Comparing Distributions with the Probability of Agreement

by

Maziar Dadbin

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis we adapt the probability of agreement (PoA) methodology for the comparison of distributions. Most of the commonly used methods for comparing distributions are rooted in hypothesis testing where decisions are made using p-values. The proposed methodology, however, provides a more context-driven comparison by accounting for practically important differences. Two situations are considered: first, the one-sample comparison problem in which we have observed one sample and interest lies in determining whether the sample comes from a given known distribution. Second, we consider the two-sample comparison of distributions in which we have observed two independent samples and interest lies in determining whether these samples have the same distribution. The Horvitz-Thompson estimator is used to estimate the cumulative distribution function(s) corresponding the sample(s) under comparison and the asymptotic normality of the Horvitz-Thompson estimator is used to estimate the PoA. Confidence intervals (CIs) are also determined for the estimated PoA so as to quantify estimation uncertainty. We develop two methods for calculating CIs: one based on asymptotic normality and the delta-method and the other based on the bootstrap. To illustrate the application and interpretation of the methodology, we consider both real world and simulated examples. We also conduct a simulation study that evaluates the bias and variance of the PoA estimator as well as the coverage of the associated CIs. Finally we propose the relative density methodology as a graphical supplement that provides further information about the similarities and differences between the distributions under comparison. In summary, the contributions of this thesis are (1), the generalization of the PoA methodology to the one- and two-sample comparison of distributions, and (2), the suggestion of using the relative density and the PoA methodologies in tandem to gain more thorough information about the similarities and differences between the distributions under comparison.

Acknowledgements

First, I would like to express my sincerest appreciation to my supervisor, Professor Nathaniel Stevens who taught me how to do academic research. Thank you for you support, correcting my mistakes, and teaching me attention to detail. I also wish to dearly thank my wife and my parents, who were supportive of me in this journey.

Table of Contents

List of Tables

List of Figures

Part I

One-Sample Problem

Chapter 1

Introduction

1.1 The Problem

The aim of this thesis is to develop a new method for the common problem of comparing distributions. This thesis is organized into two parts. In Part I we will consider the goodness-of-fit problem, in which we investigate whether a known distribution is a good fit for an observed sample that has been drawn from a population. We refer to this as the "one-sample problem". In Part II, we focus on the problem of comparing the distributions of two populations from each of which we have an observed sample. We refer to this as the "two-sample problem". The common goal of both Part I and Part II of this thesis is to understand if there is a difference between the distributions under comparison, and if there is, to characterize the nature of that difference. Both problems are among the most common problems in statistical science and have a wide variety of applications in the real world. One can easily think of countless examples for which interest lies in investigating the distribution of one or more variables. For instance, one may be interested in the distribution of the age of patients suffering from a disease, or the distribution of housing prices in the past year, or even the distribution of the number of vehicles passing a red light camera during different times of the day. In these kinds of one-sample problems one may be interested in fitting a parametric model in order to describe the variable probabilistically. This then requires determining whether a family of distributions (such as the normal, gamma, or exponential) is a good fit to a given sample. Therefore, one can see that applications of the problem of goodness-of-fit are endless; in any univariate parametric statistical analysis, evaluating the goodness of fit of a model is relevant.

Throughout the thesis, we will use a real-world dataset to showcase our methodology and to illustrate how the analyses should be conducted. We will illustrate both the one- and two-sample methods in the context of this example dataset. The data we will be using is the result of a test known as "The Programme for International Student Assessment" (PISA) which is a test given to 15-year-old students from around the world every three years. The PISA test evaluates the students' performance in mathematics, reading, and science and the goal is to compare the performance of students from different parts of the world. The dataset we will be using however, only contains reading scores for American students. It also contains information about the demographics and schools of the students although we focus attention on gender and the reading scores of the students. The interested reader can consult Kaggle[1] to access the dataset and obtain a more in-depth description of the other variables recorded.

[1] https://www.kaggle.com/econdata/pisa-test-scores

The dataset consists of 1,791 female and 1,872 male students. Below, in Figure 1.1, we can see the histogram of all 3,663 students' reading scores. One may be interested in determining, for example, whether the students' reading scores follow a normal distribution. If a hypothesis test is used to answer this question, the answer will be a simple "yes" or "no". However, a more informative analysis would reveal for which scores the two distributions are similar and for which they are different. Such an analysis can be used to examine whether the distributions are similar or different, and if they are different, whether this difference arises in the right tail, the left tail, the middle, or some combination of these.

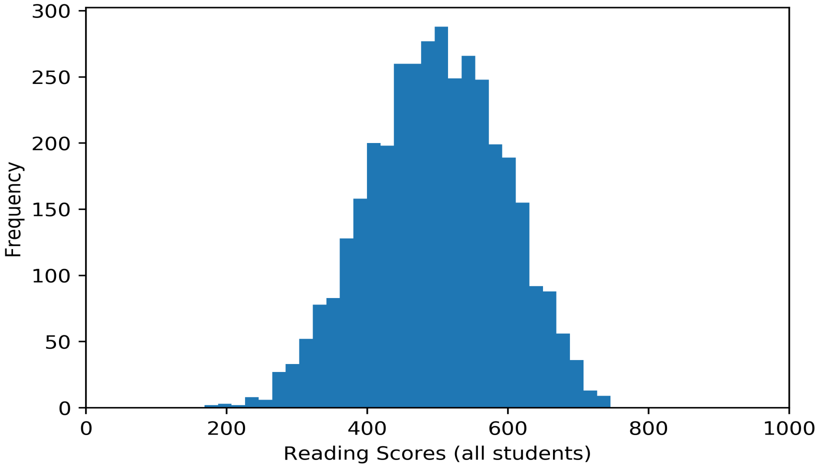


Figure 1.1: Histogram of the students' reading scores

We now introduce the one-sample goodness-of-fit problem mathematically. Consider a population $\mathcal{P}$ of size N with true cumulative distribution function (CDF) F and true probability distribution function (PDF) f. Also, consider a known distribution with CDF G and corresponding PDF g. Assume that we have observed a sample $\mathcal{S}$ of n units from

$\mathcal{P}$ on which we observe $\{x_1, x_2, \ldots, x_n\}$. Our objective is to compare F with G or, equivalently, to compare f with g. Because we do not know F or f, we must first estimate them from the sample data, obtaining $\hat{F}$ and $\hat{f}$. Note that we distinguish between estimates and estimators in our notation. While $\tilde{F}(x)$ is a random variable, $\hat{F}(x)$ is a real-valued number which is our estimate of $F(x)$. Note that throughout the thesis we refer to F as the comparison distribution and G as the reference distribution.

In Section 1.2 we review several traditional solutions for the goodness-of-fit problem. In particular, we present an overview of the general procedure common to all of the methods and then discuss some drawbacks. Then, in Section 1.3, we briefly introduce our proposed solution to the goodness-of-fit problem and we discuss its advantages relative to existing methods.

1.2 Existing Methods

The problem of comparing distributions has been at the forefront of academic research and practical data analysis since at least the first half of the twentieth century and so there exist a variety of methods for comparing distributions. Among the oldest and the most famous methods are the Shapiro-Wilk test (Shapiro and Wilk, 1965), the Kolmogorov-Smirnov test (Kolmogorov, 1933), the Anderson-Darling test (Anderson and Darling, 1952; 1954), and the Cramér–von Mises test (Cramér, 1928; von Mises, 1928). Both the Anderson-Darling and Cramér–von Mises statistics are special cases of a class of statistics called quadratic EDF statistics (Stephens, 1986) which are relevant in tests based on the empirical distribution function (EDF). The Shapiro-Wilk test is a test only for normality in the one-sample case i.e., when interest lies in comparing the distribution of an observed sample with the normal distribution. Each of the other tests, however, can be used for both one-sample and two-samples problems. We review the one-sample versions here and discuss their two-sample counterparts in Chapter 5. These tests are based on a standard hypothesis test where the null hypothesis is the equality between the true distribution of the sample and the reference distribution. More precisely, following the notation introduced in Section 1.1, the hypothesis statement is

$$H_0 : F = G \quad vs. \quad H_A : F \neq G$$

or equivalently

$$H_0 : f = g \quad vs. \quad H_A : f \neq g.$$

The decision to reject or not reject the null hypothesis is made by calculating a test statistic and a p-value, and then comparing the p-value with a significance threshold. To better understand the procedure of these traditional hypothesis tests, we will now look at the above-mentioned tests in more detail.

First, consider the Shapiro-Wilk test for normality. Given an observed sample $\{x_1, \ldots, x_n\}$, the Shapiro–Wilk test statistic is given by

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $x_{(i)}$ is the i^{th} order statistic, (i.e., the i^{th} smallest number in the sample), and $\bar{x}$ is the sample mean. The coefficients a_i are given by

$$(a_1, \ldots, a_n)^T = \frac{\boldsymbol{m}^T \boldsymbol{V}^{-1}}{(\boldsymbol{m}^T \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \mathbf{m})^{1/2}},$$

where the $n \times 1$ vector $\boldsymbol{m} = (m_1, \ldots, m_n)^T$ is composed of the expected values of the order statistics of n independent and identically distributed (IID) standard normal random variables, and $\boldsymbol{V}$ is the $n \times n$ covariance matrix of the normal order statistics (Davis and Stephens, 1977). Values for $a_1, a_2, \ldots, a_n$ are tabulated in Shapiro and Wilk (1965) for different values of n. The null distribution of the test statistic W does not have a closed form but the original paper (Shapiro and Wilk, 1965) provides a table of the quantiles of the null distribution for sample sizes smaller than 50. Note that it can be shown that if the sample is indeed from a normal distribution then the numerator and denominator of W are both, up to a constant, estimates of the normal variance σ^2. So, if H_0 is true, we would expect W to be close to 1. For non-normal samples, however, these quantities are not in general estimates of the same thing and so values very different from 1 provide evidence against H_0. For more details about this test, see Shapiro and Wilk (1965).

Next we consider the Kolmogorov-Smirnov (KS) test. The test statistic quantifies a distance between the CDF of the known reference distribution and the empirical cumulative distribution function (ECDF) of the sample which, given the observed data $\{x_1, \ldots, x_n\}$, is defined as

$$\hat{F}_n(x) = \sum_{u \in \mathcal{S}} \frac{\mathbb{I}[x_u \leq x]}{n}, \tag{1.1}$$

where $\mathbb{I}[x_u \leq x]$ is the indicator function which equals 1 if $x_u \leq x$ and is equal to 0 otherwise. The exact form of the KS test statistic is

$$D_n = sup_{x \in \mathcal{A}} |\hat{F}_n(x) - G(x)|,$$

where $\hat{F}_n(\cdot)$ is the ECDF corresponding to the sample $\mathcal{S}$, $\mathcal{A}$ is the support set of $\hat{F}_n(x)$, and $G(\cdot)$ is the known CDF. This statistic quantifies the distance between the ECDF and the known CDF at the point where they are farthest away from each other. If G is continuous and the null hypothesis is true, $\sqrt{n}D_n$ asymptotically follows the Kolmogorov distribution which is defined as the distribution of the random variable $K = \sup_{t \in [0,1]} |B(t)|$ where $B(t)$ is a Brownian bridge. The CDF of K is given by

$$P(K \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{i=1}^{\infty} e^{\frac{-(2i-1)^2 \pi^2}{(8x^2)}}.$$

The form of the KS test statistic and its asymptotic distribution under the null hypothesis was published by Kolmogorov (1933) and quantiles of the distribution were tabulated by Smirnov (1948). In summary, once the KS test statistic is calculated, one may reject the null hypothesis at level α if $\sqrt{n}D_n > K_\alpha$ where K_α is the value of the Kolmogorov distribution such that $P(K \leq K_\alpha) = 1 - \alpha$. Note however that in the case that the comparison is against a family of distributions instead of a completely specified one, the parameters of G should be estimated using the data $\{x_1, \ldots, x_n\}$. In such a case the above critical values are no longer valid and some modifications are necessary. In the special case of an exponential family distribution, details about the required changes to the test statistic and the critical values have been published (Pearson and Hartley, 1972).

To describe the Anderson-Darling and Cramér-von Mises tests we define the class of quadratic EDF statistics (Stephens, 1986) by

$$Q_n = n \int_{x \in \mathcal{A}} (\hat{F}_n(x) - G(x))^2 w(x) dF(x),$$

where $w(x)$ is a non-negative weight function which should be chosen by the experimenter to put more weight on those values of $\hat{F}_n(x) - G(x)$ where the test is desired to have more sensitivity. As before $G(\cdot)$ denotes the known CDF and $\hat{F}_n(\cdot)$ is the ECDF corresponding to the observed sample. If we choose $w(x) = 1$ all values of $x \in \mathcal{A}$ are given equal weight and the resulting statistic corresponds to the Cramér–von Mises test (Cramér, 1928; von Mises, 1928). Choosing $w(x) = [G(x)(1 - G(x))]^{-1}$ gives rise to the Anderson-Darling test statistic (Anderson and Darling, 1952; 1954). The intuition behind this choice of $w(x)$ can be explained as follows. For any given value of $x \in \mathcal{A}$, $n\tilde{F}_n(x)$ is a binomial random variable where the probability of success is $F(x)$, the true CDF of the population from which the sample was drawn. To see this, notice that to calculate $\hat{F}_n(x)$ we are effectively counting the number of data points that are less than or equal to x, and the probability that a data point is less than or equal to x is $F(x)$. Under the null hypothesis $H_0 : F(x) = G(x)$ the probability of success would be $G(x)$ and the variance would be $G(x)(1 - G(x))$. Therefore, by choosing $w(x) = [G(x)(1 - G(x))]^{-1}$, $\hat{F}_n(x) - G(x)$ is divided by its variance over the entire range of x, under the null hypothesis. This makes the test more sensitive to differences with small variance and less sensitive to differences with large variance. The quantiles of the distribution of the Anderson-Darling statistic are not available for small sample sizes. However, the asymptotic quantiles are given by Anderson and Darling (1954) (there the authors call them *significance points*). Note that these quantiles are valid when we are using a completely specified distribution function G as the reference distribution. However, when we want to test the sample against a distribution with unknown parameters that must be estimated with $\{x_1, \ldots, x_n\}$, the quantiles in Anderson and Darling (1954) are no longer valid. In this case, Andreson-Darling quantiles depend on the specific distribution G and also on the method of estimating the parameters.

It is interesting to note that despite the age of these methods (which date back to the mid-1900s) they still appear to be the most commonly used for these types of problems. To justify this claim and the corresponding lack of a contemporary literature review, in Table 1.1 we present the number of citations of the seminal Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk papers across different years and in aggregate. As we can see, the number of citations steadily increases over time, with many citations in very recent years. This suggests that no new methods have been developed in recent years that are used more widely than these traditional ones.

One issue with each of these hypothesis tests is that they begin with the assumption that the characteristics under comparison are the same, and evidence is sought to disprove this.

There are two possibilities, either we find enough evidence to reject the null hypothesis and conclude that the characteristics are not the same or we fail to find enough evidence to reject the null hypothesis. However, absence of evidence cannot be considered evidence of absence, meaning that in the case of failing to reject the null hypothesis, one cannot conclude that the null hypothesis is correct. Therefore, in applications for which we aim to prove the equality of two characteristics, traditional hypothesis testing is not appropriate. Equivalence testing is an alternative approach designed precisely for this problem. Wellek (2010) provides a thorough overview. In equivalence testing, we begin with the opposite assumption, that the two statistical characteristics under comparison are different. Then, we look for evidence of equivalence. Equivalence testing is commonly used in many fields including bioequivalence (Karalis and Macheras, 2012; Patterson and Jones, 2017) clinical psychology (McKay, 2008), and industrial engineering (Anderson-Cook and Borror, 2016; Szarka, 2014). Another important consideration in equivalence testing is that there is no need for the equivalence to be exact, meaning that one may consider some small difference between the characteristics under comparison to be practically negligible. More precisely, one may define a margin δ and an interval $(-\delta, \delta)$ within which differences may be con-

Year	Kolmogorov (1933)	Anderson and Darling (1952)	Shapiro and Wilk (1965)
$\leq$ 2000	358	589	2360
2001-2005	163	195	1080
2006-2010	274	523	2080
2011-2015	731	992	4490
2016-2020	1210	1230	7870
All Time	2791	3596	18208

Table 1.1: Citation counts for common statistical tests.

sidered practically unimportant. As long as the two characteristics are within $\pm\delta$ of each other they may be considered practically equivalent. See Wellek (2010) for a variety of one-sample goodness-of-fit equivalence tests.

It is important to emphasize that, except for the points just mentioned, traditional hypothesis tests and equivalence tests have the same procedure: in both cases we calculate a test statistic and then a p-value which is compared to a significance threshold, to decide whether to reject or not reject the null hypothesis. As we discuss in the next section, reliance on p-values can be problematic. In particular, we discuss limitations associated with p-values and we introduce a new methodology for comparing distributions that is not rooted in hypothesis testing and therefore does not suffer from these limitations.

1.3 An Alternative Approach

Here we discuss several problems associated with using hypothesis tests (either traditional or equivalence-based) for the purpose of comparing distributions. First, all of the information from either type of hypothesis test is summarised with a single number: the p-value. This represents a loss of information in the sense that even if the null hypothesis is rejected, the p-value on its own provides no information about how the two distributions disagree. For example, the distributions $F(x)$ and $G(x)$ may be similar for x in the middle of the distributions' support set but differ drastically in the tails or vice versa, but there is no way to determine this with just the p-value. On the other hand, in the case that we get a large p-value, we can neither reject nor accept the null hypothesis.

The traditional hypothesis test also has a practical issue related to sample size: no matter how similar the two distributions are, a sufficiently large sample size will result in an arbitrarily small p-value and hence the rejection of the null hypothesis. This problem is not limited to the case of comparing two distributions, it may occur in any hypothesis testing setting. This is not ideal if one prefers to ignore very small and practically unimportant differences. Equivalence tests sensibly avoid this issue by explicitly accounting for practical significance; a consideration we build into our proposed methodology.

A very real and practical problem associated with hypothesis testing (of both types) is that the interpretation of the p-value is complicated and non-trivial. Many researchers tend to interpret the p-value as the probability of the null hypothesis being true, which is of course not correct. As a result, many non-specialists misinterpret and misuse the p-value. This issue is especially problematic because hypothesis testing is so popular and widely used. It has become so controversial that the American Statistical Association (ASA) published a statement regarding the widespread misuse of p-values and urged the statistical community to develop alternatives to traditional hypothesis tests that overcome the limitations of p-values (Wasserstein and Lazar, 2016). In response, *The American Statistician* (TAS) published a special issue containing 43 articles concerned with "moving to a world beyond $p < 0.05$" (Wasserstein et al., 2019). Therefore, there is substantial interest in the development of alternative methods that account for practical significance in addition to, or instead of, statistical significance.

The probability of agreement (PoA) has been proposed as a way of quantifying the similarities between two statistical characteristics (random variables), C_1 and C_2, taking into consideration the size of a practically important difference. Generally speaking, the PoA may be defined as

$$P(|C_1 - C_2| \leq \delta),$$

where δ is a constant representing the boundary between practically important and unimportant differences. Therefore the PoA is the probability that the difference between the characteristics under comparison is practically negligible. The primary contribution of this thesis is the adaptation of the PoA methodology to facilitate the comparison of CDFs. Here, we take the characteristics of interest C_1 and C_2 to be CDFs and we develop methodology for point and interval estimates of the PoA in this new context. In Part I of the thesis one of these CDFs is known and the other will be estimated based on sample data; in Part II both CDFs will be estimated. Thus, the PoA in Part I will be defined as

$$\theta(x) = P(-\delta \leq \tilde{F}(x) - G(x) \leq \delta), \tag{1.2}$$

where δ is a constant and $\tilde{F}(\cdot)$ is the estimator of the CDF of the observed sample and $G(\cdot)$ is the known CDF. Notice that any difference smaller than δ between the two CDFs at point x will be considered practically negligible. Thus the PoA calculated at point x is the probability that the two CDFs are practically equivalent at point x. The value of δ depends on the context of the problem and should be chosen by someone who has insight into the specific application at hand. The PoA does not suffer from the abovementioned issues of hypothesis testing and it is also very straightforward to interpret. The

PoA methodology was first introduced in the context of measurement system comparison (Stevens et al., 2017; 2018a) and has since been applied to the comparison of population reliabilities (Stevens and Anderson-Cook, 2017a;b), the comparison of confirmation runs to previous experimental results (Stevens and Anderson-Cook, 2019), the comparison of generalized linear response surfaces (Stevens et al., 2019; 2018b), and the comparison of parametric (Stevens et al., 2020) and nonparametric (Stevens and Lu, 2020) survival functions.

In the next chapter, we describe the details of estimating the PoA in the one-sample case and we consider both simulated and real examples to illustrate how one should use this method and what information one can gain when conducting it. We will also explain the concept of relative density and show how one can use it to supplement the PoA analysis to gain a more in-depth understanding of the similarities and differences between the distributions under comparison. In particular, we use the PoA to compare two CDFs and we use the relative density to compare the corresponding PDFs. In Chapter 3 we present the results of a simulation study that investigates the performance and properties of our proposed methodology. That will end Part I of the thesis. In Part II we extend the PoA and relative density analyses to the problem of two-sample comparisons.

Chapter 2

Proposed Methodology for the One-Sample Scenario

Our proposed methodology consists of two parts. We suggest that when determining whether a sample can be modeled by a specific family of distributions (e.g., normal or gamma distributions) one first applies the probability of agreement (PoA) methodology as the primary analysis to quantify the (dis)similarity of the distributions under comparison. Second, for additional insight, we recommend supplementing this with an informative graphical visualization based on the concept of relative density. In Section 2.1, we generalize the PoA methodology so that it can be applied to the problem of comparing a known with an unknown (estimated) CDF. Recall the comparison of two unknown (estimated) CDFs is considered in Part II of the thesis. Then, in Section 2.2, we discuss the relative density and we show how it can be used to provide informative graphical summaries of the similarities and differences between the two CDFs under comparison. Finally, in Section 2.3, we provide several examples to illustrate how our proposed methodology should be conducted in practice and also to explain how one should interpret the results of the analysis.

2.1 The Probability of Agreement for One-Sample CDF Comparisons

2.1.1 The Probability of Agreement

In general, the PoA quantifies the similarities between two statistical characteristics, taking into account what size difference is or is not practically important. This "size" should be determined by the user and it may vary a great deal depending on the context. In this context the characteristics of interest are a known and an estimated CDF.

As in Chapter 1, we define $F(x)$ to be the true CDF of the population $\mathcal{P}$ from which we observed the sample, and let $\tilde{F}(x)$ be the corresponding estimator. Also let $G(x)$ be a known CDF. Then the PoA between $\tilde{F}(x)$ and $G(x)$ is defined as

$$\theta(x) = P(-\delta \leq \tilde{F}(x) - G(x) \leq \delta). \tag{2.1}$$

Note that the PoA is defined for all $x \in A$, the support set of $\tilde{F}(\cdot)$, which depends on the observed sample. One may interpret the interval $(-\delta, \delta)$ as a context-specific indifference region, meaning that as long as the difference between the CDFs is in this interval, we consider those CDFs to be practically equal. For instance if $\hat{F}(x) = 0.5$ and $G(x) = 0.52$, these are practically equivalent if $\delta = 0.05$ but not if $\delta = 0.01$. Therefore the PoA at point x is the probability that the CDFs, calculated at point x, are practically equivalent.

To be able to calculate the PoA at any point x we need to know the distribution of $\tilde{F}(x)$ which depends on the method of estimation. In this thesis, we will use the Horvitz-Thompson method to estimate the CDF of the observed sample. We use Horvitz-Thompson estimation because it allows us to account for different sampling protocols, because it also has desirable distributional properties that enable us to easily estimate the PoA, and because we regard samples as having been drawn from finite populations.

In the next section, we explain the Horvitz-Thompson method and then in Section 2.1.3, we show how one can estimate the PoA when we use the Horvitz-Thompson approach to estimate $F(x)$.

2.1.2 Horvitz-Thompson Estimator

The first step towards calculating the PoA is to estimate the unknown CDF of the population from which we have observed a sample. We must then determine the distribution of the estimated CDF to calculate the PoA. For the first step, we will use the method of Horvitz-Thompson (Horvitz and Thompson, 1952). As mentioned above this method accounts for the sampling design; different sampling designs will result in different estimates for the CDF. In the case of simple random sampling, i.e., when every unit in the population has the same chance of being selected into the sample, the Horvitz-Thompson estimate is the same as the empirical cumulative distribution function (ECDF) defined in Equation 1.1. However, the Horvitz-Thompson framework allows us to flexibly accommodate other sampling protocols if necessary.

Let $\mathcal{P}$ denote a population of size N and assume we have a sample $\mathcal{S} \subset \mathcal{P}$ of n units on which we observe $\{x_1, \ldots, x_n\}$. Also assume that F is the true CDF corresponding to $\mathcal{P}$ which is defined as

$$F(x) = \sum_{u \in \mathcal{P}} \frac{\mathbb{I}[x_u \leq x]}{N},$$

where $\mathbb{I}[x_u \leq x]$ is the indicator function defined as

$$\mathbb{I}[x_u \leq x] = \begin{cases} 1 & \text{if } x_u \leq x \\ 0 & \text{if } x_u > x \end{cases}.$$

Then the Horvitz-Thompson estimate of $F(x)$ is given by

$$\hat{F}_{HT}(x) = \sum_{u \in \mathcal{S}} \frac{\mathbb{I}[x_u < x]}{N\pi_u} \quad \text{for } x \in \mathcal{A}, \tag{2.2}$$

where $\pi_u = P(u \in \mathcal{S})$ is the inclusion probability for unit $u \in \mathcal{S}$ and $\mathcal{A}$ is the support set of $\hat{F}_{HT}(x)$. As already mentioned, we distinguish between estimates and estimators, and thus we denote the Horvitz-Thompson estimate of $F(x)$ by $\hat{F}_{HT}(x)$ and the Horvitz-Thompson estimator by $\tilde{F}_{HT}(x)$. The variance of $\tilde{F}_{HT}(x)$ is given by

$$Var\left[\tilde{F}_{HT}(x)\right] = \sum_{u\in\mathcal{P}}\sum_{v\in\mathcal{P}}(\pi_{uv} - \pi_u\pi_v)\frac{\mathbb{I}[x_u \leq x]}{N\pi_u}\frac{\mathbb{I}[x_v \leq x]}{N\pi_v},$$

where $\pi_{uv} = P(u \in \mathcal{S}, v \in \mathcal{S})$ is the joint inclusion probability for units $u, v \in \mathcal{P}$. This variance cannot be calculated as the whole population is typically not available. However, it can be estimated using again the Horvitz-Thompson method. The Horvitz-Thompson estimate of the above variance is given by

$$\widehat{Var}\left[\tilde{F}_{HT}(x)\right] = \sum_{u\in\mathcal{S}}\sum_{v\in\mathcal{S}}\left(\frac{\pi_{uv} - \pi_u\pi_v}{\pi_{uv}}\right)\frac{\mathbb{I}[x_u \leq x]}{N\pi_u}\frac{\mathbb{I}[x_v \leq x]}{N\pi_v}. \tag{2.3}$$

In order to calculate the PoA, we require the distribution of $\tilde{F}_{HT}(x)$. Although the exact distribution of $\tilde{F}_{HT}(x)$ is in general not known, Berger (1998) showed that

$$\frac{\tilde{F}_{HT}(x) - F(x)}{\sqrt{\widehat{Var}[\tilde{F}_{HT}(x)]}} \xrightarrow{D} N(0,1). \tag{2.4}$$

We use this asymptotic result to calculate the PoA, which is the focus of the next section.

2.1.3 Estimating the Probability of Agreement

Because interest lies in the difference between the two CDFs, the asymptotic approximation from Equation 2.4 can be used to derive the following result for the difference $\tilde{F}_{HT}(x)-G(x)$

$$\frac{\left(\tilde{F}_{HT}(x) - G(x)\right) - \left(F(x) - G(x)\right)}{\sqrt{\widehat{Var}[\tilde{F}_{HT}(x)]}} \xrightarrow{D} N(0,1). \tag{2.5}$$

Recall, $G(x)$ represents the reference distribution with which we are comparing the CDF of the sample and is therefore treated as a known constant. With this result we calculate the PoA as

$$\begin{aligned} \theta(x) &= P\left(-\delta \le \tilde{F}_{HT}(x) - G(x) \le \delta\right) \\ &\cong \Phi\left(\frac{\delta - (F(x) - G(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]}}\right) - \Phi\left(\frac{-\delta - (F(x) - G(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]}}\right), \end{aligned} \tag{2.6}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. It is important to acknowledge that the second equivalence in the above equation is an approximation due to the fact that in practice we have a finite sample size and hence the result in Equation 2.5 holds only approximately. By replacing $F(\cdot)$ with its Horvitz-Thompson estimate in Equation 2.6, we obtain the estimate of the PoA

$$\hat{\theta}(x) = \Phi\left(\frac{\delta - (\hat{F}_{HT}(x) - G(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]}}\right) - \Phi\left(\frac{-\delta - (\hat{F}_{HT}(x) - G(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]}}\right). \tag{2.7}$$

The PoA is then calculated and plotted against x for all $x \in \mathcal{A}$. The resulting PoA plot visualizes the dependence of agreement on x.

Because of the sampling variation inherent in the estimation of the PoA, we use confidence intervals (CIs) to communicate the uncertainty associated with the PoA estimate. We investigate two different approaches for constructing approximate pointwise CIs. In Section 2.1.4 we describe the details of constructing approximate CIs for the PoA using the bootstrap. In Section 2.1.5 we will use asymptotic normality and the delta method to determine the distribution of $\tilde{\theta}(x)$ and then use this distribution to calculate CIs for the PoA. In both cases we calculate and visualize the CI for each $x \in \mathcal{A}$.

2.1.4 Confidence Intervals Based on the Bootstrap

The bootstrap methodology was first published by Efron (1979). It is a technique that can be used to approximate the sampling distribution of a statistic through resampling. Here, we will use the bootstrap to construct approximate pointwise CIs for the PoA. Let $\mathcal{S}$ be the observed sample of n units. A bootstrap sample also contains n units, each of which has been selected randomly from $\mathcal{S}$. Note that this selection is done with replacement, and so a bootstrap sample is not simply a permutation of $\mathcal{S}$. Assume $\mathcal{S}_1, \ldots, \mathcal{S}_B$ are B bootstrap samples from $\mathcal{S}$. Note that B, the number of bootstrap samples, is a value that should be chosen by the user. Larger values for B result in a more precise estimate of the sampling distribution and hence lead to CIs with better coverage, but at the expense of higher computation time. Efron and Hastie (2016) suggest that $B = 2{,}000$ should be used for bootstrap-based CIs, and so we follow this recommendation.

In previous sections, we have described how the PoA at a given point $x \in \mathcal{A}$ is estimated from a sample $\mathcal{S}$, yielding $\hat{\theta}(x)$. Here we follow the exact same procedure for each of the bootstrap samples $\mathcal{S}_1, \ldots, \mathcal{S}_B$ and we calculate $\hat{\theta}_1(x), \ldots, \hat{\theta}_B(x)$ for every $x \in \mathcal{A}$. We use the standard deviation of $\hat{\theta}_1(x), \ldots, \hat{\theta}_B(x)$ to estimate the sampling variability of $\hat{\theta}(x)$, and we use it in our construction of CIs for $\theta(x)$. Note that here we consider pointwise CIs, not simultaneous ones, and so we calculate the standard deviation of $\hat{\theta}_1(x), \ldots, \hat{\theta}_B(x)$ for each $x \in \mathcal{A}$ separately. A bootstrap-based $(1 - \alpha) \times 100\%$ CI for $\theta(x)$ is given by

$$\left[\hat{\theta}(x) - z_{\alpha/2} \times SE[\hat{\theta}(x)], \hat{\theta}(x) + z_{\alpha/2} \times SE[\hat{\theta}(x)]\right],$$

where $SE[\hat{\theta}(x)]$ is the standard deviation of $\hat{\theta}_1(x), \ldots, \hat{\theta}_B(x)$ and $z_{\alpha/2}$ is the quantile of the standard normal distribution such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

We also considered "percentile method" bootstrap CIs (Efron and Hastie, 2016). However, with respect to coverage, the percentile method intervals were inferior to the naive normal theory interval defined above. Several amendments to the percentile method, such as bias-corrected intervals, have been proposed (Efron and Hastie, 2016). However, we did not consider them here because Stevens and Lu (2020) investigated them in a similar context and found that they were not helpful.

2.1.5 Confidence Intervals Based on Asymptotic Normality

The goal in this section is to determine the sampling distribution of the PoA at a given point x, and use it to calculate a CI for $\theta(x)$. Let us write the estimator version of Equation 2.7 as

$$\tilde{\theta}(x) \cong \Phi\left(U(\tilde{D}(x))\right) - \Phi\left(L(\tilde{D}(x))\right), \tag{2.8}$$

where

$$\tilde{D}(x) = \tilde{F}_{HT}(x) - G(x),$$

$$U(\tilde{D}(x)) = \frac{\delta - \tilde{D}(x)}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]}},$$

and

$$L(\tilde{D}(x)) = \frac{-\delta - \tilde{D}(x)}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]}}.$$

As we can see, $\tilde{\theta}(x)$ is a function of $\tilde{D}(x)$ because $\theta(x)$ is a function of $D(x)$. Using the distributional result for $\tilde{D}(x)$ given in Equation 2.5, we can use the delta method (Doob, 1935) to determine the asymptotic distribution of $\tilde{\theta}(x)$. According to the delta method, the variance of $\tilde{\theta}(x)$ is given by

$$\begin{aligned} Var[\tilde{\theta}(x)] &= \left[\frac{d\theta(x)}{dD(x)}\right]^2 Var[\tilde{D}(x)] \\ &= \left[-\frac{\phi(U(D(x))) - \phi(L(D(x)))}{\sqrt{Var\left[\tilde{F}_{HT}(x)\right]}}\right]^2 Var[\tilde{D}(x)] \\ &= \left[\phi(U(D(x))) - \phi(L(D(x)))\right]^2, \end{aligned}$$

where $\phi(\cdot)$ is the probability density function (PDF) of a standard normal random variable. Thus, by delta method we have

$$\frac{\tilde{\theta}(x) - \theta(x)}{[\phi(U(D(x))) - \phi(L(D(x)))]} \xrightarrow{D} N(0,1). \tag{2.9}$$

Using this asymptotic result we can build approximate pointwise CIs using the estimate $\hat{\theta}(x)$ and its corresponding standard error $\phi(U(\hat{D}(x))) - \phi(L(\hat{D}(x)))$. The approximate $100(1-\alpha)\%$ CI for $\theta(x)$ is given by

$$\left[\hat{\theta}(x) - z_{\alpha/2} \times \left(\phi(U(\hat{D}(x))) - \phi(L(\hat{D}(x)))\right), \hat{\theta}(x) + z_{\alpha/2} \times \left(\phi(U(\hat{D}(x))) - \phi(L(\hat{D}(x)))\right)\right],$$

where $z_{\alpha/2}$ is the quantile of the standard normal distribution such that $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$. Comparing this interval to the one developed in Section 2.1.4, we see that the only difference is in the standard error term. Whereas in the bootstrap approach the standard error is defined by the bootstrap standard deviation, here it is defined based on the delta method. In Chapter 3 we use simulation to compare and contrast these two approaches to confidence interval construction.

2.2 The Relative Density for One-Sample PDF Comparisons

In this section we review the relative density due to Parzen (1999) which we propose using as a graphical supplement to the PoA analysis described in Section 2.1. Unlike the PoA methodology which compares CDFs, the relative density methodology is based on a comparison of PDFs. In particular, the ratio of the PDFs is compared to the constant value 1. This ratio, which is called the *relative density function*, can be estimated directly from what is called *relative data*.

Below we define the relative density function, and we discuss how to estimate it and how to construct approximate CIs for it. In Appendix A, we briefly show how one may perform a PoA-style calculation to compare the relative density function to the constant value 1. However, for practical reasons (that are explained in the Appendix), we do not recommend using the PoA methodology for the relative density and we simply suggest that the relative density be used as a supplementary graphical tool while applying the PoA methodology described in Section 2.1.

Recall that $X \sim F$ and $Y \sim G$ denote the comparison and the reference distribution. The relative distribution comparing F to G is defined as the distribution of the random variable $R = G(X)$. A realization of R, denoted r ($0 \leq r \leq 1$), is called relative data. Let H and h be the CDF and PDF of R respectively. One can easily derive the distribution of R as follows

$$H(r) = P(R \leq r) = P(G(X) \leq r) = P(X \leq G^{-1}(r)) = F(G^{-1}(r)).$$

Notice that because $G(\cdot)$ is a CDF, it is an increasing and one-to-one function and thus it is invertible. The PDF h may be found by taking derivatives of the CDF H with respect to r, which by chain rule is given by

$$h(r) = \frac{dH(r)}{dr} = \frac{dF(G^{-1})(r)}{dr} = \frac{f(G^{-1}(r))}{g(G^{-1}(r))}. \tag{2.10}$$

Note that if $f = g$ then $h(r) = 1$ for all $0 \leq r \leq 1$. Therefore, to gain information about the possible similarities and differences between f and g, we can compare $h(r)$ with 1 for each $r \in [0, 1]$. If $h(r) > 1$ ($h(r) < 1$), it indicates that at point $G^{-1}(r)$ we have larger (smaller) density in the comparison distribution compared to the reference distribution. To be able to use the relative density in practice, we need to first estimate the function $h(\cdot)$.

Recall that here in Part I of the thesis, we treat $g(\cdot)$ as a known function. Therefore, in Equation 2.10 the only piece that needs estimation is $f(\cdot)$. A naive approach to estimate the relative density is to first estimate $f(\cdot)$ and substitute it into Equation 2.10 to obtain an estimate for $h(\cdot)$. This would work in principle in the one-sample case, but if $g(\cdot)$ also needs to be estimated, as in the two-sample case, this plug-in estimation breaks down. As such, for a unified approach, Thas (2010) suggests it is preferable to estimate $h(\cdot)$ directly from the relative data, which for $i = 1, \ldots, n$ is defined as

$$r_i = G(x_i),$$

where $\{x_1, \ldots, x_n\}$ is the observed sample from the comparison distribution. As the x_i's are independent and identically distributed (IID) realizations from $f(\cdot)$, the r_i's are also IID realizations from $h(\cdot)$. As such, we can estimate $h(\cdot)$ using a non-parametric density estimation method applied to the relative data. There are many such methods that can potentially be used. To name a few, there are estimation using histograms, kernel density estimators (KDE), and orthogonal series density estimators which have been studied by Parzen (1983), Eubank et al. (1987), Alexander (1989), Cwik and Mielniczuk (1993), Mielniczuk (1992), Li et al. (1996), and Parzen (1999).

In the relative density context it is common to use the KDE method to estimate $h(\cdot)$ because the asymptotic distribution of this estimator is known, allowing one to easily construct approximate confidence intervals for $h(\cdot)$. The KDE of the function h is given by

$$\hat{h}_n(u) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{u - r_i}{b_n}\right), \tag{2.11}$$

where b_n is the bandwidth and $K(\cdot)$ is a kernel function satisfying the following conditions

$$\int_{-1}^{1} K(u)du = 1, \quad \int_{-1}^{1} uK(u)du = 0, \text{ and } \quad \int_{-1}^{1} u^2 K(u)du = \sigma_K^2 > 0.$$

Note that in general a kernel function does not have to satisfy the above conditions but Handcock and Morris (1999) recommend using such a kernel in the context of relative density estimation because doing so facilitates the asymptotic normality of the KDE which is the basis for confidence intervals for $h(\cdot)$. Here we adopt the commonly used biweight kernel

$$K(u) = \begin{cases} \frac{15}{16}(1-u^2)^2 & \text{if } u \in [-1,1] \\ 0 & \text{if } u \notin [-1,1] \end{cases}. \tag{2.12}$$

Another important consideration in kernel density estimation is the choice of the bandwidth b_n. Many bandwidth selection methods exist. For instance, the normal reference rule (Sheather and Jones, 1991), Scott's rule (Scott, 1992), and Silverman's rule (Silverman, 1986). Through simulation and empirical investigation we found the normal reference rule works well for our purpose. As such, when building relative density plots in this thesis we calculate the bandwidth as

$$b_n = 2.778 \times \min\left(s, \frac{IQR}{1.349}\right) n^{-0.2}, \tag{2.13}$$

where s and IQR are respectively the standard deviation and interquartile range of the relative data $\{r_1, \ldots, r_n\}$.

The asymptotic distributional properties of the estimator 2.11 are proved by Handcock and Morris (1999). In particular, we have

$$\frac{\tilde{h}_n(r) - h(r)}{\sqrt{\frac{h(r)R(K)}{nb_n}}} \xrightarrow{D} N(0,1), \tag{2.14}$$

where $R(v) = \int_{-\infty}^{\infty} v(x)^2 dx$. This asymptotic result is used to build approximate CI's for the relative density. The approximate $100(1-\alpha)\%$ CI for $h(r)$ is

$$\left[\hat{h}_n(r) - z_{\alpha/2}\sqrt{\frac{\hat{h}_n(r)R(K)}{nb_n}}, \hat{h}_n(r) + z_{\alpha/2}\sqrt{\frac{\hat{h}_n(r)R(K)}{nb_n}}\right],$$

where $h(r)$ has been replaced by $\hat{h}_n(r)$ in the standard deviation term of Equation 2.14. Note that the larger the sample size the better the approximation. In small sample scenarios (i.e., $n < 30$) Handcock and Morris (1999) suggest using the bootstrap to estimate the sampling distribution of the estimated relative density.

2.3 Examples

In this section we will provide a real-world example as well as two simulated examples to demonstrate the process of comparing distributions in practice and, in particular, illustrate how one should interpret the PoA and the relative density plots.

2.3.1 PISA Dataset

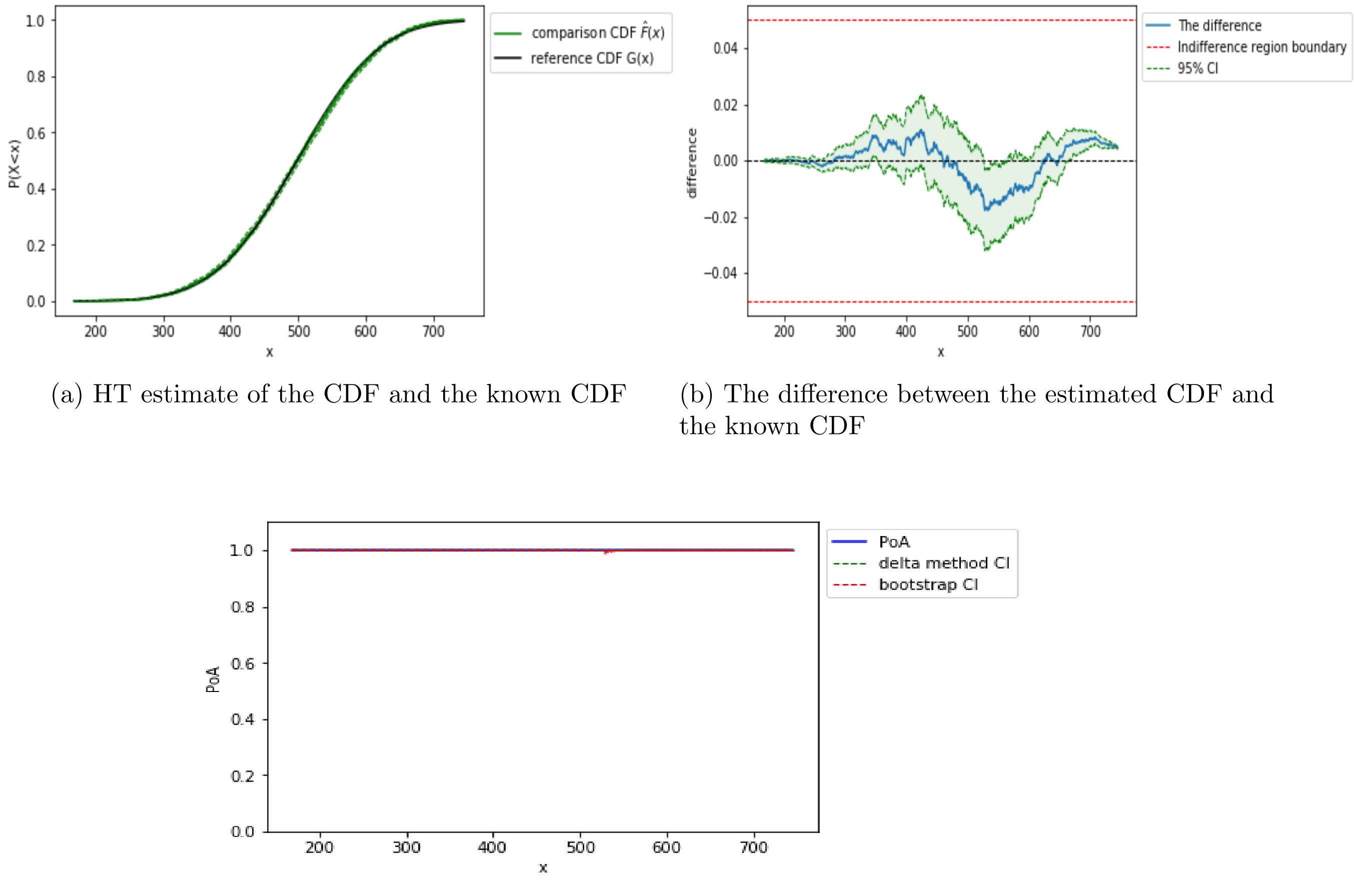
We begin by illustrating the methodology on the PISA dataset that we introduced in Section 1.1. As we explained there, the dataset contains reading scores of 15-year-old students. Looking at the histogram of the students' reading scores shown in Figure 1.1, one may wonder whether this data can be well-modeled by a normal distribution. Here we formally investigate this. Note that here we will specify the mean and variance of the normal distribution (G) to be the sample mean and the sample variance of $\{x_1, \dots x_n\}$. Therefore, the precise question we are trying to answer is whether the population of reading scores from which our sample was obtained follow a normal distribution with mean and variance equal to those of the sample.

We have plotted the estimated and known CDFs as well as the difference between them in Figure 2.1a and 2.1b respectively. Note that to be able to apply the Horvitz-Thompson method we assume for illustration that the $n = 3{,}663$ observed reading scores come from a population of size $N = 20{,}000$. We assume that simple random sampling has been used and so the marginal and joint inclusion probabilities are $\pi_u = \frac{n}{N} = \frac{3663}{20000}$ and $\pi_{uv} = \frac{n(n-1)}{N(N-1)} = \frac{3663 \times 3662}{20000 \times 19999}$. In Figure 2.1c we can see the PoA between the Horvitz-Thompson estimate of the population CDF and the normal CDF with mean and variance equal to the sample mean and variance. Recall that the relation between the PoA plot and the difference plot is that at any point x, the estimated PoA $\hat{\theta}(x)$ estimates the probability that the absolute difference is less than δ. We have used $\delta = 0.05$ for illustration. This choice implies that we believe an absolute difference of at most 0.05 between the CDFs can be considered practically unimportant.

As we can see in Figure 2.1c, the PoA is quite large for all values of x, which is unsurprising given the agreement between $\hat{F}(x)$ and $G(x)$ depicted in Figures 2.1a and 2.1b. These plots illustrate that with $\delta = 0.05$ the reading scores can be well-modeled by the normal distribution.

However, it is important to acknowledge that the conclusion of the PoA analysis depends strongly on the value of δ. To illustrate the effect the value of δ has on the estimated PoA, in Figure 2.2a we provide a contour plot of the PoA as a function of δ and x. Figure 2.2b shows several PoA curves corresponding to different values of δ. As evidenced by Figures 2.2a and 2.2b, the value of δ dramatically impacts the estimated PoA. As we can see in both Figures 2.2a and 2.2b, even in this example, when we see very strong similarities between two CDFs under comparison, choosing δ small enough will yield very small values of PoA. Likewise, choosing large enough δ values will yield PoA values close to 1 for all values of x. Therefore, given the impact the value of δ has on the outcome of the PoA analysis, we suggest that a practitioner should carefully choose a value for δ that is appropriate in their specific context.

Like the value of δ, how large $\hat{\theta}(x)$ must be to conclude practical equivalence between $\hat{F}(x)$ and $G(x)$ should also be determined by the practitioner. For some practitioners, $\hat{\theta}(x) > 0.95$ for all x might be required, whereas others may require the average $\hat{\theta}(x)$ value to be larger than 0.9, for example.

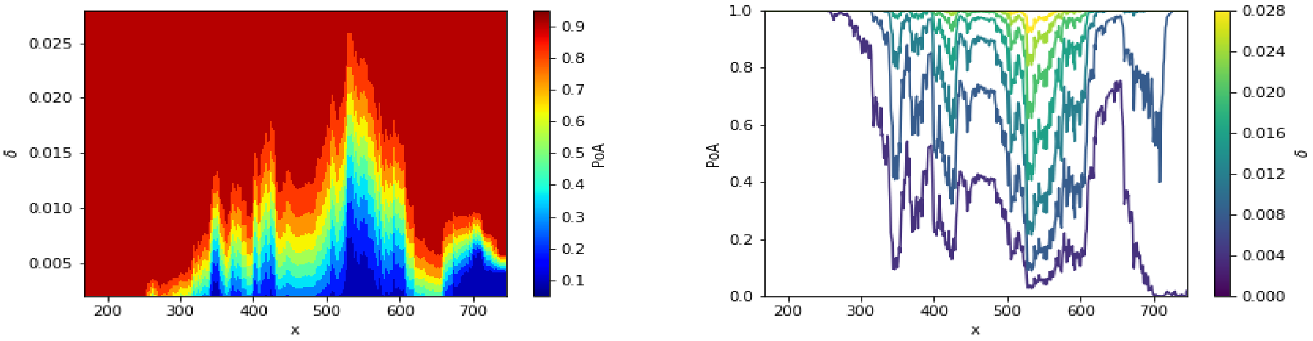


(a) HT estimate of the CDF and the known CDF

(b) The difference between the estimated CDF and the known CDF

(c) The PoA

Figure 2.1: (a) Horvitz-Thompson estimate of the CDF of the observed data and the normal CDF with mean and variance equal to those of the sample. The 95% CI is constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated and the known CDF. The 95% CI is constructed the same way as in (a). (c) Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.



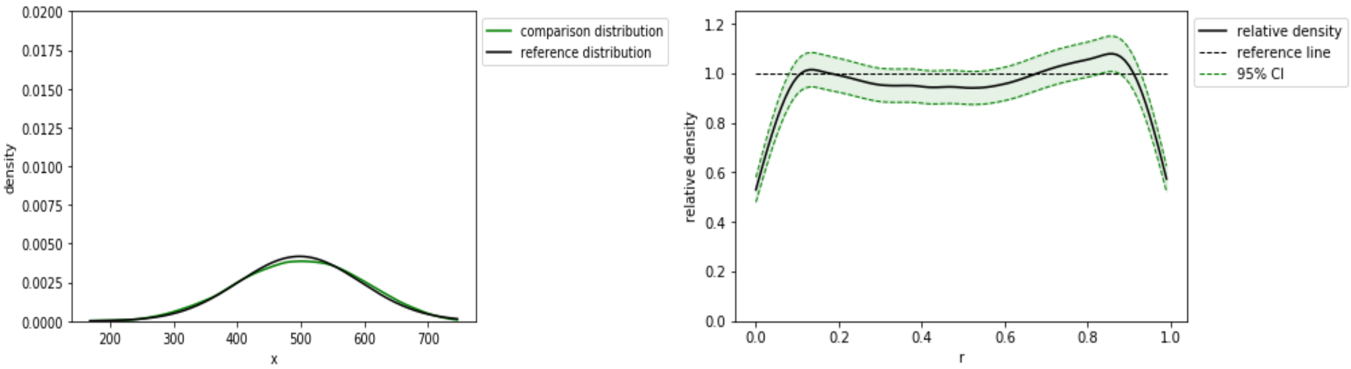
(a) contour plot of PoA

(b) PoAs with different δ values

Figure 2.2: (a) The contour plot of PoA. (b) Several PoA plots constructed with different values of δ

To gain further insight into the comparability of the distributions, we also suggest looking at the relative density plot as a supplementary graphical tool. In Figure 2.3a we can see that the KDE of the PDF of the sample is overlapping the normal reference PDF with mean and variance equal to those of the sample for almost all values of x. This is aligned with what we see in 2.3b because, apart from the tails, the relative density is quite close to 1. Recall that the relative density at point r is the ratio of the comparison distribution to the reference distribution, at the r^{th} percentile of the reference distribution. Therefore a relative density plot (i.e, a plot of $h(r)$ versus r for $r \in [0, 1]$) and a plot of the PDFs are not comparable pointwise. In fact the value of $h(\cdot)$ at point r corresponds to the ratio of the PDFs at point $G^{-1}(r)$. Therefore, the range of values of x where the comparison density is below (above) the reference density, corresponds to the range of values of r where the relative density is smaller than 1 (larger than 1). However, the fact that the value 1 is contained in the relative density's 95% pointwise CIs for almost all values of $0 < r < 1$, suggests good agreement and that the normal family should be a good model for our data.

It is well known that the KDE in this context has a downward bias near the boundaries at 0 and 1. That is why we observe values below 1 for values of r near 0 and 1 in Figure 2.3b. Intuitively speaking, this happens because the kernel density estimator does not understand the boundaries and acts as if the relative density is zero outside $[0, 1]$ while it is in fact not defined there. In Chapter 7 we discuss potential remedies for this downward bias of the KDE.



(a) The estimated and the known PDF

(b) The relative density

Figure 2.3: (a) The KDE of the PDF of the observed sample and the normal PDF with mean and variance equal to those of the sample. (b) The KDE of the relative density plot. The 95% CI is constructed based on asymtotic normality of the KDE.

2.3.2 Simulated Example (Normal Data)

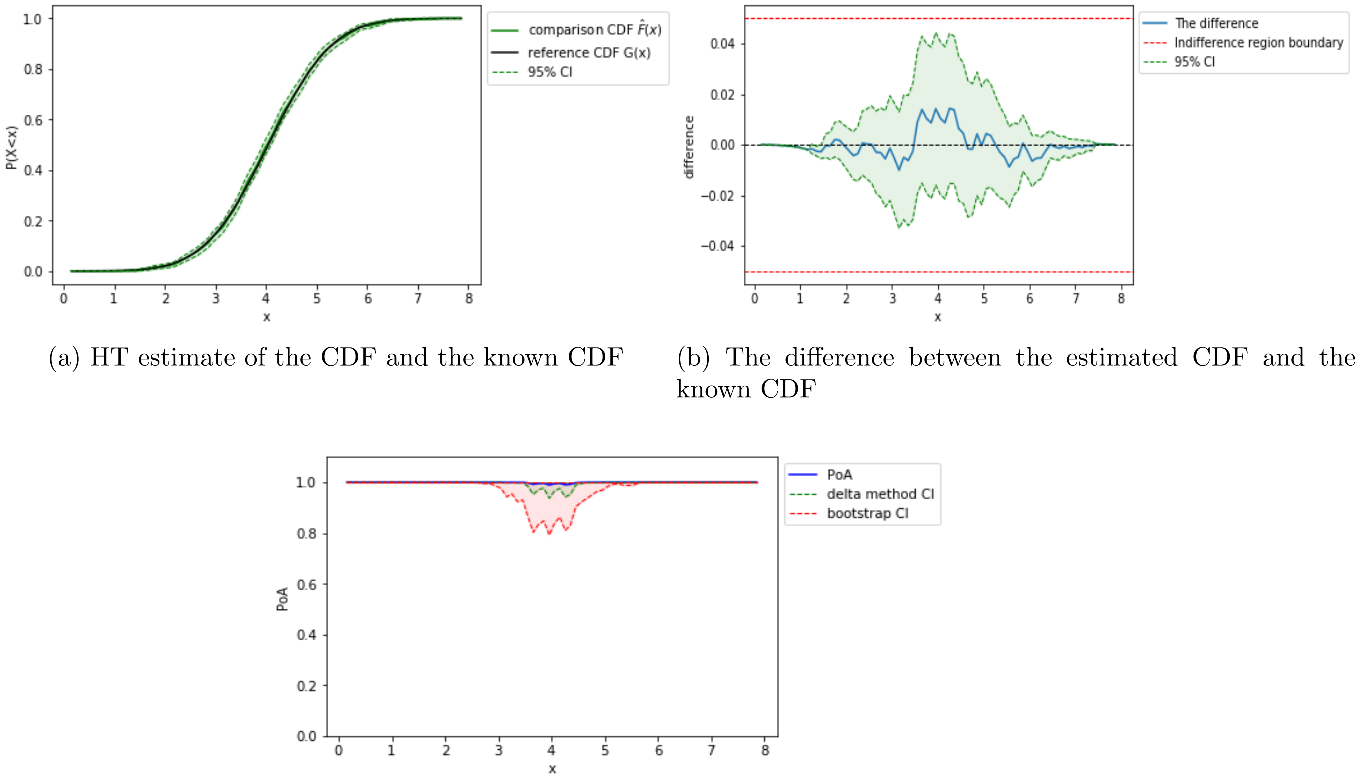
In the PISA dataset example showcased in the previous section, we did not know whether the reading scores truly followed a normal distribution or not. In this section we consider a simulated example to make sure that our methodology works as expected when the distributions under comparison really are the same. We will also illustrate the effect of sample size on the PoA which can only be done in a simulated scenario.

First assume our sample is a simple random sample of $n = 1{,}000$ observations from a population of size $N = 20{,}000$ which is itself obtained by IID draws from $N(4, 1)$. We use the proposed methodology to determine whether the sample can indeed be modeled by a normal distribution. Once again note that, like in the previous section, we specify the mean and variance of the reference distribution as the sample mean and sample variance of $\{x_1, \ldots, x_n\}$. Then we estimate the PoA between the Horvitz-Thompson estimator of the CDF for the observed sample and the normal CDF with mean and variance equal to sample mean and sample variance of the $n = 1{,}000$ sample observations. Figures 2.4a and 2.4b both indicate good agreement between the estimated CDF and the normal CDF with mean and variance equal to those of the sample. Note that the difference plot is contained in the indifference region showed by two horizontal lines at $\pm\delta = \pm 0.05$. This suggests the normal distribution is a good model for these data. In Figure 2.4c we use the PoA plot (with $\delta = 0.05$) to formalize this assessment. We see the PoA equals 1 for all value of x suggesting that for each x, the sampling distribution of the difference $\tilde{F}_{HT} - G(x)$ is entirely contained within $\pm\delta$ and hence that the sample data can indeed be well-modeled by a normal distribution. Of course, this is unsurprising since these data were drawn from a normal distribution. However, the CIs, especially the bootstrap CI, show some uncertainty in this conclusion around $x = 4$. The difference plot in Figure 2.4b offers insight into why this is: the wide CI around $x = 4$ indicates uncertainty in the estimated difference here, which is transmitted into uncertainty about the probability that the difference lies within $\pm\delta$.

Note that, as in the previous section, we have chosen $\delta = 0.05$ for illustration. Larger (smaller) values of δ would have resulted in uniformly larger (smaller) values of the estimated PoA, independent of whether the distributions under comparison are truly the same or not. This is due to the fact that it is more (less) probable for the difference between two CDFs to lie within a larger (smaller) interval regardless of their underlying similarities and differences.

Next we examine the effect of the sample size n. To that end we consider a sample 10 times smaller ($n = 100$) and a sample 10 times larger ($n =$ 10,000) than what we considered above. Both samples are drawn from the same population as before and we want to quantify the agreement between their estimated CDF and the normal family CDF. We have re-calculated the CDFs, their difference, and the corresponding PoAs for these samples and the result is given in Figure 2.5.

When we have just $n = 100$ observations the sample does not resemble a normal distribution nearly as strongly as when $n =$ 1,000. We see this in the CDF and difference plots in Figure 2.5 in the left column. The PoA calculated with $\delta = 0.05$ is much smaller and is associated with much more uncertainty than when $n =$ 1,000. This is expected; when we have a smaller sample size and hence less certainty in the sample's distribution, the variance of the Horvitz-Thompson estimate of the CDF is larger and so also is the variance of the difference between the two CDFs. When this variance is larger, we are less certain of whether the difference lies with $\pm\delta$, which corresponds to a smaller estimated PoA. Note, however, that even when $n = 100$ the PoA is close to 1 in the tails. That is because a CDF always starts at 0 and ends at 1 and therefore the variance is much smaller in the tails. This accounts for the U-shape pattern exhibited in the PoA plot.



(a) HT estimate of the CDF and the known CDF

(b) The difference between the estimated CDF and the known CDF

(c) The PoA

Figure 2.4: (a) Horvitz-Thompson estimate of the CDF of the observed data and the normal CDF with mean and variance equal to those of the sample. The 95% CI is constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated and the known CDF. The 95% CI is cunstructed the same way as in (a). (c) Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.

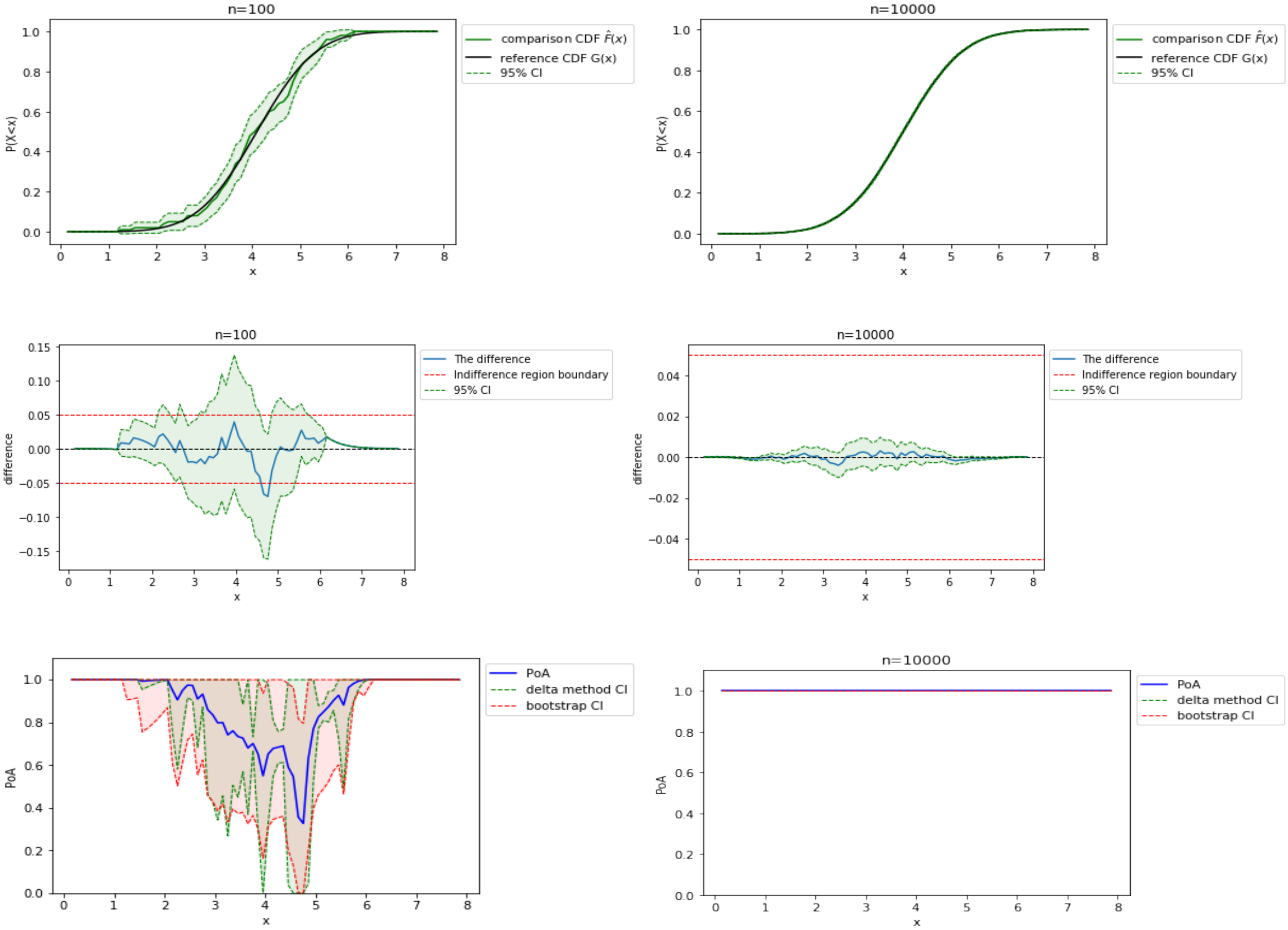


Figure 2.5: The plots in the left column are the estimated CDF overlaid on the reference CDF, the difference between the estimated and the reference CDFs, and the estimated PoA respectively, corresponding a sample of size $n = 100$. On the right hand side we have the same CDF, difference, and PoA plots but this time with a sample of size $n = 10{,}000$. Both samples are generated from N(1,4) and the PoA is calculated with $\delta = 0.05$ in both cases.

On the other hand, when we have a sample of size $n = 10{,}000$ like in the right column of Figure 2.5 we have absolutely no doubt that the difference lies within $\pm\delta$, and hence that the CDFs are practically equivalent. This is suggested by the CDF, difference, and PoA plots in the right column of Figure 2.5, where we see the PoA equals 1 for all values of x

and there is no uncertainty as the upper and lower limits of both CIs are also equal to 1. This again can be explained by the variance of the Horvitz-Thompson estimator. When we have a larger sample, the variance of the estimated CDF will be lower, the variance of the corresponding difference will therefore also be lower, and because the true difference is $F(x) - G(x) = 0$ the estimated PoA will therefore be higher.

As we did in the previous section, we can also look at the estimated PDFs and the relative density to gain more information about how the distributions are similar or different. We have illustrated the relative density and the PDF plots corresponding samples of size $n = 100, 1{,}000,$ and 10,000 in Figure 2.6.

We can see that, as the sample size grows, the known PDF becomes a better approximation for the estimated PDF and the estimated relative density gets closer to 1 for all $0 \leq r \leq 1$. For all sample sizes, the pointwise 95% CI for the relative density covers the value 1 for almost all $0 \leq r \leq 1$. However, the CIs do not cover the value 1 for r near 0 and 1 which is a result of the downward bias of the KDE mentioned previously. In summary, the relative density plots agree with our findings from the PoA analysis: the normal distribution is a good model for the observed data, and that our confidence in this conclusion increases as n increases.

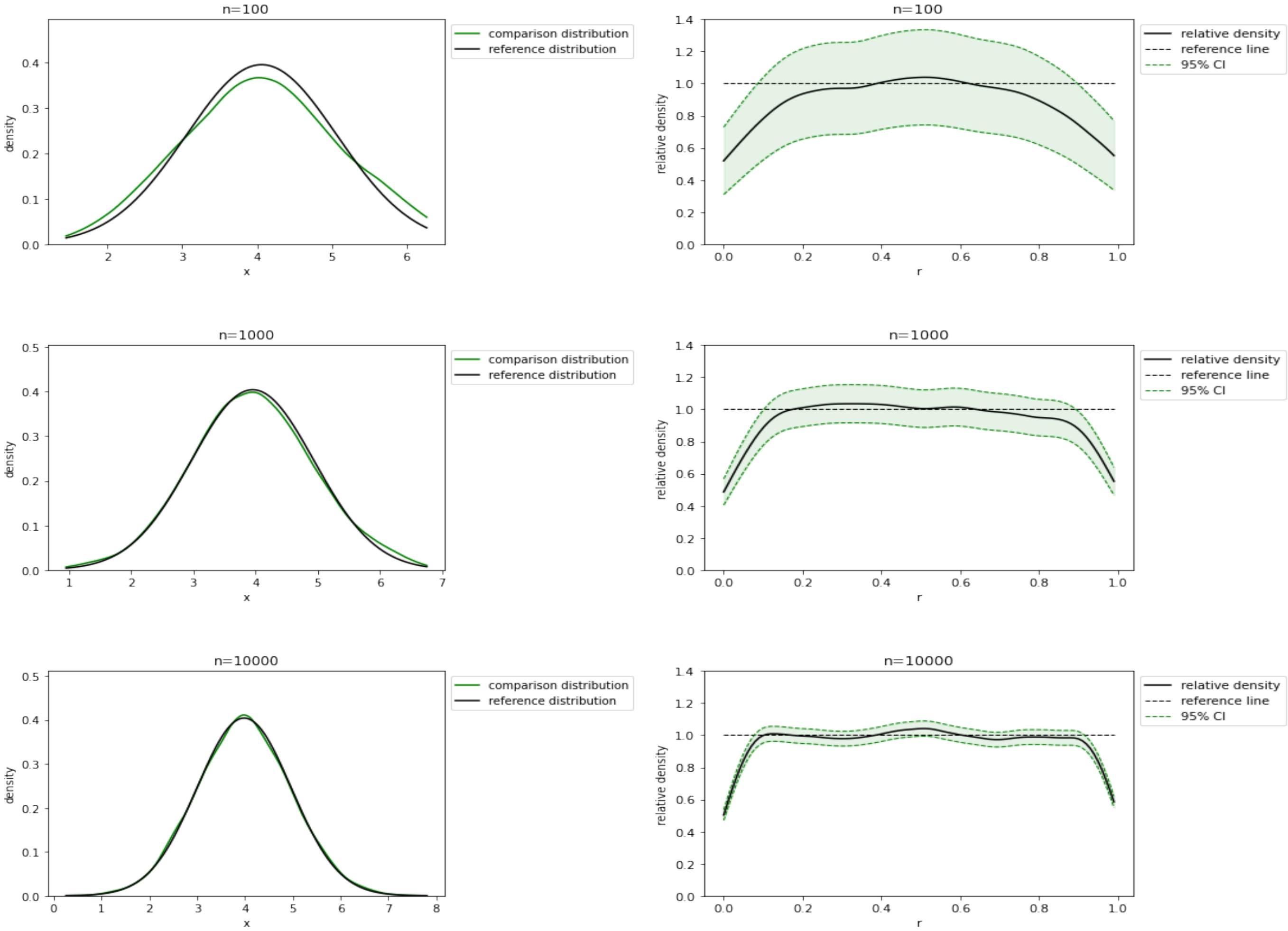


Figure 2.6: The plots on the left are the KDE of the PDF of the observed samples with sizes $n = 100, 1{,}000,$ and 10,000 respectively and also the normal PDFs with mean and variance equal to those of the samples. The plots on the right visualize the KDE of the relative density of the corresponding samples of sizes $n = 100, 1{,}000,$ and 10,000 respectively. The 95% CI is constructed based on asymtotic normality of the KDE.

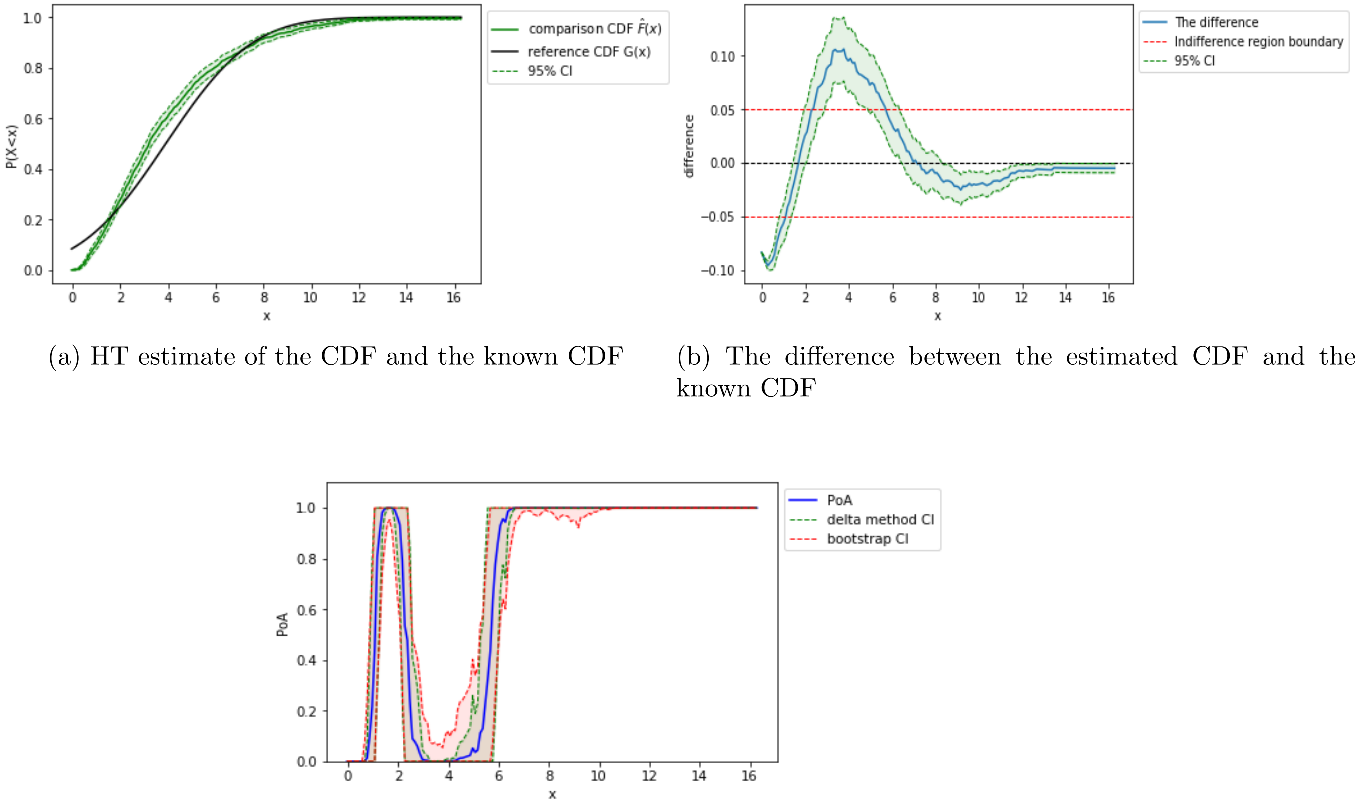
2.3.3 Simulated Example (Gamma Data)

Both of our examples so far have had two things in common: first, the conclusion of the analysis in both cases was that the distributions under comparison were practically equivalent. Second, both of the observed samples were well-modeled by a normal distribution. In this section we consider a different example. The sample is composed of $n = 1{,}000$ observations drawn from a population of size $N = 20{,}000$ which is itself obtained by IID draws from Gamma(2,2). We use the proposed methodology to determine whether this sample can be well-modeled by a normal distribution. Figures 2.7a and 2.7b show the CDF and difference plots respectively. We can see that for values of $x < 7$ the CDFs differ noticeably while for $x > 7$ they are very close to each other. Figure 2.7c shows the estimated PoA with $\delta = 0.05$. We will explain what we see in the PoA from left to right. Unlike in previous examples, the PoA is small for x close to zero, indicating a lack of agreement in the left tail of the distributions being compared. That is because the normal distribution can have negative values in its support set while a gamma distribution is defined only for positive values. Therefore the reference CDF (which is a normal CDF with mean and variance equal to those of the observed sample) has much larger values than the estimated gamma CDF near $x = 0$ because the normal distribution has density for $x < 0$.

Then we find the estimated PoA rises around $x = 2$, which corresponds to the first intersection of the estimated and the known CDFs. After that the PoA is close to zero for $3 < x < 5$ which suggests that in this interval it is quite likely the CDFs differ by more than $\delta = 0.05$. This is supported by the CDF and the difference plots where we see that the differences are not within ±δ for the same interval $3 < x < 5$. Finally, like in the previous examples, we see a large PoA values for $x > 7$ because both CDFs are close to 1 for such large x values and thus their difference is almost certainly smaller than 0.05. In summary, the PoA analysis (with $\delta = 0.05$) suggests that the distributions under comparison are practically the same for large values of x ($x > 7$) but they are different for smaller values of x.

We can also look at the relative density plot (Figure 2.8) for additional insight. We can see that for most values of $0 \leq r \leq 1$, the pointwise 95% CIs for the relative density do not cover 1, which suggests that the PDFs are not the same at those points. This is unsurprising given what we see in the overlaid densities in Figure 2.8a. More specifically, the relative density plot suggests that the comparison PDF has more density for values of r approximately between 0.1 and 0.6 which corresponds to values of x approximately between 0 and 4 (which are the 10^{th} and 60^{th} percentiles of the reference distribution). But after that, for $r > 0.6$ or equivalently $x > 4$, the reference PDF has more density than the comparison PDF.

There are two things worth mentioning when considering the plots in Figure 2.8. First, the PDFs plot and the relative density plot do not agree for large values of x (i.e., $x > 10$). In Figure 2.8a we see the comparison PDF is higher than the reference PDF in this region, which should correspond to a relative density larger than 1. However, in Figure 2.8b we see the estimated relative density is smaller than 1. This is again the effect of the downward bias of the KDE estimator near 0 and 1 which causes the relative density to be smaller than it should be for values of r near the boundary. Second, one should note that the range of x values in the PDFs plot is determined by the observed sample and because our sample is drawn from a Gamma(2,2) population there are only positive observations. However, the reference PDF is a member of the normal family and so a small part of the PDF is in the negative values of x which have not been plotted. Therefore, $x = 0$ on the PDFs plot does not correspond to $r = 0$ on the relative density plot.

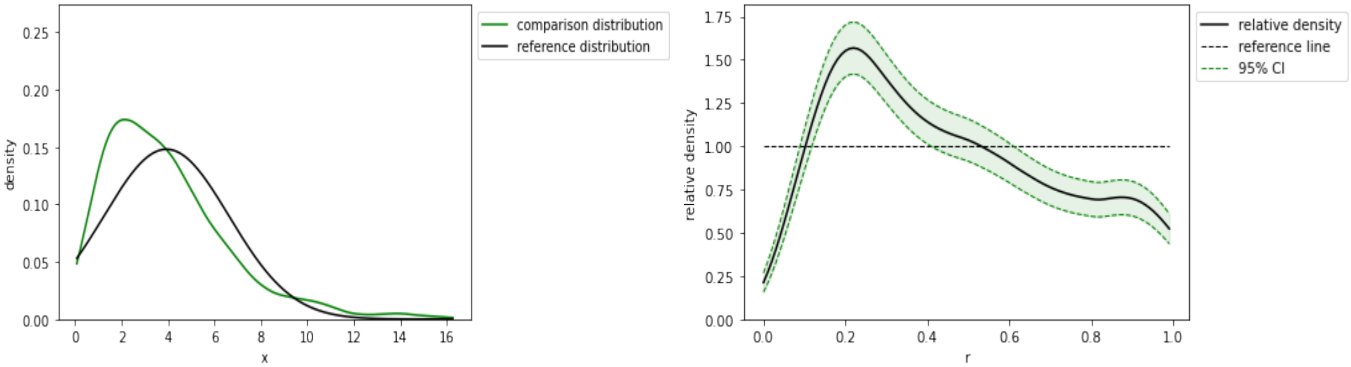


(a) HT estimate of the CDF and the known CDF

(b) The difference between the estimated CDF and the known CDF

(c) The PoA

Figure 2.7: (a) Horvitz-Thompson estimate of the CDF of the observed sample ($n = 1,000$ IID observations from Gamma(2,2)) and the normal CDF with mean and variance equal to those of the sample. The 95% CI is constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated and the known CDF. The 95% CI is constructed the same way as in (a). (c) Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.



(a) The estimated and the known PDFs

(b) The relative density

Figure 2.8: (a) The KDE of the PDF of the observed sample (n = 1,000 IID observations from Gamma(2,2)) and the normal PDF with mean and variance equal to those of the sample. (b) The KDE of the relative density plot. The 95% CI is constructed based on asymtotic normality of the KDE.

Chapter 3

One-Sample Evaluation

3.1 Explaining the Design of the Simulation Study

In this chapter we present the results of a simulation study conducted to examine the coverage of the proposed CIs for the PoA and also the bias and root mean squared error (RMSE) of the PoA estimator. We have considered several different scenarios: i. Gamma(2,2) versus normal family, ii. Gamma(2,2) versus gamma family, iii. N(4,1) versus normal family, and iv. N(4,1) versus gamma family. We also considered a fifth scenario: N(0,1) versus N(0,1), which is different from the other four scenarios because unlike those, in this one the reference distribution is completely specified and its parameters will not be estimated from the observed sample. We consider this fifth scenario because it is a special and important case that has many applications, for instance determining whether standardized residuals follow a N(0,1) distribution. In all five scenarios, we define the population $\mathcal{P}$

by taking $N = 20{,}000$ draws from the specified distribution and we take simple random samples from this population. We consider three different sample sizes, $n = 100$, 1,000, and 10,000, to examine the effect of increasing n. For each sample size, we take $J = 1{,}000$ samples from $\mathcal{P}$ and for each sample, we estimate the CDF, the PoA, and both types of confidence intervals. To estimate the coverage for each type of interval, at each point x we have $J = 1{,}000$ CIs and we calculate the proportion of these CIs which cover the *true* PoA. As Stevens and Lu (2020) do, we define the true PoA $\theta_n(x)$ using a Monte Carlo approach as the proportion of $J = 1{,}000$ estimated CDFs (estimated using the Horvitz-Thompson method) that are within $\pm\delta$ of the known CDF. In other words, the *true* PoA is defined as

$$\theta_n(x) = \frac{1}{J}\sum_{j=1}^{J}\mathbb{I}\{|\hat{F}^j_{HT}(x) - G^j(x)| \leq \delta\}, \tag{3.1}$$

Which is a Monte Carlo estimate of Equation 2.1. Note that $\hat{F}^j_{HT}(\cdot)$ is the Horvitz-Thompson estimate of the CDF for the j^{th} sample and $G^j(\cdot)$ is the known distribution being used for the j^{th} sample. Note that because the parameters of the reference distribution are estimated using the observed sample, the actual CDF being used as the known CDF may be different for each of the $J = 1{,}000$ samples. As such, $G^j(x)$ for all j will correspond to the same distributional family, but the parameters may be slightly different. Also note that the subscript n in $\theta_n(x)$ emphasizes the fact that the true PoA defined as in Equation 3.1 depends on the sample size through the estimator $\tilde{F}_{HT}(x)$. Later we will see that as the sample size grows, $\theta_n(x)$ converges to the indicator function $\mathbb{I}\{|F(x) - G(x)| \leq \delta\}$. In particular, we have

$$\lim_{n\to\infty}\theta_n(x) = \mathbb{I}\{|F(x) - G(x)| \leq \delta\},$$

where $F(\cdot)$ and $G(\cdot)$ are the true CDFs under comparison. We refer to this indicator function as the asymptotic PoA.

The bias and RMSE at a given point x are calculated as

$$bias(x) = \frac{1}{J}\sum_{j=1}^{J}\left[\hat{\theta}_{j,n}(x) - \theta_n(x)\right] \tag{3.2}$$

and

$$RMSE(x) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\hat{\theta}_{j,n}(x) - \theta_n(x)]^2}, \tag{3.3}$$

where $\hat{\theta}_{j,n}(x)$ is the PoA estimated using the j^{th} simulated sample of size n.

In what follows, we showcase the results of our simulation studies for three different scenarios, namely Gamma(2,2) vs normal family, N(4,1) vs normal family, and N(0,1) vs N(0,1). The results of other two scenarios (N(4,1) vs gamma family and Gamma(2,2) vs gamma family) are similar to the N(4,1) vs normal family case and so, for the sake of brevity, we do not include them here. They are however presented in Appendix C.

3.2 Gamma(2,2) vs Normal Family

We begin with the Gamma(2,2) vs normal family case where our samples are generated from a Gamma(2,2) population and we want to determine whether they can be appropriately modeled by a normal distribution.

Figure 3.1a shows the population CDF (Gamma(2,2)) from which the samples are drawn and the CDF of the normal distribution (N(4,8)) which has mean and variance equal to the mean and variance of the Gamma(2,2) distribution. Note however that in the simulations we do not use the N(4,8) CDF as the reference, exactly. Instead we use the normal distribution with mean and variance estimated from the observed sample. This should be close to, but not exactly the same as, N(4,8). The asymptotic PoA which is shown in Figure 3.1b with a dashed line, is calculated using the true CDFs of Figure 3.1a. This asymptotic PoA equals 1 wherever the distance between the true CDFs is smaller than $\delta = 0.05$, and equals 0 otherwise. The blue, red, and green curves in Figure 3.1b correspond to the true probability of agreement $\theta_n(x)$ for $n = 100$, 1,000, and 10,000 respectively. As we can see, the true PoA approaches the asymptotic PoA as n increases. This is expected because with a larger sample size, the actual estimated CDFs under comparison (which are the basis of the definition of the true PoA) get closer to the true CDFs of Figure 3.1a (which are the basis of the definition of the asymptotic PoA).

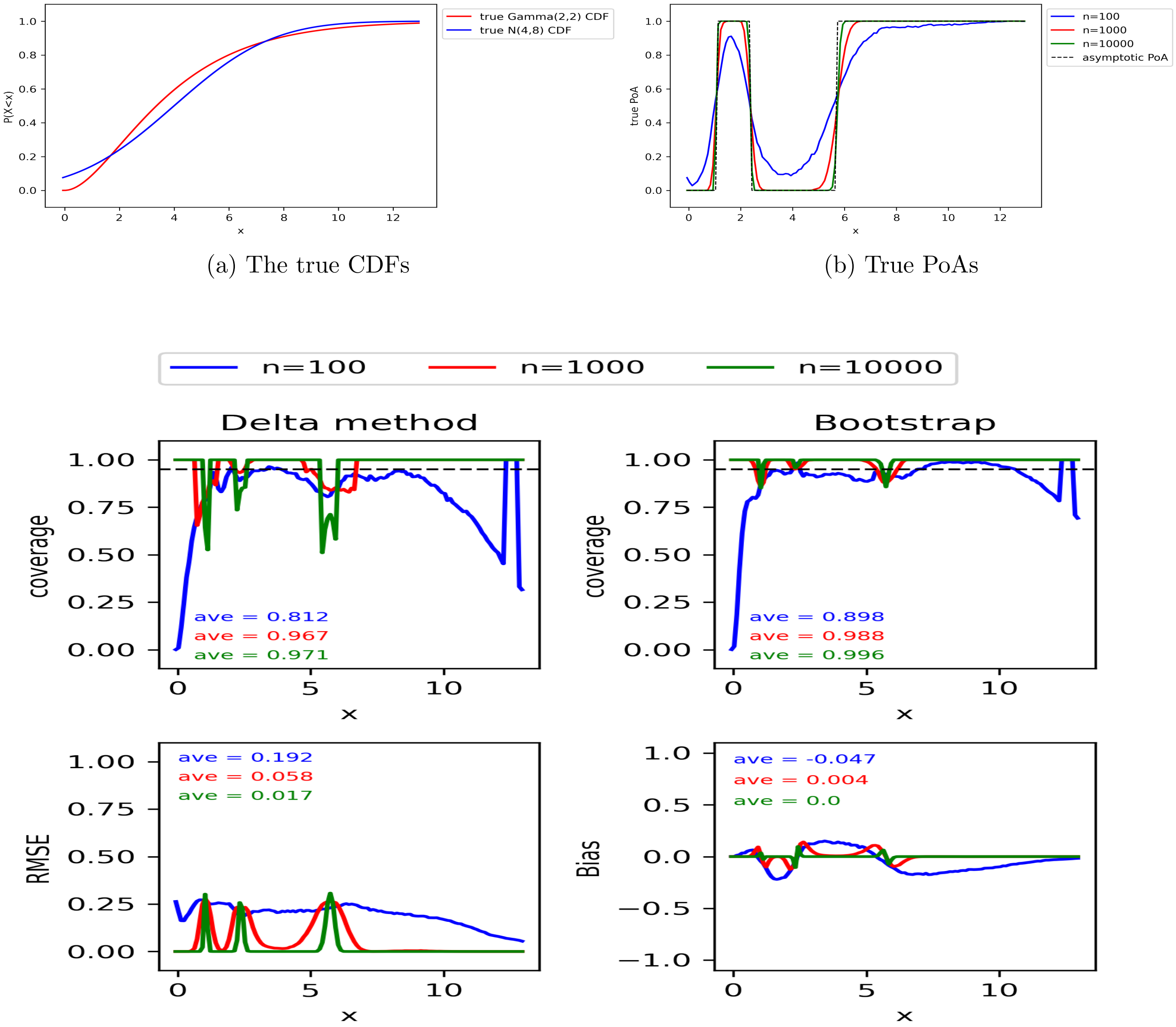
Figure 3.1c illustrates the results for coverage of the bootstrap-based and delta method-based CIs, as well as the bias and RMSE of the PoA estimator, all for $n = 100$, 1,000, and 10,000. In each plot we also display the corresponding average values across the entire range of x for each value of n. The dashed lines in the two coverage plots are drawn at 0.95 and serve as a reference. There are several interesting insights that can be observed from these plots.

Generally speaking, and unsurprisingly, the larger the sample size, the better the result. However, if we look at each of the plots more carefully, we can see that for some values of x, even for very large samples, we observe poor coverage, large bias and large variability. Figure 3.1b provides a justification. We see that it is for these same values of x that the asymptotic PoA transitions from 0 to 1 or from 1 to 0. Therefore, we conclude that the estimated PoA performs worst for values of x where $F(x)$ and $G(x)$ transition from truly being within versus not within $\pm\delta$ of each other. These are the periods of greatest uncertainty regarding whether $|F(x) - G(x)|$ is less than δ or not, and so it is unsurprising that the PoA methodology would struggle here. We call this phenomenon "the transition effect" and we will see this effect again in other simulations in this chapter and also in Part II of the thesis. Although this effect is most pronounced for large values of n, it appears to be a problem regardless of sample size. Another interesting observation concerning the transition effect is that as we increase the sample size (especially from 1,000 to 10,000) the effect's range becomes narrower but the effect itself becomes more severe. To see this, we compare the green ($n =$ 10,000) and the red ($n =$ 1,000) curves in all four plots in Figure 3.1c. We see that in the transition areas, the green curve is worst, i.e., the CI coverages are lower and the RMSE and the bias are higher. However, the range in which we observe poor performance is narrower, and this results in better performance on average, as can be seen by the average coverage, average bias, and average RMSE.

With respect to confidence interval coverage we see that the bootstrap-based and delta method-based CIs follow very similar patterns. In fact, the bootstrap coverage appears to be a scaled version of the delta method coverage with generally higher values. This phenomenon can be seen in all of the simulation scenarios considered here in Part I and also in Part II. This is to be expected given that the bootstrap CIs and delta method CIs differ only by their standard error terms, as was discussed in Section 2.1.5. The bootstrap standard error tends to be larger, which translates into wider intervals and hence the superior coverage observed in Figure 3.1c.

3.3 N(4,1) vs Normal Family

In this section our samples are drawn from a N(4,1) population and we want to check whether a normal distribution is a good model for them. The difference between this section and Section 3.2 is that here the distributions under comparison are truly the same. Therefore, in Figure 3.2a we have just one true CDF and in Figure 3.2b the asymptotic PoA (which is invisible beneath the true PoAs corresponding samples of size 1,000 and 10,000), is a constant line at 1. This makes interpretation more straightforward than the previous section because there is no longer a transition between 0 and 1 and consequently there is no transition effect in any of the plots of Figure 3.2c. Apart from the transition effect, every other aspect of the results in Figure 3.2c is similar to what we saw in Figure 3.1c. Larger sample sizes still result in better overall performance, the bootstrap coverage still looks like an upward scaled version of the delta method coverage, and bias and RMSE are negligible for reasonably large samples.

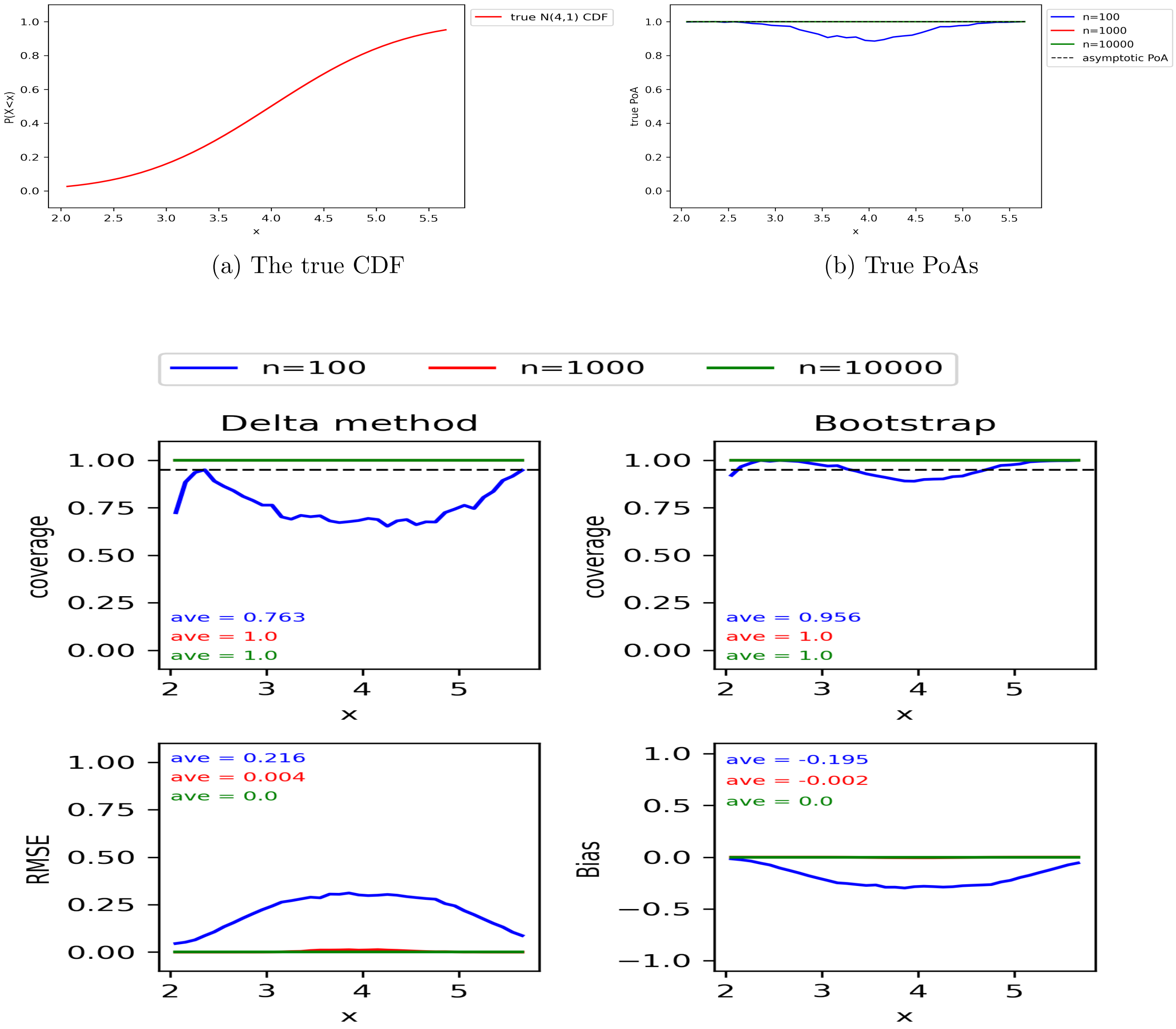


(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure 3.1: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario Gamma(2,2) vs normal family.



(a) The true CDF

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure 3.2: (a) The true versions of the CDFs under comparison (in this scenario they are overlapping). (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(4,1) vs normal family.

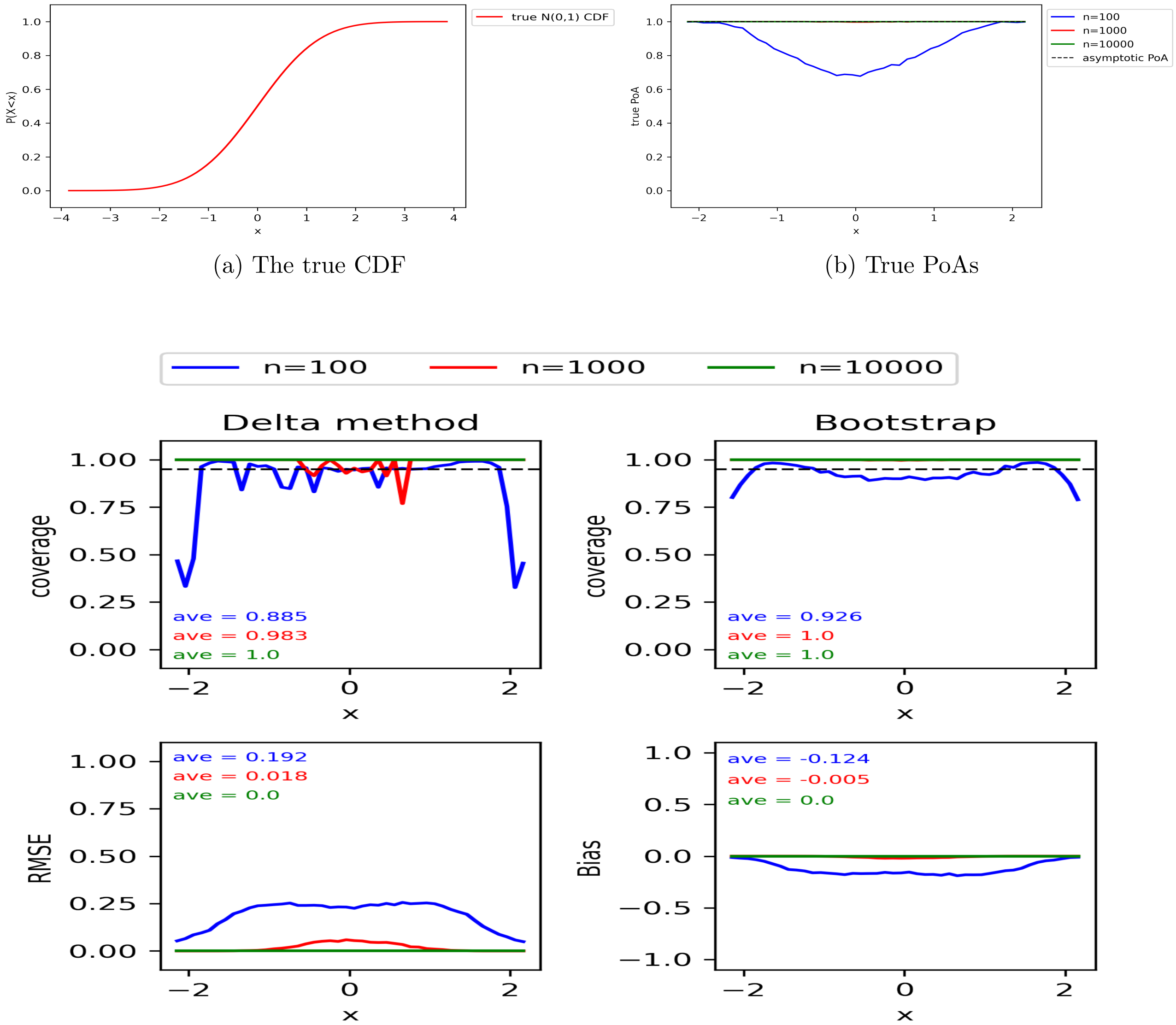
3.4 N(0,1) vs N(0,1)

In this section we consider the special scenario where we compare the estimated CDF against a completely specified distribution. More precisely, our samples are drawn from a N(0,1) population and we want to check whether N(0,1) is a good model for them. We include this scenario in our simulation study to verify that the methodology works well when the reference distribution is completely specified. As we can see, the plots in Figure 3.3 look very similar to those in Figure 3.2 in the previous section. Larger sample sizes (n = 1,000 and 10,000) still result in close to 100% coverage for both types of CIs, and the bias and the RMSE are close to zero. For the small sample size (n = 100), however, the coverage of the CIs are worse than their counterparts in Figure 3.2c. This difference is due to the variability that exists when estimating the parameters of the reference distribution, which we did in Section 3.3 but not here. We discuss this point in more detail in Chapter 7.

3.5 General Insights Drawn

The results of the three scenarios considered in this chapter (and also the two scenarios considered in Appendix C, where we find results very similar to the ones discussed here), lead us to the following conclusions:

- The bootstrap-based CI has consistently higher coverage than the delta method-based CI across different scenarios and different sample sizes.



(a) The true CDF

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure 3.3: (a) The true versions of the CDFs under comparison (in this scenario they are overlapping). (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(0,1) vs N(0,1).

- With larger sample sizes (n = 1,000 and 10,000) the coverage of the both types of CIs are reasonably high, though they may perform poorly in transition areas (if there are any). With the small sample size ($n = 100$) we still get good average coverage.

- The RMSE and bias of the PoA estimator is almost zero for large sample sizes (n = 1,000 and 10,000) meaning that the PoA estimator accurately and precisely estimates the true PoA. When the sample size is small ($n = 100$), however, we have non-zero RMSE and bias but they are still reasonably small.

Part II

Two-Sample Problem

Chapter 4

Introduction

4.1 The Problem

In Part II of the thesis, we will generalize the methods of Part I to the two-sample problem i.e., when we have two independent samples from two potentially different populations and we want to compare the distribution of the populations. Like the one-sample case, the two-sample problem also has a variety of real-world applications. For example, we encounter this problem in statistical experiments where we wish to make inferences about possible differences between subjects in different experimental conditions. Similarly, the question of whether the distributions of two groups are the same may also be of interest in two-group comparisons in observational studies.

Such comparisons could be made on the basis of summary statistics such as means or medians of the two groups. Although such comparisons reveal some information about the potential differences between the underlying groups, more insight may be gained if we compare the whole distributions rather than just summary statistics. For example, consider the PISA dataset introduced in Part I which contains reading scores of 3663 students, and assume we are investigating whether reading scores differ significantly by gender. In Figure 4.1 we can see the histograms of the girls' and boys' reading scores. Suppose we find that the mean of the reading scores of the girls and boys are practically equivalent, meaning the difference is practically negligible. Does this mean that the distribution of reading scores are the same for girls vs. boys? Of course not. A comparison of means provides information only about the location of these distributions. It is possible that they have similar means, but they may differ with respect to variability or skewness, for example. Therefore, it is evident that with a full comparison of the underlying distributions, we can answer such questions more accurately. Later, in Section 5.3, we will apply our proposed methodology to this example to investigate the similarities and differences between the distributions of the reading scores of the girls and boys.

The notation we use here in Part II will be similar to that of Part I. In particular we consider a population $\mathcal{P}$ of size N with the true CDF F and true PDF f from which we have observed a sample $\mathcal{S}$ of n units on which we observe the variable x: $\{x_1, x_2, \dots, x_n\}$. The difference between Part II and Part I is that here G and g are no longer known. In fact, we consider a second population $\mathcal{Q}$ of size M with true CDF G and true PDF g from which we have observed a sample $\mathcal{L}$ of m units on which we observe the variable y: $\{y_1, \dots y_m\}$. Our objective is again to compare the comparison distribution F with the

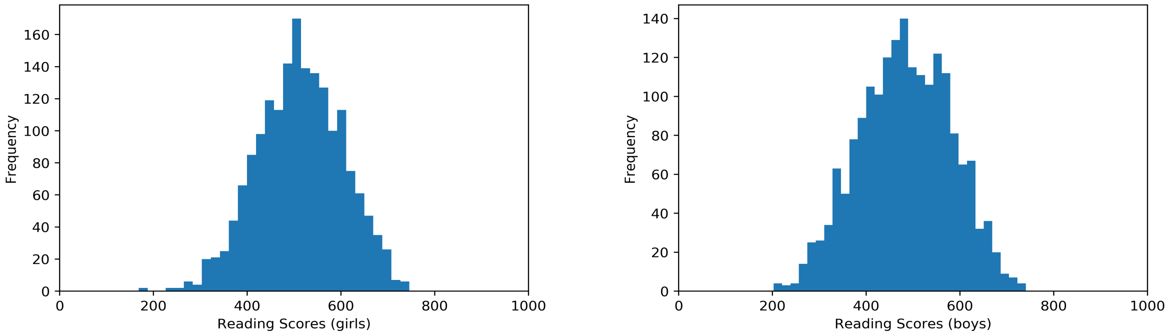


Figure 4.1: Histogram of the girls' (the left plot) and the boys' (the right plot) reading scores.

reference distribution G (and also f with g). But here, because we do not know these functions, we estimate them from the sample data, $\mathcal{S}$ and $\mathcal{L}$. Just as we did in Part I, we use the Horvitz-Thompson estimator to estimate the CDFs, and then we compare these estimates using a probability of agreement analysis that accounts for the uncertainty associated with estimating F and G. For comparing the PDFs we will again use the relative density. But here, because the functions G and g are unknown, we use a different version of the relative density from what was used in Section 2.2.

4.2 Existing Methods

Similar to the one-sample case, the common approach for comparing two unknown distributions is with a hypothesis test. Two of the most well-known tests are the Kolmogorov-Smirnov (Kolmogorov, 1933) test and the Anderson-Darling (Anderson and Darling, 1952;

1954) test. We have seen the one-sample versions in Chapter 1; here we review their two-sample counterparts. The primary difference between the one-sample and two-sample versions of these tests is in the null hypothesis and the corresponding test statistic. While in the one-sample case the null hypothesis is the equality between the distribution from which the sample was drawn and the reference distribution, here, in the two-sample case, the null hypothesis is the equality between the two distributions from which the samples were drawn. Notationally there is no difference, however. The hypothesis statement is

$$H_0 : F = G \quad vs. \quad H_A : F \neq G$$

or, equivalently,

$$H_0 : f = g \quad vs. \quad H_A : f \neq g.$$

Of course, new test statistics are needed for the two-sample case, but the decision to reject the null hypothesis or not, is still based on a p-value. The test statistics associated with the two-sample version of the Kolmogorov-Smirnov (KS) test and the two-sample Anderson Darling test still quantifiy the distance between the distributions under comparison, but here the comparison is between the ECDFs of the two samples, defined as

$$\hat{F}_n(x) = \sum_{u \in \mathcal{S}} \frac{\mathbb{I}[x_u \leq x]}{n} \text{ and } \hat{G}_m(x) = \sum_{u \in L} \frac{\mathbb{I}[y_u \leq x]}{m}, \tag{4.1}$$

where $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ are the observed data associated with samples $\mathcal{S}$ and $\mathcal{L}$. The two-sample KS test statistic is then given by

$$D_{n,m} = sup_x|\hat{F}_n(x) - \hat{G}_m(x)|.$$

For large sample sizes, the null hypothesis is rejected at level α when $D_{n,m} > \sqrt{-\frac{1}{2}\ln(\frac{\alpha}{2})\frac{n+m}{nm}}$.

The two-sample Anderson-Darling test statistic also has the same construction as its one-sample counterpart. It is a weighted sum of the distances between the two distributions under comparison which here are the two ECDFs. It has the following form

$$A^2_{nm} = \frac{nm}{n+m}\int_{-\infty}^{\infty}\frac{\left[\hat{F}_n(x) - \hat{G}_m(x)\right]^2}{\hat{H}_{n+m}(x)[1-\hat{H}_{n+m}(x)]}d\hat{H}_{n+m}(x),$$

where $\hat{H}_{n+m}(\cdot)$ is the ECDF of the combined sample

$$\hat{H}_{n+m}(x) = \frac{n\hat{F}_n(x) + m\hat{G}_m(x)}{n+m}.$$

Note that this test statistic makes the test sensitive to differences in the tails of the combined sample where $\hat{H}_{n+m}(x)[1-\hat{H}_{n+m}(x)]$ is close to zero. The distribution of A^2_{nm} is not known in closed form, but tables of critical values have been published (Pettitt, 1976).

In Part I we explained that when we aim to prove the equality of two characteristics, it may be preferable to start with the opposite assumption, i.e., to begin by assuming that the two characteristics under comparison are different, and then look for evidence of equivalence. This philosophy is also true for the two-sample problems discussed here. Wellek (2010) promotes the use of two-sample equivalence tests in which the null and alternative hypotheses are defined as

$$H_0 : \pi_+ \leq 1/2 - \epsilon_1 \text{ or } \pi_+ \geq 1/2 + \epsilon_2 \quad \text{vs.} \quad H_A : 1/2 - \epsilon_1 \leq \pi_+ \leq 1/2 + \epsilon_2,$$

where $\pi_+ = P(X > Y)$ and $X \sim F$ and $Y \sim G$. The intuition here is that if $F = G$ then $\pi_+ = P(X > Y) = \frac{1}{2}$. The constants ϵ_1 and ϵ_2 are chosen using practical considerations just like δ in the PoA analysis. Wellek (2010) proposes testing this hypothesis with a statistic related to the Mann-Whitney U-statistic.

However, just as in the one-sample case, the decision to reject or not reject the null hypothesis (in either traditional tests or equivalence tests) is based on a p-value. Therefore, all of the p-value problems discussed in Part I are still present here. In particular: the fact that a single p-value is not very informative, difficult to interpret, and can be made arbitrarily small by making the sample size arbitrarily large, are still relevant issues.

In the next section, we demonstrate that the PoA approach can be easily adapted for use in the two-sample problem and therefore provides an alternative method of comparing two unknown distributions that does not suffer from the same issues as hypothesis tests do.

4.3 An Alternative Approach

The PoA can also be used in the two-sample setting as well. We define the PoA between two unknown distributions F and G based on their estimators $\tilde{F}$ and $\tilde{G}$

$$\theta(x) = P(-\delta \leq \tilde{F}(x) - \tilde{G}(x) \leq \delta).$$

The only difference between this definition and the one-sample version in Equation 1.2 is that here G is also unknown and must be estimated. As in Part I, we suggest that an additional comparison of PDFs f and g may be beneficial because, as opposed to comparing CDFs (which have a cumulative interpretation), comparing PDFs have a pointwise interpretation that can be insightful. A graphical comparison of PDFs is facilitated by the two-sample version of the relative density. In Chapter 5 we discuss the two-sample generalization of the Horvitz-Thompson-based PoA and the use of the relative density in the two-sample setting. In Chapter 6 we report the results of a simulation study (like the one in Chapter 3) that investigates the properties and performance of the proposed PoA methodology in the two-sample setting.

Chapter 5

Proposed Methodology for the Two-Sample Scenario

5.1 The Probability of Agreement for Two-Sample CDF Comparisons

The problem of comparing the distributions of two different populations arises in practically every field of study. In this section, we adapt the probability of agreement (PoA) methodology for use in that scenario.

5.1.1 The Probability of Agreement

The definition of the PoA used in the two-sample scenario is analogous to the one-sample version in Equation 2.1. The difference is that here we consider the difference between estimators of *two* unknown CDFs instead of the difference between the estimator of one unknown CDF and one known CDF. Therefore, the PoA is defined as

$$\theta(x) = P\left(-\delta \leq \tilde{F}(x) - \tilde{G}(x) \leq \delta\right), \tag{5.1}$$

where $\tilde{F}(\cdot)$ and $\tilde{G}(\cdot)$ are the estimators for the unknown CDFs $F(\cdot)$ and $G(\cdot)$ that characterize populations $\mathcal{P}$ and $\mathcal{Q}$, respectively. The interpretation of the interval $(-\delta, \delta)$ is the same as in Part I: it is a context-specific indifference region, meaning that as long as the difference between the characteristics lies in this interval, we consider the characteristics practically equivalent. In order to calculate the PoA we must estimate the CDFs and, like in Part I, we do this using the Horvitz-Thompson method as we describe in the next subsection.

5.1.2 Applying the Horvitz-Thompson Method

To use the Horvitz-Thompson method in the two-sample setting, we mimic what was done in the one-sample case separately for both samples. Let $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ be the samples from populations $\mathcal{P}$ and $\mathcal{Q}$ that are characterized by the CDFs F and G respectively. Then we have

$$\hat{F}_{HT}(x) = \sum_{u \in \mathcal{S}} \frac{\mathbb{I}(x_u \leq x)}{N\pi_u}$$

and

$$\hat{G}_{HT}(x) = \sum_{u \in \mathcal{L}} \frac{\mathbb{I}(y_u \leq x)}{M\pi_u},$$

where N and M are the population sizes corresponding to $\mathcal{P}$ and $\mathcal{Q}$ respectively, and π_u is a marginal inclusion probability.

The estimated variances of these estimators are given by

$$\widehat{Var}\left[\tilde{F}_{HT}(x)\right] = \sum_{u \in \mathcal{S}} \sum_{v \in \mathcal{S}} \left(\frac{\pi_{uv} - \pi_u \pi_v}{\pi_{uv}} \right) \frac{\mathbb{I}(x_u \leq x)}{N\pi_u} \frac{\mathbb{I}(x_v \leq x)}{N\pi_v} \tag{5.2}$$

and

$$\widehat{Var}\left[\tilde{G}_{HT}(x)\right] = \sum_{u \in \mathcal{L}} \sum_{v \in \mathcal{L}} \left(\frac{\pi_{uv} - \pi_u \pi_v}{\pi_{uv}} \right) \frac{\mathbb{I}(y_u \leq x)}{M\pi_u} \frac{\mathbb{I}(y_v \leq x)}{M\pi_v}, \tag{5.3}$$

where π_{uv} is a joint inclusion probability. Finally, given the asymptotic normality result of Section 2.2, we have

$$\frac{\tilde{F}_{HT}(x) - F(x)}{\widehat{Var}\left[\tilde{F}_{HT}(x)\right]} \xrightarrow{D} N(0,1) \tag{5.4}$$

and

$$\frac{\tilde{G}_{HT}(x) - G(x)}{\widehat{Var}\left[\tilde{G}_{HT}(x)\right]} \xrightarrow{D} N(0,1), \tag{5.5}$$

where the variance estimates are calculated as in Equations 5.2 and 5.3.

5.1.3 Estimating the Probability of Agreement

Because interest lies in the distribution of the difference between the two CDFs, the asymptotic normality results given in Equations 5.4 and 5.5 can be used to determine the following asymptotic result for the difference

$$\frac{\left(\tilde{F}_{HT}(x) - \tilde{G}_{HT}(x)\right) - \left(F(x) - G(x)\right)}{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]} \xrightarrow{D} N(0,1). \tag{5.6}$$

Note that because the two samples $\mathcal{S}$ and $\mathcal{L}$ are independent, so also are the two estimators $\tilde{F}_{HT}(x)$ and $\tilde{G}_{HT}(x)$, and so the variance of their difference is given by the sum of their individual variances. With this result, we calculate the PoA as

$$\begin{aligned}\theta(x) &= P\left(-\delta \leq \tilde{F}_{HT}(x) - \tilde{G}_{HT}(x) \leq \delta\right)\\ &\cong \Phi\left(\frac{\delta - (F(x) - G(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}}\right) - \Phi\left(\frac{-\delta - (F(x) - G(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}}\right),\end{aligned} \tag{5.7}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. By replacing $F(\cdot)$ and $G(\cdot)$ with their Horvitz-Thompson estimates, we obtain an estimate of the PoA

$$\hat{\theta}(x) = \Phi\left(\frac{\delta - (\hat{F}_{HT}(x) - \hat{G}_{HT}(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}}\right) - \Phi\left(\frac{-\delta - (\hat{F}_{HT}(x) - \hat{G}_{HT}(x))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}}\right). \tag{5.8}$$

Like before, we can use either the bootstrap or asymptotic normality and the delta method to construct approximate CIs for $\theta(x)$. The bootstrap approach in this case is identical to the approach described in Section 2.1.4 for the one-sample case, except that $\hat{\theta}_1(x), \ldots \hat{\theta}_B(x)$ are now calculated using Equation 5.8 instead of Equation 2.7. However, the delta method approach must be generalized for the two-sample problem. We describe this generalization in the next section.

5.1.4 Confidence Intervals Based on Asymptotic Normality

In this section we use the delta method to estimate the distribution of the PoA estimator at a given point x, and we use it to calculate a pointwise CI for $\theta(x)$. Similar to the development in Section 2.1.5, we will write the estimator version of Equation 5.8 as

$$\tilde{\theta}(x) \cong \Phi\left(U(\tilde{D}(x))\right) - \Phi\left(L(\tilde{D}(x))\right), \tag{5.9}$$

where

$$\tilde{D}(x) = \tilde{F}_{HT}(x) - \tilde{G}_{HT}(x),$$

$$U(\tilde{D}(x)) = \frac{\delta - \tilde{D}(x)}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}},$$

and

$$L(\tilde{D}(x)) = \frac{-\delta - \tilde{D}(x)}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}}.$$

Regarding the PoA as a function of $\tilde{D}(x)$ and given the distributional result from Equation 5.6, we are able to use the delta method to determine the distribution of the PoA estimator, $\tilde{\theta}(x)$. Based on the delta method the variance of $\tilde{\theta}(x)$ is given by

$$\begin{aligned} Var[\tilde{\theta}(x)] &= \left[\frac{d\theta(x)}{dD(x)}\right]^2 Var[\tilde{D}(x)] \\ &= \left[-\frac{\phi(U(D(x))) - \phi(L(D(x)))}{\sqrt{\widehat{Var}\left[\tilde{F}_{HT}(x)\right] + \widehat{Var}\left[\tilde{G}_{HT}(x)\right]}}\right]^2 Var[\tilde{D}(x)] \\ &= \left[\phi(U(D(x))) - \phi(L(D(x)))\right]^2, \end{aligned}$$

where $\phi(\cdot)$ is the PDF of a standard normal distribution. Hence, by the delta method we have

$$\frac{\tilde{\theta}(x) - \theta(x)}{[\phi(U(D(x))) - \phi(L(D(x)))]} \xrightarrow{D} N(0,1). \tag{5.10}$$

Using this asymptotic result one can construct approximate CIs. The approximate $100(1-\alpha)\%$ CI for $\theta(x)$ is given by

$$\left[\hat{\theta}(x) - z_{\alpha/2}\left(\phi(U(\hat{D}(x))) - \phi(L(\hat{D}(x)))\right), \hat{\theta}(x) + z_{\alpha/2}\left(\phi(U(\hat{D}(x))) - \phi(L(\hat{D}(x)))\right)\right],$$

where $z_{\alpha/2}$ is the quantile of the standard normal distribution such that $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$, and $U(\hat{D}(x))$ and $L(\hat{D}(x))$ are the upper and lower bounds defined above, but evaluated at $\hat{D}(x) = \hat{F}_{HT}(x) - \hat{G}_{HT}(x)$.

5.2 The Relative Density for Two-Sample PDF Comparisons

Here we review the relative density methodology for the two-sample problem. Just as in Section 2.2, the goal is to compare two PDF functions, but here both PDFs are unknown and so additional care is required. In Section 5.2.1, we review the process of estimating the relative density and its associated confidence intervals in the two-sample context. In Section 5.2.2 we discuss a decomposition of the relative density which one can use in the two-sample setting to gain a more in-depth understanding of the similarities/differences between the two PDFs under comparison.

We could in theory apply the PoA methodology directly to the estimated relative density, but for the same practical issue discussed in Part I, we do not suggest using the PoA for the relative density. Nevertheless, the interested reader can refer to Appendix B to see how the PoA may be defined and calculated for the two-sample relative density.

5.2.1 Two-Sample Relative Density

As in Section 2.2 we assume that we have a random variable X with CDF F and PDF f that defines the distribution of the comparison population $\mathcal{P}$ from which we have observed a sample $\mathcal{S}$. We also assume we have a random variable Y with CDF G and PDF g that defines the distribution of the reference population $\mathcal{Q}$, from which we have observed a sample $\mathcal{L}$. The definition of the two-sample relative density is the same as in one-sample setting: the relative distribution comparing F to G is defined as the distribution of the random variable $R = G(X)$. Therefore, H and h, the CDF and PDF of R, can be found in the same manner as in Section 2.2

$$H(r) = F(G^{-1}(r))$$

and

$$h(r) = \frac{dH(r)}{dr} = \frac{f(G^{-1}(r))}{g(G^{-1}(r))}.$$

However, when estimating the relative density in the two-sample setting, we must now account for the fact that the function G is not known. We can no longer simply calculate the relative data defined as $r_i = G(x_i)$ for $i = 1, \ldots, n$; we must first estimate G. We may do so using the ECDF of G defined in Equation 4.1, and then we can calculate what is known as the quasi-relative data (Handcock and Morris, 1999).

$$q_i = \hat{G}(x_i) \quad for \quad i = 1, \ldots, n.$$

Next we apply the KDE method to the quasi-relative data to estimate the relative density function $h(r)$

$$\hat{h}_{n,m}(r) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{r - q_i}{b_n}\right), \tag{5.11}$$

where b_n is the bandwidth and $K(\cdot)$ is a kernel function. Note that the kernel density estimate of h depends on m through the quasi-relative data q_i. As in the one-sample case we use the biwieght kernel given in Equation 2.12 and we also use the same rule for selecting the bandwidth, namely the normal reference rule given in Equation 2.13.

(Handcock and Morris, 1999) provide the asymptotic distribution of the KD estimator of the relative density in the two-sample setting. They show the following asymptotic result for the estimator in Equation 5.11

$$\frac{\hat{h}_{n,m}(r) - h(r)}{\sqrt{\frac{h(r)R(K)}{nb_n} + \frac{h^2(r)R(K)}{mb_n}}} \xrightarrow{D} N(0,1), \tag{5.12}$$

where $R(v) = \int_{-\infty}^{\infty} v(x)^2 dx$.

Note that the above result holds when we are using the quasi-relative data q_i instead of relative data r_i in the kernel density estimation. The second term in the asymptotic variance of the estimator, as compared to the single variance term in the one-sample version (see Equation 2.14), may be interpreted as the cost of using the estimated $\hat{G}$ instead of the true G.

This asymptotic result can be used to calculate approximate pointwise CIs for $h(r)$. The approximate level $100(1-\alpha)\%$ CI is given by

$$\left[\hat{h}_{n,m}(r) - z_{\alpha/2}\sqrt{\frac{\hat{h}_{n,m}(r)R(K)}{nb_n} + \frac{\hat{h}^2_{n,m}(r)R(K)}{mb_n}}, \hat{h}_{n,m}(r) + z_{\alpha/2}\sqrt{\frac{\hat{h}_{n,m}(r)R(K)}{nb_n} + \frac{\hat{h}^2_{n,m}(r)R(K)}{mb_n}}\right],$$

where the standard errors are obtained by replacing $h(r)$ with $\hat{h}_{n,m}(r)$ in the variance term of Equation 5.12.

5.2.2 Decomposition of the Relative Density

In order to gain a more detailed understanding of the potential differences between the two distributions under comparison, Handcock and Morris (1999) proposed a decomposition of the relative density into factors that can be attributed to differences in mean, differences in scale, and differences in shape other than mean and scale. Assume that $y_r = G^{-1}(r)$ and consider the following identity

$$h(r) = \frac{f(y_r)}{g(y_r)} = \frac{g_L(y_r)}{g(y_r)} \times \frac{g_{LS}(y_r)}{g_L(y_r)} \times \frac{f(y_r)}{g_{LS}(y_r)}.$$

Here g_L is the density function of the random variable $Y + \alpha$ where Y has density g and α is chosen such that the mean of $Y + \alpha$ equals the mean of X. Similarly, g_{LS} is the density function of the random variable $\beta(Y + \alpha)$ where Y is as before and α and β are chosen such that the mean and variance of $\beta(Y + \alpha)$ equal those of X. This way the first term $h_L(r) = \frac{g_L(y_r)}{g(y_r)}$ only corresponds to the difference in mean, the second term $h_{LS}(r) = \frac{g_{LS}(y_r)}{g_L(y_r)}$ only corresponds to the difference in variance, and the third term (also known as the residual term) $h_R(r) = \frac{f(y_r)}{g_{LS}(y_r)}$ contains all the information about shape differences not attributed to a mean shift or a difference in variability.

It is important to justify why we did not use this decomposition in the one-sample case. In the one-sample goodness-of-fit problem we were primarily concerned with comparing F with a known distribution G whose mean and variance were equal to the mean and variance of X. In this scenario, the above decomposition gives no additional information because the reference distribution already has the same mean and variance as the comparison distribution and so the first two terms of the decomposition would be 1 and the residual term would be the same as the overall relative density.

5.3 Examples

In this section we will illustrate our proposed methodology through some examples. We start with a real world example to demonstrate the process in practice. Then we will consider several simulated examples to better illustrate the methodology and to show that the results are in line with our expectation in a simulated scenario where we know the truth about the distributions being compared.

5.3.1 PISA Dataset, Boys vs Girls

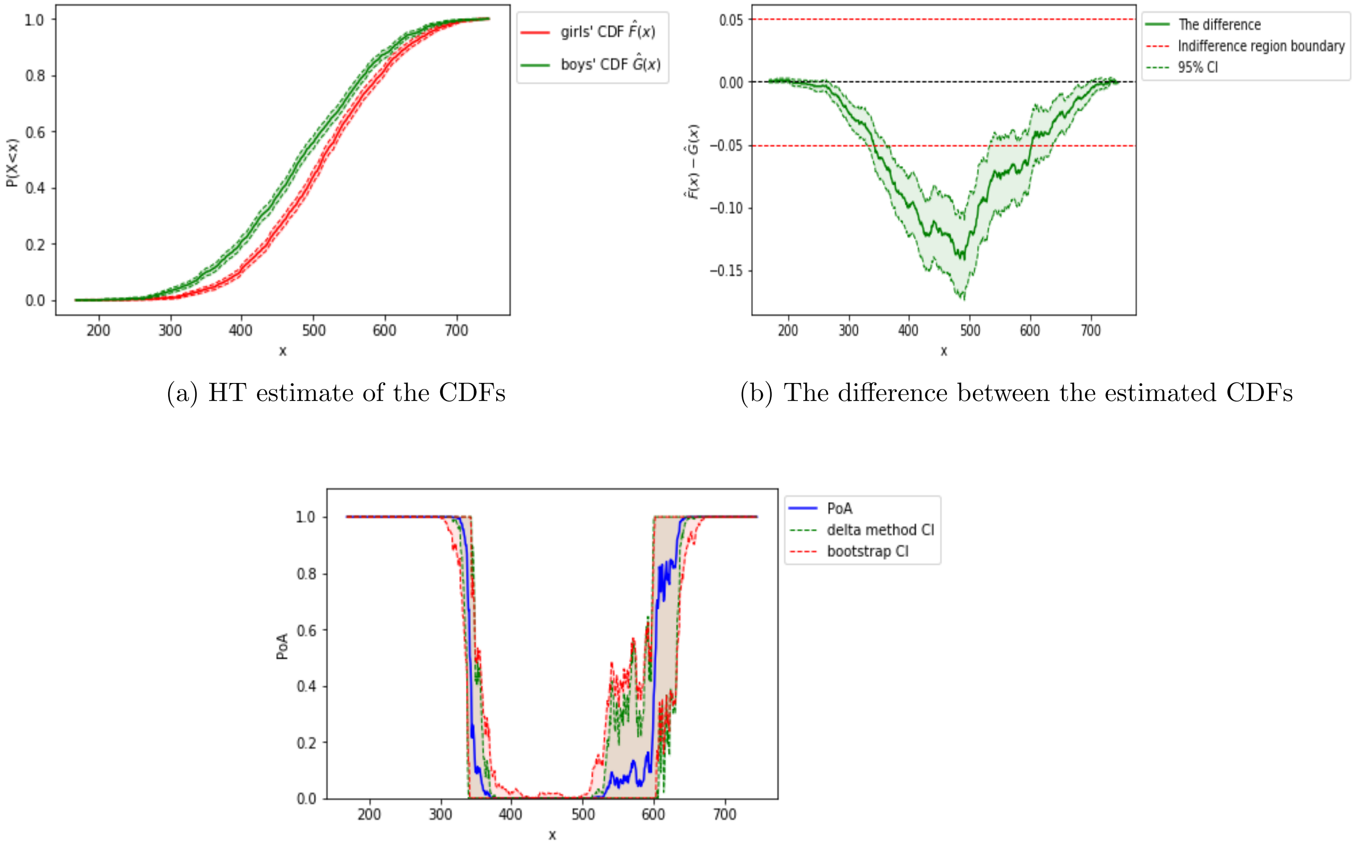
As our first example, we return to the PISA dataset that we introduced in Section 1.1. In Section 2.3.1 we investigated whether the normal distribution is a good model for the students' reading scores. Here we apply our proposed two-sample methodology to compare the distributions of boys' and girls' reading scores and to investigate whether they are practically equivalent. Figure 4.1 shows the histogram of boys' and girls' reading scores. Looking at those histograms it seems that girls' and boys' scores are similarly distributed. We now apply our proposed methodology to gain more in-depth information about the similarities and differences between these two distributions.

Figure 5.1a shows the Horvitz-Thompson estimate of the CDFs of the girls' and boys' reading scores along with the corresponding approximate 95% CIs calculated based on the asymptotic normality of the Horvitz-Thompson estimator. We assume for illustration a population of size $M = N = 10{,}000$ for both girls and boys and we assume that the observed data ($n = 1{,}791$ girls and $m = 1{,}872$ boys) represent simple random samples from them. Therefore the relevant marginal and joint inclusion probabilities are $\frac{n}{N} = \frac{1791}{10000}$, $\frac{n(n-1)}{N(N-1)} = \frac{1791 \times 1790}{10000 \times 9999}$ for the girls and $\frac{m}{M} = \frac{1872}{10000}$, $\frac{m(m-1)}{M(M-1)} = \frac{1872 \times 1871}{10000 \times 9999}$ for the boys.

We can see that the estimated CDFs are quite different except in the tails. The same phenomenon can be seen in Figure 5.1b which shows the difference between the estimated CDFs from Figure 5.1a. In Figure 5.1b we also include two horizontal lines at $\pm\delta = \pm 0.05$. We can see that for values of $350 < x < 600$ the absolute difference between the two estimated CDFs is larger than $\delta = 0.05$. The estimated PoA is shown in Figure 5.1c, and the plot illustrates that in the same range of $350 < x < 600$ the chance that the difference is smaller than $\delta = 0.05$ is very small. For other values of x we are almost certain that the absolute difference between the estimated CDFs is smaller than 0.05.

However, as was discussed in Section 2.3.1, the conclusion of the PoA analysis depends on the value of δ. Figure 5.2a shows the contour plot of PoA and Figure 5.2b shows several PoA curves corresponding to different values of δ. As is expected, the PoA is larger (smaller) with larger (smaller) values for δ. As we explained in Part I, like the value of δ, how large the estimated PoA, $\hat{\theta}(x)$, must be to conclude practical equivalence between $\hat{F}(x)$ and $\hat{G}(x)$ should also be determined by the practitioner. But given what we see in

Figures 5.1c, 5.2a, and 5.2b, it appears as though there is, generally speaking, poor agreement between these two distributions, and therefore that girls' and boys' reading scores are not practically equivalent.



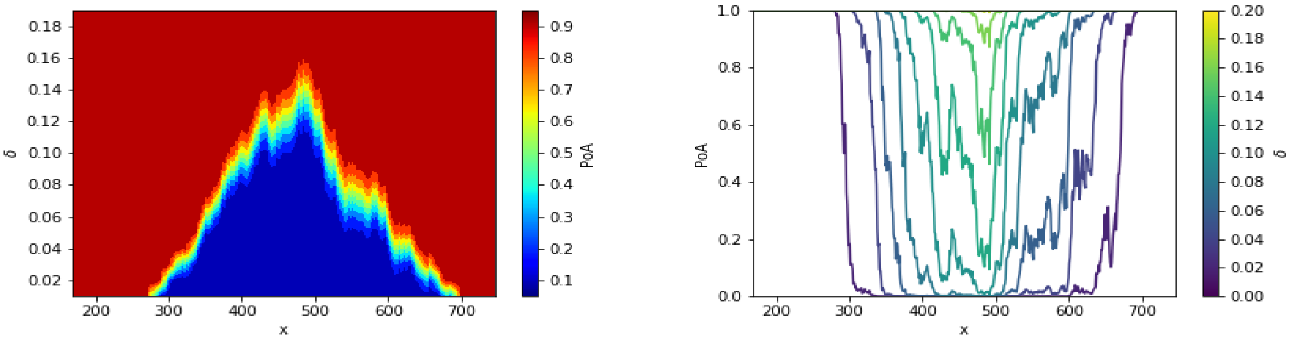
(a) HT estimate of the CDFs

(b) The difference between the estimated CDFs

(c) The PoA

Figure 5.1: (a) Horvitz-Thompson estimate of the CDFs of the girls' and boys' reading scores with their corresponding 95% CIs. The 95% CI is constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated CDFs. The 95% CI is constructed the same way as in (a). (c)Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.

Next, we look at the estimated PDFs and the relative density plots to gain more insight



(a) contour plot of PoA

(b) PoAs with different δ values

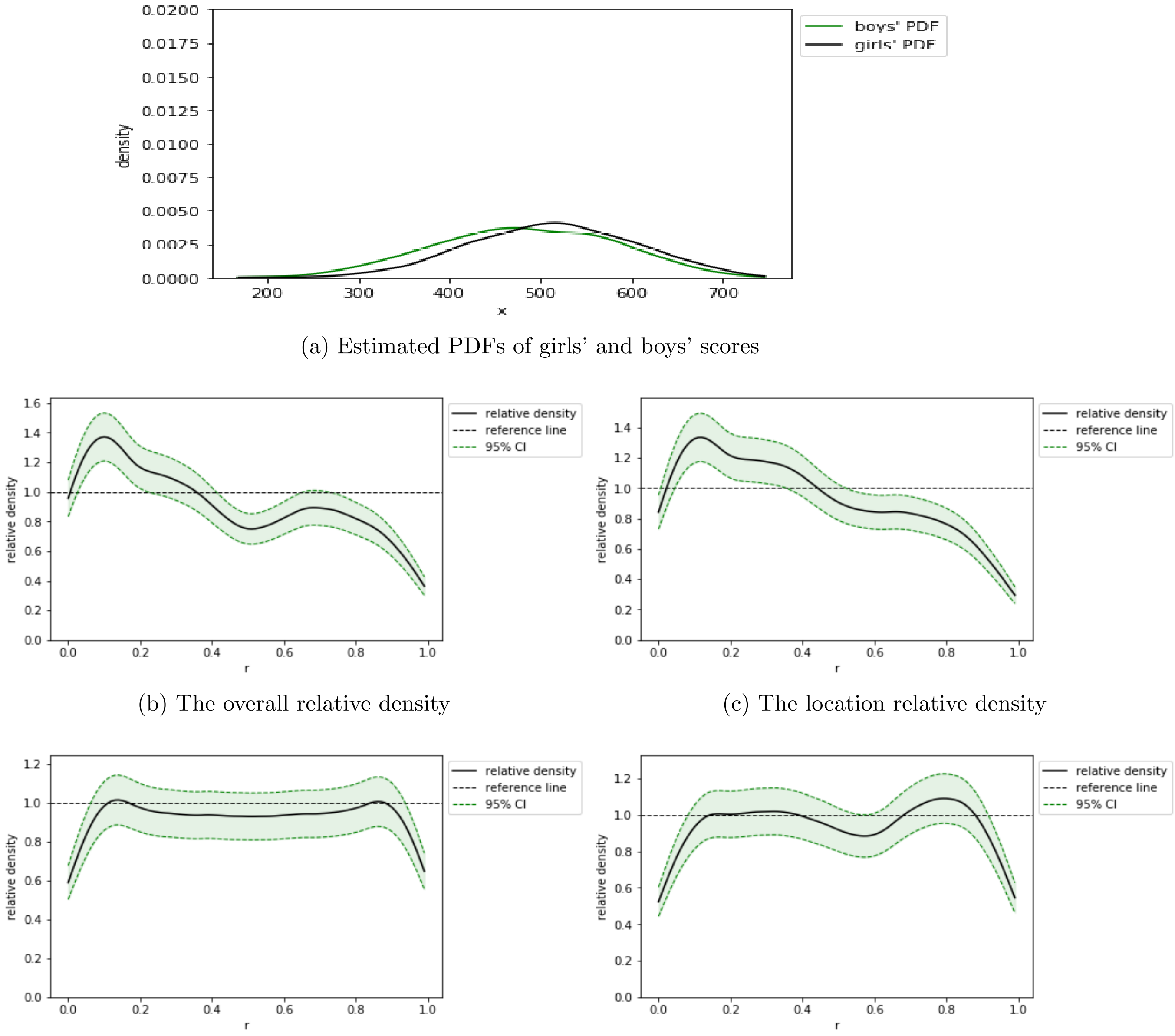
Figure 5.2: (a)The contour plot of PoA. (b) Several PoA plots constructed with different values of δ.

into how the underlying distributions differ. Figure 5.3a shows the KDE of the PDFs of the girls' and boys' reading scores. As opposed to the histograms in Figure 4.1, Figure 5.3a suggests that the girls' reading scores tend to be slightly larger than the boys. Figures 5.3b, 5.3c, 5.3d, and 5.3e show the overall, location, scale, and residual relative densities respectively. Recall that the location relative density quantifies the difference between the mean of the reference and comparison distributions while the scale relative density corresponds only to the difference in the variance of the two distributions. Finally, the residual relative density accounts for any difference not attributed to differences in mean or variance.

Let us first interpret the location relative density visualized in Figure 5.3c. When the mean of the reference distribution is larger (smaller) than the mean of the comparison, we expect to see a decreasing (increasing) location relative density plot. Apart from values of r near 0 and 1 which are affected by the downward bias of the KDE, we can see a decreasing trend which suggests that the reference (girls') distribution has larger mean than the comparison (boys') distribution.

Next we interpret the scale relative density visualized in Figure 5.3d. Generally speaking, in a scale relative density plot we expect to see a U-shape (inverted U-shape) when the variance of the reference distribution is smaller (larger) than the variance of the comparison distribution. However, we do not see either of these patterns. Instead, we see a relatively flat shape, with the 95% pointwise CIs covering 1 for all values of r except those close to 0 or 1. Recall, the 95% CI is a pointwise confidence interval and should not be interpreted as a confidence band. The decrease in relatively density values for r near 0 and 1 should not be interpreted as an inverted U- shape. This is a result of the downward bias associated with the KDE. Therefore, Figure 5.3d suggests that there is no difference in the variance of the distribution of boys' and girls' scores.

Finally we consider the residual relative density. Unlike the other two plots, here there is no general pattern that we should look for. We should only check and see if the residual plot is close to 1. A good reference for closeness is the CI; we should see if the reference value 1 is included in the pointwise CIs at each $0 < r < 1$. In Figure 5.3e we can see that the reference value 1 is included in the pointwise CIs for all values of r except for those near 0 and 1 (which are again affected by the downward bias of the KDE). The plots in Figures 5.3c, 5.3d, and 5.3e together suggest that the primary difference between the two distributions is a difference in mean. In other words, the two underlying distributions are likely just shifted versions of each other.



(a) Estimated PDFs of girls' and boys' scores

(b) The overall relative density

(c) The location relative density

(d) The scale relative density

(e) The residual relative density

Figure 5.3: (a) The KDEs of the PDFs of girls' and boys' reading scores. (b), (c), (d), and (e) are the KDEs of the overall, location, scale, and residual relative densities with corresponding 95% CIs respectively. All CIs are constructed based on the asymptotic normality of the KDE. The girls' scores are considered as the reference sample and the boys' scores are the comparison sample.

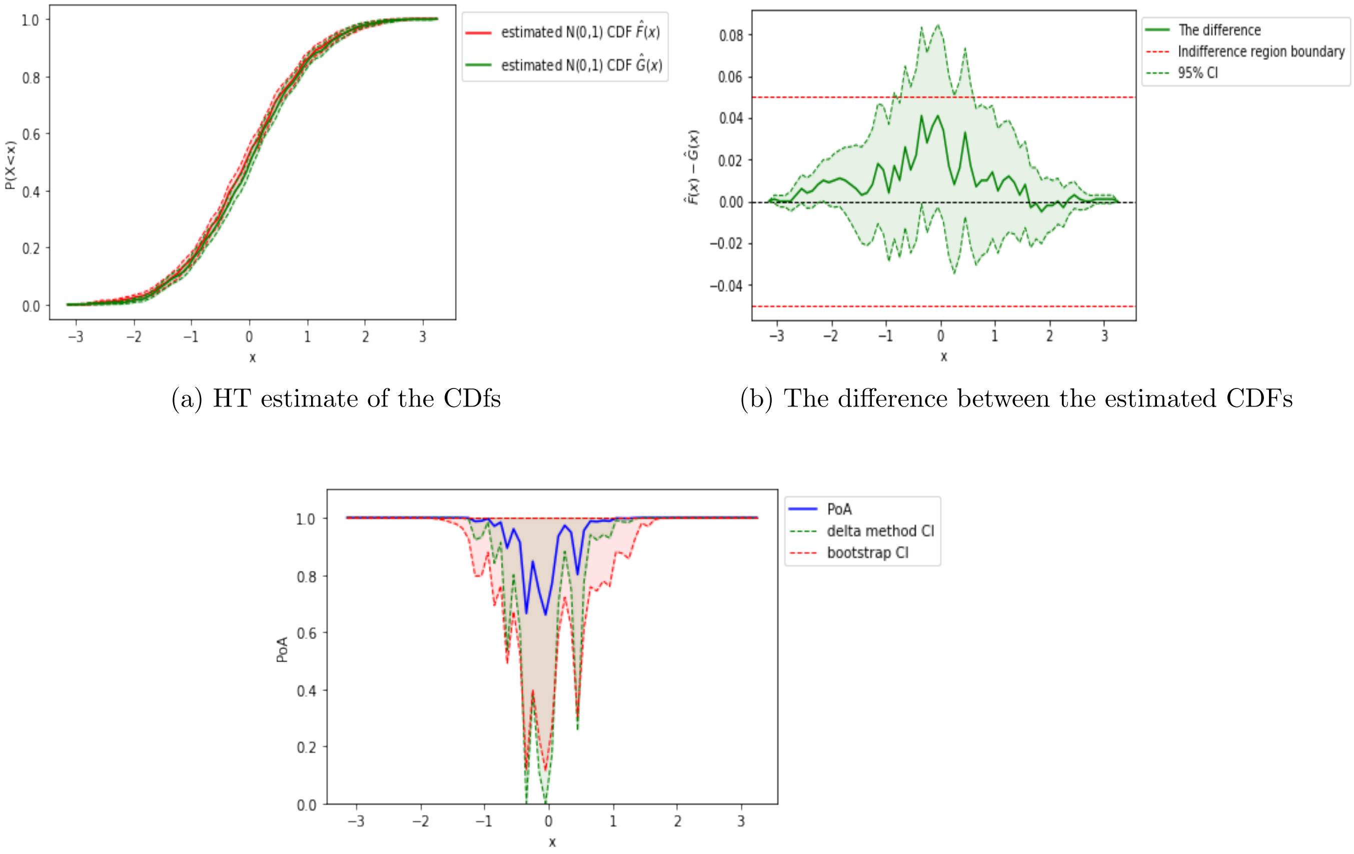
5.3.2 N(0,1) vs N(0,1)

In the boys' vs girls' test score example that we considered in the previous subsection, we did not know whether the boys' reading scores and girls' reading scores were truly from the same distribution or not. In this section we consider a simulated example where the two observed samples are truly from the same distribution.

Assume that we have two simple random samples, each composed of $m = n = 1{,}000$ observations drawb from populations $\mathcal{P}$ and $\mathcal{Q}$ which are themselves composed of $N = M = 20{,}000$ IID draws from $N(0, 1)$. The goal is to apply the proposed methodology to quantify the agreement between the observed distributions and decide whether they are practically equivalent. Note that we do not want to find a specific distribution that fits both samples, rather we just want to check whether it is reasonable to believe that they come from the same underlying model, whatever that happens to be. In our methodology, we first calculate the Horvitz-Thompson estimate of the CDF of both samples and then we calculate the PoA which is the probability that these estimated CDFs are practically equivalent across their support sets.

Figure 5.4a shows the Horvitz-Thompson estimate of the CDFs of the observed samples. As we can see for values of x near -3 and 3, the two CDFs are very close and the CIs are very narrow. This is expected because a CDF starts at 0 and increases to 1 and therefore in the tails the variance of the estimated CDF is very small. For values of x between -3 and 3, on the other hand, the variances of the estimators are larger and so the CIs are not as narrow as in the tails. The same phenomenon can be seen in the difference plot (Figure 5.4b) where we see very narrow CIs in the tails and slightly wider CIs in the middle. Note that in Figure 5.4b the difference curve is in the indifference region for all values of x, but for values of $x \in [-1, 1]$ we can see that the upper bound of the CI is above the horizontal line at $\delta = 0.05$. This is in line with the PoA plot in Figure 5.4c where we can see large PoA values in the tails and slightly smaller PoA values for $x \in [-1, 1]$. This means that we are certain that the absolute difference between the estimated CDFs is smaller than $\delta = 0.05$ for value of $x < -1$ and $x > 1$ but for values of $-1 < x < 1$ there is less certainty as evidenced by wider CIs and slightly smaller PoA values in this region. Even still, the chance of the estimated CDFs being within ± 0.05 of each other is still reasonably high even for $x \in [-1, 1]$. It is up to the practitioner to decide what size the PoA should be to conclude the estimated CDFs are practically equivalent, but the evidence in Figure 5.4c seems strong.

Notice that the patterns exhibited in these plots depend on the sample size and the value of δ. The larger the sample, the smaller the variance, and the less uncertainty we have. Here, because the true underlying distributions are the same, larger sample sizes will result in larger PoA values with narrower CIs. For the sake of brevity, we do not include plots corresponding to larger or smaller sample sizes but one can refer to Section 2.3.2 where we have illustrated the effect of the sample size in the one-sample case. The effect of sample size is analogous in the two-sample case. We do, however, explore the effect of sample size on confidence interval coverage and the bias and variance of the PoA estimator in Chapter 6.

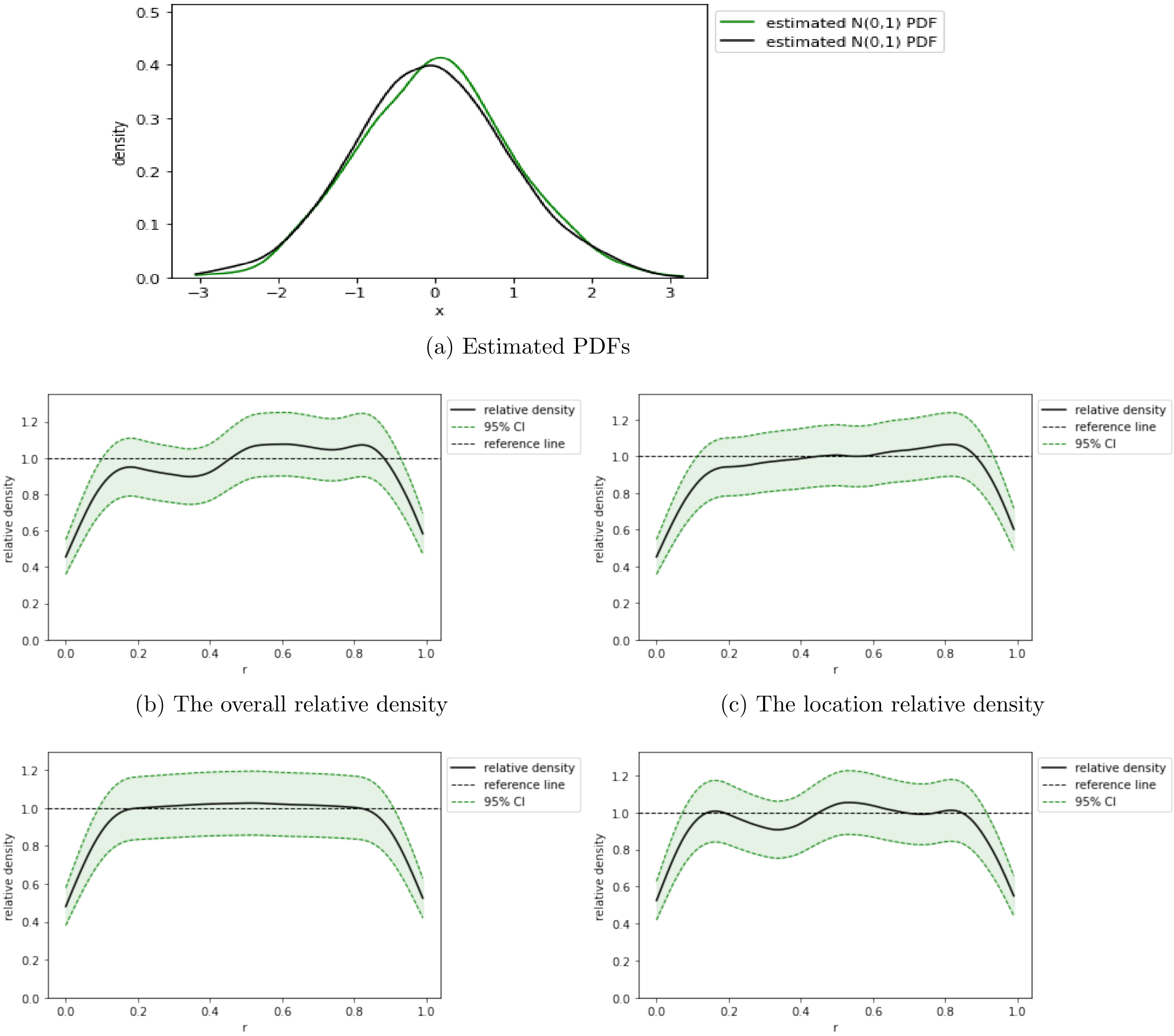


(a) HT estimate of the CDfs

(b) The difference between the estimated CDFs

(c) The PoA

Figure 5.4: (a) Horvitz-Thompson estimate of the CDFs of the observed samples ($n = m = 1{,}000$) from N(0,1) with their corresponding 95% CIs. The 95% CIs are constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated CDFs. The 95% CI is constructed the same way as in (a). (c)Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.



(a) Estimated PDFs

(b) The overall relative density

(c) The location relative density

(d) The scale relative density

(e) The residual relative density

Figure 5.5: (a) The KDEs of the PDFs of samples ($n = m = 1{,}000$) generated from N(0,1). (b), (c), (d), and (e) are the KDEs of the overall, location, scale, and residual relative densities with corresponding 95% CIs respectively. All CIs are constructed based on the asymptotic normality of the KDE.

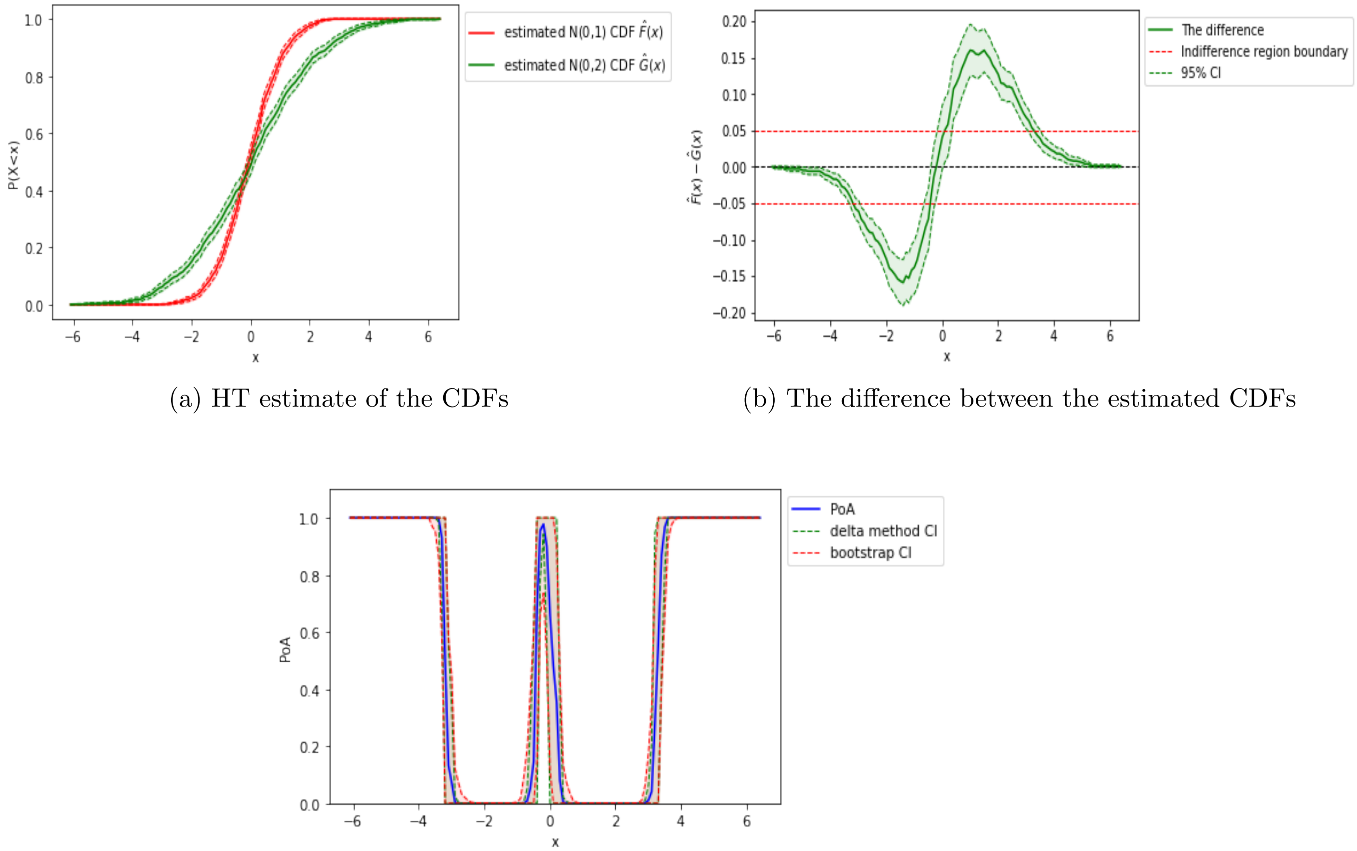
Finally, we examine the relative density plots for further insight into the comparability of the two distributions. The KDE estimate of all four types of relative density, as well as the KDE of the PDFs of the observed samples are shown in Figure 5.9. As we can see, all five plots suggest that the two samples come from the same distribution. The estimated PDFs in Figure 5.5a are very similar and the pointwise CIs in all four relative density plots contain the value 1 for all values of r except for those near 0 or 1 (which are affected by the downward bias associated with kernel density estimation as has already been discussed).

5.3.3 N(0,1) vs N(0,2)

Next we consider another simulated example where the underlying population distributions are truly different. More precisely, we consider two simple random samples of size $m = n =$ 1,000, one drawn from a population of size $N =$ 20,000 characterized by the $N(0, 1)$ distribution, and the other drawn from a population of size $M =$ 20,000 characterized by the $N(0, 2)$ distribution. As in the previous examples, the goal is to use the proposed methodology to determine whether or not the distributions that characterize the underlying populations that gave rise to the samples are practically equivalent.

Figure 5.6a shows the Horvitz-Thompson estimates of the CDFs of the observed samples along with the corresponding pointwise 95% CIs which are calculated based on the asymptotic normality of the Horvitz-Thompson estimator. As always, we see that the estimated CDFs are very close with very narrow CIs in both tails. However, unlike what we saw in the previous example, $\hat{F}(x)$ and $\hat{G}(x)$ are not close to each other in the middle of the graph. They are substantially different except for values of x near 0 where they have an intersection. Figure 5.6b visualizes the difference between $\hat{F}(x)$ and $\hat{G}(x)$ and we draw the same conclusions as in Figure 5.6a: it is almost zero with narrow CIs in the tails, it crosses 0 around $x = 0$, and it is far from zero for other values of x. The PoA plot in Figure 5.6c (with $\delta = 0.05$) visualizes the probability that these differences are within ± 0.05. The interpretation of the PoA plot matches intuition: the PoA is large (close to 1) for values of x where the differences are in the indifference region and the PoA is small (close to 0) where the difference is not contained in the indifference region. Note that the value of $\delta = 0.05$ is again chosen for illustration. In practice, the user should carefully choose a suitable value for δ which facilitates an informative comparison in the context of their problem.

As always, we suggest looking at the relative density plots to gain more information about the similarities and differences between the two distributions. Figure 5.7a shows the estimated PDFs which suggests that there is a difference in scale but the location seems to be the same. The same insight is confirmed by the location and scale relative densities in Figures 5.7c and 5.7d respectively. The location relative density does not show an increasing or decreasing trend, suggesting there is no difference in the location of the two estimated PDFs. The scale relative density shows a U-shape which means there is more variability in the comparison sample compared to the sample reference. The residual relative density in Figure 5.7e suggests that there is no difference other than a difference in location or scale. These conclusions are expected as the distribution under comparison differ only with respect to their variances.

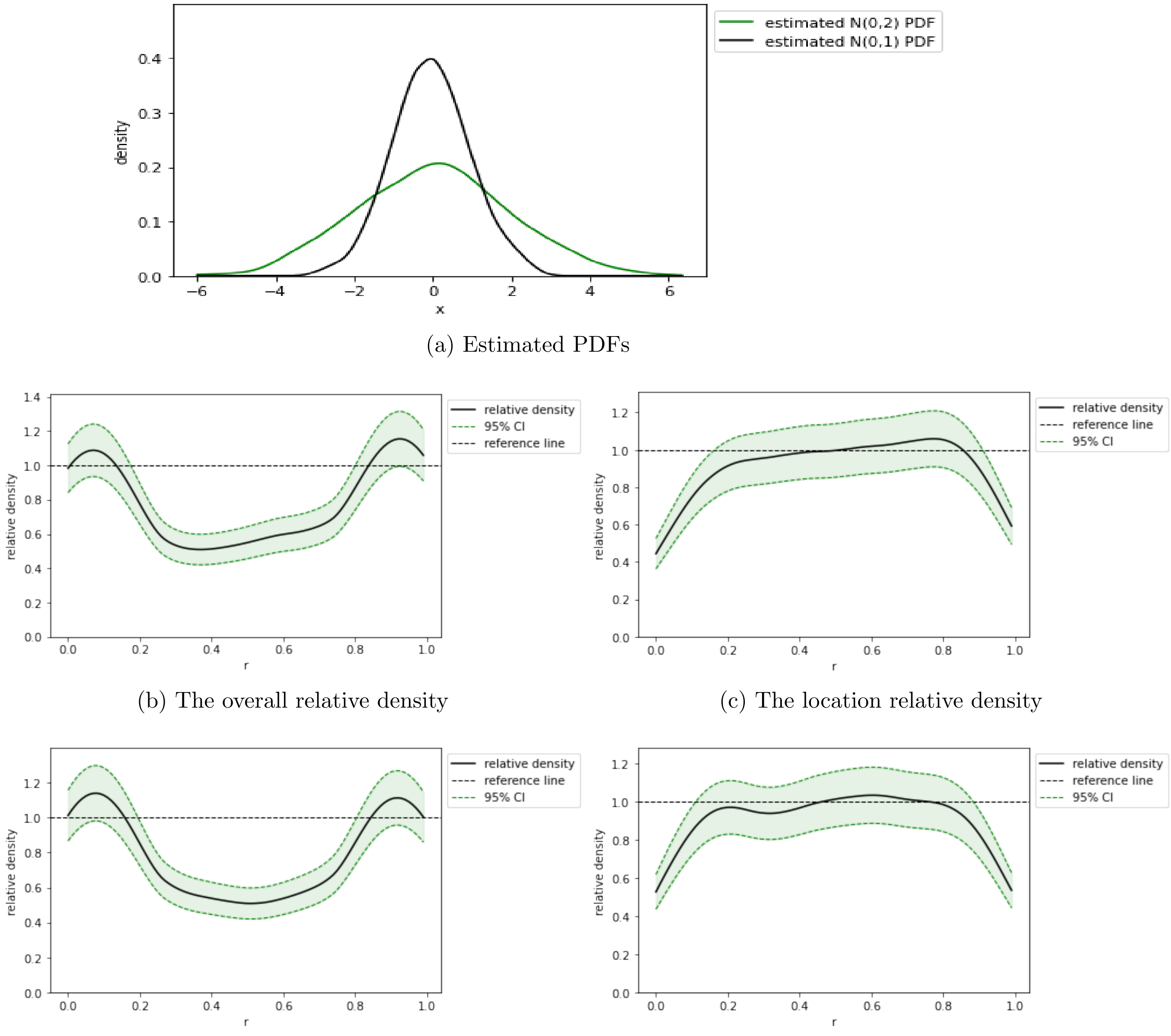


(a) HT estimate of the CDFs

(b) The difference between the estimated CDFs

(c) The PoA

Figure 5.6: (a) Horvitz-Thompson estimate of the CDFs of the observed samples ($n = m = 1{,}000$) from N(0,1) and N(0,2) with their corresponding 95% CIs. The 95% CIs are constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated CDFs. The 95% CI is constructed the same way as in (a). (c)Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.



(a) Estimated PDFs

(b) The overall relative density

(c) The location relative density

(d) The scale relative density

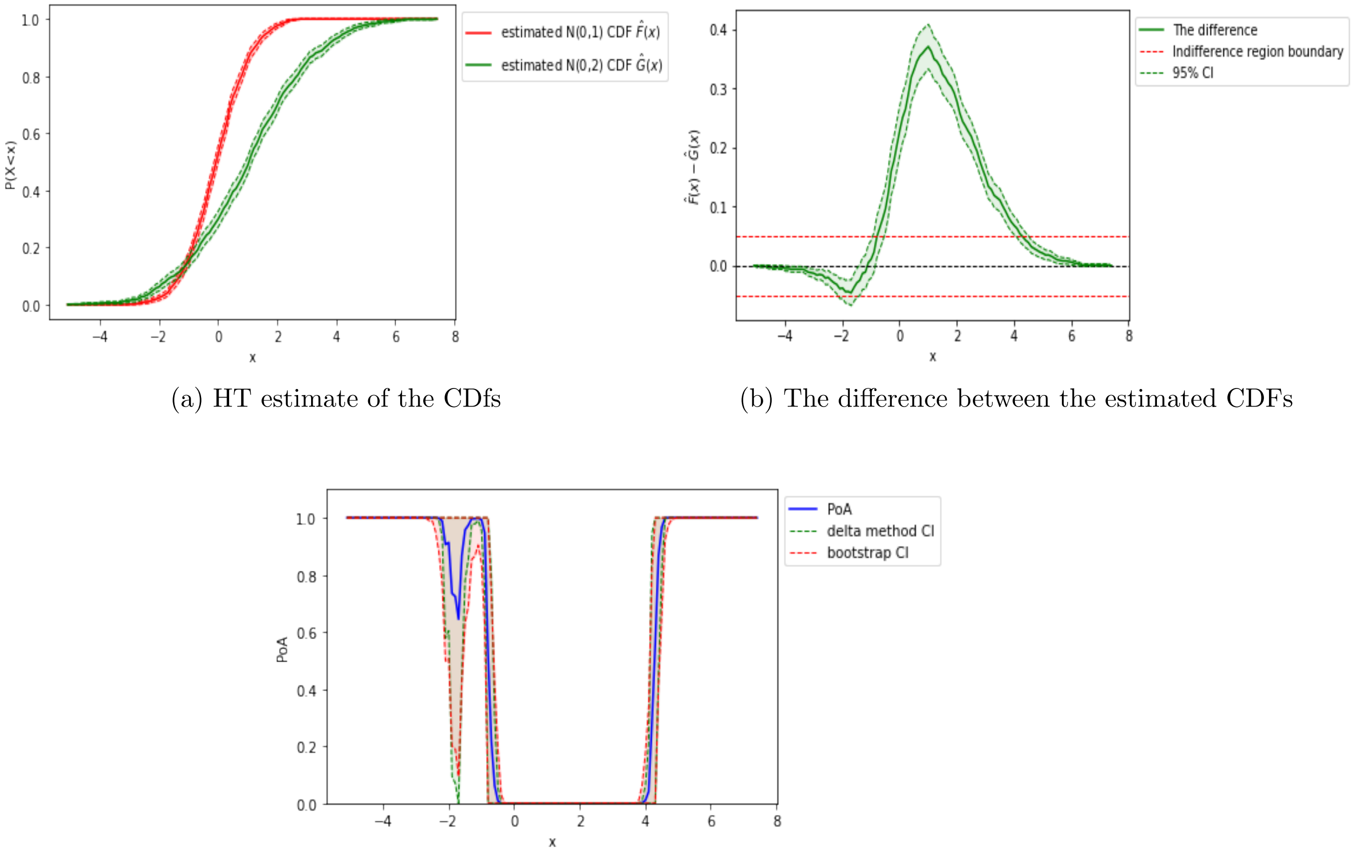
(e) The residual relative density

Figure 5.7: (a) The KDEs of the PDFs of the reference and comparison samples ($n = m = 1{,}000$) generated from from N(0,1) and N(0,2) respectively. (b), (c), (d), and (e) are the KDEs of the overall, location, scale, and residual relative densities with corresponding 95% CIs respectively. All CIs are constructed based on the asymptotic normality of the KDE.

5.3.4 N(0,1) vs N(1,2)

In our fourth and final example, we consider two normal distributions which differ both in terms of mean and variance. Here our two simple random samples of size $n = m = 1{,}000$ are drawn from a population of size $N = 20{,}000$ characterized by the N(0,1) distribution, and the other drawn from a population of size $M = 20{,}000$ characterized by the N(1,2) distribution. Figure 5.8a shows the Horvitz-Thompson estimate of the CDFs and Figure 5.8b shows the difference between the estimated CDFs. Both plots show that except in the tails, the estimated CDFs are quite different. The same interpretation can be gleaned from the PoA plot in Figure 5.8c which was calculated using $\delta = 0.05$. This plot shows almost certain practical equivalence in the tails and a very small chance of practical equivalence for other values of x. However, for x near -2 the PoA increases, suggesting that for those values of x there is roughly a 70% chance of the two estimated CDFs being within $\pm\delta$ from each other. Increased agreement for these values of x is not surprising given that it is for the same values of x that the estimated CDFs intersect. However, the CIs are very wide here, indicating increased uncertainty in this conclusion. The final decision in such areas with high uncertainty should be made by the practitioner.

Just as in the previous example, the PoA plot suggests that the distributions under comparison are practically inequivalent, so we consult the relative density plots for insight into why this might be the case. Figure 5.9a shows the KDE of the PDFs which suggest the two distributions under comparison are different both in terms of mean and variance. To calculate the relative density plots we consider the N(1,2) population the reference population and the N(0,1) population the comparison population. We made this choice to illustrate the fact that when the reference distribution has larger mean and larger variance, we will see a decreasing location relative density and an inverted U-shape scale relative density which are shown in Figures 5.9c and 5.9d respectively. Figure 5.9e is the estimated residual relative density which suggests there is no difference between the two distributions under comparison beyond a difference in means and variances. This is reassuring given that the distributions under comparison are normal distributions with different means and variances.

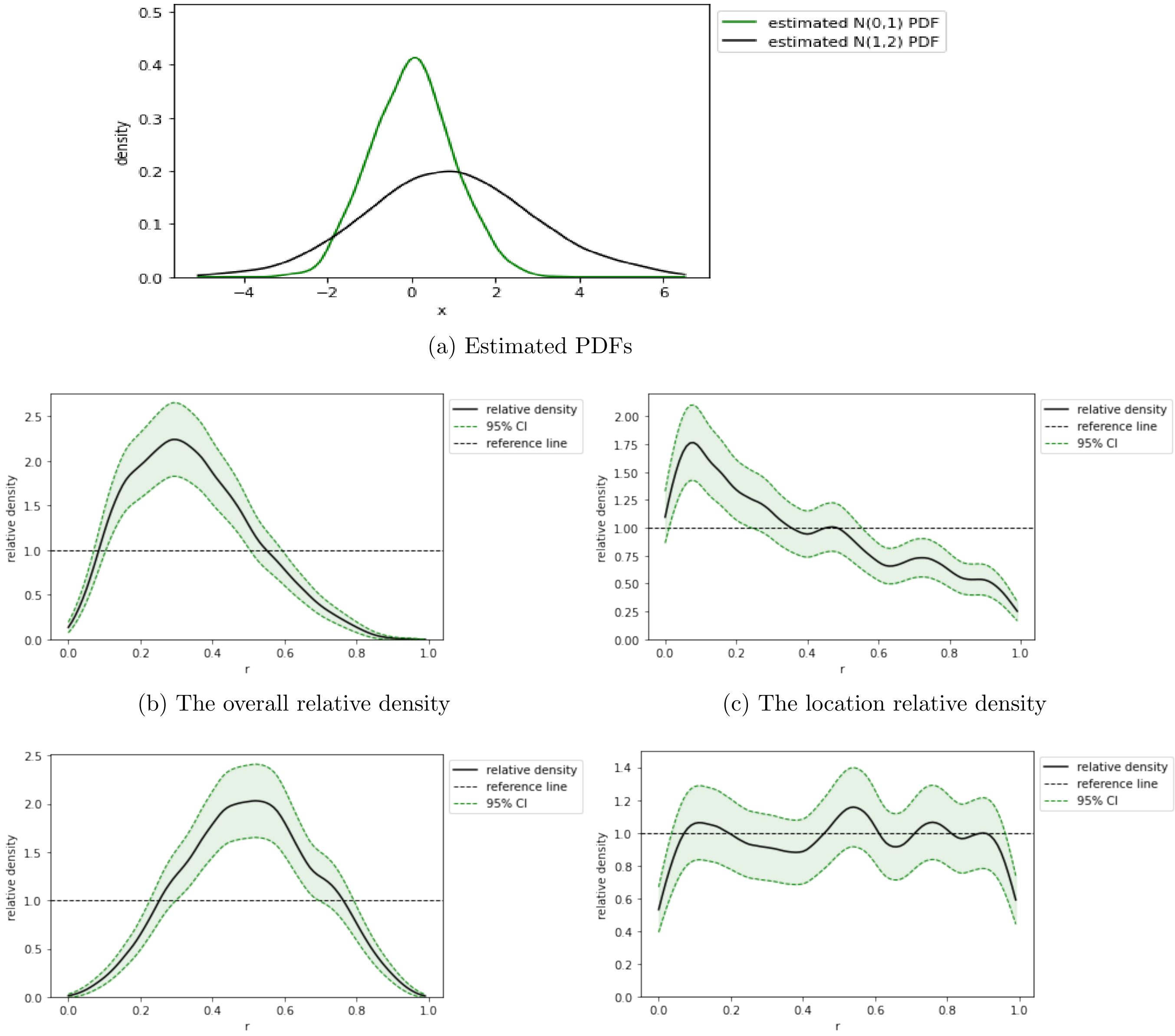


(a) HT estimate of the CDfs

(b) The difference between the estimated CDFs

(c) The PoA

Figure 5.8: (a) Horvitz-Thompson estimate of the CDFs of the observed samples ($n = m = 1{,}000$) from N(0,1) and N(1,2) with their corresponding 95% CIs. The 95% CIs are constructed based on asymptotic normality of the Horvitz-Thompson estimator. (b) The difference between the estimated CDFs. The 95% CI is constructed the same way as in (a). (c)Estimated PoA with $\delta = 0.05$ and 95% bootstrap and delta method CIs.



(a) Estimated PDFs

(b) The overall relative density

(c) The location relative density

(d) The scale relative density

(e) The residual relative density

Figure 5.9: (a) The KDEs of the PDFs of the reference and comparison samples ($n = m = 1{,}000$) generated from from N(1,2) and N(0,1) respectively. (b), (c), (d), and (e) are the KDEs of the overall, location, scale, and residual relative densities with corresponding 95% CIs respectively. All CIs are constructed based on the asymptotic normality of the KDE.

Chapter 6

Two-Sample Evaluation

6.1 Explaining the Design of the Simulation Study

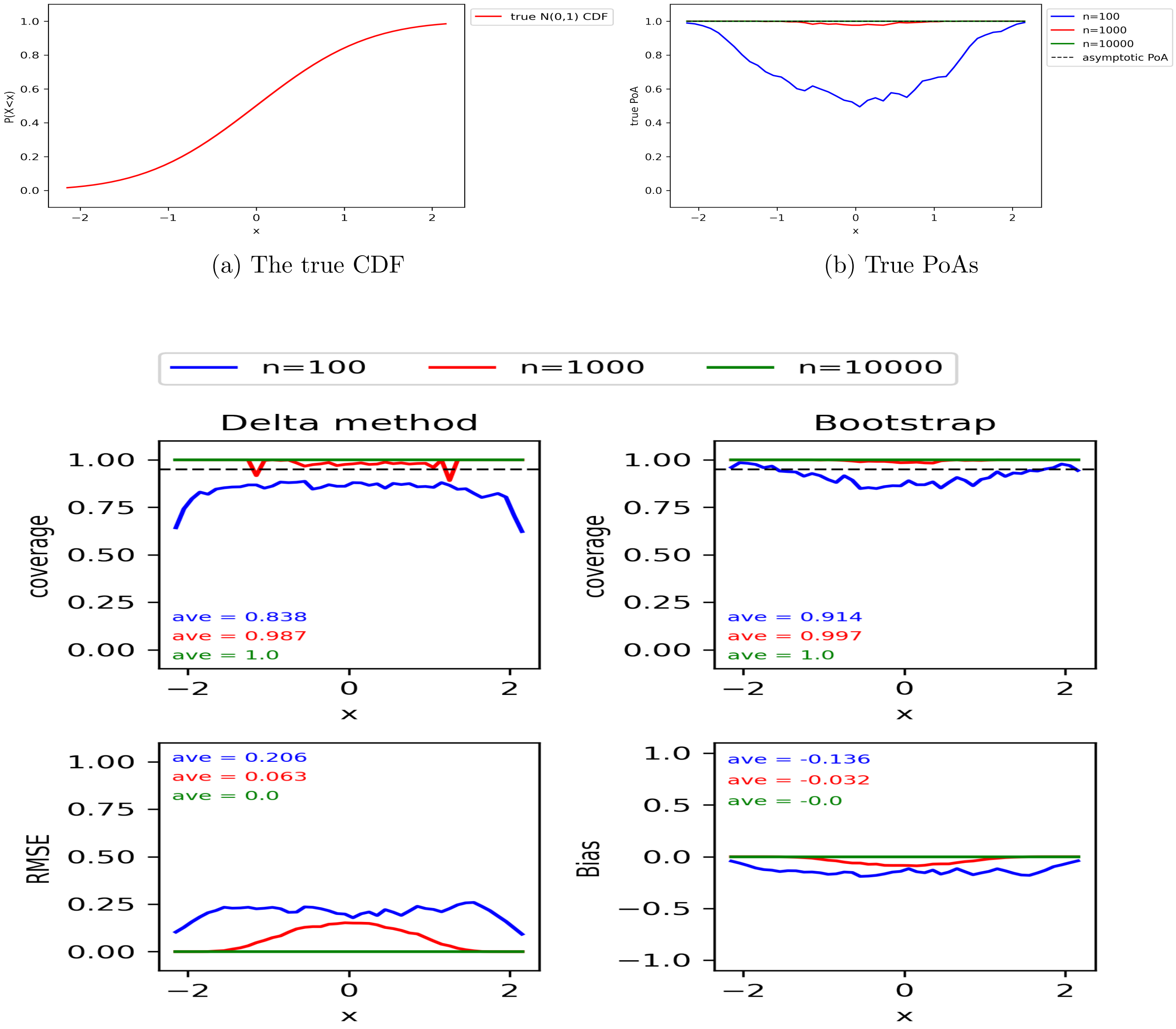
In this chapter, we conduct a similar simulation study to the one in Chapter 3, to examine the coverage of the proposed CIs for the PoA and also to study the bias and root mean squared error (RMSE) of the PoA estimator, in the two-sample setting. We have considered nine different scenarios: i. N(0,1) versus N(0,1), ii. N(0,1) versus N(1,1), iii. N(0,1) versus N(0,2), iv. N(0,1) versus N(1,2), v. Gamma(2,2) versus Gamma(2,2), vi. Gamma(4,1) versus Gamma(1,2), vii. Gamma(2,1) versus Gamma(1,2), viii. Gamma(2,2) versus Gamma(1,2), and ix. N(1,5) vs IG(1,0.2). The intuition behind choosing these scenarios is to have an example where the underlying distributions are the same, one where there is only a difference in mean, one where the only difference is in variance, and one where we have differences in both mean and variance. The last scenario N(1,5) vs IG(1,0.2) is considered as an example where the mean and variance of the underlying distributions are the same but the distributions themselves are very different. We also consider gamma distributions to verify that the methodology is not limited to normal data. We define the populations $\mathcal{P}$ and $\mathcal{Q}$ by taking $N = 20{,}000$ draws from the specified distributions and we take simple random samples from each population. As in Chapter 3, we considered three different sample sizes: $m = n = 100$, $m = n = 1{,}000$, and $m = n = 10{,}000$ to be able to examine the effect of an increase in the sample sizes. For a given sample size, we take simple random samples from $\mathcal{P}$ and $\mathcal{Q}$ and using this data we estimate the CDFs, the PoA, and the corresponding CIs. We repeat this $J = 1{,}000$ times. Thus, at a given point x, we have $J = 1{,}000$ CIs and we define coverage to be the proportion of these $J = 1{,}000$ CIs that cover the true PoA. The true PoA here is defined similar to in Chapter 3, except that we replace the known CDF in Equation 3.1 with the Horvitz-Thompson estimate of the reference CDF calculated from the second observed sample

$$\theta_n(x) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}\{|\hat{F}^j_{HT}(x) - \hat{G}^j_{HT}(x)| \leq \delta\}.$$

The bias and RMSE are calculated exactly as in Equations 3.2 and 3.3 while we consider nine different scenarios, the results obtained from these scenarios were very similar, hence we only present two of them here in the main text: N(0,1) vs N(0,1) and N(0,1) vs N(1,2). The results of the other seven scenarios are provided in Appendix D.

6.2 N(0,1) vs N(0,1)

In this scenario we have two populations, both characterized by the N(0,1) distribution, and we want to determine whether the two observed samples apear to come from a single distribution. Similar to Chapter 3, we summarize the simulation results with six plots: the true CDFs, the true PoAs, the coverage of the delta method- and bootstrap-based CIs, the RMSE, and the bias for each of the sample sizes ($m = n = 100, 1{,}000,$ and $10{,}000$). These plots, for this scenario, are shown in Figure 6.1. Note that in this scenario the underlying true CDFs are the same and thus we have only one CDF plotted in Figure 6.1a and there is no transition between 0 and 1 in Figure 6.1b. This simplifies interpretation because here we do not have the transition effect (that was explained in Section 3.2). The results in Figure 6.1c look very similar to what we saw in the one-sample evaluation study (compare Figure 6.1c and Figure 3.3c, for example). We still see improved performance in all criteria with larger sample sizes; the bootstrap coverage looks like an upward-scaled version of the delta method coverage; and bias and RMSE are negligible for reasonably large samples. The same reasoning that we provided in Section 3.2 can still be used to explain why the overall shape of the coverage plots looks so similar: the two methods of CI construction differ only in how the standard error of the PoA is estimated. As in the one-sample problem, the bootstrap approach yields wider intervals and hence higher coverage.



(a) The true CDF

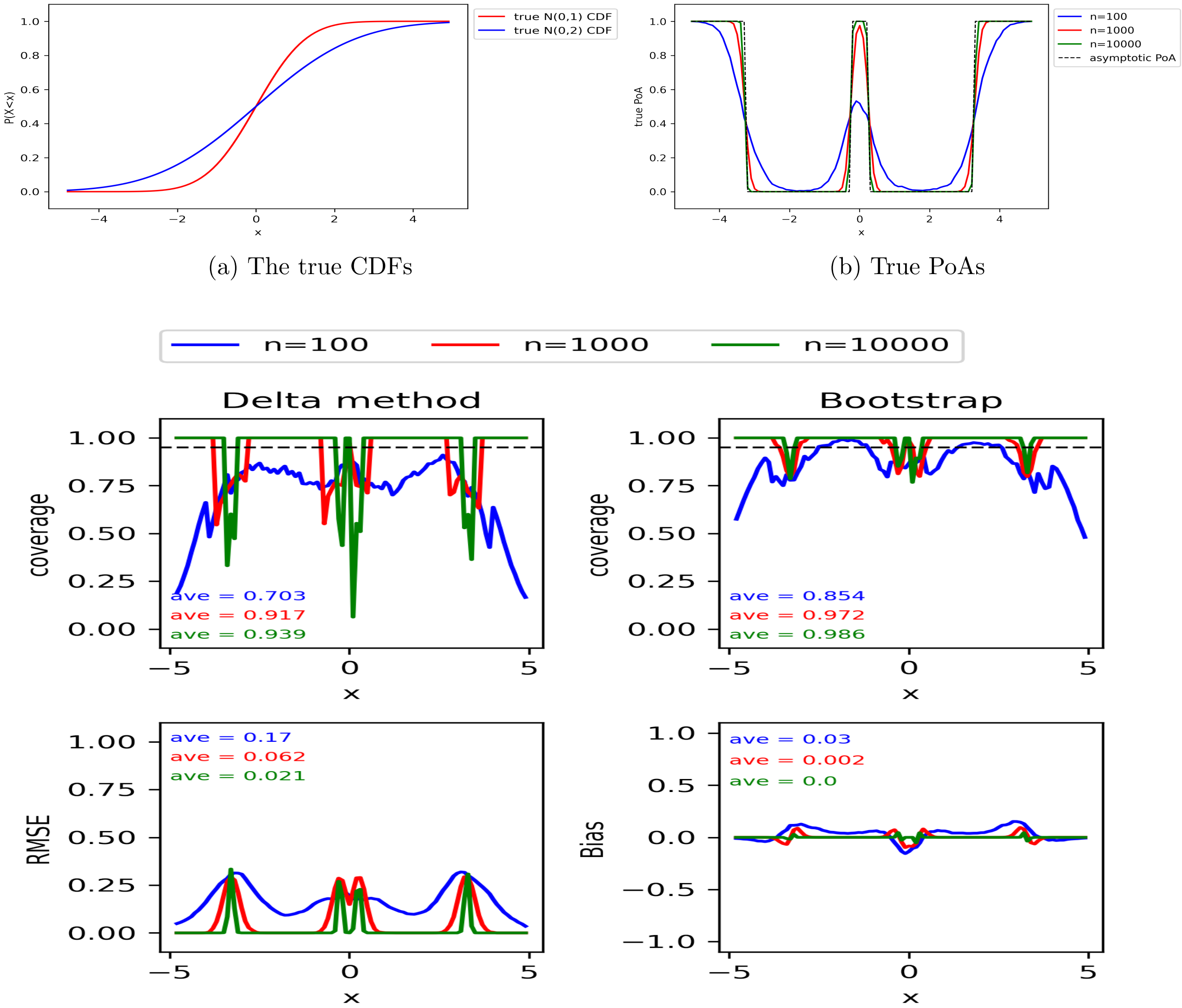
(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure 6.1: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(0,1) vs N(0,1).

6.3 N(0,1) vs N(0,2)

Here we consider a scenario in which the underlying distributions are not truly the same and the difference is in their variances. Nothing new is learned by examining this scenario, relative to what has already been learned in the simulations discussed so far. However, we include this scenario to illustrate that the transition effect is not just a characteristic of the one-sample problem and it can be seen in two-sample problems as well. If we compare the true PoAs in Figure 6.2b with the results shown in Figure 6.2c, the transition effect is clearly visible. While we have good overall performance (especially with larger sample sizes of 1,000 and 10,000), the performance is worst for values of x when $|F(x) - G(x)|$ transitions between being less than δ and greater than δ. As before, the larger the sample size the more severe but more fleeting the transition effect.



(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure 6.2: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(0,1) vs N(0,2).

6.4 General Insights Drawn

The results of the two scenarios considered in this chapter (and also the seven scenarios considered in Appendix D, where we find results very similar to the ones discussed here), lead us to the following conclusions:

- The bootstrap-based CI has consistently higher coverage than the delta method-based CI across different scenarios and different sample sizes.
- With larger sample sizes ($n = m =$ 1,000 and 10,000) the coverage of the both types of CIs are reasonably high, though they may perform poorly in transition areas (if there are any). With the small sample size ($n = m = 100$) we still get adequate average coverage.
- The RMSE and bias of the PoA estimator is almost zero for large sample sizes ($n = m =$ 1,000 and 10,000) meaning that the PoA estimator accurately and precisely estimates the true PoA. When the sample size is small ($n = m = 100$), however, we have non-zero RMSE and bias but they are still reasonably small.

Note that these are the same conclusions drawn in the one-sample evaluation. As such, we find that the properties and performance of the PoA methodology is consistent across both one- and two- sample applications.

Chapter 7

Conclusions and Future Work

In this thesis, we developed a methodology for the one- and two-sample comparison of distributions. More precisely, we have adapted the probability of agreement (PoA) methodology for use when one wants to compare the distribution of an observed sample with a known distribution (one-sample case), or when interest lies in comparing the distributions of two independently observed samples (two-sample case). Our proposed methodology overcomes the issues of the hypothesis testing-based approaches that are commonly applied in these circumstances. In particular, the PoA methodology provides more information about the similarities and differences between the distributions under comparison, its interpretation is straightforward, and it is based on practical equivalence rather than statistical significance. We also suggest using relative density plots as a supplementary graphical tool to help diagnose why the distributions under comparison are different if the PoA analysis suggest they are different. We showed how one should apply the PoA and the relative density and how one should interpret the resulting plots, through several real-world and simulated examples.

We have used a large sample normal theory approach for estimating the PoA and we have considered two different methods for calculating confidence intervals for the PoA: a large sample normal theory method in conjunction with the delta method and a bootstrap-based method. We have examined the coverage rate of both methods in several simulation studies and have found that the bootstrap-based method for constructing CIs performs consistently better in all simulation scenarios in both one- and two-sample cases, especially when the sample sizes are small (samples of size 100 in our simulation studies). For bigger sample sizes (we considered samples of size 1,000 and 10,000) both methods perform quite well. We also found that the RMSE and bias of the PoA estimator decrease as the sample sizes increases, however even with the smallest sample sizes in our simulation ($n = 100$ in one-sample case and $m = n = 100$ in two-sample case) the bias and RMSE values were reasonable. Therefore we conclude that even with small sample sizes, the PoA estimator is able to accurately and precisely identify whether the distributions under comparison are practically equivalent.

There are of course several ways that this work can be extended in the future. First, the PoA methodology could be generalized for the k-sample ($k > 2$) problem. That is, we intend to make the PoA methodology applicable for the situation in which we seek to compare k samples while accounting for the multiple comparison issue.

Another useful extension would be to adapt the methodology to potentially account for a non-constant value of δ, making the methodology suitable for a wider range of applications. This acknowledges that the indifference interval $[-\delta, \delta]$ need not be constant across the whole range of x. For instance, in our examples we saw that because all CDFs start at 0 and end at 1, the variance of an estimated CDF is smallest in the tails. Therefore, it may be reasonable to consider smaller values of δ in the tails and larger values of δ in the middle of the support set of the estimated CDFs.

Another extension, relevant just to the one-sample case, is to account for the uncertainty associated with estimating the parameters of the reference distribution. Here we use the observed sample (from the comparison distribution) to estimate the unknown parameters of the reference distribution. We then treat these values as known, and ignore the fact that they are in fact estimates. Accounting for uncertainty in this estimation is an important consideration that we intend to consider in future work. Doing so may provide a more accurate estimate of the PoA. The challenge is that, unlike in the two-sample case, the uncertainty associated with both estimated CDFs is the observed sample, so the estimates are not independent.

In Chapter 2 we justified the use of Horvitz-Thompson estimation because it could accommodate sampling mechanisms other than simple random sampling, but we have not explored that here. In future work we intend to investigate the PoA methodology with sampling designs other than simple random sampling. It will be important to conduct a proper simulation study to evaluate the performance of the PoA methodology when the sampling design is something other than simple random sampling, such as stratified or cluster-based random sampling, for example.

As mentioned several times throughout the thesis the version of the KDE used in this thesis has a downward bias near 0 and 1. There are however some ways to overcome this bias, and considering them in the future may result in a more accurate relative density plots. For example, Cwik and Mielniczuk (1993) propose a boundary kernel estimator which uses the reflection method to overcome the downward bias of the KDE. They also provide asymptotic normality results that can be used for the calculation of the CIs. Alternatively one can also construct bootstrap-based CIs.

Finally, we acknowledge that sometimes a single-number summary can be useful, and may serve as a useful supplement to the PoA plot. One such summary that can be easily calculated alongside the existing calculations, is a weighted average of the estimated PoA values across the range of $x \in \mathcal{A}$. More precisely, we could use

$$\frac{\sum_{x \in \mathcal{A}} w(x)\hat{\theta}(x)}{\sum_{x \in \mathcal{A}} w(x)},$$

where $w(x)$ is a weight function that can be chosen by the practitioner according to their specific problem, to give differential weight to different values in the support set.

References

W. Alexander. *Boundary Kernel Estimation of the Two-Sample Comparison Density Function.* PhD thesis, Texas A M University, College Station, Texas, USA,, 1989.

T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

T. W. Anderson and D. A. Darling. A test of goodness-of-fit. *Journal of the American Statistical Association*, 49:765–769, 1954.

C. M. Anderson-Cook and C. M. Borror. The difference between 'equivalent' and 'not different'. *Quality Engineering*, 28(3):249–262, 2016.

Y. G. Berger. Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference*, 67:209–226, 1998.

H. Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1: 13–74, 1928.

J. Cwik and J. Mielniczuk. Data-dependent bandwidth choice for a grade kernel estimate. *Statistics and Probability Letters*, 16:397–405, 1993.

C. S. Davis and M. A. Stephens. The covariance matrix of normal order statistics. *Communications in Statistics - Simulation and Computation*, 6:1:75–81, 1977.

J. L. Doob. The limiting distributions of certain statistics. *Annals of Mathematical Statistics*, 6:160–169, 1935.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

B. Efron and T. Hastie. *Computer Age Statistical Inference.* Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA, 2016.

R. L. Eubank, V. N. Lariccia, and R. B. Rosenstein. Test statistics derived as components of pearson's phi-squared distance measure. *Journal of the American Statistical Association*, 82:816–825, 1987.

T. Gasser and H-G. Muller. *Smoothing Techniques for Curve Estimation.* Springer-Verlag, Berlin, 1979.

M. Handcock and M. Morris. *Relative Distribution Methods in Social Siences.* SpringerVerlag, New York, USA, 1999.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

V. Karalis and P. Macheras. Current regulatory approaches of bioequivalence testing. *Expert opinion on drug metabolism toxicology*, 8(8):929–942, 2012.

A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91, 1933.

G. Li, R. Tiwari, and M. Wells. Quantile comparison functions in two-sample problems with applications to comparisons of diagnostic markers. *JASA,*, 91:689–698, 1996.

D. McKay. *Handbook of research methods in abnormal and clinical psychology.* Sage, 2008.

J. Mielniczuk. Grade estimation of kullback-leibler information number. *Probability and Mathematical Statistics*, 13:139–147, 1992.

E. Parzen. Nonparametric statistical data modeling (with discussion). *JASA,*, 74:105–131, 1979.

E. Parzen. *FUN.STAT: Quantile approach to two sample statistical data analysis.* Technical report, Texas A M University, College Station, Texas, USA,, 1983.

E. Parzen. *Statistical Methods, Mining, Two-Sample Data Analysis, Comparison Distributions, and Quantile Limit Theorems. In Asymptotic Methods in Probability and Statistics.* Elsevier, Amsterdam, The Netherlands, 1999.

S. D. Patterson and B. Jones. *Bioequivalence and statistics in clinical pharmacology.* Chapman and Hall/CRC, 2017.

E. S. Pearson and H. O. Hartley. Biometrika tables for statisticians. *Cambridge University Press*, page 117–123, 1972.

A. N. Pettitt. A two-sample anderson-darling rank statistic. *Biometrika*, 63:161–168, 1976.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York, Chichester, 1992.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika.*, 52 (3–4):591–611, 1965.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of Royal Statistical Society (Serie B)*, 53:683–690, 1991.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman Hall, 1986.

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics.*, 19(2):279–281, 1948.

M. A. Stephens. *Tests Based on EDF Statistics.* Marcel Dekker, New York, 1986.

N. T. Stevens and C. M. Anderson-Cook. Comparing the reliability of related populations with the probability of agreement. *Technometrics*, 59(3):371–380, 2017a.

N. T. Stevens and C. M. Anderson-Cook. Quantifying similarity in reliability surfaces with the probability of agreement. *Quality Engineering*, 29(3):395–408, 2017b.

N. T. Stevens and C. M. Anderson-Cook. Design and analysis of confirmation experiments. *Journal of Quality Technology*, 51(2):109–124, 2019.

N. T. Stevens and L. Lu. Comparing kaplan-meier curves with the probability of agreement. *Statistics in Medicine*, pages 4621–4635, 2020.

N. T. Stevens, S. H. Steiner, and R. J. MacKay. Assessing agreement between two measurement systems: an alternative to the limits of agreement approach. *Statistical Methods in Medical Research*, 26(6):2487–2504, 2017.

N. T. Stevens, R. J. MacKay, and S. H. Steiner. Comparing heteroscedastic measurement systems with the probability of agreement. *Statistical Methods in Medical Research*, 27(11):3420–3435, 2018a.

N. T. Stevens, S. E. Rigdon, and C. M. Anderson-Cook. Bayesian probability of predictive agreement for comparing the outcome of two separate regressions. *Quality and Reliability Engineering International*, 34(4):425–445, 2018b.

N. T. Stevens, S. E. Rigdon, and C. M. Anderson-Cook. Bayesian probability of agreement for comparing the similarity of response surfaces. *Journal of Quality Technology*, pages 67–80, 2019.

N. T. Stevens, L. Lu, C. M. Anderson-Cook, and S. E. Rigdon. Bayesian probability of agreement for comparing survival or reliability functions with parametric lifetime regression model. *Quality Engineering*, 32(3):312–332, 2020.

J. L. Szarka. Equivalence and noninferiority tests for quality, manufacturing and test engineers. *Journal of Quality Technology*, 46(4):378–380, 2014.

O. Thas. *Comparing Distributions.* Springer, 233 Spring Street, New York, NY 10013, USA, 2010.

R. E. von Mises. *Wahrscheinlichkeit, Statistik und Wahrheit.* Julius Springer., 1928.

R. L. Wasserstein and N. A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108.

R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19, 2019. doi: 10.1080/00031305.2019.1583913.

S. Wellek. *Testing statistical hypotheses of equivalence and noninferiority.* Chapman and Hall/CRC Press, 2010.

APPENDICES

Appendix A

The PoA for the One-Sample Relative Density

In this section we discuss an adaptation of the PoA methodology to the relative density. While this is theoretically feasible, as we describe below, there are some practical concerns when it comes to using the PoA for the relative density that limit its utility in practice.

We define the PoA in this setting to compare the relative density to the constant value 1 as

$$\theta(r) = P(-\delta \leq \tilde{h}_n(r) - 1 \leq \delta).$$

We can interpret $\theta(r)$ as the probability that the relative density is practically equivalent to 1 at point r. In other words, $\theta(r)$ may be interpreted as the probability that f and g are practically equivalent. We can use the asymptotic normal distribution given in Equation

2.14 to approximate the PoA

$$\begin{aligned}\theta(r) &= P(-\delta \le \tilde{h}_n(r) - 1 \le \delta) \\ &\cong \Phi\left(\frac{\delta - (h(r)-1)}{\sqrt{\frac{h(r)R(K)}{nb_n}}}\right) - \Phi\left(\frac{-\delta - (h(r)-1)}{\sqrt{\frac{h(r)R(K)}{nb_n}}}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. However, because we do not know $h(r)$, the above asymptotic approximation cannot be calculated. In practice we need to replace $h(r)$ with its kernel density estimate $\hat{h}_n(r)$ to obtain an estimate of th PoA in this case

$$\hat{\theta}(r) = \Phi\left(\frac{\delta - (\hat{h}_n(r)-1)}{\sqrt{\frac{\hat{h}_n(r)R(K)}{nb_n}}}\right) - \Phi\left(\frac{-\delta - (\hat{h}_n(r)-1)}{\sqrt{\frac{\hat{h}_n(r)R(K)}{nb_n}}}\right).$$

We can also build CI's for this version of the PoA with either the bootstrap method or the delta method, just as we did for the Horvitz-Thompson-based PoA in Sections 2.1.4 and 2.1.5. However, for the reason described below, we do not pursue the details of these methods here.

As we have shown, the PoA can easily be defined and calculated, but this relies on a practitioner being able to specify a meaningful value for δ. As we have seen, the conclusions from a PoA analysis depend dramatically on δ. Because it is difficult to choose the value of δ in an intuitive manner here, we do not recommend applying the PoA method on the relative density. Nonetheless the relative density is still a useful graphical tool that can complement a PoA analysis that compares CDFs.

Appendix B

The PoA for the Two-Sample Relative Density

Similar to the one-sample case, given the asymptotic result in Equation 5.12, we can define a relative density-based PoA for the two-sample case

$$\begin{aligned}\theta(r) &= P(-\delta \leq \tilde{h}_{n,m}(r) - 1 \leq \delta) \\ &\approx \Phi\left(\frac{\delta - (h(r)-1)}{\sqrt{\frac{h(r)R(K)}{nb_n} + \frac{h^2(r)R(K)}{mb_n}}}\right) - \Phi\left(\frac{-\delta - (h(r)-1)}{\sqrt{\frac{h(r)R(K)}{nb_n} + \frac{h^2(r)R(K)}{mb_n}}}\right),\end{aligned}$$

which is estimated by substituting the kernel density estimate $\hat{h}_{n,m}(r)$ for $h(r)$

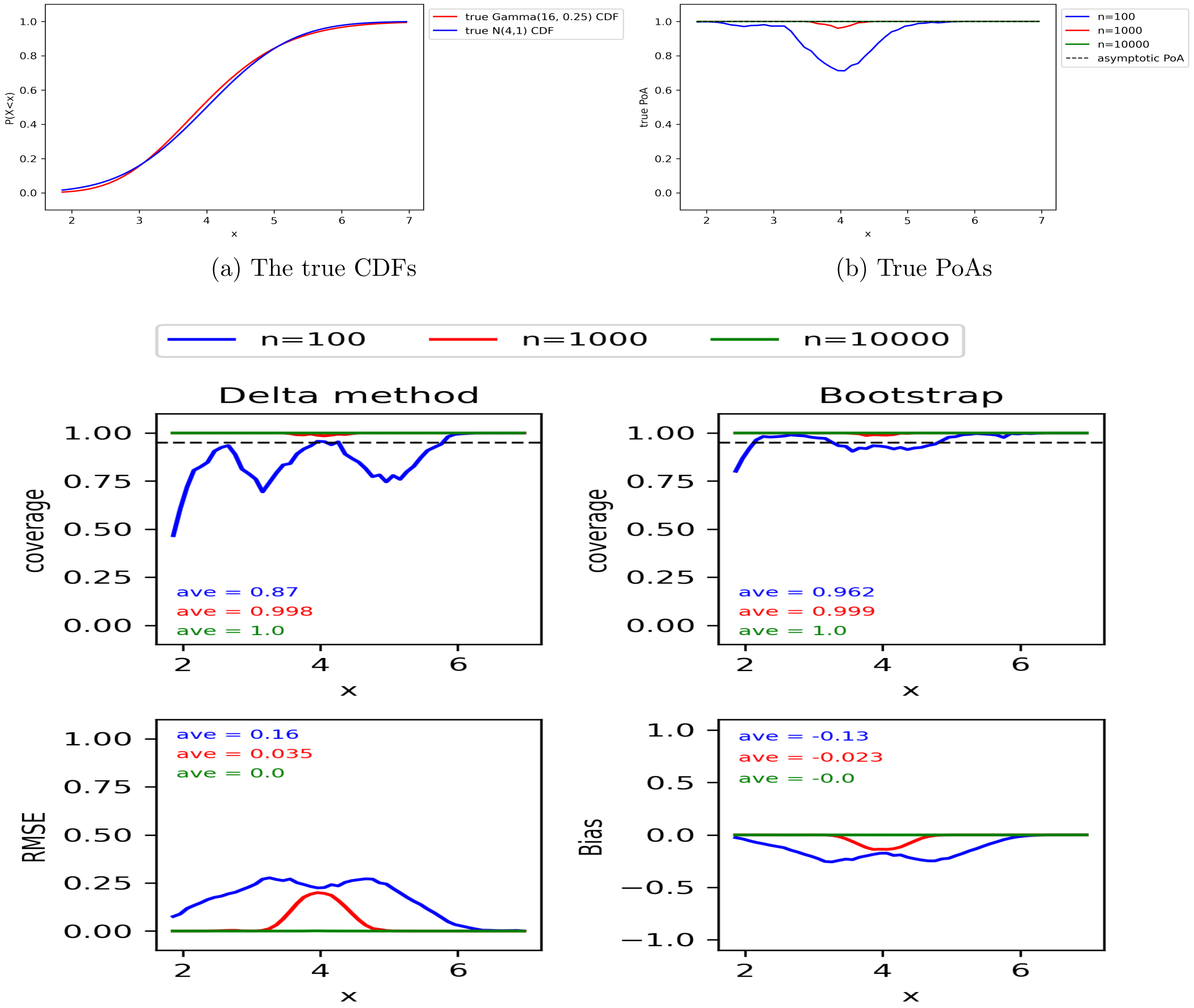
$$\begin{aligned}\hat{\theta}(r) &= P(-\delta \leq \tilde{h}_{n,m}(r) - 1 \leq \delta) \\ &\approx \Phi\left(\frac{\delta - (\hat{h}_{n,m}(r) - 1)}{\sqrt{\frac{\hat{h}_{n,m}(r)R(K)}{nb_n} + \frac{\hat{h}^2_{n,m}(r)R(K)}{mb_n}}}\right) - \Phi\left(\frac{-\delta - (\hat{h}_{n,m}(r) - 1)}{\sqrt{\frac{\hat{h}_{n,m}(r)R(K)}{nb_n} + \frac{\hat{h}^2_{n,m}(r)R(K)}{mb_n}}}\right).\end{aligned}$$

However, the same practical issue arises that was discussed in Appendix A: the choice of δ is not intuitive. Recall that, the interval $(-\delta, \delta)$ should be thought of as an indifference region and if the difference between the estimated relative density and the constant value 1 is in this interval we consider f and g practically equivalent, and the PoA calculates the probability of this practical equivalence at a given value r. However, if an appropriate value of δ is not clear to a practitioner, then the PoA methodology is not useful. Therefore, as we did in the one-sample case, we suggest that the PoA method should be used only to compare CDFs and, the relative density methodology can be used as a supplementary and useful graphical tool.

Appendix C

One-Sample Evaluation Studies

C.1 N(4,1) vs Gamma Family



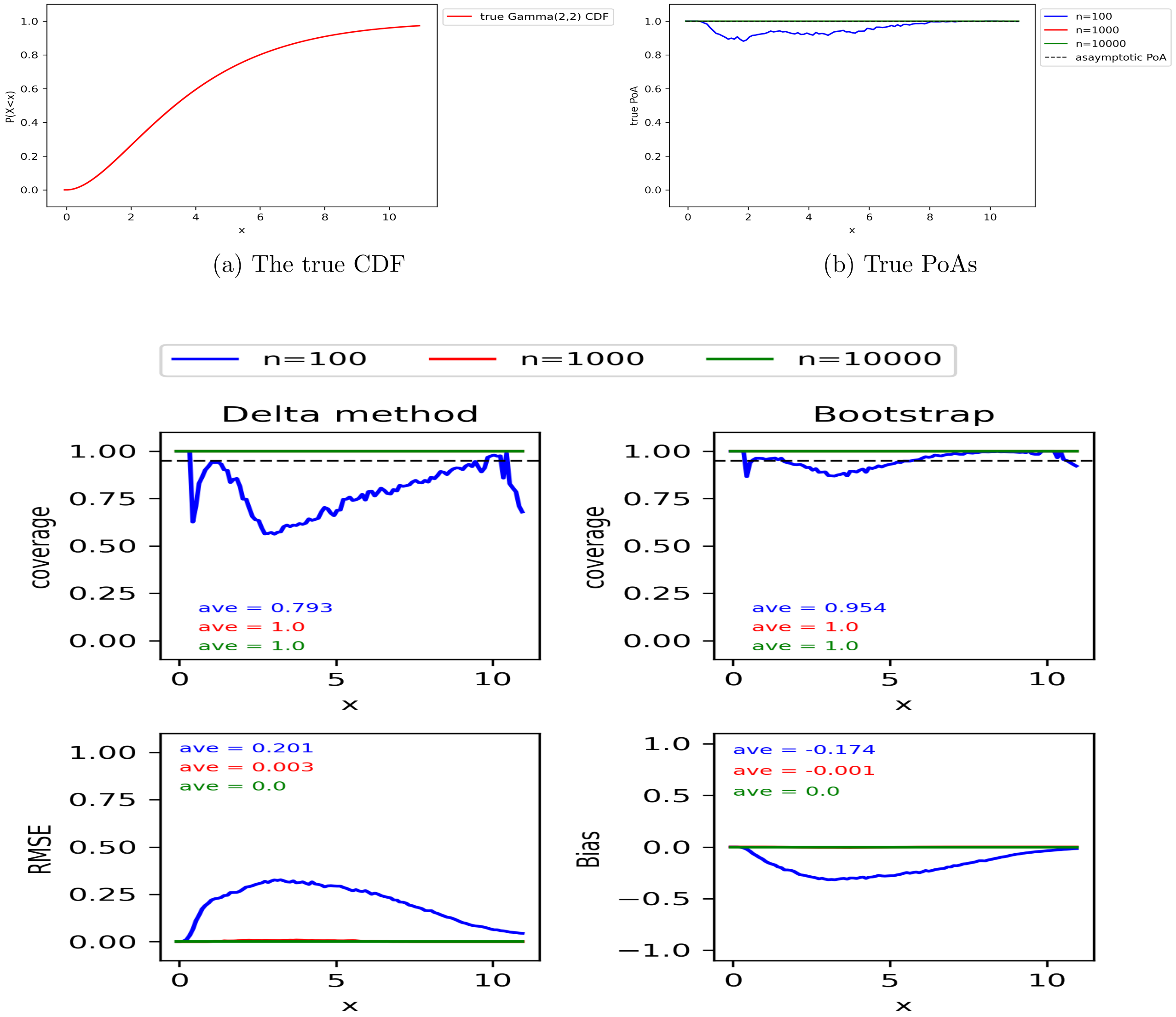
(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure C.1: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(4,1) vs gamma family.

C.2 Gamma(2,2) vs Gamma Family



(a) The true CDF

(b) True PoAs

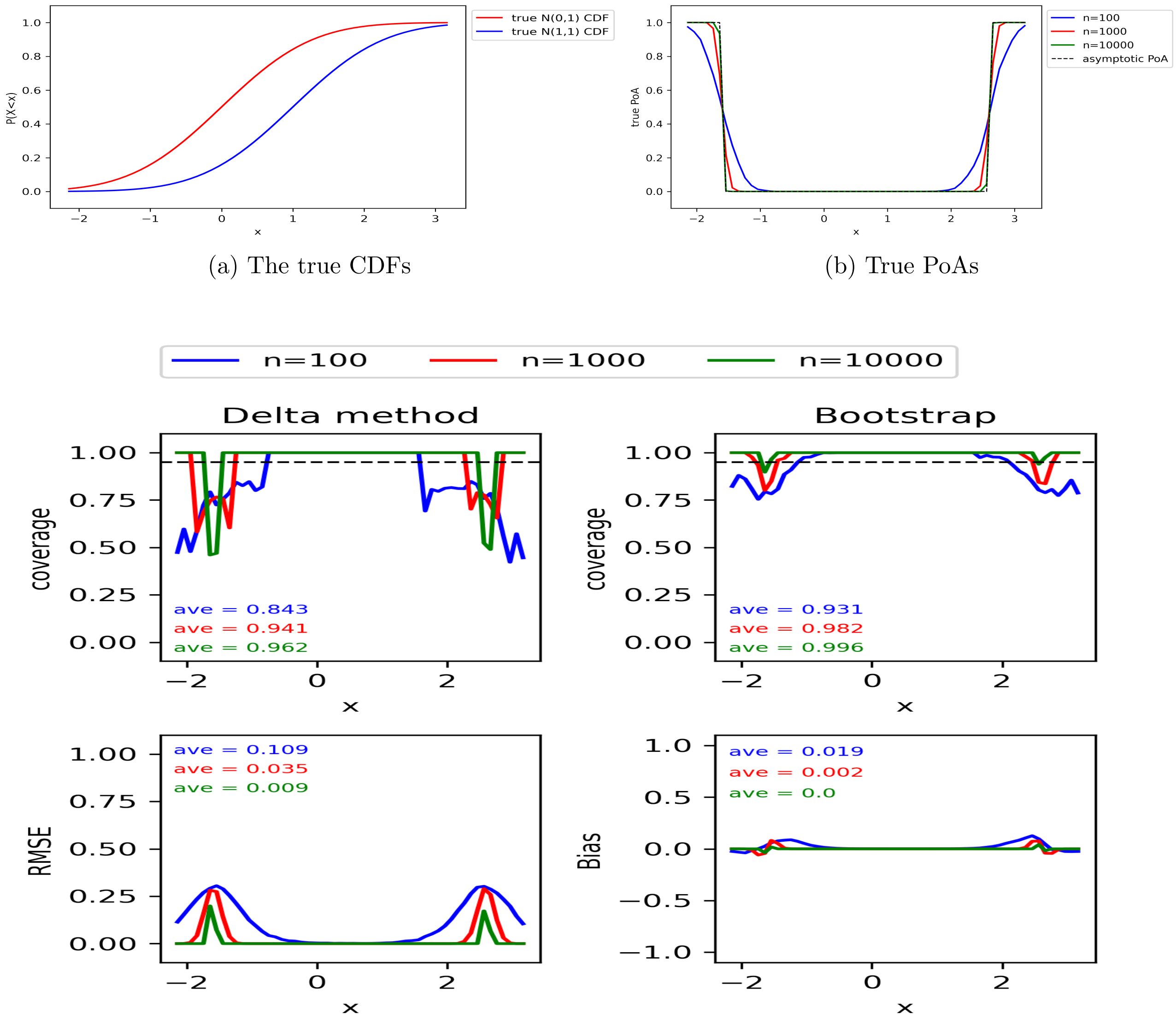
(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure C.2: (a) The true versions of the CDFs under comparison (in this scenario they are overlapping). (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario Gamma(2,2) vs gamma family.

Appendix D

Two-Sample Evaluation Studies

D.1 N(0,1) vs N(1,1)



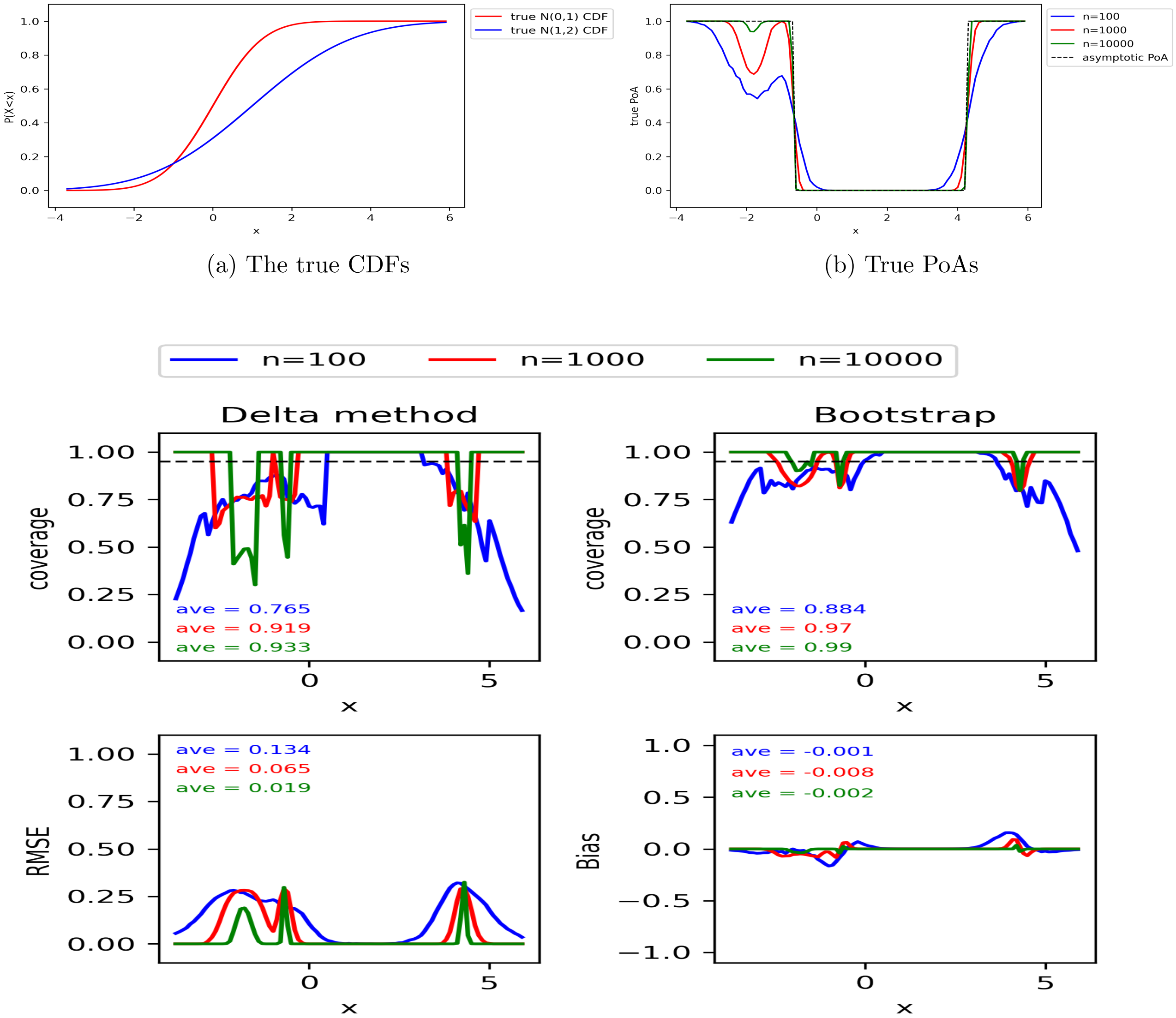
(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.1: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(0,1) vs N(1,1).

D.2 N(0,1) vs N(1,2)



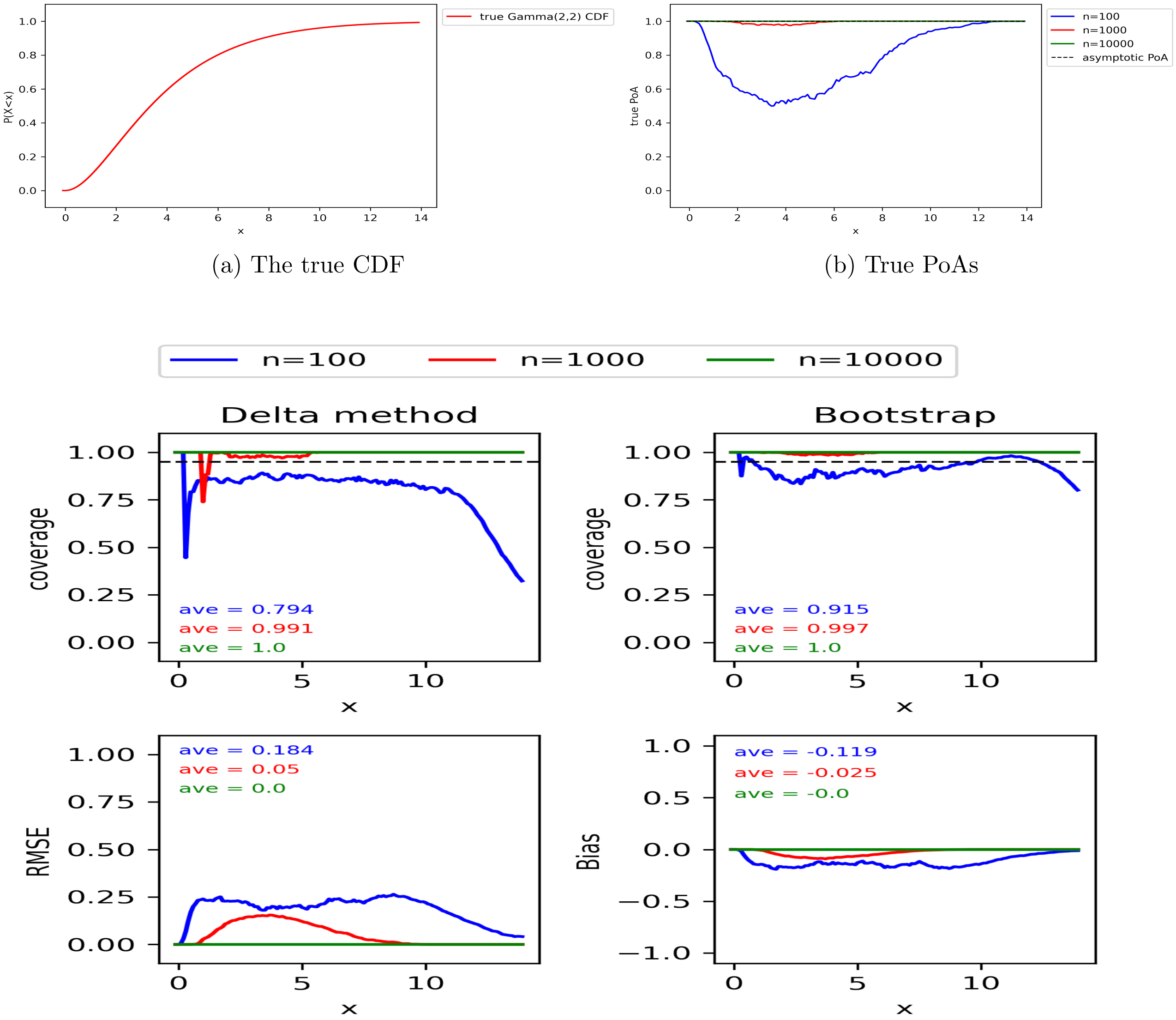
(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.2: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(0,1) vs N(1,2).

D.3 Gamma(2,2) vs Gamma(2,2)



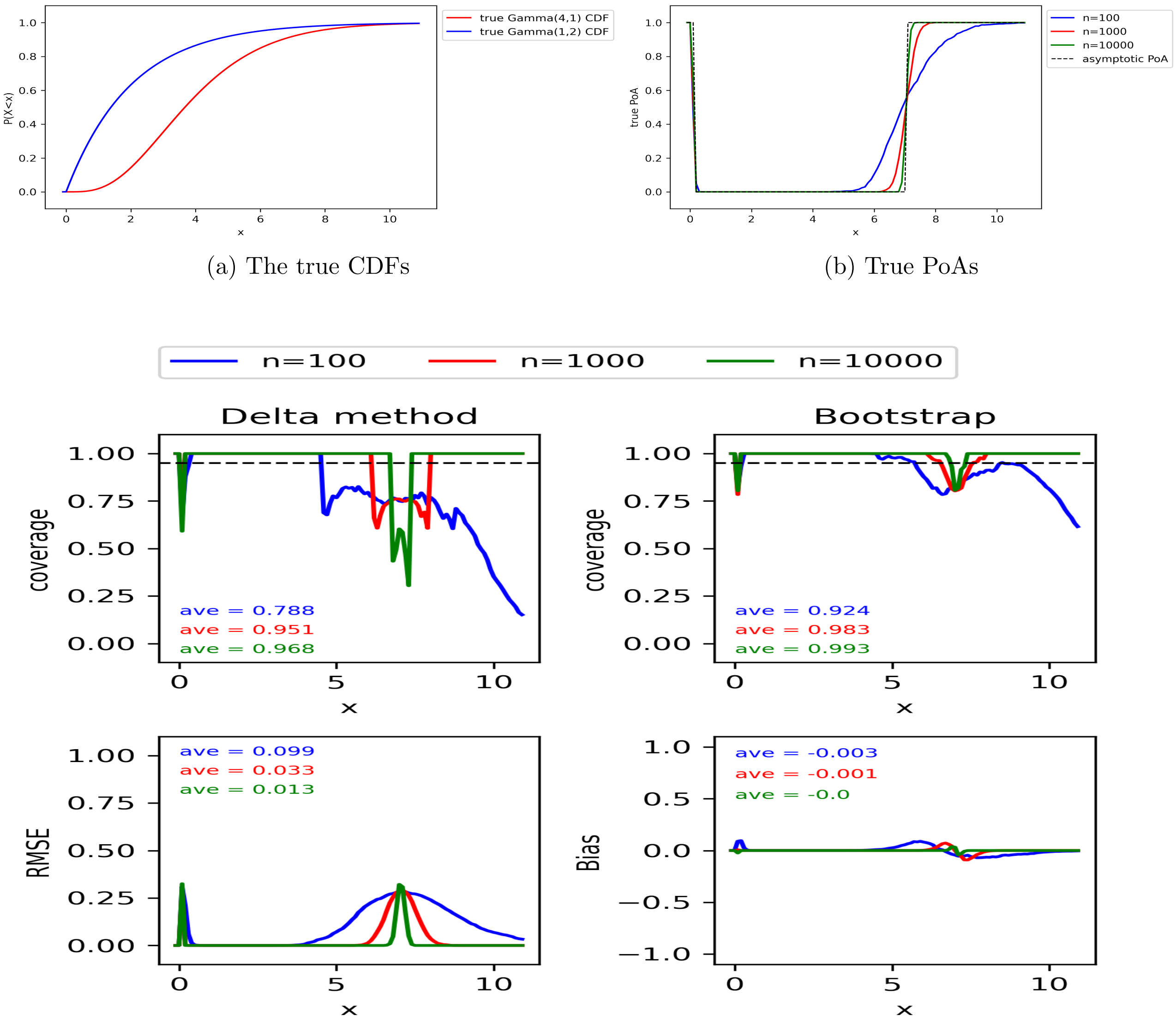
(a) The true CDF

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.3: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario Gamma(2,2) vs Gamma(2,2).

D.4 Gamma(4,1) vs Gamma(1,2)



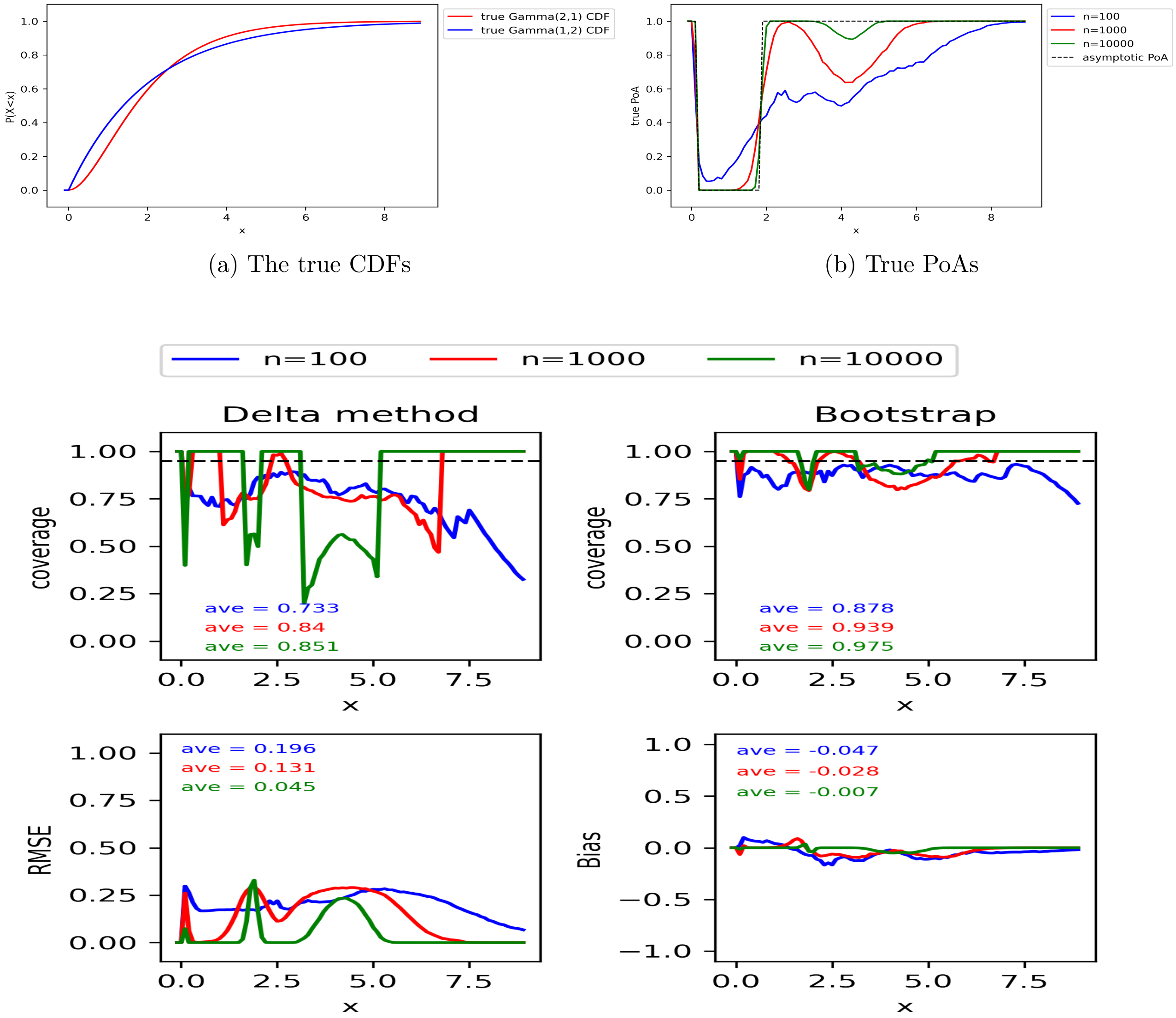
(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.4: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario Gamma(4,1) vs Gamma(1,2).

D.5 Gamma(2,1) vs Gamma(1,2)



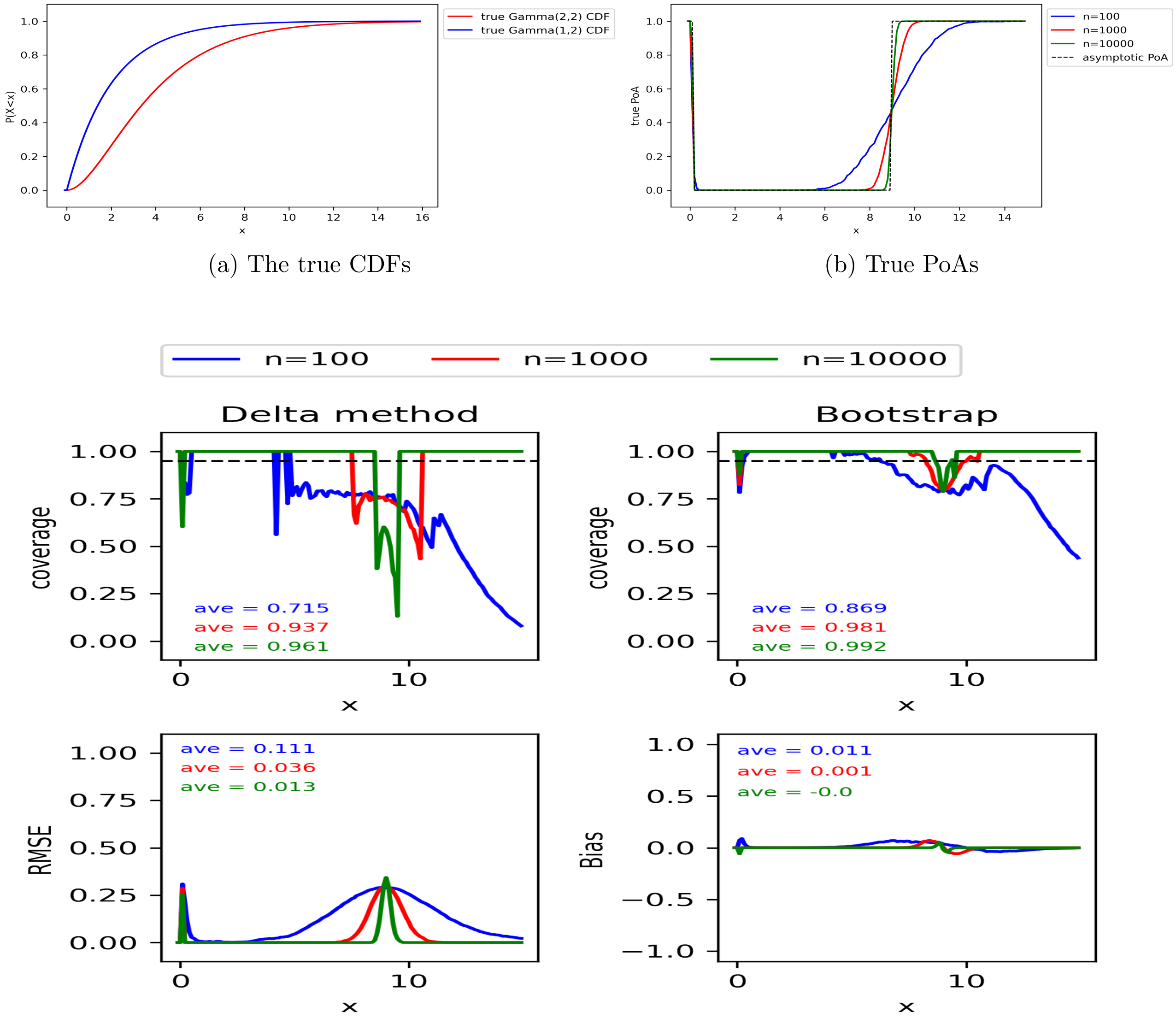
(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.5: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario Gamma(2,1) vs Gamma(1,2).

D.6 Gamma(2,2) vs Gamma(1,2)



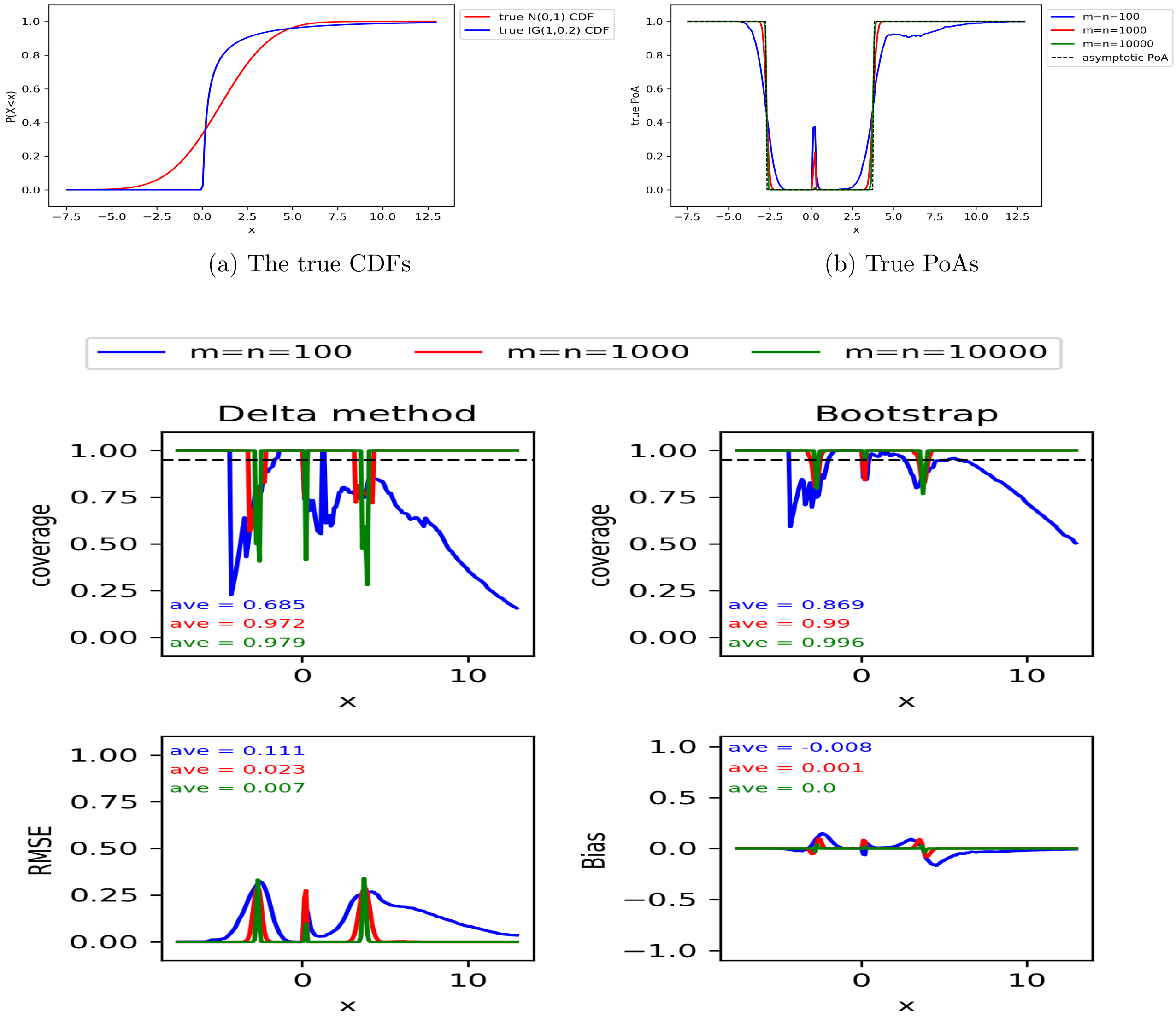
(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.6: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario Gamma(2,2) vs Gamma(1,2).

D.7 N(1,5) vs IG(1,0.2)



(a) The true CDFs

(b) True PoAs

(c) The bootstrap and delta method CI coverage, the bias, and the RMSE

Figure D.7: (a) The true versions of the CDFs under comparison. (b) The true and asymptotic PoAs. (c) The results of the simulation study of the scenario N(1,5) vs IG(1,0.2).