

國立臺灣大學生物資源暨農學院農藝學研究所生物統計組

碩士論文

Division of Biometry, Graduate Institute of Agronomy

College of Bioresources and Agriculture

National Taiwan University

Master Thesis



常態混合分佈良率之區間估計

Interval Estimation for Conformance Proportions
in Normal Mixture Distributions

黃則仿

Tse-Le Huang

指導教授：蔡欣甫 博士

Advisor: Shin-Fu Tsai, Ph.D.

中華民國 108 年 6 月

June, 2019

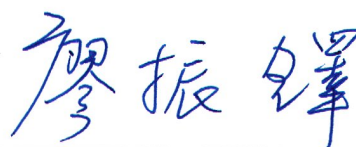
國立臺灣大學碩士學位論文
口試委員會審定書

常態混合分佈良率之區間估計
Interval Estimation for Conformance Proportions
in Normal Mixture Distributions

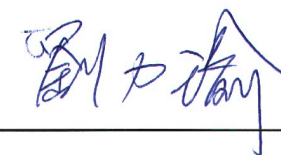
本論文係黃則仈君 (R06621208) 在國立臺灣大學農藝學研究所
生物統計組完成之碩士學位論文，於民國一百零八年六月十二日承下
列考試委員審查通過及口試及格，特此證明

口試委員：

國立臺灣大學農藝學研究所
廖振鐸 教授



國立臺灣大學農藝學研究所
劉力瑜 教授



國立臺灣大學農藝學研究所
蔡欣甫 助理教授 (指導教授)





致謝

時光飛逝，轉眼間兩年的研究生涯已經到了尾聲，感謝蔡教授對我的指導，除了學業上的幫助，還有其他做人處事的教誨也讓我受益匪淺，而教授寬容又認真的態度使我能在最小的壓力下順利完成研究，非常謝謝教授這兩年的教導。此外，還要感謝生統組的各位同學，特別是生統研究室的同學們，在我沮喪的時候聽我訴苦，當我遇到困難時不吝嗇地向我伸出援手，一起分享喜悅和歡笑，同甘共苦的這兩年將是我最珍貴的回憶。最後，感謝一直默默支持我的家人，沒有家人的陪伴我無法走到這一步，謝謝你們當我最強而有力的後盾，讓我能無後顧之憂完成學業，謝謝你們的陪伴和支持。

這本論文若缺少上述任何一人都無法完成，感謝大家對我的付出和鼓勵，謝謝。

黃則仉

謹致於臺灣大學

中華民國 108 年 6 月



摘要

良率為評估製程能力與品質的重要指標，目前被廣泛地應用於品質管制、環境監控與其他研究領域。假設目標族群服從常態混合分佈且規格已指定時，如何有系統地建構良率的信賴區間是目前尚未解決的問題。本研究提出一個針對常態混合分佈良率的區間估計方法，利用馬可夫鍊蒙地卡羅法自參數的廣義置信分佈抽樣並計算信賴區間。透過分析一筆實際環境監控的資料說明新方法的可行性，並藉由數值模擬評估新方法的表現。根據模擬結果，新方法所建構的信賴區間能提供足夠的覆蓋率。

關鍵字: 置信推論、信賴區間、潛在變數、馬可夫鍊蒙地卡羅法、品質管制。



Abstract

Conformance proportions, which are often employed in quality control, environmental monitoring, and many other areas, are important indices for evaluating product quality and process capability. When the population of interest is assumed to have a normal mixture distribution and specification limits are set by a quality engineer, estimating conformance proportions can be a practical issue. Under the framework of normal mixture distributions, a new method is proposed in this study to obtain confidence intervals for conformance proportions. More specifically, a Markov chain Monte Carlo sampler is developed to generate realizations from the generalized fiducial distributions. The required interval estimates can then be calculated by using the obtained realizations. A real-world environmental monitoring example is used to demonstrate that the proposed method is feasible in practice. Based on simulation results, it is shown that the proposed method can maintain the empirical coverage rate sufficiently close to the nominal level.

Keywords: Fiducial inference; Confidence interval; Latent variable; Markov chain Monte Carlo; Quality control.



Contents

口試委員會審定書	i
致謝	ii
摘要	iii
Abstract	iv
1 Introduction	1
2 Methods	3
2.1 Notation and Definitions	3
2.1.1 Normal Mixture Distributions	3
2.1.2 Conformance Proportions	4
2.1.3 Generalized Fiducial Inference	7
2.2 The Proposed Method	9
3 Results	13
3.1 An Application to Lake Acidity Data	13
3.2 Simulation Studies	17
4 Discussion	24
Bibliography	26



List of Figures

Figure 3.1: Histogram of lake acidity data	14
Figure 3.2: Autocorrelation functions and trace plots	16
Figure 3.3: Marginal density curves of scenarios 1 to 4	19
Figure 3.4: Marginal density curves of scenarios 5 to 8	20
Figure 4.1: Histogram of stamp thickness data	25



List of Tables

Table 3.1: Confidence intervals for lake acidity data	15
Table 3.2: Normal mixture distributions for simulation studies	18
Table 3.3: Empirical coverage rates for conformance proportions of entire population	21
Table 3.4: Empirical coverage rates for conformance proportions of subpopulation 1	22
Table 3.5: Empirical coverage rates for conformance proportions of subpopulation 2	23



Chapter 1

Introduction

A conformance proportion, which is defined as the probability of a quality characteristic within the specification limits set by a quality engineer, is an important numerical index for evaluating product quality and process capability. This numerical measure has been widely used in quality control, environmental regulation, and many other areas. To the best of my knowledge, Wang and Lam (1996) appears to be the first statistical reference regarding the construction of confidence intervals for conformance proportions. Perakis and Xekalaki (2002) and Perakis and Xekalaki (2005) studied conformance proportions under various distributional assumptions. Under the framework of linear mixed effects models, Lee and Liao (2012) and Lee and Liao (2014) developed systematic methods to obtain confidence intervals for bilateral and unilateral conformance proportions. When there are two or more quality characteristics of interest, Chen et al. (2015) conducted a series of Monte Carlo simulations to evaluate different interval estimation methods for conformance proportions of multiple quality characteristics. When prior knowledge and information are available, Perakis and Xekalaki (2015) proposed a Bayesian method to estimate conformance proportions. Recently, Lee et al. (2016) introduced the concept of reference population-based conformance proportions for the purpose of safety evaluation of genetically modified crops. The reader can consult Perakis and Xekalaki (2016) for a comprehensive introduction regarding conformance proportions and related applications.

In practice, sometimes, the population of interest can be characterized by a bimodal or multimodal distribution. Fitting a normal mixture distribution to the observed data is a straightforward strategy to characterize the target population. When the population is assumed to have a normal mixture distribution, the unknown parameters can be estimated by using the expectation-maximization algorithm developed by Dempster et al (1977). After obtaining the maximum likelihood estimates, the corresponding Fisher information matrix can be derived via the systematic procedure proposed by Louis (1982) to construct frequentist asymptotic inference procedures. On the other hand, when prior information and knowledge regarding the parameters are available, several Bayesian methods have been proposed for mixture distributions. For a comprehensive introduction to mixture distributions and related inference procedures, the reader is referred to McLachlan and Peel (2000). As far as I know, when the population is assumed to have a normal mixture distribution, there seems no literature regarding the construction of confidence intervals for conformance proportions. The aim of this study is to develop a systematic method for addressing this practical issue. In addition, the concepts of universal and individual conformance proportions are introduced to take the inherent structure of normal mixture distributions into account. This novel monitoring strategy allows us to explore the entire population and subpopulations of interest, respectively.

The remainder of this thesis is organized as follows. Chapter 2 first provides some notation and definitions. Next, when the population of interest is assumed to have a normal mixture distribution, a novel approach is proposed to obtain confidence intervals for bilateral and unilateral conformance proportions. Chapter 3 presents some numerical results to demonstrate that the proposed method can be a satisfactory solution for real-world applications. Some discussions are given in Chapter 4.



Chapter 2

Methods

On the basis of generalized fiducial inference, when the population is assumed to have a normal mixture distribution, a novel approach is proposed in this chapter to construct confidence intervals for conformance proportions.

2.1 Notation and Definitions

Before presenting the proposed method, some notation and definitions used in this study are first introduced.

2.1.1 Normal Mixture Distributions

Suppose that the population of interest can be split into J heterogeneous subpopulations, where J represents a known integer greater than or equal to two. Let X_i represent the continuous response of the i th individual sampled from the population. Assume that X_i can be characterized by the following normal mixture distribution:

$$X_i|w_i = j \sim N(\mu_j, \sigma_j^2) \quad \text{with probability } \Pr(W_i = j) = \pi_j,$$

where w_i denotes the realized value of membership indicator variable of the i th individual, the subpopulation label j is an integer ranging from 1 to J , μ_j and σ_j^2 denote the mean and

variance of the j th subpopulation, respectively, π_j represents the corresponding mixing proportion, and $\pi_1 + \pi_2 + \dots + \pi_J = 1$. The membership indicator variable W_i is often called a latent variable, whose realized value is unobserved in practice. Additionally, it is assumed that the J subpopulations are labeled properly, so that

$$\mu_1 < \mu_2 < \dots < \mu_J.$$

On the other hand, when the random variable X_i is sampled from a J -component normal mixture distribution, its marginal density function can be written as

$$f(x_i) = \sum_{j=1}^J \pi_j f(x_i | w_i = j),$$

where $f(x_i | w_i = j)$ denotes the conditional probability density function given by

$$f(x_i | w_i = j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right\}.$$

2.1.2 Conformance Proportions

Let X denote a quantitative quality characteristic of interest. When the lower specification limit L and the upper specification limit U are both set by a quality engineer, the bilateral conformance proportion can be defined as follows:

$$\Pr(L < X < U).$$

Clearly, a bilateral conformance proportion quantifies the uncertainty of X lying between the user-specified lower and upper specification limits. The reader can consult Ott et al. (2005) and Krishnamoorthy and Mathew (2009) for some real-world applications of bilateral conformance proportions. In some studies, however, only the lower or upper specification limit is used to evaluate the target quality characteristic. Consequently, the

corresponding conformance proportions are given by

$$\Pr(X > L) \text{ and } \Pr(X < U),$$



which are often called the unilateral conformance proportions. Unilateral conformance proportions also play a vital role in quality control and environmental monitoring. For example, Lee and Liao (2014) applied unilateral conformance proportions to evaluate rice quality and rain acidity.

Specifically, when the quality characteristic of interest is assumed to be a continuous random variable sampled from a normal mixture distribution, two classes of conformance proportions, including the individual and universal conformance proportions, are now introduced for real-world quality assessments. First, the individual bilateral conformance proportion of the j th subpopulation is defined as follows:

$$\begin{aligned} \theta_{bj} &= \Pr(L < X < U | W = j) \\ &= \Phi\left(\frac{U - \mu_j}{\sqrt{\sigma_j^2}}\right) - \Phi\left(\frac{L - \mu_j}{\sqrt{\sigma_j^2}}\right), \end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. On the other hand, when unilateral lower or upper specification limit is set by a quality engineer, the individual unilateral conformance proportions are given by

$$\begin{aligned} \theta_{lj} &= \Pr(X > L | W = j) \\ &= 1 - \Phi\left(\frac{L - \mu_j}{\sqrt{\sigma_j^2}}\right), \end{aligned}$$

and

$$\begin{aligned}\theta_{uj} &= \Pr(X < U|W = j) \\ &= \Phi\left(\frac{U - \mu_j}{\sqrt{\sigma_j^2}}\right),\end{aligned}$$



respectively. Individual conformance proportions are defined by using the conditional distribution for the j th subpopulation. Therefore, they offer local information regarding the j th subpopulation. On the other hand, based on the marginal distribution, the universal bilateral conformance proportion can be written as

$$\begin{aligned}\theta_b &= \Pr(L < X < U) \\ &= \sum_{j=1}^J \Pr(W = j) \Pr(L < X < U|W = j) \\ &= \sum_{j=1}^J \pi_j \theta_{bj}.\end{aligned}$$

Obviously, a universal bilateral conformance proportion is equal to the weighted sum of individual conformance proportions by mixing proportions. Similarly, universal unilateral conformance proportions can be written as

$$\begin{aligned}\theta_l &= \Pr(X > L) \\ &= \sum_{j=1}^J \Pr(W = j) \Pr(X > L|W = j) \\ &= \sum_{j=1}^J \pi_j \theta_{lj},\end{aligned}$$

and

$$\begin{aligned}\theta_u &= \Pr(X < U) \\ &= \sum_{j=1}^J \Pr(W = j) \Pr(X < U | W = j) \\ &= \sum_{j=1}^J \pi_j \theta_{uj},\end{aligned}$$



respectively. Although universal and individual conformance proportions can be used to evaluate the corresponding populations, they are complicated functions of all unknown parameters. To obtain confidence intervals for these complicated functions of parameters, generalized fiducial inference can be a feasible solution. Before presenting the proposed method, generalized fiducial inference is briefly introduced in the next section.

2.1.3 Generalized Fiducial Inference

Let \mathbf{X} represent a random vector sampled from a distribution with an unknown parameter vector ζ . Assume that \mathbf{X} can be generated by the following data-generating equation:

$$\mathbf{X} = G(\zeta, \mathbf{U}),$$

where $G(\cdot)$ denotes a measurable function, and \mathbf{U} denotes a random vector sampled from a completely known distribution. In addition, the realizations of \mathbf{X} and \mathbf{U} are denoted by \mathbf{x} and \mathbf{u} , respectively. Let

$$Q(\mathbf{x}, \mathbf{u}) = \{\zeta : \mathbf{x} = G(\zeta, \mathbf{u})\},$$

which can be regarded as an inverse mapping of the data-generating equation $G(\zeta, \mathbf{U})$. Accordingly, a generalized fiducial distribution of ζ can be defined as follows:

$$V(Q(\mathbf{x}, \mathbf{U}^*) | \{Q(\mathbf{x}, \mathbf{U}^*) \neq \emptyset\}),$$

where U^* represents an independent copy of U . Specifically, if there are two or more elements in $Q(x, U^*)$, the stochastic scheme $V(\cdot)$ can be used to randomly choose an element from $Q(x, U^*)$ for defining a generalized fiducial distribution. The reader is referred to Hannig (2009) for a rigorous definition regarding $V(\cdot)$. Typically, a random quantity sampled from a generalized fiducial distribution of ζ is said to be a generalized fiducial quantity for ζ , abbreviated as GFQ and denoted by $R_\zeta(x)$ or simply R_ζ . Below, two examples are given to show GFQs for the parameters of normal and multinomial distributions.

Example 1. Let X_1, X_2, \dots, X_n represent n independent random variables sampled from a normal distribution $N(\mu, \sigma^2)$. In addition, let \bar{x} and s^2 denote the observed sample mean and observed sample variance, respectively. Following the recipe proposed by Hannig (2009), a generalized fiducial distribution of μ and σ^2 can be obtained. Specifically, a GFQ for σ^2 is given by

$$R_{\sigma^2} = \frac{(n-1)s^2}{V},$$

where V denotes a chi-squared random variable with $n-1$ degrees of freedom. On the other hand, a GFQ for μ is given by

$$R_\mu = \bar{x} - Z\sqrt{\frac{R_{\sigma^2}}{n}},$$

where Z represents a standard normal random variable.

Example 2. Let W_1, W_2, \dots, W_n denote n independent discrete random variables which take the value j with probability

$$\Pr(W_i = j) = \pi_j,$$

where $j = 1, 2, \dots, J$. Assume that the number of occurrences of j among w_1, w_2, \dots, w_n

is equal to n_j . In addition, let U_1, U_2, \dots, U_n denote n random variables sampled from the continuous uniform distribution with bounded support between zero and one. As shown in Hannig (2009), a GFQ for π_1 is given by

$$R_{\pi_1} = U_{(t_1)} + D [U_{(t_1+1)} - U_{(t_1)}], \quad (1)$$

where $t_1 = n_1$, and D denotes a discrete uniform random variable taking value either 0 or 1. On the other hand, a GFQ for π_j is given by

$$R_{\pi_j} = U_{(t_j)} + D [U_{(t_j+1)} - U_{(t_j)}] - R_{\pi_{j-1}}, \quad (2)$$

where $t_j = n_1 + n_2 + \dots + n_j$, and $j = 2, 3, \dots, J$. Note that $U_{(0)}$ and $U_{(n+1)}$ are set to 0 and 1, respectively, in (1) and (2). In fact, there are several candidate probability models for D . The reader can consult Hannig (2009) for a fruitful discussion regarding these candidates.

2.2 The Proposed Method

Assume that X_1, X_2, \dots, X_n are n continuous random variables independently sampled from a normal mixture distribution. When the realized values of membership indicator variables w_1, w_2, \dots, w_n are all observed, that is, n_1, n_2, \dots, n_J are all observed, a GFQ for the j th subpopulation variance σ_j^2 is given by

$$R_{\sigma_j^2} = \frac{(n_j - 1) s_j^2}{V_j}, \quad (3)$$

where s_j^2 denotes the observed sample variance of the j th subpopulation, and V_j denotes a chi-squared random variable with $n_j - 1$ degrees of freedom. Similarly, a GFQ for the

j th subpopulation mean μ_j is given by

$$R_{\mu_j} = \bar{x}_j - Z_j \sqrt{\frac{R_{\sigma_j^2}}{n_j}}, \quad (4)$$



where \bar{x}_j represents the observed sample mean of the j th subpopulation, and Z_j represents a standard normal random variable. On the other hand, a GFQ for R_{π_j} can be computed by using (1) and (2). Accordingly, a GFQ for the j th individual bilateral conformance proportion θ_{bj} can be derived through

$$R_{\theta_{bj}} = \Phi \left(\frac{U - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right) - \Phi \left(\frac{L - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right). \quad (5)$$

Similarly, GFQs for the j th individual unilateral conformance proportions θ_{lj} and θ_{uj} are given by

$$R_{\theta_{lj}} = 1 - \Phi \left(\frac{L - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right), \quad (6)$$

and

$$R_{\theta_{uj}} = \Phi \left(\frac{U - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right), \quad (7)$$

respectively. On the other hand, a GFQ for the universal bilateral conformance proportion is given by

$$R_{\theta_b} = \sum_{j=1}^J R_{\pi_j} \left[\Phi \left(\frac{U - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right) - \Phi \left(\frac{L - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right) \right]. \quad (8)$$

In addition, GFQs for universal unilateral conformance proportions are given by

$$R_{\theta_l} = \sum_{j=1}^J R_{\pi_j} \left[1 - \Phi \left(\frac{L - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right) \right], \quad (9)$$



and

$$R_{\theta_u} = \sum_{j=1}^J R_{\pi_j} \Phi \left(\frac{U - R_{\mu_j}}{\sqrt{R_{\sigma_j^2}}} \right), \quad (10)$$

respectively. In practice, however, the realized values of the n membership indicator variables w_1, w_2, \dots, w_n are unobserved, with the result that all the required GFQs in (1) to (10) are unavailable. To account for the uncertainty of membership indicator variables W_1, W_2, \dots, W_n , the following Markov chain Monte Carlo (MCMC) sampler is proposed to generate realizations from the generalized fiducial distributions.

Step 1: Choose a sufficiently large number T as the length of Markov chain. Set $t = 0$. Generate an arbitrary assignment $\mathbf{w}^0 = (w_1^0, w_2^0, \dots, w_n^0)$, where w_i^0 represents a positive integer ranging from 1 to J for each i .

Step 2: Increase t by 1. Select an element, say w_i^{t-1} , from \mathbf{w}^{t-1} . Suppose that $w_i^{t-1} = a$, then randomly choose an integer, say b , from $\{1, 2, \dots, J\} \setminus a$. Propose a new assignment

$$\mathbf{w}^* = (w_1^*, w_2^*, \dots, w_n^*),$$

where $w_i^* = b$, and $w_j^* = w_j^{t-1}$ for all $j \neq i$.

Step 3: Calculate the acceptance ratio r via

$$r = \frac{(n_b - 1) \Gamma \left(\frac{n_a^* - 1}{2} \right) \Gamma \left(\frac{n_b^* - 1}{2} \right) [(n_a - 1) s_a^2]^{\frac{n_a - 2}{2}} [(n_b - 1) s_b^2]^{\frac{n_b - 2}{2}}}{(n_a^* - 1) \Gamma \left(\frac{n_a - 1}{2} \right) \Gamma \left(\frac{n_b - 1}{2} \right) [(n_a^* - 1) s_a^{*2}]^{\frac{n_a^* - 2}{2}} [(n_b^* - 1) s_b^{*2}]^{\frac{n_b^* - 2}{2}}}, \quad (11)$$

where $\Gamma(\cdot)$ represents the gamma function, s_a^2 and s_b^2 denote the sample variances of observations from the a th and b th subpopulations according to the assignment \boldsymbol{w}^{t-1} , s_a^{*2} and s_b^{*2} represent the sample variances of observations from the a th and b th subpopulations according to the assignment \boldsymbol{w}^* . Let U denote a random variable sampled from uniform $(0, 1)$. The realized value of U is denoted by u . The current assignment \boldsymbol{w}^t is set to

$$\boldsymbol{w}^t = \begin{cases} \boldsymbol{w}^* & \text{if } u \leq \min(r, 1); \\ \boldsymbol{w}^{t-1} & \text{otherwise.} \end{cases}$$

Step 4: Based on the current assignment \boldsymbol{w}^t , generate realizations of GFQs via (1) to (10).

Step 5: If $t < T$, back to Step 2. Otherwise, stop.

The number of burn-in iterations is set to $T/2$ in this study, and different values of T are set in the real data analysis and simulation studies, which will be mentioned later in the next chapter. After an adequate burn-in period, the proposed MCMC sampler can be used to generate realizations of GFQs from the generalized fiducial distributions. Specifically, the acceptance ratio r in (11) is a special case of equation (8) in Tsai (2019). This compact expression allows a practitioner to update the assignments in a computationally efficient manner.



Chapter 3

Results

Some numerical results are presented in this chapter to show that my proposal is a feasible and efficient method for real-world applications.

3.1 An Application to Lake Acidity Data

In 1983, the Environmental Protection Agency of the United States began the National Surface Water Survey to monitor water quality and acidification trends of lakes and streams in the United States. This project collected water samples from different lakes to analyze acid neutralizing capacities, pH levels, and other chemical attributes. When analyzing the observed data, most chemical attributes can be fitted by single probability models. As seen from Figure 3.1, however, it seems not adequate to fit the observed values of acid neutralizing capacities collected from north central Wisconsin by using a single probability distribution. Crawford et al. (1994) used a two-component normal mixture distribution to characterize this data set. The same probability model is applied in this study. Specifically, a lower specification limit is set to zero, which means that the lake water cannot neutralize acid. In other words, the level of acidification is rather severe.

To obtain confidence intervals for conformance proportions, the proposed MCMC sampler is implemented to generate realizations of required GFQs. The first 5,000,000 iterations are treated as burn-in samples, and then one realization is chosen in every 50

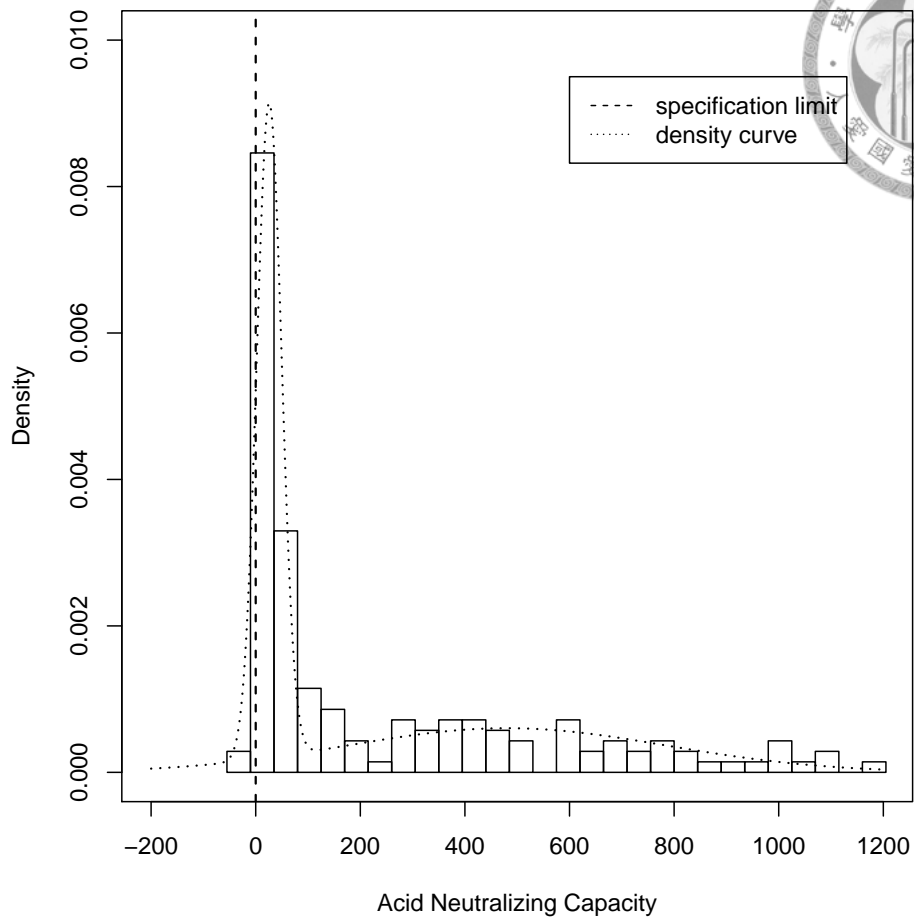


Figure 3.1: Histogram of lake acidity data

iterations for thinning the Markov chain. In total, 100,000 realizations are obtained for each GFQ. First, autocorrelation functions and trace plots are employed for diagnosing the obtained Markov samples. As shown in Figure 3.2, all autocorrelation coefficients are smaller than 0.2, so the obtained realizations are subjectively treated as independent samples in the subsequent analysis. On the other hand, there appears to be no specific pattern in each trace plot in Figure 3.2, which indicates that the obtained realizations are sampled from a stationary distribution. The formal testing procedure proposed by Geweke (1992) is further used to evaluate each Markov sequence. Note that Geweke's test statistic can be easily obtained by using the R package `coda`. When the significance level is set to 0.05, according to the p -values reported in Table 3.1, no strong evidence

against the assumption that these realizations are sampled from stationary distributions is observed. Subsequently, confidence intervals for conformance proportions are derived by using the obtained realizations of GFQs. When the lower specification limit is set to zero, as shown in Table 3.1, the true value of universal unilateral conformance proportion lies between 0.852 and 0.931 with 95% confidence. This interval estimate provides global information regarding the entire lake population. Specifically, with 95% confidence, the interval 0.789 to 0.915 contains the true value of the individual unilateral conformance proportion of the first subpopulation. Lastly, the interval 0.883 to 0.977 contains the true value of the individual unilateral conformance proportion of the second subpopulation with 95% confidence. Clearly, these confidence intervals provide local information about the two lake subpopulations. The 99% confidence intervals presented in Table 3.1 can be interpreted similarly. From this real-world application, the proposed method appears to be a feasible solution to obtain confidence intervals for universal and individual conformance proportions. In the next section, the proposed method will be thoroughly evaluated via a series of simulation studies.

Table 3.1: Confidence intervals for lake acidity data

Parameter	95% Confidence interval		99% Confidence interval		Geweke's test
	Lower limit	Upper limit	Lower limit	Upper limit	<i>p</i> -value
θ_l	0.852	0.931	0.835	0.940	0.575
θ_{l1}	0.789	0.915	0.763	0.930	0.758
θ_{l2}	0.883	0.977	0.859	0.983	0.310

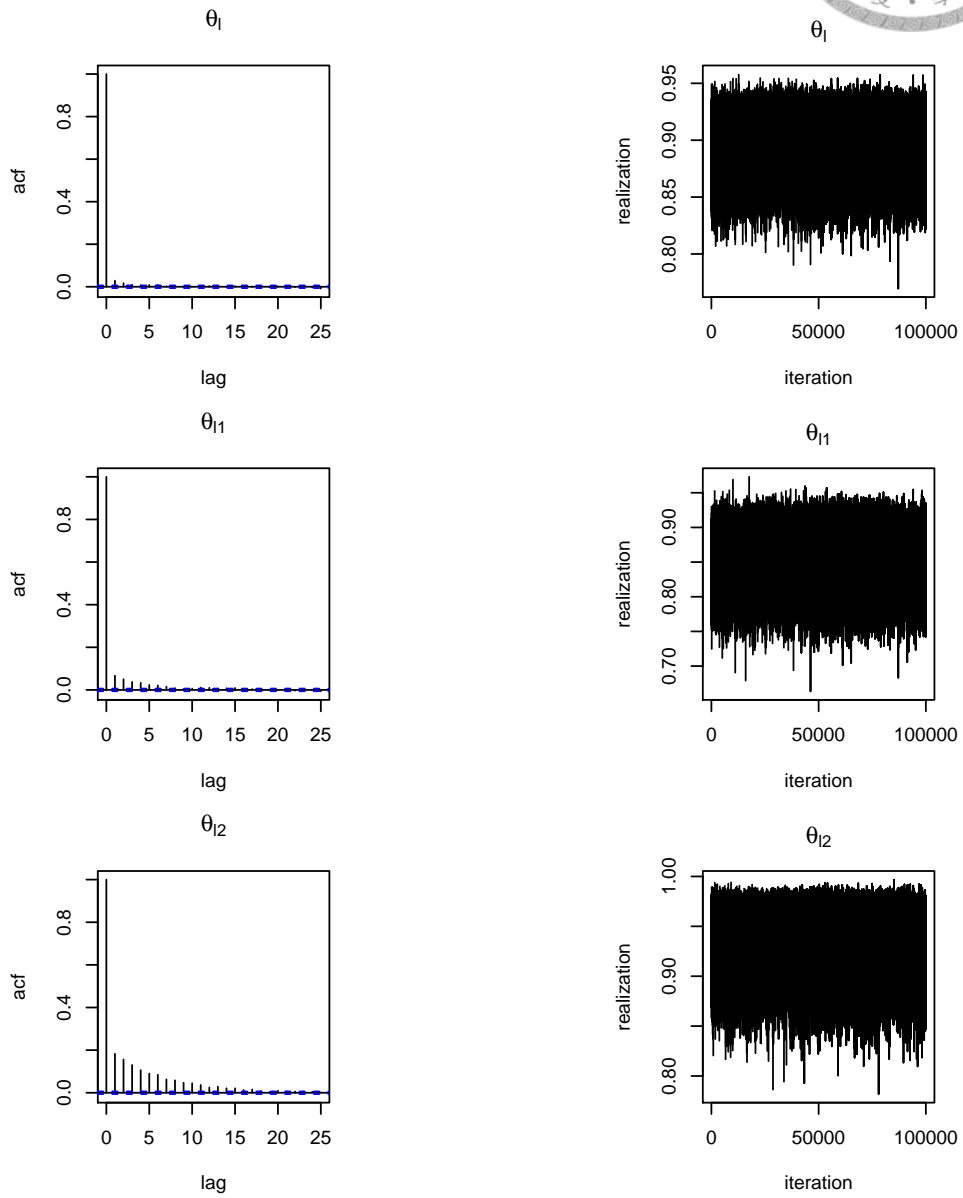
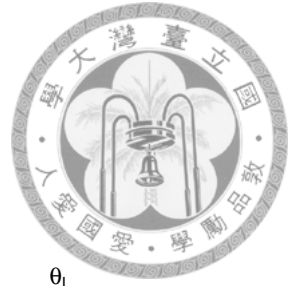


Figure 3.2: Autocorrelation functions and trace plots

3.2 Simulation Studies

To evaluate the proposed method, eight scenarios with different marginal density curves in Table 3.2 are considered. Although the number of subpopulations J is set to two for each scenario in this study, the proposed method can be readily applied to handle the cases of $J > 2$ without any difficulty. Two sets of mixing proportions $(\pi_1, \pi_2) = (0.4, 0.6)$ and $(0.2, 0.8)$ are chosen for the simulation studies. The sample sizes n are set to 50, 100, 200 and 300, respectively. In addition, the specification limits (L, U) are set to $(9, 16)$ and $(8, 17)$ for determining the true values of conformance proportions. By varying the true models exhibited in Table 3.2, n responses are generated to emulate observations from real-world applications.

To take the membership uncertainty of n individuals into account, n discrete random variables W_1, W_2, \dots, W_n are first generated via the following scheme:

$$\Pr(W_i = j) = \pi_j.$$

Next, if the realized value $w_i = 1$, then the corresponding X_i is generated from $N(\mu_1, \sigma_1^2)$. Otherwise, the corresponding X_i is simulated from $N(\mu_2, \sigma_2^2)$. For each combination of simulation parameters, including scenario and specification limits, 5000 simulated data sets are generated. Subsequently, 100,000 realizations of GFQs are generated for each simulated data set, and the required confidence intervals are then computed by using the obtained realizations. The performance of the proposed method is evaluated in terms of its empirical coverage rate at the nominal level 0.95. Specifically, the empirical coverage rate is defined as the proportion of the 5000 obtained confidence intervals containing the true conformance proportion. All simulation results are collected in Tables 3.3 to 3.5.

According to the empirical coverage rates in Tables 3.3 to 3.5, when the sample size n is small, the proposed method appears to be conservative for estimating conformance proportions. Specifically, when the specification limits (L, U) are set to $(8, 17)$ and $n =$

Table 3.2: Normal mixture distributions for simulation studies

Scenario	μ_1	σ_1^2	π_1	μ_2	σ_2^2	π_2
1	10	1	0.4	15	2	0.6
2	10	2	0.6	15	1	0.4
3	10	1	0.2	15	2	0.8
4	10	2	0.8	15	1	0.2
5	10	1	0.4	15	4	0.6
6	10	4	0.6	15	1	0.4
7	10	1	0.2	15	4	0.8
8	10	4	0.8	15	1	0.2



50 or 100, the empirical coverage rates of scenarios 5, 6, 7 and 8 in Table 3.3 are slightly lower than the nominal level 0.95, primarily due to the fact that the sample size n_i is rather small and subpopulation variance σ_i^2 is relatively large. However, when the sample size is large, the proposed method can maintain empirical coverage rate sufficiently close to the nominal level. When the sample size n is 300, the gaps between the empirical coverage rates and nominal level are negligible. In summary, the proposed method appear to be a satisfactory solution for each scenario in Table 3.2.

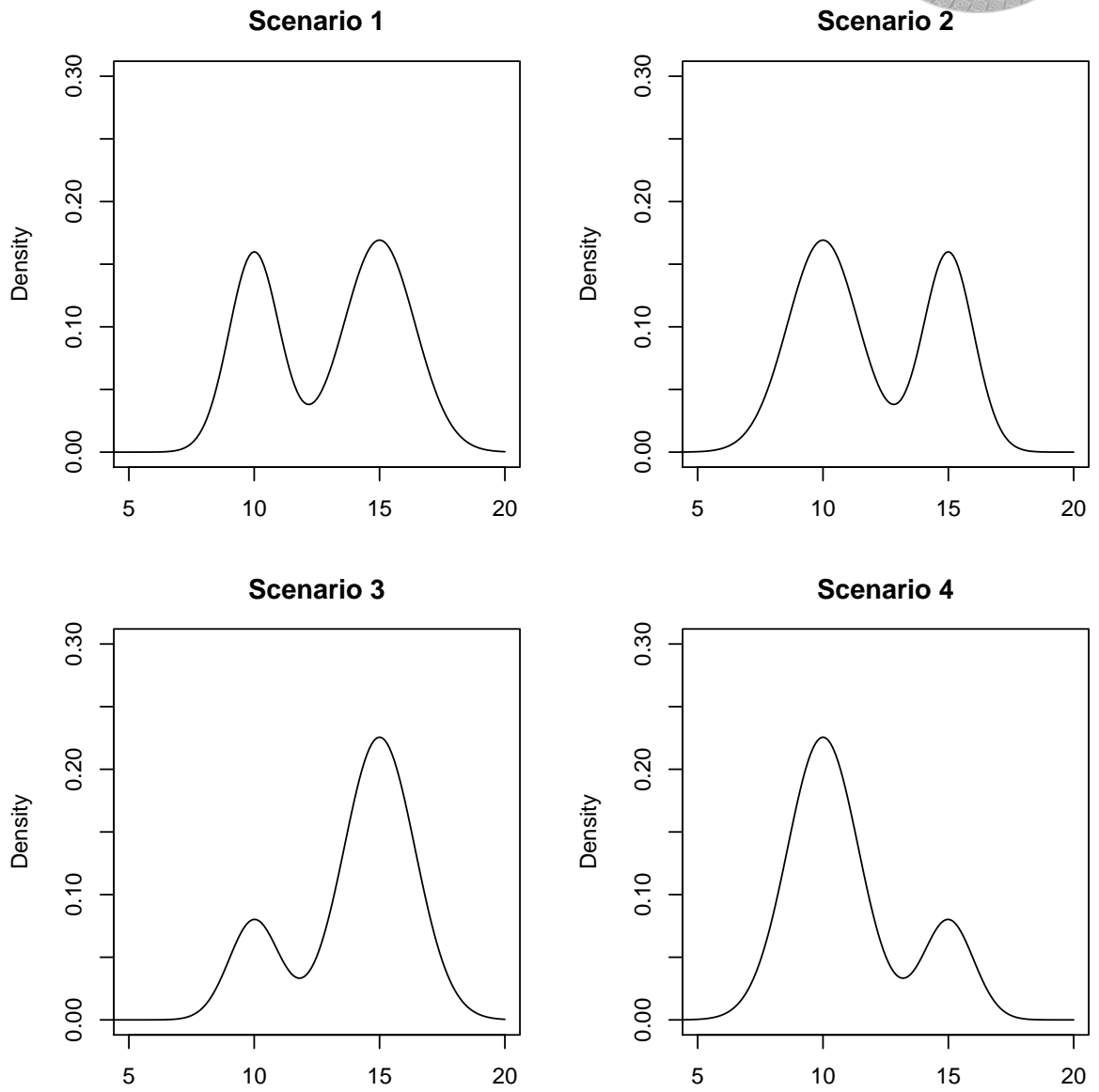


Figure 3.3: Marginal density curves of scenarios 1 to 4

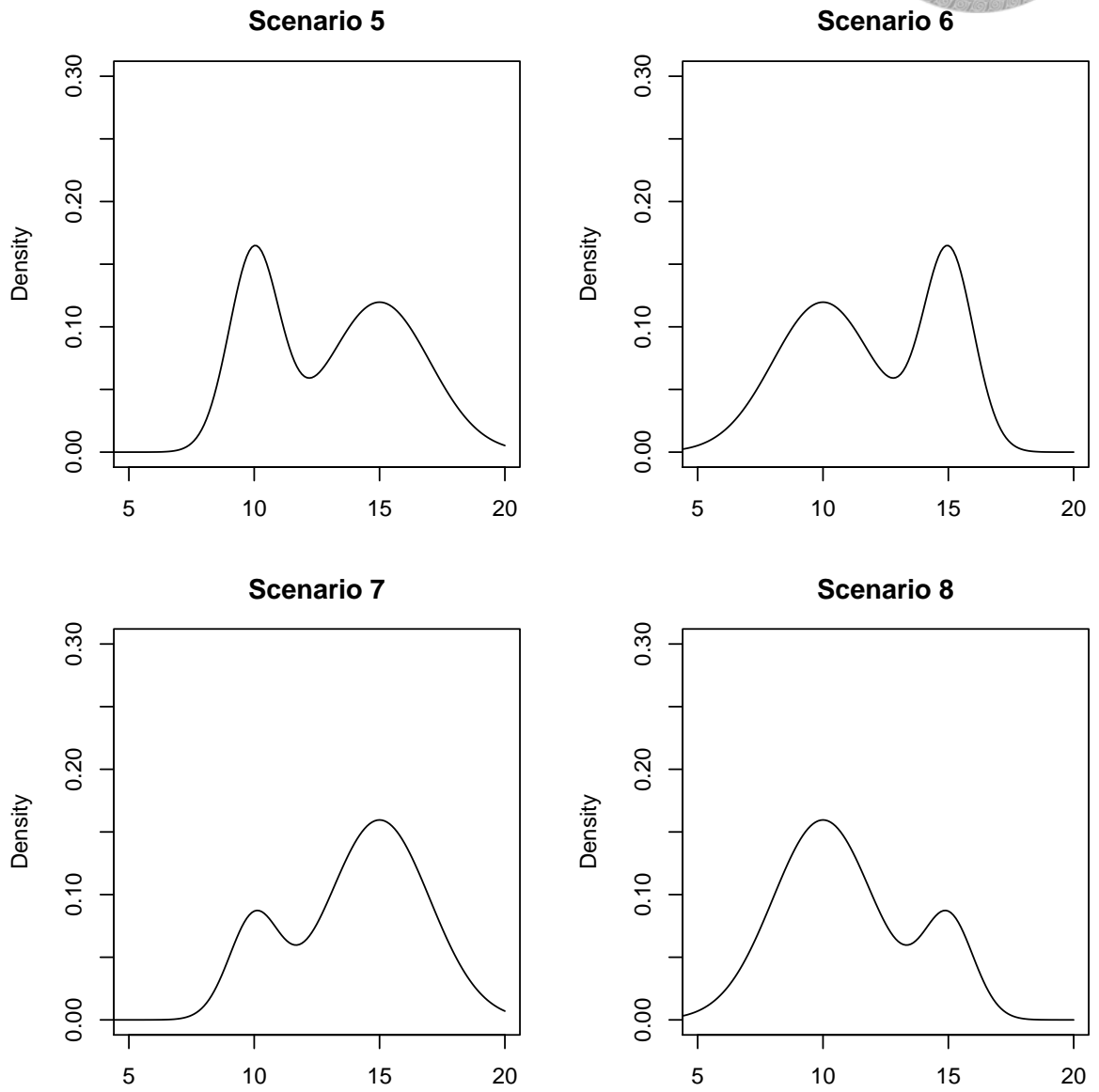


Figure 3.4: Marginal density curves of scenarios 5 to 8



Table 3.3: Empirical coverage rates for conformance proportions of entire population

Scenario	Type	$(L, U) = (9, 16)$				$(L, U) = (8, 17)$			
		$n=50$	$n=100$	$n=200$	$n=300$	$n=50$	$n=100$	$n=200$	$n=300$
1	θ_b	0.962	0.956	0.949	0.945	0.945	0.951	0.949	0.953
	θ_l	0.967	0.949	0.954	0.954	0.952	0.955	0.951	0.948
	θ_u	0.957	0.952	0.945	0.950	0.961	0.954	0.950	0.950
2	θ_b	0.965	0.949	0.947	0.949	0.939	0.948	0.950	0.952
	θ_l	0.958	0.955	0.950	0.949	0.954	0.955	0.952	0.953
	θ_u	0.968	0.948	0.957	0.950	0.954	0.947	0.947	0.953
3	θ_b	0.966	0.951	0.954	0.950	0.949	0.956	0.949	0.955
	θ_l	0.979	0.969	0.954	0.956	0.964	0.952	0.949	0.946
	θ_u	0.963	0.951	0.952	0.949	0.959	0.953	0.950	0.954
4	θ_b	0.965	0.954	0.953	0.951	0.952	0.952	0.948	0.953
	θ_l	0.963	0.952	0.948	0.948	0.959	0.954	0.946	0.947
	θ_u	0.976	0.966	0.961	0.951	0.960	0.952	0.953	0.949
5	θ_b	0.970	0.961	0.953	0.955	0.948	0.954	0.951	0.949
	θ_l	0.970	0.964	0.950	0.951	0.923	0.956	0.957	0.954
	θ_u	0.957	0.954	0.951	0.951	0.968	0.958	0.955	0.950
6	θ_b	0.966	0.961	0.959	0.955	0.948	0.950	0.947	0.948
	θ_l	0.959	0.954	0.958	0.952	0.963	0.955	0.950	0.952
	θ_u	0.972	0.963	0.951	0.950	0.924	0.950	0.957	0.955
7	θ_b	0.970	0.967	0.957	0.956	0.950	0.945	0.953	0.947
	θ_l	0.975	0.977	0.961	0.955	0.921	0.921	0.948	0.958
	θ_u	0.962	0.960	0.954	0.953	0.965	0.956	0.958	0.948
8	θ_b	0.967	0.966	0.957	0.959	0.949	0.949	0.949	0.945
	θ_l	0.962	0.956	0.950	0.951	0.967	0.958	0.953	0.947
	θ_u	0.979	0.973	0.965	0.952	0.912	0.930	0.947	0.949



Table 3.4: Empirical coverage rates for conformance proportions of subpopulation 1

Scenario	Type	$(L, U) = (9, 16)$				$(L, U) = (8, 17)$			
		$n=50$	$n=100$	$n=200$	$n=300$	$n=50$	$n=100$	$n=200$	$n=300$
1	θ_{b1}	0.970	0.951	0.960	0.952	0.961	0.954	0.948	0.943
	θ_{l1}	0.971	0.951	0.960	0.952	0.965	0.954	0.948	0.943
	θ_{u1}	0.971	0.953	0.953	0.954	0.967	0.957	0.954	0.949
2	θ_{b1}	0.967	0.957	0.948	0.950	0.962	0.955	0.952	0.950
	θ_{l1}	0.969	0.958	0.948	0.950	0.964	0.955	0.952	0.950
	θ_{u1}	0.971	0.954	0.958	0.956	0.974	0.963	0.958	0.951
3	θ_{b1}	0.983	0.973	0.957	0.952	0.962	0.955	0.947	0.946
	θ_{l1}	0.986	0.972	0.956	0.953	0.985	0.968	0.950	0.947
	θ_{u1}	0.955	0.939	0.947	0.954	0.954	0.942	0.942	0.945
4	θ_{b1}	0.971	0.959	0.954	0.950	0.973	0.961	0.945	0.948
	θ_{l1}	0.973	0.958	0.955	0.950	0.973	0.961	0.945	0.948
	θ_{u1}	0.976	0.966	0.955	0.951	0.968	0.967	0.960	0.954
5	θ_{b1}	0.980	0.962	0.954	0.954	0.981	0.967	0.958	0.951
	θ_{l1}	0.980	0.963	0.954	0.954	0.983	0.967	0.958	0.951
	θ_{u1}	0.982	0.971	0.958	0.960	0.981	0.972	0.959	0.950
6	θ_{b1}	0.981	0.971	0.959	0.953	0.977	0.962	0.954	0.952
	θ_{l1}	0.985	0.970	0.956	0.954	0.981	0.967	0.955	0.952
	θ_{u1}	0.989	0.971	0.948	0.947	0.988	0.967	0.947	0.941
7	θ_{b1}	0.991	0.986	0.969	0.960	0.985	0.978	0.962	0.959
	θ_{l1}	0.991	0.978	0.965	0.959	0.996	0.987	0.970	0.963
	θ_{u1}	0.968	0.974	0.962	0.956	0.966	0.974	0.960	0.956
8	θ_{b1}	0.984	0.980	0.963	0.959	0.985	0.981	0.966	0.953
	θ_{l1}	0.988	0.983	0.964	0.961	0.988	0.982	0.968	0.954
	θ_{u1}	0.989	0.979	0.969	0.956	0.990	0.981	0.964	0.958



Table 3.5: Empirical coverage rates for conformance proportions of subpopulation 2

Scenario	Type	$(L, U) = (9, 16)$				$(L, U) = (8, 17)$			
		$n=50$	$n=100$	$n=200$	$n=300$	$n=50$	$n=100$	$n=200$	$n=300$
1	θ_{b2}	0.970	0.956	0.945	0.952	0.963	0.953	0.949	0.952
	θ_{l2}	0.969	0.957	0.946	0.954	0.968	0.958	0.952	0.945
	θ_{u2}	0.972	0.957	0.944	0.952	0.967	0.954	0.949	0.952
2	θ_{b2}	0.968	0.954	0.952	0.955	0.962	0.948	0.945	0.953
	θ_{l2}	0.972	0.956	0.952	0.953	0.967	0.956	0.950	0.952
	θ_{u2}	0.972	0.955	0.952	0.955	0.966	0.949	0.945	0.953
3	θ_{b2}	0.970	0.956	0.953	0.951	0.973	0.956	0.948	0.952
	θ_{l2}	0.972	0.962	0.957	0.956	0.973	0.963	0.951	0.953
	θ_{u2}	0.971	0.956	0.953	0.951	0.973	0.956	0.948	0.952
4	θ_{b2}	0.985	0.976	0.960	0.949	0.962	0.955	0.953	0.950
	θ_{l2}	0.955	0.948	0.950	0.950	0.954	0.944	0.941	0.946
	θ_{u2}	0.985	0.972	0.960	0.949	0.984	0.967	0.957	0.950
5	θ_{b2}	0.984	0.970	0.959	0.959	0.982	0.964	0.957	0.951
	θ_{l2}	0.989	0.969	0.953	0.956	0.987	0.972	0.953	0.950
	θ_{u2}	0.989	0.971	0.957	0.960	0.986	0.964	0.958	0.952
6	θ_{b2}	0.976	0.962	0.953	0.956	0.982	0.965	0.958	0.954
	θ_{l2}	0.979	0.968	0.959	0.959	0.985	0.971	0.962	0.957
	θ_{u2}	0.978	0.963	0.953	0.956	0.983	0.966	0.958	0.954
7	θ_{b2}	0.985	0.980	0.964	0.956	0.986	0.978	0.966	0.954
	θ_{l2}	0.990	0.983	0.966	0.957	0.991	0.982	0.967	0.951
	θ_{u2}	0.990	0.984	0.968	0.957	0.988	0.980	0.967	0.955
8	θ_{b2}	0.994	0.986	0.969	0.958	0.983	0.981	0.963	0.959
	θ_{l2}	0.966	0.973	0.963	0.954	0.968	0.976	0.964	0.960
	θ_{u2}	0.991	0.981	0.966	0.958	0.995	0.988	0.972	0.964



Chapter 4

Discussion

When the population of interest is assumed to have a normal mixture distribution with J subpopulations, where J is a known positive integer, a new method is proposed in this study to obtain confidence intervals for conformance proportions. Based on simulation results, the proposed method can maintain the empirical coverage rate sufficiently close to the nominal level. Accordingly, it is recommended for real-world applications.

In practice, however, the number of subpopulations J could be an unknown integer, so that a practitioner needs to estimate J before applying the proposed method. Under the framework of normal mixture distributions, some statistical methods have been proposed to estimate the number of subpopulations. The proposals include those by Richardson and Green (1997) and Dellaportas and Papageorgiou (2006), among others. Sometimes, estimating the number of subpopulations J can be a challenging task. A typical example is the stamp thickness data set presented in Izeman and Sommer (1988). As seen from Figure 4.1, the number of subpopulations is not trivial. Based on the analysis results by McLachlan and Peel (2000), both three-component and seven-component normal mixture distributions can be used to fit this data set. Intuitively, if J is misspecified, the proposed method might be relatively inefficient for estimating the conformance proportions. The performance of the proposed method under a misspecified J will be addressed in a future study.

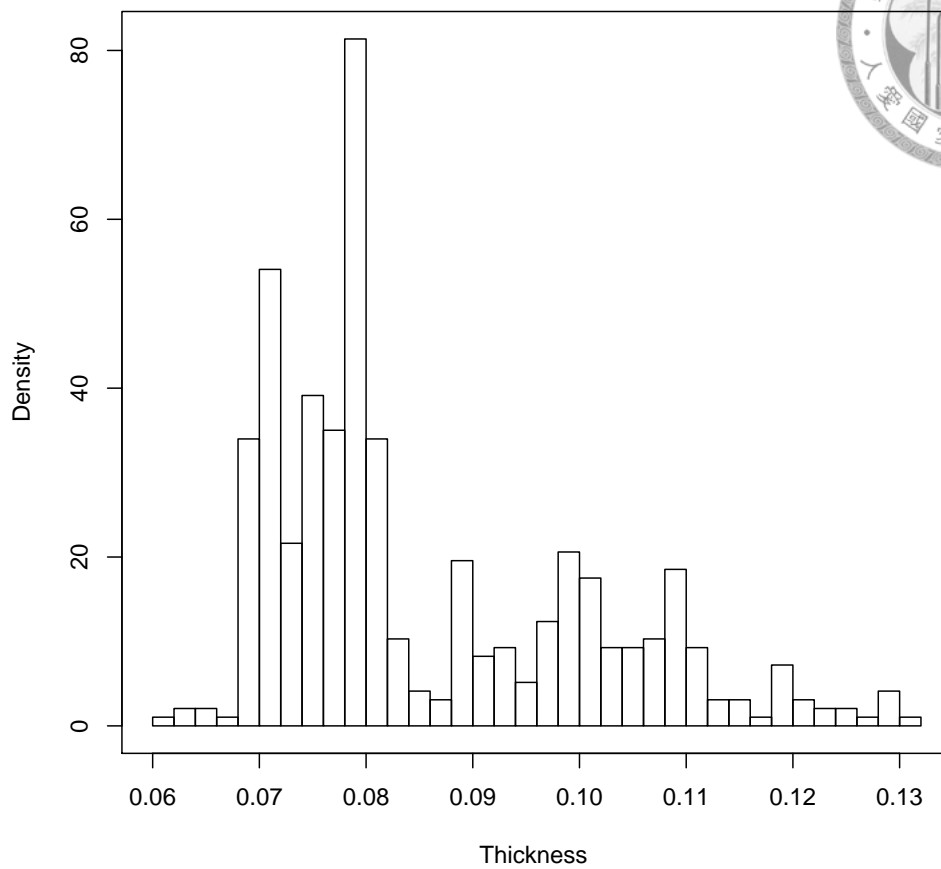


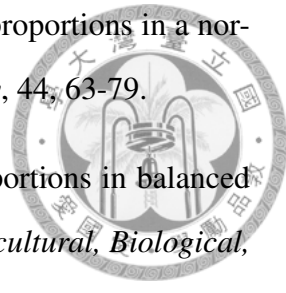
Figure 4.1: Histogram of stamp thickness data


Multivariate normal mixture distributions are commonly used to characterize multiple quality characteristics observed from different sources. Recently, Zimmer et al. (2016) introduced some applications of bivariate normal mixture distributions. An alternative research topic is to extend the proposed method to construct confidence regions under the assumption of multivariate normal mixture distributions. This interesting topic will be one of my future research projects.



Bibliography

- [1] Chen, C. L., Ou, S. L. and Liao, C. T. (2015). Interval estimation for conformance proportions of multiple quality characteristics. *Journal of Applied Statistics*, 42, 1829-1841.
- [2] Crawford, S. L., DeGroot, M. H., Kadane, J. B. and Small, M. J. (1994). Modeling lake-chemistry distributions: approximate Bayesian methods for estimating a finite-mixture model. *Technometrics*, 34, 441-453.
- [3] Dellaportas, P. and Papageorgiou, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing*, 16, 57-68.
- [4] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society. Series B*, 39, 1-38.
- [5] Gewekw, J. (1992). Evaluation the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, 4, 169-193.
- [6] Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19, 491-544.
- [7] Izenman, A. J. and Sommer, C. J. (1988). Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, 83, 941-953.
- [8] Krishnamoorthy, K. and Mathew, T. (2009). *Statistical tolerance regions*. Hoboken NJ: John, Wiley & Sons, Inc.

- 
- [9] Lee, H. I. and Liao, C. T. (2012). Estimation for conformance proportions in a normal variance components model. *Journal of Quality Technology*, 44, 63-79.
- [10] Lee, H. I. and Liao, C. T. (2014). Unilateral conformance proportions in balanced and unbalanced normal random effects models. *Journal of Agricultural, Biological, and Environment Statistic*, 19, 202-218.
- [11] Lee, H. I., Chen, H., Kishino, H. and Liao, C. T. (2016). A reference population-based conformance proportion. *Journal of Agricultural, Biological, and Environmental Statistics*, 21, 684-697.
- [12] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 44, 226-223.
- [13] McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons, Inc.
- [14] Ott, E. R., Schilling, E. G. and Neubauer, D. V. (2005). *Process quality control*, 4th Edition. Milwaukee, WI: ASQ Quality Press.
- [15] Perakis, M. and Xekalaki, E. (2002). A process capability index that is based on the proportion of conformance. *Journal of Statistical Computation and Simulation*, 72, 707-718.
- [16] Perakis, M. and Xekalaki, E. (2005). A process capability index for discrete process. *Journal of Statistical Computation and Simulation*, 75, 175-187.
- [17] Perakis, M. and Xekalaki, E. (2015). Assessing the proportion of conformance of a process from a Bayesian perspective. *Quality Reliability and Engineering International*, 31, 381-387.
- [18] Perakis, M. and Xekalaki, E. (2016). On the relationship between process capability indices and the proportion of conformance. *Quality Technology and Quantitative Management*, 13, 207-220.

- 
- [19] Richardson, S. and Green, P. J. (1997). On the Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B*, 59,731-792.
- [20] Tsai, S. F. (2019). Comparing coefficients across subpopulations in gaussian mixture regression models. *Journal of Agricultural, Biological and Environmental Statistics*, to appear.
- [21] Wang, C. M. and Lam, C. T. (1996). Confidence limits for proportion of conformance. *Journal of Quality Technology*, 28, 439-445.
- [22] Zimmer, Z., Park, D. and Mathew, T. (2016). Tolerance limits under normal mixtures: application to the evaluation of nuclear power plant safety and to the assessment of circular error probable. *Computational Statistics and Data Analysis*, 103, 304-315.