

## A TEST FOR SURVIVAL DISTRIBUTIONS USING DATA FROM COMPLEX SAMPLES

Susana Rubin Bleuer<sup>1</sup>

### ABSTRACT

The distribution of spell duration and other lifetime distributions may be estimated by applying survival analysis methods. Survival analysis methods were developed for independent identically distributed data (i.i.d.) from a super-population. When data is obtained from complex surveys the i.i.d. - assumption is not valid and standard results may not apply. In this paper, we develop a test to assess observed differences in survival curves, based on the Gehan-Wilcoxon Statistic, that can be applied to data from probability proportional to size- design.

KEY WORDS: Complex survey data; Survival analysis.

### RÉSUMÉ

La distribution des durées de périodes et d'autres distributions de temps de vie peuvent être estimées en appliquant des méthodes d'analyse de survie. Des méthodes d'analyse de survie ont été développées pour des données indépendantes identiquement distribuées (IID) d'une super-population. Quand des données sont obtenues à partir d'une méthodologie complexe de sondage, l'hypothèse des données IID n'est pas valide et les résultats standards ne peuvent pas s'appliquer. Dans cet article, nous développons un test, basé sur la statistique de Gehan-Wilcoxon, qui peut être appliqué à des données provenant de plans d'échantillonnage ave probabilité proportionnelle à la taille (PPT).

KEY WORDS: Analyse de survie; méthodologie complexe de sondage.

### 1. INTRODUCTION

The Survey of Labor and Income Dynamics (SLID) conducted by Statistics Canada provides, among other type of information, data on the movements into and out of unemployment. SLID is a longitudinal complex survey and the duration of the spells of unemployment provided by SLID is censored failure time data. There are many other examples of lifetime data from longitudinal complex surveys at Statistics Canada, like data obtained from the Workforce and Employee Survey (WES), the Longitudinal Administrative Database (LAD), etc. Here we develop a methodology for survival analysis with this type of data.

We might need, for example, to use data from SLID to compare the distribution of the length of unemployment spells in two different provinces or regions. A variety of non-parametric significance tests are used to assess observed differences in the empirical survival curves. One common non-parametric test used is based on the log-rank

statistic, but there are many other tests  $W_k$ , the so-called weighted log-rank statistics, which yield efficient tests in some situations.

These tests were developed for independent identically distributed random variables (iidrv)'s. When data is obtained from complex surveys, the i.i.d. - assumption is not valid and standard results may not apply. For example, if individuals in a survey were selected with probability proportional to size (pps) of the cluster where they live, the tests statistics could be affected by selection bias (see Pfeffermann, 1993).

In this paper, we concentrate on a non-parametric significance test to assess observed differences in empirical survival curves. We develop a test, based on the Gehan-Wilcoxon statistic, which can be applied to data from a pps- design.

In order to develop a test for complex lifetime data, we look at the sample from the point of view of the super-population theory. Many authors worked

<sup>1</sup> Susana Rubin-Bleuer, Statistics Canada, Ottawa, Canada, K1H-0T6, rubisus@statcan.ca.

under this theory, to mention just a few among them, Hartley and Sielken (1975), Fuller (1975), Binder and Roberts (1999), etc. The approach views the observed sample as the result of a 2-step process. It regards the finite populations of interest as samples of independent random variables of size  $N_1$  and  $N_2$  from two respective infinite populations, and it regards the stochastic procedure generating the observed samples from the finite populations (respective sizes  $n_1$  and  $n_2$ ) as the second phase samples of a two-phase sampling process. Thus, in terms of the super-population model, the design-probability may be viewed as being based on the conditional distribution given a particular outcome of the first phase process. Rubin-Bleuer (2000) formally defined a general space, which contains both the sampling design and the super-population that generates the finite population. This space, called the product probability space, is defined on the product of the sample space and the super-population space, with a sigma field defined by the product of the corresponding sigma fields, and a well-defined probability measure  $P_{m,d}$ .

Under this set up, if a classical test for the super-population is given by a statistic  $W_k$  which is a total or ratio of the  $N_1$  and  $N_2$  values of the respective finite populations, we may consider design-consistent sample estimators  $\hat{W}_K$  to develop tests when the data comes from a complex sample survey. These statistics are random on two accounts, the first phase randomization that yields the finite population, and given the finite population, the sampling randomization

The idea is to show that the sample estimators of the "finite population parameters" also converge to normal random variables under the null hypothesis of equality of the survival distributions. We consider  $\hat{W}_K$  as a random variable of the product probability space and prove convergence in the probability measure  $P_{m,d}$ .

The sampling design considered here is pps-sampling but the results are valid also for Poisson sampling under similar conditions.

In Section 2 we describe the Random Censorship Model under which many tests were developed for the comparison of survival curves. We also state Gill (1980)'s theorem, which is fundamental in the theory of non-parametric survival analysis. In

Section 3 we describe the sampling design that produces the observed data, and we define the product probability space that contains both the model and the design. In Section 4 we define the Gehan-Wilcoxon statistic (GW), used to compare two survival curves and is the basis of the test we are going to develop. In Section 5 we define the sample estimator of GW and show that, under the null hypothesis of equality of the survival curves, the sample estimator is asymptotically normal in the law of the product space.

## 2. THE MODEL

Most tests for comparing survival curves assume what is called the Random Censorship Model (RCM), with independent censoring. The following is a somewhat more general model that pertains to arrays of random variables (see chapter 3, Fleming & Harrington, 1991), and is used to obtain the necessary conditions to apply the central limit theorem to sequences of probability spaces.

Let  $(T_{1j}^N, U_{1j}^N)$ ,  $j=1, \dots, N_1$ , and  $(T_{2j}^N, U_{2j}^N)$ ,  $j=1, \dots, N_2$ , denote two samples of failure and censoring time pairs, one from each population, defined on an infinite probability space  $(\Omega, \mathfrak{S}, P)$ . Let  $N = N_1 + N_2$ , and we assume M1.  $N_i / N \rightarrow a_i$ ,  $i=1,2$  as  $N \rightarrow \infty$ .

Let us denote the actual observations, whether they are the end of the spell or the censoring time, by

$$X_{ij}^N = \min(T_{ij}^N, U_{ij}^N), \quad j=1, \dots, N_i, \quad i=1,2$$

Let  $T_{ij}^N \sim F_i = 1 - S_i$ ,  $j=1, \dots, N_i$ ,  $i=1,2$ , where the distribution functions  $F_i$  are absolutely continuous. Also let  $U_{ij}^N \sim L_i$ ,  $j=1, \dots, N_i$ ,  $i=1,2$ , absolutely continuous, and

$$\Lambda_i^N(t) = \Lambda_i(t) = \int_0^t \frac{dF_i(s)}{1 - F_i(s)}, \quad i=1,2. \quad \text{We assume}$$

that  $T_{ij}^N$  and  $U_{ij}^N$  are stochastically independent. Let  $\delta_{ij}^N = I(X_{ij}^N = T_{ij}^N)$  be the indicator function of a failure time actually being observed. We set  $\pi_i(t) = P(X_{ij}^N \geq t) = (1 - F_i(t))(1 - L_i(t))$ . We also assume that  $\pi_i(t) > 0 \forall 0 < t < \infty$ ,  $i=1,2$ .

Let  $Y_1(t)$  and  $Y_2(t)$  denote the number of individuals at risk at  $t$ ; that is, the number of

individuals for which we did not yet observe the end of the spell, nor it was censored by t.

We can express the statistic  $W_K$  as stochastic integral, which, under the null hypothesis of a common survival distribution, has the form

$$W_K = \int_0^\infty H_1(t) dM_2(t) - \int_0^\infty H_2(t) dM_1(t),$$

where the  $H_i$  are stochastic processes with some nice properties, and the  $M_i$  are martingales. The  $H_i$  and the  $M_i$  are sums of  $N_i, i=1,2$  random variables. Each term above is a sum of dependent random variables. The central limit theorem usually applied to prove asymptotic normality does not work here. Gill (1980) showed that under certain general conditions,  $W_K$  is asymptotically normal with mean zero and variance  $\sigma^2$ , where  $\sigma^2$  is a function of the limiting functions of the  $H_i$  and the common survival curve.

The following is a modification of Gill's theorem that can be found in Theorem 6.2.1 of Fleming and Harrington (1991). It provides conditions for weak convergence of random variables of the form

$$U_i^N = \int_0^\infty H_i^N(t) dM_i^N(t), i=1,2, \text{ and is an}$$

important tool in the development of large sample properties of one and two-sample statistics. Although it is not explicitly mentioned in the statement of the theorem, the result holds for statistics  $U_i^N, i=1,2$  that live in probability spaces that change with N. Indeed, the tightness condition is sufficient for the Lindeberg condition to hold, in which case the central limit theorem for arrays in changing probability spaces apply (see Theorem 27.2, Billingsley, 1995).

### Theorem (Gill, 1980)

Assume the random censorship model defined above, and let the statistics

$$U_i^N = \int_0^\infty H_i^N(t) dM_i^N(t), i=1,2, N=1,2,\dots$$

be based on this model, where  $H_i^N(t), i=1,2$ , are locally bounded predictable processes, and

$$M_i^N(t) = \eta_i(t) - \int_0^t Y_i(s) d\Lambda_i(s), i=1,2.$$

We assume that the following conditions hold:

$$1) \sup_{0 \leq t < \infty} |Y_i(t)/N_i - \pi_i(t)| \xrightarrow{P} 0 \text{ as } N \rightarrow \infty.$$

2) There exists a nonnegative function  $h_i(t)$  bounded on closed subintervals of  $[0, \infty)$ , such that

$$\sup_{0 \leq t < \infty} |\{H_i^N(t)\}^2 Y_i(t) - h_i(t)| \xrightarrow{P} 0 \text{ as } N \rightarrow \infty.$$

$$3) \int_0^\infty h_i(t) d\Lambda_i(t) < \infty$$

4)

$$\lim_{t \rightarrow \infty} \limsup_{N \rightarrow \infty} P \left\{ \int_0^\infty \{H_i^N(t)\}^2 Y_i(t) d\Lambda_i(t) > \varepsilon \right\} = 0$$

for any  $\varepsilon > 0$ .

Then

$$(U_1^N, U_2^N) \Rightarrow (Z_1^\infty, Z_2^\infty) \text{ in } D[0, \infty]^2,$$

where  $Z_1^\infty$  and  $Z_2^\infty$  are independent zero mean Gaussian processes with independent increments, continuous sample paths, and variance

$$\int_0^\infty h_i(t) d\Lambda_i(t).$$

### 3. THE PRODUCT SPACE

We define a space containing the model and the design spaces (following the methodology of Rubin-Bleuer (2000)), that will enable us to work with both at the same time.

Let us assume that the finite populations of interest are generated by an outcome of the ordered pairs  $(T_{ij}^N, U_{ij}^N), j=1, \dots, N_i, i=1,2$ . defined on an infinite probability space  $(\Omega, \mathfrak{S}, P)$  following the random censorship model. We select, independently, a probability proportional to size sample with replacement (pps) of size  $n_i, i=1,2$  from each finite population. The selection probabilities are  $p_{ij} > 0, \sum_j p_{ij} = 1, p_{ij} = Z_{ij} / Z_i$ , where  $Z_{ij}$  is

a measure of size of the unit  $ij$  and  $Z_i = \sum_j Z_{ij}$ .

We assume

D1.

$$\limsup_{N_i} \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{n_i p_{ij}} = O_p(1), \text{ as } N_i \rightarrow \infty \quad i=1,2.$$

D2.  $f_i = n_i / N_i, \lim f_i > 0$  as

$$N_i \rightarrow \infty, n_i \rightarrow \infty \quad i=1,2.$$

Condition D1 implies that the mean of the unit sizes is bounded as the number of units increase, and condition D2 ensures that the relationship between the sample and the population sizes (and its impact on the Statistics considered) remains the same as we increase the population size towards infinite.

Let  $p_d^i, i=1,2$  denote the respective design probability measures, which are defined given the prior information given by the unit sizes, that is,  $p_d^i(s, \omega) = p_d^i(s, Z_{i1}(\omega), \dots, Z_{iN_i}(\omega)), i=1,2$ .

Let  $S_i$  denote the collection of all possible samples under the design  $p_d^i, i=1,2$ . Let  $C(S_i)$  denote the collection of subsets of the sample space  $S_i$ . Then

$$(S_1 \times S_2 \times \Omega, C(S_1) \times C(S_2) \times \mathfrak{S}) \quad (3.1)$$

is a measurable space. We define a probability measure on the elementary rectangles of this space by

$$P_{d,m}(\{s_1\} \times \{s_2\} \times F) = \int_F p_d^1(s_1, \omega) p_d^2(s_2, \omega) dP(\omega), \quad (3.2)$$

where  $s_i \in S_i, i=1,2, F \in \mathfrak{S}$ . Thus the measurable space (3.1) and the probability measure (3.2) constitute a well-defined probability space determined by the super-population and the design (see Rubin-Bleuer, 2000). Note that this space changes as the respective population sizes change. In what follows, we denote by  $E_d, E_m, E_{d,m}$  the expectation with respect to the design probability space, the model space and the product space respectively.

#### 4. THE GEHAN WILCOXON STATISTIC

The Gehan-Wilcoxon Statistic  $GW$  is one of the weighted log-rank Statistics and it has the form

$$GW = c_N \int Y_1(t) Y_2(t) \left\{ \frac{d\eta_1(t)}{Y_1(t)} - \frac{d\eta_2(t)}{Y_2(t)} \right\},$$

where the constant is  $c_N = (N_1 N_2 (N_1 + N_2))^{-1/2}$ , the number of individuals at risk at time  $t$  from population  $i=1,2$  is  $Y_i(t) = \sum I(X_{ij}^N \geq t)$  and the

counting processes  $\eta_i(t)$  representing the number of observed distinct failures at time  $t$  is defined by the sum of the individuals failures,

$$\eta_i(t) = \sum_{j=1}^{N_i} \eta_{ij}(t), \quad \eta_{ij}(t) = I(X_{ij}^N \leq t) \delta_{ij}^N.$$

Note that we use a slightly different notation than the usual for counting processes to avoid confusion with the finite population sizes  $N_i, i=1,2$ .

Similarly, we define  $\eta_{ij}^U(t) = I(X_{ij}^N \leq t)(1 - \delta_{ij}^N), i=1,2$ .

Now let  $M_i(t) = \sum_{j=1}^{N_i} M_{ij}(t)$  where

$$M_{ij}(t) = \eta_{ij}(t) - \int_0^t I(X_{ij}^N \geq u) d\Lambda_i(u). \quad (4.1)$$

Hence  $M_i(t) = \eta_i(t) - \int_0^t Y_i(u) d\Lambda_i(u)$ . The  $M_{ij}$ 's

and the  $M_i$ 's are martingales with respect to the filtration (see theorem 1.3.2, Fleming & Harrington, 1991)

$$\{\mathfrak{S}_t, t \geq 0\}, \quad \mathfrak{S}_t = \sigma\{\eta_{ij}(t), \eta_{ij}^U(t), j=1, \dots, N_i, i=1,2\} \quad (4.2)$$

Under the null hypothesis of equality of the two survival functions, we have  $\Lambda_1 = \Lambda_2$ , and the statistic  $GW$  can be expressed by

$$GW(H_0) = c_N \int_0^t Y_2(t) dM_1(t) - Y_1(t) dM_2(t). \quad (4.3)$$

#### 5. THE SAMPLE STATISTIC

Let us consider the design-based estimator of the finite population total  $GW = GW(H_0)$  (see Sarndal et al, 1992):

$$G\hat{W} = c_N \int_0^t [Y_2(t) d\hat{M}_1(t) - Y_1(t) d\hat{M}_2(t)]$$

with  $\hat{Y}_i(t) = \frac{1}{n_i} \sum_k \left( \sum_{j=1}^{N_i} I(X_{ij} \geq t) \frac{I_{ij}^k(s_i)}{p_{ij}} \right)$  and

$\hat{M}_i(t) = \frac{1}{n_i} \sum_k \left( \sum_{j=1}^{N_i} M_{ij}(t) \frac{I_{ij}^k(s_i)}{p_{ij}} \right)$ , where

$I_{ij}^k(s_i) = 1$  if unit  $j$  of population  $i$  is selected to the sample  $s_i$  in the  $k$ -th draw, and zero otherwise.

**Theorem 5.1** Assume the conditions of the random censorship model, the design described in Section 3.

Then

$$G\hat{W} \Rightarrow N(0, \sigma^2) \text{ in } \ell(P_{d,m}). \quad (5.1)$$

where

$$\sigma^2 = \int_0^\infty \pi_1(t) \pi_2(t) (a_1 \pi_1(t) + a_2 \pi_2(t)) d\Lambda(t)$$

Proof: We apply Gill's Theorem to the stochastic integrals

$$U_1^{N_1} = c_N \int_0^\infty \hat{Y}_2(t) d\hat{M}_1(t)$$

and  $U_2^{N_2} = c_N \int_0^\infty \hat{Y}_1(t) d\hat{M}_2(t)$ .

Equation (5.1) follows from Gill's theorem, so we must verify all of its conditions. These are properties a) and b) below and Conditions 1 to 4 of Gill's theorem.

a)  $H_1^{N_1}(t) \equiv c_N \hat{Y}_2(t)$  and  $H_2^{N_2}(t) \equiv c_N \hat{Y}_1(t)$  are bounded predictable processes on  $0 \leq t < \infty$ .

b)  $\hat{M}_i(t) = \hat{\eta}_i(t) - \int_0^t \hat{Y}_i(u) d\Lambda(u)$  under the null

hypothesis  $H_0 : \Lambda = \Lambda_1 = \Lambda_2$ , are martingales ( $i = 1, 2$ ) in the product space with respect to the filtration defined by

$$\{\mathfrak{S}_t^N, t \geq 0\}, \quad \mathfrak{S}_t^N = C(S_1) \times C(S_2) \times \mathfrak{S}_t.$$

We first show a) and b):

The  $H_i^{N_i}(t, \omega, s_j) = c_N \hat{Y}_j(t)$   $i = 1, 2$   $j \equiv i + 1$  are bounded predictable processes on  $0 \leq t < \infty$ , in the product probability space, because they are linear combinations of indicator functions of predictable rectangles in  $R^+ \times S_1 \times S_2 \times \Omega$  of the form

$$(t, \infty) \times \{s_j : I_{jh}^k(s_j) = 1\} \times S_i \times \{X_{jh}^N \geq t\},$$

$h = 1, \dots, N_j$ , (see definition 1.4.2, Fleming & Harrington, 1991) •

Let us write  $J_{ij}(s_i) = \frac{1}{n_i} \sum_{k=1}^{n_i} I_{ij}^k(s_i)$ ,  $i = 1, 2$  and

$$\hat{M}_i(t) = \sum_{j=1}^{N_i} M_{ij}(t) J_{ij}(s_i) / p_{ij}.$$

Now we define  $\tilde{M}_{ij}(t) = M_{ij}(t) J_{ij}(s_i) / p_{ij}$ . The

$\tilde{M}_{ij}(t)$  are martingales in the product space with respect to the filtration  $C(S_1) \times C(S_2) \times \mathfrak{S}_t$ . Indeed, we verify the three necessary conditions for a process to be a martingale (for pertinent definitions, see Fleming & Harrington, 1991):

$\tilde{M}(t)$  is  $\mathfrak{S}_t^N$ -measurable since it is the product of  $\mathfrak{S}_t$  and  $C(S_1) \times C(S_2)$  measurable variables.

The  $M_{ij}$  are martingales in the model space, so we have

$$E_{d,m}(|\tilde{M}_{ij}(t)|) = E_m(|M_{ij}(t)|) E_d(J_{ij}) / p_{ij} < \infty.$$

$J_{ij}$  is  $C(S_1) \times C(S_2) \times \mathfrak{S}_t$ -measurable, hence

$$E_{d,m}(\tilde{M}_{ij}(t+u) | \mathfrak{S}_t^N) = (J_{ij}(s_i) / p_{ij}) E_m(M_{ij}(t+u) | \mathfrak{S}_t)$$

and since  $M_{ij}$  is a martingale with respect to  $\mathfrak{S}_t$ ,

$$\text{we have } (J_{ij}(s_i) / p_{ij}) E_m(M_{ij}(t+u) | \mathfrak{S}_t) =$$

$$(J_{ij}(s_i) / p_{ij}) M_{ij}(t) = \tilde{M}_{ij}(t).$$

Thus the  $\hat{M}_i(t)$  are sums of martingales in the product space and we have

$$\hat{M}_i(t) = \hat{\eta}_i(t) - \int_0^t \hat{Y}_i(u) d\Lambda(u) \text{ when we express the}$$

$M_{ij}(t)$  as in (4.1) •

Now we verify Conditions 1 to 4 of Gill's theorem.

**Condition 1**

$$\sup_{0 \leq t < \infty} |\hat{Y}_i(t) / N_i - \pi_i(t)| \rightarrow 0 \text{ in } P_{d,m} \quad (5.2)$$

for  $i = 1, 2$ . Indeed, by the Glivenko-Cantelli Theorem,

$$\sup_{0 \leq t < \infty} |Y_i(t) / N_i - \pi_i(t)| \rightarrow 0 \text{ in } P. \quad (5.3)$$

We also have, by lemma 5.1 below,

$$\sup_{0 \leq t < \infty} |\hat{Y}_i(t) - Y_i(t)| / N_i \rightarrow 0 \text{ in } P_{d,m} \quad (5.4)$$

Hence, definition 3.2 implies that the sequence in (5.3) converges to zero in  $P_{d,m}$  (the probability measure of the product space), respectively, and thus condition 1 holds as the population sizes increase towards infinity •

**Condition 2.**

$\sup_{0 \leq t < \infty} \left| \left\{ \int_N \hat{Y}_2(t) \right\} \hat{Y}_1(t) - a_2 \pi_2^2(t) \pi_1(t) \right| \xrightarrow{P} 0$   
for  $H_1^{N_1}(t)$ , and similarly for  $H_2^{N_2}(t)$ . It follows from (5.2) and Condition M1 of Section 2 •

**Condition 3.** It refers to the variance of  $G\hat{W}(H_0)$ , which is finite since the integrands are bounded continuous functions:

$$\int_0^{\infty} (a_2 \pi_2^2(t) \pi_1(t) + a_1 \pi_1^2(t) \pi_2(t)) d\Lambda(t) < \infty \bullet$$

**Condition 4.** It is the tightness condition. We show

$$\lim_{t \rightarrow \infty} \limsup_{N \rightarrow \infty} P_{m,d} \left\{ c_N^2 \int_t^{\infty} \left\{ \hat{Y}_i(t) \right\} \hat{Y}_k(t) d\Lambda(t) > \varepsilon \right\} = 0 \text{ for any } \varepsilon > 0. \quad (5.4)$$

The integrability of the survival distributions of the  $X_{ij}$   $i=1,2$  (Condition 3) implies that

for every  $\varepsilon > 0$  there exists a  $t_0$  such that for  $t \geq t_0$ ,

$$a_2 \int_t^{\infty} \pi_2^2(u) \pi_1(u) d\Lambda(u) < \varepsilon / 2. \quad (5.6)$$

And the uniform convergence of  $c_N^2 \hat{Y}_2^2 \hat{Y}_1$  (Condition 2) implies that

$$\left| c_N^2 \int_t^{\infty} \hat{Y}_2^2(u) \hat{Y}_1(u) d\Lambda(u) - a_2 \int_t^{\infty} \pi_2^2(u) \pi_1(u) d\Lambda(u) \right| = o_{P_{d,m}}(1) \text{ as } N \rightarrow \infty$$

Hence, there exists  $N_0 \geq 1$  such that

$$\left| c_N^2 \int_t^{\infty} \hat{Y}_2^2(u) \hat{Y}_1(u) d\Lambda(u) - a_2 \int_t^{\infty} \pi_2^2(u) \pi_1(u) d\Lambda(u) \right| < \varepsilon / 2 \text{ for all } N \geq N_0. \quad (5.7)$$

Equations (5.6) and (5.7) imply (5.5) •

**Lemma 5.1** Under conditions D1 and D2 and uniform continuity of  $\pi_i(t)$   $i=1,2$ , we have

$$\sup_{0 \leq t < \infty} |\hat{Y}_i(t) - Y_i(t)| / N_i \rightarrow 0 \text{ in } P_{d,m} \quad (5.8)$$

Without loss of generality, we omit the index  $i$  in the outline of the proof. (Here  $N = N_i, i=1,2$ .)

Let  $W(t) = |\hat{Y}(s,t) - Y(t)| / N$ . We have

$W(t) \rightarrow 0$  in  $p_d$  and thus in  $P_{d,m}$  for every  $t$ , and

$$\lim_{h \rightarrow 0} \limsup_{n \rightarrow \infty} P_{d,m}(\sup_{0 \leq t < \infty} |W(t) - W(\tau)| > \varepsilon) = 0.$$

Indeed, the first statement above follows from the consistency of the design estimator and Theorem 4.1 in Rubin-Bleuer (2000). The second statement follows from the equation

$$\sup_{|t-\tau| \leq h} |W(t) - W(\tau)| \leq A.B \text{ where}$$

$$A = \max_{1 \leq j \leq N} \left| \frac{J_j}{np_j} - 1 \right| = O_p(\log n) \text{ as } N \rightarrow \infty \quad \text{by}$$

condition D1 and

$$B = \sup_{|t-\tau| \leq h} \frac{1}{N} \sum_{j=1}^N I(t \leq X_j \leq \tau) = \frac{1}{\sqrt{N}} o_p(h)$$

as  $h \rightarrow 0, N \rightarrow \infty$ , by the uniform continuity of  $\pi(t)$ . Now we apply Skorokhod representation theorem (see Theorem 3.29, Breiman 1992) noting that  $X_j = \pi^{-1}(\xi_j)$  with  $\xi_j$  independent Uniform (0,1) random variables and obtain (5.8) •

## REFERENCES

- Billingsley, P.(1995). *Probability and Measure*, third edition, Wiley, New York.
- Binder, D. and Roberts, G. (1999). Design-based and Model-based Methods for Estimating Model Parameters. *Presented at the International Conference for Analysis of Survey Data, 24-26 August 1999, Southampton, U.K.*
- Breiman, L.(1992). *Probability*. SIAM, Philadelphia.
- Chung,K.L. (1974). *A Course in Probability Theory*, second edition. Academic Press, New York.
- Fleming,T. and Harrington, D.(1991). *Counting Processes and Survival Analysis*. Wiley, New York.
- Fuller W.A. (1975). Regression analysis for sample surveys. *Sankhya (C)*. 37, 117-132.

- Gill, R.D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.
- Hartley, H.O. and Sielken, R.L. (1975). A "super-population viewpoint" for finite population sampling, *Biometrics*, 31, 411-422.
- Pfeffermann, D. (1993). The role of Sampling weights when modeling survey data. *International Statistical Review*, 61, 317-37.
- Rubin-Bleuer, S. (2000). Some issues in the analysis of complex survey data. *Statistics Canada Series, Methodology Branch, Business Survey Methods Division*, BSMD- 20-001 E.
- Särndal, C-E, Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.