DURATION MODELS

UNIL | Université de Lausanne

Lecture Notes 2015-2016

Part 4

Non-parametric models

Frédéric PLANCHET / Olivier DURAND

Version 3.2

April 2016

Table of content

1. Introduction

1.1. General information

In some situations, one would rather not make any *a priori* assumptions on the form of the survival distribution; one therefore seeks to directly estimate this function, in a space of infinite dimension; this framework of functional estimation is the field of non-parametric estimation.

Provided one disposes of data in sufficient quantity, one can obtain reliable estimates of the survival function as well as associated expressions.

In the usual context of a non-censored iid sample $(T_1, \ldots, T_n)$, one has the empirical estimator of the distribution function $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i \le t\}}$. This estimator has a certain number of "good well-known properties": it is without bias, convergent and asymptotically Gaussian. More precisely, convergence is uniform in the almost-certain sense, and one has the following "central limit theorem":

$$\sqrt{n}\left(F_n - F\right) \to W$$

Where W is a centered Gaussian process of covariance $\rho(s,t) = F(s) \wedge F(t) - F(s)F(t)$. This result rises directly from the theorem of Donsker in the case of the uniform distribution[1] and owing to the fact that $F(T)$ follows a uniform distribution on $[0,1]$.

The purpose of empirical estimation in duration models is to seek an estimator that verifies equivalent properties in the presence of censoring. In order to do so, one commences by introducing the presentation of duration models starting from point processes, which then facilitates obtaining results *via* asymptotic results on the martingales.

1.2. Notations

In the following, one notes F the cumulative distribution function of the non-censored model, G the cumulative distribution function of censoring and $T = X \wedge C$ the censored variable. One also notes:

$$S_0(t) = P(T > t, D = 0),\ S_1(t) = P(T > t, D = 1) \text{ and}$$
$$S_c(t) = S_0(t) + S_1(t) = P(T > t) = (1 - F(t))(1 - G(t)).$$

2. Duration models and point processes

The study of a survival duration is generally carried out by studying the distribution of variable X, associated with the survival function S. One proposes here to reason differently and to consider the point process naturally associated with X, $N(t)$ equal to 0 as long as the

[1] The limiting process being then the Brownian bridge, a centered Gaussian process of covariance $s \wedge t - st$.

event has not taken place, then 1 once it has: $N(t) = 1_{\{X \le t\}}$. When taking into account censoring, one builds in the same way $N^1(t) = 1_{\{T \le t, D=1\}}$ the process of the non-censored exits[2].

The presentation made here is heuristic and its purpose is to provide an understanding of mechanisms at play. The reader interested by a rigorous mathematical formalisation of these tools may refer to the article of Gill [1980] or to the book of Fleming and Harrington [1991], or for a presentation in French to the work of Dacunha-Castelle and Duflo [1983].

This approach largely uses the theory of martingales, whose significant results are reminded hereafter.

2.1. Reminders on martingales

It is said that a process (M_t) adapted to a filtration $(F_t)_{t \ge 0}$ is a martingale if it is of continuous trajectory on the right with limits on the left (cor-lol), and verifies:

$$E(|M_t|) < \infty \ \forall t \ge 0 \text{ and } E(M_t | F_s) = M_s \ \forall s \le t.$$

A martingale can be seen as a process of errors, in the way that on the one hand its expectancy is constant (one will thus always be able to assume that it is null) and on the other hand its increments are non-correlated:

$$\mathbf{cov}(M_t - M_s, M_v - M_u) = 0, 0 \le s \le t \le u \le v.$$

If the condition of constancy of the conditional expectancy is weakened and if the process is increasing in conditional expectancy in a way that $E(M_t | F_s) \ge M_s \ \forall s \le t$, then it is said that M is a sub-martingale. By the inequality of Jensen, if M is a martingale then M^2 is a sub-martingale since $E(M_t^2 | F_s) \ge (E(M_t | F_s))^2 = M_s^2 \ \forall s \le t$.

In order to push formalisation further, it is necessary to introduce a new definition.

<u>Definition</u>: A *predictable process* is a measurable random variable defined on the combined space $(]0,+\infty] \times \Omega, \mathrm{P})$ with tribe P generated by sets of the form $]s,t] \times \Gamma$, with $\Gamma \in F_s$.

The tribe of predictable events is generated by processes adapted to filtration $(F_{t-})_{t \ge 0}$ with $F_{t-} = \underset{s<t}{\vee} F_s$ and with continuous trajectories on the left.

In an intuitive way, one can say that a predictable process is a process which value in t is known “right before” t. Thus a continuous on the left (and adapted) process is predictable because of the continuity property.

These various tools lead to the decomposition of Doob-Meyer of a cor-lol adapted[3] process X, which expresses that such a process is the difference between two (local) sub-martingales if and only if there exists a unique decomposition of X in the form $X = A + M$ with A a

[2] Notations of lecture notes on parametric models are used, with X the non-censored variable, and the couple (T,D) in right-censoring situation.

[3] See for example DACUNHA-CASTELLE and DUFLO [1983].

predictable process with limited variation (in the sense that $\int_0^t |dA_s| = \sup_D \sum \left| A_{t_i} - A_{t_{i-1}} \right| < \infty$ with D the set of subdivisions of $[0,t]$) and M a (local) centered martingale.

One can deduce of this, in particular, that if M is a martingale, M^2 has a predictable compensator, noted $\langle M \rangle$ (that one will take care not to confuse with the quadratic variation $[M]$).

2.2. Application to duration models

Let us recall the definition of a point process.

Definition: a point process $\left(N(t), t \geq 0\right)$ is an integer values process adapted to a filtration $(F_t)_{t\geq 0}$ such that $N(0) = 0$, $N(t) < \infty$ almost surely and such that the trajectories are continuous on the right, piecewise constant and only show jumps of amplitude +1. In practice one will often consider for $(F_t)_{t\geq 0}$ the natural filtration associated with *N*, that is to say $F_t = \sigma\left(N(u), 0 \leq u \leq t\right) \vee \mathrm{N}$ with N the P-negligible events.

The Poisson process provides an example of point process; the $N(t)$ process introduced above is a simple case in which the process jumps only once.

Point processes show positive and increasing trajectories, thus with limited variation, and one can then define for an adapted process $X(t)$ the integral $\int_0^t X(u)dN(u)$ as a Stieljes integral, trajectory by trajectory. For example, in the presence of censoring the process of non-censored events $N^1(t) = 1_{\{T \leq t\ ,\ D=1\}}$ can be written:

$$N^1(t) = \int_0^t C(u)dN(u)$$

with $C(u) = 1_{[0,C]}(s)$. Censoring therefore acts like a filter. Since a point process is a sub-martingale (since it is increasing), its predictable compensator is associated to it, which is therefore a predictable increasing process, so that the difference between the point process and its compensator is a martingale. In a more formal way there is the following result.

Proposal: If a point process $\left(N(t), t \geq 0\right)$ adapted to a filtration $(F_t)_{t\geq 0}$ is such that $E[N(t)] < \infty$, then there exists a unique increasing and continuous on the right process Λ, such that $\Lambda(0) = 0$, $E[\Lambda(t)] < \infty$ and $M(t) = N(t) - \Lambda(t)$ is a martingale.

When Λ can be put in the form $\Lambda(t) = \int_0^t \lambda(u)du$, the process λ is called the intensity of the point process. For example, the compensator of a homogeneous Poisson process is $\Lambda(t) = \lambda t$, or, in an equivalent way, the intensity of a homogeneous Poisson process is constant and equal to λ.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} \boldsymbol{1}_{]t;\infty[}(T_x)$$

From a heuristic point of view, the decomposition $N(t) = \Lambda(t) + M(t)$ expresses that the process N “oscillates” around the predictable trend Λ so that the difference between the process of interest N and its trend is comparable to a residue, the variations of which are controlled. The equation $N(t) = \Lambda(t) + M(t)$ can thus be read as “observations = model + term of error”. One has in particular $E(N_t) = E(\Lambda_t)$.

One now seeks to determine the predictable compensator of the process $N(t) = 1_{\{X \le t\}}$.

One notes $N(t-) = \lim\limits_{u \uparrow t} N(u)$ the limit on the left of $N(t)$ and one is interested in the distribution of the random variable $P\left(dN_t = 1 \middle| N(t-)\right)$, while having noted formally $dN_t = N(t+dt) - N(t)$ with dt “small”. The random variable $dN(t)$ can only take values 0 and 1. By definition of the survival function and hazard function, one has:

$$P\left(dN_t = 1 \middle| N(t-)\right) = h(t)dt \text{ with the probability } S(t)$$

and

$$P\left(dN_t = 1 \middle| N(t-)\right) = 0 \text{ with the probability } 1 - S(t).$$

Indeed, if $N(t-) = 1$, the exit already occurred and the process cannot jump any more. This event occurs with the probability $1 - S(t)$. The process N can only jump between t and $t + dt$ only if $N(t-) = 0$ (event of probability $S(t)$) and the probability of jump is $h(t)dt$. One poses then $\lambda(t) = h(t)1_{\{X \ge t\}}$, product of the hazard function in t and indicator of presence right before t, $Y(t) = 1_{\{X \ge t\}}$. The process $\lambda(t)$ is predictable and $Y(t) = 1$ is equivalent to $N(t-) = 0$. Therefore $P\left(dN_t = 1 \middle| N(t-)\right) = \lambda(t)dt$, or in an equivalent way $E\left(dN_t \middle| N(t-)\right) = \lambda(t)dt$. The remarks above imply that:

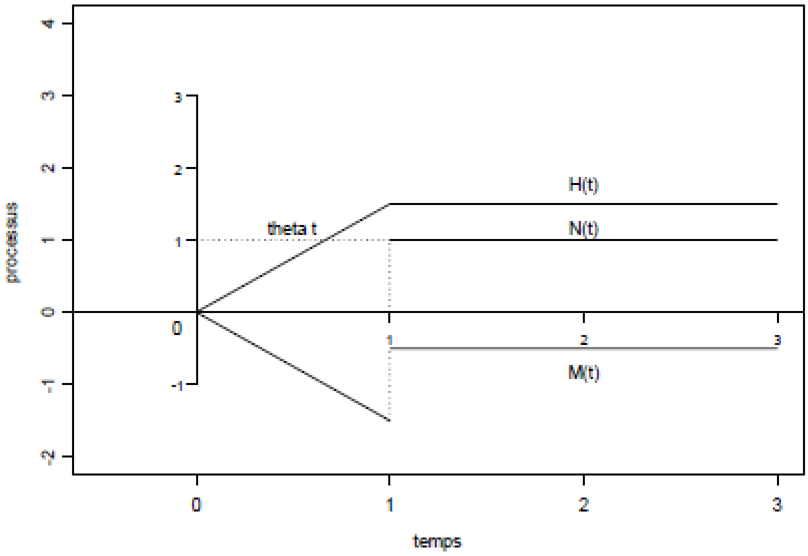
$$M(t) = N(t) - \int_0^t \lambda(u)du = N(t) - \int_0^t h(u)Y(u)du = N(t) - H(t \wedge T)$$

is a centered martingale since $E\left(dM_t \middle| N(t-)\right) = 0$ and since the intensity of process N can be calculated according to:

$$\lambda(t) = \lim_{u \to 0^+} \frac{1}{u} P\left[N(t+u) - N(t) = 1 \middle| F_{t-}\right].$$

The process $\lambda(t)$ is thus the intensity of process $N(t)$, which is random. Conditionally to the “immediate past”, the increase in $N(t)$ between t and $t + dt$ thus follows a distribution of Bernouilli of parameter $\lambda(t)dt$.

As an illustration one finds, in the case of an exponential distribution, the following paths of *N*, *M* and *H*:



One can show just as well that the predictable compensator of the process of non-censored events $N^1(t) = 1_{\{T \le t, D=1\}}$ is written:

$$\Lambda^1(t) = \int_0^t R(u)h(u)du ,$$

with $R(t) = 1_{\{T \ge t\}}$ the indicator of presence at risk before t (*i.e.* the function taking value 1 if the individual is neither dead nor censored; one indeed recalls that since $T = X \wedge C$, $\{T \ge t\} = \{X \ge t , C \ge t\}$). One thus passed from the statistical model where one gave oneself the couple (T, D) as observed information, to the model made up of (N^1, R).

In the case of a population, of which individuals are all supposed to have the same hazard function *h*, one associates to each population member a process of non-censored event $N_i^1(t) = 1_{\{T_i \le t , D_i = 1\}}$ as well as the indicator of presence under risk, counting individuals neither dead nor censored $R_i(t) = 1_{\{T_i \ge t\}}$ and one builds the aggregate processes $\overline{R}(t) = \sum_{i=1}^{n} R_i(t)$ and $\overline{N}^1(t) = \sum_{i=1}^{n} N_i^1(t)$. They respectively count population under risk and the number of non-censored events which have occurred.

One is thus in the presence of a model with “multiplicative intensity” (Aalen [1978]), in the sense that the counting process $\overline{N}^1$ has an intensity which can be put in the form:

$$\lambda(t) = \overline{R}(t)h(t)$$

with $\overline{R}$ an observable process (predictable) and *h* the hazard function, unknown and to be estimated. These processes will make it possible to introduce in a simple way the usual non-parametric estimators.

3. Non-parametric estimators in duration models

One will note as an introduction that the distribution can be, as seen above, characterised by various functions: hazard function, cumulative hazard function, cumulative distribution function, probability density function, etc. It is obvious that estimating the hazard function is of the same degree of complexity as estimating the probability density function; one will therefore naturally turn towards the empirical estimation of cumulative hazard or survival function, *a priori* more simple. Estimating the hazard function will then require to regularise the estimator of the cumulative hazard function, which will generally be discontinuous. These aspects are not covered here[4]. The two principal estimators in this context are the estimator of Nelson-Aalen of the cumulative hazard rate and the Kaplan-Meier estimator of the survival function.

3.1. The estimator of Nelson-Aalen[5] of the cumulative hazard rate

One recalls that the cumulative hazard function is defined, in the general case, by $H(t)=\int_0^t \frac{dS(u)}{S(u-)}$, expression which leads to the traditional expression in the case of a continuous model $H(t)=\int_0^t h(u)\,du$ where $h(t)=-\frac{d}{dt}\ln S(t)$.

3.1.1. General presentation

The fact that $M(t)=\bar{N}^1(t)-\int_0^t \bar{R}(u)h(u)du$ is a centered martingale suggests proposing $\bar{N}^1(t)$ as estimator of $\int_0^t \bar{R}(u)h(u)du$. But then the process $\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)}dM(u)$ is also a martingale and one has by construction of *M*:

$$\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)}dM(u)=\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)}d\bar{N}^1(u)-\int_0^t h(u)\,du=\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)}d\bar{N}^1(u)-H(t)$$

provided that *t* is such that $\bar{R}(t)>0$. Thus $\hat{H}(t)=\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)}d\bar{N}^1(u)$ is a natural estimator of *H*, called the estimator of Nelson-Aalen. It was initially proposed by Nelson [1972]. One can give another justification of it, by noticing that the cumulative hazard function verifies, by construction:

$$H(u+du)-H(u)\approx h(u)\,du$$

[4] The interested reader can refer to Droesbeke and Al [1989].

[5] The original study of Nelson-Aalen relates to the operating life of ventilators.

and $h(u)du = P(\textit{sortie entre } u \textit{ et } u+du \mid \textit{en vie en } u)$; a natural estimator of this quantity is therefore $\frac{\bar{N}^1(u+du)-\bar{N}^1(u)}{\bar{R}(u)} = \frac{d\bar{N}^1(u)}{\bar{R}(u)}$ if $\bar{R}(u) > 0$, so that while summing on a $[0,t]$ cut, sufficiently fine so that each subdivision contains no more than one jump, one obtains:

$$\hat{H}(t) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u),$$

Which is indeed the preceding expression. As the processes considered here are purely with jumps one can, while noting $\Delta\bar{N}(t) = \bar{N}(t) - \bar{N}(t-)$, put this expression in the form:

$$\hat{H}(t) = \sum_{\{i/T_i \le t\}} \frac{\Delta\bar{N}(T_i)}{\bar{R}(T_i)}$$

By posing $d(t) = \Delta\bar{N}(t)$ the number of deaths in *t* and $r(t) = \bar{R}(t)$ the population under risk right before *t*, one can thus rewrite the equation above in the following intuitive form:

$$\hat{H}(t) = \sum_{\{i/T_i \le t\}} \frac{d(T_i)}{r(T_i)} = \sum_{\{i/T_i \le t\}} \frac{d_i}{r_i} = \sum_{T_i \le t} \frac{d_i}{n-i+1},$$

the second equality being true only if there is no *ex-œquo*. The function $\hat{H}$ is continuous to the right. One can verify that this estimator is biased and underestimates in average the cumulative hazard function. Indeed,

$$\hat{H}(t) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} \left(dM(u) + \bar{R}(u)h(u)du\right).$$

Since *M* is a martingale, it comes in taking the expectancy of the two members of the equation above $E\left[\hat{H}(t)\right] = \int_0^t E\left(1_{\{\bar{R}(u)>0\}}\right)h(u)du$. But:

$$E\left[1_{\{\bar{R}(u)>0\}}\right] = P\left[\bar{R}(u) > 0\right] = 1 - P\left[\bar{R}(u) = 0\right].$$

One finally deduces:

$$E\left[\hat{H}(t)\right] = \int_0^t h(u)du - \int_0^t P\left[\bar{R}(u) = 0\right]h(u)du = H(t) - \int_0^t P\left[\bar{R}(u) = 0\right]h(u)du$$

which implies that $E\left\lfloor\hat{H}(t)\right\rfloor \le H(t)$: the estimator of Nelson-Aalen indeed tends to underestimate the model’s cumulative hazard function.

3.1.2. Variance of the estimator of Nelson-Aalen

It results from the approximation carried out in the previous section that the increase in the process $\bar{N}^1(t)$ between *t* and $t+u$ approximately follows a Poisson distribution of parameter

$\int_t^{t+u} \bar{R}(s)h(s)ds \approx \bar{R}(t)h(t)u$. Indeed, one had seen that conditionally to the "immediate past", the increase in $N^1(t)$ between t and $t+dt$ follows a distribution of Bernouilli of parameter $h(t)R(t)dt$. The sum on the various individuals thus leads to a binomial variable, which can be approached by a Poisson distribution in choosing $dt = \frac{u}{n}$. It is therefore deduced that, conditionally to $\bar{R}(t)$, $V\left(\frac{\bar{N}^1(t+u)-\bar{N}^1(t)}{\bar{R}(t)}\right) \approx \frac{h(t)u}{\bar{R}(t)}$; however one saw in the previous section that $h(t)u$ could be estimated by $\frac{\bar{N}^1(t+u)-\bar{N}^1(t)}{\bar{R}(t)}$, leading to the estimator of the variance $\hat{V}\left(\frac{\bar{N}^1(t+u)-\bar{N}^1(t)}{\bar{R}(t)}\right) \approx \frac{\bar{N}^1(t+u)-\bar{N}^1(t)}{\bar{R}(t)^2}$, which finally results in proposing as an estimator of the variance of $\hat{H}$:

$$\hat{V}\left(\hat{H}(t)\right) = \sum_{\{i/T_i \leq t\}} \frac{\Delta\bar{N}^1(T_i)}{\bar{R}(T_i)^2}$$

which can be written with simplified notations, in the absence of *ex aequo*:

$$\hat{V}\left(\hat{H}(t)\right) = \sum_{\{i/T_i \leq t\}} \frac{\Delta\bar{N}^1(T_i)}{\bar{R}(T_i)^2}.$$

3.1.3. An example

Freireich, in 1963, carried out a therapeutic test to compare remission durations, in weeks, of patients suffering from leukemia depending on whether or not they took a drug called 6 M-P; the reference group received a placebo. The results are the following[6]:

6 M-P: 6,6,6,6+, 7.9+, 10.10+, 11+, 13,16,17+, 19+, 20+, 22,23,25+, 32+, 32+, 34+, 35+.

Placebo: 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23.

Numbers followed by the sign + correspond to censored data. The application of the formulas above to these data leads to:

[6] Duration of remission, in weeks.

Relapses	t_i	r_i	d_i	d_i/r_i	$\hat{H}(t)$	d_i/r_i^2	$\sigma^2(\hat{H}(t))$	$\sigma(\hat{H}(t))$
1-2-3	6	21	3	0.143	0.143	0.007	0.007	0.082
5	7	17	1	0.059	0.202	0.003	0.010	0.101
7	10	15	1	0.067	0.268	0.004	0.008	0.089
10	13	12	1	0.083	0.352	0.007	0.011	0.107
11	16	11	1	0.091	0.443	0.008	0.015	0.123
15	22	7	1	0.143	0.585	0.020	0.029	0.169
16	23	6	1	0.167	0.752	0.028	0.048	0.220

for the group treated with 6 M-P and for the group taking the placebo one obtains:

Relapses	t_i	r_i	d_i	d_i/r_i	$\hat{H}(t)$	d_i/r_i^2	$\sigma^2(\hat{H}(t))$	$\sigma(\hat{H}(t))$
1-2	1	21	2	0.095	0.095	0.005	0.005	0.067
3-4	2	19	2	0.105	0.201	0.006	0.010	0.100
5	3	17	1	0.059	0.259	0.003	0.014	0.116
6-7	4	16	2	0.125	0.384	0.008	0.021	0.146
8-9	5	14	2	0.143	0.527	0.010	0.032	0.178
10-11-12-13	8	12	4	0.333	0.861	0.028	0.059	0.244
14-15	11	8	2	0.250	1.111	0.031	0.091	0.301
16-17	12	6	2	0.333	1.444	0.056	0.146	0.382
18	15	4	1	0.250	1.694	0.063	0.209	0.457
19	17	3	1	0.333	2.027	0.111	0.320	0.565
20	22	2	1	0.500	2.527	0.250	0.570	0.755
21	23	1	1	1.000	3.527	1.000	1.570	1.253

It is noted in particular that the cumulative hazard rate of the treated group is appreciably lower than that of the untreated group, which allows the guessing of a certain effectiveness of the treatment. This point will be shown *infra*.

3.1.4. Asymptotic properties

The estimator of Nelson-Aalen is asymptotically Gaussian; more precisely there is the following result.

Proposal: if the cumulative distribution functions of survival and censoring do not have any common discontinuity, then:

$$\sqrt{n}\left(\hat{H}-H\right) \to W_H$$

with W_H a centered Gaussian process of covariance $\rho(s,t) = \int_0^{s\wedge t} \frac{dS_1(u)}{S_c(u)^2}$ with $S_c(t) = \left(1-F(t)\right)\left(1-G(t)\right)$ and $S_1(t) = P(T > t, D = 1)$.

3.2. The Kaplan-Meier estimator of the survival function

One can notice that the estimator of Nelson-Aalen of the cumulative hazard rate leads to a natural estimator of the survival function, by using the relationship $S(t) = \mathbf{exp}\left(-H(t)\right)$; one can thus propose for estimator of the survival function:

$$\hat{S}_{HF}(t) = \mathbf{exp}\left(-\hat{H}_{NA}(t)\right).$$

This estimator is the estimator of Harrington and Fleming; its variance can be obtained by the Delta method which, under reasonable conditions of regularity of the function *f*, makes it possible to write that $V\left(f(X)\right) \approx \left(\frac{df}{dx}\left(E(X)\right)\right)^2 V(X)$. Indeed, if $X = \mu + \sigma Z$ with σ small and *Z* centered and reduced, one notices that for a sufficiently regular function $x \to f(x)$, by carrying out the limited development $f(\mu+h) \approx f(\mu) + h\frac{df}{dx}(\mu)$, one finds that $V\left(f(X)\right) \approx V\left(f(\mu) + \sigma Z \frac{df}{dx}(\mu)\right) = \sigma^2 \frac{df}{dx}(\mu)^2$. In taking $f(x) = e^{-x}$, one finds that $V\left(\hat{S}\right) \approx e^{-2E\left(\hat{H}\right)} V\left(\hat{H}\right) \approx \hat{S}^2 V\left(\hat{H}\right)$, which leads to the estimator of the variance:

$$\hat{V}\left(\hat{S}_{HF}(t)\right) = \mathbf{exp}\left(-2\sum_{\{i/t_i \le t\}} \frac{d(t_i)}{n-i+1}\right) \sum_{\{i/t_i \le t\}} \frac{d(t_i)}{(n-i+1)^2} = \hat{S}_{HF}(t)^2 \times \sum_{\{i/t_i \le t\}} \frac{d(t_i)}{(n-i+1)^2}.$$

As it was shown that $E\left[\hat{H}_{NA}(t)\right] \le H(t)$ and that the function $g(x) = e^{-x}$ is convex, one deduces that:

$$E\left(\hat{S}_{HF}(t)\right) = E\left(g\left(\hat{H}_{NA}(t)\right)\right) \ge g\left(E\left(\hat{H}_{NA}(t)\right)\right) = \mathbf{exp}\left(-E\left(\hat{H}_{NA}(t)\right)\right) \ge \mathbf{exp}\left(-H(t)\right) = S(t),$$

in other words, the estimator of Harrington-Fleming of the survival function presents an over-estimating bias.

However, this estimator can be improved, which leads to the Kaplan-Meier estimator.

3.2.1. General presentation

The Kaplan-Meier estimator (Kaplan and Meier [1958]) can be introduced *via* point processes, by noticing that the model's basic survival function is the unique solution of the following integral equation:

$$S(t) = 1 - \int_0^t S(u-)h(u)du \, .$$

The equation above simply expresses the fact that the sum of survivors in *t* and individuals who have left before *t* is constant. When the survival function is continuous, the demonstration is immediate by carrying out the variable shift $v = \mathbf{ln}\, S(u)$, $dv = -h(u)du$.

Replacing $h(u)du$ by its estimator $\frac{d\overline{N}^1(u)}{\overline{R}(u)}$ introduced in the previous section, one can propose an estimator of the survival function by seeking a solution to the equation:

$$\hat{S}(t) = 1 - \int_0^t \hat{S}(u-)\frac{d\overline{N}^1(u)}{\overline{R}(u)} \, .$$

One can show that there exists a unique solution to this equation and one then obtains the Kaplan-Meier estimator of the survival function. If the existence is not simple to prove, unicity rises directly from the remark that if two estimators are solutions of the equation above then:

$$\hat{S}_1(t) - \hat{S}_2(t) = \sum_{T_i \le t} \left[\hat{S}_1(T_i -) - \hat{S}_2(T_i -) \right] \frac{d_i}{r_i}$$

and since $\hat{S}_1(0) - \hat{S}_2(0) = 0$, by recurrence $\hat{S}_1(t) - \hat{S}_2(t) = 0$ for any *t*. This estimator can be expressed using the estimator of Nelson-Aalen in the following way:

$$\hat{S}(t) = \prod_{s \le t} \left(1 - \Delta \hat{H}(s)\right)$$

Where $\Delta \hat{H}(s) = \hat{H}(s) - \hat{H}(s-)$. One can however propose a more intuitive explicit construction of this estimator, described *infra*.

The heuristic construction of the Kaplan-Meier estimator is based on the following remark: the probability of surviving beyond $t > s$ can be written:

$$S(t) = P(T > t \,|T > s)P(T > s) = P(T > t \,|T > s)S(s) \, .$$

One can renew the operation, which reveals products of terms in $P(T > t \,|T > s)$; if one chooses as conditioning instants the moments when an event occurs (exit or censoring), one is left estimating probabilities of the form:

$$p_i = P\left(T > T_{(i)} \,\middle|\, T > T_{(i-1)}\right)$$

p_i is the probability of surviving on the interval $\left]T_{(i-1)}, T_{(i)}\right]$ knowing that one was alive at the instant $T_{(i-1)}$. A natural estimator of $q_i = 1 - p_i$ is $\hat{q}_i = \frac{d_i}{r_i} = \frac{d_i}{n-i+1}$. One then observes that at instant $T_{(i)}$, and in the absence of *ex aequo*, if $D_{(i)} = 1$ then there is exit by death thus $d_i = 1$, and in the contrary case the observation is censored and $d_i = 0$. The Kaplan-Meier estimator is thus finally written:

$$\hat{S}(t) = \prod_{T_{(i)} \le t} \left(1 - \frac{1}{n-i+1}\right)^{D_{(i)}}.$$

In practice however one is confronted with the presence of *ex æquo*; one then supposes by convention that the non-censored observations always precede the censored ones. One obtains the following expression of the estimator:

$$\hat{S}(t) = \prod_{T_{(i)} \le t} \left(1 - \frac{d_i}{r_i}\right)$$

Comment n°1: here one works with the continuous to the right version of the survival function; some authors use the continuous to the left version. In which case the expressions above remain valid by replacing the term $T_{(i)} \le t$ by $T_{(i)} < t$.

Comment n°2: should there be arrivals in the course of the period (left truncations), the expression $\hat{S}(t) = \prod_{T_{(i)} \le t} \left(1 - \frac{d_i}{r_i}\right)$ remains valid by taking it into account in the calculation of r_i; there again, the survival function only jumps at the time of non-censored exits.

3.2.2. Comparison with the estimator of Harrington and Fleming

The two estimators are written respectively, after transformation by the logarithm $\mathbf{ln}\,\hat{S}_{KM}(t) = \sum_{T_{(i)} \le t} \mathbf{ln}\left(1 - \frac{d_i}{r_i}\right)$ and $\mathbf{ln}\,\hat{S}_{HF}(t) = -\sum_{T_{(i)} \le t} \frac{d_i}{r_i}$ thus

$$\mathbf{ln}\,\hat{S}_{KM}(t) - \mathbf{ln}\,\hat{S}_{HF}(t) = \sum_{T_{(i)} \le t} \left(\mathbf{ln}\left(1 - \frac{d_i}{r_i}\right) + \frac{d_i}{r_i}\right).$$

It is easily verified that the function $f(x) = \mathbf{ln}(1-x) + x$ is always negative and thus $\hat{S}_{KM}(t) \le \hat{S}_{HF}(t)$.

3.2.3. Examples

3.2.3.1. Freireich dataset

One takes the Freireich dataset used into 3.1.3 above, and one is interested in the comparison of results obtained by Kaplan-Meier and Nelson-Aalen methods; it is found that:

Relapses	t_i	r_i	d_i	d_i/r_i	$\hat{H}_{NA}(t)$	$\hat{S}_{KM}(t)$	$-\ln \hat{S}_{KM}(t)$
1-2-3	6	21	3	0.143	0.143	0.857	0.154
5	7	17	1	0.059	0.202	0.807	0.215
7	10	15	1	0.067	0.268	0.753	0.284
10	13	12	1	0.083	0.352	0.690	0.371
11	16	11	1	0.091	0.443	0.627	0.466
15	22	7	1	0.143	0.585	0.538	0.620
16	23	6	1	0.167	0.752	0.448	0.803

It is noted that the cumulative hazard rate obtained with Kaplan-Meier is higher than the cumulative hazard rate resulting from the estimator of Nelson-Aalen. If one calculates the estimator of Harrington and Fleming of the survival function $\hat{S}(t) = \mathbf{exp}\left(-\hat{H}_{NA}(t)\right)$, one notes just as well that it is systematically higher than the Kaplan-Meier estimator. Beyond the strictly statistical aspects, prudential considerations could therefore point towards the choice of one estimator or another.

3.2.3.2.Another example

Over 10 patients suffering from bronchi cancer, one observed the following survival durations, expressed in months[7]: 3/1/4 +/5/7+/9/8/10 +/11/13+. The Kaplan-Meier estimator of the survival function $S(t)$ is calculated in the following way:

t_i	r_i	d_i	Survival	Interval
0	10	0	100.0%	[01 [
1	10	1	90.0%	[1 3 [
3	9	1	80.0%	[3 5 [
5	7	1	68.6%	[5 8 [
8	5	1	54.9%	[8 9 [
9	4	1	41.1%	[9 11 [
11	2	1	20.6%	

3.2.4. Main properties

The Kaplan-Meier estimator has a certain number of “good properties” which makes it the natural generalisation of the empirical estimator of the cumulative distribution function in the presence of censoring: it is convergent[8], asymptotically Gaussian, coherent and is also a generalised maximum-likelihood estimator. However, this estimator is positively biased. The

[7] The sign + indicates a censored observation

[8] Provided the survival function and the distribution of censoring do not have common discontinuities.

Kaplan-Meier estimator is the only coherent estimator of the survival function (see Droesbeke and *al*. [1989] for the demonstration of this property).

The concept of "maximum-likelihood" must be adapted to the non-parametric context in the following way[9].

<u>Definition</u>: let Φ be a family of probabilities on $\Re^n$ (with the Borel tribe) not dominated; $\forall\ x \in \Re^n$ and $P_1, P_2 \in \Phi$, one writes $l(x, P_1, P_2) = \frac{dP_1}{d(P_1+P_2)}(x)$, it is then said that $\hat{P}$ is GMLE for P if $l\left(x, \hat{P}, P\right) \geq l\left(x, P, \hat{P}\right)$.

One can then show that the estimator $\hat{S}$ is GMLE for *S*, provided that the distributions of the non-censored lifetime and of the censoring are diffuse, and provided that the family Φ contains the distributions of probability charging the points (T_i, D_i). The other properties are detailed hereafter.

3.2.5. Variance of the Kaplan Meier estimator

One proposes here a heuristic justification of an estimator of the variance of the Kaplan-Meier estimator, the estimator of Greenwood.

The expression $\hat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right)$ makes it possible to write[10]:

$$\mathbf{ln}\left(\hat{S}(t)\right) = \sum_{T_{(i)} \leq t} \mathbf{ln}\left(1 - \frac{d_i}{r_i}\right) = \sum_{T_{(i)} \leq t} \mathbf{ln}\left(1 - \hat{q}_i\right).$$

If, as a first approximation, one assumes the independence of variables $\mathbf{ln}\left(1 - \hat{q}_i\right)$, as the distribution of $r_i \hat{p}_i$ is binomial with parameters $\left(r_i\ , p_i\right)$, one obtains through the delta method $V\left(f(X)\right) \approx \left(\frac{df}{dx}\left(E(X)\right)\right)^2 V(X)$:

$$V\left(\mathbf{ln}\,\hat{p}_i\right) \approx V\left(\hat{p}_i\right)\left[\frac{d}{dp}\mathbf{ln}\left(\hat{p}_i\right)\right]^2 = \frac{\hat{q}_i}{r_i\left(1 - \hat{q}_i\right)}$$

which leads to suggesting as an estimator of the variance of $\mathbf{ln}\,\hat{S}(t)$:

$$\hat{V}\left(\mathbf{ln}\,\hat{S}(t)\right) = \sum_{T_{(i)} \leq t} \frac{\hat{q}_i}{r_i\left(1 - \hat{q}_i\right)} = \sum_{T_{(i)} \leq t} \frac{d_i}{r_i\left(r_i - d_i\right)}$$

By applying the delta method again with the logarithmic function for *f*, one finally obtains:

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \gamma(t)^2$$

[9] One will see in 3.2.7 the link with maximum-likelihood in a parametric context.

[10] This formula provides an estimator of the cumulative hazard function called estimator of Breslow of *H*.

with $\gamma(t) = \sqrt{\sum_{T_{(i)} \le t} \frac{d_i}{r_i(r_i - d_i)}}$. This estimator is the estimator of Greenwood. It is consistent for the asymptotic variance of the Kaplan-Meier estimator. It allows, together with the asymptotic normality[11] of the Kaplan-Meier estimator, the calculation of (asymptotic) confidence intervals whose bounds are, for the value of survival at time $T_{(i)}$:

$$S_i \times \left(1 \pm u_{1-\frac{\alpha}{2}} \gamma\left(T_{(i)}\right)\right) = S_i \times \left(1 \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{d_1}{r_1(r_1 - d_1)} + \frac{d_2}{r_2(r_2 - d_2)} + \ldots + \frac{d_i}{r_i(r_i - d_i)}}\right)$$

In this way, one builds point intervals, with *t* fixed. One can then seek to build bands of confidence for the survival function. Nair proposes in 1984 (cf. Klein and Moeschberger [2005]) linear bands of confidence of the form:

$$\hat{S}(t)\left(1 \pm c_\alpha\left(a(t_m), a(t_M)\right)\gamma(t)\right)$$

with $a(t) = \frac{n \times \gamma(t)^2}{1 + n \times \gamma(t)^2}$ and where the confidence coefficients *cα (x1,x2)* are tabulated (they are provided in appendix of Klein and Moeschberger [2005]).

One can also observe that these formulas can be used to build confidence intervals for conditional rates of exit $\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)}$; indeed, one can deduct from $\hat{S}(x) = \prod_{T_{(i)} \le x} \left(1 - \frac{d_i}{r_i}\right)$ that $1 - \hat{q}(x) = \prod_{x < T_i \le x+1} \left(1 - \frac{d_i}{r_i}\right)$ and thus:

$$\hat{V}\left(\hat{q}(x)\right) = \left(1 - \hat{q}(x)\right)^2 \sum_{x < T_i \le x+1} \frac{d_i}{r_i(r_i - d_i)}$$

from which the expression of an asymptotic confidence interval is immediately yielded:

$$\hat{q}_{\pm}(x) = 1 - \left(1 - \hat{q}(x)\right) \times \left(1 \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\sum_{x < T_i \le x+1} \frac{d_i}{r_i(r_i - d_i)}}\right).$$

3.2.6. Asymptotic properties

The Kaplan-Meier estimator is asymptotically Gaussian; precisely there is the following result.

<u>Proposal</u>: if the cumulative distribution functions of survival and censoring do not have any common discontinuity, then:

$$\sqrt{n}\left(\hat{S} - S\right) \to W_S$$

with W_S a centered Gaussian process of covariance:

[11] See 3.2.6.

$$\rho(s,t) = S(s)S(t) \int_0^{s\wedge t} \frac{dF(u)}{(1-F(u))^2 (1-G(u))}.$$

In particular, when the model is non-censored (i.e. $G(u)=0$), one finds the traditional result presented in 1.1 above. The interest of results of convergence of the process itself – rather than for one fixed instant – is that one can deduce some asymptotic bands of confidence for the Kaplan-Meier estimator.

One can find in Gill [1980] a demonstration of the asymptotic normality of $\hat{S}_{KM}$, based on the theory of point processes. By noting $F = 1 - S$ and $\hat{F} = 1 - \hat{S}_{KM}$, the band of confidence which is obtained is written:

$$\liminf_{n\to\infty} P\left\{ \sup_{s\in[0,t]} \left| \frac{\hat{F}(s) - F(s)}{1-\hat{F}(s)} \right| \le \frac{\sqrt{\hat{V}(t)}}{1-\hat{F}(t)} x \right\} \ge \sum_{k=-\infty}^{\infty} (-1)^k \left[\Phi((2k+1)x) - \Phi((2k-1)x) \right]$$

where $\hat{V}(t) = \hat{S}_{KM}^2 \int_0^t \frac{d\bar{N}^1(u)}{\bar{R}(u)\left(\bar{R}(u) - \Delta\bar{N}^1(u)\right)}$ estimates the variance of the limiting Gaussian process W_S.

3.2.7. Discretised version: link with the parametric approach

Calculation of the Kaplan-Meier estimator implies that one disposes of individual data with the exact dates of occurrence of events; in practice, in addition to the fact that for large populations calculations can be heavy, this information is not always available. One then wishes to use this approach for data gathered by period, for example considering the month as the unit of time and counting exits month by month. It is the approach followed by the BCAC in France for disability[12] (decree of 1996).

Formally, if one considers the instants $t_1 < .. < t_N$ at which the exits occur (for example integer ages of death) and if one disposes of a sample of size n for which one observed a sequence (r_i, d_i) of individuals at risk as well as deaths at dates $t_1 < .. < t_N$, one can notice that D_i the number of exits on the interval $[t_i, t_{i+1}[$ follows a binomial distribution of parameters (r_i, h_i); h_i indicating the hazard rate at time t_i (homogeneous to one q_x).

The exits in the intervals $[t_i, t_{i+1}[$ being independent from one another, it is therefore found that the likelihood of this model is written:

$$L = \prod_{i=1}^{N} C_{r_i}^{d_i} h_i^{d_i} (1-h_i)^{r_i - d_i}.$$

Hence log-likelihood is written:

[12] http://www.ressources-actuarielles.net/C1256F13006585B2/0/A0D8FE4A9807886AC1257D11002A0BE9

$$\mathbf{ln}(L)=\sum_{i=1}^{N}\left[C_{r_i}^{d_i}+d_i\,\mathbf{ln}(h_i)+(r_i-d_i)\,\mathbf{ln}(1-h_i)\right]$$

and the first order conditions $\frac{\partial}{\partial h_i}\mathbf{ln}\,L=0$ lead to the estimators:

$$\hat{h}_i=\frac{d_i}{r_i}.$$

One is reminded of the estimator introduced in 3.2.1 above. For that approach to be relevant, it is advisable to make sure that the discretisation does not generate large bias on the estimation of exit rates: for example, in the case of sick leave, it is known that the exits are very numerous during the first month (in practice approximately 50 % of sick leaves last less than 30 days). Therefore, if a monthly step is adopted, one badly takes into account the high pace of exits during the first period; it would thus be advisable here to choose a smaller discretisation step. More generally, the above reasoning is relevant as long as the length of each interval considered is “small” in the light of the variation speed of the survival function.

4. Taking into account explanatory variables

When the studied population is heterogeneous, it is important to take into account specificities of each sub-group. By assuming that heterogeneity is the consequence of a blend of subpopulations each characterised by observable variables, one looks into modelling the hazard function that integrates the effect of explanatory variables. This question was already tackled in parametric and semi-parametric contexts (Cox model) – one is interested here in the non-parametric case.

This chapter is inspired by Martinussen and Scheike [2006] to which the reader will be able to refer for demonstrations. It is also specified that the practical application of the models presented here can be carried out using the *timereg* package of the software R, developed by these authors or by using the *survival* package.

4.1. The additive model of Aalen

The hazard function is supposed to be written:

$$h(t)=X^T(t)\beta(t)$$

with $X^T(t)=\left(X_1(t),\ldots,X_p(t)\right)$ a vector of explanatory variables (predictable) and $\beta(t)$ a p-dimensional process locally integrable. One can in an equivalent way say that the intensity of the underlying counting model is written:

$$\lambda(t)=R(t)X^T(t)\beta(t).$$

One has a set of observations $\left(N_i^1(t),R_i(t),X^i(t)\right)_{1\le i\le n}$ and one seeks to estimate the vector $\beta(t)$; in practice one will be able to easily build an estimator of $B(t)=\int_0^t\beta(u)du$ through the use of the remarks which follow.

One notes to reduce the formulas $\lambda(t)=(\lambda_1(t),\ldots,\lambda_n(t))^T$ and $N^1(t)=(N_1^1(t),\ldots,N_n^1(t))^T$, then $X(t)=(R_1(t)X^1(t),\ldots,R_n(t)X^n(t))^T$, which is a matrix of dimension $n \times p$. With these notations one gets, while defining by $\Lambda(t)=\int_0^t \lambda(u)du$ the vectorial process of the n cumulated intensities, $M(t)=N^1(t)-\Lambda(t)$ which is a martingale. By observing that:

$$dN^1(t)=X(t)\beta(t)dt+dM(t)=X(t)dB(t)+dM(t)$$

since the term $dM(t)$ is centered and the increments of the martingale are non-correlated, one can seek to estimate the increments $dB(t)$ by traditional techniques of linear regression. For that, one writes:

$$X^-(t)=\left(X^T(t)X(t)\right)^{-1}X^T(t),$$

if $X^T(t)X(t)$ is invertible and 0 if not. $X^-(t)$ is called the generalised inverse of *X*, which is a matrix of dimension $p \times n$ that verifies $X^-(t)X(t)=J(t)I_p$ with $J(t)$ taking value 1 if the inverse exists, and 0 if not. In practice when $X(t)$ is of full rank $X^T(t)X(t)$ is invertible and one simply has $X^-(t)X(t)=I_p$. It is then natural to propose for estimator of *B* the following process:

$$\hat{B}(t)=\int_0^t X^-(u)dN^1(u).$$

The fact that $\hat{B}(t)=\int_0^t J(s)dB(s)+\int_0^t X^-(s)dM(s)$ ensures that $\hat{B}$ estimates *B* essentially without bias; and one can moreover show under certain not very restrictive technical conditions that $\sqrt{n}\left(\hat{B}-B\right)$ converges in distribution as a process towards a centered Gaussian process which function of covariance can be calculated.

The calculation of the estimator $\hat{B}(t)=\int_0^t X^-(u)dN^1(u)$ consists in calculations of discrete sums at the instants of jump of the process $N^1(t)$. In a more precise way $\hat{B}(t)$ is a vector of size *p* and:

$$\hat{B}_j(t)=\sum_i \int_0^t X_{ji}^-(u)dN_i^1(u)$$

But $N_i^1(t)$ jumps no more than once at time T_i and the increment at this instant is 1 (if a jump occurs). The following expression is deduced:

$$\hat{B}_j(t)=\sum_{T_i \le t} X_{ji}^-(T_i)\times D_i.$$

Calculation thus requires the determination of $X^-(T_i) = \left(X^T(T_i)X(T_i)\right)^{-1} X^T(T_i)$ for all the non-censored exits.

4.2. Semi-parametric alternative: the model of Lin and Ying

In insurance situations, the explanatory variables are in general constant in the course of time (typically they are associated with a characteristic such as gender, professional status, contract type, *etc.*).

This results in the constancy of variables $X_j(t)$. This typical case leads to a semi-parametric model, and the methods described above are slightly modified. Among these models one can mention in particular the model of Lin and Ying [1994], in which the hazard function is supposed of the following form:

$$h(t \mid Z = z) = h_0(t) + \gamma^T z .$$

Lin and Ying [1994] and Klein and Moeschberger [2005] show that starting from the martingale decomposition of the Poisson process, the estimator of the coefficients of the model is:

$$\gamma = A^{-1}B ,$$

where $A = \sum_{i=1}^{D} \sum_{j \in R_i} \left(z_j - \overline{z}_i\right)^T \left(z_j - \overline{z}_i\right)$, $B = \sum_{i=1}^{n} d_i \left(z_i - \overline{z}_i\right)$ and $\overline{z}_i = \frac{1}{R_i} \sum_{j \in R_i} z_j$.

The global significance of the model can be appreciated through the statistics of Wald which follows a Chi-squared distribution with p degrees of freedom (p being the dimension of Z representing the explanatory variables of the model) under the assumption $H_0 : \gamma = 0$, that is to say:

$$\chi_W^2 = \gamma^T V^{-1} \gamma ,$$

where $V = A^{-1}CA^{-1}$ with $C = \sum_{i=1}^{n} d_i \left(z_i - \overline{z}_i\right)^T \left(z_i - \overline{z}_i\right)$. In the case of the significance test of a parameter, one tests the null hypothesis for each parameter γ_j (with $j = 1, \ldots, p$ and $\gamma = \left(\gamma_1, \ldots, \gamma_p\right)$), and one thus considers $H_0 : \gamma_j = 0$, hence $\chi_{W_j}^2 = \gamma_j^2 / V_{jj}$.

5. Sample comparison: non-parametric approach

Imagine the situation in which one wishes to compare the respective lifetimes of two independent samples. More precisely, one disposes of two independent samples, possibly censored, and one wishes to test the null hypothesis of equality of survival functions in the two samples. In the absence of censoring, one can use the traditional rank tests (Wilcoxon test, Savage test), which one will adapt to the presence of censoring.

5.1. Reminder: principle of rank tests[13]

One disposes of two series of observations, E_1 and E_2, of respective sizes n_1 and n_2; one notes $n = n_1 + n_2$; one arranges the sequence of observed values $(x_1, \ldots, x_n)$ in ascending order:

$$x_{(1)} < \ldots < x_{(n)}.$$

The principle of linear rank statistics is to grant a weight (a score) α_i to observation $x_{(i)}$ of rank *i* in the common classification of both samples. Two statistics are then built:

$$R_1 = \sum_{i \in E_1} \alpha_i \text{ and } R_2 = \sum_{i \in E_2} \alpha_i.$$

As $R_1 + R_2 = \sum_{i=1}^{n} \alpha_i$, which is known and deterministic, working on one or the other statistics makes no difference; in practice one retains the one associated with the smallest sample. In choosing $\alpha_i = i$, one obtains the Wilcoxon test; the Savage test being associated with the choice $\alpha_i = 1 - \sum_{j=1}^{i} \frac{1}{n - j + 1}$.

Lastly, the choice of a test rather than another can be guided by the form of the alternative, by retaining the (locally) most powerful test for a given alternative.

5.2. Adaptation of rank tests to the censored case[14]

The adaptation of the preceding tests to the censored case leads to introducing the ordered series of observed moments of deaths (non-censored) into the common sample, which one will note $t_1 < \ldots < t_N$. At each moment t_i, d_{ij} indicates the number of deaths and r_{ij} the population under risk in the group *j*. Population under risk is calculated before exits in t_i, so that the "survivors" after t_i are of headcount $r_{ij} - d_{ij}$. One can summarise this in the following table:

	Death in t_i	Survivors afterwards	Total
Group n°1	d_{i1}	$r_{i1} - d_{i1}$	r_{i1}
Group n°2	d_{i2}	$r_{i2} - d_{i2}$	r_{i2}
All groups	d_i	$r_i - d_i$	r_i

Under the null hypothesis of equality of the survival distributions in the two groups, at every moment one must have equality of the proportions of death in the two groups, which has as a consequence of independence of the lines and the columns in the above table. One is therefore

[13] For developments on the subject, refer to Capéraà and Van Cutsem [1988].

[14] See for example Hill and *al* [1996] for further developments.

in the case of a contingency table with fixed margins, and then the random variable d_{ij} is distributed according to a hypergeometric distribution[15] $H\left(r_i, d_i, \frac{r_{ij}}{r_i}\right)$ (since one counts the number of deaths in the selected group n°*j* among the total deaths d_i, the probability of belonging to group n°*j* being $p = \frac{r_{ij}}{r_i}$ and population size being r_i). One concludes that the expectancy and the variance of d_{ij} are: $E\left(d_{ij}\right) = d_i \frac{r_{ij}}{r_i}$ and $V\left(d_{ij}\right) = d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}$. These observations lead to building statistics based on weighed sums of $d_{ij} - E(d_{ij})$, which are asymptotically Gaussian. By noting (w_i) the selected weights, one finally uses statistics of the form:

$$\phi_j = \frac{\left[\sum_{i=1}^{N} w_i \left(d_{ij} - d_i \frac{r_{ij}}{r_i}\right)\right]^2}{\sum_{i=1}^{N} w_i^2 d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}}$$

which asymptotically follows $\chi^2(1)$. In what follows one will note $\sigma^2 = \sum_{i=1}^{N} w_i^2 d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}$.

5.2.1. The log-rank test

The most simple choice one can think of for the weights is $w_i = 1$, which leads to the test known as the “log-rank test”. In this case the numerator of the statistics of test φ_j is the square of the difference between observed and theoretical counts of deaths, under the null hypothesis:

$$\phi_j = \frac{\left(D_j^{th} - D_j^{obs}\right)^2}{\sigma^2}.$$

This test generalises Savage test to the case with censored data. One can note that under the null hypothesis $D_1^{obs} + D_2^{obs} = D_1^{th} + D_2^{th}$, in other words the value of the test statistics does not depend on the group on which one evaluates it. The statistics form suggests the following approximate formula:

$$\phi = \frac{\left(D_1^{th} - D_1^{obs}\right)^2}{D_1^{th}} + \frac{\left(D_2^{th} - D_2^{obs}\right)^2}{D_2^{th}}$$

[15] It is reminded that the hypergeometric distribution $H(n, k, p)$ is the distribution of *k* successes in *n* draws, without replacement, from a finite population containing successes in proportion *p*.

which one can show that it is lower than that of the log-rank (*cf* Peto and Peto [1972]). Its form evokes that of a usual fitting Chi-squared. The log-rank test is the most frequently used one.

5.2.2. The Gehan test

Gehan (Gehan E.A. [1965]) proposes to retain $w_i = r_i$, which results in weighing more strongly the earliest deaths. This test generalises Wilcoxon test to the case of censored data. This test statistics does not admit a simplified expression like the log-rank. It presents the disadvantage of depending rather strongly on the distribution of censoring.

5.2.3. Example: application to the Freireich dataset

One takes again the two groups of the protocol used by Freireich. Calculations of the test statistics can be carried out based on the following table:

	6-MP		**Placebo**					
Durations	n_{i1}	d_{i1}	n_{i2}	d_{i2}	n_i	d_i	$E(d_{i2})$	$V(d_{i2})$
1	21	0	21	2	42	2	1.00	0.49
2	21	0	19	2	40	2	0.95	0.49
3	21	0	17	1	38	1	0.45	0.25
4	21	0	16	2	37	2	0.86	0.48
5	21	0	14	2	35	2	0.80	0.47
6	21	3	12	0	33	3	1.09	0.65
7	17	1	12	0	29	1	0.41	0.24
8	16	0	12	4	28	4	1.71	0.87
10	15	1	8	0	23	1	0.35	0.23
11	13	0	8	2	21	2	0.76	0.45
12	12	0	6	2	18	2	0.67	0.42
13	12	1	4	0	16	1	0.25	0.19
15	11	0	4	1	15	1	0.27	0.20
16	11	1	3	0	14	1	0.21	0.17
17	10	0	3	1	13	1	0.23	0.18
22	7	1	2	1	9	2	0.44	0.30
23	6	1	1	1	7	2	0.29	0.20

One gets the results summarised as follows:

	log-rank			Gehan		
Durations	Weighting	Coefficient	Variance	Weighting	Coefficient	Variance
1	1.00	1.00	0.49	42	42.00	860.49
2	1.00	1.05	0.49	40	42.00	777.54
3	1.00	0.55	0.25	38	21.00	357.00
4	1.00	1.14	0.48	37	42.00	653.33
5	1.00	1.20	0.47	35	42.00	570.71
6	1.00	-1.09	0.65	33	-36.00	708.75
7	1.00	-0.41	0.24	29	-12.00	204.00
8	1.00	2.29	0.87	28	64.00	682.67
10	1.00	-0.35	0.23	23	-8.00	120.00
11	1.00	1.24	0.45	21	26.00	197.60
12	1.00	1.33	0.42	18	24.00	135.53
13	1.00	-0.25	0.19	16	-4.00	48.00
15	1.00	0.73	0.20	15	11.00	44.00
16	1.00	-0.21	0.17	14	-3.00	33.00
17	1.00	0.77	0.18	13	10.00	30.00
22	1.00	0.56	0.30	9	5.00	24.50
23	1.00	0.71	0.20	7	5.00	10.00
		105.07	6.26		73441.00	5457.11
			$\varphi_2 = 16{,}79$			$\varphi_2 = 13{,}46$

One finds in both cases very weak p-*values*, which confirms the different behaviour of the two groups, which had already been highlighted at the time of the study of the respective cumulated risk functions.

5.3. Approach through point processes

In the same manner that estimators of cumulative hazard or survival function can be obtained in a “natural” way within the framework of point processes, this formalism can be applied to the tests presented above. This method is detailed in Gill [1980].

Imagine the situation where two groups are observed, and one disposes of the two processes of non-censored events $\overline{N}_1^1(t)$ and $\overline{N}_2^1(t)$. The assumption is made that the two processes do not jump at the same time (which translates the orthogonality of martingales M_1 and M_2, $< M_1, M_2 >= 0$,). The idea is, for a predictable and positive process *K*, to consider the following process:

$$\Delta(t) = \int_0^t K(u) \frac{d\overline{N}_1^1(u)}{\overline{R}_1(u)} - \int_0^t K(u) \frac{d\overline{N}_2^1(u)}{\overline{R}_2(u)}$$

The process $M(t) = \int_0^t K(u) \frac{dM_1(u)}{\bar{R}_1(u)} - \int_0^t K(u) \frac{dM_2(u)}{\bar{R}_2(u)}$ is a martingale and also verifies:

$$M(t) = \Delta(t) - \int_0^t K(u)\left(h_1(u) - h_2(u)\right) du .$$

Lastly, under the null hypothesis of identity of the underlying distribution of both populations $M(t) = \Delta(t)$.

The traditional tests are then obtained by specifying the process K. Thus $K(u) = R_1(u)R_2(u)$ leads to the statistics of Wilcoxon-Gehan and $K(u) = \frac{R_1(u)R_2(u)}{R_1(u)+R_2(u)}$ to the statistics of the log-rank. General results on point processes make it possible to obtain the limiting distribution of $\Delta(t)$ under the null hypothesis; more precisely, it is shown that $\Delta(t)$ converges in distribution towards a centered normal distribution of variance $\sigma^2(t)$; a convergent estimator of the variance is given by the quadratic variation of the martingale $\Delta(t)$:

$$< \Delta , \Delta >_t = \int_0^t \left[\frac{K(u)}{R_1(u)} \right]^2 d\bar{N}_1^1(u) + \int_0^t \left[\frac{K(u)}{R_2(u)} \right]^2 d\bar{N}_2^1(u) .$$

6. References

Aalen O. [1978] « Non-parametric inference for a family of counting processes ». *Ann. Stat.* 6, 710-726.

Borgan O. [2014] « Kaplan-Meier Estimator », *Wiley StatsRef: Statistics Reference Online*, doi: 10.1002/9781118445112.stat06033

Capéraà P., Van Cutsem B. [1988] *Méthodes et modèles en statistique non paramétrique*, Presses de l'Université Laval, Paris : Dunod.

Dacunha-Castelle D., Duflo M. [1983] *Probabilités et Statistiques*. Vol. 1 et 2, Paris : Masson.

Droesbeke J.J., Fichet B., Tassi P. [1989] *Analyse statistique des durées de vie* , Paris : Economica.

Gehan E.A. [1965] « A generalized Wilcoxon test for comparing arbitrarily singly-censored samples ». *Biommetrika*, 41, 361-372.

Gill R.D. [1980] « Censoring and stochastic Integrals ». *Mathematical Centre Tracts*, n°124, Amsterdam : Mathematische Centrum.

Fleming T.R., Harrington D.P. [1991] *Counting processes and survival analysis*, Wiley Series in Probability and Mathematical Statistics. New-York : Wiley.

Hill C., Com-Nougué C. [1996] *Analyse statistique des données de survie*, Médecine-Sciences, Paris : Flammarion.

Kaplan E.L., Meier P. [1958] « Non-parametric estimation from incomplete observations ». *Journal of the American Statistical Association*, 53, 457-481.

Klein J. P., Moeschberger M. L. [2005] « Survival Analysis – Techniques for Censored and Truncated Data », *Springer, 2nd edition*.

Lin D. Y., Ying Z. [1994] « Semiparametric analysis of the additive risk model », *Biometrika, n. 81*.

Martinussen T., Scheike T. [2006] *Dynamic regression models for survival data*, New-York: Springer.

NAIR V. N. [1984] « Confidence Bands for Survival Functions with Censored Data: A Comparative Study », Technometrics 14: 945-965.

NELSON W.B. [1972] « Theory and applications of hazard plotting for censored data ». *Technometrics*, 14, 945-965.

PETO R., PETO J. [1972] « Asymptotically efficient rank invariant test procedures (with discussion) ». *J. R. Stat. Soc. A*, 135, 185-207.