Saegusa, Takumi

Article

Confidence bands for a distribution function with merged data from multiple sources

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Saegusa, Takumi (2020) : Confidence bands for a distribution function with merged data from multiple sources, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 21, Iss. 4, pp. 144-158, https://doi.org/10.21307/stattrans-2020-035

This Version is available at:
https://hdl.handle.net/10419/236785

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 144–158, DOI 10.21307/stattrans-2020-035
Received – 31.01.2020; accepted – 30.06.2020

Confidence bands for a distribution function with merged data from multiple sources

Takumi Saegusa[1]

ABSTRACT

We consider nonparametric estimation of a distribution function when data are collected from multiple overlapping data sources. Main statistical challenges include (1) heterogeneity of data sets, (2) unidentified duplicated records across data sets, and (3) dependence due to sampling without replacement from a data source. The proposed estimator is computable without identifying duplication but corrects bias from duplicated records. We show the uniform consistency of the proposed estimator over the real line and its weak convergence to a Gaussian process. Based on these asymptotic properties, we propose a simulation-based confidence band that enjoys asymptotically correct coverage probability. The finite sample performance is evaluated through a simulation study. A Wilms tumor example is provided.

Key words: confidence band, data integration, Gaussian process.

1. Introduction

We consider nonparametric estimation of a distribution function F of a random variable X when data are collected from multiple overlapping data sources. Inference on F is a rather simple problem if data are independent and identically distributed (i.i.d.). When data sets are merged from various sources, this basic question faces a significant challenge from both theoretical and methodological perspectives. Statistical issues we address in this paper is (1) heterogeneity of data sources, (2) unidentified duplicated records in multiple data sets, and (3) finite population sampling from each data source. Without proper care, these issues yield bias in estimation and wrong quantification of uncertainty.

The following setting (schematically shown in Figure 1) is considered:

• The variables of interest for data integration is a random vector $W = (X, Y)$ taking values in a measurable space $(\mathscr{W}, \mathscr{A})$. In this paper, we focus on inference regarding X but inference on X and Y is of general interest in data integration.

• Let $V = (\tilde{W}, Z) \in \mathscr{V}$ where $\tilde{W}$ is a coarsening of W and Z is a vector of auxiliary variables. The variables Z do not involve inference on W but help to create data sources. The space $\mathscr{V}$ is composed of J overlapping population data sources $\mathscr{V}^{(1)}, \ldots, \mathscr{V}^{(J)}$ with $\mathscr{V} = \cup_j \mathscr{V}^{(j)}$ and $\mathscr{V}^{(j)} \cap \mathscr{V}^{(k)} \neq \emptyset$ for some (j,k). Values of V determine membership of data sources.

• Data collection is carried out in a two-stage framework. First, a large i.i.d. sample of $V_1, \ldots, V_N$ is collected from a population. The unit i is distributed to data source j if

[1]University of Maryland. USA. E-mail: tsaegusa@umd.edu. ORCID: https://orcid.org/0000-0001-6869-2451.

$V_i \in \mathscr{V}^{(j)}$. Because data sources overlap, the unit i may belong to multiple sources. The sample size of data source $\mathscr{V}^{(j)}$ is denoted as $N^{(j)}$.

• Next, a random sample of size $n^{(j)}$ is selected without replacement from data source $\mathscr{V}^{(j)}$. The selection probability for this data source is $\pi^{(j)}(V_i) = (n^{(j)}/N^{(j)})I\{V_i \in \mathscr{V}^{(j)}\}$ where I is the indicator function. For selected items, variables W_i are observed.

• The above procedure is repeated for all data sources. Data sets from each data source are then combined and statistical analysis is conducted. If the unit i is selected multiple times, its duplication is not identified.

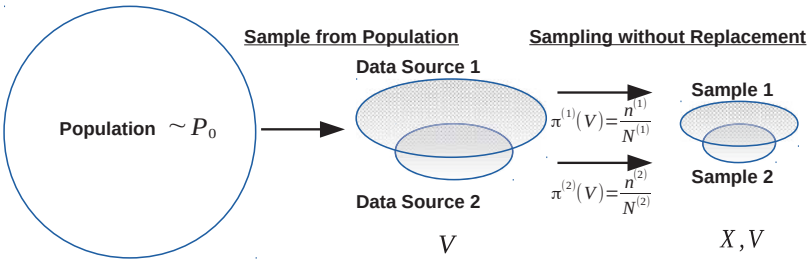

Figure 1: Sampling scheme for merged data from multiple sources with $J = 2$.

This two-stage formulation is essential in describing duplicated records in multiple data sets. Duplication naturally occurs in public health data integration. Clinical studies have their own target populations defined by the inclusion and exclusion criteria. When these studies are combined with national disease registries, a patient in a study is also in a national database. Duplicated records are difficult to identify in practice because key identifiers such as names and addresses are often not disclosed for privacy protection in public health data. Instead, the membership of selected items in the final sample is assumed known (e.g., the selected item i from source $\mathscr{V}^{(j)}$ is also known to belong to $\mathscr{V}^{(k)}$). This is plausible because one can compare inclusion and exclusion criteria. For more detailed discussion on practical issues of our setting, see SAEGUSA (2019).

The final sample is a biased and dependent sample with duplication. There are two sources of bias in our setting. Certain data sources are over/under-represented in the final sample due to biased sampling with different selection probabilities $\pi^{(j)}$. Duplicated records from overlapping data sources enter statistical analysis without identification. Dependence also comes from two sources. Multiple data sets are dependent through duplicated records while items in the same data source are dependent due to sampling without replacement. These characteristics well capture the challenging issue of heterogeneity in data integration problems. Our framework covers the number of examples including opinion polls (BRICK et al., 2006), public health surveillance (HU et al., 2011), and health interview surveys (CERVANTES et al., 2006), and the synthesis of existing clinical and epidemiological studies with surveys, disease registries, and healthcare databases (CHATTERJEE et al., 2016; KEIDING and LOUIS, 2016; METCALF and SCOTT, 2009).

In this paper, we propose and study a nonparametric estimator of the distribution

function F. Our estimator is motivated by Hartley's estimator for multiple-frame surveys in sampling theory (HARTLEY, 1962, 1974). We provide a rigorous asymptotic theory to its uniform consistency over the real line and weak convergence to a Gaussian process. Based on the limiting distribution, we propose a Monte Carlo based method to construct confidence bands for F. We verify the validity of our methodology theoretically and through a simulation study for both continuous and discrete random variables.

Recently SAEGUSA (2019) studied the same data integration setting and derived the law of large numbers and the central limit theorem. Asymptotic results are then applied to infinite-dimensional M-estimation to study the Cox proportional hazards model (COX, 1972). These results are useful to compute the limiting distribution of our estimator but not sufficient for constructing confidence bands.

Typically, confidence bands for F are obtained from a rather simple limiting distribution or bootstrap. In the i.i.d. setting, the Kolmogorov-Smirnov statistic is used to compute confidence bands for continuous random variables. Its limiting distribution is the supremum of Brownian bridge, whose quantile is analytically obtained (KOLMOGOROV, 1933; SMIRNOV, 1944). For non-continuous random variables, confidence bands can be obtained by inverting the Dvoretzky–Kiefer–Wolfowitz inequality (DVORETZKY et al., 1956) with a tight constant obtained by MASSART (1990). An alternative way explored by BICKEL and FREEDMAN (1981) is to bootstrap the Kolmogorov-Smirnov statistic to estimate its quantiles. For stratified sampling from a finite population where X_i is treated as fixed, BICKEL and KRIEGER (1989) apply bootstrap methods for finite population sampling to the weighted Kolmogorov-Smirnov statistic to obtain valid confidence bands. These bootstrap methods cover the distribution function for non-continuous random variables.

In our data integration setting, randomness comes from (1) sampling from population and (2) subsequent sampling from data sources. A valid confidence band should reflect both types of uncertainty. The previous methods described above focus on randomness due to either sampling from population or finite population sampling, and cannot be applied to our data integration problem. The corresponding limiting distribution in our setting is the supremum of the linear combination of independent Gaussian processes. This process cannot be reduced to other well-known processes in general. Also, our formulation of the data integration problem is rather new and a valid bootstrap method is not available.

Methods for confidence bands for the distribution function have been studied in various ways other than analytical computation of quantiles of the limiting distribution and bootstrap. Confidence bands for parametric models are considered for normal distributions (KANOFSKY and SRINIVASAN, 1972), Weibull distributions (SCHAFER and ANGUS, 1979), and the location scale parameter model (CHENG and ILES, 1983). Bayesian approach with the Dirichlet prior was studied by BRETH (1978). OWEN (1995) considered inverting a nonparametric likelihood test of uniformity by BERK and JONES (1978). FREY (2008) proposed the narrowness criterion to derive optimal confidence bands. WANG et al. (2013) developed a smooth confidence band based on the kernel smoothed estimator of a distribution function.

The rest of the paper is organized as follows. In Section 2, we introduce our esti-

mator of F and derive its limiting distribution. We present the algorithm to compute the confidence band and study its asymptotic property in Section 3. We extend our methodology to conditional distribution functions in Section 4. The performance of the proposed method is evaluated through a simulation study in Section 5. We discuss a data example from the national Wilms tumor study in Section 6. All proofs are deferred to the appendix.

2. Estimator and its asymptotic properties

We introduce additional notation for our estimator. Let $R_i^{(j)} \in \{0,1\}$ be the selection indicator from data source $\mathscr{V}^{(j)}$. The item i has a vector of selection indicators $R_i = (R_i^{(1)}, \ldots, R_i^{(J)})$ but $R_i^{(j)} = 0$ if the item i does not belongs to source $\mathscr{V}^{(j)}$. For the items i in data source j (i.e., $V_i \in \mathscr{V}^{(j)}$), $R_i^{(j)}$s follow the distribution of sampling without replacement where $n^{(j)}$ is selected out of $N^{(j)}$. Since data collection procedures are carried out independently, selection indicators $(R_1^{(j)}, \ldots, R_N^{(j)})$ and $(R_1^{(k)}, \ldots, R_N^{(k)})$ are conditionally independent given $V_1, \ldots, V_N$ if $j \neq k$. For $V \in \mathscr{V}^{(j)}$, we assume the selection probability $\pi^{(j)}(V) = n^{(j)}/N^{(j)}$ converges to $p^{(j)} > 0$ as $N \to \infty$. We write the membership probability in source $\mathscr{V}^{(j)}$ as $\nu^{(j)} = P(V \in \mathscr{V}^{(j)})$ and the conditional expectation given membership in source $\mathscr{V}^{(j)}$ as $E^{(j)}$.

The desirable properties that an estimator of F in our data integration setting should satisfy are that (1) the estimator corrects bias due to biased sampling and duplication, and that (2) the estimator is computable without identification of duplicated records. To describe our estimator, we begin with $J = 2$ data sources. The key component of our estimator is

$$\rho(v) = (\rho^{(1)}(v), \rho^{(2)}(v)) \equiv \begin{cases} (1,0) & \text{if} \quad v \in \mathscr{V}^{(1)} \text{ and } v \notin \mathscr{V}^{(2)}, \\ (0,1) & \text{if} \quad v \notin \mathscr{V}^{(1)} \text{ and } v \in \mathscr{V}^{(2)}, \\ (c^{(1)}, c^{(2)}) & \text{if} \quad v \in \mathscr{V}^{(1)} \cap \mathscr{V}^{(2)}, \end{cases}$$

for positive constants $c^{(1)}, c^{(2)}$ with $c^{(1)} + c^{(2)} = 1$. The evaluation of this function only requires the membership in the mutually exclusive subsets of $\mathscr{V}$ based on data sources $\mathscr{V}^{(1)}$ and $\mathscr{V}^{(2)}$. We can compute the value of ρ for selected items because we assume information on data source membership is available for selected items. The choice of ρ is at the disposal of a data analyst. The optimal choice of ρ is considered by SAEGUSA (2019) and we use them in a simulation study and data example below.

Using the function ρ, we propose the following estimator of F given by

$$\mathbb{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) \right) I\{X_i \leq x\}.$$

Here we use the convention $0/0 = 0$ for the inverse probability weighting $R^{(j)}/\pi^{(j)}(V)$. This estimator is unbiased for F because inverse probability weighting $R^{(j)}/\pi^{(j)}(V)$ has conditional expectation 1 given $V_1, \ldots, V_N$ and $X_1, \ldots, X_N$ and because $\rho^{(1)}(v) + \rho^{(2)}(v) =$

1 for every v. Moreover, the estimator can be computed separately based on two subsamples through the expression

$$\mathbb{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i^{(1)}\rho^{(1)}(V_i)}{\pi^{(1)}(V_i)} I\{X_i \le x\} + \frac{1}{N}\sum_{i=1}^{N} \frac{R_i^{(2)}\rho^{(2)}(V_i)}{\pi^{(2)}(V_i)} I\{X_i \le x\}.$$

The proposed estimator can be considered as the weighted empirical distribution with weights computed from the selection probability and the function ρ. A difference from the empirical distribution is that our estimator may not have $\mathbb{F}_N(x) = 1$ for x greater than the largest selected X_i unless all the items i in $\mathcal{V}^{(1)} \cap \mathcal{V}^{(2)}$ selected from source $\mathcal{V}^{(1)}$ are also selected from $\mathcal{V}^{(2)}$. If $\mathbb{F}_N(x) > 1$ we can modify our estimator to $\tilde{\mathbb{F}}_N(x) = \min\{\mathbb{F}_N(x), 1\}$. For brevity of the presentation, we study $\mathbb{F}_N(x)$ but all properties below are satisfied for $\tilde{\mathbb{F}}_N(x)$.

The extension to more than two data sources is straightforward. Let $\rho = (\rho^{(1)}, \ldots, \rho^{(J)}) : \mathcal{V} \mapsto [0,1]^J$ where

$$\rho^{(j)}(v) = \begin{cases} 1, & v \in \mathcal{V}^{(j)} \cap \left(\cup_{m \ne j} \mathcal{V}^{(m)}\right)^c, \\ c^{(j)}_{k_1,\ldots,k_l}, & v \in \mathcal{V}^{(j)} \cap \left(\cap_{m=1}^{l} \mathcal{V}^{(k_m)}\right) \cap \left(\cup_{m \notin \{j,k_1,\ldots,k_l\}} \mathcal{V}^{(m)}\right)^c, \\ 0, & v \notin \mathcal{V}^{(j)}, \end{cases}$$

with $j, k_1, \ldots, k_l$ all different and $\sum_{j=1}^{J} \rho^{(j)}(v) = 1$. The proposed estimator is

$$\mathbb{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} \frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} I\{X_i \le x\}.$$

Now, we develop asymptotic properties of our estimator. As the uniform consistency of the empirical distribution follows from the Glivenko-Cantelli theorem, the uniform consistency for our estimator follows from the uniform law of large numbers for data integration (SAEGUSA, 2019).

Theorem 2.1. *The estimator $\mathbb{F}_N$ is uniformly consistent for F over $\mathbb{R}$. That is,*

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_N(x) - F(x)| \to_P 0.$$

As the Donsker theorem yields the weak convergence of the empirical distribution to the Brownian bridge process, the weak convergence for our estimator follows from the uniform central limit theorem for data integration (SAEGUSA, 2019). Its limiting distribution is still a Gaussian process, but not the Brownian bridge process.

Theorem 2.2. *Let $D(\mathbb{R})$ be the class of cadlag functions on $\mathbb{R}$. Our estimator $\sqrt{N}(\mathbb{F}_N - F)$ weakly converges to the Gaussian process $\mathbb{G}$ in $D(\mathbb{R})$ given by*

$$\mathbb{G} = \mathbb{G}_0 + \sum_{j=1}^{J} \sqrt{\nu^{(j)}} \sqrt{\frac{1-p^{(j)}}{p^{(j)}}} \mathbb{G}_j,$$

where $\mathbb{G}_j, j = 0,1,\ldots,J$, *are independent Gaussian processes with covariance functions*

$$\begin{aligned} k^{(0)}(s,t) &= F(s\wedge t) - F(s)F(t), \\ k^{(j)}(s,t) &= E^{(j)}\left[\left\{\rho^{(j)}(V)\right\}^2 I\{X \le s\wedge t\}\right] \\ &\quad - E^{(j)}\left[\rho^{(j)}(V) I\{X \le s\}\right] E^{(j)}\left[\rho^{(j)}(V) I\{X \le t\}\right], \end{aligned}$$

for $s,t \in \mathbb{R}$ *and* $j = 1,\ldots,J$.

An immediate consequence of this theorem is that $\sqrt{N}(\mathbb{F}_N(x) - F(x))$ converges in distribution to the zero-mean normal random variable with variance as the sum of $P(X \le x)\{1 - P(X \le x)\}$ and

$$\sum_{j=1}^{J}\left(\nu^{(j)}\frac{1-p^{(j)}}{p^{(j)}}E^{(j)}\left[\left\{\rho^{(j)}(V)\right\}^2 I\{X \le x\}\right] - \left\{E^{(j)}\left[\rho^{(j)}(V) I\{X \le x\}\right]\right\}^2\right).$$

Note that $P(X \le x)\{1 - P(X \le x)\}$ is asymptotic variance which we would obtain from the analysis of i.i.d. data. Merging samples from overlapping sources increases additional uncertainty in our estimator. If we select all items from each source without identifying duplication, then $p^{(j)} = 1, j = 1,\ldots,J$, yield the same variance as in the i.i.d. case. Hence, we see that the additional variance comes from additional selection, not duplication. The effect of duplication appear only through the variable $\rho^{(j)}(V)$. Uncertainty in large data source (i.e., $\nu^{(j)} = P(V \in \mathscr{V}^{(j)})$ contributes more to the asymptotic variance.

3. Confidence band

The basic idea to obtain a confidence band is to obtain $q_{1-\alpha}$ such that

$$P\left(\sup_{x\in\mathbb{R}} \sqrt{N}\,|\mathbb{F}_N(x) - F(x)| \le q_{1-\alpha}\right) \to 1-\alpha, \quad n \to \infty,$$

from which the large sample $100(1-\alpha)\%$ confidence band is obtained as

$$\mathbb{F}_N(x) - q_{1-\alpha}/\sqrt{N} \le F(x) \le \mathbb{F}_N(x) + q_{1-\alpha}/\sqrt{N}, \quad \text{all } x \in \mathbb{R}.$$

One potential approach is to use an analytical expression of quantiles of the limiting distribution of $\sup_{x\in\mathbb{R}} \sqrt{N}\,|\mathbb{F}_N(x) - F(x)|$ but this limiting distribution $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$ obtained from Theorem 2.1 is the supremum of the complicated Gaussian process whose quantiles cannot be analytically derived in general. Another approach is to estimate $q_{1-\alpha}$ by nonparametrically bootstrapping $\sup_{x\in\mathbb{R}} \sqrt{N}\,|\mathbb{F}_N(x) - F(x)|$ but there is no known valid bootstrap method for our setting. Generating data from $\mathbb{F}_N$ would be another alternative but it is not clear how to simultaneously generate V to mimic the data integration process.

The proposed methodology does not analytically compute $q_{1-\alpha}$ from the limiting distribution nor simulating data generating mechanism. Instead, we directly simulate

the limiting distribution to estimate its quantiles. The distribution of the zero-mean Gaussian process $\mathbb{G}$ is completely determined by the unknown covariance function

$$k(s,t) = k^{(0)}(s,t) + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} k^{(j)}(s,t).$$

We estimate this covariance function $k(s,t)$ as follows. For data source membership probability $\nu^{(j)}$ and selection probability $p^{(j)}$, we estimate them by $N^{(j)}/N$ and $n^{(j)}/N^{(j)}$ respectively. For $k^{(0)}(s,t)$, an obvious estimator is $\mathbb{F}_N(s \wedge t) - \mathbb{F}_N(s)\mathbb{F}_N(t)$. For $k^{(j)}(s,t)$, conditional expectations given membership in $\mathscr{V}^{(j)}$ are estimated by inverse probability weighting based on a sample selected from source $\mathscr{V}^{(j)}$ (i.e., items i with $R_i^{(j)} = 1$). Specifically, the first term in $k^{(j)}(s,t)$ is estimated by

$$\frac{1}{N^{(j)}} \sum_{i=1}^{N} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \{\rho^{(j)}(V_i)\}^2 I\{X_i \le s \wedge t\},$$

and the second term in $k^{(j)}(s,t)$ is estimated by

$$\left\{ \frac{1}{N^{(j)}} \sum_{i=1}^{N} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i) I\{X_i \le s\} \right\} \left\{ \frac{1}{N^{(j)}} \sum_{i=1}^{N} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i) I\{X_i \le t\} \right\}.$$

We denote our estimator of $k(s,t)$ by $\hat{k}_N(s,t)$.

The zero-mean Gaussian process $\hat{\mathbb{G}}_N$ with covariance function $\hat{k}_N(s,t)$ weakly converges to the limiting process $\mathbb{G}$. However, the supremum of $|\mathbb{G}(x)|$ may have a jump at the lower end of the support of X (TSIRELSON, 1975). To avoid the possibility that the jump occurs at its $100(1-\alpha)$%tile, we assume the following condition. The same condition is imposed by BICKEL and KRIEGER (1989) for finite population sampling.

Condition 3.1. *The distribution of* $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$ *is continuous.*

Under this condition, we have the following result.

Theorem 3.1. *Let* $q \in \mathbb{R}$. *Let* $\hat{\mathbb{G}}_N$ *be the zero-mean Gaussian process with covariance function* $\hat{k}_N(s,t)$, *Under Condition 3.1, as* $N \to \infty$,

$$P\left(\sup_{x\in\mathbb{R}} |\hat{\mathbb{G}}_N(x)| \le q \right) \to P\left(\sup_{x\in\mathbb{R}} |\mathbb{G}(x)| \le q \right).$$

We propose the following procedure to construct a confidence band of F:

- Generate the first zero-mean Gaussian process $\hat{\mathbb{G}}_N$ with covariance function $\hat{k}_N(s,t)$, and compute the supremum s_1 of $|\hat{\mathbb{G}}_N|$
- Repeat this procedure B times to obtain $s_1, \ldots, s_B$, and compute their $100(1-\alpha)$%tile $\hat{q}_{1-\alpha}$.

- Compute the $100(1-\alpha)\%$ confidence band of F by

$$\mathbb{F}_N(x) - \hat{q}_{1-\alpha}/\sqrt{N} \le F(x) \le \mathbb{F}_N(x) + \hat{q}_{1-\alpha}/\sqrt{N}, \quad \text{all } x \in \mathbb{R}. \tag{1}$$

The proposed confidence band has the correct coverage probability asymptotically.

Theorem 3.2. *Under Condition 3.1, as* $N, B \to \infty$,

$$P\left(\mathbb{F}_N(x) - \hat{q}_{1-\alpha}/\sqrt{N} \le F(x) \le \mathbb{F}_N(x) + \hat{q}_{1-\alpha}/\sqrt{N}, \quad \textit{all } x \in \mathbb{R}\right) \to 1-\alpha.$$

4. Extension to conditional distribution given discrete variables

In practice, it is of interest to compare different groups through graphical comparison of distribution functions. An extension of our method to conditional distributions given a discrete random variable is straightforward. Let U be a discrete random variable. First, we estimate the sub-distribution function $F(x,u) = P(X \le x, U = u)$ by

$$\mathbb{F}_N(x,u) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)\right) I\{X_i \le x, U_i = u\}.$$

The limiting distribution is similar to the one in Theorem 2.2 but covariance functions are now

$$\begin{aligned} k_u^{(0)}(s,t) &= P(X \le s \wedge t, U = u) - P(X \le s, U = u)P(X \le t, U = u), \\ k_u^{(j)}(s,t) &= E^{(j)}\left[\left\{\rho^{(j)}(V)\right\}^2 I\{X \le s \wedge t, U = u\}\right] \\ &\quad - E^{(j)}\left[\rho^{(j)}(V) I\{X \le s, U = u\}\right] E^{(j)}\left[\rho^{(j)}(V) I\{X \le t, U = u\}\right]. \end{aligned}$$

This covariance function can be similarly estimated and the same procedure described above yields the confidence band given by

$$\mathbb{F}_N(x,u) - \hat{q}_{1-\alpha,u}/\sqrt{N} \le F(x,u) \le \mathbb{F}_N(x,u) + \hat{q}_{1-\alpha,u}/\sqrt{N}, \quad \text{all } x \in \mathbb{R}.$$

Since $F(x|u) = P(X \le x|U = u) = P(X \le x, U = u)/P(U = u)$, we estimate $p_u = P(U = u)$ by a consistent estimator

$$\hat{p}_u = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)\right) I\{U = u\}.$$

Now we propose the confidence band for $F(x|u)$ given by

$$\frac{\mathbb{F}_N(x,u)}{\hat{p}_u} - \frac{\hat{q}_{1-\alpha,u}}{N^{1/2}\hat{p}_u} \le F(x|u) \le \frac{\mathbb{F}_N(x,u)}{\hat{p}_u} + \frac{\hat{q}_{1-\alpha,u}}{N^{1/2}\hat{p}_u}, \quad \text{all } x \in \mathbb{R}. \tag{2}$$

Table 1: Sample sizes for three scenarios.

	Scenario 1 & 2			Scenario 3		
N	100	250	500	100	250	500
$N^{(1)}$	79	197	395	78	197	395
$N^{(2)}$	51	127	255	51	127	255
$N^{(3)}$				28	70	141
$n^{(1)}$	16	40	80	16	40	80
$n^{(2)}$	16	38	77	16	38	77
$n^{(3)}$				14	35	71
Duplication (2 sources)	2	5	9	6	2	27
Duplication (3 sources)				0	1	1

This confidence band has the correct coverage probability asymptotically. The proof is similar to that of Theorem 3.2, and omitted.

5. Simulation study

We carry out a simulation study to evaluate the finite-sample performance of the proposed confidence band. We consider three different scenarios. The first two scenarios concern two partially overlapping data sources. The third scenario deals with three data sources with one data source contained in other two. The distributions considered are mixtures of beta distributions, Poisson distributions, and normal distributions, respectively.

In the first scenario, the variable Y is a Bernoulli random variable with $p = 0.3$. The variable X of interest follows the beta distribution with $\alpha = 5$ and $\beta = 2$ if $Y = 0$ and the beta distribution with $\alpha = 2$ and $\beta = 5$ if $Y = 1$. The variables $W = (X, Y)$ are not available at the first stage of sampling. The auxiliary binary variable V is correlated with Y with sensitivity 0.9 and specificity 0.9. Data sources are created by values of V. If $V = 0$, the item belongs to data source 1 and if $V = 1$ it belongs to data source 2. In both situations, the item belongs to the intersection of two data sources with probability 0.3. Selection probabilities are 0.2 from data source 1 and 0.3 from data source 2. The second scenario is the same as the first except that the variable X follows the Poisson distribution with $\lambda = 2$ if $Y = 0$ and the Poisson distribution with $\lambda = 4$ if $Y = 1$. In the third scenario, variables Y and V and data sources 1 and 2 are similarly generated as in the other two cases. The variable X follows the normal distribution with $\mu = 1$ and $\sigma^2 = 1$ if $Y = 0$ and the normal distribution with $\mu = 3$ and $\sigma^2 = 1$ if $Y = 1$. The data source 3 consists of items with $X \in [1, 2]$. Selection probabilities are 0.2 from data source 1, 0.3 from data source 2, and 0.5 from data source 3.

Data were generated 500 times in each scenario with sample size $N = 100$, $N = 250$, and $N = 500$. In each data set, the 95% confidence band was constructed based on 2000 simulated Gaussian processes with the formula (1). Table 1 summarizes average sample sizes for each data source before and after the selection into the final sample. Note that the proposed estimator is based on 30 items for scenarios 2 and 3, and 40

Table 2: Simulated coverage probabilities for the confidence bands.

	Scenario 1		Scenario 2		Scenario 3	
	Coverage	Width	Coverage	Width	Coverage	Width
$N = 100$	0.940	0.454	0.936	0.442	0.920	0.464
$N = 250$	0.944	0.295	0.954	0.286	0.956	0.304
$N = 500$	0.952	0.211	0.944	0.203	0.956	0.217

items for scenario 3 on average without duplication when $N = 100$. Table 2 shows simulated coverage probabilities and average width based on 500 data simulated data sets. Coverage probabilities are close to the nominal level in all scenarios when N is greater than 250 while we see under-coverage when $N = 100$. Confidence bands are wide for $N = 100$ but the width becomes reasonable as N increases. Overall, our methodology shows reasonable performance for a practical use.

6. Application

We illustrate the proposed method using data from the national Wilms tumor study (D'ANGIO et al., 1989). Wilms tumor is a rare kidney cancer for children. The predictor of relapse includes histology of cancer, age at diagnosis, and tumor diameter. Data for all 3915 patients are available and were used to compare different designs (BRESLOW and CHATTERJEE, 1999; BRESLOW et al., 2009; SAEGUSA, 2019). In our analysis, we check if the empirical distributions based on the entire cohort are contained in the proposed confidence bands based on a smaller biased sample with duplication. Three data source are deceased patients, patients with unfavorable histology measured at the hospital, and the entire cohort. Selection probabilities 100%, 50%, and 10%, respectively, yielding the sample size 1027 in the final sample (885 patients without duplication). For selected patients, tumor diameter is measured and histology is re-examined at the central reference laboratory. Our goal is to create two distribution functions of tumor diameter based on the histology information measured at the second time. Among selected patients, 646 (603 without duplication) patients have favorable histology and 382 (282 without duplication) patients have unfavorable histology.

Figure 2 shows the confidence bands for the conditional distributions of tumor diagmeter given histology based on the formula (2). The solid line is smoothed empirical distribution based on the entire cohort of size 3915. Our estimators are close to empirical distributions. Moreover, the proposed confidence bands successfully contain empirical distributions. The difference in sample sizes based on histology is reflected in the difference of widths. The confidence band for favorable histology has width 0.133 while the band for unfavorable histology has width 0.307. Graphical comparison of both estimators with the help of confidence bands shows that there is no striking difference between distributions of tumor diameter in different histology groups. In fact, empirical quartiles of tumor diameter for both groups agree well. A similar analysis (not shown here) conditional on survival status led to the same conclusion. In the proportional hazards regression analysis, SAEGUSA (2019) shows that tumor diameter has a small effect on

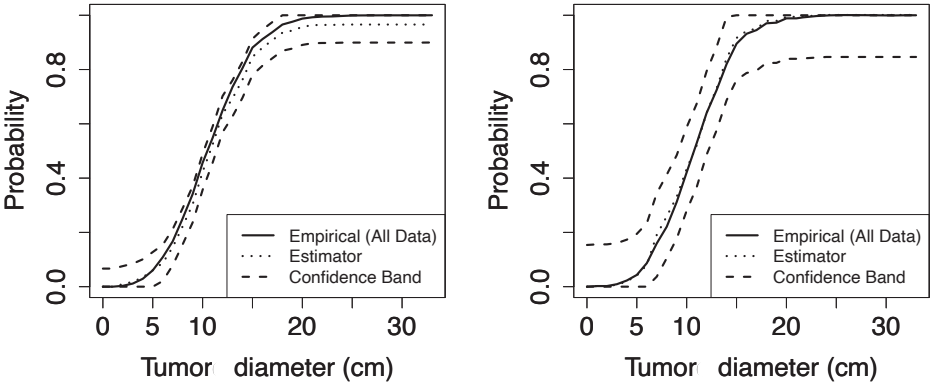

Figure 2: Confidence bands for conditional distribution functions of tumor diameter given favorable histology (left panel) and unfavorable histology (right panel).

tumor relapse while histology is statistically significant.

Acknowledgements

We thank Partha Lahiri for helpful discussions and encouragement for this project.

References

BERK, R. H. JONES, D. H., (1978). Relatively optimal combinations of test statistics. *Scand. J. Statist.*, 5(3), pp. 158–162.

BICKEL, P. J. FREEDMAN, D. A., (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6), pp,1196–1217.

BICKEL, P. J. KRIEGER, A. M., (1989). Confidence bands for a distribution function using the bootstrap. *J. Amer. Statist. Assoc.*, 84(405), pp. 95–100.

BRESLOW, N. E. CHATTERJEE, N., (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4), pp. 457–468.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C., CHAMBLESS, L., KULICH, M., (2009). Using the whole cohort in the analysis of case-cohort data. *American J. Epidemiol.*, 169, pp. 1398–1405.

BRETH, M., (1978). Bayesian confidence bands for a distribution function. *Ann. Statist.*, 6(3), pp. 649–657.

BRICK, J. M., DIPKO, S., PRESSER, S., TUCKER, C., YUAN, Y., (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *The Public Opinion Quarterly*, 70(5), pp. 780–793.

CERVANTES, I., JONES, M., ROJAS, L., BRICK, J., KURATA, J., GRANT, D., (2006). A review of the sample design for the california health interview survey. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 3023–3030.

CHATTERJEE, N., CHEN, Y.-H., MAAS, P., CARROLL, R. J., (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Amer. Statist. Assoc.*, 111(513), pp. 107–117.

CHENG, R. C. H. ILES, T. C., (1983). Confidence bands for cumulative distribution functions of continuous random variables. *Technometrics*, 25(1), pp.77–86.

COX, D. R., (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34, pp. 187–220.

D'ANGIO, G. J., BRESLOW, N., BECKWITH, J. B., EVANS, A., BAUM, H., DE-LORIMIER, A., FERNBACH, D., HRABOVSKY, E., JONES, B., KELALIS, P., (1989). Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. *Cancer*, 64(2), pp. 349–360.

DVORETZKY, A., KIEFER, J., WOLFOWITZ, J., (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27, pp. 642–669.

FREY, J., (2008). Optimal distribution-free confidence bands for a distribution function. *J. Statist. Plann. Inference*, 138(10), pp. 3086–3098.

GINÉ, E. NICKL, R., (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

HARTLEY, H. O., (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203–206.

HARTLEY, H. O., (1974). Multiple frame methodology and selected applications. *Sankhyā Ser. C*, 36, pp. 99–118.

HU, S. S., BALLUZ, L., BATTAGLIA, M. P., FRANKEL, M. R., (2011). Improving public health surveillance using a dual-frame survey of landline and cell phone numbers. *American Journal of Epidemiology*, 173(6), pp. 703–711.

KANOFSKY, P. SRINIVASAN, R., (1972). An approach to the construction of parametric confidence bands on cumulative distribution functions. *Biometrika*, 59, pp. 623–631.

KEIDING, N. LOUIS, T. A., (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2), pp. 319–376.

KOLMOGOROV, A. N., (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, pp. 83–91.

MASSART, P., (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3), pp. 1269–1283.

METCALF, P. SCOTT, A., (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28(10), pp. 1512–1523.

OWEN, A. B., (1995). Nonparametric likelihood confidence bands for a distribution function. *J. Amer. Statist. Assoc.*, 90(430), pp. 516–521.

SAEGUSA, T., (2019). Large sample theory for merged data from multiple sources. *Ann. Statist.*, 47(3), pp. 1585–1615.

SAEGUSA, T. WELLNER, J. A., (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.*, 41(1), pp. 269–295.

SCHAFER, R. E. ANGUS, J. E., (1979). Estimation of weibull quantiles with minimum error in the distribution function. *Technometrics*, 21(3), pp. 367–370.

SMIRNOV, N. V., (1944). Approximate laws of distribution of random variables from empirical data. *Uspehi Matem. Nauk*, 10, pp. 179–206.

TSIRELSON, V. S., (1975). The density of the distribution of the maximum of a Gaussian process. *Theory of Probability and its Applications*, 20, pp. 847–865.

WANG, J., CHENG, F., YANG, L., (2013). Smooth simultaneous confidence bands for cumulative distribution functions. *J. Nonparametr. Stat.*, 25(2), pp. 395–407.

APPENDIX

Proof of Theorem 2.1. Because the class of functions $\mathscr{F} = \{f_t(x) = I(x \le t) : t \in \mathbb{R}\}$ is a Glivenko-Cantelli class, apply the uniform law of large numbers for data integration (Theorem 3.1 of SAEGUSA (2019)) to $\mathscr{F}$ to obtain the desired result. □

Proof of Theorem 2.2. Because the class of functions $\mathscr{F} = \{f_t(x) = I(x \le t) : t \in \mathbb{R}\}$ is also a Donsker class, apply the uniform central limit theorem for data integration (Theorem 3.2 of SAEGUSA (2019)) to $\mathscr{F}$. The computation of the covariance function is straightforward. □

Proof of Theorem 3.1. We show the weak convergence of $\hat{\mathbb{G}}_N$ to $\mathbb{G}$. First we consider the finite dimensional convergence of $\hat{\mathbb{G}}_N$ to $\mathbb{G}$. As in the proof of Theorem 2.1, the law of large numbers for data integration yields

$$\sup_{s,t\in\mathbb{R}} |\mathbb{F}_N(s \wedge t) - \mathbb{F}_N(s)\mathbb{F}_N(t) - k^{(0)}(s,t)| \to_P 0.$$

For $k^{(j)}(s,t), j = 1,\ldots,J$, the law of large numbers for sampling without replacement (Theorem 5.1 of SAEGUSA and WELLNER (2013)) yields the uniform consistency over $s,t \in \mathbb{R}$. Since $n^{(j)}/N^{(j)} \to p^{(j)}$ by assumption and $N^{(j)}/N \to_P \nu^{(j)}$ by the weak law of large numbers, we conclude

$$\sup_{s,t\in\mathbb{R}} |\hat{k}_N(s,t) - k(s,t)| \to_P 0.$$

This implies the desired finite dimensional convergence.

Second, we consider asymptotic equicontinuity and total boundedness of $\mathbb{R}$ with respect to a constant multiple of

$$d^{(0)}(s,t) = k^{(0)}(s,s) + k^{(0)}(t,t) - 2k^{(0)}(s,t).$$

Note that the intrinsic metric $d(s,t) = k(s,s) + k(t,t) - 2k(s,t)$ to the limiting process $\mathbb{G}$ is equivalent to $d^{(0)}(s,t)$ (i.e.,$C_1 d(s,t) \le d^{(0)}(s,t) \le C_2 d(s,t)$ for some constants $C_1, C_2 > 0$) because $\rho^{(j)}(v)$ is bounded. Also, on the event A that $\sup_{s,t\in\mathbb{R}} |\hat{k}_N(s,t) - k(s,t)| < C_3$ for some small fixed constant $C_3 > 0$, $\hat{d}_N(s,t) = \hat{k}_N(s,s) + \hat{k}_N(t,t) - 2\hat{k}_N(s,t)$ is equivalent to $d(s,t)$ since $d(s,t)$ is bounded over $\mathbb{R}^2$. These observations imply that the process $\mathbb{G}$ and $\hat{\mathbb{G}}_N$ are sub-Gaussian processes with respect to $Cd^{(0)}(s,t)$ for some constant $C > 0$ on the event A. As a consequence, the property of the sub-Gaussian process (see e.g. Theorem 2.3.7 of GINÉ and NICKL (2016)) implies that

$$E\left[\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \varepsilon \middle| A\right] \le K \int_0^\delta \sqrt{\log 2N(\mathbb{R}, Cd^{(0)}, \varepsilon)}\, d\varepsilon \tag{3}$$

for some constant $K > 0$ as long as the integral on the right hand side is finite. Here $N(\mathbb{R}, Cd^{(0)}, \varepsilon)$ is the covering number of $\mathbb{R}$ with respect to the metric $Cd^{(0)}$ with radius ε.

For asymptotic equicontinuity, let $\eta > 0$ be arbitrary. We have

$$\limsup_{n\to\infty} P\left(\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \eta \right)$$
$$\le \limsup_{n\to\infty} P\left(\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \eta, A \right) + P(A^c).$$

where A^c is the complement of A. Since $P(A^c) \to 0$, we bound the first term by the Markov inequality and the inequality (3) to obtain

$$\limsup_{n\to\infty} P\left(\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \eta \,\middle|\, A \right) P(A)$$
$$\le \limsup_{n\to\infty} \eta^{-1} K \int_0^\delta \sqrt{\log 2N(\mathbb{R}, Cd^{(0)}, \varepsilon)}\, d\varepsilon \to 0, \quad \text{as } \delta \downarrow 0,$$

assuming the integral on the right hand side is finite for any δ, which we will show next.

To compute the covering number with radius ε, create l subintervals $[I_i, I_{i+1}]$ of [0,1] with length less than ε with $I_0 = 0 < I_1 < \cdots < I_{l+1} = 1$. Note that we do not consider $\varepsilon \ge 1$ since we take $\delta \downarrow 0$. Let $q_i = F^{-1}(I_i)$. Then $F(q_{i+1}) - F(q_i) \le \varepsilon$. If $t \in [I_i, I_{i+1})$, we have

$$d^{(0)}(q_i, t) = F(q_i)\{1 - F(q_i)\} + F(t)\{1 - F(t)\} - 2\{F(q_i) - F(q_i)F(t)\} \le 4\varepsilon.$$

This means t is in the $d^{(0)}$-ball with center q_i and radius 4ε. This implies that the covering number with radius ε is proportional to $1/\varepsilon$, and hence the entropy integral converges. This computation also shows that $\mathbb{R}$ is totally bounded with respect to $d^{(0)}$. Because asymptotic equicontinuity and total boundedness imply asymptotic tightness, we now conclude the weak convergence of $\hat{\mathbb{G}}_N$ to $\mathbb{G}$.

The continuous mapping theorem yields that $\sup_{x\in\mathbb{R}} |\hat{\mathbb{G}}_N(x)|$ converges in distribution to $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$. Thus, the desired result follows from Condition 3.1. □

Proof of Theorem 3.2. Theorem 2.2 and continuous mapping theorem imply $\sup_{x\in\mathbb{R}} \sqrt{N}|\mathbb{F}_N(x) - F(x)|$ converges in distribution to $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$. Theorem 3.1 implies that $\hat{q}_{1-\alpha}$ converges in probability to the $100(1-\alpha)$%tile $q_{1-\alpha}$ of $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$. Combining these results completes the proof. □