2016

STATISTICAL METHODS FOR ENVIRONMENTAL EXPOSURE DATA SUBJECT TO DETECTION LIMITS

Yuchen Yang
University of Kentucky, yuchen.y@uky.edu
Author ORCID Identifier:
http://orcid.org/0000-0001-7821-6993

Digital Object Identifier: https://doi.org/10.13023/ETD.2016.467

Yuchen Yang, Student

Dr. Richard J. Kryscio, Major Professor

Dr. Constance L. Wood, Director of Graduate Studies

STATISTICAL METHODS FOR ENVIRONMENTAL EXPOSURE DATA SUBJECT TO DETECTION LIMITS

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of
Philosophy in the College of Arts and Sciences
at the University of Kentucky

By

Yuchen Yang

Lexington, Kentucky

Directors: Dr. Richard J. Kryscio, Professor of Statistics

and Dr. Li Chen, Associate Professor of Biostatistics

Lexington, Kentucky

2016

ABSTRACT OF DISSERTATION

STATISTICAL METHODS FOR ENVIRONMENTAL EXPOSURE DATA SUBJECT TO DETECTION LIMITS

In this dissertation, we develop unified and efficient nonparametric statistical methods for estimating and comparing environmental exposure distributions in presence of detection limits. In the first part, we propose a kernel-smoothed nonparametric estimator for the exposure distribution without imposing any independence assumption between the exposure level and detection limit. We show that the proposed estimator is consistent and asymptotically normal. Simulation studies demonstrate that the proposed estimator performs well in practical situations. A colon cancer study is provided for illustration. In the second part, we develop a class of test statistics to compare exposure distributions between two groups by using the integrated weighted difference in the kernel-smoothed estimator proposed in the first part. We study the conditions on the weight function such that the test statistics are stable, i.e. the asymptotic variances are finite. Simulation studies demonstrate that the proposed tests preserve type I errors regardless whether the distributions of the detection limit in the two groups differ or not and are more efficient than current methods in certain situations. A colon cancer study is provided for illustration. In the third part, we extend the estimation and testing methods developed in the part one and two to survey data by incorporating sampling weights. The results of several simulation studies are reported to demonstrate the performance of the proposed methods. The Jackknife method is utilized for the variance estimation to account for complex sample designs.

KEYWORDS: detection limits; left-censored data; environmental exposure; kernel smoothing; nonparametric estimator; integrated weighted difference; sampling weight; jackknife

Yuchen Yang

December 9, 2016

STATISTICAL METHODS FOR ENVIRONMENTAL EXPOSURE DATA SUBJECT TO DETECTION LIMITS

By

Yucheng Yang

Director of Dissertation: Dr. Richard J. Kryscio

Co-Director of Dissertation: Dr. Li Chen

Director of Graduate Studies: Dr. Constance Wood

Date: December 9, 2016

ACKNOWLEDGMENTS

I would like to thank everybody who has supported, helped and contributed to my graduate studies in the past five years. First of all, I want to express appreciations to my committee chair, Dr. Richard Kryscio, for his tremendous support throughout my graduate school life.. Without Dr. Kryscios guidance and persistent help this dissertation would not have been possible.

Secondly, I would like to express the greatest thanks and deepest appreciation to my committee co-chair, Dr. Li Chen, for her tremendous patience and continuous support during my graduate study and research. Dr. Chen has been an excellent advisor, professional mentor, and great friend. Her patience, motivation, and immense knowledge in statistics helped me in all the time of research and writing of this dissertation. Thanks to Dr. Chen for all the weekend nights she has worked on guiding my dissertation. I could not have imagined having a better advisor and mentor for my Ph.D study.

Thirdly, besides Dr. Kryscio and Dr. Chen, I would like to thank the rest of my dissertation committee: Dr. Brent Shelton, Dr. William Griffith and Dr. Yanbing Zheng for their insightful comments and encouragement, but also for the wonderful questions which have incented me to widen my research from various perspectives.

Finally, I would like to say thank you to my wife Ran for her love and company, and for being my closest friend. Also, I want to thank my parents for loving me and raising me. Without them, I could not have accomplished what has been done. Now, a new life has been expanding in front of me. I want to thank them again for being there in my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Motivation

As researchers investigate the relationship between diseases and exposures to environmental chemicals such as trace elements, pesticides, and dioxins, they often find concentrations that are lower than limits deemed reliable enough to report as numerical values. A detection limit (DL) is “a threshold below which measured values are not considered significantly different from a blank signal, at a specified level of probability” [1]. Therefore, the exposure level of a chemical for a sample is only reported when its value is not less than the DL and otherwise is reported as a less than value or non-detect. The DL may be a fixed number in some studies, but it can also vary widely from sample to sample in other studies. For the latter, the DL may be associated with the exposure level, as observed in a colon cancer study in Kentucky[2]. The data subject to DLs present challenges for data analysis and interpretation. In this dissertation we focus on two important statistical problems encountered in the analysis of data from environmental epidemiologic studies: (a) estimation of the chemical distribution in a specific group; and (b) comparison of distributions among groups. For these two problems, ad hoc, parametric, and nonparametric methods have been proposed. Ad hoc methods are ill-advised unless there are relatively few measurements below DLs; and parametric methods can lead to markedly biased results when the parametric model is misspecified [3, 4]. Nonparametric methods have received increasing attention in recent years because of their robustness. However, current nonparametric methods simply borrow the commonly used methods for right-censored survival data, and do not take into account the following two unique characteristics of environmental exposure data with DLs: (a) it is not meaningful to define the hazard function for an exposure measurement; and (b) DL values are observable for all subjects including those whose actual exposure levels are detected. In addition, current nonparametric methods do not allow for sampling weights, which are typically present in survey data such as the National Health and Nutrition Examination Survey (NHANES). Due to these issues, current nonparametric methods may lead to the following four problems for the analysis of environmental exposure data with DLs: (a) lack of meaningful interpretation; (b) inefficient results; (c) inability to deal with the situation that the exposure level and DL are associated; and (d) inability to handle survey data with sampling weights. To address the aforementioned problems, we will develop unified and efficient nonparametric estimation and testing methods that can (a) deal with possible association between the exposure level and DL; (b) incorporate sampling weights. We will utilize state-of-the-art methods for censored survival data and tailor them to environmental exposure data with DLs. The proposed methods will be applied to data from a colon cancer case-control study in Kentucky and the NHANES data.

1.2 Colon Cancer Data

Kentucky has the nation's highest colon cancer incidence rate [5]. Appalachian Kentucky, which has a unique geology that contains high-quality bituminous coal naturally rich in trace elements, has an even higher rate of colon cancer compared to other regions of the state. A case-control study was conducted to explore the association between environmental exposures to trace elements such as arsenic (As), chromium (Cr) and nickel (Ni) and colon cancer and whether exposures to these trace elements contribute to the elevated colon cancer rate in Appalachian Kentucky [2, 6]. For this purpose, 274 colon cancer cases and 253 controls were selected from 23 contiguous rural counties in Kentucky (Appalachian region) and Jefferson County, the largest, most urban county in Kentucky (non-Appalachian region). Among 247 subjects from the Appalachian region, 145 were cases and 102 were controls; among

280 from the non-Appalachian region, 129 were cases and 151 were controls. Toenail samples from these subjects were collected, and the concentrations of 12 trace elements were measured as markers of long-term environmental exposures to these trace elements. The DL varies from one subject to another for these trace element concentrations as a function of the toenail mass. We found at least 6 trace elements with significant association between the exposure level and DL in cases from the Appalachian region.

1.3 National Health and Nutrition Examination Survey (NHANES) Data

The NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. Starting in 1999, NHANES became a continuous, ongoing annual survey of the noninstitutionalized civilian resident population of the United States. About 12,000 persons per 2-year cycle were asked to participate in NHANES. Response rates varied by year, but an average of 10,500 persons out of the initial 12,000 agreed to complete a household interview. A four-stage sampling design was used: (i) selection of Primary Sampling Units (PSUs), which are counties or small groups of contiguous counties; (ii) selection of segments within PSUs that constitute a block or group of blocks containing a cluster of households; (iii) selection of specific households within segments; and (iV) selection of individuals within a household. A weight was assigned to each respondent. Weighting took into account several features of the survey: the differential probabilities of selection for the individual domains; nonresponse to survey instruments; and differences between the final sample and the total population[7]. Masked Variance Strata and Masked Variance Units or MVUs are used to protect the confidentiality of information provided by survey participants and to reduce disclosure risks. The variance estimates that are produced, using the Masked strata and MVUs, closely approximate the variances that would have been estimated using the true sample design variance units that are based on the actual survey sample strata and PSUs[8].

1.4 Current methods

Ad hoc methods, such as substituting the DL, DL/2, or DL/$\sqrt{2}$ for values below the DL, are widely used in environmental science literatures for the analysis of the data subject to DLs. However, these methods have no theoretic basis and are ill-advised unless relatively few measures fall below DLs [3, 4]. Parametric models for left-censored data, such as the Tobit model and the lognormal model, can be used since the data subject to DLs can also be treated as left-censored data [1]. The caution of using these methods is that the validity of the results depends on the correct specification of the parametric model. Recently nonparametric methods have received increasing attention because they do not require distributional assumptions, and thus may be a safe choice for data analysis [1, 9, 10].

The nonparametric reverse Kaplan-Meier (RKM) estimator, which mimics the Kaplan-Meier (KM) estimator for right-censored survival data with the scale reversed, has been recommended for estimating the exposure distribution in presence of DLs [9]. Let $\widetilde{T}$ and D be random variables for the exposure level and DL, respectively. Let $T = max(\widetilde{T}, D)$ and $\delta = I(\widetilde{T} \geq D)$, where δ indicates whether T is an exposure level value or a DL value. For data subject to DL, we observe (T, δ, D) for each subject. Suppose the data consist of n replicates $\{(T_i, \delta_i, D_i)\colon\ i = 1, \cdots, n\}$. We then define two counting processes $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(T_i \leq t)$. The RKM estimator can be written as

$$\hat{F}_{RKM}(t) = \prod_{s>t} \left\{ 1 - \frac{\sum_{j=1}^{n} dN_j(s)}{\sum_{j=1}^{n} Y_j(s)} \right\}, \quad t \geq \tau_n, \tag{1.1}$$

where $\tau_n = \min_{i=1,\ldots,n}\{T_i\}$. In addition, when the smallest observation is uncensored, $\hat{F}_{RKM}(t) = 0$ for $t \in (0, \tau_n)$. When the smallest observation is censored, $\hat{F}_{RKM}(t)$ is undefined for $t \in (0, \tau_n)$. Standard errors of RKM estimates are readily available using the greenwood formula[11, 12] that was developed for the KM estimator [13].The RKM estimator requires the independence assumption between the exposure level and DL.

Nonparametric methods outperform ad hoc and parametric methods for comparing exposure distributions[10]. Current nonparametric methods transform (flip) data subject to DLs to right-censored and then apply the log-rank or Wilcoxon test (4). However, these methods do not account for the unique characteristics of environmental data compared to survival data and can cause several problems. First, these methods lack epidemiological interpretation because they are constructed by comparing hazard functions which are not meaningful quantities for environmental data. Second, these tests are not efficient for detecting the absolute difference in environmental exposure distributions. Third, the validity of these tests depends on the assumption that distributions of the DL in two groups are identical[14, 15]

In survival analysis, the weighted Kaplan-Meier (WKM) statistics have been proposed as alternatives to the log-rank or Wilcoxon test for comparing the absolute difference in two survival distributions[16]. The WKM statistics consider the integrated weighted difference in Kaplan-Meier estimates for the two groups and are defined as

$$\sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{w}(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt,$$

where $T_c = \sup\{t : \min(\hat{C}_1(t), \hat{C}_2(t))\}$, n_i is the sample size in group i, $n = n_1 + n_2$, $\hat{C}_i$ is the Kaplan- Meier estimator of the censoring survival function in group i, $\hat{S}_i$ is the Kaplan- Meier estimator of the survival function in group i, $\hat{w}(\cdot)$ is a random weight function estimating a deterministic function $w(\cdot)$ and $i = 1$ or 2. The variability of $\hat{S}_1(t) - \hat{S}_2(t)$ tends to be large for t close to T_c. In this case, the WKM statistics may have unstable results without using an appropriate weight function $w(\cdot)$. To remedy instability of the WKM statistic, $w(\cdot)$ needs to downweigh the contribution of $\hat{S}_1(t) - \hat{S}_2(t)$ over larger t. Under the null hypothesis, the WKM statistics are asymptotically normal when $w(\cdot)$ meets certain constraints. Small-sample simulation studies showed that the WKM statistics may perform better than the log-rank or Wilcoxon test under the crossing hazards alternative. However, the WKM statistics are for right-censored survival data and have not been extended to data subject to DLs.

1.5 Outline of the Dissertation

The remainder of this dissertation is organized as follows.

In chapter 2, we propose a kernel-smoothed nonparametric estimator for the exposure distribution without imposing any independence assumption between the exposure level and DL. We show that the proposed estimator is consistent and converges weakly to a Gaussian process. The results of several simulation studies are reported to demonstrate the performance of the estimator comparing to the RKM estimator and the parametric estimator based on a lognormal exposure distribution. A colon cancer study is provided for illustration.

In chapter 3, we develop a class of test statistics to compare exposure distributions between two groups by using the integrated weighted difference in the proposed estimators for the two groups. We study the condition of the weight function so that the propsed test statistics arevasymptotically normal. The results of several simulation studies are reported to demonstrate the performance of the test statistics. A colon cancer study is provided for illustration.

In chapter 4, we extend the proposed estimator and test statistics to complex survey data by incorporating sampling weights. The results of several simulation studies are reported to demonstrate the performance of the proposed methods. The Jackknife method is utilized for the variance estimation to account for complex sample designs. The NHANES data is provided for illustration.

In chapter 5, we implement the aforementioned methods to an R package 'krkm'.

Chapter 2 Estimation of Exposure Distribution Adjusting for Association between Exposure Level and Detection Limit

2.1 Introduction

2.2 Introduction

In environmental exposure studies, one fundamental question is to estimate distributions of environmental chemicals, such as trace elements and pesticides, in a certain population. However, it is very common to observe a portion of exposure measurements to fall below experimentally determined detection limits (DLs). A detection limit (DL) is "a threshold below which measured values are not considered significantly different from a blank signal, at a specified level of probability" [1]. Therefore, the exposure level of a chemical for a sample is only reported when its value is not less than the DL and otherwise is reported as a less than value or non-detect. The DL itself can depend on the mass/volume of the analyzed sample and/or on the mass/volume of adjustment factors such as lipid content. The laboratory may report a common DL for all samples or different DLs for different samples. When the latter occurs, it is mostly because the mass/volume of the obtained sample and/or any adjustment factor differs for each individual, and the exposure level and DL may be associated in this case. For example, in the colon cancer study measuring trace element accumulation in toenails [2], we observed a statistically significant association between the exposure level and DL in Appalachian cancer cases for at least 6 trace elements (Table 2.4). This may be because trace elements can cause adverse effects on metabolism and therefore lead to slow growth rate of toenails [17]. As a result, toenail samples obtained from individuals with high exposure to trace elements tend to have low masses. In addition, a higher toenail mass results in a lower DL (i.e., a better ability to detect low levels of metal accumulation). Therefore, the exposure level and DL may be associated because both may be associated with the toenail sample mass.

Ad hoc methods, such as substituting DL, DL/2, or DL/$\sqrt{2}$ for the value below a DL, are widely used in environmental science literature to estimate the exposure distribution for the data subject to DLs. However, these methods have no theoretical basis and are ill-advised unless relatively few measures fall below DLs [3, 4]. To appropriately account for values below DLs, parametric models for left-censored data, such as the lognormal model [1], can be used since the data subject to DLs can also be treated as left-censored data [1]. But these parametric methods can lead to markedly biased results when the parametric form of the exposure distribution is misspecified [1, 4]. Recently nonparametric methods have received increasing attention because they do not require distributional assumptions, and thus may be a safer choice for data analysis. The reverse Kaplan-Meier (RKM) estimator, which mimics the Kaplan-Meier (KM) estimator for right-censored survival data with the scale reversed, has been recommended [9]. Note that both the RKM estimator and the aforementioned parametric methods require the independence assumption between the exposure level and DL. To our knowledge, there are no appropriate statistical methods available to deal with the case when the exposure level and DL are associated.

In this chapter, we utilize a two-step strategy and the kernel smoothing technique to develop a nonparametric consistent estimator for the exposure distribution allowing for the situation when the exposure level and DL are dependent. We first estimate the conditional exposure distribution given the DL by adding kernel weights into the RKM estimator and then obtain the average of the estimated conditional distribution over all DL values in the sample to estimate the marginal exposure distribution. The proposed method does not require any independence assumption between the exposure level and DL and any distributional assumption about the exposure level.

In Section 2.3, we propose the estimator and show that it is consistent and converges weakly to a Gaussian process. In Section 2.4, the results of several simulation studies are reported to demonstrate the performance of the estimator comparing to the RKM estimator and the parametric estimator assuming a lognormal exposure distribution. In Section 2.5, a colon cancer study is provided for illustration. Finally, Section 2.6 contains discussions and some extensions.

2.3 Methods

Let $\widetilde{T}$ and D be random variables for the exposure level and DL, respectively, and $F(\cdot)$ be the cumulative distribution function (CDF) of the exposure level. Let $T = max(\widetilde{T}, D)$ and $\delta = I(\widetilde{T} \geq D)$. Here δ indicates whether T is an exposure level value or a DL value. For data subject to DL, only (T, δ, D) are observable for each subject. Suppose the data consist of n replicates $\{(T_i, \delta_i, D_i)\colon\ i = 1, \cdots, n\}$. Note that the method proposed below requires the DL to be known for each subject in the data.

It is useful to adopt the counting process notation. Analogous to the observed counting process and at-risk process for right censored survival data, we define two counting processes, $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(T_i \leq t)$, for the data subject to DLs. Then the RKM estimator can be rewritten as

$$\hat{F}_{RKM}(t) = \prod_{s>t} \left\{ 1 - \frac{\sum_{j=1}^{n} dN_j(s)}{\sum_{j=1}^{n} Y_j(s)} \right\}, \quad t \geq \tau_n, \tag{2.1}$$

where $\tau_n = \min_{i=1,\ldots,n}\{T_i\}$. In addition, when the smallest observation is uncensored, $\hat{F}_{RKM}(t) = 0$ for $t \in (0, \tau_n)$. When the smallest observation is censored, $\hat{F}_{RKM}(t)$ is undefined for $t \in (0, \tau_n)$. This estimator mimics the KM estimator for right-censored survival data with the scale reversed. Similar to the independence assumption between the survival time and censoring time for the KM estimator, the RKM estimator requires the independence assumption between the exposure level and DL and is not a consistent estimator when this assumption is violated.

To develop a consistent estimator for the exposure distribution allowing for the association between the exposure level and DL, we propose a two-step strategy based on the statistical fact that $F(t) = E_D\{F(t; D)\}$, where $F(t; d)$ is the conditional CDF of the exposure level given the DL, i.e. $F(t; d) = \Pr(\widetilde{T} \leq t \mid D = d)$, and E_D is the expectation with respect to D. In the first step, we obtain a consistent estimator for the conditional CDF of the exposure level, denoted by $\hat{F}(t; d)$. In the second step, we estimate $F(t)$ by the average of estimated conditional CDF values over all DL values in the sample, i.e. $\hat{F}(t) = n^{-1} \sum_{i=1}^{n} \hat{F}(t; D_i)$. Specifically, we estimate the conditional CDF by adding kernel weights into the RKM estimator in equation (2.1), i.e.

$$\hat{F}(t; d) = \prod_{s>t} \left[1 - \frac{\sum_{j=1}^{n} K\{(D_j - d)/h\} dN_j(s)}{\sum_{j=1}^{n} K\{(D_j - d)/h\} Y_j(s)} \right], \quad t \geq \tau_n,$$

where $K(\cdot)$ is a kernel function, and h is a bandwidth such that $nh \to \infty$ and $nh^4 \to 0$ as $n \to \infty$. Similar to the RKM estimator, when the smallest observation is uncensored, $\hat{F}(t; d) = 0$ and $\hat{F}(t) = 0$ for $t \in (0, \tau_n)$. When the smallest observation is censored, $\hat{F}(t; d)$ and $\hat{F}(t)$ are undefined for $t \in (0, \tau_n)$. The above estimator for the conditional CDF borrows the idea of the kernel conditional KM estimator which adds kernel weights into the KM estimator to estimate the conditional survival function for right-censored survival data [18]. In the following, the proposed estimator $\hat{F}(t)$ will be referred to as the KRKM estimator. Through the above two-step strategy, in order for $\hat{F}(t)$ to be a consistent estimator for the marginal CDF of the exposure level, we only need the estimator for the conditional CDF given the DL to be a consistent estimator. The latter only requires the conditional independence between the exposure level and DL given the DL. Since it is true that the exposure level and DL are independent given the DL, the KRKM estimator is consistent without requiring any independence assumption between the exposure level and DL. We show in Appendix A that $\sqrt{n}\{\hat{F}(t) - F(t)\}$ converges weakly to a zero-mean Gaussian process and is asymptotically equivalent to the process $n^{-1/2}\sum_{i=1}^{n}\xi_i(t)$, where

$$\xi_i(t) = F(t; D_i) - F(t) - F(t; D_i)\left\{\frac{\delta_i I(T_i \geq t)}{F(T_i; D_i)} + 1 - \frac{1}{F(max(T_i, t); D_i)}\right\}. \quad (2.2)$$

The above theoretic result does not require the kernel function to have any special shape. But numerically, because the kernel function appears in the denominator of the proposed estimator, standard kernel functions, such as Gaussian kernel with fixed standard deviation and Triangular kernel, can produce extremely small kernel weights and thus cause unstable results. Therefore, to ensure computational stability, we suggest using the following modified Silverman kernel [19], which is flatter and less likely to produce extremely small kernel weights,

$$K(u) = \frac{|\frac{1}{2}e^{\frac{-|u|}{\sqrt{2}}}\sin(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4})|}{\int_{-\infty}^{\infty}|\frac{1}{2}e^{\frac{-|u|}{\sqrt{2}}}\sin(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4})|du}.$$

For the bandwidth, we suggest using $\hat{\sigma}n^{-1/3}$, where $\hat{\sigma}^2$ is the sample variance of the DL. This choice satisfies the conditions that $nh \to \infty$ and $nh^4 \to 0$ as $n \to \infty$. Based on the formula in (2.2), the variance of the KRKM estimator can be estimated by $n^{-2}\sum_{i=1}^{n}\hat{\xi}_i^2(t)$, where $\hat{\xi}_i(t)$ is obtained by replacing $F(\cdot; D_i)$ and $F(\cdot)$ by $\hat{F}(\cdot; D_i)$ and $\hat{F}(\cdot)$. The log-log transformed 95% confidence intervals for $F(t)$ can then be calculated as that for the survival function in survival analysis. This will be referred to as formula-based variance estimation method. Another approach to estimate the variance is to use the bootstrap method. Similar log-log transformed 95% confidence intervals can be obtained. This approach will be referred to as bootstrap-based variance estimation method. The formula-based variance estimation method is computationally faster than the bootstrap-based method, but may underestimate the variance and thus yield poor coverage probabilities at the points below which there are few observations, as shown in simulation studies of Section 2.4.

2.4 Simulation studies

To assess the performance of the proposed KRKM estimator under the situation that the exposure level and DL are associated, we mimicked the cadmium (Cd) and nickel (Ni) data in Appalachian cases from the colon cancer study in Section 4. We generated the DL for each trace element based on their empirical distributions in the data and the exposure level for each trace element from the lognormal regression model: $\log(\widetilde{T}) = \mu + \beta \log(D) + \sigma\varepsilon$, where ε follows a standard normal distribution. The parameters μ, β, σ are estimated based on the data for each trace element, which are -3.05, 0.42, and 1.21 for Cd (setting 1) and 0.16, 0.34, and 1.62 for Ni (setting 2). The non-detect rate of the simulated data is 76% and 25% for the above two settings, respectively. We compared the KRKM estimator, with both bootstrap-based and formula-based variance estimation, to the RKM estimator and the parametric estimator assuming a lognormal distribution for the exposure level. The latter two estimators were obtained from NADA R package [20]. Table 2.1 summarizes the results for the above three estimators of $F(t)$ at t = 1st, 2nd and 3rd quartiles based on 1000 replicates and 500 bootstraps for both settings. The proposed KRKM estimator with the bootstrap-based variance estimation performs very well except for t = 1st quartile in setting 1: the biases are small and the confidence intervals have proper coverage probabilities. At t = 1st quartile in setting 1, the coverage probability is lower than the nominal value due to the very high non-detect rate of 76%. Compared to the bootstrap-based variance estimation, the formula-based variance estimation for the KRKM estimator is computationally faster. But at the points below which there are few observations, e.g. t = 1st and 2nd quartiles in setting 1, the formula-based variance estimation tends to underestimate the variance and thus yield poor coverage probabilities. In contrast to the KRKM estimator, the RKM estimator has large biases and poor coverage probabilities, especially when the sample size increases, due to its inability to account for the association between the exposure level and DL. Likewise, the lognormal estimator also has large biases and low coverage probabilities, resulting from not accounting for the association between the exposure level and DL and possibly misspecified exposure distribution. To further unravel the impact of not accounting for the association between the exposure level and DL for the lognormal estimator, we considered additional simulations where the DL for each trace element was generated from a lognormal distribution with parameters estimated from the colon cancer data. Under this scenario, the marginal distribution of the exposure level is guaranteed to follow a lognormal distribution so that the parametric distribution is correctly specified for the lognormal estimator. However, as shown in Table 2.2, the lognormal estimator still yields large biases and poor coverage probabilities.

To compare the performance of the KRKM, RKM and lognormal estimators under the situation that the exposure level and DL are independent, we adopted the above set-up but set $\beta = 0$. The non-detect rate of the simulated data is 78% and 31% for the two settings, respectively. Table 2.3 summarizes the results for the KRKM, RKM and lognormal estimators of $F(t)$ at $t =$ 1st, 2nd and 3rd quartiles based on 1000 replicates and 500 bootstraps. For all the estimators, the biases are very small, the variance estimators are accurate and the confidence intervals have proper coverage probabilities. The KRKM estimator obtains comparable results as the RKM estimator when the exposure level and DL are independent. The lognormal estimator yields slightly smaller variances than the KRKM and RKM estimators, which is expected since the exposure level and DL are independent and the exposure distribution is lognormal under this set-up.

2.5 Example

Kentucky has the nation's highest colon cancer incidence rate [5]. Appalachian Kentucky, which has a unique geology that contains high-quality bituminous coal

Table 2.1: Comparison of simulation results for the KRKM, RKM and lognormal estimators when the exposure level and DL are associated and the DL was generated from the empirical distribution of the colon cancer cases. True, the true CDF value; Bias, the sampling bias; SSE, the sampling standard error; SEE, the sampling mean of the standard error estimator; CP, the coverage probability of the 95% confidence interval. A subscript of B pertains to the bootstrap-based variance estimation and a subscript of F pertains to the formula-based variance estimation. Each entry is based on 1000 replicates and 500 bootstraps.

Setting 1		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.050	.101	.095	.901	.074	.678	.081	.099	.011	.743	.081	.083	.081	.871
	.50	.003	.056	.055	.952	.039	.765	.091	.055	.060	.888	.091	.058	.060	.702
	.75	.002	.038	.036	.946	.030	.938	.048	.034	.036	.843	.048	.036	.036	.700
500	.25	.044	.061	.062	.866	.033	.746	.102	.063	.071	.298	.102	.048	.047	.524
	.50	.003	.036	.035	.948	.025	.801	.096	.035	.037	.752	.096	.035	.036	.346
	.75	.002	.023	.024	.953	.019	.942	.048	.021	.022	.626	.048	.022	.022	.397
1000	.25	.031	.043	.044	.899	.028	.739	.105	.044	.051	.243	.105	.033	.032	.389
	.50	.003	.025	.025	.953	.015	.876	.095	.026	.025	.511	.095	.026	.026	.367
	.75	.001	.019	.018	.953	.015	.946	.049	.016	.015	.607	.049	.016	.015	.327
Setting 2		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.012	.041	.040	.922	.036	.942	.025	.041	.045	.902	.023	.034	.033	.925
	.50	.006	.041	.040	.941	.039	.944	.025	.040	.041	.912	.020	.033	.032	.911
	.75	.002	.032	.032	.939	.030	.951	.018	.031	.032	.926	.008	.024	.025	.929
500	.25	.008	.025	.026	.935	.022	.941	.026	.025	.027	.821	.029	.020	.021	.814
	.50	.005	.027	.025	.944	.026	.945	.024	.024	.025	.796	.023	.020	.020	.783
	.75	.002	.021	.021	.949	.021	.944	.014	.019	.020	.842	.009	.016	.015	.859
1000	.25	.009	.018	.018	.940	.016	.940	.024	.018	.019	.808	.028	.015	.015	.529
	.50	.005	.019	.018	.942	.018	.950	.022	.018	.018	.759	.023	.015	.015	.657
	.75	.002	.016	.015	.945	.016	.945	.016	.014	.014	.816	.010	.014	.015	.821

Table 2.2: Comparison of simulation results for the KRKM, RKM and lognormal estimators when the exposure level and DL are associated and the DL was generated from a lognormal distribution. True, the true CDF value; Bias, the sampling bias; SSE, the sampling standard error; SEE, the sampling mean of the standard error estimator; CP, the coverage probability of the 95% confidence interval. A subscript of B pertains to the bootstrap-based variance estimation and a subscript of F pertains to the formula-based variance estimation. Each entry is based on 1000 replicates and 500 bootstraps.

Setting 1		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.032	.100	.101	.902	.074	.692	.165	.104	.110	.561	.117	.075	.074	.683
	.50	.005	.064	.067	.958	.052	.877	.094	.068	.066	.724	.076	.058	.059	.592
	.75	.004	.035	.033	.947	.028	.933	.039	.036	.034	.701	.054	.032	.031	.663
500	.25	.020	.093	.095	.899	.077	.734	.153	.092	.091	.420	.101	.073	.073	.563
	.50	.005	.045	.045	.943	.036	.925	.089	.046	.047	.399	.093	.038	.039	.521
	.75	.004	.027	.025	.931	.025	.952	.038	.026	.025	.467	.046	.024	.022	.429
1000	.25	.017	.076	.074	.901	.059	.781	.121	.078	.078	.663	.094	.057	.056	.334
	.50	.006	.030	.029	.946	.025	.934	.086	.031	.029	.581	.084	.027	.027	.221
	.75	.003	.016	.016	.954	.014	.946	.040	.015	.015	.396	.046	.014	.013	.196
Setting 2		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.006	.038	.038	.943	.034	.928	.034	.036	.038	.812	.036	.028	.027	.875
	.50	.005	.031	.032	.940	.029	.937	.029	.030	.033	.793	.029	.029	.028	.858
	.75	.004	.020	.021	.947	.019	.952	.023	.023	.023	.871	.020	.019	.020	.802
500	.25	.006	.023	.024	.951	.022	.932	.037	.022	.020	.662	.035	.018	.018	.671
	.50	.005	.020	.020	.945	.019	.951	.024	.021	.021	.759	.030	.017	.019	.821
	.75	.004	.013	.013	.955	.012	.953	.017	.014	.013	.837	.018	.012	.013	.833
1000	.25	.006	.017	.017	.946	.014	.941	.032	.016	.018	.663	.033	.014	.012	.669
	.50	.005	.014	.014	.945	.013	.945	.026	.014	.014	.589	.030	.011	.011	.571
	.75	.004	.012	.012	.943	.012	.947	.016	.012	.013	.743	.017	.010	.011	.605

Table 2.3: Comparison of simulation results for the KRKM, RKM and lognormal estimators when the exposure level and DL are independent and the DL was generated from the empirical distribution of the colon cancer cases. True, the true CDF value; Bias, the sampling bias; SSE, the sampling standard error; SEE, the sampling mean of the standard error estimator; CP, the coverage probability of the 95% confidence interval. A subscript of B pertains to the bootstrap-based variance estimation and a subscript of F pertains to the formula-based variance estimation. Each entry is based on 1000 replicates and 500 bootstraps.

Setting 1		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.005	.040	.039	.939	.025	.897	.005	.039	.044	.938	.000	.035	.035	.944
	.50	.000	.041	.039	.947	.027	.918	.000	.039	.041	.943	-.003	.034	.034	.954
	.75	.008	.032	.032	.942	.028	.940	.008	.031	.033	.937	-.006	.026	.026	.937
500	.25	.004	.024	.028	.960	.017	.926	.003	.024	.034	.958	.001	.021	.022	.960
	.50	.002	.026	.025	.941	.023	.942	.002	.025	.026	.946	-.002	.021	.021	.948
	.75	.007	.021	.021	.946	.018	.945	.008	.020	.025	.941	.001	.017	.016	.931
1000	.25	.004	.017	.018	.945	.015	.943	.002	.017	.018	.957	-.003	.015	.015	.953
	.50	.001	.018	.018	.948	.018	.946	.003	.018	.018	.942	.001	.015	.015	.952
	.75	.008	.015	.015	.951	.014	.944	.007	.015	.014	.936	.006	.012	.012	.937
Setting 2		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.003	.048	.046	.933	.039	.922	.004	.048	.054	.933	-.003	.054	.049	.931
	.50	-.000	.047	.046	.943	.042	.939	-.000	.045	.048	.944	-.003	.050	.046	.941
	.75	.007	.035	.035	.938	.031	.941	.008	.034	.036	.951	.005	.032	.030	.933
500	.25	.005	.028	.030	.946	.024	.944	.003	.028	.031	.948	.001	.030	.030	.955
	.50	-.002	.028	.030	.957	.023	.944	-.001	.028	.029	.945	-.002	.028	.029	.944
	.75	.006	.023	.023	.942	.021	.939	.006	.022	.022	.948	.005	.020	.020	.940
1000	.25	.003	.020	.021	.954	.016	.943	.001	.018	.020	.951	.002	.021	.021	.942
	.50	-.000	.022	.021	.941	.022	.953	-.004	.018	.020	.949	-.002	.021	.020	.941
	.75	.008	.017	.017	.939	.015	.945	.007	.015	.016	.938	.006	.015	.014	.934

naturally rich in trace elements, has an even higher rate of colon cancer compared to other regions of the state. A case-control study was conducted to explore the association between environmental exposures to trace elements such as arsenic (As), chromium (Cr) and nickel (Ni) and colon cancer and whether exposures to these trace elements contribute to the elevated colon cancer rate in Appalachian Kentucky [2, 6]. For this purpose, 274 colon cancer cases and 253 controls were selected from 23 contiguous rural counties in Kentucky (Appalachian region) and Jefferson County, the largest, most urban county in Kentucky (non-Appalachian region). Among 247 subjects from the Appalachian region, 145 were cases and 102 were controls; among 280 from the non-Appalachian region, 129 were cases and 151 were controls. Toenail samples from these subjects were collected, and the concentrations of 12 trace elements were measured as markers of long-term environmental exposures to these trace elements. The DL varies from one subject to another for these trace element concentrations as a function of the toenail mass. For illustration purposes, we only focus on the Appalachian region. The proportion below the DL is over 20% for most trace elements and is as high as 79% and 83% for Cd in Appalachian cases and controls, respectively (Table 2.4).

We first examine the independence assumption between the exposure level and DL for each trace element using the following three methods. In the first method, we fitted a lognormal accelerated failure time (AFT) model [21] with the left-censored exposure level as the outcome and the log-transformed DL as a covariate. Under this model, the independence assumption between the exposure level and DL was examined by testing whether the coefficent is equal to 0 and the Pearson's correlation coefficient between the exposure level and DL (both log-transformed) was estimated by $\hat{\beta}/\sqrt{\hat{\beta}^2 + \hat{\sigma}^2/\hat{\sigma}_1^2}$, where $\hat{\beta}$, $\hat{\sigma}$ are the estimators of the coefficient and scale parameters in the lognormal AFT model and σ_1^2 is the sample variance of $\log(D)$. In the second method, the Pearson's correlation coefficient between the exposure level and

DL (both log-transformed) and the corresponding p-value were calculated based on the “clikcorr” R package, which assumes a bivariate normal distribution for the two variables and uses a profile likelihood method [22]. In the third method, the nonparametric Kendall’s tau correlation coefficient [23] and the corresponding p-value were calculated based on the “cenken” function in the NADA R package [20]. The results based on the above three methods are reported in Table 2.4. The results from the first two parametric methods are very close for all trace elements except for Cd in controls, where the non-detect rate is as high as 83%. For colon cancer cases, there is a statistically significant association between the exposure level and DL for all 12 trace elements based on the two parametric methods. The nonparametric Kendall’s tau method, which appears more conservative, identifies 6 trace elements with a significant association between the exposure level and DL. For controls, there is only one trace element showing a significant association between the exposure level and DL based on the three methods.

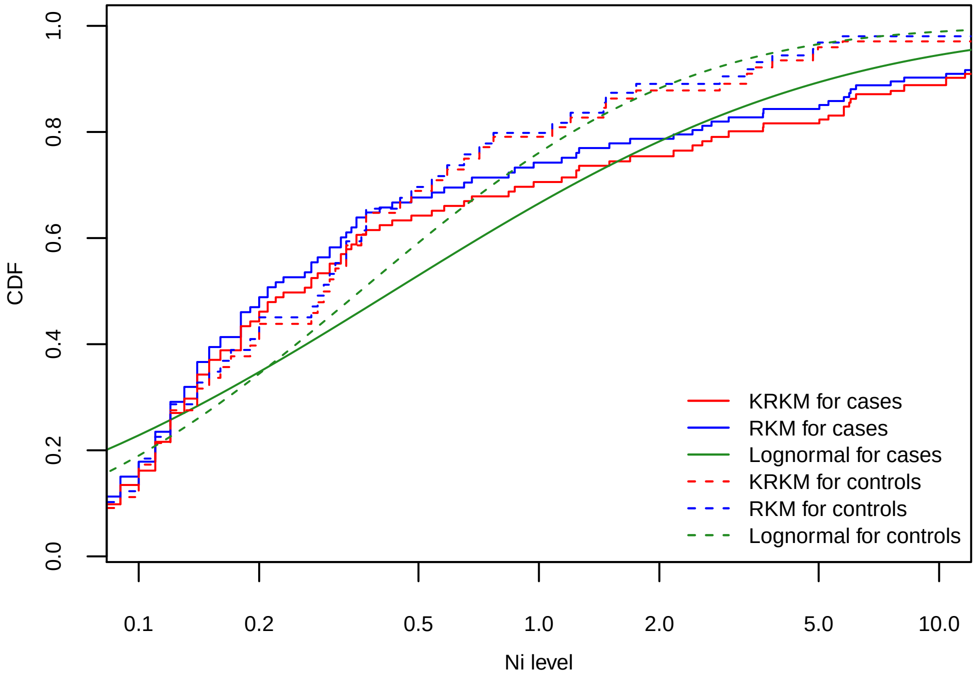
We then use the trace element Ni to demonstrate our proposed KRKM estimator, comparing to the RKM estimator and the parametric estimator. For cases, the Ni level ranges from 0.02 to 624.4 and the DL ranges from 0.004 to 24.84; for controls, the Ni level ranges from 0.04 to 39.37 and the DL ranges from 0.01 to 38.38. Table 2.4 shows that for Ni there is a signifcant association between the exposure level and DL for cases but no signficant association for controls. We estimated the exposure distributions of Ni level for cases and controls, respectively. The lognormal distribution was selected for the distributions of Ni for both cases and controls by the Akaike information criterion (AIC) [24] among a number of candidate distributions, including normal, lognormal, Weibull and loglogistic. Figure 2.1 displays the CDF estimates for colon cancer cases and controls based on the KRKM, RKM and lognormal estimators, and Figure 2.2 displays the differences in CDF estimates between the KRKM estimator and the latter two estimators along with 95% confidence limits. These figures

Table 2.4: The non-detect rate, correlation coefficient between the exposure level and DL and the corresponding p value

	Non-detect rate(%)											
	Ni	Cd	As	Cr	Pb	Co	Al	Mn	Fe	Cu	Zn	Se
case	23	79	43	13	7	73	7	51	21	0	0	3
control	48	83	45	42	33	79	14	59	37	4	3	8
	Correlation coefficient and p value for colon cancer cases											
	Ni	Cd	As	Cr	Pb	Co	Al	Mn	Fe	Cu	Zn	Se
AFT_cor	.513	.369	.29	.638	.436	.538	.562	.332	.446	.320	.202	.174
AFT_p	<.001	.010	.003	<.001	<.001	<.001	<.001	.005	<.001	<.001	.013	.034
Clik_cor	.512	.368	.289	.636	.435	.537	.561	.331	.445	.319	.201	.174
Clik_p	<.001	.035	.007	<.001	<.001	<.001	<.001	.012	<.001	<.001	.014	.036
Ken_cor	.134	.009	.077	.365	.254	.070	.344	.07	.222	.110	.090	-.066
Ken_p	.016	.863	.166	<.001	<.001	.205	<.001	.207	<.001	.050	.108	.241
	Correlation coefficient and p value for controls											
	Ni	Cd	As	Cr	Pb	Co	Al	Mn	Fe	Cu	Zn	Se
AFT_cor	.081	-.254	.375	.092	.105	.102	.106	-.199	.141	.097	-.098	-.055
AFT_p	.636	.350	.001	.466	.391	.629	.311	.360	.320	.337	.330	.614
Clik_cor	.068	-.493	.346	.092	.104	.102	.105	-.198	.140	.096	-.097	-.054
Clik_p	.683	.055	.005	.477	.405	.646	.317	.334	.334	.340	.330	.612
Ken_cor	-.032	-.032	.173	.042	.069	-.006	.047	-.030	.039	.046	.003	.024
Ken_p	.633	.629	.009	.533	.302	.932	.481	.650	.561	.491	.968	.717

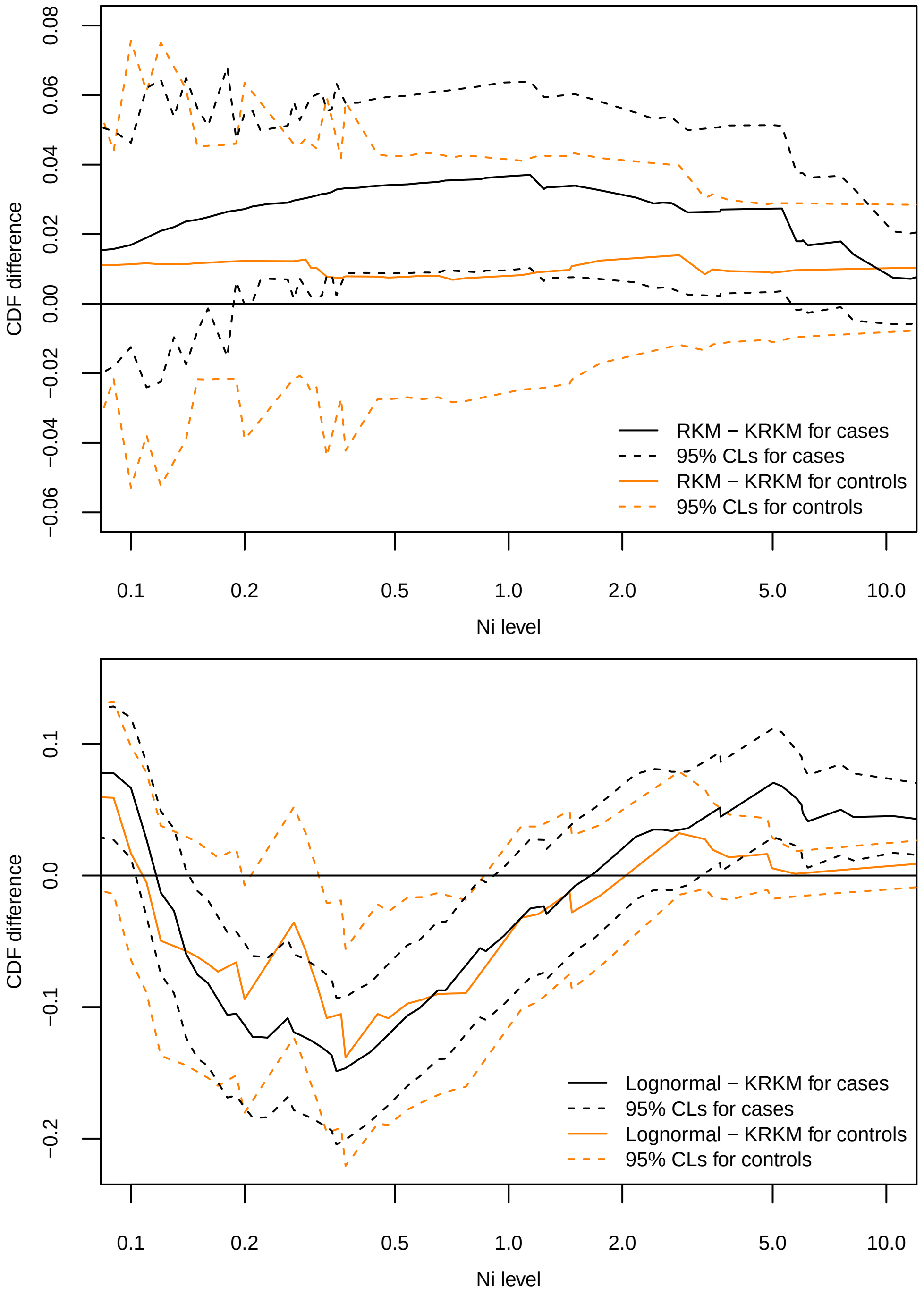
Note: AFT_cor and Clik_cor are the estimates of the Pearson's correlation coefficient between the exposure level and DL (both log-transformed) based on the lognormal AFT model method and the "clikcorr" R package, respectively. Ken_cor is the Kendall's tau correlation coefficient estimate. AFT_p, Clik_p and Ken_p are the corresponding p-values.

Figure 2.1: CDF estimates of Ni exposure distribution for colon cancer cases and controls in the Appalachian region based on the KRKM, RKM, and lognormal estimators. Solid curves pertain to cases and dotted curves pertain to controls. The red curves represent the KRKM estimator, the blue curves represent the RKM estimator, and the green curves represent the lognormal estimator.



show that for cancer casesthe RKM estimator significantly overestimates the CDF for the Ni levels between 0.21 and 5.29 compared to the proposed KRKM estimator. This may be because of the significant association between the exposure level and DL. In contrast, there is no significant difference between the two estimators for controls, which may be because of the insignificant association between the exposure level and DL. As a result, the RKM estimator significantly underestimates the difference between the cases and controls compared to the KRKM estimator. Figures 2.1 and 2.2 also show remarkable difference between the lognormal and KRKM estimators for cases, most likely due to the significant association between the exposure level and DL. The difference between these two estimators is smaller for controls.

Figure 2.2: Differences in CDF estimates between the RKM and KRKM estimators (upper panel) and between the lognormal and KRKM estimators (lower panel). The solid curves are for the point estimates of differences , and the dotted curves are the corresponding 95% bootstrapped confidence limits (CLs). The black curves pertain to the cases and the orange ones petain to the controls.



2.6 Discussion

We have developed a consistent nonparametric estimator for the exposure distribution without requiring any independence assumption between the exposure level and DL. Our proposed estimator outperforms the RKM estimator and the parametric estimator when the exposure level and DL are associated because the latter two estimators are not consistent in that situation. In the case of a common DL, our estimator reduces to the RKM estimator; and in the case of varying DLs but the exposure level and DL are independent, our estimator can obtain comparable results as the RKM estimator. Thus, our estimator provides a unified nonparametric approach for estimating the exposure distribution regardless whether the exposure level and DL are independent or not and whether the association between the exposure level and DL are linear, curvilinear, or step function, etc. Therefore, the user does not have to test whether the exposure level and DL are associated before using our method, which is an advantage over the RKM method whose validity depends on the test results.

We have utilized a two-step strategy and kernel smoothing technique along with a special feature of data subject to DLs, i.e. the DL is observable for each subject, to completely eliminate the independence assumption between the exposure level and DL. In contrast, the consistent estimators developed based on similar two-step strategies for the marginal survival function for right-censored survival data need to find a set of covariates and require the independence assumption between the censoring time and survival time conditional on those covariates [25, 26]. In our approach, we take advantage of the data characteristic that the DL is observable for each subject and utilize the DL as the conditioning covariate. As a statistical fact, the independence assumption between the DL and exposure level given the DL automatically holds. Therefore, our estimator is free of any independence assumption between the exposure level and DL.

In survival analysis, another approach dealing with dependent censoring for estimating the survival function is the inverse probability of censoring weighting (IPCW) KM estimator [27, 28]. This weighted version of the KM estimator assigns a weight, inversely proportional to an estimate of the conditinal survival function of the censoring time given a set of covariates, to each subject. Under the condition that the censoring time and survival time are independent given that set of covariates, the IPCW KM estimator is consistent. By borrowing this idea, one can construct an IPCW RKM estimator for the exposure distribution by adding subject-specific weights, proportional to each subject's conditional CDF of the DL given a set of covariates, in the RKM estimator. The consistency of this estimator requires that the exposure level and DL are independent given that set of covariates. To obtain an IPCW RKM estimator not requiring any independence assumption between the exposure level and DL as the proposed KRKM estimator, we need to choose DL as the covariate. However, the conditional CDF of the DL given DL can only take values 0 or 1 and thus cannot be used as an inverse weight. Therefore, the IPCW KM estimator cannot be extended to the data subject to DLs without imposing certain conditional independence assumptions between the exposure level and DL.

A key issue in our two-step strategy is how to estimate the conditional CDF of the exposure level given the DL for the data subject to DL . To address this issue, we have added kernel weights into the RKM estimator. The use of the kernel technique assures our estimator is purely nonparametric and free of any distributional assumption. Importantly, our estimator does not suffer the curse of dimensionality of the kernel method because we only need to condition on a one-dimensional variable, i.e. the DL, for estimating the conditional CDF. In addition, our estimator is robust to the choice of bandwidth. Besides the bandwidth of $\hat{\sigma}n^{-1/3}$ presented in the paper, we also conducted simulation studies using several other bandwidths including $\hat{\sigma}n^{-7/24}$, $\hat{\sigma}n^{-2/5}$, and $\hat{\sigma}n^{-1/2}$, which yielded very similar results (data not shown). As an alternative to the kernel method, one can use a parametric AFT model with the DL as a covariate to estimate the conditional CDF. Additional simulation studies reveal that this alternative method performs well and has smaller variance than the proposed estimator when the model is correctly specified but can lead biased results when the model is misspecified (data not shown).

In this chapter, we highlight the critical need to account for the association between the exposure and DL and the consequences of ignoring it. This problem of association between the exposure and DL may sometimes be alleviated by improving the design of sample collection. For example, samples can be collected from multiple toes or at multiple time points if time and resources allow. Such strategies can increase the toenail mass, lowers the DL and thus possibly reduces the association. However, the obtained toenail mass may still be low for some subjects due to slow toenail growth or noncompliant toenail cutting. It is therefore difficult to eliminate the association problem. In presence of varying DLs, appropriate statistical methods should be used to deal with the possible association between the exposure level and DL so that unbiased analysis results can be obtained.

There are at least two extensions of the proposed method. First, the proposed KRKM estimator requires the data come from a simple random sample of the underlying population. One can extend the proposed estimator to survey data by incorporating sampling weights. Second, our estimator can serve as the building block for a formal test to compare the exposure distributions between two groups by considering the cumulative weighted difference in CDF estimates for the two groups, analogous to the weighted KM statistics for right-censored data [16]. However, it will be more complex than the latter because the proposed KRKM estimator is more complicated than the KM estimator and does not have a martingale representation like the KM estimator. Of further interest is to incorporate the adjustment of confounding factors in the comparison between two groups. Current literature [29, 30] considered logistic regression models with exposure(s) and confounding factors as covariates and the disease status as the outcome and used the maximum likelihood method to make inferences. However, these methods require the independence assumption between the exposure level and DL. One possible approach to account for the association between the exposure level and DL is to use multiple imputation to impute exposure values below DLs based on our kernel-smoothed conditional CDF given the DL. Since our kernel-smoothed conditional CDF is undefined in $(0, \tau_n)$ when the smallest observation is censored, additional distributional assumptions are needed for that region in order to perform the imputation under this situation.

APPENDIX A.

Weak convergence of $\sqrt{n}\{\hat{F}(t) - F(t)\}$

In this section, we prove the weak convergence of $\sqrt{n}\{\hat{F}(t) - F(t)\}$ through the modern empirical process theory. Let P_n and P denote the empirical measure and the distribution under the true model, respectively. For a measurable function f and measure Q, the integral $\int f dQ$ is abbreviated as Qf. Specifically, $P_n f(T, \delta, D) = n^{-1} \sum_{i=1}^{n} f(T_i, \delta_i, D_i)$, $P\{f(T, \delta, D)$ is the expectation of $f(T, \delta, D)$, and $P\{f(T, \delta, D)|D\}$ is the conditional expectation of $f(T, \delta, D)$ given D. We express $\sqrt{n}\{\hat{F}(t) - F(t)\}$ as

$$\sqrt{n}(P_n - P)\{F(t; D)\} + \sqrt{n}P\{\hat{F}(t; D) - F(t; D)\} + \sqrt{n}(P_n - P)\{\hat{F}(t; D) - F(t; D)\}. \tag{2.3}$$

To study the second term in (2.3), we define

$$R(t; d) \quad = \quad \int_t^{\infty} \frac{dF(u; d)}{F(u; d)}.$$

By some algebras we obtain $R(t; d) = -\log F(t; d)$, which is analogous to the conditional cumulative hazard function in survival analysis but with the conditional survival function replaced by the conditional CDF. We first study

$$\hat{R}(t;d) \quad = \quad \int_t^\infty \frac{\sum_{j=1}^n K\{(D_j-d)/h\}dN_j(s)}{\sum_{j=1}^n K\{(D_j-d)/h\}Y_j(s)}.$$

Let $N(t) = I(T \le t, \delta = 1)$ and $Y(t) = I(T \le t)$. We express $\hat{R}(t;d) - R(t;d)$ as

$$\begin{aligned}
&P_n\left[\frac{K\{(D-d)/\}\delta I(T\ge t)}{P_n[K\{(D-d)/h\}Y(u)]\mid_{u=T}}\right] - P\left[\frac{I(\tilde{T}\ge t)}{P\{Y(u)\mid D=d\}\mid_{u=\tilde{T}}}\middle| D=d\right] \\
&= (P_n-P)\left[\frac{K\{(D-d)/h\}\delta I(T\ge t)}{P_n(K\{(D-d)/h\}Y(u)\mid_{u=T})}\right] \\
&\quad - P\left[\frac{K\{(D-d)/h\}\delta I(T\ge t)(P_n-P)(K\{(D-d)/h\}Y(u)\mid_{u=T})}{P(K\{(D-d)/h\}Y(u)\mid_{u=T})P_n(K\{(D-d)/h\}Y(u)\mid_{u=T})}\right] \\
&\quad + \left(P\left[\frac{K\{(D-d)/h\}\delta I(T\ge t)}{P[K\{(D-d)/h\}Y(u)\mid_{u=T}]}\right] - P\left[\frac{I(\tilde{T}\ge t)}{P\{Y(u)\mid D=d\}\mid_{u=\tilde{T}}}\middle| D=d\right]\right) \\
&= (P_n-P)\left[\frac{K\{(D-d)/h\}\delta I(T\ge t)}{P(K\{(D-d)/h\}Y(u)\mid_{u=T})}\right] \\
&\quad - P\left(\frac{K\{(D-d)/h\}\delta I(T\ge t)(P_n-P)[K\{(D-d)/h\}Y(u)\mid_{u=T}]}{P^2[K\{(D-d)/h\}Y(u)\mid_{u=T}]}\right) \\
&\quad + \left(P\left[\frac{K\{(D-d)/h\}\delta I(T\ge t)}{P(K\{(D-d)/h\}Y(u)\mid_{u=T})}\right] - P\left[\frac{I(\tilde{T}\ge t)}{P\{Y(u)\mid D=d\}\mid_{u=\tilde{T}}}\middle| D=d\right]\right) + o_p(n^{-1/2}).]
\end{aligned} \tag{2.4}$$

It's straightforward to show that the first term on the right side of (2.4) is equal to

$$(P_n - P)\int_t^\infty \frac{[K\{(D-d)/h\}dN(u)]}{P[K\{(D-d)/h\}Y(u)]}.$$

By Lemma 1 and some algebras, the second term on the right side of (2.4) is equal to

$$(P_n - P) \int_t^\infty \frac{K\{(D-d)/h\}Y(u)dR(u;d)}{P[K\{(D-d)/h\}Y(u)]} + O(h^2).$$

By Lemma 1 and the statistical fact that $\tilde{T}$ and D is independent given D, the third term on the right side of (2.4) can be shown to be $O(h^2)$. Therefore, we obtain that $\hat{R}(t;d) - R(t;d)$ is equal to

$$(P_n - P)\left(K\{(D-d)/h\} \int_t^\infty \frac{dN(u) + Y(u)dR(u;d)}{P[K\{(D-d)/h\}Y(u)]}\right) + O(h^2) + o_p(n^{-1/2}).$$

By the condition that $\sqrt{n}h^2 = o_p(1)$, the Duhamel equation and Lemma 1, we obtain that the second term on the right side of (2.3) is asymptotically equivalent to

$$\sqrt{n}(P_n - P)\left(P_{D^*}\left[-F(t;d)K\{(D-d)/h\} \int_t^\infty \frac{dN(u) + Y(u)dR(u;d)}{P[K\{(D-d)/h\}Y(u)]}\right]\Bigg|_{d=D^*}\right)$$

$$=\sqrt{n}(P_n - P)\left[-F(t;D) \int_t^\infty \frac{dN(u) + Y(u)dR(u;D)}{P\{Y(u) \mid D\}}\right] + o_p(1),$$

where D^* is a random variable with the same distribution as D, and P_{D^*} denotes expectation only respective to D^*.

Similarly, we can verify that $P\{\hat{F}(t;D) - F(t;D)\}^2 \longrightarrow_p 0$ uniformly for $t \in [0, \infty]$ and that $\hat{F}(t;D), F(t;D)$ belong to a P- Donsker class. It then follows that the third term of (2.3) converges uniformly to zero in probability by Lemma 19.24 of[31].

Combining the aforementioned results, we conclude that $\sqrt{n}(\hat{F}(t) - F(t))$ is asymptotically equivalent to the process

$$\sqrt{n}(P_n - P)\left\{F(t;D) - F(t;D) \int_t^\infty \frac{dN(u) - Y(u)d\log F(u;D)}{F(u;D)I(D \le u)}\right\}$$

$$=n^{-1/2} \sum_{i=1}^n \left[F(t;D_i) - F(t) - F(t;D_i)\left\{\frac{\delta_i I(T_i \ge t)}{F(T_i;D_i)} + 1 - \frac{1}{F(max(T_i,t);D_i)}\right\}\right].$$

Lemma 1. Let $f_D(d)$ be the probability density function of D, then

$$P[h^{-1}K\{(D-d)/h\}\delta I(T \geq t)] = P\{\delta I(T \geq t) \mid D = d\}f_D(d) + O(h^2)$$

$$P[h^{-1}K\{(D-d)/h\}Y(u)] = P\{Y(u) \mid D = d\}f_D(d) + O(h^2)$$

Proof: We have

$$P[h^{-1}K\{(D-d)/h\}\delta I(T \geq t)] = \int h^{-1}K\{(x-d)/h\}P[\delta I(T \geq t) \mid D = x]f_D(x)dx. \tag{2.5}$$

Let $g(x) = P[\delta I(T \geq t) \mid D = x]f_D(x)$. Using a simple transformation $s = (x-d)/h$ and the Taylor expansion of $g(d+sh)$ at d, we obtain the right side of (2.5) is equal to

$$\int K(s)g(d)ds + \int sK(s)g^{'}(d)ds + O(h^2). \tag{2.6}$$

Because $\int K(s)ds = 1$ and $\int sK(s)ds = 0$, we then obtain the first equation. Similarly, we can obtain the second equation.

Chapter 3 Comparison of Exposure Distributions Between Two Groups

3.1 Introduction

In environmental exposure studies, comparing two groups is a basic design: whether the distributions of environmental chemicals, such as heavy metals and pesticides, vary between treatment and control groups is of particular interest. However, it is very common to observe a portion of exposure measurements to fall below experimentally determined detection limits (DLs). A detection limit (DL) is "a threshold below which measured values are not considered significantly different from a blank signal, at a specified level of probability" [1]. Therefore, the exposure level of a chemical for a sample is only reported when its value is not less than the DL and otherwise is reported as a less than value or non-detect. Due to the this problem, the standard two-sample t test fails and several methods have been developed in past decades. When there is only one reported DL, Mann-Whitney test can be directly applied to the data with DLs, i.e. all values below the DL are considered tied. It will efficiently capture the information of the data, including the proportion of non-detects[32]. Zhang et al. performed a large set of simulations comparing 14 methods when exposure measurements below a common fixed DL[10]. In other cases, the laboratory may report different DLs for different samples. Parametric methods, such as the Tobit model [33], work for data with multiple DLs. The caution of these methods is that the validity of their results depends on choosing the correct distribution. Nonparametric methods are widely used for DLs data since they do not require an assumption that data follow a specific distribution. The log-rank test[34] and the Peto-Peto modification of the Gehan-Wilcoxon test[14] are the most common two in the right-censored survival data and can be applied to left-censored DLs data with the scale reversed. The later one will be referred to as Peto-Peto test. In additional, the Peto-Peto test is the most appropriate test for left-censored log-normal data[1]. However, there are still several limitations for this approach. Firstly, the test statistics based on these tests essentially estimate integrated weight difference in hazard function. Left-censored DLs data lack meaningful environmental interpretation since there is no concept corresponding to the hazard function. Secondly, though these tests are sensitive to alternatives of ordered hazard function, they are not to alternatives of ordered CDFs[16], such as the absolute difference between CDFs. Thirdly, the asymptotically efficiency of these tests depends on the assumption that the distributions of DLs in two groups are identical[14, 15]. If the distributions of DLs in the two groups are heterogeneous, type I error rate is inflated. The heterogeneity of DLs commonly occurs in environmental exposure studies. For example, in the colon cancer study measuring heavy metal accumulation in toenails [2], DLs distributions depended on cases and controls for 10 heavy metals($p \leq 0.003$). Therefore, aforementioned tests are not appropriate in this case.

In previous work , we proposed a Kernel reverse Kaplan-Meier(KRKM) estimator for the exposure distribution without imposing any independence assumption between the exposure level and DL. In this chapter, we develop a class of test statistics to compare exposure distributions between two groups by using the integrated weighted difference in the KRKM estimator. In section 3.2, we propose the class of statistics and study the condition of weight function to satisfy asymptotic normality. In Section 3.3, the results of several simulation studies are reported to demonstrate the performance of the test statistics. In Section 3.4, a colon cancer study is provided for illustration. Finally, Section 3.5 contains discussions and some conclusions.

3.2 Methods

Kernel reverse Kaplan-Meier (KRKM) estimator

Let $\widetilde{T}$ and D be random variables for the exposure level and DL, respectively, and $F(\cdot)$ be the cumulative distribution function (CDF) of the exposure level. Let $T = max(\widetilde{T}, D)$ and $\delta = I(\widetilde{T} \geq D)$. Here δ indicates whether T is an exposure level value or a DL value. For data subject to DL, only (T, δ, D) are observable for each subject. Suppose the data consist of n replicates $\{(T_i, \delta_i, D_i)\colon\ i = 1, \cdots, n\}$. It is useful to adopt the counting process notation. Analogous to the observed counting process and at-risk process for right censored survival data, we define two counting processes, $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(T_i \leq t)$, for the data subject to DLs.

In the previous chapter, we utilized a two-step strategy and the kernel smoothing technique to develop a nonparametric consistent estimator for the exposure distribution. In the first step, we obtained a consistent estimator for the conditional CDF of the exposure level,denoted by $\hat{F}(t; d)$,i.e.

$$\hat{F}(t; d) = \prod_{s>t} \left[1 - \frac{\sum_{j=1}^{n} K\{(D_j - d)/h\} dN_j(s)}{\sum_{j=1}^{n} K\{(D_j - d)/h\} Y_j(s)} \right],$$

where $K(\cdot)$ is a kernel function, and h is a bandwidth such that $nh \to \infty$ and $nh^2 \to 0$ as $n \to \infty$. To ensure computational stability, a modified Silverman kernel [19] is suggested, which is flatter and less likely to produce extremely small kernel weights,

$$K(u) = \frac{|\frac{1}{2} e^{\frac{-|u|}{\sqrt{2}}} \sin(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4})|}{\int_{-\infty}^{\infty} |\frac{1}{2} e^{\frac{-|u|}{\sqrt{2}}} \sin(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4})| du}.$$

For the bandwidth, $\hat{\sigma} n^{-1/3}$ is suggested, where $\hat{\sigma}^2$ is the sample variance of the DL. In the second step, we estimated $F(t)$ by the average of estimated conditional CDF values over all DL values in the sample, i.e. $\hat{F}(t) = n^{-1} \sum_{i=1}^{n} \hat{F}(t; D_i)$. $\hat{F}(t)$ is the kernel reverse Kaplan-Meier (KRKM).

Weighted kernel reverse Kaplan-Meier statistics

To develop a test statistic comparing the CDFs of the exposure level between two groups, we consider a cumulative weighted difference in the CDF estimates for the two groups. Suppose n_1 and n_2 are the sample sizes in the two groups, $n = n_1 + n_2$ and $\hat{F}_1(t)$ and $\hat{F}_2(t)$ are the CDF estimates for the exposure level in the two groups obtained by the KRKM estimators. We propose the following class of test statistics

$$U = \sqrt{\frac{n_1 n_2}{n}} \int_0^\infty \hat{w}(t)\{\hat{F}_1(t) - \hat{F}_2(t)\}dt,$$

where $\hat{w}(\cdot)$ is a random weight function that estimates a deterministic function $w(\cdot)$. The statistics U will be referred to as a weighted kernel reverse Kaplan-Meier (WKRKM) statistic. Because the variability of the difference $\hat{F}_1(t) - \hat{F}_2(t)$ is large at values close to 0 where the probability of below DL is large, it is critical to choose an appropriate weight function to down-weigh the difference at these values in the integrand so that the corresponding test statistic has finite asymptotic variance. We will first study the conditions under which the weight function ensures the stability of the proposed WKRKM statistic and then discuss the choice of the weight function later.

The proposed class of statistics is based directly on the difference between two CDFs, so it will be sensitive to the alternative hypothesis of ordered CDFs and the absolute difference between two CDFs. In contrast, the log-rank test essentially estimates integrated weighted differences in hazard functions and thus is not necessarily sensitive to the alternative hypothesis of ordered CDFs [16]. In addition, by plugging in the KRKM estimator proposed in Chapter 1, the proposed statistics can handle the correlation between the exposure level and DL. The idea of using the integrated weighted difference in distribution estimates to compare the distributions between two groups was first proposed by [16, 35]. However, it was for right-censored survival data and only considered the simple KM estimator and thus cannot deal with dependent censoring.

Choice of the weight function

To study the conditions for the weight function such that the corresponding WKRKM statistic is stable, we need to study the asymptotic variance of the WKRKM statistic. To this end, we first study the asymptotic distribution of $\sqrt{n}\int_0^\infty w(t)[\hat{F}(t) - F(t)]dt$. As shown in the Appendix,

$$\sqrt{n}\int_0^\infty w(t)[\hat{F}(t) - F(t)]dt \xrightarrow{d} N(0, \sigma^2)$$

where

$$\sigma^2 = \int_0^\infty \int_0^\infty w(t)w(s)\left[\int_0^\infty \{F(t \mid u) - F(t)\}\{F(s \mid u) - F(s)\}dG(u)\right] dtds$$

$$+ \int_0^\infty w^2(t)\left[\int_0^\infty \left\{F^2(t \mid s)[1 - \frac{1}{F(max(t, s) \mid s)}]\right\} dG(s)\right] dt,$$

and $G(\cdot)$ is the CDF of the DL.

To ensure that σ^2 is finite for all choices of the unknown underlying exposure and DL distributions, a sufficient and almost necessary condition to be satisfied by $w(\cdot)$ is that $w^2(t)/G(t)$ should be bounded uniformly in t on $[0, \infty)$. We can replace the deterministic weight function $w(t)$ by a random weight function $\hat{w}(t)$ if

$$\sup_{t\in[0,\infty)} \frac{\hat{w}(t) - w(t)}{G(t)^{1/2}} \xrightarrow{p} 0$$

$$|w(t)| \leq \mathit{\Gamma} G(t)^{1/2+\delta}$$

$$|\hat{w}(t)| \leq \mathit{\Gamma} \hat{G}(t)^{1/2+\delta}$$

for some $\Gamma \geq 0$ and $\delta \geq 0$. Then

$$\sqrt{n}\int_0^\infty \hat{w}(t)\{\hat{F}(t) - F(t)\}dt \xrightarrow{d} N(0, \sigma^2).$$

Assume that the above conditions are satisfied for both groups. Then under the null hypothesis $H_0 : F_1(t) = F_2(t)$, we have

$$\sqrt{\frac{n_1 n_2}{n}}\int_0^\infty \hat{w}(t)\{[\hat{F_1}(t) - \hat{F_2}(t)]\}dt \xrightarrow{d} N(0, \sigma^2_{WKRKM}).$$

where

$$\sigma^2_{WKRKM} =$$

$$p_2 \int_0^\infty \int_0^\infty w(t)w(s)\left[\int_0^\infty \{F_1(t \mid u) - F_1(t)\}\{F_1(s \mid u) - F_1(s)\}dG_1(u)\right] dtds$$

$$+ \int_0^\infty w^2(t)\left[\int_0^\infty \left\{F_1^2(t \mid s)[1 - \frac{1}{F_1(max(t,s) \mid s)}]\right\} dG_1(s)\right] dt$$

$$+ p_1 \int_0^\infty \int_0^\infty w(t)w(s)\left[\int_0^\infty \{F_2(t \mid u) - F_2(t)\}\{F_2(s \mid u) - F_2(s)\}dG_2(u)\right] dtds$$

$$+ \int_0^\infty w^2(t)\left[\int_0^\infty \left\{F_2^2(t \mid s)[1 - \frac{1}{F_2(max(t,s) \mid s)}]\right\} dG_2(s)\right] dt,$$

and p_1 and p_2 are the limits of n_1/n and n_2/n, respectively.

Integrated difference is sensitive to the range of exposure level. To ensure the stability, we may exclude the outliers in the exposure level. Therefore, one choice of the weight function can be

$$\hat{w}(t) = \begin{cases} \frac{\hat{G_1}(t)\hat{G_2}(t)}{(n_1/n)\hat{G_1}(t) + (n_2/n)\hat{G_2}(t)}, & t \leq q \\ 0, & t > q \end{cases},$$

where $\hat{G_1}(\cdot)$ and $\hat{G_2}(\cdot)$ are the CDF estimates of the DL in these two groups, $q = Q_3 + 3(Q_3 - Q_1)$, Q_1 and Q_3 are the lower and upper quartiles of exposure in two groups combined respectively [36]. Any exposure level greater than q is considered to be an outlier. This weight function is akin to a geometric average of two CDF estimates of the DL in the two groups, and satisfies the conditions of the weight function. With this weight function the corresponding WKRKM statistic becomes the difference in means when there are no non-detects. Therefore, the WKRKM statistics can be regarded as a generalization of the two-sample z-test to data subject to DLs.

3.3 Simulation studies

Size properties

To access the performance under null hypothesis, we analyze size properties in this section. We mimicked exposure levels in colon cancer study through $\hat{F}(\cdot)$, where $\hat{F}(\cdot)$ is the KRKM estimetors of two groups combined. We considered two configurations: the distributions of DLs in both groups are identical or not. When the distributions of DLs are identical, DLs were generated from $\hat{G}(\cdot)$, which is estimated from two groups combined. When the distributions of DLs are different, DLs were generated from $\hat{G}_1(\cdot)$ and $\hat{G}_2(\cdot)$, i.e. the empirical CDFs in cases and controls. We first generated two standard uniform random variables X and Y, then exposure levels of two groups are generated from $\hat{F}^{-1}(X)$, DLs are generated from $\hat{G}^{-1}(Y)$, $\hat{G_1}^{-1}(Y)$ and $\hat{G_2}^{-1}(Y)$ dependent on different configurations. X and Y can be either dependent or independent. Three correlations (0, 0.3, 0.5) are taken into account. We considered the setting that mimics Nickel (Ni) from Appalachian sample under each configuration. Table3.1 summarizes the size simulation results for the WKRKM test, the Log-normal test, the Peto-Peto test and the log-rank test based on 1000 replicates and 500 bootstraps with different correlations. Different correlations may cause different non-detect rates. Correlation with 0, 0.3, 0.5 will cause non-detect rate 39, 36 and 34 per cent. As expected, all the tests are valid when the distributions of DLs are identical. When the distributions of DLs are not identical, the empirical levels of

Table 3.1: Size simulation results for the Log-rank test, the Peto-Peto test, the log-normal test and the WKRKM test.

Identical DLs	Cor	Log-rank	Peto-Peto	Log-normal	WKRKM
$n_1 = n_2 = 50$	0	0.042	0.042	0.036	0.052
	0.3	0.058	0.072	0.082	0.051
	0.5	0.046	0.064	0.072	0.058
$n_1 = n_2 = 100$	0	0.050	0.058	0.064	0.047
	0.3	0.054	0.034	0.050	0.052
	0.5	0.052	0.040	0.054	0.051
$n_1 = n_2 = 150$	0	0.048	0.048	0.054	0.050
	0.3	0.050	0.050	0.056	0.042
	0.5	0.050	0.056	0.048	0.047
Different DLs	Cor	Log-rank	Peto-Peto	Log-normal	WKRKM
$n_1 = n_2 = 50$	0	0.044	0.042	0.046	0.047
	0.3	0.058	0.076	0.072	0.048
	0.5	0.090	0.094	0.080	0.055
$n_1 = n_2 = 100$	0	0.040	0.058	0.066	0.045
	0.3	0.080	0.072	0.066	0.045
	0.5	0.150	0.154	0.106	0.046
$n_1 = n_2 = 150$	0	0.048	0.050	0.066	0.050
	0.3	0.110	0.116	0.074	0.043
	0.5	0.190	0.188	0.114	0.047

Note: Cor denotes the correlation between the exposure level and DL.

the Peto-Peto test, the log-rank test and the log-normal test could be substantially higher than the nominal levels in correlated setting. With the increase of sample size, such elevation gets even worse. In contrast, the empirical levels of the WKRKM test are very close to the nominal levels across both configurations regardless of the distributions of DLs and the correlation between exposure level and DL. We can draw the conclusion that all the aforementioned tests work well when the the distributions of DLs are identical or exposure level and DL are uncorrelated. The WKRKM test is the only valid test when DLs distributions are different and the exposure level and DL are correlated.

Power properties

In this part, we conducted simulation studies of power. We adopt the above setups but generated exposure level from $\hat{F}_1(\cdot)$ and $\hat{F}_2(\cdot)$, i.e. the KRKM estimators of cases and controls. In the simulated setting, the exposure difference of two groups does not manifest itself until larger exposure level as shown in the left panel of figure 3.1. Non-detect rates are 39, 35 and 33 per cent for the correlation with 0, 0.3, 0.5. Table 3.2 summarizes the power simulation results for the WKRKM test, the Log-normal test, the Peto-Peto test and the log-rank test based on 2000 replicates and 500 bootstraps. When the DLs distributions are identical, it indicates that the superior performance of the the WKRKM test over the Peto-Peto test, the log-rank test and the log-normal test regardless the correlation between the exposure level and DL. The WKRKM test, which places more weight at larger exposure level rather than smaller exposure level, is quite well suited to this alternative. When DLs distributions are different and the exposure level and DL are correlated, the power of the WKRKM test achieve the same range compared to other tests even though these tests are most likely to be inflated.

3.4 Example

Kentucky has the nation's highest colon cancer incidence rates [5]. Appalachian Kentucky, which has a unique geology that contains high-quality bituminous coal naturally rich in trace elements, has even higher rates of colon cancer compared to other regions of the state. A case-control study was conducted to explore the association between environmental exposure to trace elements such as arsenic (As), chromium (Cr) and nickel (Ni) and colon cancer and whether exposure to these trace elements contributes to the elevated colon cancer rate in Appalachian Kentucky [2, 6]. For this purpose, 274 colon cancer cases and 253 controls were selected from 23 contiguous rural counties in the Appalachian region of Kentucky and Jefferson County, the

Table 3.2: Power simulation results at significance level $\alpha = 0.05$ for the Log-rank test, the Peto-Peto test, the log-normal test and the WKRKM test.

Identical DLs	Cor	Log-rank	Peto-Peto	Log-normal	WKRKM
$n_1 = n_2 = 50$	0	0.074	0.108	0.216	0.524
	0.3	0.048	0.106	0.234	0.506
	0.5	0.050	0.104	0.260	0.494
$n_1 = n_2 = 100$	0	0.058	0.142	0.360	0.584
	0.3	0.052	0.152	0.430	0.608
	0.5	0.058	0.166	0.486	0.626
$n_1 = n_2 = 150$	0	0.064	0.210	0.530	0.734
	0.3	0.040	0.220	0.576	0.708
	0.5	0.052	0.248	0.638	0.750
Different DLs	Cor	Log-rank	Peto-Peto	Log-normal	WKRKM
$n_1 = n_2 = 50$	0	0.070	0.106	0.204	0.486
	0.3	0.062	0.180	0.288	0.470
	0.5	0.110	0.276	0.428	0.478
$n_1 = n_2 = 100$	0	0.056	0.144	0.360	0.564
	0.3	0.088	0.308	0.528	0.596
	0.5	0.164	0.496	0.680	0.630
$n_1 = n_2 = 150$	0	0.054	0.226	0.528	0.714
	0.3	0.120	0.450	0.710	0.714
	0.5	0.254	0.680	0.858	0.736

Note: Cor denotes the correlation between the exposure level and DL.

largest, most urban county in Kentucky as a comparison to the Appalachian region (henceforth referred to as non-Appalachian region). Among 274 cancer cases, 145 were from Appalachian and 129 from non-Appalachian; Among 253 controls, 102 were from Appalachian and 151 from non-Appalachian. Toenail samples from these subjects were collected, and 12 trace elements concentrations in toenail samples were measured as markers of long-term environmental exposure to these trace elements.

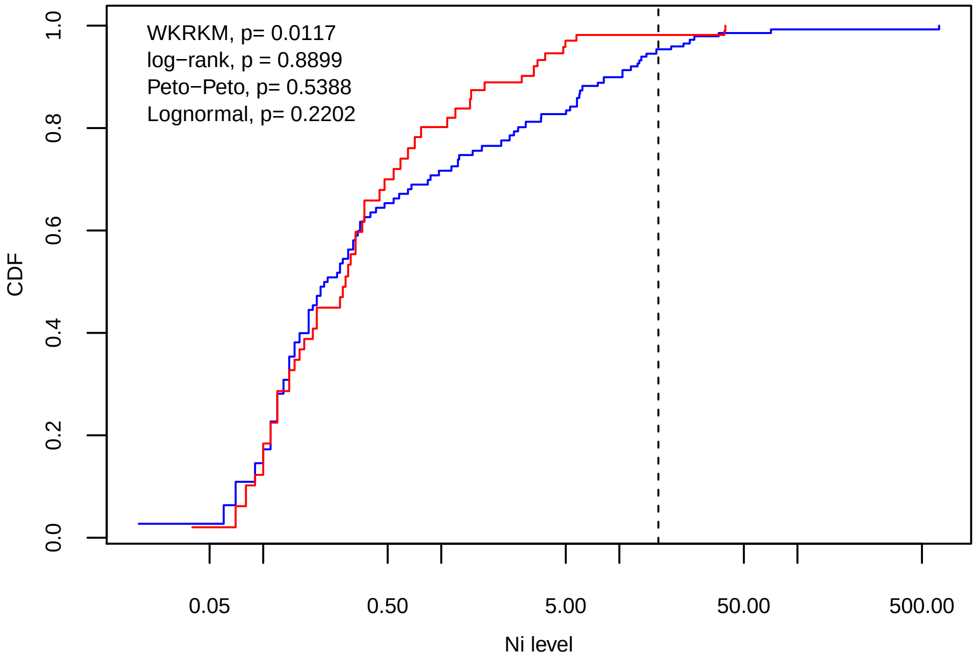
We first examine the homogeneity of DL distributions in cases and controls by region. The DLs in cases and controls for 10 metals from the Appalachian sample are drawn from different distributions. Seven metals have significantly heterogeneous results from the Non-appalachian sample.

We then use the metal Ni from Appalachian sample to demonstrate the WKRKM test. Exposure levels are estimated by KRKM estimators as shown in Figure 3.1.

Table 3.3: P value of Kolmogorov-Smirnov test between cases and controls by region

	Ni	Cd	As	Cr	Pb	Co	Al	Mn	Fe	Cu	Zn	Se
Appalachian	<.001	<.001	.331	.035	<.001	.042	<.001	.035	<.001	<.001	.005	.262
Non-appalachian	<.001	<.001	.681	.081	.162	.015	<.001	<.001	<.001	.190	<.001	<.001

Figure 3.1: KRKM Estimation of exposure distribution and 4 tests comparison in colon cancer study from Appalachian sample. The blue curves are the estimations of cases; the red curves are the estimations of controls; the dotted black curves is q in $\hat{w}(t)$ of WKRKM test statistics.



There are 145 subjects in cases and 102 subjects in controls from Appalachian sample. The non-detect rates for Ni are 0.23, 0.48 in cases and controls respectively. We compare the WKRKM test to the Log-rank, the Peto-Peto test and the Log-normal test. Estimated exposure distributions manifest in the same range for small exposure level but differ for large exposure level. Dls distributions from cases and controls in this setting are significant different($p < .001$). From the above simulation results, the Log-rank, the Peto-Peto test and the Log-normal test are tend to inflate type I error when DLs distributions are different. Although all three tests tend to reject the null hypothesis, only the WKRKM test is significant, which is consistent with the demonstration in Figure 3.1. As a result, the WKRKM test is more powerful to detect the difference at large exposure compared to Peto-Peto and Log-rank tests since it places more weight at large exposure rather than small exposure.

3.5 Discussion

To compare the exposure distributions between two groups, we have developed a class of test statistics(WKRKM) by using the integrated weight difference in the KRKM estimator. The WKRKM test statistics is based on two exposure level CDFs and has several advantages. Firstly, it will be sensitive to the alternative hypothesis of ordered CDFs and the absolute difference between two CDFs. Secondly, comparing to the tests based on hazard function, it has a more meaningful epidemiology interpretation. Thirdly, by relying on the KRKM estimator, it can deal with the correlation between the exposure level and DL. Under null hypothesis, when the distributions of DLs differ in two groups, the empirical levels of the WKRKM test are close to nominal levels, while the rank-based tests inflate the type I errors. In the cases that rank-based tests are valid, the WKRKM test are powerful to detect the difference occurs at large exposure level.

There are several extensions of the WKRKM statistics. First, the WKRKM test statistics is based on the KRKM estimator, which requires the data come from a simple random sample of the underlying population. Once we adjust the KRKM estimator by incorporating sampling weight, we can form the adjust the WKRKM statistics in the same way. Then we can extend the WKRKM test statistics to complex survey data. Second, one can extend the WKRKM test statistics to paired data. In this case, variance estimator is more complicated than the unpaired case but the bootstrap method can be used.

APPENDIX

In Chapter 1, we utilized a two-step strategy and kernel techniques to develop a nonparametric consistent estimator for the exposure distribution without imposing any independence assumption between the exposure level and DL, i.e.

$$\hat{F}(t;d) = \prod_{s>t}\left[1 - \frac{\sum_{j=1}^{n} K\{(D_j - d)/h\}dN_j(s)}{\sum_{j=1}^{n} K\{(D_j - d)/h\}Y_j(s)}\right]$$

$$\hat{F}(t) = n^{-1}\sum_{i=1}^{n}\hat{F}(t;D_i).$$

And we showed that the process $\sqrt{n}(\hat{F}(t) - F(t))$ converges weakly to a zero-mean Gaussian process and is asymptotically equivalent to the process

$$\sqrt{n}(P_n - P)\left\{F(t;D) - F(t;D)\int_t^{\infty}\frac{dN(u) - Y(u)d\ln F(u;D)}{F(u;D)I(D \leq u)}\right\}.$$

Applying functional delta method, we can easily show that $\sqrt{n}\int_0^{\infty} w(t)[\hat{F}(t) - F(t)]dt = \sqrt{n}(P_n - P)\int_0^{\infty} w(t)\xi(t)dt$, where

$$\xi(t) = \left\{F(t;D) - F(t;D)\int_t^{\infty}\frac{dN(u) - Y(u)d\ln F(u;D)}{F(u;D)I(D \leq u)}\right\}.$$

The asymptotic variance σ^2 would be the variance of $\int_0^{\infty} w(t)\xi(t)dt$. Let $g(T) = \int_0^{\infty} w(t)\eta(t)dt$, where $\eta(t) = \xi(t) - F(t)$. Since $E(\eta(t)) = 0$ for all the t, then $E(g(T)) = 0$, $\sigma^2 = E(g(T)^2) = E_D[E(g(T)^2 \mid D)]$. To study the variance, we fist study

$$\int_t^{\infty}\frac{dN(u) - Y(u)d\ln F(u;d)}{F(u;D)I(D \leq u)}.$$

By lemma 1, we can show that

$[N(t) - \int_t^{\infty} Y(u)d\ln F(u;d) \mid D] = [N(t) + \int_t^{\infty} Y(u)dR(u;d) \mid D]$ is a martingale. Also $dR(u;D) = -\ln F(u;d)$.Then

$$E\left[\int_t^{\infty}\frac{dN(u) - Y(u)d\ln F(u;d)}{F(u;D)I(D \leq u)} \mid D\right] = 0$$

$$Var\left[\int_t^{\infty}\frac{dN(u) - Y(u)d\ln F(u;d)}{F(u;D)I(D \leq u)} \mid D\right] = \int_t^{\infty}\frac{dR(u;D)}{F(u;D)I(D \leq u)}.$$

Apply the lemma of[37] and the properties of martingale,

$$E(g(T)^2 \mid D) =\{\int_0^\infty w(t)(F(t \mid D) - F(t))dt\}^2$$

$$+\int_0^\infty w^2(t)F^2(t \mid D)(\int_t^\infty \frac{dR(u;D)}{F(u;D)I(D \le u)})dt.$$

Then

$$E_D(\left[\int_0^\infty w(t)\{F(t \mid D) - F(t)\}dt\right]^2)$$

$$= \int_0^\infty \left[\int_0^\infty \int_0^\infty w(t)w(s)\{F(t \mid u) - F(t)\}\{F(s \mid u) - F(s)\}dtds\right] dG(u)$$

$$= \int_0^\infty \int_0^\infty w(t)w(s)\left[\int_0^\infty \{F(t \mid u) - F(t)\}\{F(s \mid u) - F(s)\}dG(u)\right] dtds.$$

clearly, this term is bounded.

$$E_D\left\{\int_0^\infty w^2(t)F^2(t \mid D)(\int_t^\infty \frac{dR(u;D)}{F(u;D)I(D \le u)})dt\right\}$$

$$= E_D\left\{[\int_0^\infty w^2(t)F^2(t \mid D)(\int_t^\infty \frac{-dF(u;D)}{F^2(u;D)I(D \le u)})dt\right\}$$

$$\le E_D\left\{\int_0^\infty w^2(t)F^2(t \mid D)\frac{\int_t^\infty dF(u;D)}{-F^2(t;D)I(D \le t)}dt\right\}$$

$$= E_D\left\{\int_0^\infty w^2(t)\frac{(1-F(t;D))}{-I(D \le t)}dt\right\}$$

$$= \int_0^\infty w^2(t)\left\{\int_0^\infty \frac{(1-F(t;s))}{-I(s \le t)}dG(s)\right\}dt$$

$$\le \int_0^\infty \frac{w^2(t)}{G(t)}dt \qquad (Jensen's \quad inequality)$$

Thus this term would be bounded if $|w(t)| \le \mathit{\Gamma} G(t)^{1/2+\delta}$ for some $\mathit{\Gamma} \ge 0$ and $\delta \ge 0$.

Combining the aforementioned results σ^2 is finite when this condition holds,where

$$\sigma^2 = \int_0^\infty \int_0^\infty w(t)w(s) \left[\int_0^\infty \{F(t \mid u) - F(t)\}\{F(s \mid u) - F(s)\} dG(u) \right] dtds$$

$$+ \int_0^\infty w^2(t) \left[\int_0^\infty \left\{ F^2(t \mid s)[1 - \frac{1}{F(max(t,s) \mid s)}] \right\} dG(s) \right] dt,$$

Lemma 1. $[M(t) \mid D] = [N(t) + A(t) \mid D]$ is a martingale, where $A(t) = \int_t^\infty Y(u)dR(u;d)$

Proof:

$$r(t;d)dt \approx P(t - \triangle t \leq T \leq t \mid Y \leq t)$$

$$= P(t - \triangle t \leq T \leq t \mid Y \leq t, D \leq t, D)$$

$$E(dN(t) \mid T \leq t, D) = r(t;d)dtI(T \leq t) = -dA(t) \mid D$$

$$E(dA(t) \mid T \leq t, D) = E[-r(t;d)dtI(T \leq t) \mid T \leq t, D]$$

$$= -r(t;d)dtI(T \leq t) \mid D = dA(t) \mid D$$

$$E[dM(t) \mid T \leq t, D] = E[d(N(t) + A(t) \mid T \leq t, D] = 0$$

Chapter 4 Estimation and Comparison of Exposure Distributions Adjusting for Complex Sample Designs

4.1 Introduction

In environmental exposure studies, researchers are interested in investigating the relationship between cancer and exposure to environmental chemicals such as trace elements, pesticides, and dioxins. To achieve this goal, there are two fundamental questions: (i) estimate exposure distributions under various situations, (ii) compare exposure distributions between two groups. It is very common to observe a portion of exposure measurements to fall below experimentally determined detection limits (DLs). A detection limit (DL) is "a threshold below which measured values are not considered significantly different from a blank signal, at a specified level of probability" [1]. Therefore, the exposure level of a chemical for a sample is only reported when its value is not less than the DL and otherwise is reported as a less than value or non-detect.

Both parametric and nonparametric methods have been developed to estimate and compare exposure distributions. Parametric methods, such as the Tobit model [33], assuming a normal distribution for the residual, and the accelerated failure time (AFT) models [21] for left-censored data, including the log-normal regression model as a special case, can be used since the data subject to DLs can also be treated as left-censored data [1]. But these approaches require assumptions about the underlying distribution of the exposure distribution. Nonparametric methods are widely used for DLs data since they do not require an assumption that data follow a specific distribution. The Reverse Kaplan-Meier (RKM) estimator, which mimics the Kaplan-Meier (KM) estimator for right-censored survival data with the scale reversed has been used to estimate exposure distribution. But it requires the independence assumption between the exposure level and DL. Log-rank test[34] and Peto-Peto test[14] are the most common two in the right-censored survival data and can be applied to left-censored DLs data with the scale reversed. However, the test statistics based on these tests essentially estimate integrated weight difference in hazard function and lack meaningful environmental interpretation since there is no concept that corresponds to hazard function. And the asymptotically efficiency of these tests depends on the assumption that the distributions of DLs in two groups are identical[14, 15]. To address these limitations, we proposed a kernel reverse Kaplan-Meier (KRKM) estimator for the exposure distribution without imposing any independence assumption between the exposure level and DL and a class of nonparmetric test (WKRKM) statistics by considering the integrated weighted difference in KRKM estimators of the two groups. Both KRKM estimator and WKEKM test statistics assume data come from simple random sampling.

Statisticians usually use sampling weighted estimators and the unweighted estimators tend to introduce basis. When the sample design is simple, we can use linearizion as a method to estimate the variance of an estimator that is a function of a set of simpler estimator, e.g. weighted sums. However, the variances are usually underestimated for surveys with complex sample designs when the sample design is not taken into account[38]. For example, for clustering in multistage designs, extra variability occurs because of the correlation of observations within each sampled cluster. As a result, ignoring the sample design can underestimate the variance of an estimator. Therefore, the aforementioned methods of estimations of exposure distributions and exposure difference, as well as their variances that assume simple random sampling are not suitable for data from complex sample designs. Rader and Andrew extended the log-rank test and the Peto-Peto test to complex survey data[39]. Lumley et al. implemented and updated this approach in their R package 'survey' [40].

In this chapter, we develop appropriate nonparametric methods for estimating the exposure distributions and exposure difference, as well as their variances under stratified multistage cluster samples. In section 4.2, we extended the KRKM estimator and the WKRKM test statistics to complex survey data by incorporating sampling weights. In Section 4.3, the results of several simulation studies are reported to demonstrate the performance of the proposed methods. Jackknife method is utilized for variance estimation of our proposed estimators that accounts for complex sample design and sampling weight. In Section 4.4, a National Health and Nutrition Examination Survey (NHANES) data is provided for illustration.

4.2 Methods

Kernel reverse Kaplan-Meier (KRKM) estimator

Let $\widetilde{T}$ and D be random variables for the exposure level and DL, respectively, and $F(\cdot)$ be the cumulative distribution function (CDF) of the exposure level. Let $T = \mathrm{m}ax(\widetilde{T}, D)$ and $\delta = I(\widetilde{T} \geq D)$. Here δ indicates whether T is an exposure level value or a DL value. For data subject to DL, only (T, δ, D) are observable for each subject. Suppose the data consist of n replicates $\{(T_i, \delta_i, D_i)\colon\ i = 1, \cdots, n\}$. It is useful to adopt the counting process notation. Analogous to the observed counting process and at-risk process for right censored survival data, we define two counting processes, $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(T_i \leq t)$, for the data subject to DLs.

In previous work, we utilized a two-step strategy and the kernel smoothing technique to develop a nonparametric consistent estimator for the exposure distribution. In the first step, we obtained a consistent estimator for the conditional CDF of the exposure level, denoted by $\hat{F}(t; d)$, i.e.

$$\hat{F}(t; d) = \prod_{s>t} \left[1 - \frac{\sum_{j=1}^{n} K\{(D_j - d)/h\} dN_j(s)}{\sum_{j=1}^{n} K\{(D_j - d)/h\} Y_j(s)} \right],$$

where $K(\cdot)$ is a kernel function, and h is a bandwidth such that $nh \to \infty$ and $nh^2 \to 0$ as $n \to \infty$. To ensure computational stability, following modified Silverman kernel [19] is suggested, which is flatter and less likely to produce extremely small kernel weights,

$$K(u) = \frac{|\frac{1}{2}e^{\frac{-|u|}{\sqrt{2}}}\sin(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4})|}{\int_{-\infty}^{\infty}|\frac{1}{2}e^{\frac{-|u|}{\sqrt{2}}}\sin(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4})|du}.$$

For the bandwidth, $\hat{\sigma}n^{-1/3}$ is suggested, where $\hat{\sigma}^2$ is the sample variance of the DL.In the second step, we estimated $F(t)$ by the average of estimated conditional CDF values over all DL values in the sample, i.e. $\hat{F}(t) = n^{-1}\sum_{i=1}^{n}\hat{F}(t; D_i)$. $\hat{F}(t)$ is the KRKM estimator. Though it is a consistent nonparametric estimator for the exposure distribution without requiring any independence assumption between the exposure level and DL, it requires the data come from a simple random sampling(SRS) of the underlying population. To apply the KRKM estimator to complex sample survey data we need to take into account differential sampling rates and other aspects of the sample designs.

Estimation of exposure level for stratified multistage cluster sampling

For stratified simple random sampling(SSRS), each unit in the population is categorized into disjoint and exhaustive strata prior to sampling. Simple random sampling is then done independently in each of the strata, with the sample size for each stratum set by the sampler. SSRS not only increases the statistical efficiency of estimators but also permits the calculation of accurate estimates for strata. Stratified multistage cluster sampling, a multistage version of SSRS is a commonly used complex design for national household surveys. There are two major reasons for using multistage cluster sampling. First, it minimizes the travel costs of interviews in household survey. Second, a sampling frame may not exist for individuals in the target population, but may be constructed sequentially as needed[38]. However, the clustering in multistage designs tends to increase the variability of estimators because of the correlation of observation within each cluster. Without loss of generality, we will consider stratified two-stage cluster sampling (STSCS) design. SRS is conducted independently for each stratum. The sampling rates within stratum can vary depending on the survey requirements. For example, at the first stage of clusters, called primary sample units(PSUs), a sample PSUs with probability proportional-to-size (population size of the cluster) sampling (PPS) of the PSUs is taken from within each stratum. At second stage, units are selected by SRS from the sample PSUs. This type of sample design will result in a self-weighted sample, i.e. where all the sampled second stage units have the same probability of inclusion in the sample.

In particular, we assume the finite population has L strata with K_h PSUs in the hth stratum. From the hth stratum, k_h PSUs are sampled with inclusion probabilities $\pi_{h1}, \pi_{h2}, \cdots, \pi_{hk_h}$. When the sampling is complete, there are n_{hi} units that have been sampled from the ith sampled PSU of stratum h, with sample weights ω_{hij}, $j = 1, 2, \cdots, n_{hi}$. The sample weights are the inverse of joint inclusion probabilities, i.e. $\omega_{hij} = [\pi_{hi}(n_{hi}/N_{hi})]^{-1}$, where N_{hi} is the population size of the ith sampled PSU in the hth stratum. Then we can modify KRKM estimator by incorporating sample weights, i.e.

$$\hat{F}_w(t;d) = \prod_{s>t}\left[1 - \frac{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} \omega_{hij} K\{(D_j - d)/h\} dN_{hij}(s)}{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} \omega_{hij} K\{(D_j - d)/h\} Y_{hij}(s)}\right],$$

Then $\hat{F_w}(t) = n^{-1}\sum_{i=1}^{n}\hat{F}_w(t; D_i)$. $\hat{F_w}(t)$ will be referred to as a sampling weight-adjusted KRKM (KRKM^A) estimator. The KRKM^A estimator can be easily extended for PPS sampling of the PSUs by replacing the inverse of the single inclusion probabilities and joint inclusion probabilities according to the PPS sampling scheme that is used. In the cases of a common DL, the KRKM^A estimator reduces to the RKM estimator.

Adjust weighted kernel reverse Kaplan-Meier statistics

In the previous chapter, we developed a class of test statistics, the WKRKM statistics, by using the integrated weighted difference in KRKM^A estimator for the two groups. Following the same strategy, we propose the adjusted WKRKM (WKRKM^A) statistics through the KRKM^A estimator, i.e.

$$U = \sqrt{\frac{n_1 n_2}{n}} \int_0^\infty \hat{W}(t)\{\hat{F}_{w1}(t) - \hat{F}_{w2}(t)\}dt,$$

where n_1 and n_2 are the sample sizes in the two groups, $n = n_1 + n_2$, $\hat{F}_{w1}(t)$ and $\hat{F}_{w2}(t)$ are the CDF estimates for the exposure level in the two groups obtained by the KRKM^A estimators and $\hat{W}(\cdot)$ is a random weight function that estimates a deterministic function $W(\cdot)$. In practice, the weight function can be

$$\hat{W}(t) = \frac{\hat{G}_1(t)\hat{G}_2(t)}{(n_1/n)\hat{G}_1(t) + (n_2/n)\hat{G}_2(t)},$$

where $\hat{G}_1(\cdot)$ and $\hat{G}_2(\cdot)$ are the CDF estimates of the DL in the two groups. Analogous to the WKRKM statistics, first, the WKRKM^A statistic is epidemiologically meaningful and sensitive to the alternative hypothesis of ordered CDFs and the absolute difference between two CDFs. Second, it can handle the correlation between the exposure level and DL. In additional, the WKRKM statistics can be used to test two groups exposure difference when data come from complex survey.

Variance estimation

We use jackknife leaving-one-out method for variance estimation of the KRKM^A estimator and the WKRKM^A test statistics. A jackknife variance estimator for data from the STSCS design is given by

$$\widehat{var}_{JK}(\hat{\theta}) = \sum_{h=1}^{L} \frac{k_h - 1}{k_h} \sum_{i=1}^{k_h} (\hat{\theta}_{(hi)} - \hat{\theta})^2,$$

Table 4.1: Simulation results of KRKM^A and jackknife standard error estimator when exposure level and DL are dependent. Bias, the sampling bias; SSE, the sampling standard error ; JK, the sampling mean of jackknife standard error estimator; CP, the coverage probability of the 95% confidence interval. Each entry is based on 1000 replicates and 500 bootstraps.

	True	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
			$\rho = 0$				$\rho = 0.2$		
n=160	0.25	.003	.062	.068	.951	.006	.065	.070	.924
	0.50	.005	.062	.068	.933	.007	.064	.069	.911
	0.75	.003	.049	.055	.935	.006	.051	.059	.917
n=400	0.25	.003	.040	.043	.922	.005	.046	.048	.934
	0.50	.004	.038	.044	.936	.006	.047	.051	.934
	0.75	.002	.031	.036	.949	.005	.036	.041	.947
n=800	0.25	.002	.028	.030	.957	.005	.035	.038	.947
	0.50	.003	.028	.030	.933	.006	.038	.042	.953
	0.75	.002	.022	.025	.950	.005	.029	.033	.959

where K_h PSUs are sampled from hth stratum and $\hat{\theta}_{(hi)}$ are the estimators of the same functional form as $\hat{\theta}$, but computed from the reduced sample by omitting the ith sampled PSU from stratum h[41].

4.3 Simulation Studies

KRKM^A estimator and jacknife variance estimator under STSCS designs

To access the performance of the KRKM^A estimator and jacknife variance estimator, we mimicked the trace metal cadmium (Cd) from NHANES 2005-2010. Without loss of generality, STSCS designs are used in all the situations. A finite population size of $T = 200,000$ is generated with H $= 8$ strata. The population size of each stratum is set to be the same size ($T_h = 25,000$). Each stratum is composed of 10 unequal-sized PSUs. Intra-cluster correlations (ρ) for the PSUs vary from $\rho = 0$ or 0.2. In each PSU. we generated DLs from log-normal distribution $\ln N(\mu, \sigma)$ and exposure levels from the log-normal regression model: $\ln(\widetilde{T}) = \mu' + \beta \log(D) + \rho\sigma' s + \sqrt{1-\rho^2}\sigma'\varepsilon$, where s is a fixed value sample from standard normal distribution across with PSU and ε follows a standard normal distribution. The parameters $\mu, \sigma, \mu', \beta, \sigma'$ are -1.61, 2.22, -0.79, 0.21, and 1.95. The non-detect rate of the simulated data is 44%. The total sample size draw from the population are, $N = 160$, 400, or 800 . At the first stage, 2 out of 10 clusters are selected without replacement from each stratum by using PPS, where the size measure is the PSU population size and a specified number of subjects are chosen from the selected PSUs within each stratum by SRS. Table 4.1 summarizes the results of the KRKM^A and the jackknife standard error estimator when exposure level and DL are dependent. Since the KRKM^A is a weighted version of the KRKM, the biases are small and the coverage probabilities are accurate regardless of intra-cluster correlation. The jackknife standard error estimators are slightly larger than the sampling standard errors.

To compare the performance of the KRKM^A estimator and the jacknife variance estimator along with the RKM estimator under situation that the the exposure level and DL are independent, we adopted the above set-ups but set $\beta = 0$. The non-detect rate of the simulated data is 48%. In additional, common DLs (0.061) are taken into account with non-detect rate 40%. Table 4.2 shows the comparison of the KRKM^A (with the jackknife standard error estimator) and the RKM when the exposure level and DL are independent. When there is no intra-cluster correlation, for both KRKM^A and RKM estimators, the biases are very small, the variance estimators are accurate and the confidence intervals have proper coverage probabilities. When intra-cluster correlation exists, RKM estimators have relatively large biases. Especially when the DLs are common, the biases of RKM estimators result in inappropriate coverage probabilities. In all the settings, The jackknife standard error estimators of KRKM^A are slightly larger than the sampling standard errors. With increasing sample size, the difference decreases. As expected, the variance estimators of the KRKM^A are slightly and consistently larger than variance estimators of the RKM because sample weighting usually increases the variance.

Table 4.2: Comparison of simulation results of KRKM^A (with jackknife standard error estimator) and RKM when exposure level and DL are independent. Bias, the sampling bias; SSE, the sampling standard error of both KRKM^A and RKM; JK, the sampling mean of jackknife standard error estimator; SEE, the sampling mean of the standard error estimator of RKM; CP, the coverage probability of the 95% confidence interval. Each entry is based on 1000 replicates and 500 bootstraps.

		KRKM^A				RKM			
	True	Bias	SSE	JK	CP	Bias	SSE	SEE	CP
				Various DLs					
$\rho = 0$									
n=160	0.25	.000	.055	.062	.935	.003	.038	.039	.936
	0.50	.000	.060	.066	.953	.001	.041	.042	.941
	0.75	.001	.049	.055	.945	.002	.035	.035	.934
n=400	0.25	.000	.035	.039	.950	.001	.024	.024	.943
	0.50	.000	.037	.041	.948	.001	.025	.026	.942
	0.75	.000	.030	.034	.946	.002	.022	.022	.935
n=800	0.25	.000	.025	.028	.952	.002	.017	.017	.931
	0.50	.000	.027	.030	.954	.001	.018	.019	.941
	0.75	.000	.021	.023	.952	.002	.015	.016	.929
$\rho = 0.2$									
n=160	0.25	.005	.057	.066	.952	.005	.040	.042	.942
	0.50	.005	.066	.075	.938	.006	.043	.047	.955
	0.75	.005	.055	.059	.958	.003	.036	.040	.948
n=400	0.25	.001	.040	.045	.945	.005	.028	.030	.949
	0.50	.001	.045	.050	.955	.007	.031	.034	.948
	0.75	.001	.037	.041	.953	.004	.027	.029	.941
n=800	0.25	.000	.033	.035	.946	.005	.022	.025	.959
	0.50	.000	.037	.040	.946	.005	.026	.030	.949
	0.75	.000	.030	.033	.952	.003	.022	.025	.947
				Common DLs					
$\rho = 0$									
n=160	0.25	.001	.048	.054	.952	.001	.033	.035	.949
	0.50	.000	.054	.060	.956	.003	.038	.040	.945
	0.75	.002	.049	.054	.952	.002	.035	.035	.934
n=400	0.25	.002	.031	.035	.942	.002	.022	.022	.931
	0.50	.001	.036	.040	.957	.002	.024	.026	.946
	0.75	.000	.031	.034	.951	.001	.021	.022	.949
n=800	0.25	.002	.021	.024	.958	.002	.015	.015	.945
	0.50	.002	.024	.027	.955	.003	.017	.018	.930
	0.75	.001	.022	.023	.956	.002	.015	.016	.948
$\rho = 0.2$									
n=160	0.25	.009	.052	.059	.935	.017	.036	.037	.904
	0.50	.011	.060	.066	.946	.020	.043	.044	.912
	0.75	.007	.051	.059	.961	.015	.036	.038	.946
n=400	0.25	.010	.036	.041	.929	.018	.025	.025	.865
	0.50	.012	.043	.047	.947	.022	.030	.031	.890
	0.75	.007	.037	.041	.946	.016	.026	.027	.917
n=800	0.25	.011	.029	.032	.926	.017	.019	.021	.853
	0.50	.012	.035	.037	.944	.021	.023	.026	.875
	0.75	.008	.028	.033	.948	.015	.020	.022	.903

Size properties of WKRKM^A test under STSCS designs

To access the performance of the WKRKM^A test under null hypothesis, we analyze size properties in this section. We adopt the previous STSCS designs setting. In each PSU, we mimicked the trace metal cobalt(Co) of cancer-free responders from NHANES 2005-2010. Exposure levels were generated from log-normal distribution $\ln N(-0.95, 0.79)$ and truncated under the 3σ rule. Common Dls(0.040) with non-detect rate 6% are used. To consider the situation that Dls are varying, we generated the Dls from log-normal distribution $\ln N(-1.18, 0.79)$. Then non-detect rate is 36%. In both situations, we consider two settings. Setting 1: in the sample drawn from the finite population, the sizes of two groups are equal. Setting 2: one group always has size of 36, whatever the total sample size drawn from the population. The latter setting is aimed to mimic the rare cancer cases in population. Table 4.3 summarizes the size simulation results for the Peto-Peto test, the log-rank test, the weighted Peto-Peto test, the weighted log-rank test and the WKRKM^A test. The WKRKM^A test, the Peto-Peto test and the log-rank test are valid and the empirical levels are close to nominal levels across a broad range of situations. The Peto-Peto test and the log-rank test are more conservative compared to the WKRKM^A test. The weighted Peto-Peto test and the weighted Log-rank test are not valid when the group sizes are unbalanced. When the tests are valid, intra-cluster correlation does not affect the performance of these tests.

Power properties of WKRKM^A test under STSCS designs

In this part, we conducted simulation studies of power. We adopt the previous STSCS designs setting. In each PSU, exposure levels for two groups were mimicked from trace metal cobalt(Co) of responders in controls and colon cancer cases from NHANES 2005-2010 and were generated from log-normal distribution $\ln N(-0.95, 0.79)$ and $\ln N(-1.13, 0.79)$. DLs are common (0.31) for two groups. The

Table 4.3: Size simulation results under STSCS designs

	Common DLs									
	Setting 1					Setting 2				
n	WKRKM^A	Peto	Logrank	Peto^W	Logrank^W	WKRKM^A	Peto	Logrank	Peto^W	Logrank^W
$\rho = 0$										
160	.062	.045	.046	.054	.068	.064	.066	.072	.057	.080
400	.061	.040	.055	.063	.078	.051	.060	.055	.057	.075
800	.056	.058	.047	.048	.060	.062	.064	.054	.054	.055
$\rho = 0.2$										
160	.064	.047	.049	.052	.056	.056	.048	.056	.052	.056
400	.056	.044	.060	.060	.058	.053	.039	.041	.073	.085
800	.059	.053	.049	.053	.064	.060	.057	.057	.103	.125

	Various DLs									
	Setting 1					Setting 2				
n	WKRKM^A	Peto	Logrank	Peto^W	Logrank^W	WKRKM^A	Peto	Logrank	Peto^W	Logrank^W
$\rho = 0$										
160	.062	.045	.046	.060	.049	.056	.057	.053	.060	.061
400	.060	.046	.047	.059	.057	.060	.060	.058	.058	.056
800	.059	.046	.051	.048	.053	.062	.050	.047	.063	.060
$\rho = 0.2$										
160	.059	.051	.046	.068	.067	.059	.059	.060	.069	.070
400	.062	.054	.050	.069	.057	.065	.065	.049	.059	.060
800	.064	.046	.041	.050	.054	.062	.049	.051	.066	.072

Note: N: the total sample size of two groups combined

WKRKM^A, adjust WKRKM test; Peto^W, weighted Peto-Peto test, Logrank^W, weighted Log-rank test.

non-detect rate of the simulated data are 38% and 43% in two groups. Two settings of sample sizes allocation for two groups are considered. The ratio of cases to controls are 1:1 and 1:4 in the two settings, corresponding to setting 1 and setting 2. Table 4.4 summarizes the power simulation results for the Peto-Peto test, the log-rank test, the weighted Peto-Peto test, the weighted Log-rank test and the WKRKM^A test. Five tests are more powerful when two group sizes are balanced. The WKRKM^{A}test, the Peto-Peto test and the log-rank test attain high efficiency over the weighted Peto-Peto test, the weighted Log-rank test across a broad range of situations. When two group sizes are balanced, the WKRKM^A test is more powerful for small sample size and less power with the increasing of sample size compared to the Peto-Peto test and log-rank test. When two group sizes are unbalanced, it indicates that the superior performance of the WKRKM^A test over the Peto-Peto test and the log-rank test. Intra-cluster correlation does not affect the performance.

4.4 Example

The NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. Starting in 1999, NHANES became a continuous, ongoing annual survey of the noninstitutionalized civilian resident population of the United States. About 12,000 persons per 2-year cycle were asked to participate in NHANES. Response rates varied by year, but an average of 10,500 persons out of the initial 12,000 agreed to complete a household interview. A four-stage sampling design was used: (i) selection of PSUs, which are counties or small groups of contiguous counties; (ii) selection of segments within PSUs that constitute a block or group of blocks containing a cluster of households; (iii) selection of specific households within segments; (iiii) selection of individuals within a household. A sample weight was assigned to each sample person. Weighting took into account several features of the survey: the differential probabilities of selection for the individual domains;

Table 4.4: Power simulation results at significance level $\alpha = 0.05$ under STSCS designs

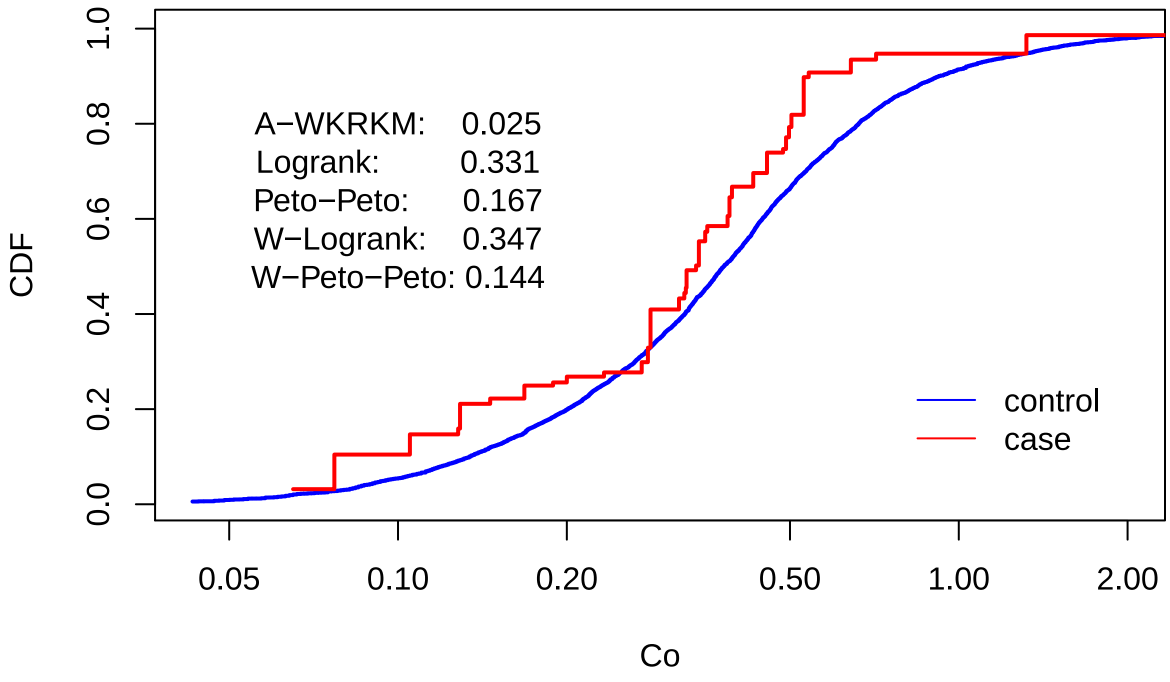
n	WKRKM^A	Peto	Logrank	Peto^W	Logrank^W
			Setting 1		
$\rho = 0$					
160	.322	.286	.241	.183	.161
400	.518	.599	.550	.371	.315
800	.760	.885	.842	.627	.567
$\rho = 0.2$					
160	.335	.331	.299	.209	.203
400	.518	.645	.583	.362	.348
800	.714	.879	.840	.582	.520
			Setting 2		
$\rho = 0$					
160	.315	.153	.130	.162	.123
400	.502	.410	.361	.301	.259
800	.688	.677	.663	.503	.448
$\rho = 0.2$					
160	.261	.174	.171	.159	.133
400	.452	.434	.404	.272	.247
800	.633	.612	.622	.447	.405

Note: N: the total sample size of two groups combined
WKRKM^A, adjust WKRKM test; Peto^W, weighted Peto-Peto test, Logrank^W, weighted Log-rank test.

nonresponse to survey instruments; and differences between the final sample and the total population[7]. Masked Variance Strata and Masked Variance Units or MVUs are used to protect the confidentiality of information provided by survey participants and to reduce disclosure risks. The variance estimates that are produced, using the Masked strata and MVUs, closely approximate the variances that would have been estimated using the true sample design variance units that are based on the actual survey sample strata and PSUs[8].

From NHANES 2005-2010, inductively coupled plasma-mass spectrometry (ICP-MS) method is used to measure the following 12 elements in urine: beryllium (Be), cobalt (Co), molybdenum (Mo), cadmium (Cd), antimony (Sb), cesium (Cs), barium (Ba), tungsten (W), platinum (Pt), thallium (TI), lead (Pb), and uranium (U)[42, 43].

Figure 4.1: KRKM^A estimation of exposure distribution and p values of tests comparison in cobalt (Co).



The detection limits were constant for all of the heavy metals. The NHANES medical conditions questionnaire (MCQ) section is generally modeled on the ”Medical Conditions” questionnaire section of the U.S. National Health Interview Survey. It provides self-reported personal interview data on a broad range of health conditions for both children and adults. Among the responders in MCQ section from NHANES 2005-2010, there were 8353 participants whose laboratory urine samples were available. 36 were self reported as colon cancer cases. There were 46 pseudo-strata each with two pseudo-PSUs of the NHANES 2005-2010.

We then use cobalt (Co) to demonstrate the KRKM^A estimator and the WKRKM^A test. We classify the participants by colon cancer. Figure 4.1 displays CDF estimates of the KRKM^A and 5 tests comparison.

Chapter 5 R package: krkm

5.1 Document

Package ‘krkm’

December 9, 2016

Type Package

Title Statistical Methods for Environmental Exposure Data Subject to Detection Limits

Version 1.0

Date 2016-05-13

Author Yuchen Yang

Maintainer Yuchen Yang <`yuchen.y@uky.edut`>

Description Estimation of Exposure Distribution Adjusting for Dependence between Exposure Level and Detection Limit.Comparison of Exposure Level Distributions Between Two Groups.

License GPL-2

`krkm-package` *Statistical Methods for Environmental Exposure Data Subject to Detection Limits*

Description

Estimation of Exposure Distribution Adjusting for Dependence between Exposure Level and Detection Limit.

Comparison of Exposure Level Distributions Between Two Groups.

Details

Package:	krkm
Type:	Package
Version:	1.0
Date:	2016-05-13
License:	GPL-2

Author(s)

Yuchen Yang

Maintainer: Yuchen Yang <yuchen.y@uky.edu>

`KRKM` *Calculate exposure distribution for detection limit data*

Description

Estimate exposure level by kernel reverse Kaplan-Meier(KRKM) estimator.

Usage

```
KRKM(obs, lod, wei = rep(1, length(obs)), bandwidth = NULL,

psu = NULL, stra= NULL, b = 1000, var = FALSE)
```

Arguments

`obs`	Observed exposure levels
`lod`	Detection limit
`wei`	Sampling weight. The default value has the same sampling weight and assumes data come from simple random sampling.

`bandwidth` Bandwidth of Silverman kernel. The default value use $\hat{\sigma}n^{-1/3}$, where $\hat{\sigma}^2$ is the sample variance of the detection limit.

`psu` Identifier for primary sampling units. The default value is NULL and assumes data come from simple random sampling.

`stra` Identifier for sampling strata. The default value is NULL and assumes data come from simple random sampling.

`b` Number of bootstrap replicates to calculate variance when data come from simple random sampling. The default value is 1000.

`var` Indicator for variance calculation. The default value is false.

Details

This function calculates exposure distribution for detection limit data. It can handel the correlation between the exposure level and detection limit when data either come from simple random sampling or complex survey design.

Value

`obs` Unique observed exposure

`n.risk` Number of subjects that exposure levels are greater then current exposure level

`n.event` Number of subjects that exposure levels are observed at current exposure level

`prob` Estimates of exposure level

`sd` The standard errors of the estimated exposure level

`lower.cl` The 95% lower confidence limits of exposure level

`upper.cl` The 95% upper confidence limits of exposure level

Note

Bootstrap method is used to estimate variance when data come from simple random sampling. Jackknife method is used to estimate variance when data come from complex survey design.

Author(s)

Yuchen Yang

See Also

plot.KRKM, summary.KRKM

Examples

```
data(nhanes)
t=as.numeric(nhanes[,'co'])
data1=nhanes[!is.na(t),]
wei=as.numeric(data1[,2])
t=as.numeric(data1[,'co'])
lod=as.numeric(data1[,'co_dl'])
stra=as.numeric(data1[,'stra'])
psu=as.numeric(data1[,'psu'])

fit<-KRKM(obs=t,lod=lod,wei=wei,psu=psu,stra=stra)
summary(fit)
```

`nhanes` *NHANES*

Description

NHANES 2005-2010 Laboratory Data

Usage

```
data("nhanes")
```

Format

A data frame with 8353 observations on the following 58 variables.

Details

It cintains following 4 elements in urine:cobalt (Co), cadmium (Cd), lead (Pb) and Arsenics (As)

Source

http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm

References

Zipf, G., Chiappa, M., Porter, K., Ostchega, Y., Lewis, B., and Dostal, J. (2013).National health and nutrition examination survey: plan and operations, 1999-2010.Vital and health statistics. Ser. 1, Programs and collection procedures, (56):1-37.

Examples

```
data(nhanes)
```

`plot.KRKM` *Plot method for KRKM objects*

Description

Plot the estimated exposure level obtained by the KRKM function

Usage

```
plot.KRKM(t, ...)
```

Author(s)

Yuchen Yang

See Also

KRKM

Examples

```
data(nhanes)
t=as.numeric(nhanes[,'co'])
data1=nhanes[!is.na(t),]
wei=as.numeric(data1[,2])
t=as.numeric(data1[,'co'])
lod=as.numeric(data1[,'co_dl'])
stra=as.numeric(data1[,'stra'])
psu=as.numeric(data1[,'psu'])

fit<-KRKM(obs=t,lod=lod,wei=wei,psu=psu,stra=stra)
plot(fit)
```

`plot.WKRKM` *Plot method for KRKM objects*

Description

Plot exposure level estimates of two groups by KRKM estimator.

Usage

```
plot.WKRKM(t, ...)
```

Note

WKRKM

Author(s)

Yuchen Yang

See Also

WKRKM

Examples

```
data(nhanes)
t=as.numeric(nhanes[,'co'])
data1=nhanes[!is.na(t),]
wei=as.numeric(data1[,2])
t=as.numeric(data1[,'co'])
lod=as.numeric(data1[,'co_dl'])
stra=as.numeric(data1[,'stra'])
psu=as.numeric(data1[,'psu'])


group=data1[,'cancer']=='lung'
fit<-WKRKM(t=t,lod=lod,wei=wei,group=group,psu=psu,stra=stra)

plot(fit)
```

WKRKM *Compare two exposure level distributions*

Description

Comparison of exposure level distributions between two groups by WKRKM test statistics.

Usage

```
WKRKM(t, lod, wei = rep(1, length(t)), group, bandwidth = NULL,
 psu = NULL, stra= NULL, b = 1000, cl = FALSE, weight =
function(x, g1, g2, n1, n2) {
```

```
out = (g1(x) * g2(x))/(n1/(n1 + n2) * g1(x) + n2/(n1 + n2) * g1(x))
return(out)
})
```

Arguments

`t` Observed exposure levels

`lod` Detection limit

`wei` Sampling weight. The default value has the same sampling weight and assumes data come from simple random sampling.

`group` Two group indicator

`bandwidth` Bandwidth of Silverman kernel of KRKM estimator. The default value use $\hat{\sigma} n^{-1/3}$, where $\hat{\sigma}^2$ is the sample variance of the detection limit.

`psu` Identifier for primary sampling units. The default value is NULL and assumes data come from simple random sampling.

`stra` Identifier for sampling strata. The default value is NULL and assumes data come from simple random sampling.

`b` Number of bootstrap replicates to calculate variance when data come from simple random sampling. The default value is 1000.

`cl` Indicator for variance calculation of KRKM estimator. The default value is false. If TRUE, the plot.WKRKM() would contain confidence limits.

`weight` Weight function for WKRMK statistics. The default function is $\hat{w}(t) = \frac{\hat{G}_1(t)\hat{G}_2(t)}{(n_1/n)\hat{G}_1(t)+(n_2/n)\hat{G}_2(t)}$, where $\hat{G}_1(\cdot)$ and $\hat{G}_2(\cdot)$ are the CDF estimates of the DL in the two groups.

Value

`z`	Test statistics
`pvalue`	P-value
`group1`	KRKM object for the first group
`group2`	KRKM object for the second group

Note

Bootstrap method is used to estimate variance when data come from simple random sampling. Jackknife method is used to estimate variance when data come from complex survey design.

Author(s)

Yuchen Yang

See Also

plot.WKRKM

Examples

```
data(nhanes)
t=as.numeric(nhanes[,'co'])
data1=nhanes[!is.na(t),]
wei=as.numeric(data1[,2])
t=as.numeric(data1[,'co'])
lod=as.numeric(data1[,'co_dl'])
stra=as.numeric(data1[,'stra'])
psu=as.numeric(data1[,'psu'])

group=data1[,'cancer']=='lung'
WKRKM(t=t,lod=lod,wei=wei,group=group,psu=psu,stra=stra)
```

5.2 Source codes

```
silverman<-function(x)

{

abs(0.5*exp(-abs(x)/sqrt(2))*sin(abs(x)/sqrt(2)+pi/4))/1.140086

}

ker2<-function(obs,lod,wei,bandwidth)

{

n=length(lod)

censored= obs>lod

if(is.null(bandwidth)){

h=1.059*sd(lod)*n^(-1/3)

}else{

h=bandwidth

}

ind<-order(obs)

data=cbind(obs,censored)

data=data[ind,]

wd=lod[ind]

obs=obs[ind]

we=wei[ind]

dn<-data[,2]

unique.obs=unique(obs)

if(length(unique(obs))!=length(obs)){

een=tn=en=ttn=rep(NA,length(unique.obs))

for (i in 1:length(unique.obs))

{

ttn[i]=sum(obs==unique.obs[i])
```

```
tn[i]=sum(  we[obs==unique.obs[i]] )
indi=which(obs==unique.obs[i])
en[i]=sum(dn[indi]*we[indi])
een[i]=sum(dn[which(obs==unique.obs[i])])
}
}else{
tn=we
en=dn*we
ttn=rep(1,length(obs))
een=dn
}
if (length(unique(lod))==1){
num=en
deno=cumsum(tn)
pro=1-(num/deno)
out=rev(cumprod(rev(pro[2:length(unique.obs)])))
prob=c(out,1)
}else
{
record=rep(NA,length(unique.obs))
num=deno=temp.num=temp.deno=matrix(0,length(obs),length(unique.obs))
for(i in 1: length(obs)){
if(obs[i]<=max(unique.obs)){
record=min(which(unique.obs>=obs[i]))
temp.deno[i,record:length(unique.obs)]=1
if(dn[i]==1){
temp.num[i,record]=1
}
}
}
dist=t(kronecker(t(wd),rep(1,n))-wd)/h
temp=silverman(dist)
sw=t(kronecker(t(we),rep(1,length(unique.obs))))
temp.num=temp.num*sw
temp.deno=temp.deno*sw
for(i in 1:length(lod))
{
for(j in 1:length(unique.obs))
{
num[i,j]=sum(temp[i,]*temp.num[,j])
deno[i,j]=sum(temp[i,]*temp.deno[,j])
}
}
pro=1-(num/deno)
cond=t(apply(pro,1,function (x){
out=rev(cumprod(rev(x[2:length(unique.obs)])))
out=c(out,1)
return(out)
}  ))
prob=apply(cond,2,mean)
}
out<-cbind(obs=unique.obs,n.risk=cumsum(ttn),n.event=een,prob)
out=subset(out,out[,3]>0)
return(out)
```

```
}
KRKM <- function(obs, lod,wei=rep(1,length(obs)),bandwidth=NULL,
psu=NULL,stra=NULL,b=1000,var=FALSE) UseMethod("KRKM")
KRKM.default<-function(obs,lod,wei=rep(1,length(obs)),
bandwidth=NULL,psu=NULL,stra=NULL,b=1000,var=FALSE)
{
if(var==0){
output=ker2(obs,lod,wei,bandwidth)
}else{
if(length(unique(wei))==1){
out=ker2(obs,lod,wei,bandwidth)
temp=matrix(NA,dim(out)[1],b)
for(i in 1:b)
{
ind=sample(1:length(lod),replace=TRUE)
obs.b=obs[ind]
lod.b=lod[ind]
wei.b=wei[ind]
x=out[,c(1,4)]
y=ker2(obs.b,lod.b,wei.b,bandwidth)[,c(1,4)]
temp[,i]=merge(x,y,by='obs',all.x=TRUE)[,3]
}
sd=apply(temp,1,sd,na.rm=TRUE)
lower.cl=pmax(rep(0,length(sd)),out[,4]-1.96*sd)
upper.cl=pmin(rep(1,(length(sd))),out[,4]+1.96*sd)
var=cbind(sd,lower.cl,upper.cl)
output=cbind(out,var)
```

```
}else{
out=ker2(obs,lod,wei,bandwidth)
temp.stra=matrix(NA,dim(out)[1],length(unique(stra)))
for( j in 1:length(unique(stra)))
{
ind3=stra==unique(stra)[j]
kh=length(unique(psu[ind3]))
temp.psu=matrix(NA,dim(out)[1],kh)
for(i in 1:kh)
{
ppsu=psu[stra==unique(stra)[j]]
ind1=stra==unique(stra)[j]&psu==unique(ppsu)[i]
ind2=stra==unique(stra)[j]&psu==unique(ppsu)[3-i]
wei[ind2]=kh/(kh -1 )*wei[ind2]
tjk=t[!ind1]
lodjk=lod[!ind1]
weijk=wei[!ind1]
x=out[,c(1,4)]
y=ker2(tjk,lodjk,weijk,bandwidth)[,c(1,4)]
temp=merge(x,y,by='obs',all.x=TRUE)
temp.psu[,i]=(kh-1)/kh*(temp[,2]-temp[,3])^2
}
temp.stra[,j]=apply(temp.psu,1,sum,na.rm=TRUE)
}
sd=sqrt(apply(temp.stra,1,sum,na.rm=TRUE))
lower.cl=pmax(rep(0,length(sd)),out[,4]-1.96*sd)
upper.cl=pmin(rep(1,(length(sd))),out[,4]+1.96*sd)
```

```
var=cbind(sd,lower.cl,upper.cl)
output=cbind(out,var)
}
}
class(output) <- 'KRKM'
output
}
summary.KRKM<-function(t,...)
{
print(t[,])
}
print.KRKM<-function(t,...)
{
n=tail(t[,2],1)
event=sum(t[,3])
output=cbind(n,event)
colnames(output)=c('N','event')
print(output)
}
plot.KRKM<-function(t,...)
{
options( warn = -1 )
if(dim(t)[2]==4){
plot(t[,1],t[,4],type='s',col='black',lty=1,log="x",ylab='CDF'
,xlab='',lwd=1)
}else{
plot(t[,1],t[,4],type='s',col='black',lty=1,log="x",ylab='CDF'
,xlab='',lwd=1)
points(t[,1],t[,6],type='s',col='black',lty=2,log="x",lwd=1)
points(t[,1],t[,7],type='s',col='black',lty=2,log="x",lwd=1)
}
}
testest<-function(t,lod,group,wei,bandwidth,cl,weight=
function(x,g1,g2,n1,n2){
out=(g1(x)*g2(x))/( n1/(n1+n2)*g1(x)  +n2/(n1+n2)*g1(x))
return( out )
}
)
{
names=unique(group)
ind=group==names[1]
t1=t[ind]
t2=t[!ind]
lod1=lod[ind]
lod2=lod[!ind]
wei1=wei[ind]
wei2=wei[!ind]
g1=ecdf(lod1)
g2=ecdf(lod2)
n1=length(t1)
n2=length(t2)
bound=max(t1,t2,lod1,lod2)
temp1=KRKM(t1,lod1,wei1,bandwidth,var=cl)
temp2=KRKM(t2,lod2,wei2,bandwidth,var=cl)
```

```
f1=stepfun(temp1[,1],c(0,temp1[,4]))
f2=stepfun(temp2[,1],c(0,temp2[,4]))
t=sort(c(temp1[,1],temp2[,1]))
wet=weight(t,g1,g2,n1,n2)
wet[is.na(wet)]=0
f12=f1(t)-f2(t)
t.gap=c(diff(t),bound-max(t))
est=sum(  (wet*f12)*t.gap)
return(list(est=est,group1=temp1,group2=temp2,names=names))
}
WKRKM<-function(t,lod,wei=rep(1,length(t)),group,bandwidth=NULL,
psu=NULL,stra=NULL,b=1000,cl=FALSE,weight=function(x,g1,g2,n1,n2)
{
out=(g1(x)*g2(x))/( n1/(n1+n2)*g1(x)  +n2/(n1+n2)*g1(x))
return( out )
}) UseMethod("WKRKM")
WKRKM.default<-function(t,lod,wei=rep(1,length(t)),group,
bandwidth=NULL,psu=NULL,stra=NULL,b=1000,cl=FALSE,weight=
function(x,g1,g2,n1,n2){
out=(g1(x)*g2(x))/( n1/(n1+n2)*g1(x)  +n2/(n1+n2)*g1(x))
return( out )
})
{
temp.test=testest(t,lod,group,wei,bandwidth,cl,weight=weight)
est=temp.test$est
if(length(unique(wei))!=1){
temp.stra=rep(NA,length(unique(stra)))
```

```
rec=NULL
for( j in 1:length(unique(stra)))
{
ind3=stra==unique(stra)[j]
kh=length(unique(psu[ind3]))
temp.psu=rep(NA,kh)
for(i in 1:kh)
{
ppsu=psu[stra==unique(stra)[j]]
ind1=stra==unique(stra)[j]&psu==unique(ppsu)[i]
ind2=stra==unique(stra)[j]&psu==unique(ppsu)[3-i]
wei[ind2]=kh/(kh -1 )*wei[ind2]
tjk=t[!ind1]
lodjk=lod[!ind1]
weijk=wei[!ind1]
groupjk=group[!ind1]
temp=testest(tjk,lodjk,groupjk,weijk,bandwidth,cl,weight=weight)$est
rec=c(rec,temp)
temp.psu[i]=(kh-1)/kh*(temp-est)^2
}
temp.stra[j]=sum(temp.psu)
}
sd=sqrt(sum(temp.stra)/(kh^2))
}else
{
temp=rep(NA,b)
for(i in 1:b)
```

```
{
ind=sample(1:length(lod),replace=TRUE)
t.b=t[ind]
lod.b=lod[ind]
wei.b=wei[ind]
group.b=group[ind]
temp[i]=testest(t.b,lod.b,group.b,wei.b,bandwidth,cl=FALSE,
weight=weight)$est
sd=sd(temp,na.rm=TRUE)
}
}
stat=abs(est/sd)
output=list(z=est/sd,pvalue=2*(1-pnorm(stat)),group1=temp.test$group1,
group2=temp.test$group2,names=temp.test$names)
class(output) <-'WKRKM'
output
}
print.WKRKM<-function(t,...)
{
cat('z: \n')
print(round(t$z,3))
cat('p-value: \n')
print(round(t$pvalue,4))

n1=tail(t$group1,1)[2]
n2=tail(t$group2,1)[2]
event1=sum(t$group1[,3])
```

```
event2=sum(t$group2[,3])
output=cbind(c(n1,n2),c(event1,event2))
colnames(output)=c('N','event')
rownames(output)=c(paste('group=',t$names[1]),
paste('group=',t$names[1]))
print(output)
}
plot.WKRKM<-function(t,...)
{
xl=min(t$group1[,1],t$group2[,1])
xu=max(t$group1[,1],t$group2[,1])
options( warn = -1 )
if(dim(t$group1)[2]==4){
plot(t$group1[,1],t$group1[,4],type='s',col='blue',lty=1,
log="x",ylab='CDF',xlab='',lwd=1,xlim=c(xl,xu))
points(t$group2[,1],t$group2[,4],type='s',col='red',log="x",
lty=1,lwd=1)
}else
{
plot(t$group1[,1],t$group1[,4],type='s',col='blue',lty=1,log="x",
ylab='CDF',xlab='',lwd=1,xlim=c(xl,xu))
points(t$group1[,1],t$group1[,6],type='s',col='blue',log="x",
lty=2,lwd=1)
points(t$group1[,1],t$group1[,7],type='s',col='blue',log="x",
lty=2,lwd=1)
points(t$group2[,1],t$group2[,4],type='s',col='red',log="x",
lty=1,lwd=1)
points(t$group2[,1],t$group2[,6],type='s',col='red',log="x",
lty=2,lwd=1)
points(t$group2[,1],t$group2[,7],type='s',col='red',log="x",
lty=2,lwd=1)
}
}
```

Bibliography

[1] Dennis R Helsel. *Nondetects and data analysis. Statistics for censored environmental data.* Wiley-Interscience, 2005.

[2] Nancy Johnson, Brent J Shelton, Claudia Hopenhayn, Thomas T Tucker, Xianglin Shi, Jason M Unrine, Bin Huang, W Jay Christian, Zhuo Zhang, and Li Li. Concentrations of arsenic, chromium, and nickel in toenail samples from appalachian kentucky residents. *Journal of Environmental Pathology, Toxicology and Oncology*, 30(3):213–223, 2011.

[3] Dennis R Helsel. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, 65(11):2434–2439, 2006.

[4] Jay H Lubin, Joanne S Colt, David Camann, Scott Davis, James R Cerhan, Richard K Severson, Leslie Bernstein, and Patricia Hartge. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*, 112(17):1691–1696, 2004.

[5] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 62(1):10–29, 2012.

[6] Li Li, Sarah J Plummer, Cheryl L Thompson, Thomas C Tucker, and Graham Casey. Association between phosphatidylinositol 3-kinase regulatory subunit p85α met326ile genetic polymorphism and colon cancer risk. *Clinical Cancer Research*, 14(3):633–637, 2008.

[7] G Zipf, M Chiappa, KS Porter, Y Ostchega, BG Lewis, and J Dostal. National

health and nutrition examination survey: plan and operations. *Vital and Health Statistics. Ser. 1, Programs and Collection Procedures*, (56):1–37, 2013.

[8] CL Johnson, R Paulose-Ram, CL e Ogden, MD Carroll, D Kruszon-Moran, SM Dohrmann, and LR Curtin. National health and nutrition examination survey: analytic guidelines. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (161):1–24, 2013.

[9] Brenda W Gillespie, Qixuan Chen, Heidi Reichert, Alfred Franzblau, Elizabeth Hedgeman, James Lepkowski, Peter Adriaens, Avery Demond, William Luksemburg, and David H Garabrant. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. *Epidemiology*, 21(4):S64–S70, 2010.

[10] Donghui Zhang, Chunpeng Fan, Juan Zhang, and Cun-Hui Zhang. Nonparametric methods for measurements below detection limit. *Statistics in Medicine*, 28(4):700, 2009.

[11] Bruce W Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345): 169–173, 1974.

[12] Major Greenwood. Reports on public health and medical subjects no. 33, appendix 1: the errors of sampling of the survivorship tables. *London: Her Majesty's Stationary Office*, 1926.

[13] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[14] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135:185–207, 1972.

[15] Peter R Cox. *Life Tables*. Wiley Online Library, 1972.

[16] Margaret Sullivan Pepe and Thomas R Fleming. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 45(2): 497–507, 1989.

[17] Ab Razak, Nurul Hafiza, Sarva Mangala Praveena, and Zailina Hashim. Toenail as a biomarker of heavy metal exposure via drinking water: a systematic review. *Reviews on Environmental Health*, 30(3):1–7, 2014.

[18] Dorota M Dabrowska. Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics*, 17(3):1157–1167, 1989.

[19] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. CRC press, 1986.

[20] Lopaka Lee. *NADA: Nondetects And Data Analysis for Environmental Data*, 2013. URL `https://CRAN.R-project.org/package=NADA`. R package version 1.5-6.

[21] David Collett. *Modelling Survival Data in Medical Research*. CRC press, 2015.

[22] Yanming Li, Brenda W Gillespie, Kerby Shedden, and John A Gillespie. Calculating profile likelihood estimates of the correlation coefficient in the presence of left, right or interval censoring and missing data. *Technical Report*, 2015.

[23] Michael G Akritas, Susan A Murphy, and Michael P LaValley. The theil-sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association*, 90(429):170–177, 1995.

[24] Hirotugu Akaike. Prediction and entropy. *A Celebration of Statistics*, 1(1):1–24, 1985.

[25] S Murray and AA Tsiatis. A nonparametric approach to incorporating prognostic longitudinal covariate information in survival estimation. *Biometrics*, 52(1):137–151, 1996.

[26] Li Chen, DY Lin, and Donglin Zeng. Attributable fraction functions for censored event times. *Biometrika*, 97(3):713–726, 2010.

[27] James M Robins and Dianne M Finkelstein. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[28] James M Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, pages 24–33, 1993.

[29] Stephen R Cole, Haitao Chu, Lei Nie, and Enrique F Schisterman. Estimating the odds ratio when exposure has a limit of detection. *International Journal of Epidemiology*, 38(6):1674–1680, 2009.

[30] Ryan C May, Joseph G Ibrahim, and Haitao Chu. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine*, 30(20):2551–2561, 2011.

[31] AW Van der Vaart and JA Wellner. *Weak Convergence and Empirical Processes.* Springer, New York, 1998.

[32] Dennis R Helsel. *Statistics for censored environmental data using Minitab and R.* John Wiley & Sons, 2011.

[33] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 26(1):24–36, 1958.

[34] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[35] Margaret Sullivan Pepe and Thomas R Fleming. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):341–352, 1991.

[36] John Tukey. Exploratory data analysis. *Reading: Addison-Wesley*, 1977.

[37] James H Ware and David L Demets. Reanalysis of some baboon descent data. *Biometrics*, 32(2):459–463, 1976.

[38] Edward L Korn and Barry I Graubard. *Analysis of health surveys.* John Wiley & Sons, 2011.

[39] Kevin Andrew Rader. *Methods for Analyzing Survival and Binary Data in Complex Surveys.* PhD thesis, 2014.

[40] Thomas Lumley et al. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004.

[41] Kirk Wolter. *Introduction to variance estimation.* Springer Science & Business Media, 2007.

[42] Antonio R Damasio. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1(1):123–132, 1989.

[43] Kevin J Mulligan, Timothy M Davidson, and Joseph A Caruso. Feasibility of the direct analysis of urine by inductively coupled argon plasma mass spectrometry for biological monitoring of exposure to metals. *Journal of Analytical Atomic Spectrometry*, 5(4):301–306, 1990.

Vita

- Place of birth: Hangzhou, China

- Educational institutions attended and degrees already awarded

  PhD student in Statistics: University of Kentucky, 2013-now

  MS in Statistics: University of Kentucky, 2011-2013

  BS in mathematics: Lanzhou University, 2007-2011

- Employment:

  Research Assistant, University of Kentucky, 2012-2016

  Teaching Assistant, University of Kentucky, 2011-2012

- Yuchen Yang