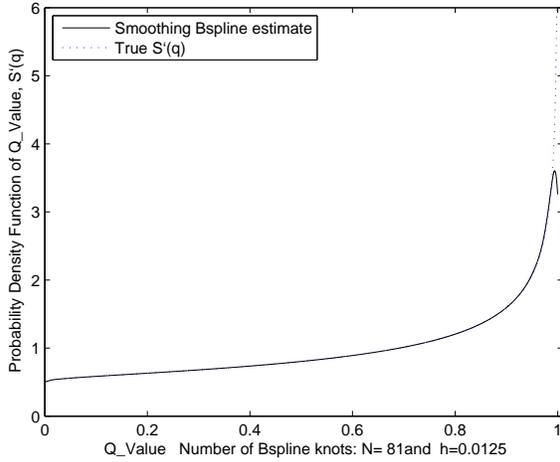


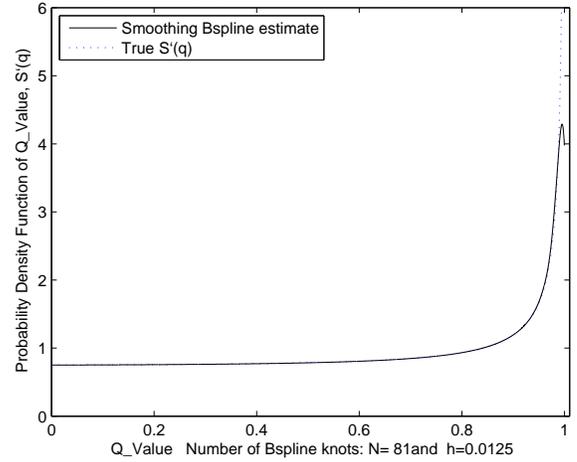
Figure 3.8: Estimated PDF of  $Q$ -value,  $S'(q) =$  derivative of estimated CDF Order 3, is the  $2^{nd}$ -order B-splines as shown in Lemma (3.6).

Size 1000 Sample, 50%  $N(1,1)$ , Order 3 Smoothing B spline and True PDF of Q\_Value



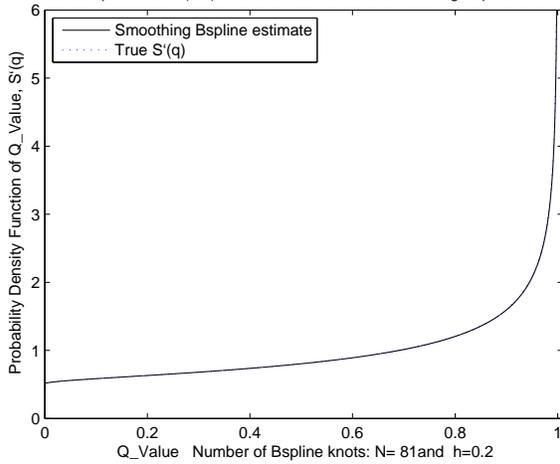
(a) one additional knot at  $q = 0$  and  $q = 1,50\%N(1, 1)$

Size 1000 Sample, 25%  $N(2,1)$ , Order 3 Smoothing B spline and True PDF of Q\_Value



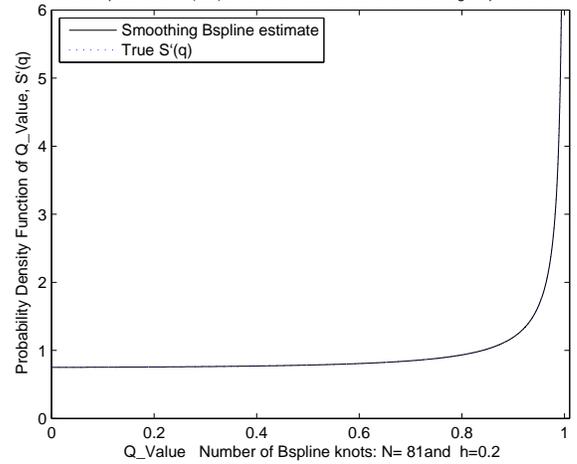
(b) one additional knot at  $q = 0$  and  $q = 1,25\%N(2, 1)$

Size 1000 Sample, 50%  $N(1,1)$ , Transformed Order 3 Smoothing B spline and True PDF



(c) Transform the  $Q$ -values to whole line and back

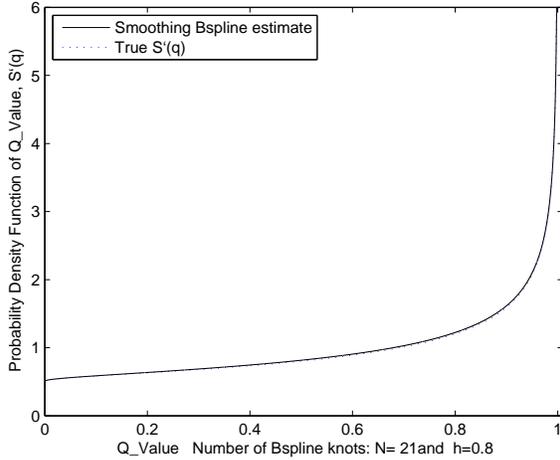
Size 1000 Sample, 25%  $N(2,1)$ , Transformed Order 3 Smoothing B spline and True PDF



(d) Transform the  $Q$ -values to whole line and back

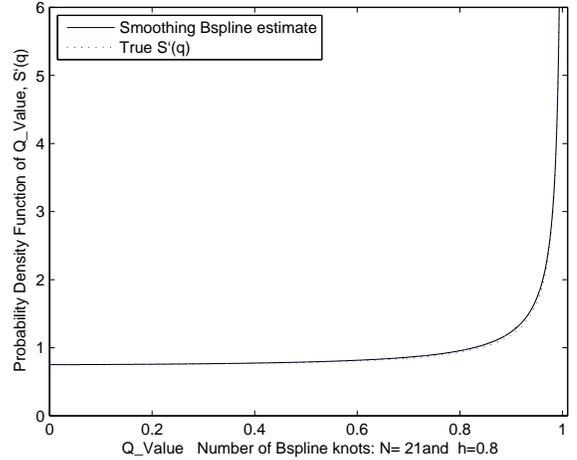
Figure 3.9: Estimated probability density  $S'(q)$  by Order 3 smoothing B splines.  $Q$ -values are sampled from the mixed population with the sample size  $n = 1000$ , of which in the left panel 500 are simulated from  $N(0, 1)$  and 500 are simulated from  $N(1, 1)$ ; in the right panel 750 are from  $N(0, 1)$  and 250 are from  $N(2, 1)$ . The two graphs in the upper panel show the estimated PDF (solid lines) compared with the true PDF (dotted lines) of  $Q$ -values before the transformation with the two additional knots as described in equation (3.15); and the two graphs in the lower panel show the estimated PDF respectively using the transformation  $\log \frac{q}{1-q}$  to the whole real line, and then transforming back.

Size 1000 Sample, 50%  $N(1,1)$ , Transformed Order 3 Smoothing Bspline and True PDF



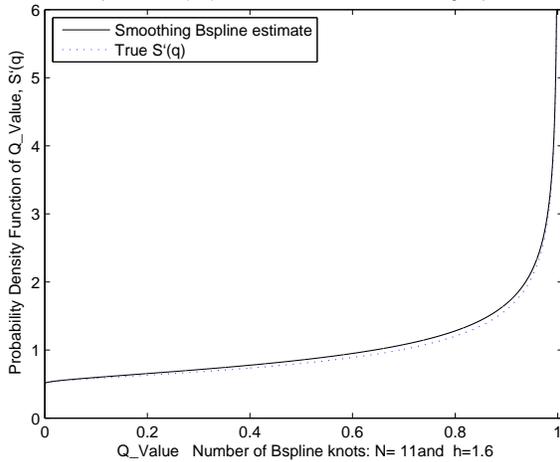
(a) 50% $N(1, 1)$  Bspline knots  $N = 21$

Size 1000 Sample, 25%  $N(2,1)$ , Transformed Order 3 Smoothing Bspline and True PDF



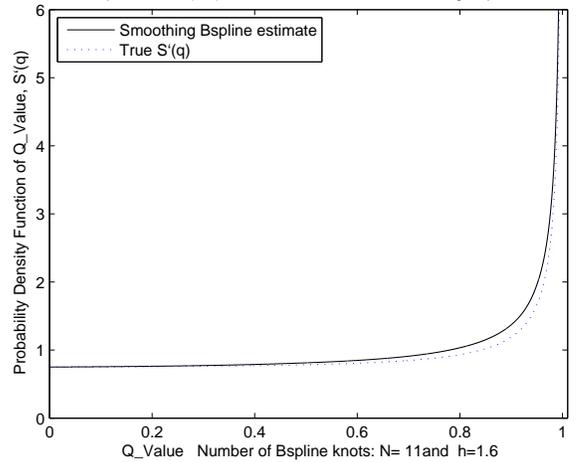
(b) 25% $N(2, 1)$  Bspline knots  $N = 21$

Size 1000 Sample, 50%  $N(1,1)$ , Transformed Order 3 Smoothing Bspline and True PDF



(c) 50% $N(1, 1)$  Bspline knots  $N = 11$

Size 1000 Sample, 25%  $N(2,1)$ , Transformed Order 3 Smoothing Bspline and True PDF



(d) 25% $N(2, 1)$  Bspline knots  $N = 11$

Figure 3.10: Estimated probability density  $S'(q)$  by using Order 3 smoothing Bsplines with different numbers of knots and by using the transformation  $\log \frac{q}{1-q}$ , and then transforming back.  $Q$ -values are sampled from the mixed population with the sample size  $n = 1000$ , of which in the left panel 500 are simulated from  $N(0, 1)$  and 500 are simulated from  $N(1, 1)$ ; in the right panel 750 are from  $N(0, 1)$  and 250 are from  $N(2, 1)$ . The two graphs in the upper panel show the estimated PDF with the number of Bspline knots, 21 (solid lines) compared with the true PDF (dotted lines) of  $Q$ -values; and the two graphs in the lower panel show the estimated PDF with the number of Bspline knots, 11. The Bspline estimators with more knots in the upper panel are more approximate to the true PDF than those in the lower panel, respectively. Hence, the precision of the estimators can be improved by increasing the number of knots as described in Theorem (3.3).

# CONCLUSION & FUTURE WORK

In this thesis a survey of methodologies to tackle multiplicity have been discussed. Despite the disagreement among the three inference schools, this thesis focuses on calibrating  $P$ -values from the empirical distribution of  $P$ -values from both frequentist and Bayesian perspectives.

A noninterpolatory and shape-preserving estimator based on B-splines as smoothing functions has been developed. According to Theorem (3.1) and (3.2), this shape-preserving B-spline approximator maps a convex or concave function to a convex or concave function. Therefore, the probability density estimator is increasing or decreasing based on the monotonicity of the estimated probability density function. Although any continuous function can be approximated to any desired precision by increasing the number of knots of B-splines based on Theorem (3.3), the accuracy of spline approximation can be achieved with an optimal variable knots placement unequally spaced such as Chebyshev points as knots, which will be an interesting topic for future research.

# Bibliography

- [1] Abramowitz, Milton and Stegun, Irene A., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables; New York: Wiley, (1972).
- [2] Ahlberg J. H., Nilson E. N., and Walsh J. L., The Theory of Splines and their Applications, New York: Academic Press, (1967).
- [3] Bickis, Mikelis (2004), Coping with multiplicity by exploiting the empirical distribution of p-values, The 6th World Congress of the Bernoulli Society and 67th Annual Meeting of the Institute of Mathematical Statistics, Barcelona, Catalunya, Spain.
- [4] Bickis, Mikelis, Bleuer S., and Krewski D. (1996), On the estimation of the proportion of positives in a sequence of screening experiments, The Canadian Journal of Statistics, Vol. 24, No. 1, Pages 1-15.
- [5] Bickis, Mikelis and Krewski D. (1989), Statistical issues in the analysis of the long term carcinogenicity bioassay in small rodents: An empirical evaluation of statistical decision rules, Fundamental and Applied Toxicology, 12, Pages 202-221.
- [6] Bickis, Mikelis, Statistical Inference, Lecture Notes Manuscript, (2007).
- [7] Bain, Lee J. and Engelhardt, Max, Introduction to probability and mathematical statistics, Pacific Grove, Calif.: Duxbury/Thomas Learning, (1992).
- [8] Berger, James O. (2003), Could Fisher, Jeffreys and Neyman have agreed on testing?, Statistical Science, Vol. 18, No. 1, Pages 1-32.
- [9] Berger J. O., Brown L. D., and Wolpert R. L.(1994), A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing, The Annals of Statistics, Vol. 22, Pages 1787-1807.
- [10] Benjamini Y. and Hochberg Y. (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society, Series B, Vol. 57, No. 1, Pages 289-300.

- [11] Boneva, Liliana I., Kendall, David, and Stefanov, Ivan (1971), Spline transformations: three new diagnostic aids for the statistical data analyst, *Journal of the Royal Statistical Society*, Vol. 33, No. 1, Pages 1-71.
- [12] Bowman A.W. and Azzalini A., *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, (1997).
- [13] Burden, Richard L. and Faires, Douglas J., *Numerical Analysis*, Belmont, CA: Thomson Brooks/Cole, (2005).
- [14] Chambers J. M. (1970), *Computers in statistical research: simulation and computer-aided mathematics*, American Statistical Association and American Society for Quality, Pages 1-15.
- [15] De Boor, Carl, *A Practical Guide to Splines*, New York: Springer-Verlag, (1978).
- [16] Eubank, Randall L., *Spline Smoothing and Nonparametric Regression*, New York: M. Dekker, (1988).
- [17] Fisher, Sir Ronald A., *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, (1970).
- [18] Fisher, Sir Ronald A., *Statistical Methods and Scientific Inference*, Edinburgh, Oliver and Boyd, (1959).
- [19] Fishburn, Peter C, *Mathematics of Decision Theory*, The Hague, Mouton, (1973).
- [20] Green P.J. and Silverman B.W., *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London; New York: Chapman and Hall, (1994).
- [21] Greville, T. N. E., *Theory and Applications of Spline Functions*, New York: Academic Press, (1969).
- [22] Harlow, Lisa L., Mulaik, Stanley A., and Steiger, James H., *What If There were No Significance Tests?*, Lawrence Erlbaum Associates Publishers, (1997).
- [23] Hastie T.J. and Tibshirani R.J., *Generalized Additive Models*, London; New York: Chapman and Hall, (1990).
- [24] Iacobucci, Dawn (2005), From the editor on p-values, *Journal of Consumer Research*.
- [25] Jaynes E.T., *Probability Theory: The Logic of Science*, Cambridge University Press, (2003).
- [26] Jeffreys, Harold, *Theory of Probability*, Oxford: Clarendon Press, (1961).
- [27] Krantz, David H. (1999), The null hypothesis testing controversy in psychology, *Journal of the American Statistical Association*, Vol.94, No. 448, Pages 1372-1381.

- [28] Kincaid, David and Ward, Cheney, Numerical Analysis : Mathematics of Scientific Computing, Australia; Pacific Grove, CA : Brooks/Cole/Thomson Learning, (2002).
- [29] Marsden, Martin J. (1970), An identity for spline functions with applications to variation-diminishing spline approximation, Journal of Approximation Theory, Vol. 3, No. 1, Pages 7-49.
- [30] Mayo, Deborah G., Error and the Growth of Experimental Knowledge, Chicago: University of Chicago Press, (1996).
- [31] Migon H.S. and Gamerman D., Statistical Inference: an Integrated Approach, London; New York: Arnold; New York: Copublished in the United States of America by Oxford University Press, (1999).
- [32] Press, William H. ... [et al.], Numerical Recipes in C: the Art of Scientific Computing, Cambridge; New York: Cambridge University Press, (1992).
- [33] Schilling, Robert J. and Harris, Sandra L., Applied Numerical Methods for Engineers Using MATLAB and C, Pacific Grove, Calif.: Brooks/Cole, (2000).
- [34] Schimek, Michael G., Smoothing and Regression: Approaches, Computation, and Application, New York: Wiley, (2000).
- [35] Shaffer, Juliet Popper (1995), Multiple hypothesis testing, Annual Reviews, Vol. 46, Pages 561-584.
- [36] Sheather, Simon J. (2004), Density estimation, Statistical Science, Vol. 19, No. 4, Pages 588-597.
- [37] Schoenberg I. J., Cardinal Spline Interpolation, Philadelphia, Society for Industrial and Applied Mathematics, (1973).
- [38] Schoenberg I. J., Approximations, with Special Emphasis on Spline Functions, New York: Academic Press, (1969).
- [39] Schultz, Martin H., Spline Analysis, Englewood Cliffs, N.J., Prentice-Hall, (1973).
- [40] Sellke, Thomas, Bayarri, M. J., and Berger, James O. (2001), Calibration of p-values for testing precise null hypotheses, The American Statistician, Vol. 55, No. 1, Pages 62-71.
- [41] Silverman B.W. (1984), Spline smoothing: the equivalent variable kernel method, The Annals of Statistics, Vol. 12, No. 3, Pages 898-916.
- [42] Silverman B.W. (1985), Some aspects of the spline smoothing approach to non-parametric regression curve fitting, Journal of the Royal Statistical Society, Vol. 47, No. 1, Pages 1-52.

- [43] Silverman B.W., Density Estimation for Statistics and Data Analysis, London; New York: Chapman and Hall, (1986).
- [44] Storey, John D. (2003), The positive false discovery rate: a Bayesian interpretation and the q-value, *Ann. Statist.*, 31, Pages 2013-2035.
- [45] Thisted, Ronald A., Elements of Statistical Computing, New York: Chapman and Hall, (1988).
- [46] Wegman, Edward J. and Wright, Ian W. (1983), Splines in statistics, *Journal of the American Statistical Association*, Vol. 78, No. 382, Pages 351-365.
- [47] Wood, John B. (1970), The p-value as an estimate of the posterior error probability, *Technometrics*, Vol. 12, No. 1, Pages 191-206.
- [48] Wood, John B., The Bayesian Probability of Accepting an Incorrect Directional Conclusion and its Relation to the P-value, PhD Dissertation, (1968).

# APPENDIX

```
function C(pi0,l,a,dis)
%
% Figure (1.3) and (1.4) simulate the proportion of tests having true nulls
% when p = 0.05 or p = 0.01.
% pi0 is the initial percentage of true nulls,
% a is relevant to the theta1, the means under the alternatives.
% l is the total number of tests,
% l0 is the number of tests under the nulls;
% l1 is the number of tests under the alternatives.
% The six options for alternatives means are accessed by setting dis
% equal to 1, 2, 3, 4, 5, and 6 respectively.
%
sigma=1
n=20
l0=round(l*pi0)
l1=l-l0
x0=sigma/sqrt(n)*randn(l0,1)
if dis==1
    x1=a+sigma/sqrt(n)*randn(l1,1);
    t0=abs(x0)/(sigma/sqrt(n));
    t1=abs(x1)/(sigma/sqrt(n));
    xx1=t1>1.96 & t1<=2;
    xx0=t0>1.96 & t0<=2;
    nu(1,1)=sum(xx0);
    nu(1,2)=sum(xx1);
    pr(1)=1/(1+sum(xx1)/sum(xx0));
    xx1=t1>2.576 & t1<=2.616;
    xx0=t0>2.576 & t0<=2.616;
```

```

nu(2,1)=sum(xx0);
nu(2,2)=sum(xx1);
pr(2)=1/(1+sum(xx1)/sum(xx0));
bar(nu);
title 'Size 100,000 with 50%  $X_{\{0\}} \sim N(0,1/20)$  and 50%  $X_{\{1\}} \sim N(\theta_{\{1\}},1/20)$ ,
\theta_{\{1\}}=a';
xlabel(['Proportion of Tests Having True Nulls: ', num2str(pr(1)),'& ',
num2str(pr(2)),' a=', num2str(a)]);
ylabel('Numbers of True Nulls and True Alternatives p=0.05 & 0.01');
end
if dis==2
theta1=a*randn(1,1);
x1=theta1+sigma/sqrt(n)*randn(1,1);
t0=abs(x0)/(sigma/sqrt(n));
t1=abs(x1)/(sigma/sqrt(n));
xx1=t1>1.96 & t1<=2;
xx0=t0>1.96 & t0<=2;
nu(1,1)=sum(xx0);
nu(1,2)=sum(xx1);
pr(1)=1/(1+sum(xx1)/sum(xx0));
xx1=t1>2.576 & t1<=2.616;
xx0=t0>2.576 & t0<=2.616;
nu(2,1)=sum(xx0);
nu(2,2)=sum(xx1);
pr(2)=1/(1+sum(xx1)/sum(xx0));
bar(nu);
title 'Size 100,000 with 50%  $X_{\{0\}} \sim N(0,1/20)$  and 50%  $X_{\{1\}} \sim N(\theta_{\{1\}},1/20)$ ,
\theta_{\{1\}} \sim N(0,a)';
xlabel(['Proportion of Tests Having True Nulls: ', num2str(pr(1)),'& ',
num2str(pr(2)),' a=', num2str(a)]);
ylabel('Numbers of True Nulls and True Alternatives p=0.05 & 0.01');
end
if dis==3
theta1=a*randn(1,1);
x1=abs(theta1)+sigma/sqrt(n)*randn(1,1);
t0=abs(x0)/(sigma/sqrt(n));

```

```

t1=abs(x1)/(sigma/sqrt(n));
xx1=t1>1.96 & t1<=2;
xx0=t0>1.96 & t0<=2;
nu(1,1)=sum(xx0);
nu(1,2)=sum(xx1);
pr(1)=1/(1+sum(xx1)/sum(xx0));
xx1=t1>2.576 & t1<=2.616;
xx0=t0>2.576 & t0<=2.616;
nu(2,1)=sum(xx0);
nu(2,2)=sum(xx1);
pr(2)=1/(1+sum(xx1)/sum(xx0));
bar(nu);
title 'Size 100,000 with 50%  $X_{0} \sim N(0, 1/20)$  and 50%  $X_{1} \sim N(|\theta_{1}|, 1/20)$ ,
\theta_{1} \sim N(0, a)';
xlabel(['Proportion of Tests Having True Nulls: ', num2str(pr(1)), '& ',
num2str(pr(2)), ', a=', num2str(a)]);
ylabel('Numbers of True Nulls and True Alternatives p=0.05 & 0.01');
end
if dis==4
theta1=-a + 2*a*rand(1,1);
x1=theta1+sigma/sqrt(n)*randn(1,1);
t0=abs(x0)/(sigma/sqrt(n));
t1=abs(x1)/(sigma/sqrt(n));
xx1=t1>1.96 & t1<=2;
xx0=t0>1.96 & t0<=2;
nu(1,1)=sum(xx0);
nu(1,2)=sum(xx1);
pr(1)=1/(1+sum(xx1)/sum(xx0));
xx1=t1>2.576 & t1<=2.616;
xx0=t0>2.576 & t0<=2.616;
nu(2,1)=sum(xx0);
nu(2,2)=sum(xx1);
pr(2)=1/(1+sum(xx1)/sum(xx0));
bar(nu);
title 'Size 100,000 with 50%  $X_{0} \sim N(0, 1/20)$  and 50%  $X_{1} \sim N(\theta_{1}, 1/20)$ ,
\theta_{1} \sim UNIF(-a, a)';

```

```

xlabel(['Proportion of Tests Having True Nulls: ', num2str(pr(1)),'& ',
num2str(pr(2)),' a=', num2str(a)]);
ylabel('Numbers of True Nulls and True Alternatives p=0.05 & 0.01');
end
if dis==5
theta1=exprnd(a,l1,1);
theta2=(-1).^(rand(l1,1)<0.5);
theta1=theta1.*theta2;
%hist(theta1);
x1=theta1+sigma/sqrt(n)*randn(l1,1);
t0=abs(x0)/(sigma/sqrt(n));
t1=abs(x1)/(sigma/sqrt(n));
xx1=t1>1.96 & t1<=2;
xx0=t0>1.96 & t0<=2;
nu(1,1)=sum(xx0);
nu(1,2)=sum(xx1);
pr(1)=1/(1+sum(xx1)/sum(xx0));
xx1=t1>2.576 & t1<=2.616;
xx0=t0>2.576 & t0<=2.616;
nu(2,1)=sum(xx0);
nu(2,2)=sum(xx1);
pr(2)=1/(1+sum(xx1)/sum(xx0));
bar(nu);
title 'Size 100,000 with 50%  $X_{\{0\}} \sim N(0,1/20)$  and 50%  $X_{\{1\}} \sim N(\theta_{\{1\}},1/20)$ ,
|\theta_{\{1\}}| \sim \text{EXP}(a)';
xlabel(['Proportion of Tests Having True Nulls: ', num2str(pr(1)),'& ',
num2str(pr(2)),' a=', num2str(a)]);
ylabel('Numbers of True Nulls and True Alternatives p=0.05 & 0.01');
end
if dis==6
theta1=exprnd(a,l1,1)-a;
x1=theta1+sigma/sqrt(n)*randn(l1,1);
t0=abs(x0)/(sigma/sqrt(n));
t1=abs(x1)/(sigma/sqrt(n));
xx1=t1>1.96 & t1<=2;
xx0=t0>1.96 & t0<=2;

```

```

nu(1,1)=sum(xx0);
nu(1,2)=sum(xx1);
pr(1)=1/(1+sum(xx1)/sum(xx0));
xx1=t1>2.576 & t1<=2.616;
xx0=t0>2.576 & t0<=2.616;
nu(2,1)=sum(xx0);
nu(2,2)=sum(xx1);
pr(2)=1/(1+sum(xx1)/sum(xx0));
bar(nu);
title 'Size 100,000 with 50%  $X_{\{0\}} \sim N(0, 1/20)$  and 50%  $X_{\{1\}} \sim N(\theta_{\{1\}}, 1/20)$ ,
 $\theta_{\{1\}} \sim \text{EXP}(a) - a$ ';
xlabel(['Proportion of Tests Having True Nulls: ', num2str(pr(1)), '& ',
num2str(pr(2)), ', a=', num2str(a)]);
ylabel('Numbers of True Nulls and True Alternatives p=0.05 & 0.01');
end
%
% Figure (2.5) is implemented using kernel density estimation.
% The ksdensity function, ksdensity, does this by using a kernel smoothing function
% and an associated bandwidth to estimate the density.
% call to ksdensity returns the default bandwidth, u, of the kernel
% smoothing function.
%savefile = 'kernel75.mat';
%save(savefile, 'knots', 'qcdf')
clear;
sigma=1;
n=1;
l=1000;
pi0=0.5;
l0=round(l*pi0);
kappa=l-l0;
mu0=0.;
mu1=1.;
x0=mu0+sigma/sqrt(n)*randn(l0,1)
x1=mu1+sigma/sqrt(n)*randn(kappa,1);
t0=x0/(sigma/sqrt(n));
t1=x1/(sigma/sqrt(n));

```

```

lambda=mu1-mu0;
if lambda>0
    q0=normcdf(t0);
    q1=normcdf(t1);
else
    q0=1-normcdf(t0);
    q1=1-normcdf(t1);
end
qq=[q0',q1'];
q1=reshape(qq,1000,1);
q_value=sort(q1);
for i=1:1000
y(i)=log(q_value(i)/(1-q_value(i)));
end
[f,x,u] = ksdensity(y);
for i=1:length(x)
q(i)=1-(1/(exp(x(i))+1));
end
for i=1:length(x)
ff(i)=f(i)*1/(q(i)*(1-q(i)));
end
savefile = 'kernel50_2.mat';
save(savefile, 'q_value','q','ff');
plot(q,ff,'k');
hold on;
x=0.001:0.001:0.999;
for i=1:999
    sq1(i+1)=s_prime(x(i),pi0,mu1,sigma);
end;
sq1(1)=pi0;sq1(1001)=2*sq1(1000)-sq1(999);
x=0:0.001:1;
plot(x,sq1,':');
set(gca,'ylim',[-0.01 5]);
set(gca,'xlim',[-0.01 1.01]);
title(['Size 1000 Sample, ',num2str((1-pi0)*100),'% N(',num2str(mu1),',',1),
Kernel Probability Density Estimation']);

```

```

xlabel('Q\_Value');
ylabel('Kernel Density Estimate of Q\_Values, S'(q)');
h = legend('Kernel Density Estimate','True Density',2);
set(h,'Interpreter','none');

%
% Figure (3.5) is implemented using Order~k Bsplines.
% pi0 is the initial percentage of true nulls,
% l is the total number of tests,
% l0 is the number of tests under the nulls;
% l1 is the number of tests under the alternatives.
% n: # of spline knots or bins=n-1
% k: the degree of B-spline
clear;
sigma=1;
n=1;
l=1000;
pi0=0.5;
l0=round(l*pi0);
kappa=l-l0;
mu0=0.;
mu1=1;
x0=mu0+sigma/sqrt(n)*randn(l0,1);
x1=mu1+sigma/sqrt(n)*randn(kappa,1);
t0=x0/(sigma/sqrt(n));
t1=x1/(sigma/sqrt(n));
lambda=mu1-mu0;
if lambda>0
    q0=normcdf(t0);
    q1=normcdf(t1);
else
    q0=1-normcdf(t0);
    q1=1-normcdf(t1);
end
q=[q0',q1'];
q1=reshape(q,1000,1);

```

```

q_value=sort(q1);
% n: # of spline knots or bins=n-1
n=21;
edges(1)=0.;
edges(n)=1.;
h=(edges(n)-edges(1))/(n-1);
for i=2:n-1
    edges(i)=edges(1)+(i-1)*h
end;
qfreq=hist(q_value,edges+h/2.);
knots=edges+h/2;
%cdfplot(q_value);
hold on;
qcdf(1)=qfreq(1);
for i=2:n-1
    qcdf(i)=qcdf(i-1)+qfreq(i);
end;
qcdf(n)=1.;
qcdf=qcdf./1000;
qcdf(n)=1.;
knots(n+1)=knots(n)+h;
qcdf(n+1)=2*qcdf(n)-qcdf(n-1);
%k: Bspline Order
k=2;
x=0:0.0001:1;
for i=1:10001
    y(i)=sum_base(x(i)-h/2.,k,knots',qcdf');
end;
plot(x,y,'k');
x=0:0.0001:1;
for i=1:10001
    sq(i)=s(x(i),pi0,mu1,sigma);
end;
plot(x,sq,':');
%scatter(knots,qcdf,'p');
title(['Size 1000 Sample, ',num2str(pi0*100),'% N(',num2str(mu1),'',1),

```

```

    Order ',num2str(k), ' Smoothing B spline Estimate and True CDF']);
xlabel(['Q\_Value   Number of B spline knots: N= ',
       num2str(n), 'and h=', num2str(h)]);
ylabel('S(q)');
h = legend('Order 2 Smoothing B spline Estimate','TrueCDF',2);
set(h,'Interpreter','none');
savefile = 'test324_June16.mat';
save(savefile, 'knots', 'qcdf');

```

```

function f=sum_base(x,dis,xxx,yyy)
%
%   This is based on equation (3.8) as an expectation.
%
f=0.;
n=length(xxx);
bins=n-1;
h=(xxx(n)-xxx(1))/bins;
x=(x-xxx(1))/h;
for i=1:n
    f=f+yyy(i)*pdf_Bspline(x-i+1,dis);
end;

```

```

function f=s(x,pi0,mu,sigma)
%
%   This is based on equation (2.10).
%
f=pi0*x+(1-pi0)*normcdf(norminv(x),mu,sigma);

```

```

function f=pdf_Bspline(x,dis)
%
%   This function is the base of Order k=0,1,2,3 Bsplines.
%
if dis==0
    if x>.5 |x<-.5
        f=0;
    end
end

```

```

    if x>-.5 & x<.5
        f=1;
    end
    if x==.5 | x==-.5
        f=0.5;
    end
end
if dis==1
    if x>=1 | x<=-1
        f=0;
    else
        f=1-abs(x);
    end
end
if dis==2
    if x>=1.5 | x<=-1.5
        f=0;
    end
    if x>-.5 & x<.5
        f=-x^2+.75;
    end
    if abs(x)>=.5 & abs(x)<=1.5
        f=0.5*x^2-1.5*abs(x)+9./8.;
    end
end
if dis==3
    if x>=2 | x<=-2
        f=0;
    end
    if x>=-1 & x<=1
        f=.5*abs(x*x*x)-x*x+2./3.;
    end
    if abs(x)>1 & abs(x)<2
        f=-1./6.*abs(x*x*x)+x*x-2*abs(x)+4./3.;
    end
end
end

```