Risques concurrents en Statistique des durées de vie

Jean-Yves Dauxois (Univ. Franche-Comté)

En collaboration avec
Laurent Bordes (Univ. Pau), Agathe Guilloux (Univ. Paris VI), Pierre Joly (Univ. Bordeaux II et INSERM) and Sophie Sencey (INSEE)

4 décembre 2008

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Let T be a lifetime with support $\mathbb{R}^+$, with c.d.f. F, Survival Function $\bar{F} = 1 - F$ and Cumulative Hazard Function:

$$\Lambda(t) = \int_{]0,t]} \frac{dF(s)}{\bar{F}(s^-)},$$

for all $t > 0$.

If T is absolutely continuous, we define the (instantaneous) hazard rate function, for $t > 0$, by:

$$\lambda(t) = \lim_{h \longrightarrow 0^+} \frac{P(T \in [t, t+h[| T \geq t)}{h} = \frac{f(t)}{\bar{F}(t)},$$

where f is the p.d.f. of T.

Inferential statistic based on the observation of a sample $T_1, \ldots, T_n$ is straightforward (undergraduate exercise!)...

Inferential statistic based on the observation of a sample $T_1, \ldots, T_n$ is straightforward (undergraduate exercise!)...

Censoring

But often (quite often in fact) we observe a sample $(X_i, \delta_i)_{i=1,\ldots,n}$ of

$$\left\{ \begin{array}{rcl} X & = & T \wedge C \\ \delta & = & I(\{T \leq C\}) \end{array} \right. .$$

The r.v. C is often supposed to be independent of T and represents the censoring mechanism. Write G its c.d.f. and $\bar{G}$ its survival function.

Nonparametric Inference

Define the counting process $N.$ and the at risk process Y, for $t > 0$, by respectively:

$$N.(t) = \sum_{i=1}^{n} I(\{X_i \leq t, \delta_i = 1\})$$

and

$$Y(t) = \sum_{i=1}^{n} I(\{X_i \geq t\}).$$

The Nelson-Aalen estimator $\widehat{\Lambda}$ of the cumulative hazard rate function is defined, for all $t > 0$, by:

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN_.(s)}{Y(s)} = \sum_{i:X_{(i)} \leq t} \frac{\Delta N_.(X_{(i)})}{Y(X_{(i)})},$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the n order statistics.

The Nelson-Aalen estimator $\widehat{\Lambda}$ of the cumulative hazard rate function is defined, for all $t > 0$, by:

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN_.(s)}{Y(s)} = \sum_{i:X_{(i)} \leq t} \frac{\Delta N_.(X_{(i)})}{Y(X_{(i)})},$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the n order statistics.

The Kaplan-Meier estimator $\widehat{\bar{F}}$ of the Survival function is defined by:

$$\widehat{\bar{F}}(t) = \prod_{s \leq t} (1 - \widehat{\lambda}(s)) = \prod_{i:X_{(i)} \leq t} \left(1 - \frac{\Delta N_.(X_{(i)})}{Y(X_{(i)})}\right).$$

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Suppose that there is p ($p \geq 2$) different causes of death (failure), denoted by $D_1, \ldots, D_p$.

Suppose that there is p ($p \geq 2$) different causes of death (failure), denoted by $D_1, \ldots, D_p$.

Let T_j, for $j = 1, \ldots, p$, be the lifetime of the individual in the hypothetic situation where there is only one cause D_j.

Suppose that there is p ($p \geq 2$) different causes of death (failure), denoted by $D_1, \ldots, D_p$.

Let T_j, for $j = 1, \ldots, p$, be the lifetime of the individual in the hypothetic situation where there is only one cause D_j.

Assuming that death (or failure) can be due to only one cause, we only observe

$$\left\{ \begin{array}{lcl} T & = & \min_{1 \leq j \leq p} T_j \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \end{array} \right. ,$$

where d is the cause of death indicator.

Without any other assumption, we can not estimate the functions λ_{T_j} et F_{T_j}, for $j = 1, \ldots, p$.

Without any other assumption, we can not estimate the functions λ_{T_j} et F_{T_j}, for $j = 1, \ldots, p$.

Two solutions (among others)

Without any other assumption, we can not estimate the functions λ_{T_j} et F_{T_j}, for $j = 1, \ldots, p$.

Two solutions (among others)

- Either we consider inference on the Cumulative Incidence Functions (CIF)

$$F_j(t) = P(T \leq t, d = j),\ j = 1, \ldots, p; 0 < t < +\infty.$$

or the cause specific hazard rate functions

$$\lambda_j(t) = \lim_{h \longrightarrow 0^+} \frac{P(T \in [t, t+h[, d = j | T \geq t)}{h},\ j = 1, \ldots, p.$$

Without any other assumption, we can not estimate the functions λ_{T_j} et F_{T_j}, for $j = 1, \ldots, p$.

Two solutions (among others)

- Either we consider inference on the Cumulative Incidence Functions (CIF)

$$F_j(t) = P(T \leq t, d = j),\ j = 1, \ldots, p; 0 < t < +\infty.$$

or the cause specific hazard rate functions

$$\lambda_j(t) = \lim_{h \longrightarrow 0^+} \frac{P(T \in [t, t+h[, d = j | T \geq t)}{h},\ j = 1, \ldots, p.$$

- Either we suppose that $T_1, \ldots, T_p$ are independent and we are able to estimate the hazard rate λ_{T_j} (and thus the c.d.f. F_{T_j}) of the r.v. T_j, for $j = 1, \ldots, p$.

We often also assume the presence of independent censoring still denoted by C and with c.d.f. G.
Thus, instead of observing (T, d), we only observe

$$\left\{ \begin{array}{lcl} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{array} \right. ,$$

where C is independent from $T_1, \ldots, T_p$.

We often also assume the presence of independent censoring still denoted by C and with c.d.f. G.
Thus, instead of observing (T, d), we only observe

$$\begin{cases} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} jI(\{T = T_j\}) \text{ si } \delta \neq 0 \end{cases},$$

where C is independent from $T_1, \ldots, T_p$.
One can show that we have in this case, for all $t \geq 0$:

$$\Lambda_j^C(t) = \Lambda_j(t),$$

allowing us to estimate Λ_j and F_j, for $j = 1, \ldots, K$, and the survival function $\bar{F}$ of T.

Nonparametric inference

Nonparametric inference

Write, for $j = 1, \ldots, p$,

$$N_j(t) = \sum_{i=1}^{n} I\{X_i \leq t, d_i = j\} \text{ and } Y(t) = \sum_{i=1}^{n} I\{X_i \geq t\}.$$

The Nelson-Aalen type estimators of the cumulative hazard rate functions are, for $j = 1, \ldots, p$:

$$\hat{\Lambda}_j(t) = \int_0^t \frac{dN_j(u)}{Y(u)}.$$

Nonparametric inference

Write, for $j = 1, \ldots, p$,

$$N_j(t) = \sum_{i=1}^{n} I\{X_i \leq t, d_i = j\} \text{ and } Y(t) = \sum_{i=1}^{n} I\{X_i \geq t\}.$$

The Nelson-Aalen type estimators of the cumulative hazard rate functions are, for $j = 1, \ldots, p$:

$$\hat{\Lambda}_j(t) = \int_0^t \frac{dN_j(u)}{Y(u)}.$$

The Aalen-Johansen of the CIF are, for $j = 1, \ldots, p$:

$$\widehat{F}_j(t) = \int_0^t \widehat{\bar{F}}(u^-) \frac{dN_j(u)}{Y(u)}.$$

Theorem

(Aalen, 1978, Aalen & Johansen, 1978) *For all $\tau < \infty$, the following weak convergence holds $\mathbb{D}^3[0,\tau]$, when $n \to \infty$:*

$$\sqrt{n}(\widehat{\bar{F}} - \bar{F}, \widehat{F}_1 - F_1, \widehat{F}_2 - F_2) \xrightarrow{\mathcal{D}} Z = (Z_0, Z_1, Z_2),$$

where $Z = (Z_0, Z_1, Z_2)$ is a mean zero gaussian process defined by

$$Z_0 = \bar{F} U_0;\ Z_k(\cdot) = \int_0^{\cdot} (F_k(\cdot) - F_k(u)) dU_0(u) + \int_0^{\cdot} \bar{F}(u) dU_k,$$

for $k = 1, 2$, and U_1 and U_2 are mean zero gaussian, orthogonal and locally square integrable martingales with covariance functions

$$< U_k(s), U_k(t) >= \int_0^{s \wedge t} \frac{dF_k(u)}{\bar{F}^2(u)\bar{G}(u-)} \ et\ U_0 = -(U_1 + U_2).$$

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation unit between 1995 and 1999.

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation unit between 1995 and 1999.

- There are 67 different types of infection. The most frequent are: urinary tracts, pneumonias or septicaemias and herpes.

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation unit between 1995 and 1999.

- There are 67 different types of infection. The most frequent are: urinary tracts, pneumonias or septicaemias and herpes.
- Each patient can develop several infections, the maximum number observed being 13, each type of infection being able to occur several times.

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation unit between 1995 and 1999.

- There are 67 different types of infection. The most frequent are: urinary tracts, pneumonias or septicaemias and herpes.
- Each patient can develop several infections, the maximum number observed being 13, each type of infection being able to occur several times.
- The end of hospitalization in the reanimation unit can be due to death or censoring, death being clearly dependent of the experienced infections.

Our aim

Two questions are of interest in this area.

Our aim

Two questions are of interest in this area.

- Is the occurrence rate of a given type of events increasing with time?

Our aim

Two questions are of interest in this area.

- Is the occurrence rate of a given type of events increasing with time?
- Is the instantaneous probability of experiencing an event of a given type always greater than the one of an other type?

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Notation

We consider a setting of competing risks for recurrent events in presence of a terminal event. That is, each type is concurring with the others and we focus on the frequency of each type.

Notation

We consider a setting of competing risks for recurrent events in presence of a terminal event. That is, each type is concurring with the others and we focus on the frequency of each type.

To simplify notation we restrict our attention to the case where only two different types of event are concurring.

Notation

We consider a setting of competing risks for recurrent events in presence of a terminal event. That is, each type is concurring with the others and we focus on the frequency of each type.

To simplify notation we restrict our attention to the case where only two different types of event are concurring.

Write $N_j^*(t) =$ total number of events of type j experienced by an individual up to time t, for $j = 1, 2$.

Individuals are subject to a terminal event at time D, such that they cannot experience any further event after. This time, with survival function $S(t) = \mathbb{P}(D > t)$, is supposed to be dependent from the recurrent event processes $N_j^*(t)$, $j = 1, 2$.

Individuals are subject to a terminal event at time D, such that they cannot experience any further event after. This time, with survival function $S(t) = \mathbb{P}(D > t)$, is supposed to be dependent from the recurrent event processes $N_j^*(t)$, $j = 1, 2$.

Finally, the presence of an independent random right-censoring mechanism is also allowed and denoted by the random variable C.

This means that the observation are:

$$\begin{aligned} N_j(t) &= N_j^*(t \wedge C), \\ X &= D \wedge C, \end{aligned}$$

for $j = 1, 2$, and

$$\delta = I(D \leq C).$$

This means that the observation are:

$$\begin{aligned} N_j(t) &= N_j^*(t \wedge C), \\ X &= D \wedge C, \end{aligned}$$

for $j = 1, 2$, and

$$\delta = I(D \leq C).$$

The overall recurrent event process $N = N_1 + N_2$ counts the number of events of all types experienced by a subject up to time t.

Now, let us define the specific (resp. overall) mean function, for $j = 1, 2$, by:

$$\mu_j(t) = \mathbb{E}(N_j^*(t)) \text{ (resp. } \mu(t) = \mathbb{E}(N^*(t))).$$

One can define the specific mean frequency functions by

$$\begin{aligned} \mu_j'(t) &= \lim_{\Delta_t \to 0} \frac{1}{\Delta_t} \mathbb{P}\left(N_j^*(t+\Delta_t) - N_j^*(t) = 1\right) \\ &= \text{infinitesimal probability of observing an event} \\ &\quad \text{of type } j \text{ at time } t \end{aligned}$$

And the overall mean frequency function is $\mu'(t) = \mu_1'(t) + \mu_2'(t)$

The probability of observing an event of type j at time t, $j = 1, 2$, given that an event occurs at time t, is given by

$$p_j(t) = \frac{\mu_j'(t)}{\mu'(t)}.$$

Finally note that we have:

$$\mu_j(t) = \int_0^t S(u-) dR_j(u), \tag{1}$$

where $dR_j(t) = \mathbb{E}(dN_j^*(t)|D \geq t)$, for $j = 1, 2$.

Estimators and their weak convergence

Suppose that we observe a sample $(N_{1i}, N_{2i}, X_i, \delta_i)$, $i = 1, \ldots, n$.

The survival function of D is easily estimated by the Kaplan-Meier estimator $\widehat{S}$.

$$\widehat{S}(t) = \prod_{i:X_{(i)} \leq t} \left(1 - \frac{\sum_{j=1}^{n} \mathbb{1}_{\{X_j = X_{(i)}, \delta_j = 1\}}}{\bar{Y}(X_{(i)})} \right),$$

where

$$\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t) \equiv \sum_{i=1}^{n} I(X_i \geq t).$$

Moreover, a Nelson-Aalen type estimator of R_j is given by

$$\widehat{R}_j(t) = \sum_{i=1}^{n} \int_0^t J(u) \frac{dN_{ji}(u)}{\bar{Y}(u)},$$

where $J(t) = I(\bar{Y}(t) > 0)$.

Thus, estimators of the specific mean functions are, for $j = 1, 2$:

$$\widehat{\mu}_j(t) = \int_0^t \widehat{S}(u-) d\widehat{R}_j(u).$$

Let $\pi(t) = \mathbb{P}(X > t)$, $t \geq 0$, and define τ such that $\pi(\tau-) > 0$.

Theorem

As $n \to \infty$,

$$n^{1/2}\left(\widehat{\mu}_1 - \mu_1, \widehat{\mu}_2 - \mu_2\right) \xrightarrow{\mathcal{D}} (G_1, G_2)$$

in $D^2[0,\tau]$, *where* $G = (G_1, G_2)$ *is a mean-zero Gaussian process with covariance function* $\xi_{jk}(s,t) = cov(G_j(s), G_k(t))$

The covariance function $\xi_{jk}(s,t) = cov(G_j(s), G_k(t))$ of the limiting process can be consistently estimated.

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

In this section, our aim is to derive a test of

$$H_0 : p_1(t) \text{ is constant}$$

against

$$H_1 : p_1(t) \text{ is an increasing function of } t.$$

The alternative hypothesis

$$H_2 = \bar{H}_0 : p_1(t) \text{ is not constant}$$

will also be considered.

One can show that

- for $t \in [0, \tau]$,

$$U(t) = \frac{\mu_1(t)\mu_2(t)}{2} - \int_0^t \mu_1(s) d\mu_2(s).$$

is a measure of the deviation from the null hypothesis H_0 on the interval $[0, t]$, when H_1 is in mind.

One can show that

- for $t \in [0, \tau]$,

$$U(t) = \frac{\mu_1(t)\mu_2(t)}{2} - \int_0^t \mu_1(s) d\mu_2(s).$$

  is a measure of the deviation from the null hypothesis H_0 on the interval $[0, t]$, when H_1 is in mind.
- A measure of deviation from H_0 can be defined by

$$\sup_{u \in [0,t]} |U(u)|$$

  when the alternative H_2 is in mind.

A plug in estimator of U is given by

$$\widehat{U}(t) = \left[\frac{\widehat{\mu}_1(t)\widehat{\mu}_2(t)}{2} - \int_0^t \widehat{\mu}_1(s) d\widehat{\mu}_2(s) \right], \ t \geq 0.$$

Theorem

For τ with $\pi(\tau) > 0$,

$$\sqrt{n}\left(\widehat{U}(.) - U(.)\right) \xrightarrow{\mathcal{D}} W_U$$

in $D[0,\tau]$ where W_U is a mean-zero Gaussian process defined, for all t, by:

$$\begin{aligned} W_U(t) &= \frac{\mu_2(t)G_1(t)}{2} - \frac{\mu_1(t)G_2(t)}{2} \\ &+ \int_0^t G_1(s)d\mu_2(s) - \int_0^t G_2(s-)d\mu_1(s). \end{aligned}$$

Our first test statistic, testing H_0 against H_1 (i.e. detecting if p_1 is increasing), is:

$$T_{1n} = \sqrt{n}\widehat{U}(\tau).$$

Since $U(\tau)$ is null under H_0, Theorem 2.2 leads to

$$T_{1n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{V}(W_U(\tau))).$$

Our second test statistic, detecting if p_1 is constant or not is defined by

$$T_{2n} = \sqrt{n} \sup_{t \in [0,\tau]} \left| \widehat{U}(t) \right|.$$

Still under H_0, one obtains

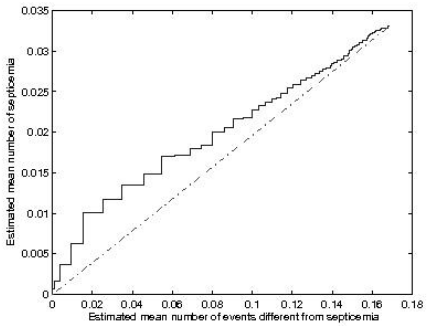
$$T_{2n} \xrightarrow{\mathcal{D}} \sup_{t \in [0,\tau]} |W_U(t)|.$$

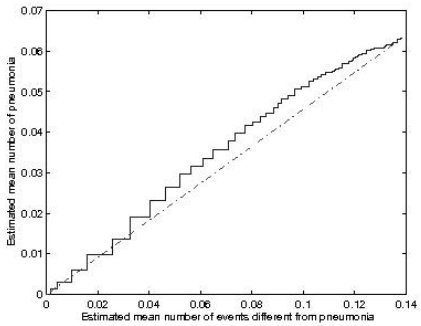
Application to the nosocomial infections data set

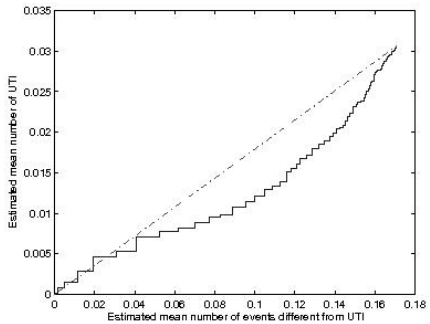
Recall: 7867 patients hospitalized in a reanimation service between 1995 and 1999. They may have developed nosocomial infections of different importance, as septicemia, pneumonia, herpes or urinary tract infections.

Our aim is to detect if the probability of contracting a particular type of event, knowing that the patient experienced an event at time t, is decreasing with time.

Tests on Nosocomial Infections data set of increasing (or decreasing) occurrence rates of : 1) Septicemia, 2) Pneumonia, 3) Urinary Tract.







If septicemia is the cause of primary interest, we obtain a p-value of nearly 0 in the test of $H_0 : p_{sept}\ is\ constant$ against $H_3 : p_{sept}\ is\ decreasing$

If septicemia is the cause of primary interest, we obtain a p-value of nearly 0 in the test of $H_0 : p_{sept}\ is\ constant$ against $H_3 : p_{sept}\ is\ decreasing$

If we consider pneumonia, our test of $H_0 : p_{pneu}\ is\ constant$ against $H_3 : p_{pneu}\ is\ decreasing$ gives a p-value of 0.001.

If septicemia is the cause of primary interest, we obtain a p-value of nearly 0 in the test of $H_0 : p_{sept}\ is\ constant$ against $H_3 : p_{sept}\ is\ decreasing$

If we consider pneumonia, our test of $H_0 : p_{pneu}\ is\ constant$ against $H_3 : p_{pneu}\ is\ decreasing$ gives a p-value of 0.001.

Finally, if we consider urinary tract infections (UTI), we obtain a p-value of nearly 0 in our test of $H_0 : p_{UTI}\ is\ constant$ against $H_3 : p_{UTI}$ *is increasing*

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Now our aim is to derive a test of

$$H_0 : \mu_1'(t) = \mu_2'(t), \ \forall t$$

against

$$H_1 : \mu_1'(t) \geq \mu_2'(t), \ \forall t.$$

With the notation

$$\bar{F}_j = \exp(-\mu_j),$$

for $j = 1, 2$, we can reformulate our hypotheses as

$$H_0 : \frac{\bar{F}_1(t)}{\bar{F}_2(t)} = 1$$

and

$$H_1 : \frac{\bar{F}_1(t)}{\bar{F}_2(t)} \text{ is an increasing function.}$$

Let

$$q(k_1, k_2) = \psi_1(k_1)\psi_2(k_2) - \psi_1(k_2)\psi_2(k_1),$$

with

$$\psi_j(k_i) = \int_0^\tau k_i(s)\bar{F}_j(s)ds$$

where k_1 and k_2 are positive weight function such that k_1/k_2 is an increasing function.

One can show that $q(k_1, k_2)$ is a measure of the deviation from H_0. Indeed the quantity is null under H_0 and positive under H_1.

Let K_i be an estimator of k_i, and let us estimate $\bar{F}_j(t)$ by

$$\widehat{\bar{F}}_j(t) = \exp(-\widehat{\mu}_j(t)).$$

We can then estimate ψ_{ji} by

$$\widehat{\psi}_j(K_i) = \int_0^\tau K_i(s)\widehat{\bar{F}}_j(s)ds.$$

Then a plug in estimator of the non-proportionality measure q is given by

$$\widehat{Q}(K_1, K_2) = \widehat{\psi}_1(K_1)\widehat{\psi}_2(K_2) - \widehat{\psi}_1(K_2)\widehat{\psi}_2(K_1).$$

Theorem

Assume that

$$\sup_{t\in[0,\tau]} |K_i(t) - k_i(t)| \xrightarrow{\mathbb{P}} 0.$$

Then

$$\sqrt{n}(\widehat{Q}(K_1, K_2) - q(k_1, k_2)) \xrightarrow{\mathcal{D}} N$$

in $D[0,\tau]$, where N is a mean-zero Gaussian random variable

From this theorem one can use the statistic

$$T_{3n} = \sqrt{n}\widehat{Q}(K_1, K_2)$$

for testing H_0 against H_1.

Application to the nosocomial infections data set

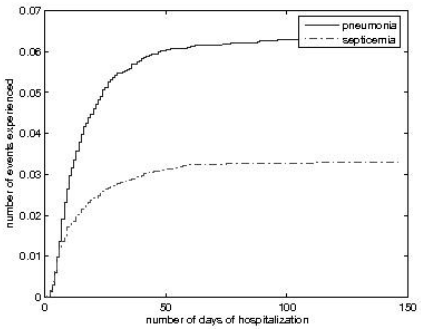


Figure: Plot of μ_{pneum} and μ_{septi}

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Suppose that

- $T_1, \ldots, T_p$ are independent.

Suppose that

- $T_1, \ldots, T_p$ are independent.
- Semiparametric model : additive risk. The conditional distribution of T_j, given a vector of covariates $Z \in \mathbb{R}^k$, is such that

$$\lambda_j(t|Z) = \lambda_{0j}(t) + \beta_j^T Z, \quad t \geq 0,$$

where λ_{0j} is an unspecified baseline hazard function and $\beta_j \in \mathbb{R}^k$ is a vector of regression parameter associated to the jth cause (under the constraint that $\lambda_j(t|Z) \geq 0$).

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Suppose furthermore that

- There is right censoring C. Instead of observing (T, d), we observe

$$\begin{cases} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{cases},$$

  where C is (conditionally to Z) independent from $T_1, \ldots, T_p$.

Suppose furthermore that

- There is right censoring C. Instead of observing (T, d), we observe

$$\begin{cases} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{cases},$$

  where C is (conditionally to Z) independent from $T_1, \ldots, T_p$.
- The cause of death (or failure) d is sometimes missing (even when T is uncensored).

Hypothesis on the missing mechanism

Hypothesis on the missing mechanism

- Write

$$M = \begin{cases} 1 & \text{when the cause } d \text{ is known} \\ 0 & \text{otherwise} \end{cases}$$

Hypothesis on the missing mechanism

- Write

$$M = \begin{cases} 1 & \text{when the cause } d \text{ is known} \\ 0 & \text{otherwise} \end{cases}$$

- We suppose that:

$$P(M = 1|X, Z, \delta = 1) = P(M = 1|\delta = 1) = \alpha \in [0, 1]$$

and

$$P(M = 0|X, Z, \delta = 0) = P(M = 0|\delta = 0) = 1,$$

Finally the observation are (X, δ, D, Z) where

$$D = \delta M \sum_{j=1}^{p} j 1(T_j = T).$$

The random vector (X, δ, D) can be seen, conditionally to Z, as the realization of an inhomogeneous Markov process with (p+3) states all being absorbant except state 0.

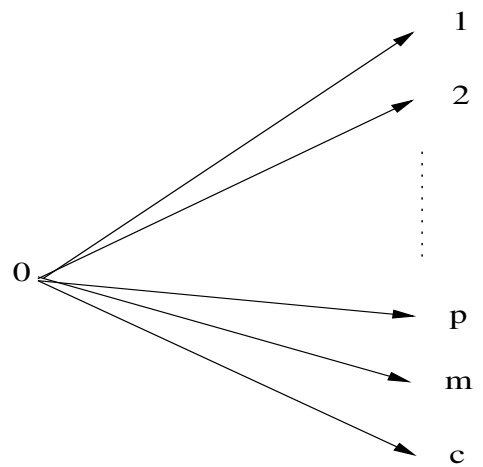


Figure: Markov graph associated to (X, δ, D)

Let us denote by $\bar{\lambda}_{0x}(\cdot|Z)$ the transition rate for $0 \longrightarrow x$ $(x \in \{1, \ldots, p, m, c\})$, conditionally to Z, of this process. We have

$$\left\{ \begin{array}{lcl} \bar{\lambda}_{0j}(t) & = & \alpha\lambda_j(t|Z) = \alpha(\lambda_{0j}(t) + \beta_j^T Z) \text{ for } j \in \{1, \ldots, p\}, \\ \bar{\lambda}_{0m}(t) & = & (1-\alpha)\sum_{j=1}^p \lambda_j(t|Z) = (1-\alpha)\left(\lambda_m(t) + \beta_m^T Z\right), \\ \bar{\lambda}_{0c}(t) & = & \lambda_C(t), \end{array} \right.$$

where $\lambda_m = \sum_{j=1}^p \lambda_{0j}$ and $\beta_m = \sum_{j=1}^p \beta_j$.

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

- Write $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ an n sample of (X, δ, D, Z).

- Write $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ an n sample of (X, δ, D, Z).
- For $j \in \{1, \ldots, p, m\}$, define the counting processes:

$$\begin{aligned} N_{ij}(t) &= 1(X_i \leq t, D_i = j) \text{ for } j \neq m, \\ N_{im}(t) &= 1(X_i \leq t, \delta_i = 1, D_i = 0). \end{aligned}$$

- Write $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ an n sample of (X, δ, D, Z).
- For $j \in \{1, \ldots, p, m\}$, define the counting processes:

$$\begin{aligned} N_{ij}(t) &= 1(X_i \leq t, D_i = j) \text{ for } j \neq m, \\ N_{im}(t) &= 1(X_i \leq t, \delta_i = 1, D_i = 0). \end{aligned}$$

- In order to simplify the notation, write $m \equiv p + 1$ and write Y_i at risk process defined by $Y_i(t) = 1(X_i \geq t)$.

- Write $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ an n sample of (X, δ, D, Z).
- For $j \in \{1, \ldots, p, m\}$, define the counting processes:

$$\begin{aligned} N_{ij}(t) &= 1(X_i \leq t, D_i = j) \text{ for } j \neq m, \\ N_{im}(t) &= 1(X_i \leq t, \delta_i = 1, D_i = 0). \end{aligned}$$

- In order to simplify the notation, write $m \equiv p + 1$ and write Y_i at risk process defined by $Y_i(t) = 1(X_i \geq t)$.
- For $1 \leq i \leq n$ and $j \in \{1, \ldots, p + 1\}$ the processes M_{ij} defined by

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(s) \bar{\lambda}_{0j}(s) ds, \quad t \geq 0,$$

are $\mathbb{F}$-martingales with respect to the filtration $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ defined by:

$$\mathcal{F}_t = \sigma\{N_{ij}(s), Y_i(s); s \leq t; 1 \leq i \leq n, j \in \{1, \ldots, p + 1\}\}.$$

First estimation of the regression parameters

If τ is the right bound of the interval of estimation, one can easily estimate α by:

$$\hat{\alpha} = \hat{\alpha}(\tau) = \frac{\sum_{i=1}^{n} 1(D_i > 0; X_i \leq \tau)}{\sum_{i=1}^{n} 1(\delta_i = 1; X_i \leq \tau)} = \frac{\sum_{j=1}^{p} N_{.j}(\tau)}{N_{..}(\tau)},$$

Transitions $0 \longrightarrow j$, for $j = 1, \ldots, p$, allow us to estimate β_j.

Transitions $0 \to j$, for $j = 1, \ldots, p$, allow us to estimate β_j. Following Lin et Ying (1994), an estimator of β_j is the explicit solution of the equation $\mathcal{U}_j(\beta, \tau) = 0$ where

$$\mathcal{U}_j(\beta, \tau) = \sum_{i=1}^{n} \int_0^{\tau} \left[Z_i - \bar{Z}(s) \right] \left[dN_{ij}(s) - \hat{\alpha}\beta^T Z_i Y_i(s) ds \right],$$

with

$$\bar{Z}(s) = \frac{\sum_{i=1}^{n} Y_i(s) Z_i}{\sum_{i=1}^{n} Y_i(s)}.$$

Also $0 \longrightarrow p+1$ one can estimate β_{p+1} with $\hat{\beta}_{p+1}$ solution of $\mathcal{U}_m(\beta, \tau) = 0$ where

$$\mathcal{U}_m(\beta, \tau) = \sum_{i=1}^{n} \int_0^{\tau} \left[Z_i - \bar{Z}(s) \right] \left[dN_{i,p+1}(s) - (1 - \hat{\alpha}) \beta^T Z_i Y_i(s) ds \right].$$

Improving the estimation of the regression parameters

We have obtained an estimator $\hat{\beta}_j$ of β_j, for $j = 1, \ldots, p, p+1$.

Improving the estimation of the regression parameters

We have obtained an estimator $\hat{\beta}_j$ of β_j, for $j = 1, \ldots, p, p+1$.

But $\hat{\beta}_{p+1}$ is an estimator of $\beta_{p+1} = \beta_1 + \cdots + \beta_p$. Thus, it seems natural to use it in order to improve the estimation of the β_js.

Improving the estimation of the regression parameters

We have obtained an estimator $\hat{\beta}_j$ of β_j, for $j = 1, \ldots, p, p+1$.

But $\hat{\beta}_{p+1}$ is an estimator of $\beta_{p+1} = \beta_1 + \cdots + \beta_p$. Thus, it seems natural to use it in order to improve the estimation of the β_js.

We are looking for an estimator $(\tilde{\beta}_1^T, \ldots, \tilde{\beta}_p^T)^T$ such that:

$$\begin{pmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_p \end{pmatrix} = H \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \\ \hat{\beta}_{p+1} \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} & \cdots & H_{1p+1} \\ H_{21} & H_{22} & \cdots & H_{2p+1} \\ \vdots & \vdots & & \vdots \\ H_{p1} & H_{p2} & \cdots & H_{pp+1} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \\ \hat{\beta}_{p+1} \end{pmatrix},$$

where, for $1 \leq i \leq p$ and $1 \leq j \leq p+1$, the H_{ij} are real matrices with dimension $p \times p$.

How to choose H?

How to choose H?

H has to be such that :

$$H \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

for all $1 \leq i \leq p$ and $\beta_i \in \mathbb{R}^p$.

How to choose H?

H has to be such that :

$$H \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

for all $1 \leq i \leq p$ and $\beta_i \in \mathbb{R}^p$.

Moreover we want H to minimize the function $\hat{q}(H)$ defined by

$$\hat{q}(H) = \text{trace}(H\hat{\Sigma}H^T),$$

where $\hat{\Sigma}$ is an estimator of the asymptotic covariance of $(\hat{\beta}_1^T, \ldots, \hat{\beta}_p^T, \hat{\beta}_{p+1}^T)^T$.

Let us denote by $\hat{H}$ the matrix which minimize $\hat{q}(H)$ and

$$\tilde{\beta}_i = \sum_{j=1}^{p+1} \hat{H}_{ij}\hat{\beta}_j$$

the final estimator of the β_is, for $i = 1, \ldots, p$.

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Regression parameters

Theorem

Under some hypothesis, the random vector $\sqrt{n}(\hat{\beta} - \beta)$ *is asymptotically mean zero gaussian with* $\Sigma(\tau)$.

Regression parameters

Theorem

Under some hypothesis, the random vector $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically mean zero gaussian with $\Sigma(\tau)$.

With the notation $\tilde{\beta} = \hat{H}\hat{\beta}$ and $\beta^* = (\beta_1^T, \ldots, \beta_p^T)^T$ we have the following result concerning the improve estimator.

Theorem

Under the same hypothesis, $\sqrt{n}(\tilde{\beta} - \beta^*)$ is asymptotically mean zero gaussian with minimal covariance matrix.

Functional parameters

We know how to estimate the cumulative hazard rate function as well as the survival function associated to each lifetimes T_i. For example, we can estimate Λ_j by

$$\hat{\Lambda}_j(t) = \frac{1}{\hat{\alpha}} \int_0^t \frac{dN_{ij}(s)}{Y(s)}$$

and $\Lambda_m(t) = \sum_{j=1}^p \Lambda_j(t)$ by

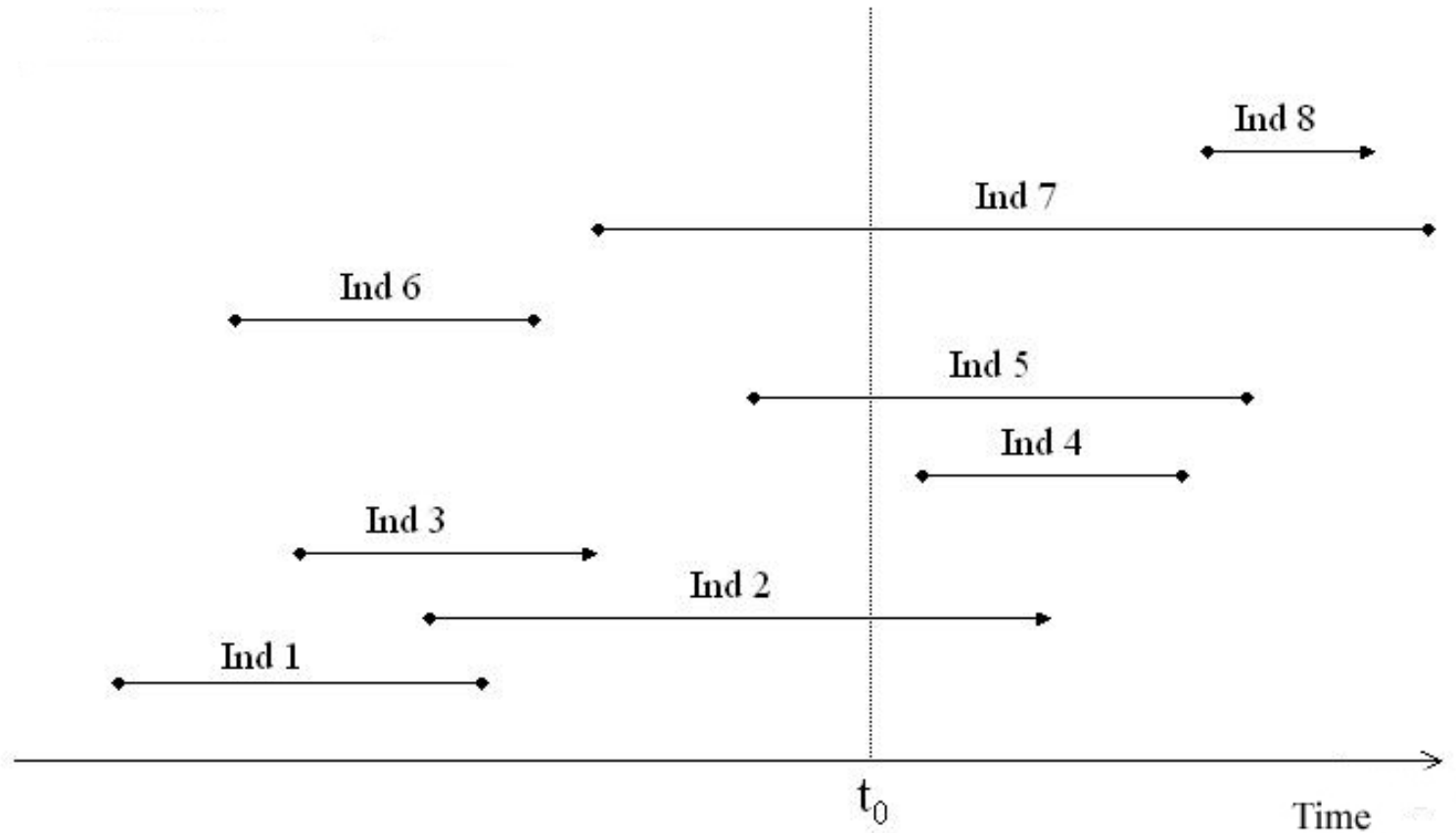
$$\hat{\Lambda}_m(t) = \frac{1}{1-\hat{\alpha}} \int_0^t \frac{dN_{im}(s)}{Y(s)},$$

where $Y(s) = \sum_{i=1}^n 1(X_i \geq s)$ and $\hat{\alpha}$ is as previously.
The large sample behaviour of these estimators and of their improved versions are also obtained.

Plan

1. Some reminders on Survival Analysis
   - Classical Survival Analysis
   - Survival Analysis with Competing Risks
2. Recurrent events and competing risks
   - Motivation
   - A recurrent events model with competing risks
   - Test of an increasing occurrence rate
   - Testing equality between two mean frequency functions
3. Competing risks, Additive risks and missing causes
   - Additive risks model
   - Censoring and missing causes of death
   - Estimation
   - Large sample behaviour
4. Length biased observation

Instead of following all the individuals, we decide (or we are only able) to follow individuals alive at a given time t_0.



What is the matter with such observations?

Let us denote by X_j^{In}, for $j = 1, \ldots, p$, (resp. X_j, for $j = 1, \ldots, p$) the lifetimes associated to the p causes of death in the initial population (resp. the sampled population). Thus,

$$\mathcal{L}(X_1, ..., X_p) \neq \mathcal{L}(X_1^{In}, ..., X_p^{In})$$

Our aim: Estimating the functions of interest of the initial population based of its length biased observation

Key relations

One can show that, under some conditions, we have:

$$F_j^{In}(t) = -\frac{\int_0^t \frac{1}{x} dF_j(x)}{\int_0^\infty \frac{1}{x} d\bar{F}(x)},$$

$$F^{In}(t) = \frac{\int_0^t \frac{1}{x} d\bar{F}(x)}{\int_0^\infty \frac{1}{x} d\bar{F}(x)},$$

for $j = 1, \ldots, p$.

Estimation

Plug-in estimation are easy

$$\hat{F}_j^{In}(t) = -\frac{\int_0^t \frac{1}{x} d\hat{F}_j(x)}{\int_0^\infty \frac{1}{x} d\hat{\bar{F}}(x)}$$

$$\hat{F}^{In}(t) = \frac{\int_0^t \frac{1}{x} d\hat{\bar{F}}(x)}{\int_0^\infty \frac{1}{x} d\hat{\bar{F}}(x)},$$

where $\hat{\bar{F}}$ and $\hat{F}_j$ are the estimators based on the length biased observation of respectively the survival function and the CIF.

Theorem

Assume
$\mathcal{A}$: $\int_0^\infty \frac{dF(x)}{\bar{G}(x-)} < \infty$.
We have the following convergence in $\mathbb{D}^3[0,\infty]$

$$\sqrt{n}\begin{pmatrix} \hat{\bar{F}} - \bar{F} \\ \hat{F}_1 - F_1 \\ \hat{F}_2 - F_2 \end{pmatrix} \xrightarrow{\mathcal{D}} Z^\infty = \begin{pmatrix} Z_0^\infty \\ Z_1^\infty \\ Z_2^\infty \end{pmatrix}$$

where Z^∞ *is the process* Z *extended* $[0,\infty]$.

Theorem

Under hypothesis $\mathcal{A}$, we have the following weak convergence in $\mathbb{D}^3[0,\infty]$:

$$\sqrt{n}(\widehat{\bar{F}}^{In} - \bar{F}^{In}, \widehat{F}_1^{In} - F_1^{In}, \widehat{F}_2^{In} - F_2^{In}) \xrightarrow{\mathcal{D}} L = (L_0, L_1, L_2),$$

where the limiting process is a mean zero gaussian defined by:

$$L_0(\cdot) = \frac{\int_0^{\cdot} \frac{1}{x} dZ_0^{\infty}(x)}{\int_0^{\infty} \frac{1}{x} d\bar{F}(x)} - F^{In}(\cdot) \frac{\int_0^{\infty} \frac{1}{x} dZ_0^{\infty}(x)}{\int_0^{\infty} \frac{1}{x} d\bar{F}(x)}$$

et, pour $k = 1, 2$,

$$L_k(\cdot) = F_k^{In}(\cdot) \frac{\int_0^{\infty} \frac{1}{x} dZ_0^{\infty}(x)}{\int_0^{\infty} \frac{1}{x} d\bar{F}(x)} - \frac{\int_0^{t} \frac{1}{x} dZ_k^{\infty}(x)}{\int_0^{\infty} \frac{1}{x} d\bar{F}(x)}.$$