Journal of Multivariate Analysis 129 (2014) 95–109

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Using linear interpolation to reduce the order of the coverage error of nonparametric prediction intervals based on right-censored data

E. Beutner [a,∗], E. Cramer [b]

[a] Department of Quantitative Economics, Maastricht University, P.O. Box 616, NL-6200 MD Maastricht, The Netherlands

[b] Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany

ARTICLE INFO

Article history:
Received 11 March 2013
Available online 18 April 2014

AMS subject classifications:
62G20
62N01

Keywords:
Nonparametric prediction intervals
Right-censored data
Asymptotic refinements

ABSTRACT

We prove a general result showing that a simple linear interpolation between adjacent random variables reduces the coverage error of nonparametric prediction intervals for a future observation from the same underlying distribution function from $O(n^{-1})$ to $O(n^{-2})$. To illustrate the result we show that it can be applied to various scenarios of right censored data including Type-II censored samples, pooled Type-II censored data, and progressively Type-II censored order statistics. We further illustrate the result by simulations indicating that the desired level of significance is almost attained for moderate sample sizes.

1. Introduction

Interest in nonparametric prediction intervals traces back at least to Wilks [25]. Given two random variables X and Y, an interval of the form $(-\infty, Y]$ is called a (conservative) one-sided nonparametric prediction interval for X at the level p if

$$\mathbb{P}(X \le Y) \ge p.$$

Here, X represents the future observation we want to predict. If we observe an i.i.d. sample of size n from the same distribution as X, then Y can be chosen as an appropriate order statistic from the sample of size n (see, for example, [11] Section 7.3). Of course, we do not need a complete sample of size n to construct a nonparametric prediction interval for X. Guilbaud [12] has used progressively Type-II censored data to construct nonparametric prediction intervals. Balakrishnan et al. [3] have constructed nonparametric prediction intervals based on a sample obtained by pooling two independent Type-II right-censored samples. An efficient computational procedure for this type of data based on a branch and bound approach has been proposed in [4]. The multi-sample case has been addressed in [23]. An extension to doubly censored data has been established by Volterman et al. [24].

Before motivating our main result and to ease its motivation, we give a brief description of the models just mentioned.

Model 1 (*Type-II Censored Data*)**.** *Denote the order statistics of an i.i.d. sample $X_1, \ldots, X_n$ from the distribution function (df) F by $X_{1:n} \le \cdots \le X_{n:n}$. If the data consist only of the first $m(n)$, $m(n) < n$, order statistics $X_{1:n} \le \cdots \le X_{m(n):n}$ and the information that $X_{i:n} \ge X_{m(n):n}$, $m(n) < i \le n$, the sample is called Type-II right-censored.*

∗ Corresponding author.
E-mail addresses: e.beutner@maastrichtuniversity.nl (E. Beutner), erhard.cramer@rwth-aachen.de (E. Cramer).

http://dx.doi.org/10.1016/j.jmva.2014.04.007
0047-259X/

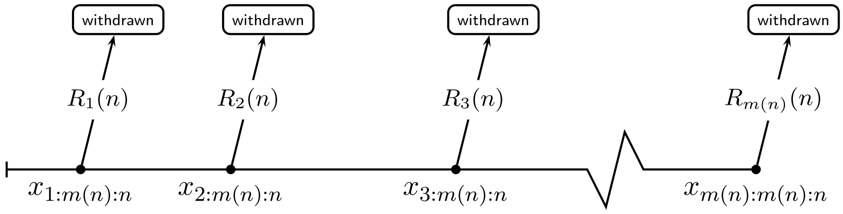

Fig. 1. Generation process of progressively Type-II censored order statistics.

Model 2 (*Progressively Type-II Censored Data*). *Progressively Type-II censored order statistics are defined as follows; (see, e.g., [2]): Given integers n and $m(n)$, $1 \le m(n) < n$, and a vector $\mathcal{R}(n) = (R_1(n), \dots, R_{m(n)}(n))$ of non-negative integers with $n = m(n) + \sum_{i=1}^{m(n)} R_i(n)$ the following censoring mechanism is employed. Suppose that n specimens with random lifetimes $X_1, \dots, X_n$ are put on a life test. Then, at the time of the first failure denoted by $X_{1:m(n):n} \equiv X_{1:n}$, $R_1(n)$ items are randomly drawn from the $n-1$ surviving units and withdrawn from the experiment. Then, the life test is continued with the remaining $n-1-R_1(n)$ specimens. At the time of the next failure $X_{2:m(n):n}$, the same procedure is carried out with $R_2(n)$ instead of $R_1(n)$. This censoring process is continued until $m(n)$ failure times $X_{1:m(n):n} \le \cdots \le X_{m(n):m(n):n}$ have been observed. This notation commonly used in the literature for this kind of ordered data motivates the notation in formula (3.1). Further, we refer to $\mathcal{R}(n) = (R_1(n), \dots, R_{m(n)}(n))$ as the censoring scheme. The generation procedure of progressively Type-II censored data is depicted in Fig. 1.*

It should be mentioned that a Type-II right censored sample can be seen as a progressively Type-II censored sample with censoring plan $\mathcal{R}(n) = (0, \dots, n - m(n))$. Thus, to unify the notation, we denote an order statistic $X_{j:n}$ also by $X_{j:m(n):n}$.

An important submodel of Model 2 is given by one-step progressive censoring schemes which turn out to be optimal design plans in several settings (see, e.g., [2,5,10,20]). Moreover, these plans are important for ordered pooled Type-II censored samples (see Section 3.4).

Model 3 (*Progressive Censoring with One-Step Censoring Plans*). *Progressive censoring with one-step censoring plans results by choosing the censoring plan $\mathcal{R}(n)$ such that $R_j(n)$ is positive for exactly one index $j \in \{1, \dots, m(n)\}$. This corresponds to a progressive censoring procedure where items are withdrawn at most at one failure time (cf. Fig. 1). For such a censoring procedure given n and $m(n)$, there is an index $\nu(n)$, $1 \le \nu(n) \le m(n)$, such that $R_{\nu(n)}(n) = n - m(n)$ and $R_j(n) = 0$ for $j \ne \nu(n)$. Such a censoring plan will be denoted by $\mathcal{O}_{\nu(n)}^{n-m(n)}$.*

Now, we start to motivate our main result given in Section 2. To do so, consider a Type-II right censored sample of size n from a continuous df F. Using the notation for Type-II censored order statistics explained at the end of the description of Model 2, we have

$$\mathbb{P}(X \le X_{i:m(n):n}) = \frac{i}{n+1}, \quad 1 \le i \le m(n), \tag{1.1}$$

where X also has df F and is independent of $X_{i:m(n):n}$, $1 \le i \le m(n)$. Eq. (1.1) is an immediate consequence of the probability integral transform and the expectation of uniform order statistics (see also [11] (7.3.2) with $r = m = 1$ and $t = i$). It is apparent from Eq. (1.1) that a conservative prediction interval at level p is given by $(-\infty, X_{\lceil (n+1)p\rceil:m(n):n}]$ (if $\lceil (n+1)p\rceil \le m(n)$; $\lceil x\rceil$ denotes the smallest integer exceeding x). Notice also that for Type-II censored data the maximum coverage probability is $m(n)/(n+1)$ which might be less than the desired level of p. A similar remark applies to the other censoring schemes discussed above. Moreover, it is easily seen that the difference between the level of $(-\infty, X_{\lceil (n+1)p\rceil:m(n):n}]$ and p is of order $O(n^{-1})$ (provided $\lceil (n+1)p\rceil \le m(n)$ at least for n large).

Similar results apply to other censoring schemes. For instance, for the one-step censoring plans of Model 3 with $m(n) = n-1$, we have for $1 \le \ell \le n-1$

$$\mathbb{P}(X \le X_{\ell:n-1:n}) = w(\ell)\frac{\ell}{n+1} + (1-w(\ell))\frac{\ell+1}{n+1}, \tag{1.2}$$

with an appropriate weight $0 \le w(\ell) \le 1$, $1 \le \ell \le n-1$. Eq. (1.2) follows from the mixture representation of $X_{\ell:n-1:n}$, which is an immediate consequence of the results of Guilbaud [13]. The difference between the level of $(-\infty, X_{\ell:n-1:n}]$ and p is also of order $O(n^{-1})$. Obviously, if the level of $(-\infty, X_{\ell-1:n-1:n}]$ is less than p and the level of $(-\infty, X_{\ell:n-1:n}]$ is larger than p, then the coverage accuracy of the 'linearly interpolated interval'

$$(-\infty, q \cdot X_{\ell-1:n-1:n} + (1-q) \cdot X_{\ell:n-1:n}], \quad 0 < q < 1,$$

might be better than the coverage accuracy of any of the two foregoing intervals.

However, the question is whether this improvement in coverage accuracy is significant. Beran and Hall [6] addressed this issue for the particular case when the data is an i.i.d. sample of size n. They proved that in this case the coverage error of the prediction interval $(-\infty, qX_{r:n} + (1-q)X_{r+1:n}]$ at the level p, where r is such that $r < p(n+1) \le r+1$