Exercises and Lecture Notes, STK 4080, Autumn 2018

Version 0.33, 10-ix-2018

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract

Exercises and Lecture Notes collected here are indeed for the Survival and Event History Analysis course STK 4080 / 9080, autumn semester 2018. The exercises will complement those given in the course book Aalen, Borgan, Gjessing, *Survival and Event History Analysis: A Process Point of View*, Springer, 2008.

1. Ancient Egyptian lifelengths

How long is a life? A unique set of lifelengths in Roman Egypt was collected by W. Spiegelberg in 1901 *(Ägyptische und griechische Eigennamen aus Mumienetiketten der römischen Kaiserzeit)* and analysed by (the very famous) Karl Pearson (1902) in the very first volume of (the very famous) Biometrika. The data set contains the age at death for 141 Egyptian mummies in the Roman period, 82 men and 59 women, dating from the last century b.C. The lifelengths vary from 1 to 96 years, and Pearson argued that these can be considered a random sample from one of the better-living classes in that society, at a time when a fairly stable and civil government was in existence (as we recall, the violent 'tax revolt' with ensuing long-lasting complications took place under Antoninus Pius later, in 139 AD). To access the data, go to `egypt-data` at the course website, reading them into your computer via

```
tt <- scan("egypt-data",skip=5)
```

Pearson did not attempt to fit any parametric models for these data, but discussed differences between the Egyptian age distribution and that of England 2000 years later. The purpose of the present exercise is to analyse aspects of the data by comparing the nonparametric survival curve (here a simplified version of the Kaplan–Meier curves, since there is no censoring; all the old Egyptians are dead) with a couple of parametric curves, in particular the Weibull.

(a) We start with the natural nonparametric estimate of the survival curve $S(t) = \Pr\{T \geq t\}$. Let the data be $t_1, \ldots, t_n$ (either the full set, or the subset for men, or that of the women). Since this is just a binomial probability, for each fixed t, we may put up the empirical survival function

$$S_{\rm emp}(t) = (1/n) \sum_{i=1}^{n} I\{t_i \geq t\} \quad \text{for } t > 0.$$

Show that $\mathrm{E}\, S_{\rm emp}(t) = S(t)$ and that $\mathrm{Var}\, S_{\rm emp}(t) = (1/n)S(t)\{1 - S(t)\}$.

(b) Compute the empirical survival curves, for men and for women, say $S_{m,\mathrm{emp}}(t)$ and $S_{w,\mathrm{emp}}(t)$, and display them in the same diagram, cf. Figure 0.1 below.

(c) Then consider the two-parameter Weibull model [note the Swedish pronunciation], which has a cumulative distribution of the form

$$F(t, a, b) = 1 - \exp\{-(at)^b\} \quad \text{for } t > 0,$$

with a and b positive parameters (typically unknown). (i) Find a formula for the median of the distribution. (ii) Show that the probability of surviving age t, given that one has survived up to t_0, is $\exp[-\{(at)^b - (at_0)^b\}]$, for $t > t_0$. (iii) Show that the density can be expressed as

$$f(t, a, b) = \exp\{-(at)^b\}a^b bt^{b-1} \quad \text{for } t > 0.$$

(d) Find formulae for the 0.20- and 0.80-quantiles, and set these equal to the observed 0.20- and 0.80-quantiles for the data. This yields two equations with two unknowns, which you can solve. In this fashion, find estimates $(\widetilde{a}, \widetilde{b})$ for the men and for the women.

(e) While quantile fitting is a perfectly sensible estimation method, a more generally versatile method is that of maximum likelihood (ML), which will also be used later on in the course. By definition, the ML estimates $(\widehat{a}, \widehat{b})$ are the parameter values maximising the log-likelihood function

$$\ell_n(a, b) = \sum_{i=1}^n \log f(t_i, a, b) = \sum_{i=1}^n \{-(at_i)^b + b \log a + \log b + (b - 1) \log t_i\}.$$

This can be maximised numerically, as soon as you can programme the log-likelihood function. With data stored in your computer, called `tt`, try this, using R's powerful non-linear minimiser `nlm`:

```
logL <- function(para)
{
a <- para[1]
b <- para[2]
hei <- -(a*tt)^b + b*log(a) + log(b) + (b-1)*log(tt)
sum(hei)
}
# then:
minuslogL <- function(para)
{-logL(para)}
# then:
nils <- nlm(minuslogL2,c(0.20,1.00),hessian=T)
ML <- nils$estimate
```

It gives you the required ML estimates $(\widehat{a}, \widehat{b})$. Carry out this estimation scheme, for the men and the women separately.

(f) I find $(0.0270, 1.3617)$ for the men and $(0.0347, 1.5457)$ for the women. Display the two estimated Weibull survival curves, perhaps along with the two nonparametric ones, as in my Figure 0.1 here. Compute the estimatead median lifelengths, for men and for women, and comment.

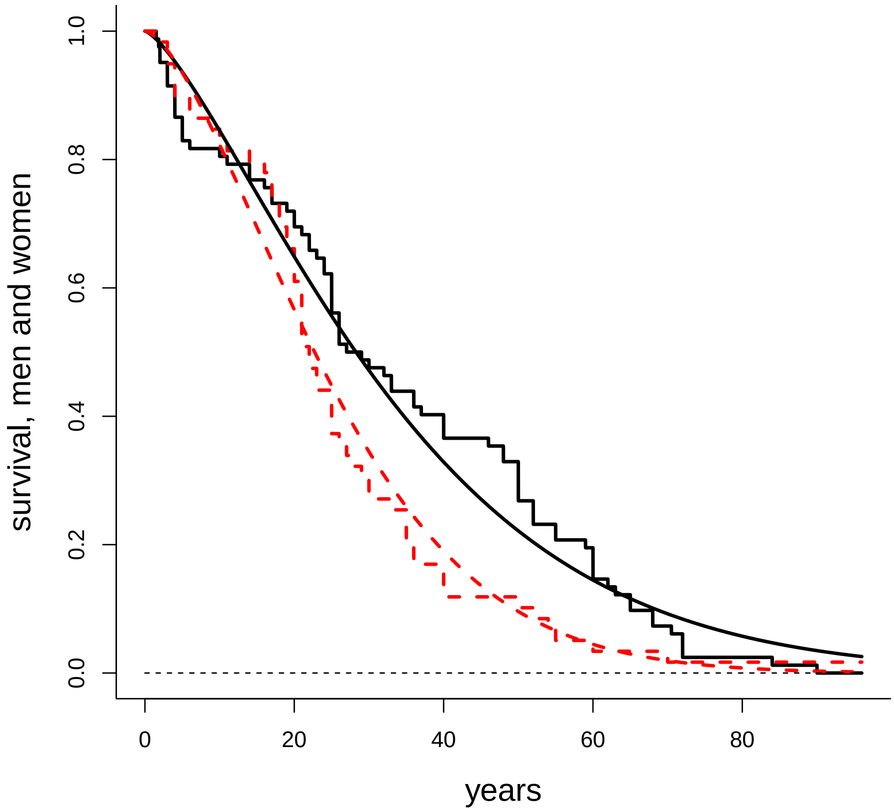

Figure 0.1: Survival curves from Roman era Egypt, two for men (black) and two for women (red). The step-functions are the empirical survival curves, type Kaplan–Meier; the continuous curves are the fitted Weibull curves.

(g) Compute and display also the estimated Weibull hazard rates, for men and for women. Comment on what you find.

(h) Considerations above invite statistical testing of the hypothesis H_0 that men and women of Roman era Egypt had the same lifelength distributions. Compute and display the 90% confidence bands

$$S_{m,\mathrm{emp}}(t) \pm 1.645\,\widehat{\tau}_m(t), \quad S_{w,\mathrm{emp}}(t) \pm 1.645\,\widehat{\tau}_w(t),$$

where

$$\begin{aligned}\widehat{\tau}_m(t)^2 &= (1/n_m)S_{m,\mathrm{emp}}(t)\{1 - S_{m,\mathrm{emp}}(t)\},\\ \widehat{\tau}_w(t)^2 &= (1/n_w)S_{w,\mathrm{emp}}(t)\{1 - S_{w,\mathrm{emp}}(t)\},\end{aligned}$$

the estimated variances. (We shall learn formal tests along such lines in the course.)

(i) Above I've forced you through the loops of things for one particular parametric model, namely the Weibull. Now do all these things for the Gamma(a, b) model too, with density $\{b^a/\Gamma(a)\}t^{a-1}\exp(-bt)$. Part of the point here is that this does not imply a doubling of your work efforts; you may edit your computer programmes, at low work cost, to accommodate other parametric models, once you've been through one of them. The Weibull does a slightly better job than the Gamma, it turns out.

"Either man is constitutionally fitter to survive to-day [than two thousand years ago], or he is mentally fitter, i.e. better able to organise his civic surroundings. Both conclusions point perfectly definitely to an evolutionary progress." – *Karl Pearson*, 1902.

2. Did men live longer than women in Ancient Egypt?

As a follow-up to the Ancient Egypt analysis of Exercise 1, consider the following attempt to quantify more accurately the extent to which men and women had different lifelengths then.

(a) Plot the difference in survival function $D(t) = S_{m,\mathrm{emp}}(t) - S_{w,\mathrm{emp}}(t)$, and also the ratio funcion $S_{m,\mathrm{emp}}(t)/S_{w,\mathrm{emp}}(t)$. Comment on what these plots indicate.

(b) Find an expression for the variance $\kappa(t)^2$ of $D(t)$. Then construct and compute an empirical estimate, say $\widehat{\kappa}(t)$.

(c) Plot both $D(t)$ and the band $D(t)\pm 1.645\,\widehat{\kappa}(t)$. What is the interpretation of this band? What are your conclusions, regarding lifelengths in ancient Egypt? What are the likely reasons for differences you spot?

3. Survival functions and hazard rates

Consider a lifetime variable T with density f and cumulative distribution function F on the halfline (so, in particular, the distribution is continuous). Define the hazard rate function α as

$$\alpha(t)\,\mathrm{d}t = \Pr\{T \in [t, t + \mathrm{d}t] \,|\, T \geq t\},$$

for a small time window $[t, t + \mathrm{d}t]$; more formally,

$$\alpha(t) = \lim_{\varepsilon \to 0} (1/\varepsilon) \Pr\{T \in [t, t + \varepsilon] \,|\, T \geq t\}.$$

(a) First define the survival function as

$$S(t) = \Pr\{T \geq t\} = 1 - F(t).$$

What are its basic properties?

(b) Show that in fact $\alpha(t) = f(t)/S(t)$. So from knowledge of f we can always find the hazard rate from $\alpha = f/(1 - F)$.

(c) Define also the cumulative hazard rate function as $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. Show that

$$F(t) = 1 - \exp\{-A(t)\} \quad \text{and} \quad f(t) = \alpha(t)\exp\{-A(t)\}.$$

(d) Let T have the exponential distribution with density $f(t, \theta) = \theta\, \exp(-\theta t)$. Find its survival function and hazard rate.

(e) For the Weibull distribution, with $F(t) = 1 - \exp\{-(at)^b\}$, with the hazard rate function, and display it in a plot, for $a = 3.33$ and b equal to 0.9, 1.0, 1.1.

(f) Consider the Gamma distribution with parameters (a, b), which has the density

$$f(t, a, b) = \frac{b^a}{\Gamma(a)} t^{a-1} \exp(-bt) \quad \text{for } t > 0.$$

Show that the mean and variance are a/b and a/b^2. Take $b = 2.22$, compute the hazard rates for a equal to 0.8, 1.0, 1.2, and display these in a diagram. Give explicit formulae for the survival function and hazard rate for the case of $a = 2$.

(g) Consider a lifetime distribution with hazard rate $\alpha(t) = 1/(1+t)$. Find its survival function and density.

4. Maximum likelihood estimation with censored data

If we have observed independent lifetime data $t_1, \ldots, t_n$, from a suitable parametric density $f(t, \theta)$, the ML estimator is found by maximising the log-likelihood function $\sum_{i=1}^n \log f(t_i, \theta)$. This exercise looks into the required amendments in the case of censored data, say $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, with

$$\delta_i = I\{t_i \text{ is the observed lifetime}\} = \begin{cases} 1 & \text{if } t_i \text{ is the observed lifelength,} \\ 0 & \text{if } t_i \text{ is the censored value.} \end{cases}$$

So, in the case of $\delta_i = 0$, this means that the real lifetime, say t_i^0, is at least as large as t_i, but we do not know more than that.

(a) Assume that the parametric model is given and perhaps primarily thought about via its hazard rate function, say $\alpha(t, \theta)$. Assume first that all t_i correspond to genuinely observed lifetimes, i.e. that there is no censoring. Show that the log-likelihood function above can be expressed as

$$\ell_n(\theta) = \sum_{i=1}^n \{\log \alpha(t_i, \theta) - A(t_i, \theta)\},$$

with $A(t, \theta)$ the cumulative hazard function for the model.

(b) For the very simple exponential model, with $\alpha(t, \theta) = \theta$, write up the log-likelihood function from the exression under (a), and show that the ML estimator is $\widehat{\theta} = n / \sum_{i=1}^n t_i = 1/\bar{t}$.

(c) Then consider the general case with censoring, i.e. some of the δ_i are equal to zero. Show that the log-likelihood function can be written

$$\ell_n(\theta) = \sum_{\delta_i = 1} \log f(t_i, \theta) + \sum_{\delta_i = 0} \log S(t_i, \theta) = \sum_{i=1}^n \{\delta_i \log \alpha(t_i, \theta) - A(t_i, \theta)\}.$$

Sometimes the first expression is more practical to work with, sometimes the second; also, as will be seen later, the second expression lends itself more easily to general counting process models.

(d) For the simple exponential model, again, but now with censoring, show that the ML estimator is $\widehat{\theta} = \sum_{i=1}^n \delta_i / \sum_{i=1}^n t_i$, generalising the non-censored case above.

(e) Generalise the previous situation to the model where the hazard rate is $\alpha(t) = \alpha_0(t)\theta$, with a known basis hazard function α_0 but with unknown multiplicative parameter θ.

5. Counting process, at-risk process, intensity process, martingale

The eternal golden braid, for modelling and analysing survival and event history data, is the quadruple (N, Y, λ, M)! The ingredients are the counting process N, the at-risk process Y, the intensity process λ, and the martingale M. These matters are of course tended to in the ABG book, in several chapters, with various setups, specialisations, and generalisations. This particular exercise gives a separate, brief, and partial introduction to these four items, in the context of survival data. These are of the form $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, as for Exercise 4. The distribution is continuous, so there are no ties among the t_i.

(a) The counting process N counts the number of observed events, over time:

$$N(t) = \#\{\text{observed events over } [0,t]\} = \sum_{i=1}^n I\{t_i \le t, \delta_i = 1\}.$$

It starts out at zero, at time zero, and then increases with jump size 1 each time a new observation is recorded.

(b) The at-risk process counts those individuals who are still at tisk, for each given time point:

$$Y(t) = \#\{\text{individuals at risk just before time } t\} = \sum_{i=1}^n I\{t_i \ge t\}.$$

The 'just before' thing can be formalised, e.g. via left continuity. The point is that an individual belongint to the risk set at time t, with this definition, can have his or her event in the time window $[t, t+\varepsilon]$. Note that $Y(t)$ counts both those t_i with $\delta_i = 1$ and those with $\delta_i = 0$ (since we do not know yet when events occur, or when censoring might occur).

(c) The intensity process $\lambda(s)$ can be defined in severeal ways, cf. the ABG book, and with somewhat different, but related, motivations and interpretations. The simplest way, in this framework, might be

$$\lambda(s)\,\mathrm{d}s = \Pr\{N[s, s+\mathrm{d}s] = 1 \,|\, \mathcal{F}_{s-}\}.$$

First,

$$\mathrm{d}N(s) = N[s, s+\mathrm{d}s] = N(s+\mathrm{d}s) - N(s-)$$

is the numer of observed events inside the small time window $[s, s+\mathrm{d}s]$. Second, $\mathcal{F}_{s-}$ is the full history of everything that has been observed up to just before time s, i.e. over $[0, s)$. In the present setup of survival data (i.e. without complications of more complex event history constructions), the relevant information in all of $\mathcal{F}_{s-}$ is simply 'how many are still at risk', i.e. $Y(s)$.

(d) In this setup, show that

$$\mathrm{d}N(s) \,|\, \mathcal{F}_{s-} \sim \mathrm{Bin}(Y(s), \alpha(s)\,\mathrm{d}s),$$

a simple binomial situation with $Y(s)$ at risk and with a small probability $\alpha(s)\,\mathrm{d}s$. Show that

$$\begin{aligned}
\Pr\{\mathrm{d}N(s) = 0 \,|\, \mathcal{F}_{s-}\} &= 1 - Y(s)\alpha(s)\,\mathrm{d}s + O((\mathrm{d}s)^2), \\
\Pr\{\mathrm{d}N(s) = 1 \,|\, \mathcal{F}_{s-}\} &= Y(s)\alpha(s)\,\mathrm{d}s + O((\mathrm{d}s)^2), \\
\Pr\{\mathrm{d}N(s) \ge 2 \,|\, \mathcal{F}_{s-}\} &= O((\mathrm{d}s)^2),
\end{aligned}$$

with order notation $g(\varepsilon) = O(\varepsilon^2)$ meaning that $g(\varepsilon)$ is of order ε^2 (more precisely, is not of a bigger order), defined as $g(\varepsilon)/\varepsilon^2$ remaining bounded as $\varepsilon \to 0$. The above means that all the action here is in 0 (high chance) and 1 (slim chance, but important, and sooner or later it will kick in). In particular, show from (c) that

$$\lambda(s) = Y(s)\alpha(s).$$

This is a special case of Aalen's multiplicative intensity model (stemming from his Berkeley PhD thesis 1975, then from his Annals of Statistics paper 1978, and further discussed and used and generalised in dozens of books and a few hundreds of journal articles, etc.).

(e) Then consider the random process

$$M(t) = N(t) - \int_0^t \lambda(s)\,\mathrm{d}s = N(t) - \int_0^t Y(s)\alpha(s)\,\mathrm{d}s.$$

Demonstrate that it has the magical martingale property,

$$\mathrm{E}\,\{\mathrm{d}M(s)\,|\,\mathcal{F}_{s-}\} = 0,$$

with $\mathrm{d}M(s) = M(s + \mathrm{d}s) - M(s)$ the martingale increment.

(f) Show that the process

$$K(t) = M(t)^2 - \int_0^t Y(s)\alpha(s)\,\mathrm{d}s$$

also is a martingale. We shall see later, in the course and in exercises, that various central properties flow from these martingales, including results on limiting normality for classes of estimators.

(g) Consider again the situation of Exercise 4, with log-likelihood functions for censored data. With the golden quadruple on board, show that the log-likelihood function also can be expressed as

$$\ell_n(\theta) = \sum_{i=1}^n \{\delta_i \log \alpha(t_i, \theta) - A(t_i, \theta)\} = \int_0^\tau \{\log \alpha(s, \theta)\,\mathrm{d}N(s) - Y(s)\alpha(s, \theta)\,\mathrm{d}s\}.$$

Here the integral of a function with respect to a counting process is defined, simply, as

$$\int_0^\tau g(s)\,\mathrm{d}N(s) = \sum_{i=1}^n g(t_i)\delta_i,$$

a sum of the function evaluated precisely at the observed timelengths.

6. A parametric step-function for the hazard rate

Consider independent lifetime data of the form $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, as met with in Exercises 4 and 5, and assume they stem from a common distribution with hazard rate $\alpha(s)$. We wish to estimate the cumulative hazard rate $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. This can famously be done using the Nelson–Aalen estimator, see the following exercise, but I start out working through a parametric version, via a step-function. When the windows become small, this parametric estimator will actually converge to the Nelson–Aalen.

(a) Consider the parametric model where the hazard rate is a step-function, i.e. constant over windows. Suppose $[0,\tau]$ is the full time window of relevance (i.e. with a sufficiently big end-point τ), with windows $W_j = [s_{j-1}, s_j)$ for $j = 1,\ldots,k$, and $0 = s_0 < s_1 < \cdots < s_k = \tau$. The model is then of the form

$$\alpha(s) = \alpha_j \text{ on time window } W_j \quad \text{for } j = 1,\ldots,k.$$

Using the log-likelihood expression of Exercise 5, show that the log-likelihood function can be written as

$$\ell_n(\alpha_1,\ldots,\alpha_k) = \sum_{j=1}^k \int_{W_j} \{\log\alpha(s)\,\mathrm{d}N(s) - Y(s)\alpha(s)\,\mathrm{d}s\} = \sum_{j=1}^k (\Delta N_j \log\alpha_j - R_j\alpha_j),$$

in which

$$\Delta N_j = N(W_j) = N(s_j) - N(s_{j-1}) \quad \text{and} \quad R_j = \int_{W_j} Y(s)\,\mathrm{d}s.$$

So ΔN_j is the number of observed events, and R_j is the 'total time at risk', over window W_j.

(b) Show the ML estimators for the local constants become

$$\widehat{\alpha}_j = \frac{\Delta N_j}{R_j} = \frac{\Delta N_j}{\int_{W_j} Y(s)\,\mathrm{d}s}$$

for $j = 1,\ldots,k$. The ML estimator of the full cumulative hazard rate is hence the integral of the estimated step-function, which becomes the piecewise linear

$$\widehat{A}(t) = \begin{cases} \widehat{\alpha}_1 t & \text{for } t \in W_1, \\ \widehat{\alpha}_1 s_1 + \widehat{\alpha}_2(t - s_1) & \text{for } t \in W_2, \\ \widehat{\alpha}_1 s_1 + \widehat{\alpha}_2(s_2 - s_1) + \widehat{\alpha}_3(t - s_2) & \text{for } t \in W_3, \\ \text{etc.} \end{cases}$$

(c) Find the Hessian matrix

$$J(\alpha) = -\partial^2\ell_n(\alpha_1,\ldots,\alpha_k)/\partial\alpha\partial\alpha^{\mathrm{t}},$$

and show that it is diagonal. I write α for the full parameter vector $(\alpha_1,\ldots,\alpha_k)$ where convenient. Find also 'the observed information', which is this minus the second order derivative matrix computed at the ML. Show indeed that

$$\widehat{J} = J(\widehat{\alpha}) = \mathrm{diag}(R_1^2/\Delta N_1,\ldots,R_k^2/\Delta N_k).$$

Large-sample theory, dealt with later in the course and in ABG Ch. 5, says that

$$\widehat{\alpha} \approx_d \mathrm{N}_k(\alpha, \widehat{J}^{-1}).$$

Show that this translates to the $\widehat{\alpha}_j$ being approximately unbiased, normal, and independent, with

$$\mathrm{Var}\,\widehat{\alpha}_j \doteq \Delta N_j/R_j^2.$$

(d) Argue, perhaps heuristically, that when the time windows become carefully small, then the above $\widehat{A}(t)$ is in effect a nonparametric estimator of the cumulative hazard function, approximately unbiased and normal, and with variance estimable by

$$\widehat{\kappa}(t)^2 = \sum_{\text{windows left of } t} \frac{\Delta N_j}{(R_j/d_j)^2},$$

with $d_j = s_j - s_{j-1}$ the width of window W_j. Since $R_j/d_j = (1/d_j)\int_{W_j} Y(s)\,\mathrm{d}s$, this is close to $Y(s_j^*)$, with s_j^* the mid-point of W_j. A further approximation, which becomes correct in a fine-tuned large-sample setup with cells becoming small at the right rate, is then

$$\widehat{\kappa}(t)^2 = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)^2}.$$

Via the step-function model, and some extra analysis, we have essentially reinvented the Nelson–Aalen estimator, along with its properties; see Hermansen and Hjort (2015).

7. The Nelson–Aalen estimator

Consider again independent lifetime data of the form $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, as met with in Exercises 4, 5, 6, and assume they stem from a common distribution with hazard rate $\alpha(s)$. In the previous exercise I set up a step-function model for $\alpha(s)$, which led to an almost nonparametric estimator for the cumulative $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. The canonical nonparametric estimator is indeed this fine-tuned limit, namely the Nelson–Aalen estimator.

(a) We start with the definition, using the counting process and at-risk process notation of Exercise 5 (and used in the book). The Nelson–Aalen estimator for $A(t)$ is

$$\widehat{A}(t) = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)} = \sum_{t_i \le t} \frac{\delta_i}{Y(t_i)}.$$

Again, the integral with respect to a counting process is simply the finite sum over the appropriate integrand, over the observed event times. It is easy to make a programme computing $\widehat{A}(t)$. Do this, for a dataset of your choice, perhaps simulated. Often one is content to compute and plot $\widehat{A}(t)$ just at the observed values t_i, in which case a simpler programme than the one below can be put up, but in various contexts it is useful to compute, plot, compare for a full fine grid of values, say, as here. This little programme requires that the (t_i, δ_i) are predefined as `tt` and `delta`.

```
eps <- 0.001
tval <- seq(0,20,by=eps)
Yval <- 0*tval
DeltaNval <- 0*tval
# then:
for (j in 1:length(tval))
{
tnow <- tval[j]
Yval[j] <- sum(1*(tt >= tnow))
ok <- 1*(tt >= tnow)*(tt < tnow+eps)*delta
```

```
DeltaNval[j] <- sum(ok)
}
# then:
jumps <- DeltaNval/Yval
Ahat <- cumsum(jumps)
matplot(tval,Ahat,type="l",xlab="time",ylab="look at Nelson-Aalen")
```

(b) Then we ought to spend a few minutes thinking about why the Nelson–Aalen $\widehat{A}(t)$ is a natural estimator of $A(t)$. Using the martingale M of Exercise 5, we may write

$$\mathrm{d}N(s) = Y(s)\alpha(s)\,\mathrm{d}s + \mathrm{d}M(s) = \text{structure} + \text{random fluctuations},$$

which implies

$$\mathrm{d}N(s)/Y(s) = \alpha(s)\,\mathrm{d}s + \text{noise}.$$

Argue that this points to the Nelson–Aalen.

(c) With a bit of heuristics, we have

$$\widehat{A}(t) - A(t) = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)} - A(t) \doteq \int_0^t \frac{\mathrm{d}M(s)}{Y(s)} = \int_0^t \frac{1}{\widehat{y}(s)} \frac{\mathrm{d}M(s)}{n},$$

where $\widehat{y}(s) = Y(s)/n$ is a steadily more precise estimate of its limit in probability, say $y(s) = \Pr\{T \ge s, C \ge s\}$, with C the censoring mechanism. It follows that

$$\sqrt{n}\{\widehat{A}(t) - A(t)\} \doteq \int_0^t \frac{1}{\widehat{y}(s)} \frac{\mathrm{d}M(s)}{\sqrt{n}}.$$

But $M(t)/\sqrt{n} \to_d V(t)$, say, a zero-mean Gaussian martingale with incremental variances $\mathrm{Var}\,\mathrm{d}V(s) = y(s)\alpha(s)\,\mathrm{d}s$, by results in ABK (Chs. 4, 5). This, at least heuristically, is seen to imply

$$Z_n(t) \to_d Z(t) = \int_0^t \frac{1}{y(s)} \mathrm{d}V(s),$$

which is another zero-mean Gaussian martingale with incremental variances

$$\mathrm{Var}\,\mathrm{d}Z(s) = \mathrm{Var}\,\frac{\mathrm{d}V(s)}{y(s)} = \frac{\alpha(s)\,\mathrm{d}s}{y(s)}.$$

(d) So the Nelson–Aalen is for large samples approximately unbiased, approximately normal, and with variance

$$\sigma(t)^2 = \mathrm{Var}\,\widehat{A}(t) \doteq \frac{1}{n} \int_0^t \frac{\alpha(s)\,\mathrm{d}s}{y(s)}.$$

Give arguments supporting the estimator

$$\widehat{\sigma}(t)^2 = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)^2}.$$

So a programme for the Nelson–Aalen just needs a few lines more to produce also $\widehat{\sigma}(t)$. In particular, confidence bands are now easy to construct, say

$$\widehat{A}(t) \pm 1.645\,\widehat{\sigma}(t) \quad \text{for } t \in [0, \tau],$$

where $[0, \tau]$ is a relevant time window for the data. Try to show that this band contains the true $A(t)$ with probability converging to 0.90, for each fixed t.

8. IUD expulsion

Data have been collected for IUD use for $n = 100$ women (I believe they stem from a Stanford PhD 1975, with data later on forwarded to and worked with by Aalen, then to Borgan and myself). The `iud-data` file has three columns: the index $i = 1, \ldots, n$; the time t_i to 'event', measured in days, from the first day of use; and an index for 'event', from 1 (she's pregnant!, which however does not happen here), to 2 (expulsion), to 3 and 4 (removal for pains, or bleeding, or other medical reasons), to yet other categories 5, 6, 7, 8, 9 of less interest here.

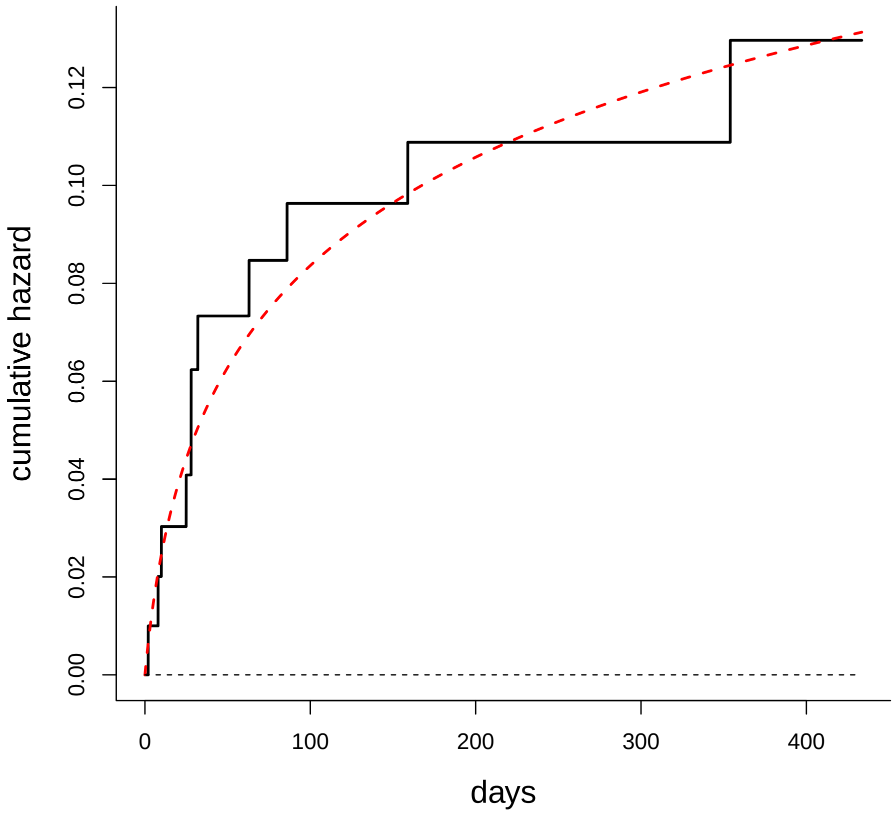

Figure 0.2: Estimated cumulative hazard rate for time to expulsion of IUD, via the nonparametric Nelson–Aalen estimator and the parametric frailty model.

(a) Fit first the simple model that takes the hazard rate to be a constant θ. Under this model, what is the estimated median time to expulsion, for women using IUD (supposing they do not quit on their own)? (I have no idea whether these 1975 IUD data would look very differently now.) Compute also $\ell_{n,0,\max} = \ell_{n,0}(\widehat{\theta})$, the attained log-likelihood maximum for that model.

(b) Then assume that each woman has an exponential IUD expulsion time, say θ, but that this parameter varies from woman to woman, according to a Gamma distribution (a, b). Show that the survival function in the population then becomes

$$S(t) = \Pr\{\text{IUD still in place at time } t\} = \frac{1}{(1 + t/b)^a} = \exp\{-a \log(1 + t/b)\}.$$

(c) Show that the ensuing hazard rate function becomes

$$\alpha(t) = \frac{a/b}{1 + t/b} = \frac{\theta_0}{1 + t/b},$$

writing for emphasis $\theta_0 = a/b$ for the mean value of the Gamma distribution of the women's random intensities. If b is large, the variance of the random θ is small, and we're back to the simpler model with a common θ_0 for all IUD users.

(d) Fit the expulsion data to this two-parameter model. Produce a version of Figure 0.2, with both the parametric and nonparametric Nelson–Aalen estimates. Does the model appear to fit? Under this two-parameter model, what is the estimated median time until expulsion (again, assuming the woman does not quit on her own)? Compute also $\ell_{n,\max} = \ell_n(\widehat{a}, \widehat{b})$, the attained log-likelihood maximum for this model, and compare to the corresponding number for the simpler model.

(e) In addition to producing a version of Figure 0.2, pertaining to cumulative hazard, make a simular figure for the estimated survival functions (parametric and nonparametric), i.e. the probability that the IUD is not yet expulsed.

9. More on martingales

We've met the full eternal golden quadruple (N, Y, λ, M) in Exercise xx. Here I go through more details regarding the martinale machinery and its basic properties. The standard setup so far is for the survival data framework of data $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, as with Exercises xx xx, for which $M(t) = N(t) - \int_0^t Y(s)\alpha(s)\,\mathrm{d}s$, but more general versions will be met later on in the course of the course.

(a) We start out with the notion of 'a growing history of information', formalised as $\mathcal{F}_t$ being the sigma-algebra of all available information for the time window $[0, t]$. Formally, a sigma-algebra is a set of sets, (i) containing the empty-set; (ii) containing all complements (so if B is in, then so is B^c); (iii) containing all countable unions (so if $B_1, B_2, \ldots$ are in, then $\cup_{j=1}^{\infty} B_j$ is in). Similarly, $\mathcal{F}_{t-}$ is all information available 'a milli-second before t', formally the limit of $\mathcal{F}_{t-\varepsilon}$ as $\varepsilon \to 0$. Thus $\mathcal{F}_{s-}$ contains the value of $Y(s - 0.33)$ and $N(s - 0.11)$, and even $Y(s)$, but not $N(t + 0.07$, and not $\mathrm{d}N(s) = N(s + \mathrm{d}s) - N(s)$.

(b) A process $M = \{M(t)\colon t \geq 0\}$ is a martingale, with respect to the filtration $\{\mathcal{F}_t\colon t \geq 0\}$, provided $M(0) = 0$ and

$$\mathrm{E}\,\{\mathrm{d}M(s) \,|\, \mathcal{F}_{s-}\} = 0 \quad \text{for each } s,$$

where $\mathrm{d}M(s) = M(s + \mathrm{d}s) - M(s)$ is a small increment for M. Show that $\mathrm{E}\,\mathrm{d}M(s) = 0$ and that $M(t) = \int_0^t \mathrm{d}M(s)$ also has mean zero.

(c) In the survival setup of Exercise xx, where $\mathrm{d}M(s) = \mathrm{d}N(s) - Y(s)\alpha(s)\,\mathrm{d}s$, show again that M is a martingale, and find $\mathrm{Var}\,\{\mathrm{d}M(s) \,|\, \mathcal{F}_{s-}\}$.

(d) Next we need the *(predictable) variance process*, say $\langle M, M\rangle(t)$, defined via

$$\mathrm{d}\langle M, M\rangle(s) = \mathrm{Var}\,\{\mathrm{d}M(s) \,|\, \mathcal{F}_{s-}\}.$$

Integrating up, or summing over a million small cells, gives

$$\langle M, M\rangle(t) = \int_0^t \mathrm{Var}\,\{\mathrm{d}M(s)\,|\,\mathcal{F}_{s-}\}.$$

Note that this is a *random process*, summing up a host of small conditional variances. For the survival analysis setup, show that

$$\langle M, M\rangle(t) = \int_0^t Y(s)\alpha(s)\,\mathrm{d}s.$$

For this setup, the variance process is hence identical to the so-called compensator $\int_0^t Y\alpha\,\mathrm{d}s$ of the counting process N.

(e) Now consider any function $H = \{H(s)\colon s \geq 0\}$, and form its integral with respect to the martingal M:

$$K(t) = \int_0^t H(s)\,\mathrm{d}M(s) \quad \text{for } t \geq 0.$$

It may be defined generally as the fine limit of Riemann type sums $\sum_j H(s_j)\{M(s_{j+1}) - M(s_j)\}$, when the cells $[s_j, s_{j+1})$ become smaller, but for the present purposes of the survival setup it is sufficient to agree on

$$\int_0^t H\,\mathrm{d}M = \int_0^t H(s)\{\mathrm{d}N(s) - Y(s)\alpha(s)\,\mathrm{d}s\} = \sum_{t_i \leq t} H(t_i)\delta_i - \int_0^t H(s)Y(s)\alpha(s)\,\mathrm{d}s.$$

Now show that $K = \int H\,\mathrm{d}M$ is also a martingale, provided H is *previsible*, in the sense that the value of $H(s)$ is known when $\mathcal{F}_{s-}$ is known. Examples would be $Y(s-0.14)^{1/3}$ and $Y(s)^{1/2}$, but not, for example, $Y(s+0.03)$.

(f) When H is previsible (with respect to the same filtration of growing history), such that $K = \int_0^\cdot H\,\mathrm{d}M$ is another martingale, show that

$$\langle \int_0^\cdot H\,\mathrm{d}M, \int_0^\cdot H\,\mathrm{d}M\rangle(t) = \int_0^t H(s)^2\,\mathrm{d}\langle M, M\rangle(s).$$

(g) For the survival data setup, consider the random function

$$K(t) = \int_0^t \frac{J(s)}{Y(s)^{1/2}}\,\mathrm{d}M(s),$$

where $J(s) = I\{Y(s) \geq 1\}$ is equal to 1 with very high probability (unless s becomes large). The point is that $1/Y(s)$ isn't defined when $Y(s) = 0$, and we take $J(s)/Y(s)$ to be zero in case of $Y(s) = 0$, which also means $J(s) = 0$. Show that K is a martingale, with variance process

$$\langle K, K\rangle(t) = A^*(t) = \int_0^t J(s)\alpha(s)\,\mathrm{d}s,$$

which with very high probability is equal to $A(t)$ itself.

(h) Consider now two martingales, say M_1 and M_2, with respect to the same filtration. We define their *(predictable) covariance process* $\langle M_1, M_2\rangle$ via

$$\mathrm{d}\langle M_1, M_2\rangle(s) = \mathrm{cov}\{\mathrm{d}M_1(s), \mathrm{d}M_2(s)\,|\,\mathcal{F}_{s-}\},$$

again with notation $\mathrm{d}M_j(s) = M_j(s + \mathrm{d}s) - M_j(s)$. If

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha(s)\,\mathrm{d}ss$$

is the little martingale associated with individual i only, so that $N_i(t) = I\{t_i \le t, \delta_i = 1\}$ and $Y_i(t) = I\{t_i\ get\}$, taking on values 0 and 1 only, show that

$$\langle M_i, M_j\rangle(t) = 0 \quad \text{for } i \ne j,$$

whereas $\langle M_i, M_i\rangle(t) = Y_i(t)\alpha(s)\,\mathrm{d}s$.

(i) Finally, at this stage, consider two martingales M_1 and M_2, along with two previsible processes H_1 and H_2, such that $K_1 = \int_0^\cdot H_1\,\mathrm{d}M_1$ and $K_2 = \int_0^\cdot H_2\,\mathrm{d}M_2$ become martingales. Show that

$$\langle K_1, K_2\rangle(t) = \langle \int_0^\cdot H_1\,\mathrm{d}M_1, \int_0^\cdot H_2\,\mathrm{d}M_2\rangle(t) = \int_0^t H_1 H_2\,\mathrm{d}\langle M_1, M_2\rangle(s).$$

In particular, if M_1 and M_2 are orthogonal, meaning that their covariance process is zero, then also $\int_0^\cdot H_1\,\mathrm{d}M_1$ and $\int_0^\cdot H_2\,\mathrm{d}M_2$ are orthogonal, even if H_1 and H_2 might be dependent in complicated ways. It suffices that M_1 and M_2 have uncorrelated increments.

10. Central limit theorems and their partial sums processes

Ceci n'est pas une pipe and this is not a course on advanced limit theorems from probability theory. We nevertheless need results on approximate normality of various important estimators and test statistics, and insights into why and how such approximations hold help us also in constructing yet new estimators and tests. I therefore include two exercises on limit theorems for 'sums of small variables', with the following exercise, pertaining to martingales and weighted martingales, of particular relevance for the course and its curriculum. The present exercise relates to the simpler universe of independent summands, where we're in the realm of classical Central Limit Theorems (from de Moivre and Laplace, around 1740, to Lindeberg 1922, Alan Turing 1925, Donsker 1950, and onwards, with several hundreds of books and several thousands of ournal articles). A point conveyed in this exercise, and more fully needed and relied upon in the following exercise, is that we care not only about a sum $\sum_{i=1}^n X_i$ being approximately normal, but about the full partial-sum process $\sum_{i\le[nt]} X_i$ being close to a Gaußian martingale.

The *statistical practical use* of these mathematical and probabilistical theorems of proofs consists in translating 'the variable or process M_n tends to a Gaußian variable or process V when n travels all the way to infinity' to 'the distribution of M_n is approximately that of a normal or of a full Gaußian process', and then to translate this further to practical confidence intervals, confidence bands, tests with a given significance level like 0.05, etc.

(a) Let me begin with a simple setup, involving a sequence $X_1, X_2, \ldots$ of i.i.d. random variables, with mean zero and finite standard deviation σ. Consider the cumulative sum process

$$M_n(t) = \sum_{i\le[nt]} X_i \quad \text{for } t \ge 0.$$

Here $[nt]$ is the largest integer smaller than or equal to nt. The M_n process is 0 on $[0, 1/n)$, is X_1 on $[1/n, 2/n)$, is $X_1 + X_2$ on $[2/n, 3/n)$, etc.; also, $M_n(1) = \sum_{i=1}^n X_i$. Show that M_n is a martingale, with $\operatorname{Var} M_n(t) = [nt]\sigma^2$.

(b) Show that $\langle M_n, M_n\rangle(t) = [nt]\sigma^2$, and consequently that the scaled process, $M_n/\sqrt{n}$, also a martingale, has variance process

$$\langle M_n/\sqrt{n}, M_n/\sqrt{n}\rangle(t) = ([nt]/n)\sigma^2 \to \sigma^2 t.$$

Verify that the famous *central limit theorem* (CLT) implies that

$$M_n(1)/\sqrt{n} = n^{-1/2}\sum_{i=1}^n X_i \to_d \mathrm{N}(0,\sigma^2).$$

(c) Rather more generally, there is a famous generalisation of the central limit theorem to the full random cum-sum process M_n, called Donsker's Theorem (from around 1950, see Billingsley 1968), which says that

$$M_n(\cdot)/\sqrt{n} \to_d V(\cdot),$$

where the limit is a Gaußian process with mean zero, independent increments, and $\mathrm{Var}\, V(t) = \sigma^2 t$. In fact, such a V is the same as a Brownian motion process W, scaled with σ. The Brownian motion is defined as a zero-mean process where increments are independent and with $W(t_2) - W(t_1) \sim \mathrm{N}(0, t_2 - t_1)$. The limit in distribution takes places inside the space $D[0,\tau]$ of all right-continuous functions $x\colon [0,\tau] \to \mathcal{R}$ with left-hand limits, and with a certain (natural) topology, that of Skorohod. Look into the 'this is saying much more' statement, and give an example. The point is partly that from $M_n \to_d V$ follows

$$H_n = h(M_n) \to_d H = h(V),$$

for every continuous $h\colon D[0,\tau] \to \mathcal{R}$, like $h_1(x) = \max|x(t)|$, $h_2(x) = \max x(t) - \min x(t)$, $h_3(x)$ the amount of time x is above zero, etc.

(d) Before I jump into martingales with dependence on the past, in the next exercise, let me point to the variation of the above where the random variables being summed are independent, but with different distributions. This is the important extension of the classical i.i.d. CLT to the Lindeberg (or Lyapunov) case – after J.W. Lindeberg, Finnish farmer and mathematician, who wrote up his famous paper in 1922, with what is known later on as 'Lindeberg conditions', translatable as 'weak conditions securing that nothing goes wrong, so that the limit is normal'. So consider independent $X_1, X_2, \ldots$, independent with zero means, but perhaps different distributions, and standard deviations $\sigma_1, \sigma_2, \ldots$. Form as above the process

$$M_n(t) = \sum_{i\le[nt]} X_i \quad \text{for } t \ge 0,$$

where the variance is $\sum_{i\le nt}\sigma_i^2$. Show that M_n is a marginale, with

$$\langle M_n/\sqrt{n}, M_n/\sqrt{n}\rangle(t) = (1/n)\sum_{i\le nt}\sigma_i^2.$$

The Lindeberg theorem says, or, rather, implies, in this context, that if the variances are such that this function tends to a positive limit $v(t)$, and if the Lindeberg condition holds, then $M_n(t)/\sqrt{n} \to_d \mathrm{N}(0, v(t))$ for each t, and there is also full process convergence

$$M_n(\cdot)/\sqrt{n} \to_d V(\cdot),$$

where $V(\cdot)$ is a Gaußian martingale, with variance $v(t)$. The Lindeberg condition, in this case, is that

$$L_n(\varepsilon) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\, X_i^2 I\{|X_i| \ge \varepsilon\sqrt{n}\} \to 0 \quad \text{for each } \varepsilon > 0.$$

There are various alternatives and generalisations and extensions and modifications, explaining why 'Lindeberg condition' is a portmanteau word. I record one of these, for the case where $B_n^2 = \sum_{i=1}^n \sigma_i^2$ is not of the order $O(n)$. Then $Z_n = \sum_{i=1}^n X_i/B_n$ tends to the standard normal, provided

$$L_n^*(\varepsilon) = \sum_{i=1}^n \mathrm{E}\,\Big|\frac{X_i}{B_n}\Big|^2 I\Big\{\Big|\frac{X_i}{B_n}\Big| \ge \varepsilon\Big\} \to 0 \quad \text{for each } \varepsilon > 0.$$

(e) [xx just a bit about the proof, with moment generating functions; can be extended to martingales, via clever enough conditioning, etc.]

(f) Let $Y_1, Y_2, \ldots,$ be independent Bernoulli 0–1 variables, with probabilities $p_i = \Pr\{Y_i = 1\}$, and consider the normalised sum

$$Z_n = \frac{\sum_{i=1}^n (Y_i - p_i)}{\{\sum_{i=1}^n p_i(1-p_i)\}^{1/2}}.$$

Show that $Z_n \to_d \mathrm{N}(0,1)$ if and only if $\sum_{i=1}^\infty p_i = \infty$.

11. Martingales have Gaußian process limits

Here I go through a few things having to do with the remarkable and powerful machinery of *martingale limit theorems*, but without the finer details. Such finer details are partly in the ABG book's Section 2.3; see also their appendix B.3, and the Helland (1982) journal article, for clear (but demanding) accounts. The overall message is that yes, lo & behold, a martingale $M_n(t)$, indexed by sample size n, is approximately normal, for each t, when n grows. *Even more*, the full random process $M_n = \{M_n(t)\colon t \ge 0\}$ tends in distribution, when properly scaled, to a full Gaußian martingale process, with independent and normally distributed increments (which is saying much more than merely 'for each t, the distribution of $M_n(t)$ is close to a normal'). All these results are important and for this course very useful generalisations of those briefly surveyed in the previous exercise, which is concerned with *independent summands*; for survival analysis models and methods we very much need results with even complicated dependencies on the past.

(a) Let $X_1, X_2, \ldots$ be a sequence of variables for which

$$\mathrm{E}\,(X_n \,|\, \mathcal{F}_{n-1}) = 0 \quad \text{for } n = 2, 3, \ldots,$$

where $\mathcal{F}_{n-1}$ means the previous history $(X_1, \ldots, X_{n-1})$. Show that

$$M_n(t) = \sum_{i \le [nt]} X_i \quad \text{for } t \ge 0$$

is a martingale. Note that X_n can depend on the past in even complicated ways, as long as its conditional mean is zero.

(b) Show that

$$\langle M_n, M_n\rangle(t) = \sum_{i\le[nt]} V_i, \quad \text{where } V_i = \mathrm{Var}\,(X_i \,|\, \mathcal{F}_{i-1}).$$

In the easier special case of independence, as with the previous exercise, the $\langle M_n, M_n\rangle(t)$ is just the sum of the variances of the X_i; here it is rather the sum of the *conditional variances* (and these are random).

(c) There are now various theorems which say that if (i) $\langle M_n, M_n\rangle(t) \to_{\rm pr} v(t)$ for each t and (ii) some Lindeberg type condition holds, then there is full process convergence $M_n(t) \to_d V(t)$, a Gaußian martingale with variance $v(t)$. Show that $V(t)$ has the same distribution as $W(v(t))$, with W standard Brownian motion.

(d) Now consider a more complicated process, namely

$$K_n(t) = \sum_{i\le[nt]} H_i X_i \quad \text{for } t \ge 0,$$

where the sequence of $H_1, H_2, \ldots$ are *previsible*, meaning that H_i is known once $\mathcal{F}_{i-1}$ is known. Show that K_n is a martingale, with

$$\langle K_n, K_n\rangle(t) = \sum_{i\le[nt]} H_i^2 \Delta\langle M_n, M_n\rangle_i = \sum_{i\le[nt]} H_i^2 \mathrm{Var}\,(X_i \,|\, \mathcal{F}_{i-1}).$$

These are parallels to what we've seen and worked with in Exercise 9.

(e) Since a martingale limit theorem is a martingale limit theorem, deduce that as long as (i) $\langle K_n, K_n\rangle(t) \to_{\rm pr} q(t)$ for each t and (ii) some Lindeberg type condition holds for K_n, then there is full process convergence $K_n(t) \to_d Q(t) = W(q(t))$, a Gaußian martingale with variance $q(t)$. Note that such K_n processes can be much more complicated than simpler sums of independent components processes.

(f) [xx indication of proof. an example. xx]

(g) [xx one or two more points here, the typical use of these theorems, exemplified by nelson–aalen normality below. xx]

12. More on the Nelson–Aalen estimator

Here are a few more technical details and supplementing remarks regarding the Nelson–Aalen estimators, compared to statements reached in Exercise 7. So we work, again, with

$$\widehat{A}(t) = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)} = \sum_{t_i\le t} \frac{\delta_i}{Y(t_i)}.$$

We define 0/0 as 0 when time has run long enough to have $Y(s) = 0$, i.e. nobody left.

(a) Note that this construction makes sense also in event history analysis setups, for each intensity direction, so to speak. There would then be Nelson–Aalen estimators of the form

$$\widehat{A}_{i,j}(t) = \int_0^t \frac{\mathrm{d}N_{i,j}(s)}{Y_i(s)},$$

from box i to box j on the map of possible stations to occupy, with $Y_i(s)$ the number at risk in box i at time just before s, and $\mathrm{d}N_{i,j}(s)$ the number of those for which an event occurs inside the short time interval $[s, s + \mathrm{d}s)$. – For the rest of this exercise, we're in the simpler survival setup, though.

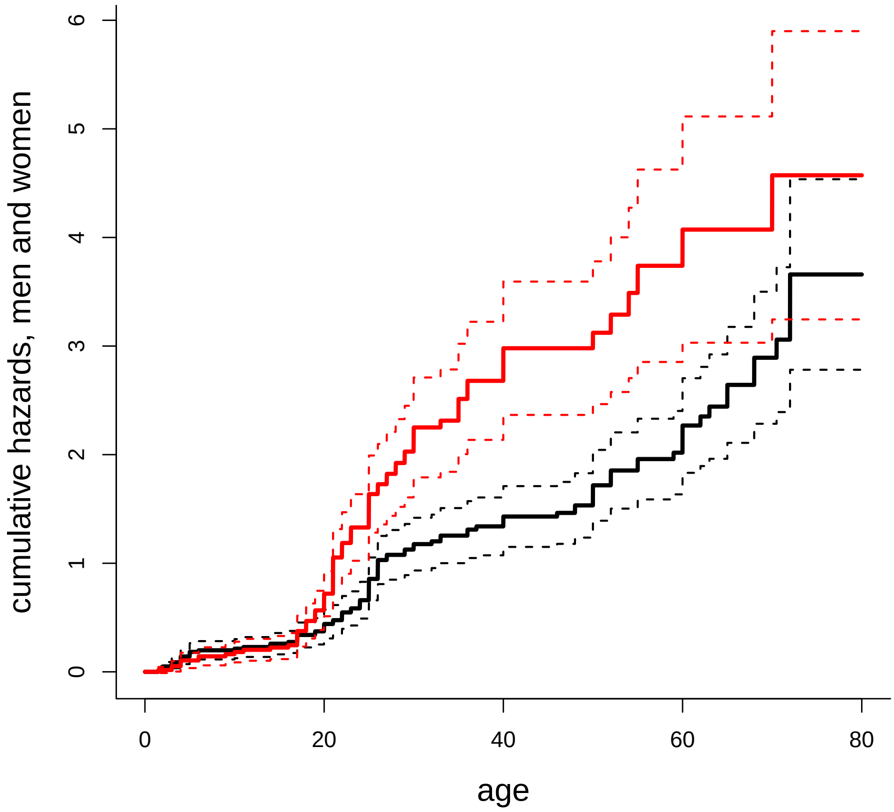

Figure 0.3: Estimated cumulative hazard rates for lives lived in Ancient Egypt, for men (black) and women (red). The full, fat curves are the Nelson–Aalen estimates, the dotted lines are approximate 90% pointwise confidence bands.

(b) With $M(t) = N(t) - \int_0^t Y(s)\alpha(s)\,\mathrm{d}s$ the martingale, show that

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)}\,\mathrm{d}N(s) = A^*(t) + \int_0^t \frac{J(s)}{Y(s)}\,\mathrm{d}M(s),$$

where $J(s) = I\{Y(s) \geq 1\}$. Find a formula for $\Pr\{J(s) = 1\}$, and show that it is very cose to 1, unless s is growing big.

(c) Show that $\widehat{A} - A^*$ is a martingale, with variance process

$$\langle \widehat{A} - A^*, \widehat{A} - A^* \rangle(t) = \int_0^t \frac{J(s)}{Y(s)^2}\mathrm{d}\langle M, M\rangle(s) = \int_0^t \frac{J(s)}{Y(s)}\alpha(s)\,\mathrm{d}s.$$

(d) Show then that the variance of $\widehat{A}(t) - A^*(t)$, which is very nearly the same as the variance of $\widehat{A}(t)$, can be expressed as

$$\sigma(t)^2 = \mathrm{E}\int_0^t \frac{J(s)}{Y(s)}\alpha(s)\,\mathrm{d}s.$$

Argue that a natural estimator of

$$\widehat{\sigma}(t)^2 = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)^2} = \sum_{t_i \leq t} \frac{\delta_i}{Y(t_i)^2}.$$

This is about the same as covered inside Exercise 7, but now with a bit more detail.

(e) A somewhat more elaborate version of the variance estimator above is

$$\widehat\sigma(t)^2=\sum_{t_i\le t}\frac{1}{Y(t_i)}\widehat p_i(1-\widehat p_i)=\sum_{t_i\le t}\frac{1}{Y(t_i)}\frac{\Delta N(t_i)}{Y(t_i)}\Big\{1-\frac{\Delta N(t_i)}{Y(t_i)}\Big\}.$$

First, we use $\Delta N(t_i)$, the number of events observed inside $[t_i,t_i+\varepsilon)$, for a small ε, in the case of ties (the theory says that we should not have two events at the very same time, but in practice data are not always given with very fine time precision). Second, going through arguments above one learns that the variance of $\mathrm{d}N(s)$ given the past enters the variance process arguments, and this conditional variance is $Y(s)p(1-p)$, with the small $p=\alpha(s)\,\mathrm{d}s$. In the mathematical fine limit, where $\mathrm{d}s$ becomes infinitesimal, $p(1-p)=p$, so to speak, but it is possible that the above variance estimator is just slightly better when $Y(s)$ is relatively small; then the estimate $\widehat p=\Delta N(t)/Y(t)$ is not so small, etc.

(f) But we can't quite live on with just a sensible estimator of the sensible variance of a sensible estimator; we need approximate normality, in order to construct confidence intervals and bands, tests, comparisons, etc. So let us prove the $\sqrt{n}\{\widehat A(t)-A(t)\}$ has a normal limit process. First, show that

$$\sqrt{n}\{A^*(t)-A(t)\}\to_{\rm pr}0,\quad\text{for each } t,$$

which means that it is enough to find the limit distribution in question for the simpler

$$Z_n(t)=\sqrt{n}\{\widehat A(t)-A^*(t)\}.$$

But this is a martingale. Show that

$$\widehat y(s)=Y(s)/n\to_{\rm pr}y(s),\quad\text{for each } s,$$

under mild conditions on the censoring distribution. Show that

$$\langle Z_n,Z_n\rangle(t)=\int_0^t\frac{nJ(s)}{Y(s)}\alpha(s)\,\mathrm{d}s\to_{\rm pr}v(t)=\int_0^t\frac{1}{y(s)}\alpha(s)\,\mathrm{d}s.$$

This secures, with a small extra technical argument having to do with Lindeberg conditions (see ABG, Ch. 2, or Helland, 1982, or Hjort, 1990b), that

$$Z_n(\cdot)\to_d V(\cdot)=W(v(\cdot)),$$

a time-transformed Brownian motion, with variance $\mathrm{Var}\,V(t)=v(t)$.

(g) The final statement we need under our belts is that $\widehat\sigma(t)$ is consistent for $\sigma(t)$, or more properly that $n\widehat\sigma(t)^2$ converges in probability to the limit of $n\sigma(t)^2$, which is the $v(t)$ above. Try to prove this.

(h) From these statements prove that

$$\frac{\widehat A(t)-A(t)}{\widehat\sigma(t)}=\frac{\sqrt{n}\{\widehat A(t)-A(t)\}}{\sqrt{n}\sigma(t)}\frac{\sigma(t)}{\widehat\sigma(t)}\to_d\mathrm{N}(0,1),$$

for each t. In particular, show from this that

$$\Pr\{A(t)\in\widehat A(t)\pm z\widehat\sigma(t)\}\to\Pr\{-z\le\mathrm{N}(0,1)\le z\},$$

yielding pointwise confidence bands, tests for hypotheses of the type $A=A_0$, etc.

13. More on Ancient Egypt

Slip into your Wellsian time machine and go back in time to Roman Era Egypt (cf. Exercises 1, 2). Compute and display the Nelson–Aalen estimators for the cumulative hazards for men and women. Supplement these with approximate pointwise 90% confidence bands. In other words, attempt to reproduce Figure 0.3. Also, plot the function

$$D(t) = \widehat{A}_w(t) - \widehat{A}_m(t),$$

estimated cumulative hazard difference, unfortunate women minus fortunate men (this was a time of relative peace and no wars, before the later tax revolt etc.), along with an approximate 90% confidence band. What are your conclusions?

14. The Kaplan–Maier estimator

As ABG argue, the cumulative hazards of the world are more versatile and useful tools, particularly in more complex event history setups, than 'only' the task of estimating the survival curve $S(t) = \Pr\{T \geq t\}$ for survival data $(t_1, \delta_1), \ldots, (t_n, \delta_n)$. The canonical nonparametric estimator for $S(t)$ remains however an important quantity, and this is the Kaplan–Meier estimator (from JASA, 1958). Its definition is

$$\widehat{S}(t) = \prod_{[0,t]} \Big\{1 - \frac{\mathrm{d}N(s)}{Y(s)}\Big\} = \prod_{t_i \leq t} \Big\{1 - \frac{\delta_i}{Y(t_i)}\Big\}.$$

(a) For non-censored data, say $t_1 < \cdots < t_n$, show that $\widehat{S}(t)$ becomes the simpler empirical survival function

$$\widehat{S}_{\rm emp}(t) = (1/n) \sum_{i=1}^n I\{t_i \geq t\} = 1 - F_{\rm emp}(t).$$

(b) We do have the easy formula $S = \exp(-A)$, binding together the survival curve and the cumulative hazard, so it is not at all forbidden to start with the Nelson–Aalen and then exp its minus to arrive at

$$\widetilde{S}(t) = \exp\{-\widehat{A}(t)\}.$$

The $A = -\log(1-F)$ formula is valid only for continuous distributions, however, so that particular connection is not as straightforward for non-continuous step-function type estimators as $\widehat{A}$ and $\widehat{S}$. You may however attempt to prove that the Kaplan–Meier $\widehat{S}$ and Aalen-related estimator are quite close; in particular, when n increases,

$$\sqrt{n}[\widehat{S}(t) - \exp\{-\widehat{A}(t)\}] \to_{\rm pr} 0.$$

This means that these two related estimators have the same large-sample properties. Proving the above is easier when working on the log-scale; attempt to show that

$$\sqrt{n}[\widehat{A}(t) + \log \widehat{S}(t)\}] \to_{\rm pr} 0.$$

This has to do with $-\log(1-x) = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \cdots$, which is very close to simply x for small x, etc.

(c) [xx briefly on how to assess and estimate the variance of $\widehat{S}(t)$. the Greenwood formula (from 1926!). limiting normality and confidence bands. xx]

15. The Hjort estimators of A and S, from Bayesian nonparametrics

The Nelson–Aalen and Kaplan–Meier estimators have nice generalisations to the Bayesian nonparametrics setting. Hjort (1985a, 1990b) introduced *Beta processes* as the natural class of priors for cumulative hazard rates. The starting point is to work with

$$\mathrm{d}A(s) = \frac{\mathrm{d}F(s)}{F[s,\infty)} \quad \text{and} \quad F(t) = 1 - \prod_{[0,t]}\{1 - \mathrm{d}A(s)\},$$

and then to form a prior process $A(t) = \int_0^t \mathrm{d}A(s)$ with independent and almost Beta distributed increments,

$$\mathrm{d}A(s) \sim \mathrm{Beta}[c(s)\,\mathrm{d}A_0(s), c(s)\{1 - \mathrm{d}A_0(s)\}].$$

Here $A_0(t) = \int_0^t \alpha_0(s)\,\mathrm{d}s$ is the prior mean and $c(s)$ a strength-of-prior function (e.g. a constant). Existence of such a process is non-trivial, as sums of Beta variables are not Beta distributed, but such Beta processes are shown to exist as proper time-coninuous limits.

Hjort (1985a, 1990b) shows that $A\,|\,\mathrm{data}$ is a new and updated Beta process, with $c(s)$ updated to $c(s) + Y(s)$ and the prior mean $A_0(t)$ updated to the Bayesian nonparametrics estimator

$$\widehat{A}_B(t) = \mathrm{E}\{A(t)\,|\,\mathrm{data}\} = \int_0^t \frac{c(s)\alpha_0(s)\,\mathrm{d}s + \mathrm{d}N(s)}{c(s) + Y(s)}.$$

Similarly, there is a natural Bayesian nonparametrics estimator of the survival curve,

$$\widehat{S}_B(t) = \mathrm{E}\{S(t)\,|\,\mathrm{data}\} = \prod_{[0,t]}\Big\{1 - \frac{c(s)\alpha_0(s)\,\mathrm{d}s + \mathrm{d}N(s)}{c(s) + Y(s)}\Big\}.$$

When $c(s)$ becomes small, or the data volume grows, these are then close to the Nelson–Aalen and Kaplan–Meier estimators. In yet other words, the Nelson–Aalen $\widehat{A}$ and Kaplan–Meier $\widehat{S}$ may be given the interpretation of being Bayesian nonparametrics estimators under a non-informative Nils-Beta process prior, where $c(s) \to 0$.

(a) Choose your own prior parameters $\alpha_0(s)$ and $c(s)$ for the Ancient Egypt data, perhaps the same for men and women, and plot the resulting Bayes estimators $\widehat{A}_B(t)$ and $\widehat{S}_B(t)$, for men and women.

(b) Hjort (1990b) shows that

$$\mathrm{Var}\,\{A(t)\,|\,\mathrm{data}\} = \int_0^t \frac{\mathrm{d}\widehat{A}_B(s)\{1 - \mathrm{d}\widehat{A}_B(s)\}}{c(s) + Y(s) + 1}.$$

For the Ancient Egypt data, again, plot the band

$$\widehat{A}_B(t) \pm 1.645\,\{\mathrm{Var}\,\{A(t)\,|\,\mathrm{data}\}\}^{1/2},$$

for men and for women, and comment.

(c) If you wish you may also simulate say 50 full realisations of $A\,|\,\mathrm{data}$, or $S\,|\,\mathrm{data}$ (or for any other quantity in which you may take an interest), via independent small increments

$$\mathrm{d}A(s)\,|\,\mathrm{data} \sim \mathrm{Beta}[c(s)\,\mathrm{d}A_0(s) + \mathrm{d}N(s), c(s)\{1 - \mathrm{d}A_0(s)\} + Y(s) - \mathrm{d}N(s)],$$

and display these in a diagram. See the Nils Exercises and Lecture Notes (Spring 2018) from the Bayesian Nonparametrics course STK 9190, e.g. Exercise 28 with figures.

16. Maximum likelihood estimators are approximately unbiased, approximately multinormal, and their variance matrix can be estimated too!

[xx well: good news for mankind. i also make the point that i can write a somewhat different book (if i take the time), partly with the same material as ABG, but i can start with a healthy amount of parametrics, and then go to nonparametrics afterwards; cf. the exercise with a parametric step function leading to nelson–aalen etc. xx]

(a) [xx first $\psi(s,\theta)=\partial\log\alpha(s,\theta)/\partial\theta$ and a bit on

$$U_n=\int_0^\tau \psi(s,\theta_0)\{\mathrm{d}N(s)-Y(s)\alpha(s,\theta_0)\,\mathrm{d}s\}=\int_0^\tau \psi(s,\theta_0)\,\mathrm{d}M(s)$$

which becomes a martingale with clear variance process. xx]

(b) [xx then $\ell_n(\theta)$ and its derivative. xx]

(c) [xx and its second order derivative, the Hessian, and two estimators for $J=J(\theta_0)$. xx]

(d) [xx then $\sqrt{n}(\widehat\theta-\theta_0)\to_d \mathrm{N}(0,J^{-1})$, as for ordinary ML. mention Borgan (1984), Hjort (1986). xx]

(e) [xx an illustration. xx]

References

Aalen, O.O. (1975). Statistical Inference for a Family of Counting Processes. PhD thesis, Department of Statistics, University of Berkeley, California.

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

Aalen, O.O., Borgan, Ø., and Gjessing, H.K. (2008). *Survival and Event History Analysis: A Process Point of View.* Springer.

Aalen, O.O. and Hjort, N.L. (2002). Frailty models that yield proportional hazarads. *Statistics and Probability Letters* **58**, 335–342.

Andersen, P.K., Borgan, Ø., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics* **11**, 1–16.

Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival curve based on transformations. *Scandinavian Journal of Statistics* **17**, 35–41.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

Cunen, C., Hjort, N.L., and Nygård, H. (2018). Statistical sightings of better angels. Submitted for publication.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press, Cambridge.

Gjessing, H.K., Aalen, O.O. and Hjort, N.L. (2003). Frailty models based on Lévy processes. *Advances in Applied Probability* **35**, 532–550.

Greenwood, M. Jr. (1926). The natural duration of cancer. *Reports of Public Health and Related Subjects* **33**, HMSO, London.

Helland, I.S. (1982). Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics* **9**, 79–94.

Hermansen, G.H. and Hjort, N.L. (2015). Bernshteĭn–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach. *Journal of Statistical Planning and Inference* **166**, 138–157.

Hjort, N.L. (1985a). Discussion contribution to P.K. Andersen and Ø. Borgan's 'Counting process models for life history data: A review'. *Scandinavian Journal of Statistics* **12**, 141–151.

Hjort, N.L. (1985b). An informative Bayesian bootstrap. Technical Report, Department of Statistics, Stanford University.

Hjort, N.L. (1986). Discussion contribution to P. Diaconis and D. Freedman's paper 'On the consistency of Bayes estimators'. *Annals of Statistics* **14**, 49–55.

Hjort, N.L. (1986). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics* **13**, 63–75.

Hjort, N.L. (1990a). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Annals of Statistics* **18**, 1221–1258.

Hjort, N.L. (1990b). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hjort, N.L. (1991). Bayesian and empirical Bayesian bootstrapping. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. (2003). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort, S. Richardson). Oxford University Press, Oxford.

Hjort, N.L. (2018). Towards a More Peaceful World [Insert '!' or '?' Here]. FocuStat Blog Post.

Hjort, N.L., Holmes, C.C., Müller, P., and Walker, S.G. (2010). *Bayesian Nonparametrics.* Cambridge University Press, Cambridge.

Hjort, N.L. and Petrone, S. Nonparametric quantile inference using Dirichlet processes. In *Festschrift for Kjell Doksum* (ed. V. Nair).

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **15**, 211–225.

Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika* **1**, 261–264.

Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability.* Cambridge University Press, Cambridge.