应用概率统计　第 40 卷
第 2 期　2024 年 4 月

Chinese Journal of Applied Probability and Statistics
Apr., 2024, Vol. 40, No. 2, pp. 229-263
doi: 10.3969/j.issn.1001-4268.2024.02.002

Selective Review of Biased Sampling Problems with Applications in Modern Statistics

QIN Jing
(National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD 20892, USA)

Abstract: Biased sampling is a pervasive issue that transcends various disciplines, impacting fields such as econometrics, epidemiology, medicine, survey research, and more recently, machine learning and artificial intelligence (AI). This ubiquitous challenge arises when the selection of data points for analysis or research introduces systematic biases, potentially compromising the accuracy and reliability of research outcomes. In this paper, our objective is to provide a comprehensive overview of the foundational concepts related to biased sampling problems and the methods of inference. Furthermore, we aim to establish a connection between biased sampling issues and the more recent discussions in machine learning regarding distribution shift problems. Additionally, we will delve into the latest advancements in biased sampling, particularly within the context of transfer learning and conformal inference for predictive confidence intervals. Our ultimate goal is to present this material in a manner that is accessible to graduate students, enabling them to identify applications of biased sampling problems within their own research endeavors.
It is with deep respect and gratitude that we dedicate this paper to the memory of the late Professor Shisong Mao, whose guidance and wisdom have been invaluable throughout the years.

Keywords: biased sampling problems; casual inference; conformal predictive interval; distributional shift; transfer learning; in memory Professor Shisong Mao

2020 Mathematics Subject Classification: 62D20; 62G20; 62G05

Citation: QIN J. Selective review of biased sampling problems with applications in modern statistics [J]. Chinese J Appl Probab Statist, 2024, 40(2): 229–263.

1 Interactions with Professor Shisong Mao

In the world of academia, the passing of a distinguished professor is a poignant moment, marked not only by the loss of an esteemed scholar but also by the legacy he or she leaves behind. It is a time when we come together to remember, honor, and celebrate the profound contributions of a remarkable individual whose work has indelibly shaped the field of statistics. This paper is a tribute to the enduring influence of Professor Shisong Mao, whose wisdom, dedication, and innovative thinking continue to inspire us. Professor

E-mail: jingqin@niaid.nih.gov.

Received November 9, 2023. Revised January 29, 2024.

Mao was not merely a statistician but a luminary whose work transcended the boundaries of traditional statistical analysis. With a passion for both the art and science of statistics, he broke new ground in research, teaching, and mentorship. His career was characterized by a relentless pursuit of knowledge, a commitment to statistical rigor, and an unwavering belief in the power of data to illuminate the mysteries of the world.

In the traditional academic setting, it's customary for each professor to mentor only a limited number of graduate students due to various constraints. Recognizing the critical shortage of statistics educators, the Chinese Department of Education took a proactive step by introducing a two-year program in 1984 to enroll college graduates for the study of statistics. This initiative aimed to address the growing demand for skilled statisticians in various sectors. Under the visionary leadership of Professor Mao, the Department of Statistics at East China Normal University embraced this educational challenge. In an exemplary move, the department opened its doors to a cohort of 24 special graduate students specializing in statistics. This significant addition complemented Professor Mao's ongoing mentorship of his two regular three-year graduate students. This bold and forward-thinking decision not only expanded the horizons for aspiring statisticians but also underscored Professor Mao's commitment to nurturing future talent in the field of statistics. It reflected his dedication to bridging the gap in statistics education and inspiring the next generation of statisticians in China.

My initial encounter with Professor Mao occurred during the spring of 1984. I had just received the news that I had successfully passed the preliminary graduate student entry examination, which marked the beginning of a transformative journey. My destination was East China Normal University, where I was to face the second round of oral examinations. This momentous occasion held great significance for me, not only in terms of my academic aspirations but also because it marked my first venture from a remote and relatively small town in Sichuan Province. The town, known as Wanzhou, would later become part of Chongqing Special District due to the construction of the Three Gorges Dam. It was a place where life moved at a slower pace, far removed from the bustling metropolis that awaited me in Shanghai, China's largest city. The transition from a tranquil and close-knit community to the dynamic and sprawling urban landscape of Shanghai was a monumental shift.

As an unsophisticated young middle school math teacher stationed in an isolated township, I embarked on my first-ever solo journey to a dynamic urban center. The mix of excitement and nervousness was palpable. The prospect of studying in Shanghai and pursuing my academic dreams was exhilarating, but it was also accompanied by a sense of the unknown. The city's vastness and the anonymity of its busy streets were both thrilling and intimidating. It was a stark contrast to the closely bonded community I was accustomed to. The cacophony of traffic, the towering skyscrapers, and the neon lights painted a picture of a world entirely different from what I had known. My situation indeed resonates with the description in the famous Chinese novel "Dream of the Red Chamber", where Grandma Liu's entry into the grand house brings her face-to-face with an entirely unfamiliar and overwhelming environment. In my case, the parallel may be even more pronounced, given the added challenge of grappling with a different dialect, the Shanghai accent. Much like the character in the story, my experience of joining the platform created by Professor Mao, with its innovative approach and a diverse group of graduate students, might have felt like stepping into a world filled with novel experiences and opportunities. Unquestionably, this adventure was a vital step in my personal and professional growth. It symbolized not only a geographical transition but a leap into the uncharted waters of higher education and self-discovery. The small-town math teacher was on the brink of a new chapter, eager to embrace the challenges and opportunities that the big city had to offer.

Professor Mao's reputation extends far and wide, earning him the respect and recognition of colleagues from universities across the academic landscape, especially within the field of statistics. As we had the privilege of meeting Professor Mao, I observed that many of my fellow students, representing various universities apart from East China Normal University, could extend warm regards from their own professors. However, when it came to my turn to exchange handshakes with Professor Mao, I found myself in a unique position. None of my professors had a prior connection with him, primarily due to the unfamiliarity between him and my undergraduate college. In light of this, I chose to express my personal admiration and warm regards, addressing Professor Mao with sincerity, "Professor Mao, please accept my heartfelt greetings." In response, Professor Mao warmly reciprocated with a smile, creating a brief yet meaningful connection that exceeded any prior lack of familiarity.

During my tenure at East China Normal University from 1984 to 1988, I frequently heard Professor Mao underscore the significance of nurturing a profound passion for the art of data collection. He advocated treating data with the same care and devotion one might reserve for a loved one. Moreover, he encouraged individuals to delve deeply into the data, allowing the information to organically weave its own narrative. In the autumn of 1987, just like any other departments, the statistics department was bustling with preparations for the school's upcoming anniversary celebration. As tradition dictated, a compelling lecture based on one's own research was the chosen way to contribute to this special occasion. However, during this particular period, I found myself immersed in a different endeavor — completing the rigorous application process for graduate school in North America. My days were filled with the arduous task of completing numerous application forms and preparing for the TOFF (Test of Foreign Language) examination. One day, while I was engrossed in these preparations, Professor Mao, a respected figure in the department, approached me with a question about my research. I couldn't help but feel a wave of embarrassment wash over me, for I had no research work to speak of at that point. It was an awkward moment as I explained my current circumstances to Professor Mao. In response, Professor Mao, a seasoned and understanding mentor, offered a reassuring gesture. "I can excuse you this time," he said with a kind smile, "but I hope this is the last time." This encounter with Professor Mao served as a pivotal moment in my academic journey. It was a gentle nudge, a reminder of the importance of research and the academic commitment that lay ahead. From that day forward, I embarked on a quest to delve into the world of statistical research, determined to ensure that it would indeed be the last time I found myself unprepared in the presence of my academic peers and mentors. Little did I know that this experience would serve as the catalyst for a rich and rewarding academic journey, one that would ultimately lead me to make significant contributions to the field of statistics. In hindsight, I am grateful for Professor Mao's guidance and understanding, as it ignited a passion for research that continues to shape my career in the world of statistics.

In late 1991 and early 1992, as I was deeply immersed in my Ph.D. journey at the University of Waterloo, a special and cherished connection blossomed in my academic life — a close bond with Professor Mao and Professor Jixiang Zhou, another distinguished professor from East China Normal University. Their visit to our university during that period marked a pivotal and memorable chapter in my academic and personal development. Professors Mao, Zhou, and I engaged in numerous discussions that covered a wide spectrum of subjects, ranging from the intricacies of statistical methodologies to the broader tapestry of life itself. These conversations were not confined to lecture halls or meeting rooms; we extended our discussions beyond the academic sphere. In fact, many evenings found us teaming up in the kitchen to prepare dinner together. These shared culinary experiences transcended mere meal preparation; they served as an extension of our intellectual and personal connection. As we chopped, stirred, and simmered, we continued our exchanges on topics that spanned from academic challenges to the joys and complexities of life.

This period of close interaction with Professor Mao left an in-erasable mark on my academic and personal journey. It was more than just a professor-student relationship; it was a mentorship filled with profound insights, camaraderie, and shared experiences. The wisdom I gained from these discussions, both academic and personal, has continued to shape my path in profound ways. As I reflect on those shared dinners and conversations, I am reminded of the lasting impact of those moments and the invaluable guidance that Professor Mao provided during my formative years in academia. His visit to the University of Waterloo was not just an academic event; it was a transformative experience that continues to inspire and influence my academic and personal pursuits to this day.

Professor Mao emphasized the paramount importance of mastering the art of data collection in the realm of scientific research. To advance our understanding and insights, we must dedicate our utmost efforts to this fundamental aspect. However, the reality of practical applications often brings forth an inescapable challenge: the specter of selection bias. In this paper, we embark on a comprehensive exploration of the concept of biased sampling. We delve into the nuances of this topic, dissecting its implications, and tracing its impact on the fields of modern statistics and machine learning. Our objective is to unravel the multifaceted nature of biased sampling, shed light on its real-world consequences, and investigate its relevance in the contemporary landscape of data-driven disciplines. Through this discourse, we aim to foster a deeper appreciation for the intricacies of data collection, the challenges it presents, and the innovative solutions that arise in the face of selection bias.

2 Introduction on Biased Sampling and Distribution Shift in Machine Learning

Biased sampling is a phenomenon that arises when an investigator collects samples from a population in a way that the resulting sampling distribution differs from the characteristics of the target population. This disparity in distribution occurs because, under the chosen sampling method, not all units within the population have an equal opportunity to be included in the sample. In other words, the natural sampling plan, while convenient and often intuitive, inadvertently introduces disparities in the representation of different segments of the population, thus leading to a biased sample. This occurrence can significantly impact the validity and generalizability of the research findings, underscoring the importance of understanding and mitigating bias in the sampling process. Biased sampling issues are indeed widespread, transcending various domains such as survey sampling, epidemiology studies, econometrics, and recently in machine learning literature. These challenges are not limited to one specific field but have the potential to affect the quality and validity of research in a wide array of disciplines. As pointed out by Professor James Heckman[1], the 2000 Nobel Laureate in Economics, "Sample selection bias may arise in practice for two reasons. First, there may be self selection by the individuals or data units being investigated. Second, sample selection decisions by analysts or data processors operate in much the same fashion as self selection".

In the ever-evolving landscape of machine learning, data is the lifeblood that fuels the algorithms driving everything from recommendation systems to image recognition and natural language processing. Data, however, is rarely perfect; it's often messy, incomplete, and, significantly, subject to a hidden peril — selection bias in sampling. Selection bias in machine learning refers to the systematic distortion of a dataset due to the non-random or biased selection of samples. This bias can emerge at various stages of data collection, from the initial acquisition to the curation and preprocessing phases. Its impact is far-reaching and can have profound consequences on the performance, fairness, and generalizability of machine learning models. As machine learning increasingly becomes an integral part of decision-making processes in industries ranging from healthcare and finance to criminal justice, it's imperative to understand and address the insidious effects of selection bias. Left unattended, selection bias can lead to dire consequences, from perpetuating unfair discrimination in algorithmic decisions to the creation of models that are not representative of the real-world conditions they're intended to address.

Through this review, we aim to provide a deeper understanding of the implications of selection bias and explore the techniques, methodologies, and tools available to detect, quantify, and mitigate its impact. We will discuss the realms of reweighting, resampling, and causality to uncover the innovative strategies that researchers are employing to address selection bias in machine learning datasets. Readers with an interest in delving deeper into biased sampling problems can turn to Qin[2] for a comprehensive exploration of this topic. This paper is organized as follows:

Section 3 focuses on one-sample and two-sample biased sampling problems and nonparametric maximum likelihood estimation. Section 4 discusses biased sampling issues in observational study where the baseline covariates are not balanced between treatment and control groups. In Section 5, we study how to effectively combine training data and test data together for an enhanced inference for the underlying conditional parameters in the presence of either covariate shift or prior probability shift. Section 6 explores the interconnections between general shift problems and biased sampling, shedding light on shared principles and techniques. Section 7 presents a specific example of how transfer learning can be conducted using shape-restricted generalized linear models. Section 8 discusses the latest developments in conformal predictive inference, a methodology focused on predicting future outcomes based on training data and provided future covariates. We conclude this paper with some discussions in Section 9.

3 One-Sample and Two-Sample Biased Sampling Problems

Let's start from the one-sample biased sampling problem. This scenario presents a challenge when direct data collection from the true distribution function, denoted as F is not possible. Instead, we work with a dataset $X_1, X_2, \cdots, X_n$ sampled from a distribution $G(x)$, which is related to F through a known non-negative function $w(x)$,

$$\mathrm{d}G(x) = \frac{w(x)\mathrm{d}F(x)}{\Delta}, \qquad \Delta = \int w(x)\mathrm{d}F(x).$$

The objective remains estimating the distribution function F. Easily we can use the nonparametric maximum likelihood estimate, i.e., the empirical distribution function

$$\widehat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x)$$

to estimate $G(x)$. Then we can solve

$$\mathrm{d}G_n(x) = \frac{w(x)\mathrm{d}F(x)}{\Delta}$$

for F. Easily we have

$$\mathrm{d}\widehat{F}(x) = w^{-1}(x)\mathrm{d}G_n(x)\Delta.$$

Noting

$$1 = \int \mathrm{d}\widehat{F}(x) = \int w^{-1}\mathrm{d}G_n(x)\Delta,$$

we have

$$\mathrm{d}\widehat{F}(x) = \frac{w^{-1}(x)\mathrm{d}G_n(x)}{\int w^{-1}(x)\mathrm{d}G_n(x)} = \frac{w^{-1}(x)\mathrm{d}G_n(x)}{n^{-1}\sum_{i=1}^{n} w^{-1}(x_i)},$$

or

$$\widehat{F}(x) = \frac{\sum_{i=1}^{n} w^{-1}(x_i) I(x_i \leqslant x)}{\sum_{i=1}^{n} w^{-1}(x_i)}.$$

The is called inverse weighting estimate or Hàjek estimator discussed in survey sampling literature, for example, [3].

As an alternative approach, we can write out the likelihood

$$L = \prod_{i=1}^{n} \frac{w(x_i)\mathrm{d}F(x_i)}{\int w(x)\mathrm{d}F(x)}.$$

When maximizing it with respect to F, we only need to consider those F with jumps at all observed data points; otherwise, the likelihood is 0. Denote $\mathrm{d}F(x_i) = p_i$, $i = 1, 2, \cdots, n$ as the jumps. As a consequence the integral $\int w(x)\mathrm{d}F(x)$ becomes the summation $\sum_{j=1}^n p_j w(x_j)$. Next we need to maximize

$$L = \prod_{i=1}^n \frac{w(x_i)p_i}{\sum_{j=1}^n p_j w(x_j)}$$

subject to constraints

$$p_i \geqslant 0, \quad i = 1, 2, \cdots, n; \qquad \sum_{j=1}^n p_j = 1.$$

Easily we can show that the above likelihood is maximized when

$$p_i = \frac{w(x_i)}{\sum_{j=1}^n w(x_j)}, \qquad i = 1, 2, \cdots, n.$$

Therefore the inverse weighting estimate $\widehat{F}(x)$ is the nonparametric maximum likelihood estimate (MLE).

Let's now turn our attention to the realm of two-sample biased sampling problems. In this scenario, we find ourselves confronted with two distinct groups of data. The first group comprises direct observations stemming from the distribution F, while the second group consists of observations that are inherently biased when drawn from the same distribution F. This intriguing two-sample problem presents us with a unique opportunity to explore the disparities between these two sets of observations and to discern the underlying factors contributing to this bias. By uncovering into the intricacies of these biased observations, we can gain valuable insights into the challenges and complexities that arise when dealing with such data.

Suppose we have observations

$$X_1, X_2, \cdots, X_{n_0} \sim \mathrm{d}F(x),$$

and

$$X_{n_0+1}, X_{n_0+2}, \cdots, X_n \sim \frac{w(x)\mathrm{d}F(x)}{\int w(x)\mathrm{d}F(x)}.$$

Denote $p_i = \mathrm{d}F(x_i)$, $i = 1, 2, \cdots, n$ as the jump size of F at each of the observed data points. Vardi [4,5] maximized

$$\prod_{i=1}^{n_0} p_i \prod_{i=n_0+1}^{n} \frac{w(x_i)p_i}{\sum_{i=1}^n p_i w(x_i)}$$

with respect to p_i, $i = 1, 2, \cdots, n$ subject to the constraints

$$p_i \geqslant 0, \qquad \sum_{i=1}^{n} p_i = 1.$$

The resultant distribution function estimate is

$$\widehat{F}(x_i) = \sum_{i=1}^{n} \widehat{p}_i I(x_i \leqslant x).$$

Here, we'll avoid delving into the intricate mathematical details, as they can be regarded as a specialized case discussed below, especially when the weight function $w(x) = w(x, \theta)$ might rely on an unknown parameter.

Let's consider a scenario with a disease indicator denoted as D where 1 represents an individual having the disease of interest, and 0 signifies the absence of the disease. Additionally, we have X, which represents a vector of p covariates. The standard logistic regression model is often employed and can be expressed as follows:

$$\mathsf{P}(D = 1 \mid x) = \frac{\exp(\alpha^* + x^{\mathsf{T}}\beta)}{1 + \exp(\alpha^* + x^{\mathsf{T}}\beta)} \equiv \pi(x), \qquad X \sim f(x). \tag{1}$$

Here, α^* is an intercept parameter, β is a $p \times 1$ vector parameter and X^{T} is the transpose of X. The marginal density $f(x)$ remains unspecified.

In the field of epidemiological studies, particularly in the context of cancer research, retrospective sampling, also known as case-control sampling, stands as one of the most prevalent methods. This approach is favored for its convenience, cost-effectiveness, and efficiency. This becomes especially pertinent when investigating rare diseases, where obtaining a substantial number of cases through prospective sampling may not be practical.

In case-control sampling, a predetermined number of cases n_1 and controls n_0 are gathered retrospectively from separate case and control populations. Typically, this is achieved by selecting cases from hospitals and controls from the general disease-free population. However, it's important to note that the disease prevalence $\pi = \mathsf{P}(D = 1)$ in the overall population may differ from the disease proportion $n_1/(n_0 + n_1)$ within the collected samples. This distinction holds relevance, as it can influence the analysis and interpretation of the study's results.

Let $X_1, X_2, \cdots, X_{n_0}$ be a random sample from $F(x \mid D = 0)$ and, independently, let $X_{n_0+1}, X_{n_0+2}, \cdots, X_n$ be a random sample from $F(x \mid D = 1)$, where $F(x \mid D = 0)$ and $F(z \mid D = 1)$ are, respectively, the covariate distribution functions for controls and cases. The corresponding density functions of the covariates are denoted as, respectively, $f(x \mid D = i) = \mathrm{d}F(x \mid D = i)/\mathrm{d}x$, $i = 0, 1$. The Bayes' rule gives

$$f(x \mid D = 1) = \frac{\pi(x)}{\pi} f(x), \qquad f(x \mid D = 0) = \frac{1 - \pi(x)}{1 - \pi} f(x).$$

It is seen that

$$\frac{f(x\,|\,D=1)}{f(x\,|\,D=0)}=\frac{1-\pi}{\pi}\frac{\pi(x)}{1-\pi(x)}.$$

Let $g(x)=f(x\,|\,D=0)$ and $h(x)=f(x\,|\,D=1)$. The corresponding cumulative distribution functions are denoted as $G(x)$ and $H(x)$, respectively. Then

$$h(x)=f(x\,|\,D=1)=\frac{1-\pi}{\pi}\frac{\pi(x)}{1-\pi(x)}g(x)=\exp(\alpha+x^{\mathsf{T}}\beta)g(x),$$

where $\alpha=\alpha^{*}+\ln\{(1-\pi)/\pi\}$. As a result, we arrive at the following two-sample exponential tilting model or density ratio model in which $(X_1,X_2,\cdots,X_{n_0})$ and $(X_{n_0+1},X_{n_0+2},\cdots,X_n)$ are independent and

$$\begin{aligned}&X_1,X_2,\cdots,X_{n_0}\text{ are independent with density }g(x),\\&X_{n_0+1},X_{n_0+2},\cdots,X_n\text{ are independent with density }h(x)=\exp(\alpha+x^{\mathsf{T}}\beta)g(x).\end{aligned}\tag{2}$$

This scenario can be described as a biased sampling model with a weight function denoted as $\exp(\alpha+x^{\mathsf{T}}\beta)$, and this weight function depends on the unknown parameters α and β. It's important to note that the specific form of $g(x)$ remains unspecified.

In this context, performing statistical inferences based on the exponential tilting model proves to be more robust than relying on a full parametric model where the form of $g(x)$ is assumed to be known. The flexibility of the exponential tilting model makes it a valuable tool in scenarios where the exact functional form of the weight function is uncertain. It's worth mentioning that the exponential tilting model encompasses a wide range of common probability distributions, including exponential distributions with varying rates and normal distributions with a common variance but different means. This versatility makes it a valuable framework for handling various practical scenarios in statistical modeling.

Next we discuss maximum semiparametric likelihood estimation. Using the exponential tilting model, we can write the likelihood as

$$\mathscr{L}(\alpha,\beta,G)=\prod_{i=1}^{n_0}\mathrm{d}G(x_i)\prod_{j=n_0+1}^{n}w(x_j)\mathrm{d}G(x_j)=\Big(\prod_{i=1}^{n}p_i\Big)\Big[\prod_{j=n_0+1}^{n}w(x_j)\Big],\tag{3}$$

where $w(x)=\exp(\alpha+x^{\mathsf{T}}\beta)$ and $p_i=\mathrm{d}G(x_i)$, $i=1,2,\cdots,n$, are (nonnegative) jumps with total unit mass.

The first step is, for fixed (α,β), to maximize $\mathscr{L}$ with respect to p_i, $i=1,2,\cdots,n$, subject to constraints $\sum p_i=1$, $p_i\geqslant 0$, $\sum p_i[w(x_i)-1]=0$, where the last constraint reflects the fact that $\int w(x)\mathrm{d}G(x)=1$. The maximum value of $\mathscr{L}$ is attained at $p_i=n_0^{-1}[1+\rho\exp(\alpha+x_i^{\mathsf{T}}\beta)]^{-1}$ by using a Lagrange multiplier method, where $\rho=n_1/n_0$.

Therefore, ignoring constants, the log-likelihood function is

$$l(\alpha,\beta)=\sum_{j=n_0+1}^{n}(\alpha+x_j^{\mathsf{T}}\beta)-\sum_{i=1}^{n}\ln\{1+\rho\exp(\alpha+x_i^{\mathsf{T}}\beta)\}. \tag{4}$$

Next we maximize l over (α,β). Let $(\widetilde{\alpha},\widetilde{\beta})$ satisfy the following system of score equations:

$$\begin{aligned}\frac{\partial l(\alpha,\beta)}{\partial\alpha}&=n_1-\sum_{i=1}^{n}\frac{\rho\exp(\alpha+t_i\beta)}{1+\rho\exp(\alpha+x_i^{\mathsf{T}}\beta)}=0,\\ \frac{\partial l(\alpha,\beta)}{\partial\beta}&=\sum_{j=n_0+1}^{n}x_j-\sum_{i=1}^{n}\frac{x_i\rho\exp(\alpha+x_i^{\mathsf{T}}\beta)}{1+\rho\exp(\alpha+x_i^{\mathsf{T}}\beta)}=0.\end{aligned}$$

Then we have

$$\widetilde{p}_i=\{n_0[1+\rho\exp(\widetilde{\alpha}+x_i^{\mathsf{T}}\widetilde{\beta})]\}^{-1}. \tag{5}$$

It's important to note that the score equations discussed above are identical to those introduced by Prentice and Pyke[6]. The application of this maximum semiparametric likelihood approach has allowed us to reaffirm a well-established result. Specifically, it highlights that the prospective likelihood can be effectively employed for making inferences regarding the log odds ratio parameter β, even in situations where the sampling design is retrospective. This underscores the robustness and adaptability of the statistical methods employed in this context, bridging the gap between different study designs and data collection methods.

On the basis of the $\widetilde{p}_i$'s, naturally we can estimate $G(t)$ by[7]

$$\widetilde{G}(x)=\sum_{i=1}^{n}\widetilde{p}_iI(x_i\leqslant x)=\frac{1}{n_0}\sum_{i=1}^{n}\frac{I(x_i\leqslant x)}{1+\rho\exp(\widetilde{\alpha}+x_i^{\mathsf{T}}\widetilde{\beta})}, \tag{6}$$

where $X_i\leqslant x$ implies the inequality applied component-wise. When we consider $\widehat{G}(x)$ as the empirical distribution function derived from the control data only, we can reasonably anticipate that $\widetilde{G}(x)$ would demonstrate greater efficiency than $\widehat{G}(x)$. This increased efficiency arises from the fact that the former incorporates both case and control data, leveraging a larger dataset for a more comprehensive and robust estimation of the distribution function.

4 Biased Sampling Problem based on Observational Data

Clinical trials and observational studies represent two common research approaches in medical and scientific investigations, each characterized by distinct design, objectives, and control levels over study conditions. In a clinical trial, patients are randomly allocated to either the treatment or control group, ensuring the creation of a balanced baseline in terms of covariates between the two groups. Conversely, in observational studies, the assignment to treatment and control groups lacks randomness and may depend on baseline information or other factors, often resulting in an imbalance in covariate information. This situation is reminiscent of "covariate shift" in machine learning terminology. Within the context of covariate shift, it is imperative to acknowledge and address variations in covariate distributions between the groups to ensure valid inferences and predictions, especially when estimating treatment effects or outcomes. Various statistical techniques and machine learning methods are employed to account for covariate shift and its potential impact on study outcomes in observational research.

A pivotal element in clinical trials is the imperative need for balanced baseline covariates between treatment and control groups. When the outcome endpoint is free from censoring, straightforward approaches, such as the two-sample t-test or the Wilcoxon test, are utilized to evaluate treatment effects. In cases involving right censoring of the outcome endpoint, on the other hand, the log-rank test serves as an appropriate statistical tool. In observational study, however, these simple statistical methods can yield biased results, and we will investigate the specifics of this issue in the subsequent discussion.

Consider the treatment or control indicator, denoted as D, where $D = 1$ signifies treatment, and $D = 0$ represents control. In the realm of causal inference, the fundamental assumption revolves around the propensity score[8], denoted as $\pi(x)$, and expressed as:

$$\mathsf{P}(D = 1 \,|\, x) = \pi(\alpha + x\beta) = \frac{\exp(\alpha + x^{\mathsf{T}}\beta)}{1 + \exp(\alpha + x^{\mathsf{T}}\beta)}.$$

Now, let Y_1 and Y_0 denote the potential outcomes for the treatment and control groups, respectively. It's important to note that only one of these outcomes, either Y_1 (if $D = 1$) or Y_0 (if $D = 0$), is observed for each individual. We define the observed outcome, Y, as:

$$Y = DY_1 + (1 - D)Y_0.$$

At the heart of causal inference lies a fundamental assumption: the absence of unmeasurable confounding. In other words, we express this assumption as follows:

$$(Y_1, Y_0) \perp D \,|\, X.$$

This expression signifies that, considering the covariate set X, the outcomes Y_0 and Y_1 are statistically independent of the treatment assignment D. This assumption is a critical pillar in the field of causal inference, underpinning our efforts to establish causal relationships and draw meaningful conclusions from observational or experimental data.

Biased sampling arises when examining the conditional density of (Y, X) given $D = 1$ or $D = 0$ as follows:

$$(Y, X) \mid D = 1 \sim (Y_1, X) \mid D = 1 \sim \frac{\mathsf{P}(D = 1 \mid x)\mathsf{P}(X = x, Y_1 = y)}{\mathsf{P}(D = 1)} = \frac{\pi(x) f_1(x, y)}{\int \pi(x) \mathrm{d}F_1(x, y)}$$

and

$$(Y, X) \mid D = 0 \sim (Y_0, X) \mid D = 0 \sim \frac{\mathsf{P}(D = 0 \mid x)\mathsf{P}(X = x, Y_0 = y)}{\mathsf{P}(D = 0)} = \frac{[1-\pi(x)] f_0(x, y)}{\int [1-\pi(x)] \mathrm{d}F_0(x, y)},$$

where $f_i(x, y)$, $i = 0, 1$ is the joint density of (Y_i, X), $i = 0, 1$ for control and treatment, respectively. If we decompose the joint density (Y_1, X) backwardly,

$$f_1(x, y) = f_1(x \mid y) f_1(y) \qquad \text{and} \qquad f_0(x, y) = f_0(x \mid y) f_0(y),$$

we can write

$$\int \pi(x) f_1(x \mid y) \mathrm{d}x =: w_1(y), \qquad \int [1 - \pi(x)] f_0(x \mid y) \mathrm{d}x =: w_0(y).$$

We can easily observe that

$$Y_1 \mid D = 1 \sim \frac{w_1(y) f_1(y)}{\int w_1(y) f_1(y) \mathrm{d}y} \qquad \text{and} \qquad Y_0 \mid D = 0 \sim \frac{w_0(y) f_0(y)}{\int w_0(y) f_0(y) \mathrm{d}y}.$$

It's evident that the observed treatment and control observations represent biased versions of the marginal densities of $f_1(y)$ and $f_0(y)$. If we were to directly compare Y_1 and Y_0, the results would inevitably be biased. However, in a special scenario where $\pi(x)$ remains a constant, we can achieve unbiased results.

It is not easy to identify the selection biases $w_1(y)$ and $w_0(y)$, as an alternative approach one may examine the conditional likelihood

$$Y_1, X \mid D = 1 \sim \frac{\pi(x) \mathrm{d}F_1(y, x)}{\int \pi(x) \mathrm{d}F_1(y, x)}, \qquad Y_0, X \mid D = 0 \sim \frac{[1 - \pi(x)] \mathrm{d}F_0(y, x)}{\int [1 - \pi(x)] \mathrm{d}F_0(y, x)}.$$

Let (X_i, Y_i, D_i), $i = 1, 2, \cdots, n$ be the observed data. Easily we can use the logistic log likelihood

$$\ell = \sum_{i=1}^{n} D_i(\alpha + x_i \beta) - \sum_{i=1}^{n} \ln\{1 + \exp(\alpha + x_i^{\mathsf{T}} \beta)\}$$

to estimate (α, β), denoted the maximum likelihood estimate as $(\widehat{\alpha}, \widehat{\beta})$. Without loss of generality, we assume $D_1 = D_2 = \cdots = D_{n_1} = 1$, and $D_{n_1+1} = 0, D_{n_1+2} = 0, \cdots, D_n = 0$. Then easily we can estimate $F_1(x, y)$ and $F_0(x, y)$ by using the inverse weighting estimators

$$\widehat{F}_1(x, y) = \frac{\sum_{i=1}^{n_1} I(x_i \leqslant x, y_i \leqslant y)/\pi(\widehat{\alpha} + x_i^{\mathsf{T}} \widehat{\beta})}{\sum_{i=1}^{n_1} 1/\pi(\widehat{\alpha} + x_i^{\mathsf{T}} \widehat{\beta})},$$

$$\widehat{F}_0(x,y)=\frac{\sum_{i=n_1+1}^{n} I(x_i\leqslant x, y_i\leqslant y)/[1-\pi(\widehat{\alpha}+x_i^{\mathsf{T}}\widehat{\beta})]}{\sum_{i=n_1+1}^{n} 1/[1-\pi(\widehat{\alpha}+x_i^{\mathsf{T}}\widehat{\beta})]}.$$

Equivalently, they can also be written as

$$\begin{aligned}\widehat{F}_1(x,y)&=\frac{n^{-1}\sum_{i=1}^{n}[D_i/\pi(x_i)]I(X_i\leqslant x, Y_i\leqslant y)}{n^{-1}\sum_{i=1}^{n} D_i/\pi(\widehat{\alpha}+x_i^{\mathsf{T}}\widehat{\beta})},\\ \widehat{F}_0(y,x)&=\frac{n^{-1}\sum_{i=1}^{n}\{(1-D_i)/[1-\pi(x_i)]\}I(X_i\leqslant x, Y_i\leqslant y)}{n^{-1}\sum_{i=1}^{n}(1-D_i)/[1-\pi(\widehat{\alpha}+x_i^{\mathsf{T}}\widehat{\beta})]}.\end{aligned}$$

This estimate, however, did not use the information the $F_1(x,\infty)=F_0(x,\infty)$, i.e., the marginal distribution of X is the same. As an improvement, one may consider

$$\widehat{F}_1(x,y)+\lambda_1[\widehat{F}_1(x,\infty)-\widehat{F}_0(x,\infty)] \qquad \text{and} \qquad \widehat{F}_0(x,y)+\lambda_0[\widehat{F}_1(x,\infty)-\widehat{F}_0(x,\infty)].$$

Clearly both of them asymptotically unbiased for any constant choice of λ_1 and λ_0. To find the best one, we may minimize the asymptotic variances with respect to λ_1 and λ_0, separately.

As an alternative approach, we let $p_i=\mathrm{d}F_1(y_i,x_i)$, $i=1,2,\cdots,n_1$ be the jumps of $F_1(x,y)$. We can maximize the conditional likelihood defined as:

$$L_1=\prod_{i=1}^{n_1}\frac{\pi(\widehat{\alpha}+x_i^{\mathsf{T}}\widehat{\beta})p_i}{\sum_{j=1}^{n_1}p_j\pi(\widehat{\alpha}+x_j^{\mathsf{T}}\widehat{\beta})}.$$

This maximization is done with respect to the variables p_i, while subject to the following constraints:

$$p_i\geqslant 0, \qquad \sum_{i=1}^{n_1}p_i=1, \qquad \sum_{i=1}^{n_1}\psi(x_i)p_i=\frac{1}{n}\sum_{i=1}^{n}\psi(x_i).$$

Here, $\psi(x)$ can take any form as a function of X, although its optimal selection depends on the estimation process. The first two constraints ensure that $F_1(x,y)$ behaves as a proper distribution function. The last constraint is designed to align the moment constraint with the population moment constraints, providing calibration to the overall statistical framework. We can estimate the marginal distribution $F_1(y)$ by

$$\widehat{F}_{C1}(y)=\sum_{i=1}^{n_1}\widehat{p}_i I(Y_i\leqslant y).$$

More details can be found in [9].

Another popular approach is the so called double robust estimator [10] defined as

$$\widehat{F}_1(y)=n^{-1}\sum_{i=1}^{n}\frac{D_i}{\pi(x_i)}I(Y_{1i}\leqslant y)-\frac{D_i-\pi(x_i)}{\pi(x_i)}F_W(y\,|\,x_i),$$

where $F_W(y\,|\,x)$ is the user specified "working conditional distribution of Y_1 given x, and Y_{1i}, $i = 1, 2, \cdots, n$ be the treatment outcomes, even though $Y_{1(n_1+1)}, Y_{1(n_1+2)}, \cdots, Y_{1n}$ are not available. It is a consistent estimator of $F_1(y)$ if either $F_W(y\,|\,x)$ is correctly specified or the propensity score is correctly specified. We can observe this easily by re-writing

$$\begin{aligned}\widehat{F}_1(y) &= n^{-1}\sum_{i=1}^{n}\frac{D_i-\pi_0(x_i)}{\pi(x_i)}[I(Y_{1i}\leqslant y)-F_1(y\,|\,x_i)]\\ &\quad + n^{-1}\frac{D_i-\pi_0(x_i)}{\pi(x_i)}[F_1(y|x_i)-F_W(y\,|\,x_i)]\\ &\quad + n^{-1}\sum_{i=1}^{n}\Big\{\Big[\frac{\pi_0(x_i)}{\pi(x_i)}-1\Big][I(Y_{1i}\leqslant y)-F_W(y\,|\,x_i)]\Big\} + n^{-1}\sum_{i=1}^{n}I(Y_{1i}\leqslant y).\end{aligned}$$

It's important to note that the first three terms in the equation tend to converge to zero when either $\pi(x) = \pi_0(x)$ or $F_W(y\,|\,x) = F_1(y\,|\,x)$. This suggests that these terms become negligible under certain conditions. On the other hand, the last term, which is expressed as $n^{-1}\sum_{i=1}^{n} I(Y_{1i} \leqslant y)$, serves as an unbiased estimator of $F_1(y)$ even if it's not available for all instances of Y_{1i}, where i ranges from 1 to n.

In practical applications, it is common to employ a parametric model to estimate $I(Y_{1i} \leqslant y)$ in relation to $F_W(y\,|\,x_i)$, where i spans from 1 to n_1. This approach helps in making estimates and drawing inferences in situations where full data might not be available.

5 Data Shift Problem in Machine Learning

The "data shift problem" in machine learning refers to a situation where the statistical properties of the training data differ significantly from those of the data the model will encounter during testing or deployment. This shift can manifest in various ways, such as differences in data distribution, feature space, or class balance. Data shift can adversely affect the performance and generalizability of machine learning models. Data shift can occur for several reasons, including changes in the data collection process, variations in the data source, or over time due to evolving conditions. It can lead to model performance degradation and affect the reliability of machine learning systems. Moreno-Torres et al.[11] introduced a unifying framework by conducting a comprehensive review and comparison of several pivotal distributional shift problems found in the literature. These encompass:

Covariate shift: This occurs when the distribution of feature variables in the training data differs from that in the test data. Models trained on covariate-shifted data may struggle to generalize to the testing data because they have not seen these variations during training.

Concept shift: Concept shift is related to changes in the underlying data-generating process. This can be caused by factors such as shifts in user behavior, external events, or changes in the environment. Models may fail to adapt to these shifts and may provide inaccurate predictions.

Prior probability shift or label shift (when outcome is a binary variable): Label shift is when there is a change in the distribution of the target variable (labels). In the presence of label shift, model predictions can be biased, as the model may rely on outdated label distributions from the training data.

Next, we will study the specifics of various shift problems. We will work under the assumption that we have access to two distinct datasets: one referred to as the "training data", and the other as the "test data". Let's explore the concepts of training data and test data in the context of a statistical framework. This framework is based on the joint density $p_0(x,y)$ for training data and $p_1(x,y)$ for test data.

Training data: The joint density $p_0(x,y)$ can be decomposed as

$$(X_{0i}, Y_{0i}), i = 1, 2, \cdots, n_0 \sim p_0(y \,|\, x, \theta) p_0(x) = p_0(x \,|\, y) p_0(y).$$

Test data: Similarly the joint density $p_0(x,y)$ can be written as

$$(X_{1i}, Y_{1i}), i = 1, 2, \cdots, n_1 \sim p_1(y \,|\, x, \theta) p_1(x) = p_1(x \,|\, y) p_1(y),$$

where $p_i(y \,|\, x)$, $i = 0, 1$ is the conditional density of Y given X, and $p_i(x)$, $i = 0, 1$ is the marginal density of X. Similar notation applies to $p_i(x \,|\, y) p_i(y)$, $i = 0, 1$. The covariate-shift assumption is

$$p_0(y \,|\, x) = p_1(y|x), \qquad p_0(x) \neq p_1(x).$$

If one assumes a conditional parametric model

$$p_0(y \,|\, x) = p_1(y \,|\, x) = p(y \,|\, x, \theta),$$

then one may simply maximize the joint log likelihood

$$\ell = \sum_{i=1}^{n_0} \ln p(y_{0i} \,|\, x_{0i} \theta) + \sum_{i=1}^{n_1} \ln p(y_{1i} \,|\, x_{1i}, \theta)$$

with respect to θ. In other words, it is straightforward to make statistical inference.

To quantify the shift function, the most popular assumption is

$$p_1(x) = r(x) p_0(x),$$

where $r(x)$ is the density ratio, typically takes the form

$$r(x) = \exp(\alpha + x^{\mathsf{T}}\beta).$$

In order to test whether there exists a covariate shift, as discussed in Section 3 we can use logistic regression log-likelihood ratio statistic

$$R(0) = 2[\max_{\beta} \ell(\beta) - \ell(0)],$$

where

$$\begin{aligned}\ell(\beta) = & \sum_{i=1}^{n_1} (\alpha + x_{1i}\beta) - \sum_{i=1}^{n_0} \ln\{1 + \rho \exp(\alpha + x_{0i}^{\mathsf{T}}\beta)\} \\ & - \sum_{i=1}^{n_1} \ln\{1 + \rho \exp(\alpha + x_{1i}^{\mathsf{T}}\beta)\}, \qquad \rho = n_1/n_0,\end{aligned}$$

is the retrospective logistic log likelihood given in Equation (4). In the absence of any shift under the null hypothesis, the function $R(0)$ approaches the standard chi-square distribution, and the degrees of freedom for this distribution are determined by the dimension of the parameter β.

Unlike covariate shift, the commonly used label shift or prior probability shift assumption in machine learning involves decomposing the joint density in a different manner. Specifically, it is expressed as:

$$p_0(x \,|\, y) = p_1(x \,|\, y), \qquad p_0(y) \neq p_1(y).$$

In simpler terms, when conditioning on the outcome variable Y, the conditional density of the covariates remains consistent between the training and test data. However, there may be differences in the marginal densities.

When we consider the relationship between the disease status (Y) and the symptoms (X), our primary interest lies in predicting the disease status based on the presented symptoms. In the context of machine learning literature, it's common to make an anticausal assumption that the disease status (Y) causes the symptoms (X). Under the label shift assumption, the conditional probability distribution $\mathsf{P}(X \,|\, Y)$ remains constant across different studies or scenarios. However, the marginal distribution of the disease status $\mathsf{P}(Y)$ can vary among these different studies. For instance, the symptoms associated with a specific cancer stage will typically remain consistent. However, the prevalence of that cancer may differ across various countries or regions.

If the binary label Y is available in the test data, we can easily determine whether there is a label shift by comparing $\mathsf{P}_0(Y = 0)$ in the training data and $\mathsf{P}_1(Y = 1)$ in the test data. However, it becomes challenging when only covariate information X is available in the test data. In the absence of knowledge of Y, the covariate density X in the test data is a mixture:

$$q(x) = \mathsf{P}_1(Y = 1)p_1(x \,|\, Y = 1) + \mathsf{P}_1(Y = 0)p_1(x \,|\, Y = 0).$$

Under the assumption of exponential tilting shift between $Y = 0$ and $Y = 1$ in the training group, we have a semiparametric model:

$$p_0(x \,|\, Y = 1) = p_0(x \,|\, Y = 0)\exp(\alpha + x^{\mathsf{T}}\beta),$$

where the baseline density $p_0(x \,|\, Y = 0)$ is not specified. Due to the label shift assumption that the conditional density of X conditioning on Y remains the same in training and test data, we have:

$$p_1(x \,|\, Y = i) = p_0(x \,|\, Y = i), \qquad i = 0, 1.$$

As a consequence, in the test data where only Xs are available, we have:

$$X_i \sim q(x) = [\lambda\exp(\alpha + x^{\mathsf{T}}\beta) + (1 - \lambda)]p_0(x \,|\, Y = 0), \qquad i = n_0 + 1, n_0 + 2, \cdots, n.$$

Without loss of generality, we assume the first m_0 observations in training data with $Y_i = 0$, and denote $p_0(x \,|\, Y = 0) = p_0(x)$, we end up to three groups of data

$$X_i \,|\, Y_i = 0 \sim p_0(x), \qquad i = 1, 2, \cdots, m_0,$$
$$X_i \,|\, Y_i = 1 \sim p_0(x)\exp(\alpha + x^{\mathsf{T}}\beta), \qquad i = m_0 + 1, m_0 + 2, \cdots, n_0,$$

and

$$X_i \sim p_1(x) = [\lambda\exp(\alpha + x^{\mathsf{T}}\beta) + (1 - \lambda)]p_0(x), \qquad i = n_0 + 1, n_0 + 2, \cdots, n.$$

This leads to a three-sample biased sampling problem with weights $w_1(x) = 1$, $w_2(x) = \exp(\alpha + x^{\mathsf{T}}\beta)$ and $w_3(x) = \lambda\exp(\alpha + x^{\mathsf{T}}\beta) + (1 - \lambda)$. Similar to the discussion in Section 3 on the two-sample biased sampling problem, it is possible to estimate $\mathsf{P}_1(Y = 1)$ by profiling out $p_0(x)$. More details can be found in [12] and [13].

If Y is continuous and available in the test data, then we can directly compare observed y_i's in the training data and in the test data, say by two sample t-test or Wilcoxon test to determine whether there is a prior probability shift.

If Y is not available, then it becomes challenging! The observed covariate X in the test data have density

$$\int p_1(x \,|\, y)p_1(y)\mathrm{d}y.$$

We need to test $p_0(y) = p_1(y)$ but no direct observations from $p_1(y)$! If we assume a parametric model for $p_1(x \mid y) = p_0(x \mid y) = p(x \mid y, \theta)$, this becomes a convolution problem as discussed in [14].

If indeed there is label shift, next we discuss how to combine training data and test data together to get an enhanced inference for the underlying parameter in $p_0(y \mid x^{\mathsf{T}}\beta)$. To achieve this, we adopt an approach that combines prospective likelihood based on the training data:

$$L_0 = \prod_{i=1}^{n_0} p_0(y_{0i} \mid x_{0i}^{\mathsf{T}}\beta) p_0(x_{0i}),$$

and retrospective likelihood based on the test data:

$$L_1 = \prod_{i=1}^{n_1} p_1(x_{1i} \mid y_{1i}).$$

One key advantage of this approach is that it avoids the need to explicitly model $p_1(y)/p_0(y)$. Additionally, it's worth noting that:

$$p_1(x \mid y) = p_0(x \mid y) = \frac{p_0(y \mid x^{\mathsf{T}}\beta) p_0(x)}{\int p_0(y \mid x^{\mathsf{T}}\beta) p_0(x) \mathrm{d}x}$$

by using the prior probability shift assumption. This equation illustrates the relationship between $p_1(x \mid y)$ and $p_1(y \mid x)$, showing their connection through the marginal density $p_0(x)$ of X.

In a manner akin to Vardi's approach from 1985, we can demonstrate that the nonparametric maximum likelihood estimate of $\mathrm{d}\mathsf{P}_0(x)$ exhibits discontinuities only at the observed data points.

The overall likelihood is

$$L = \prod_{i=1}^{n_0} \mathrm{d}\mathsf{P}_0(x_i) p_0(y_i \mid x_i^{\mathsf{T}}\beta) \prod_{j=n_0+1}^{n} \frac{p_0(y_j \mid x_j^{\mathsf{T}}\beta) \mathrm{d}\mathsf{P}_0(x_j)}{\sum_{i=1}^{n} p_0(y_j \mid x_i^{\mathsf{T}}\beta) \mathrm{d}\mathsf{P}_0(x_i)}.$$

For convenience we let $p_i = \mathrm{d}\mathsf{P}_0(x_i) \geqslant 0$, $i = 1, 2, \cdots, n$.

1) Clearly $\mathrm{d}\mathsf{P}_0(x)$ must jump at each of the observe data points $\{x_1, x_2, \cdots, x_n\}$ since otherwise the likelihood is 0.

2) If there exists an additional jump point at x_0 for $\mathrm{d}\mathsf{P}_0(x)$, where x_0 is not in the set $\{x_1, x_2, \cdots, x_n\}$, the jump size is denoted as $p_0 = \mathrm{d}\mathsf{P}_0(x_0)$. The likelihood is

$$L(p_0, p_1, \cdots, p_n) = \prod_{i=1}^{n_0} p_i p_0(y_i \mid x_i^{\mathsf{T}}\beta) \prod_{j=n_0+1}^{n} \frac{p_0(y_j \mid x_j^{\mathsf{T}}\beta) p_j}{\sum_{i=1}^{n} p_0(y_j \mid x_i^{\mathsf{T}}\beta) p_i + p_0(y_j \mid x_0^{\mathsf{T}}\beta) p_0},$$

where

$$p_0 + \sum_{i=1}^{n} p_i = 1, \qquad p_i \geqslant 0, \ \ i = 0, 1, 2, \cdots, n.$$

We can define a new probability distribution that has jumps only at observed data points, represented by masses $q_i = p_i / \sum_{j=1}^n p_i$, where $i = 1, 2, \cdots, n$. The likelihood becomes

$$L(q_1, q_2, \cdots, q_n) = \Big[\prod_{i=1}^{n_0} q_i p_0(y_i \,|\, x_i\beta)\Big]\Big[\prod_{j=n_0+1}^{n} \frac{p_0(y_j \,|\, x_j^{\mathsf{T}}\beta)q_j}{\sum_{i=1}^n p(y_j \,|\, x_i^{\mathsf{T}}\beta)q_i}\Big].$$

Note $\sum_{i=1}^n p_i \leqslant 1$, therefore

$$q_i = p_i / \sum_{j=1}^n p_j \geqslant p_i, \quad i = 1, 2, \cdots, n, \qquad \sum_{i=1}^n q_i = 1, \quad q_i \geqslant 0.$$

Moreover,

$$\begin{aligned}\prod_{j=n_0+1}^{n} \frac{p_0(y_j \,|\, x_j^{\mathsf{T}}\beta)q_j}{\sum_{i=1}^n p_0(y_j \,|\, x_i^{\mathsf{T}}\beta)q_i} &= \prod_{j=n_0+1}^{n} \frac{p_0(y_j \,|\, x_j^{\mathsf{T}}\beta)p_j}{\sum_{i=1}^n p_0(y_j \,|\, x_i^{\mathsf{T}}\beta)p_i} \\ &\geqslant \prod_{j=n_0+1}^{n} \frac{p_0(y_j \,|\, x_j^{\mathsf{T}}\beta)p_j}{\sum_{i=1}^n p_0(y_j \,|\, x_i^{\mathsf{T}}\beta)p_i + p_0(y_j \,|\, x_0)p_0} \\ &= L(p_0, p_1, \cdots, p_n).\end{aligned}$$

As a consequence we have shown that the extra mass $p_0 = \mathsf{d}\mathsf{P}_0(x_0)$ at x_0 would not increase the likelihood.

In general, the values of p_i, $i = 1, 2, \cdots, n$ cannot be expressed as a simple function of the available data and a finite set of parameters. The process of determining these p_i values often requires iterative solutions, and it becomes more intricate when considering large sample sizes.

In the realm of machine learning literature, “concept drift”, often simply referred to as “drift”, is a prevalent form of distributional shift problem. It represents a transformation in data that renders the existing data model invalid. Concept drift occurs when the statistical characteristics of the target variable that the model aims to predict change over time in unexpected and unpredictable ways. This creates challenges because as time progresses, the model’s predictions become less accurate. In statistical terms, the concept change in machine learning can be expressed as:

$$p_0(y \,|\, x) \neq p_1(y \,|\, x).$$

In other words, the conditional probability densities differ between the training data and the test data.

Let $D = 0, 1$ be training data or test data indicator, respectively. To test the null hypothesis $H_0 : p_0(y \,|\, x) = p_1(y \,|\, x)$, it is equivalent to testing

$$Y \perp D \,|\, X.$$

This can be further expressed as:

$$p(Y \mid X, D) = p(Y \mid X).$$

Hu and Lei[15] have leveraged the concept of conformal inference (to be discussed in Section 8) to examine this hypothesis in the presence of covariate shift. In the literature, various other test statistics are available, such as those proposed by Thams et al.[16].

Under the null hypothesis $H_0 : p_0(y \mid x) = p_1(y \mid x)$, $p_1(x) = p_0(x)w(x)$ for any function of (X, Y), denoted as $\psi(X, Y)$, we can express the expectation as:

$$\mathsf{E}_1[\psi(Y, X)] = \mathsf{E}_0[\psi(Y, X)w(X)],$$

where E_1 and E_0 denote expectations with respect to the joint distributions in test and training data, respectively. One approach to construct a test is the Wilcoxon rank sum test:

$$\sum_{j=n_0+1}^{n} \sum_{i=1}^{n_0} [I(\psi(X_i, Y_i) < \psi(Y_j, X_j))w(X_i) - 0.5].$$

It's worth noting that the choice of $\psi(x, y)$ can significantly impact the test's power. Hu and Lei[15] have recommended using the conditional likelihood ratio statistic:

$$\psi(y \mid x) = \frac{p_1(y \mid x)}{p_0(y \mid x)}.$$

However, it's essential to recognize that, in general, this statistic is not readily available unless strong assumptions are made about the parametric forms of $p_1(y \mid x)$ and $p_0(y \mid x)$. Instead, Hu and Lei[15] have employed machine learning methods to estimate the conditional likelihood ratio, making it a more practical and versatile choice for this type of testing. In cases where both the training and test sample sizes are small, the performance of their method remains uncertain or unclear.

6 Connections between Selection Bias Sampling and Shift Problems

In this section, we will establish a connection between traditional selection bias sampling problems and shift problems in the fields of machine learning and artificial intelligence.

We assume the training data with joint density

$$(X_i, Y_i) \sim p_0(x, y), \qquad i = 1, 2, \cdots, n_0$$

and the test data with joint density

$$(X_i, Y_i) \sim p_1(x, y) = \frac{\pi(x, y)p_0(x, y)}{\Delta}, \quad \Delta = \iint \pi(x, y)p_0(x, y)\mathrm{d}x\mathrm{d}y, \ i = n_0+1, n_0+2, \cdots, n.$$

The term $\pi(x,y)$ represents the selection bias. For the sake of simplicity, we can assume without loss of generality that:

$$0 \leqslant \pi(x,y) \leqslant 1.$$

This is because the density remains unchanged when a constant is divided both in the numerator and denominator. We will explore several cases based on these assumptions.

1) If $\pi(x,y) = \pi(x)$, then
$$p_1(x,y) = \frac{\pi(x)p_0(y\,|\,x)p_0(x)}{\Delta}.$$
The marginal density of X and conditional density of Y given X in the test sample are, respectively,
$$p_1(x) = \frac{\pi(x)p_0(x)}{\int \pi(x)p_0(x)\mathrm{d}x} \neq p_0(x)$$
and
$$p_1(y\,|\,x) = \frac{\pi(x)p_0(y\,|\,x)p_0(x)}{\pi(x)p_0(x)} = p_0(y\,|\,x).$$
This becomes the covariate shift problem.

2) If $\pi(x,y) = \pi(y)$, then
$$p_1(x,y) = \frac{\pi(y)p_0(x,y)}{\Delta}.$$
The marginal density of Y and conditional density of X given Y in the test sample are, respectively,
$$p_1(y) = \int \frac{\pi(y)p_0(y)p_0(x\,|\,y)}{\Delta}\mathrm{d}x = \frac{\pi(y)p_0(y)}{\int \pi(y)p_0(y)\mathrm{d}y} \neq p_0(y)$$
and
$$p_1(x\,|\,y) = \frac{\pi(y)p_0(x,y)}{\pi(y)p_0(y)} = p_0(y\,|\,x).$$
This is the prior probability shift problem, or label shift problem if Y is a binary variable. The conditional density X for given Y remains to be the same in both data sets.

3) If the selection probability $\pi(x,y)$ depends on both variables X and Y and the specific form of $\pi(x,y)$ is entirely unknown, this presents a challenging scenario. In such a case, even if test data are available, they don't provide any meaningful information. It's important to consider that:
$$p_1(x,y) = \frac{\pi(x,y)p_0(x,y)}{\Delta},$$

here, $\Delta = \iint \pi(x,y)p_0(x,y)\mathrm{d}x\mathrm{d}y$, represents the normalization factor obtained through the double integration of the product of $\pi(x,y)$ and $p_0(x,y)$. Since the form of $\pi(x,y)$ is entirely unspecified, even if we have complete knowledge of $p_0(x,y)$, the test data essentially possess an arbitrary joint density. In essence, $\pi(x,y)$ consumes and dominates the joint density of $p_0(x,y)$, rendering the information from $p_0(x,y)$ effectively inaccessible.

When we have complete knowledge of $\pi(x,y) = \pi_0(x,y)$, the relationship between $p_1(x,y)$ and $p_0(x,y)$ becomes more informative. In this scenario:

$$p_1(x,y) = \frac{\pi(x,y)p_0(x,y)}{\Delta}.$$

In this case, the information contained in $p_1(x,y)$ can provide meaningful insights into the parameter within $p_0(x,y)$.

Moreover, when our knowledge of $\pi(x,y)$ is limited to certain unspecified finite parameters, the information contained within $p_1(x,y)$ is notably weakened in its capacity to inform us about the parameter within $p_0(x,y)$. In such cases, the uncertainty surrounding the finite parameters in $\pi(x,y,\theta)$ hinders the strength of the information transfer from $p_1(x,y)$ to the parameter of interest in $p_0(x,y)$.

4) If $\pi(x,y)$ can be expressed as a product of $\pi_0(x)$ and $\pi_1(y)$, the relationship between $p_1(x,y)$ and $p_0(x,y)$ takes on the following form:

$$p_1(x,y) = \frac{\pi_0(x)\pi_1(y)p_0(y\,|\,x)p_0(x)}{\Delta}.$$

However, it's important to recognize that the two marginal densities differ from their respective counterparts in $p_0(x,y)$:

$$\begin{aligned} p_1(x) &= \frac{\pi_0(x)\int \pi_1(y)p_0(y\,|\,x)\mathrm{d}y p_0(x)}{\Delta} \neq p_0(x), \\ p_1(y) &= \frac{\int \pi_0(x)\pi_1(y)p_0(x\,|\,y)\mathrm{d}x p_0(y)}{\Delta} \neq p_0(y). \end{aligned}$$

The two conditional densities also deviate from those in $p_0(x,y)$:

$$\begin{aligned} p_1(y\,|\,x) &= \frac{\pi_1(y)p_0(y\,|\,x)}{\int \pi_1(y)p_0(y\,|\,x)\mathrm{d}y} \neq p_0(y|x), \\ p_1(x\,|\,y) &= \frac{\pi_0(x)p_0(x\,|\,y)}{\int \pi_0(x)p_0(x\,|\,y)\mathrm{d}x} \neq p_0(x,y). \end{aligned}$$

In this scenario, both the marginal and conditional densities have undergone changes. Notably, even without knowledge of the exact forms of $\pi_0(x)$ and $\pi_1(y)$, the test data still retain some level of information regarding the underlying parameters in $p_0(y \mid x^\mathsf{T}\theta)$ if a parametric model is assumed, albeit in a relatively weak form.

In fact, if we condition on $X_{n_0+1}, X_{n_0+2}, \cdots, X_n$ and order statistics $y_{(n_0+1)}, y_{(n_0+2)}, \cdots, y_{(n)}$ of $(Y_{n_0+1}, Y_{n_0+2}, \cdots, Y_n)$, the observed data has likelihood [17]

$$\begin{aligned}&f(y_{n_0+1}, y_{n_0+2}, \cdots, y_n \mid x_1, x_2, \cdots, x_n; y_{(n_0+1)}, y_{(n_0+2)}, \cdots, y_{(n)})\\ =\ &\frac{\prod_{i=n_0+1}^{n} p_0(y_i \mid x_i^\mathsf{T}\theta)}{\sum_{i_1,i_2,\cdots,i_m}\prod_{j=1}^{m} p_0(y_{i_j} \mid x_j^\mathsf{T}\theta)},\end{aligned}$$

where $m = n - n_0$, $i_1, i_2, \cdots, i_m$ are all possible permutations of $n_0+1, n_0+2, \cdots, n$. Note that $\pi_0(x)$, $\pi_1(y)$, $p_0(x)$ and the normalized constant $\Delta(\theta)$ are cancelled out. In order to alleviate the computational burden associated with permutations, Liang and Qin [18] have utilized the pairwise conditional technique, which serves to eliminate the unknown selection function and enhance the efficiency of the analysis.

To sum up, when amalgamating information from various data sources, a comprehensive grasp of the data generation process is paramount. This comprehension facilitates the judicious choice of suitable modeling techniques, thereby guaranteeing more precise and dependable results. It is imperative to acknowledge that employing a simplistic aggregation approach, treating distinct source data as identically distributed, can lead to biased outcomes.

7 Transfer Learning

The key idea of statistical transfer learning is grounded in the understanding that the world is rich with data, and data in one context often carries valuable insights that can be harnessed to solve problems in another domain. Whether it's image recognition, natural language processing, recommendation systems, or a myriad of other applications, the ability to capitalize on information from related or unrelated domains can significantly boost the performance of machine learning models. The relevance of statistical transfer learning spans various domains, from healthcare and finance to robotics and natural language understanding. By fostering the reuse of knowledge, it offers the potential to reduce the need for extensive labeled data and accelerate the development of models, ultimately leading to more effective and efficient machine learning systems. While the notion of transfer learning holds significant appeal, its practical application presents a challenge in quantifying the shift function between different studies.

Here, we present an example illustrating the seamless transfer of data from one model to another. Our focus is on datasets originating from separate but interrelated distributions. In this scenario, we assume that the training data adheres to a conventional generalized linear model, while the testing data exhibits a connection to the training data, driven by a prior probability shift assumption. Through our exploration, we uncover an inherent relationship between the conditional means of these two distinct samples, governed by an undisclosed, monotonically increasing function. To tackle these intricacies, we harness the combined power of generalized estimating equations and the shape-constrained score function. This approach allows us to construct a robust framework for enhancing inference regarding the underlying parameters, facilitating a deeper understanding of the data interplay between the two models.

Consider the training sample $(X_1, Y_1), (X_2, Y_2), \cdots, (X_{n_0}, Y_{n_0})$ drawn from the joint density $f(y \mid x)g(x)$, where the conditional density of Y given covariate X belongs to a canonical exponential family

$$f(y \mid x) = \exp\{[\theta(x)y - b(\theta(x))]/a(\phi) + c(y, \phi)\}$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. The parameter $\theta(\cdot)$ is called the canonical parameter and ϕ is called the dispersion parameter. Under this model, easily we have (for example, [19])

$$\mu(x) = \mathsf{E}(Y \mid x) = b'(\theta(x)), \qquad V(Y \mid x) = a(\phi)b''(\theta(x)),$$

where the primes denote differentiation with respect to θ. Sine $V(Y \mid x) \geqslant 0$, this implies that $b''(\cdot) \geqslant 0$, or equivalently $b'(\cdot)$ is a monotone non-decreasing function. In parametric generalized linear models, the unknown regression function $\mu(x)$ is modelled linearly via a link function η

$$\eta[\mu(x)] = x^{\mathsf{T}}\beta.$$

If η is the inverse function of b', i.e., $\eta = (b')^{-1}$, then η is called the canonical link function since in that case $\eta[\mu(x)]$ is the canonical parameter in the exponential family. In order to estimate β, one may use the optimal estimating equation

$$\sum_{i=1}^{n} x_i(b'')^{-1}(\theta(x_i))[y_i - \mu(x_i)]\mu'(x_i) = 0.$$

In practical applications, one needs to choose the link function η. McCullagh and Nelder [19] listed the possible choice of η for commonly used parametric families, such as normal, Poisson, Binomial, Gamma and inverse Gaussian etc.

Subsequently, we make the assumption that the test data $(X_{n_0+1}, Y_{n_0+1}), (X_{n_0+2}, Y_{n_0+2}), \cdots, (X_n, Y_n)$ is drawn from the distribution given by

$$\frac{\pi(y)f(y\,|\,x)g(x)}{\int \pi(y)f(y\,|\,x)g(x)\mathrm{d}y\mathrm{d}x}.$$

Easily we can show that for $j = n_0+1, n_0+2, \cdots, n$, the conditional density of Y_j given X_j is

$$\begin{aligned} Y_j\,|\,X_j &\sim \frac{\pi(y)f(y\,|\,x)}{\int \pi(y)f(y\,|\,x)\mathrm{d}y} \\ &= \frac{\pi(y)\exp\{c(y,\phi)\}\exp\{[\theta(x)y - b(\theta(x))]/a(\phi)\}}{\int \pi(y)\exp\{c(y,\phi)\}\exp\{[\theta(x)y - b(\theta(x))]/a(\phi)\}\mathrm{d}y} \\ &= \frac{h(y)\exp\{\theta(x)y/a(\phi)\}}{\int h(y)\exp\{\theta(x)y/a(\phi)\}\mathrm{d}y}, \end{aligned}$$

where $h(y) = \pi(y)\exp\{c(y,\phi)\}$.

Put simply, the test data can be viewed as a version of the training data that has undergone a prior probability shift. It can be demonstrated that this shifted data still belongs to the exponential family. However, when we lack knowledge of $\pi(y)$, it leads to the following relationships:

With training data:

$$\mathsf{E}^{\mathrm{Tr}}(Y\,|\,x) = \mu(x^{\mathsf{T}}\beta),$$

and with test data

$$\mathsf{E}^{\mathrm{Test}}(Y\,|\,x) = g(x^{\mathsf{T}}\beta),$$

where E^{Tr} and $\mathsf{E}^{\mathrm{Test}}$ are expectations with respect to training and test data, respectively. In these equations, $\mu(\cdot)$ represents a monotone function with a known form, while $g(\cdot)$ is a monotone function with an unknown specific form. Our primary objective is to minimize the following expression while considering constraints that require $g(\cdot)$ to be a monotone non-decreasing function:

$$\sum_{i=1}^{n_0}[Y_i - \mu(X_i^{\mathsf{T}}\beta)]^2 + \sum_{i=n_0+1}^{n}[Y_i - g(x_i^{\mathsf{T}}\beta)]^2.$$

Let's break this down into two steps:

Step 1 For fixed β, minimizing the second term with respect to g is a standard isotonic regression problem. One may use the well known pool adjacent violator algorithm to accomplish the minimization, for example [20] and [21]. Let $\widehat{g}_\beta$ be the minimizer.

Step 2 The next step involves minimizing the following expression with respect to β:

$$\sum_{i=1}^{n_0}[Y_i - \mu(X_i^{\mathsf{T}}\beta)]^2 + \sum_{i=n_0+1}^{n}[Y_i - \widehat{g}_\beta(x_i^{\mathsf{T}}\beta)]^2.$$

It's important to note that, in general, the large sample theory for shape-restricted inference entails advanced empirical process theory and can be quite intricate.

Maity et al.[22] have assumed the following transfer learning models for two sample binary outcome data. Assume that for given X, the outcome models are

$$\mathsf{P}_0(Y = 1 \,|\, x) = \eta_0(x), \qquad \mathsf{P}_1(Y = 1 \,|\, x) = \eta_1(x)$$

for training and test data, respectively. They have assumed

$$\gamma(\eta_1(x)) = \gamma(\eta_0(x)) + \phi^{\mathsf{T}}(x)\theta,$$

where $\gamma(\cdot)$ is a link function and $\phi(x)$ is a known function of X. For example, γ is a logit transformation.

In the broader context of data analysis and transfer learning, it's crucial to approach the modeling of the transferring function from one dataset to another with great care. This process involves understanding and characterizing the relationship between datasets. While the mathematical and statistical tools are vital, sometimes, incorporating insights from the physical or biological domain can be invaluable. In certain cases, there may be physical or biological evidence that provides crucial guidance in identifying the underlying structure governing data relationships. This evidence can serve as a compass, helping us navigate the intricate landscape of transfer learning and model building. It offers a path to uncover the hidden connections and dependencies between datasets.

It's important to note that simply merging different datasets without distinguishing their inherent differences can lead to biased or erroneous results. Each dataset may have its unique characteristics, sources of variation, and biases. Neglecting these distinctions can lead to misinterpretations and hinder the discovery of meaningful patterns. In summary, careful consideration and domain-specific insights are essential in the process of modeling the transfer function between datasets. By acknowledging the uniqueness of each dataset and leveraging any available physical or biological evidence, we enhance our ability to uncover the intricate structure that connects data, ultimately leading to more robust and accurate results.

8 Conditional Approach in Conformal Inference in the Presence of Covariate Shift

Conformal prediction (CP) is a powerful concept introduced by Vovk et al.[23] in 2005, and it was further elucidated in a tutorial by Shafer and Vovk[24] in 2008. CP comprises a suite of algorithms designed to evaluate the uncertainty associated with predictions generated by machine learning models. One of the compelling attributes of conformal inference is its distribution-free nature. It is capable of generating predictive intervals that provide non-asymptotic guarantees of coverage accuracy, even in the absence of specific distributional assumptions or model assumptions. This flexibility allows CP to be seamlessly integrated with a wide range of pre-trained models, including random forests, various machine learning algorithms, and neural networks, making it a versatile tool for enhancing prediction accuracy. In recent years, the adoption of conformal prediction has surged in the fields of computer science and statistics. Notable references include the works of [25–28], among others. For a comprehensive review of the latest developments in conformal inference, readers are encouraged to explore the work of [29]. This method has garnered substantial attention due to its unique ability to provide reliable and interpretable predictions while accommodating diverse modeling scenarios, making it a valuable asset in the realm of machine learning and statistical analysis. To better understand conformal predictive interval, we start from briefly reviewing the traditional prediction interval approach and then pointing the connections between them.

Suppose $(Y_1, X_1), (Y_2, X_2), \cdots, (Y_n, X_n)$ are independent and identically distributed observations from a random vector (Y, X), where given $X = x$, Y follows a parametric distribution $F(y \,|\, x^{\mathsf{T}}\theta)$. We are interested in predicting the future Y_{n+1} for a given covariate X_{n+1}. If θ is known, it is well known that $U(\theta) = F(Y \,|\, X^{\mathsf{T}}\theta)$ follows a uniform distribution on $[0, 1]$. For $\alpha \in (0, 0.5)$, we can construct a $100(1-2\alpha)\%$ prediction interval for Y_{n+1} by solving $\alpha \leqslant F(y \,|\, X_{n+1}^{\mathsf{T}}\theta) \leqslant 1-\alpha$ with respect to y.

In general, θ is unknown and can be estimated based on the observed data (X_i, Y_i), $i = 1, 2, \cdots, n$. Denote $\widehat{\theta}$ as an estimator, the most popular one is the maximum likelihood estimate. If it is consistent, then when n is large, $U(\widehat{\theta})$ should approximately follow a uniform distribution. When n is small, however there may have a gap between them. Let $G(u)$ be the cumulative distribution of $U(\widehat{\theta})$,

$$G(u) = \mathsf{P}\{U(\widehat{\theta}) \leqslant u\} = \mathsf{P}\{F(Y \,|\, X^{\mathsf{T}}\widehat{\theta}) \leqslant u\} = \mathsf{P}\{Y \leqslant F^{-1}(u \,|\, X^{\mathsf{T}}\widehat{\theta})\}.$$

In this definition, the randomness of the plugging in estimator $\widehat{\theta}$ is considered. Define the predictive distribution as

$$\widetilde{F}(y \mid x) = G(F(y \mid X^{\mathsf{T}}\widehat{\theta}).$$

The corresponding density is

$$\widetilde{f}_p(y \mid x) = g(F(y \mid X^{\mathsf{T}}\widehat{\theta}))f(y \mid X^{\mathsf{T}}\widehat{\theta}) = g(U(\widehat{\theta}))f(y \mid X^{\mathsf{T}}\widehat{\theta}),$$

when U is exactly pivotal. Harris[30] and Lawless and Fredette[31] showed that $\widetilde{f}_p(y \mid x)$ dominates the 'plug-in' predictive density $\widehat{f}_p(y \mid x) = f(y \mid x^{\mathsf{T}}\widehat{\theta})$ in terms of average Kullback-Leibler distance. Indeed if n is in the range between $10-50$, their simulation results show that the confidence intervals derived from the predictive distribution have better coverage than that from the plugging in method. On the other hand, if n is moderately large, these two methods are almost indistinguishable.

In traditional prediction approach, one needs to find $L(x)$ and $R(x)$ such that

$$\mathsf{P}\{L(X_{n+1}) \leqslant Y_{n+1} \leqslant R(X_{n+1}) \mid X_{n+1}\} = 100(1-2\alpha)\%.$$

In this situation, to get an accurate result, the conditional distribution $F(y \mid x^{\mathsf{T}}\theta)$ has to be correctly specified.

In contrast to the conditional probability calculation, in the latest conformal inference, one needs to find $L(x)$ and $R(x)$ such that

$$\mathsf{P}\{L(X_{n+1}) \leqslant Y_{n+1} \leqslant R(X_{n+1})\} = 100(1-2\alpha)\%,$$

where the probability is unconditional probability. The fundamental difference is that the conditional model $F(y \mid x^{\mathsf{T}}\theta)$ becomes a "working model", which implies that the conformal confidence interval is valid even if $F(y \mid x^{\mathsf{T}}\theta)$ may be misspecified. Of course, a correctly specified model can lead to a better confidence interval in the sense of shorter length.

If we arrange the observed outcomes in ascending order as follows:

$$Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}.$$

We can define $\widehat{C}_1$ and $\widehat{C}_2$ as the lower and upper order statistics corresponding to the $\lfloor n\alpha/2 \rfloor$-th and $\lceil n(1-\alpha/2) \rceil$-th positions, respectively. It's evident that the interval $[\widehat{C}_1, \widehat{C}_2]$ is a trivial conformal confidence interval and has an asymptotic coverage of $(1-\alpha)\%$. However, this confidence interval may lack essential information since it doesn't utilize covariate information. A desirable predictive confidence interval should combine accuracy in coverage and a compact length.

Taking on a more challenging perspective, we aim to build a predictive confidence interval that accounts for distributional shifts in the test sample. Let's assume we have training data $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ drawn from the joint density $f(y, x)$, and a test data point (X_{n+1}, Y_{n+1}) drawn from the joint density $g(x, y)$, where $g(x, y) = w(x)f(x, y)$ is a biased version of $f(x)$. We assume that $w(x)$ is completely known. Given the event A: the observed values $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ and (x_{n+1}, y_{n+1}) (note that y_{n+1} is not available), and the information that there are n individuals from f and one individual from g, the conditional probability is

$$\begin{aligned}
&\mathsf{P}\{(X_{n+1}, Y_{n+1}) = (x_i, y_i) \,|\, A\} \\
&= \frac{g(x_i, y_i) \prod_{j \neq i} f(x_j, y_j)}{g(x_1, y_1) \prod_{j=2}^{n} f(x_j, y_j) + \cdots + \prod_{j=1}^{n} f(x_j, y_j) g(x_{n+1}, y_{n+1})} \\
&= \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)}, \qquad i = 1, 2, \cdots, n+1.
\end{aligned}$$

Let $S(X_i, Y_i)$ be the score function, for example $S(X_i, Y_i) = |Y_i - \mu(X_i)|$, $i = 1, 2, \cdots, n+1$. The following conditional probability can be evaluated through

$$\mathsf{P}\{S(X_{n+1}, Y_{n+1}) = S(x_i, y_i) \,|\, A\} = \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)}, \qquad i = 1, 2, \cdots, n+1.$$

As a consequence,

$$\mathsf{P}\{S(X_{n+1}, Y_{n+1}) \leqslant c\} = \sum_{i=1}^{n+1} \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)} I(S(x_i, y_i) \leqslant c).$$

Since y_{n+1} is unknown, this probability is upper bounded by

$$\sum_{i=1}^{n} \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)} I(S(x_i, y_i) \leqslant c) + \frac{w(x_{n+1})}{\sum_{j=1}^{n+1} w(x_j)}$$

and lower bounded by

$$\sum_{i=1}^{n} \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)} I(S(x_i, y_i) \leqslant c).$$

In fact the lower bound is equivalent to defining $S(x_{n+1}, y_{n+1}) = \infty$ and is conservative.

We may choose L and R such that

$$\mathsf{P}\{S(X_{n+1}, Y_{n+1}) \leqslant L\} = \sum_{i=1}^{n} \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)} I(S(x_i, y_i) \leqslant c) \leqslant \alpha/2$$

and

$$\mathsf{P}\{S(X_{n+1}, Y_{n+1}) \leqslant R\} = \sum_{i=1}^{n+1} \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)} I(S(x_i, y_i) \leqslant R) + \frac{w(x_{n+1})}{\sum_{j=1}^{n+1} w(x_j)} \geqslant 1 - \alpha/2.$$

Finally we can solve

$$L < S(X_{n+1}, Y_{n+1}) \leqslant R$$

for Y_{n+1} to get a $1-\alpha$ conformal predictive confidence interval.

Remark 1 Tibshirani et al.[32] proposed the data splitting conditional method in the case that the shift function $w(x)$ is known completely. In practical applications, this might not be true. We have to use training data and test data to estimate the underlying parameters in w. As a consequence, the conformal confidence intervals discussed above only have correct coverage asymptotically. The application of their method into casual inference is studied recently by Lei and Candés[27].

Remark 2 Note that

$$\frac{1}{n_0}\sum_{i=1}^{n_0} w(x_i) \to \mathsf{E}_F[w(X)] = \Delta,$$

$$\frac{1}{n_0}\sum_{i=1}^{n_0} w(x_i) I(S(x_i, y_i) \leqslant c) \to \mathsf{E}_F[w(X) I(S(X,Y) < c)].$$

Therefore

$$\frac{\sum_{i=1}^{n_0} w(x_i) I(S(x_i, y_i) \leqslant c)}{\sum_{i=1}^{n_0} w(x_i)} \to \frac{\mathsf{E}_F[w(X) I(S(X,Y) < c)]}{\mathsf{E}_F[w(X)]} = \mathsf{E}_G[I(S(X,Y) < c)],$$

where F and G are distribution functions corresponding to f and g, respectively. Given the absence of direct observations from $g(x,y) = w(x)f(x,y)$, the conditional approach described above essentially utilizes training data to construct a weighted estimator of the probability function.

Conformal inference, while valuable, primarily ensures marginal coverage, which might not be the most appealing aspect for those more concerned with conditional coverage. Achieving conditional accuracy coverage is contingent on correctly specifying the conditional distribution of Y given X. If the dimension of X is relatively low, nonparametric estimation of the conditional density is feasible through kernel methods. However, in cases where the dimension of X is high, this approach becomes impractical due to the well-known curse of dimensionality problem.

It's important to note that while conformal inference can provide accuracy on average marginally, its performance can still vary substantially between subpopulations. For instance, consider an algorithm deployed for recidivism prediction, which may exhibit significantly higher false positive rates for African-American parolees compared to Caucasian parolees[33]. This highlights the need for fairness audits in statistics to address and rectify such disparities in real-world applications, especially in contexts where automated systems and algorithms significantly impact people's lives.

In practical applications, ensuring conditional coverage holds true for specific pre-specified subgroups is often necessary. It's essential to address fairness and equity not only at the aggregate level but also within subpopulations.

Consider the following:

Initially, we aim to ensure that the probability of Y_{n+1} falling within $\widehat{C}(X_1, Y_1, X_2, Y_2, \cdots, X_n, Y_n; X_{n+1})$ is $1-\alpha$. This can be expressed as:

$$\mathsf{E}[I(Y_{n+1} \in \widehat{C}) - (1-\alpha)] = 0.$$

If

$$\mathsf{E}\{[I(Y_{n+1} \in \widehat{C}) - (1-\alpha)] \,|\, X_{n+1}\} = 0,$$

then

$$\mathsf{E}\{h(X)[I(Y_{n+1} \in \widehat{C}) - (1-\alpha)]\} = 0$$

for any measurable function. If we let $h_i(x) = I(X \in G_k)$, $i = 1, 2, \cdots, k$, we need to impose

$$\mathsf{E}\{I(X_{n+1} \in G_i)[I(Y_{n+1} \in \widehat{C}) - (1-\alpha)]\} = 0, \qquad i = 1, 2, \cdots, k.$$

This implies that

$$\mathsf{P}\{I(Y_{n+1} \in \widehat{C}) \,|\, X_{n+1} \in G_i\} = 1-\alpha.$$

To make the above equality to be true simultaneous, one has to solve β and γ in the following equation,

$$\sum_{i=1}^{n_0} h(x_i)[I(Y_i \leqslant \beta + h(x_i)\gamma) - (1-\alpha)]w(x_i) = 0,$$

$w(x)$ is used to adjust the covariate shift problem. Mathematically this is equivalent to minimizing the check function [34]

$$\sum_{i=1}^{n_0} w(x_i)\rho_{1-\alpha}(Y_i - \beta - h^{\mathsf{T}}(x_i)\gamma)$$

with respect to β and γ, where

$$\rho_\tau(u) = u[\tau - I(u \leqslant \tau)].$$

While it is ideal to apply a comprehensive set of subgroup constraints, the feasibility of doing so may be limited when dealing with a small sample size. In real-world scenarios, it becomes necessary to strike a delicate balance between the number of constraints and the available sample size.

9 Concluding Remarks

Throughout the pages of this paper, we have embarked on an intellectual journey, delving deep into the intricate realms of biased sampling, navigating the challenges of distributional shift problems, exploring the nuances of transfer learning, understanding the pivotal role of conformal inference, and underscoring the urgent need for fairness audits in our increasingly data-driven world. Our objective has been to demystify these complex concepts and render them accessible to graduate students. Our aspiration is that, upon reading this paper, graduate students will not only gain a good understanding of these concepts but also be empowered to apply this knowledge within their own research domains. Much like the ancient Chinese proverb suggests, we hope that the skill gained from this paper will enable them to "cast a brick and attract jade", meaning that by addressing biased sampling issues and embracing these critical concepts, they will discover invaluable insights and solutions in their own work.

In conclusion, this paper serves as a tribute and a heartfelt homage to the memory of Professor Shisong Mao, whose indelible impact on the fields of statistics, education, and mentorship has left an enduring legacy. In our endeavor to honor and perpetuate Professor Mao's memory, we rekindle our fervor for the art of data collection, the pursuit of knowledge, and an unswerving devotion to mentorship and education. As we carry forward his legacy, we do so with a profound sense of appreciation for the wisdom he imparted and the profound impact he has had on all our lives, allowing his memory to remain as a seamless thread in the fabric of our academic and personal journey.

Acknowledgements I extend my heartfelt gratitude to Professors Yingchun Zhou for extending the invitation to contribute to this paper in honor of the late Professor Mao. It is an immense privilege to be part of this tribute to his legacy. I am also deeply thankful to Professors Jinglong Wang, Jixiang Zhou and Yukun Liu for their invaluable insights and constructive comments on this paper. Their expertise and feedback have enriched the content and quality of this work.

References

[1] HECKMAN J J. Sample selection bias as a specification error [J]. *Econometrica*, 1979, **47(1)**: 153–161.

[2] QIN J. *Biased Sampling, Over-identified Parameter Problems and Beyond* [M]. Singapore: Springer, 2017.

[3] COCHRAN W G. *Sampling Techniques* [M]. 3rd ed. New York: Wiley, 1977.

[4] VARDI Y. Nonparametric estimation in the presence of length bias [J]. *Ann Statist*, 1982, **10(2)**: 616–620.

[5] VARDI Y. Empirical distributions in selection bias models [J]. *Ann Statist*, 1985, **13(1)**: 178–203.

[6] PRENTICE R L, PYKE R. Logistic disease incidence models and case-control studies [J]. *Biometrika*, 1979, **66(3)**: 403–411.

[7] QIN J, ZHANG B. A goodness-of-fit test for logistic regression models based on case-control data [J]. *Biometrika*, 1997, **84(3)**: 609–618.

[8] ROSENBAUM P R, RUBIN D B. The central role of the propensity score in observational studies for causal effects [J]. *Biometrika*, 1983, **70(1)**: 41–55.

[9] QIN J, ZHANG B. Empirical-likelihood-based inference in missing response problems and its application in observational studies [J]. *J R Stat Soc Ser B Stat Methodol*, 2007, **69(1)**: 101–122.

[10] ROBINS J M, ROTNITZKY A, ZHAO L P. Estimation of regression coefficients when some regressors are not always observed [J]. *J Amer Statist Assoc*, 1994, **89(427)**: 846–866.

[11] MORENO-TORRES J G, RAEDER T, ALAIZ-RODRÍGUEZ R, et al. A unifying view on dataset shift in classification [J]. *Pattern Recognit*, 2012, **45(1)**: 521–530.

[12] ANDERSON J A. Multivariate logistic compounds [J]. *Biometrika*, 1979, **66(1)**: 17–26.

[13] QIN J. Empirical likelihood ratio based confidence intervals for mixture proportions [J]. *Ann Statist*, 1999, **27(4)**: 1368–1384.

[14] LINDSAY B G. *Mixture Models: Theory, Geometry and Applications* [M]. Institute of Mathematical Statistics, American Statistical Association, 1995.

[15] HU X Y, LEI J. A two-sample conditional distribution test using conformal prediction and weighted rank sum [J/OL]. *J Amer Statist Assoc*, 2023 [2023-3-8]. https://doi.org/10.1080/01621459.2023.2177165.

[16] THAMS N, SAENGKYONGAM S, PFISTER N, et al. Statistical testing under distributional shifts [J]. *J R Stat Soc Ser B Stat Methodol*, 2023, **85(3)**: 597–663.

[17] KALBFLEISCH J D. Likelihood methods and nonparametric tests [J]. *J Amer Statist Assoc*, 1978, **73(361)**: 167–170.

[18] LIANG K Y, QIN J. Regression analysis under non-standard situations: a pairwise pseudolikelihood approach [J]. *J R Stat Soc Ser B Stat Methodol*, 2000, **62(4)**: 773–786.

[19] MCCULLAGH P, NELDER J A. *Generalized Linear Models* [M]. 2nd ed. London: Chapman and Hall, 1989.

[20] BARLOW R E, BARTHOLOMEW D J, BREMNER J M, et al. *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression* [M]. New York: Wiley, 1972.

[21] GROENEBOOM P, JONGBLOED G. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics* [M]. Cambridge: Cambridge University Press, 2014.

[22] MAITY S, DUTTA D, TERHORST J, et al. A linear adjustment-based approach to posterior drift in transfer learning [J]. *Biometrika*, 2024, **111(1)**: 31–50.

[23] VOVK V, GAMMERMAN A, SHAFER G. *Algorithmic Learning in a Random World* [M]. New York: Springer, 2005.

[24] SHAFER G, VOVK V. A tutorial on conformal prediction [J]. *J Mach Learn Res*, 2008, **9**: 371–421.

[25] LEI J, WASSERMAN L. Distribution-free prediction bands for non-parametric regression [J]. *J R Stat Soc Ser B Stat Methodol*, 2014, **76(1)**: 71–96.

[26] LEI J, G'SELL M, RINALDO A, et al. Distribution-free predictive inference for regression [J]. *J Amer Statist Assoc*, 2018, **113(523)**: 1094–1111.

[27] LEI L H, CANDÈS, E J. Conformal inference of counterfactuals and individual treatment effects [J]. *J R Stat Soc Ser B Stat Methodol*, 2021, **83(5)**: 911–938.

[28] BARBER R F, CANDÈS E J, RAMDAS A, et al. The limits of distribution-free conditional predictive inference [J]. *Inf Inference*, 2021, **10(2)**: 455–482.

[29] ANGELOPOULOS A N, BATES S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification [OL]. 2021 [2021-7-15]. https://arxiv.org/abs/2107.07511.

[30] HARRIS I R. Predictive fit for natural exponential families [J]. *Biometrika*, 1989, **76(4)**: 675–684.

[31] LAWLESS J F, FREDETTE M. Frequentist prediction intervals and predictive distributions [J]. *Biometrika*, 2005, **92(3)**: 529–542.

[32] TIBSHIRANI R J, BARBER R F, CANDÈS E J, et al. Conformal prediction under covariate shift [OL]. 2019 [2019-4-12]. https://arxiv.org/abs/1904.06019.

[33] CHERIAN J J, CANDÈS E J. Statistical inference for fairness auditing [OL]. 2023 [2023-5-5]. https://arxiv.org/abs/2305.03712.

[34] GIBBS I, CHERIAN J J, CANDÈS E J. Conformal prediction with conditional guarantees [OL]. 2023 [2023-5-22]. https://arxiv.org/abs/2305.12616.

偏倚抽样问题的选择性评论及其在现代统计学中的应用

秦进

(美国国家卫生研究院国家过敏和传染病研究所, 贝塞斯达, MD 20892)

摘　要: 偏倚抽样是一个普遍存在的问题, 跨越各个学科领域, 影响着计量经济学、流行病学、医学、调查研究, 以及最近的机器学习和人工智能 (AI) 等领域. 当选择用于分析或研究的数据点引入系统性偏倚时, 这种无处不在的挑战可能会影响研究结果的准确性和可靠性. 本文的目标是全面介绍与偏倚抽样问题相关的基础概念和推理方法. 此外, 我们还旨在建立偏倚抽样问题与机器学习中关于分布转移问题的最新讨论之间的联系. 我们还将深入探讨偏倚抽样的最新进展, 特别是在转移学习和预测置信区间的符合推理方面. 我们的最终目标是以一种对研究生易于理解的方式呈现这些材料, 使他们能够在自己的研究工作中识别偏倚抽样问题的应用.

我们怀着深深的敬意和感激之情, 将本文献给已故的茆诗松教授, 他多年来的指导和智慧对我们至关重要.

关键词: 偏倚抽样问题; 因果推断; 符合预测区间; 分布转移; 迁移学习; 茆诗松教授的追忆

中图分类号: O212.1; O212.7