Exercises and Lecture Notes, STK 4080, Autumn 2018

Version 0.59, 2-x-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 $S(t) = \Pr\{T \ge t\}$. 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

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

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.

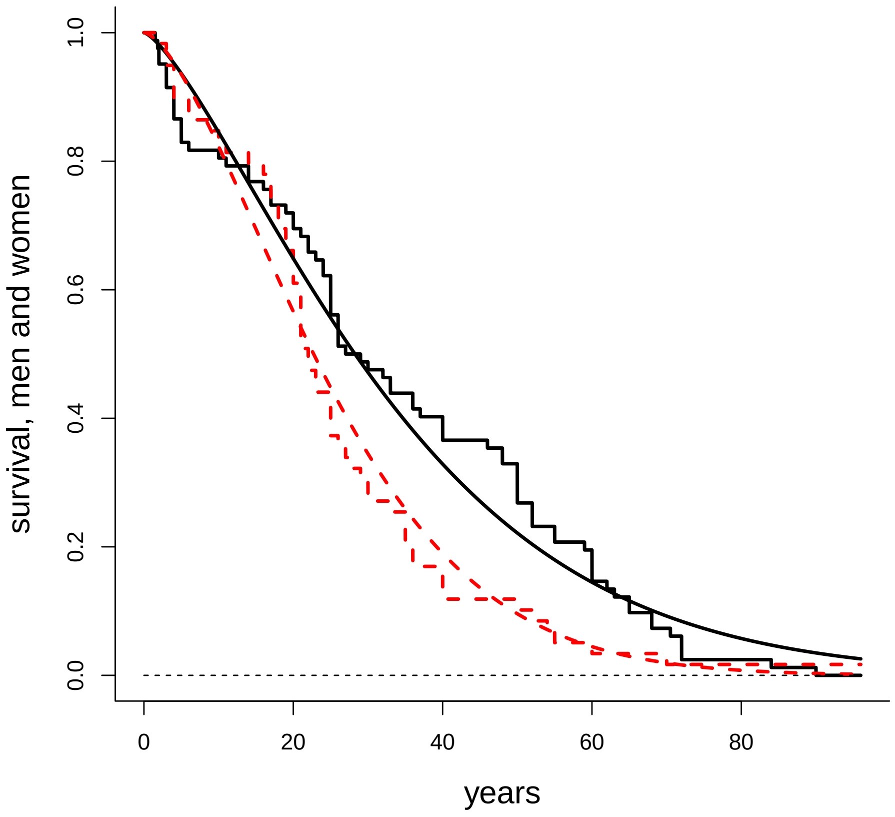

Figure 0.1: Survival curves from Roman era Egypt, two for men (black) and two for women (red). The step-functions are the empirical survival curves, type Kaplan–Meier; the continuous curves are the fitted Weibull curves.

(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

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

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

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

(a) First define the survival function as

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

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 $y(s) = \Pr\{T \geq s, C \geq s\}$, with C the censoring mechanism. It follows that

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

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.

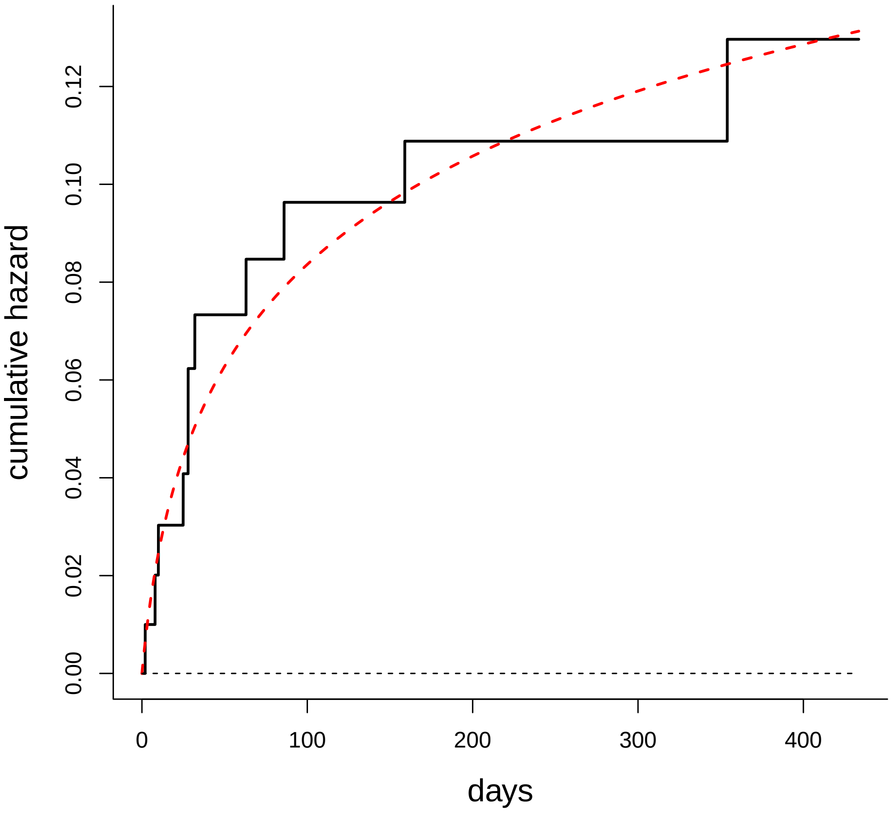

Figure 0.2: Estimated cumulative hazard rate for time to expulsion of IUD, via the nonparametric Nelson–Aalen estimator and the parametric frailty model.

(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. Convergence in probability

I've scissored in Exercises 9, 10, 11, 12, 13, 14, 15, 16 from a Nils Collection from the course STK 4011, Autumn 2014. These exercises will not be dicussed in the present STK 4080-9080 course, except perhaps in passing. They might be useful so some of the students for making certain details clearer, regarding the machinery of large-sample approximations (involving various aspects of convergence in distribution – what it is, why being interested in it, how to prove it, how to use it). So students are advised to glance through these exercises. In Exercises 17, 18, 19 and yet others, the STK 4011 type material is extended to the STK 4080 world of counting processes and martingales.

So, consider a sequence of random variables $V_1, V_2, \ldots$. We say that V_n converges in probability to the constant a, and write $V_n \to_{\rm pr} a$, if

$$\Pr(|V_n - a| \le \varepsilon) \to 1 \quad \text{for all } \varepsilon > 0$$

as $n \to \infty$. The definition extends easily to the case where the limit in probability is a random variable V rather than a constant, and is also equivalent to

$$\Pr(|V_n - V| \ge \varepsilon) \to 0 \quad \text{for all } \varepsilon > 0.$$

For most of our applications inside the STK 4011 course the probability limit will in fact be a constant, however, i.e. not a random variable per se.

(a) Show that if $V_n \to_{\rm pr} a$ and $h(v)$ is a function continuous at a, then $h(V_n) \to_{\rm pr} h(a)$.

(b) Extend the previous result to the case where the probability limit is a random variable, i.e. if $V_n \to_{\rm pr} V$ and $h(v)$ is continuous on the domain of V, then $h(V_n) \to_{\rm pr} h(V)$. (Explain also why the proof indicated in the book's exercises is not fully correct [tsk tsk].)

(c) Suppose $A_n \to_{\rm pr} a$ and $B_n \to_{\rm pr} b$. Show that $A_n + B_n \to_{\rm pr} a + b$ and that $A_nB_n \to_{\rm pr} ab$. Attempt to generalise these results; in effect, $h(A_n, B_n) \to_{\rm pr} h(a, b)$ provided h is continuous at position (a, b).

10. The Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. variables, with $\mathrm{E}\,X_i = \xi$ and $\mathrm{Var}\,X_i = \sigma^2$.

(a) Show that the sequence of averages $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ converges in probability to ξ, i.e. that $\bar{X}_n \to_{\mathrm{pr}} \xi$. You may use Chebyshov's inequality (неравенство Чебышёва). The Law of Large Numbers (LLN) says that we still have $\bar{X}_n \to_{\mathrm{pr}} \xi$, even without further assumptions that the mean is finite, i.e. even if the variance is infinite; the proof becomes more complicated, however.

(b) Suppose the variance σ^2 is finite. Show that

$$S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \to_{\mathrm{pr}} \sigma^2.$$

Explain why this also implies that $S_n \to_{\mathrm{pr}} \sigma$. We say that S_n is a consistent estimator for the parameter σ; similarly, $\bar{X}_n$ is consistent for the mean parameter ξ.

(c) Suppose that also the third moment is finite. Show that

$$T_n = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^3 \to_{\mathrm{pr}} \gamma_3 = \mathrm{E}\,(X_i - \xi)^3,$$

and that the so-called empirical skewness converges to the theoretical skewness:

$$\widehat{\kappa}_3 = n^{-1}\sum_{i=1}^n \Big(\frac{X_i - \bar{X}_n}{S_n}\Big)^3 = \frac{T_n}{S_n^3} \to_{\mathrm{pr}} \kappa_3 = \mathrm{E}\Big(\frac{X_i - \xi}{\sigma}\Big)^3.$$

(d) Generalise the above to the case of higher order moments.

11. Convergence in distribution

Let $V_1, V_2, \ldots$ be a sequence of random variables. We say that V_n converges in distribution to V, and write $V_n \to_d V$ to indicate this, if

$$F_n(t) = \Pr(V_n \le t) \to F(t) = \Pr(V \le t) \quad \text{for all } t = C_F$$

as $n \to \infty$, where C_F is the set of points at which the cdf F of the limit distribution is continuous. In particular, if this limit distribution is continuous, $V_n \to_d V$ if $F_n(t) \to F(t)$ for all t.

(a) Show that if $V \to_d V$, then

$$\Pr(V_n \in (a,b]) \to \Pr(V \in (a,b])$$

for all intervals $(a,b]$ for which a, b are continuity points. If $V_n \to_d \mathrm{N}(0,1)$, where this is accepted and traditional short-hand notation for the more cumbersome '$V_n \to_d V$, where $V \sim \mathrm{N}(0,1)$', etc., then $\Pr(|V_n| \le 1.96) \to 0.95$, etc.

(b) For an i.i.d. sample $U_1, \ldots, U_n$ from the uniform distribution on $(0,1)$, let $M_n = \max_{i \le n} U_i = U_{(n)}$. Find the limit distribution of $V_n = n(1 - M_n)$.

(c) Suppose the V_n and the V have distributions on the integers $0, 1, 2, \ldots$, with probabilities $\Pr(V_n = j) = f_n(j)$ and $\Pr(V = j) = f(j)$ for $j = 0, 1, 2, \ldots$. Prove that $V_n \to_d V$ is equivalent to convergence of these probabilities, i.e. $f_n(j) \to f(j)$ for each j.

(d) Suppose V_n is a binomial (n, p_n) where $np_n \to \lambda$, a positive parameter. Show that $V_n \to_d \text{Pois}(\lambda)$. This is how the Poisson distribution first saw light, in 1837 (though a much earlier account, containing more or less the same approximation results, is by de Moivre in 1711).

(e) Generalise the above result to the following 'law of small numbers'. Let $X_1, X_2, \ldots$ be independent binomials $(1, p_i)$ with small probabilities $p_1, p_2, \ldots$, and consider $V_n = \sum_{i=1}^n X_i$, the number of events after n trials. Show that if $\sum_{i=1}^n p_i \to \lambda$ and $\delta_n = \max_{i \le n} p_i \to 0$, then $V_n \to_d \text{Pois}(\lambda)$. Show also that these conditions are also necessary for convergence to a Poisson.

12. Convergence of densities

Suppose that V_n and V have densities f_n and f.

(a) Show that if $f_n(v) \to f(v)$ for all v, then there is also convergence of their cumulatives, i.e. $F_n(v) \to F(v)$ for all v. In other words, convergence of density functions implies convergence in distribution.

(b) If $f_n \to f$ as above, show the somewhat stronger result

$$\int |f_n(v) - f(v)| \, \mathrm{d}v \to 0.$$

This is called 'L_1 convergence', and is also equivalent to convergence in the supremum probability difference metric,

$$\Delta(P_n, P) = \sup_{\text{all } A} |P_n(A) - P(A)| \to 0.$$

(c) Work with the density of the t_m, the t distribution with m degrees of freedom, and show that it converges to the famous $\mathrm{N}(0,1)$ density as $m \to \infty$.

(d) For an i.i.d. sample $U_1, \ldots, U_n$ from the uniform distribution on the unit interval, consider the median M_n, where we for simplicity take $n = 2m+1$ to be odd, so that $M_n = U_{(m+1)}$. Work out the density for M_n and then the density $g_n(v)$ for $V_n = \sqrt{n}(M_n - \frac{1}{2})$. Show that in fact

$$g_n(v) \to \frac{1}{\sqrt{2\pi}} 2 \exp(-2v^2),$$

where you may need Stirling's formula, $m! \doteq m^m \exp(-m)\sqrt{2\pi m}$. Thus $\sqrt{n}(M_n - \frac{1}{2}) \to_d \mathrm{N}(0, \frac{1}{4})$.

(e) Give an approximation formula for $\Pr(0.49 \le M_n \le 0.51)$, and determine how big n needs to be in order for this probability to be at least 0.99.

13. The portmanteau theorem for convergence in distribution

The definition of convergence in distribution given above, in therms of their cumulative distribution functions, is somewhat cumbersome and not easy to work with, so we need reformulations and alternative conditions.

For random variables V_n and V with cumulative distribution functions F_n and F, corresponding also to probability measures $P_n(A) = P(V_n \in A)$ and $P(A) = P(V \in A)$ (where the point is that also more complicated sets A may be worked with than only intervals), consider the following statements:

(i) $V_n \to_d V$, i.e. $F_n(v) \to F(v)$ for continuity points v, as defined above.

(ii) $\liminf P_n(O) \ge P(O)$ for all open sets O.

(iii) $\limsup P_n(F) \le P(F)$ for all closed sets F.

(iv) $\lim P_n(A) = P(A)$ for all sets A for which its boundary set $\partial(A) = \bar{A} - A^o$ has P-probability zero. Here $\bar{A}$ is the smallest closed set containing A and A^0 is the biggest open set inside A; thus $\partial(A)$ for the interval (a, b) would be the two-point set $\{a, b\}$, and likewise for $[a, b]$, $(a, b]$, $[a, b)$.

(v) $\mathrm{E}\, h(V_n) \to_d \mathrm{E}\, h(V)$ for each continuous and bounded $h\colon \mathcal{R} \to \mathcal{R}$.

The purpose of this exercise is to show that in fact (i) $\iff$ (ii) $\iff$ (iii) $\iff$ (iv) $\iff$ (v), i.e. these five conditions are equivalent. This is the 'portmanteau theorem' for convergence in distribution, due, I believe, to Aleksandrov (1943).

(a) Show that (i) $\Rightarrow$ (ii). Use the mathematical analysis fact that a given open set O may be represented as a finite or countable union of disjoint open intervals (a_i, b_i).

(b) Show that (ii) $\Rightarrow$ (iii), by using the fact that a set F is closed if and only if its complement F^c is open. This also gives (iii) $\Rightarrow$ (ii).

(c) Show that (iii) $\Rightarrow$ (iv).

(d) Show that (iv) $\Rightarrow$ (v), as follows. Take a bounded continuous function h, and for simplicity stretch and scale it so that it lands inside $[0, 1]$. Then argue that

$$\mathrm{E}\, h(V_n) = \int_0^1 P(h(V_n) \ge x)\,\mathrm{d}x \quad \text{and} \quad \mathrm{E}\, h(V) = \int_0^1 P(h(V) \ge x)\,\mathrm{d}x.$$

This is related to the general fact that for any nonnegative random variable Y with cumulative distribution function G, say, we have

$$\mathrm{E}\, Y = \int_0^\infty \{1 - G(y)\}\,\mathrm{d}y = \int_0^\infty P(Y \ge y)\,\mathrm{d}y.$$

Convergence of $\mathrm{E}\, h(V_n)$ to $\mathrm{E}\, h(V)$ then follows by showing that $P(h(V_n) \ge x)$ converges to $P(h(V) \ge x)$ for all x except for at most a countable number of exceptions. Lebesgue's theorem on convergence of integrals may be called upon.

(e) Finally show that (e) $\Rightarrow$ (a). For given v at which F is continous, build a continuous bounded function h_ε so that $h_\varepsilon(x) = 1$ for $x \le v$ and $h_\varepsilon(x) = 0$ for $x \ge v + \varepsilon$, where ε is positive and small. Play a similar game with another function being 1 to the left of $v - \varepsilon$ and 0 to the right of v.

14. The continuity theorem

Show that if $V_n \to_d V$ and g is continuous, then $g(V_n) \to_d g(V)$. The g function here may be unbounded, so $\exp(V_n) \to_d \exp(V)$ etc.

(a) Suppose $V_n \to_d \mathrm{N}(0, \sigma^2)$. Show that $V_n^2/\sigma^2 \to_d \chi_1^2$. What is the limit of $|V_n|/\sigma$?

(b) Assume that nonnegative variables $X_1, X_2, \ldots$ are such that the sequence of geometric means converges in distribution, say $G_n = (X_1 \cdots X_n)^{1/n} \to U$. Show that

$$n^{-1} \sum_{i=1}^{n} \log X_i \to_d V,$$

and identify the limit V.

(c) Suppose again that $V_n \to_d V$. Show that $\exp(tV_n) \to_d \exp(tV)$, for each given t. When can we expect this to lead to

$$M_n(t) = \mathrm{E}\, \exp(tV_n) \to M(t) = \mathrm{E}\, \exp(tV)\,?$$

(d) One can indeed show a counterpart to the above, stated and used in the book without a proof: If $M_n(t) \to M(t)$, for each t in some neighbourhood $(-\delta, \delta)$ around zero, then $V_n \to_d V$. A full proof of this may be found in 'Hjorts lille grønne' from 1979 ('Kompendium for sannsynlighetsregning III', used in a course on large-sample theory for probability and statistics here at the Department of Mathematics at the University of Oslo for some fifteen years), or in e.g. Billingsley's Convergence of Probability Measures (1999). It involves characteristic functions and inversion formuale, giving us formulae for distributions in terms of such functions.

15. Slutsky–Cramér Rule

Certain very useful rules, sometimes called the Slutsky Rules, but equally due to Harald Cramér, rule. They can be presented in various ways, depending also on what precisely one has learned in advance.

(a) If $X_n \to_d X$ and $Y_n \to_{\mathrm{pr}} 0$, show that $X_nY_n \to_{\mathrm{pr}} 0$. To prove this, start from

$$\begin{aligned}\Pr(|X_nY_n| \geq \varepsilon) &= \Pr(|X_nY_n| \geq \varepsilon, |X_n| \leq M) + \Pr(|X_nY_n| \geq \varepsilon, |X_n| > M)\\ &\leq \Pr(|Y_n| \geq \varepsilon/M) + P(|X_n| > M),\end{aligned}$$

from which it follows that $\limsup P(|X_nY_n| \geq \varepsilon) \leq r(M)$, where

$$r(M) = \limsup P(|X_n| > M).$$

Show from convergence in distribution that $r(M)$ may be made arbitrarily small; hence $X_nY_n \to_{\mathrm{pr}} 0$.

(b) If $X_n \to_d X$ and $Y_n \to_{\mathrm{pr}} 0$, show that $X_n + Y_n \to_d X$.

(c) Now change the above assumptions to $X_n \to_d X$ and $Y_n \to_{\mathrm{pr}} y$, with a y non-zero constant. Use the above to show that $X_nY_n \to_d Xy$, $X_n + Y_n \to_d X + y$ and $X_n/Y_n \to_d X/y$.

(d) Try also to show that as long as $g(x', y')$ is continuous on the domain of X and at position y, then $g(X_n, Y_n) \to_d g(X, y)$. Explain how this generalises the previous results.

16. The Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. and for simplicity here with mean zero and standard deviation one. Consider

$$Z_n = \sqrt{n}\bar X_n = n^{-1/2}\sum_{i=1}^n X_i,$$

where it is to be noted that Z_n has mean zero and variance one, for each n. The Central Limit Theorem (the CLT) says that $Z_n \to_d \mathrm{N}(0,1)$, i.e. that

$$P(a \le \sqrt{n}\bar X_n \le b) \to P(a \le \mathrm{N}(0,1) \le b) \quad \text{for all intervals } (a,b).$$

A full proof, without further assumptions, needs e.g. characteristic functions, see 'Hjorts lille grønne' (1979) or Billingsley (1999). A satisfactory proof may however be given for the case of X_i having a moment-generating function $M(t) = \mathrm{E}\,\exp(tX)$ being finite in a neighbourhood around zero, appealing to the result about convergence of moment-generating functions discussed in Exercise xx.

Under the above conditions, show that

$$M(t) = 1 + \tfrac12 t^2 + \tfrac16 \mathrm{E}, X_i^3 t^3 + \tfrac{1}{24}\mathrm{E}\,X_i^4 t^4 + \cdots = 1 + \tfrac12 t^2 + r(t),$$

say, where $r(t)$ is small enough to make $r(t)/t^2 \to 0$ as $t \to 0$. Now work through the details to learn that

$$M_n(t) = \mathrm{E}\,\exp(tZ_n) = M(t/\sqrt{n})^n = \{1 + \tfrac12 t^2/n + r(t/\sqrt{n})\} \to \exp(\tfrac12 t^2) = \mathrm{E}\,\exp(tZ),$$

where $Z \sim \mathrm{N}(0,1)$.

Show from the CLT that if X_n is binomial (n,p), then

$$\frac{X_n - np}{\{np(1-p)\}^{1/2}} \to_d \mathrm{N}(0,1),$$

and that if Y_n is $\mathrm{Pois}(n)$, then

$$\frac{Y_n - n}{\sqrt{n}} \to_d \mathrm{N}(0,1).$$

Show finally that if $Z_n \sim \chi^2_n$, then

$$\frac{Z_n - n}{\sqrt{2n}} \to_d \mathrm{N}(0,1).$$

17. The delta method

The delta method is a very useful and largely easy to use class of tools for approximating the distribution of variables which are functions of simpler variables. So, if W_n is a complicated creature, but after all a function $h(X_n, Y_n, Z_n)$ of simpler fellows X_n, Y_n, Z_n, the idea is to first work out things for these, and then go back to W_n afterwards. This is particularly fruitful when the X_n, Y_n, Z_n in question have tight distributions, with relatively small variances, say around a, b, c. Then a Taylor argument says

$$\begin{aligned} W_n &= h(X_n, Y_n, Z_n) \\ &= h(a,b,c) + h_1^*(a,b,c)(X_n - a) + h_2^*(Y_n - b) + h_3^*(Z_n - c) + \text{smaller order terms}, \end{aligned}$$

where h_1^*, h_2^*, h_3^* are the partial derivatives of $h(x,y,z)$, computed at position (a,b,c). A further highly useful application of this is when (X_n, Y_n, Z_n) is exactly or approximately multinormal, since this implies that the sum on the right hand side is approximately normal, with variance

$$\tau_n^2 = (h^*)^{\mathrm{t}} \Sigma_n h^*,$$

in terms of $h^* = (h_1^*, h_2^*, h_3^*)^{\rm t}$ and the variance matrix Σ_n of (X_n, Y_n, Z_n).

When you've understood the essence of the above (model this as a survival time via a parametric model, estimate its parameters, and calculate the median time until you've comprehended this essence), you've also understood the basics of the delta method.

(a) Show that if a vector Y_n has mean ξ_n and variance matrix Σ_n, then the transformed variable

$$W_n = h^{\rm t} Y_n = h_1 Y_{n,1} + \cdots + h_p Y_{n,p}$$

has mean $h^{\rm t}\xi_n$ and variance matrix $h^{\rm t}\Sigma_n h$ (with p the length of the Y_n variable). So far there is no approximation going on.

(b) Assume now that $W_n = h(Y_n)$, with a function h which is not linear, but perhaps approximately linear in a neighbourhood of $\xi_n = {\rm E}\, Y_n$, and that Y_n has reasonably high probability of being inside this neighbourhood. Argue that

$${\rm E}\, W_n \approx (h^*)^{\rm t}\xi_n \quad \text{and} \quad {\rm Var}\, W_n \approx (h^*)^{\rm t}\Sigma_n h^*,$$

with h^* is the partial derivatives vector $\partial h(y)/\partial y$, evaluated at position ξ_n.

(c) Results of type (b) are very useful, but not very precise. The delta method, in one of its several versions, is a precise limit distribution version of the above. Consider the one-dimensional situation, and assume that $\sqrt{n}(Y_n - \xi) \to_d Z$, for an appropriate limit variable Z. Show that if $h(y)$ is smooth, with a derivative at ξ, then

$$\sqrt{n}\{h(Y_n) - h(\xi)\} \to_d h'(\xi) Z.$$

Show in addition that if Z is normal, say $Z \sim {\rm N}(0, \tau^2)$, then the limit is a ${\rm N}(0, h'(\xi)^2\tau^2)$.

(d) Then generalise to the vector case. Show that

$$\sqrt{n}(Y_n - \xi) \to_d Z \quad \text{implies} \quad \sqrt{n}\{h(Y_n) - h(\xi)\} \to_d (h^*)^{\rm t} Z,$$

if $h(y) = h(y_1, \ldots, y_p)$ has a derivative h^* at position ξ. Show in particular that

$$\sqrt{n}(Y_n - \xi) \to_d {\rm N}_p(0, \Sigma) \quad \text{implies} \quad \sqrt{n}\{h(Y_n) - h(\xi)\} \to_d (h^*)^{\rm t} Z \sim {\rm N}(0, (h^*)^{\rm t}\Sigma h^*).$$

(e) An easy example: When X_n is binomial (n, p) it's been known since around 1738 that $X_n \approx_d {\rm N}(np, np(1-p))$, which also means, in the language limit theorems, that $\sqrt{n}(X_n/n - p) \to_d {\rm N}(0, p(1-p))$. Find the limit distributions of

$$\sqrt{n}\{\exp(4.44\, X_n/n) - \exp(4.44, p)\} \quad \text{and} \quad \sqrt{n}(2 \arcsin \sqrt{X_n/n} - 2 \arcsin \sqrt{p}).$$

How can this second result be used to set a confidence interval for p?

(f) Then go back to the Nelson–Aalen estimator of Exercise 7 (and met frequently later on, also in this collection of exercises). Show that

$$\sqrt{n}\{\log \widehat{A}(t) - \log A(t)\} \to_d W(t),$$

where $W(t)$ is a zero-mean normal with variance

$${\rm Var}\, W(t) = \frac{1}{A(t)^2} \int_0^t \frac{\alpha(s)\,{\rm d}s}{y(s)}.$$

Explain in detail how this may be used to set a confidence interval for first $\log A(t)$, and then, by exp-ing, for $A(t)$.

18. 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 $M = \{M(t)\colon t \ge 0\}$ is a martingale, with respect to the filtration $\{\mathcal{F}_t\colon t \ge 0\}$, 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 $H = \{H(s)\colon s \ge 0\}$, and form its integral with respect to the martingal M:

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

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

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

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 $J(s) = I\{Y(s) \ge 1\}$ 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.

19. 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 $\mathrm{Var}\, M_n(t) = [nt]\sigma^2$.

(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 $\sum_{i=1}^\infty p_i = \infty$.

20. 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) $\langle M_n, M_n \rangle(t) \to_{\mathrm{pr}} v(t)$ 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]

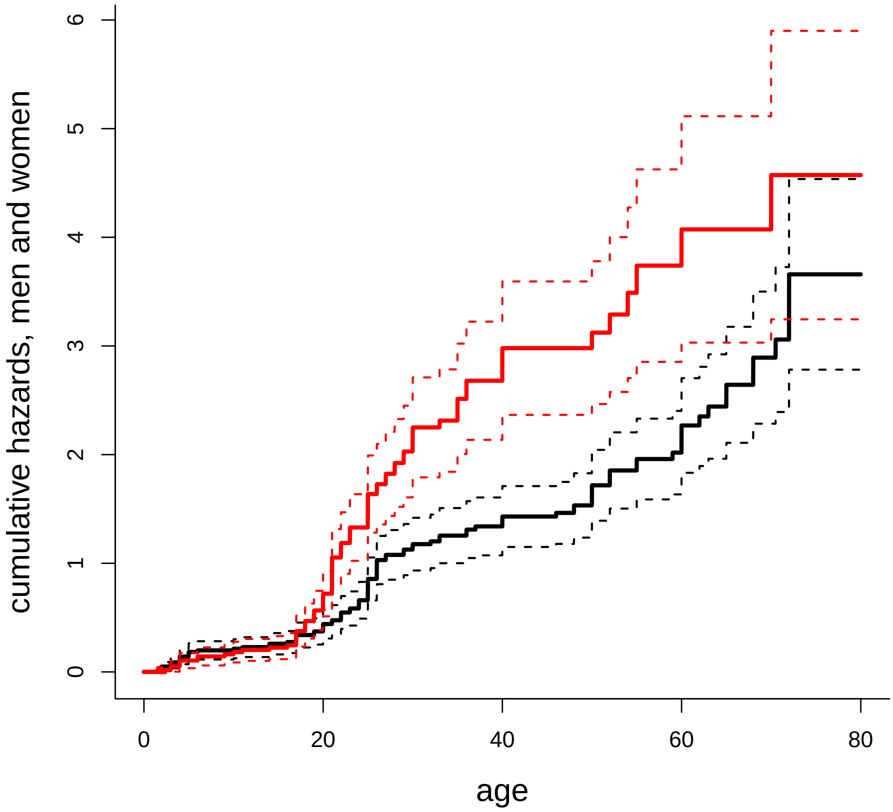

Figure 0.3: Estimated cumulative hazard rates for lives lived in Ancient Egypt, for men (black) and women (red). The full, fat curves are the Nelson–Aalen estimates, the dotted lines are approximate 90% pointwise confidence bands.

21. 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

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

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.

(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 $J(s) = I\{Y(s) \ge 1\}$. 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

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

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

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

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 $\operatorname{Var} V(t) = v(t)$.

(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

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

for each t. In particular, show from this that

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

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

22. 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?

23. 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 $S(t) = \Pr\{T \ge t\}$ 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

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

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

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

(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,

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

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

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

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]

24. 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 $A \,|\, \text{data}$ 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

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

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

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

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

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

For the Ancient Egypt data, again, plot the band

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

for men and for women, and comment.

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

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

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.

(e) [xx i include a figure here, with simulated realisations of A and of S, given data, perhaps Old Egypt. xx]

25. Wald ratios

It is useful to learn about the basic machinery involved in what is often called ‘Wald ratios’ or ‘Wald tests’ (after Abraham Wald, decision theorist, co-inventor of sequential testing, minimaxologist, admissibilist, etc.; born 1902 in Transylvania, died 1950 in a plane-crash in India). Suppose $\widehat{\theta}$ is an estimator of a parameter θ. The basic idea is then to work with

$$W = \frac{\widehat{\theta}-\theta}{\widehat{\tau}},$$

with $\widehat{\tau}$ an estimate of the standard deviation of $\widehat{\theta}$, sometimes called ‘the standard error’, even though the ‘the’ in question is problematic, since there might be several ways in which to arrive at the $\widehat{\tau}$ estimate for the underlying standard deviation τ. The slightly more pedantic label would or could be ‘a standard error for $\widehat{\theta}$’, but this is not often used.

(a) With n the sample size in question, suppose that $\sqrt{n}(\widehat\theta-\theta)\to_d \mathrm{N}(0,\kappa^2)$ (this happens very often, across a wide range of statistical situations). Show that $\sqrt{n}(\widehat\theta-\theta)/\kappa\to_d \mathrm{N}(0,1)$. In the notation above, $\kappa/\sqrt{n}$ is then an approximation to the standard deviation $\tau=\tau_n$ of $\widehat\theta$.

(b) Assume in addition that $\widehat\kappa$ is a consistent estimator of κ. Show that $\widehat\tau_n=\widehat\kappa/\sqrt{n}$ has the property that $\widehat\tau_n/\tau_n\to_{\rm pr} 1$, and that this implies

$$W_n=\frac{\widehat\theta-\theta}{\widehat\tau_n}=\frac{\sqrt{n}(\widehat\theta-\theta)}{\widehat\kappa}\frac{\kappa}{\widehat\kappa}\to_d \mathrm{N}(0,1).$$

(c) Consider the confidence interval

$$\mathrm{CI}_n=\widehat\theta\pm 1.96\,\widehat\tau_n=[\widehat\theta-1.96\,\widehat\kappa/\sqrt{n},\widehat\theta+1.96\,\widehat\kappa/\sqrt{n}].$$

Show that

$$\Pr\{\theta\in\mathrm{CI}_n\}\to 0.95,$$

and give a clear interpretation of the use of this statement.

(d) So Wald ratios lead to confidence intervals – and also to tests. Suppose you need to test the null hypothesis H_0 that $\theta=0$ (or any other fixed null value). Consider the test statistic

$$Z_n=\frac{\widehat\theta}{\widehat\tau_n}=\frac{\sqrt{n}\widehat\theta}{\widehat\kappa}.$$

If you decide to reject H_0 if and only if $|Z_n|\ge 2.576$, what is the (approximate, or limiting) significance level of your test?

(e) The Wald ratio test used above may also be used to read off a p-value:

$$p=\Pr\{|Z_n|\ge|Z_{n,\rm obs}|\,|\,H_0\}\doteq\Pr\{|\mathrm{N}(0,1)|\ge|Z_{n,\rm obs}|\},$$

where $Z_{n,\rm obs}$ is the actually observed value of the Z_n statistic. Note that the defining probability is computed under the null hypothesis. Show that the test which rejects H_0 if and only if $p\le 0.01$ is equivalent to the test above.

(f) There's a classic correspondence between confidence intervals, on one hand, and tests (well, two-sided tests, primarily), on the other. (i) First, suppose you have constructed a clever 95% confidence interval for θ, say CI_n. Then decide to reject $H_0\colon\theta=\theta_0$ by accepting it if θ_0 is inside, and rejecting it if θ_0 is outside. Show that this test has level 0.05. (ii) Second, turn the mirror: Suppose you have first constructed a 0.05 test for each $\theta=\theta_0$ null hypothesis. Then collect all the accepted ones together in a set, say C_n. Show that $\Pr_\theta\{\theta\in C_n\}=0.95$, for each θ. – Most often, this C_n is an interval, but in some cases it might be e.g. a union of three subintervals. We may still call it a confidence set, or region.

(g) From these generalities we may turn back to STK 4080-9080 matters. The points above may be broadly applied, in model after model, and are indeed used, in chapter after chapter, in the ABG book. The ingredients behind a successful Wald ratio operation, so to speak, in a situation where θ is a sufficiently interesting parameter, are (i) construction of a good estimator $\widehat\theta$; (ii) showing that $\widehat\theta$ is approximately normal, and approximately unbiased; (iii) finding a good estimator $\widehat\tau$ for the standard deviation τ of $\widehat\theta$; (iv) showing consistency, in the sense of $\widehat\tau/\tau\to_{\rm pr} 1$. Then you're very much in business, and can apply the machinery above – without having to reinvent the gutenbergian printing machine of 1449 each time.

26. Inference for the median (and other quantiles)

Consider survival data $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, with Nelson–Aalen estimator $\widehat{A}$ and Kaplan–Meier estimator $\widehat{S}$. It is useful to have a machinery for inference for the median (or other quantiles).

(a) Suppose we need to test that the median $m = \mathrm{med}(S) = S^{-1}(\frac{1}{2})$ is equal to given m_0. The Wald ratio approach leads to a statistic of the form

$$Z_n = \frac{\widehat{m} - m_0}{\widehat{\tau}},$$

where $\widehat{m} = \widehat{S}^{-1}(\frac{1}{2})$ is the median computed via the Kaplan–Meier, and $\widehat{\tau}$ is a proper estimate of the standard deviation of $\widehat{m}$. First note that the $\widehat{S}$ moves in jumps, so a bit of care is required to define the median; we typically take

$$\widehat{m} = \min\{t\colon \widehat{S}(t) \le \tfrac{1}{2}\}.$$

Secondly, we need to understand the mathematics of the (approximate) variance τ_n^2 of $\widehat{m}$, and after that an estimate $\widehat{\tau}$. This is absolutely possible [xx may return to this in later exercise xx], but it is easier to circumvent the problem via a transformation.

(b) Explain that testing $m = m_0$ is the same as testing $S(m_0) = \frac{1}{2}$, which again is the same as testing $A(m_0) = \log 2$.

(c) But testing $A(m_0) = \log 2$ (or another fixed value) is easy, via a Wald ratio:

$$Z_n(m_0) = \frac{\widehat{A}(m_0) - \log 2}{\widehat{\sigma}(m_0)}.$$

Explain that the test rejecting $m = m_0$ if $|Z_n(m_0)| \ge 1.645$ has (approximate) level 0.10.

(d) Instead of testing some given m_0 value, construct the set C_n of all m_0 for which $Z_n(m_0) \in [-1.645, 1.645]$. Show how this can be done, and explain that this constitues a 90% confidence interval for the unknown median $m = S^{-1}(\frac{1}{2})$.

(e) Generalise the above to the case of a general quantile, say $\mu(q) = S^{-1}(q)$ for a given q inside $(0, 1)$.

(f) Go again back in time, to Ancient Egypt, of Exercise 1 etc. Use the machinery above to test the hypothesis that the median for the men's distribution is equal to 25.0 years, and then the same task for the women's distribution. Then find 90% confidence intervals for these two medians. (This particular dataset is simpler than in the general case, in that there is no censoring, with all $\delta_i = 1$; use the general machinery from this exercise, though.)

27. Confidence bands for the cumulative hazard and survival functions

Consider again survival data of the familiar form $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, leading to the Nelson–Aalen estimator $\widehat{A}$ and Nelson–Aalen estimator $\widehat{S}$. These are nonparametric estimators of the cumulative hazard rate $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ and survival curve $S(t)$. Here I discuss how to construct *confidence bands* for these two crucial quantities.

(a) For notation and some properties of the $\widehat{A}$, see Exercise 20. We have seen there that

$$\widehat{A}(t) \approx_d \mathrm{N}(A(t), \widehat{\sigma}(t)),$$

in the precise sense that

$$Z_n(t) = \frac{\widehat{A}(t) - A(t)}{\widehat{\sigma}(t)} \to_d Z(t) \sim \mathrm{N}(0,1).$$

Show that this implies that the band

$$B_n(t) = \widehat{A}(t) \pm 1.96\,\widehat{\sigma}(t) = [\widehat{A}(t) - 1.96\,\widehat{\sigma}(t), \widehat{A}(t) + 1.96\,\widehat{\sigma}(t)]$$

is an approximate 95% pointwise confidence band, in the sense of

$$\Pr\{A(t) \in B_n(t)\} \to 0.95 \quad \text{for each } t.$$

(b) Explain how this can be used to test the null hypothesis $H_0(t)$ that $A(t)$ is equal to some given $A_0(t)$.

(c) Sometimes we need something more, however, namely a somewhat bigger band, say $B_n^*(t)$, which covers the full $A(t)$ across some time interval $[a, b]$, as opposed to merely at a given value t. Such bands may be constructed in different ways. Hjort (1985a, inside a long discussion contribution to the SJS Lecture 1984 by P.K. Andersen and Ø. Borgan, and with various other points) proposed the following. I work with

$$B_n^*(t) = \widehat{A}(t) \pm c\,\widehat{\sigma}(t) = [\widehat{A}(t) - c\,\widehat{\sigma}(t), \widehat{A}(t) + c\,\widehat{\sigma}(t)],$$

and wish to scale c such that it succeeds in being a 95% *simultaneous* confidence band over the time window $[a, b]$. What is needed is then

$$\Pr\{A(t) \in B_n^*(t) \text{ for all } t \in [a, b]\} \to 0.95.$$

Show that this corresponds to

$$\Pr\{M_m \le c\} \to 0.95,$$

where

$$M_n = \max_{t \in [a,b]} \Big| \frac{\widehat{A}(t) - A(t)}{\widehat{\sigma}(t)} \Big|.$$

(d) From earlier results, of Exercise 20 and elsewhere, show that

$$Z_n(t) = \frac{\widehat{A}(t) - A(t)}{\widehat{\sigma}(t)} = \frac{\widehat{A}(t) - A(t)}{\sigma_n(t)} \frac{\sigma_n(t)}{\widehat{\sigma}(t)} \to_d Z(t) = \frac{W(v(t))}{\sqrt{v(t)}},$$

where $W(v(t))$ is a Gaußian martingale with independent increments and variance

$$v(t) = \int_0^t \frac{\alpha(s)\,\mathrm{d}s}{y(s)}.$$

Argue that this implies

$$M_n = \max_{t \in [a,b]} |Z_n(t)| \to_d M = \max_{t \in [a,b]} |Z(t)|.$$

The threshold level c can then, in principle, be read off from the limit distribution, i.e. that of M.

(e) Here we're helped by first showing that

$$M = \max_{t\in[a,b]} \Big|\frac{W(v(t))}{\sqrt{v(t)}}\Big| = \max_{s\in[v(a),v(b)]} \Big|\frac{W(s)}{\sqrt{s}}\Big|.$$

Show that this lead to the following recipe: first, obtain estimates $\widehat{v}(a)$ and $\widehat{v}(b)$, via

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

Then simulate say 10^5 paths of Brownian motion $W(s)$ over the interval

$$[\widehat{v}(a), \widehat{v}(b)] = [n\widehat{\sigma}(a)^2, n\widehat{\sigma}(b)^2],$$

and record for each run the value of M, the maximal value of $W(s)/\sqrt{s}$ over that interval. You've now simulated the M distribution, and can read off c as the 0.95 quantile. Congratulations, you've now found the c which works for problem (c). [xx we'll do this in an exercise later on, with a given dataset. xx]

(f) For a Brownian motion $W = \{W(s)\colon s \ge 0\}$, consider the rescaled and time-transformed process W^*, with $W^*(s) = W(cs)/\sqrt{c}$, where c is a positive constant. Take a cup of tea (as the preeminent botanist RObert Brown did in 1827, indirectly pointing to a problem which was not mathematically solved until Einstein wrote a paper on this in 1905). Show that W^* is another Brownian motion. Show also that the variable

$$M = M_{a,b} = \max_{a\le s\le b} |W(s)/\sqrt{s}|$$

above has the property that $M_{a,b}$ has the same distribution as $M_{ca,cb}$, for any positive scaling factor c. So $M_{1.13,2.23}$ has the same distribution as $M_{113,223}$, etc.

(g) Explain how you can use such a plot to test whether $A(t)$ is equal to a given $A_0(t)$ over a given time interval.

(h) Explain finally how you can translate pointwise and simultaneous confidence bands for $A(t)$ to such bands for the survival curve $S(t) = \exp\{-A(t)\}$.

28. The pornoscope data

Consider the pornoscope data of ABG's Example 3.6, given in their Table 3.1. The Drosophila are the Lords of the Flies (but don't read the 1954 book who won its author the Nobel in 1983; it's bleaker than Bleak House). Compute and display the Nelson–Aalen estimators for the time until mating, along with 90% confidence bands. Construct first the relatively easy 90% pointwise confidence bands, and then try to follow the recipe of Exercise 26 to construct also 90% simultaneous confidence bands, valid for the time window [10 minutes, 40 minutes] (yes, these creatures jump to sex pretty quickly).

[xx more here, also on 'survival', which here means 'no sex please'. add figures: the two $\widehat{A}(t)$, with bands, and the running test of $A_1(t)-A_2(t)$. and $S_1(t)$ and $S_2(t)$ to illustrate that the median time to sex is high. xx]

29. Australian drug addicts in clinics

Access the dataset `heroin2-data` and how it downloaded to your computer. It consists of data $(t, \delta, x_1, x_2, x_3)$ for $n = 238$ Australian drug users, with t the time spent in clinic (where I've converted the original scale of days to years); δ the usual indicator for non-censoring; x_1 the daily methadone dosage (here converted to a scale from zero to about 1.1); x_2 equal to 1 or 0 depending on whether the person has been to jail or not; and x_3 equal to 1 or 2 reflecting the person is in clinic 1 or clinic 2.

The Cox regression model is that of proportional hazards, with hazard rates

$$\alpha_i(s) = \alpha_0(s)\exp(x_{i,1}\beta_1 + x_{i,2}\beta_2 + x_{i,3}\beta_3) \quad \text{for } i = 1, \ldots, n,$$

for the at the outset random time t_i a user will stay in the clinic.

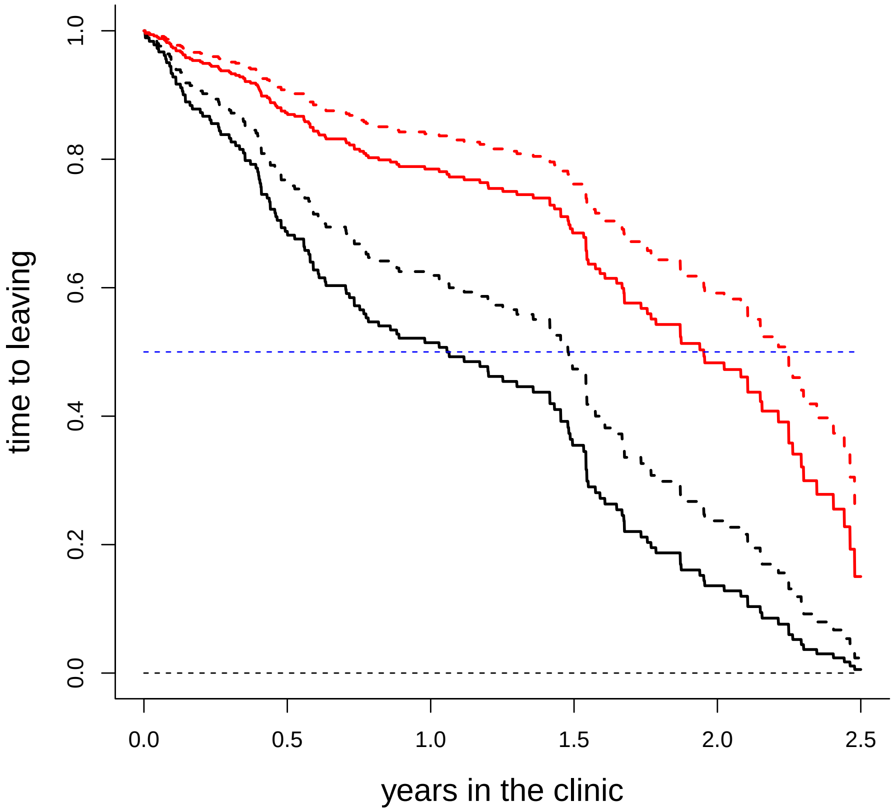

Figure 0.4: Estimated curves for staying yet longer in the clinic, for four types of users, all with methadone dosage $x_1 = 0.60$ (which is close to the average value). The two black curves are for clinic 1, the two red curves are for clinic 2; also, the full curves are for users who have been to prison, the slanted curves for users who have not been to prison.

(a) Carry out basic Cox regression analysis for the β_j coefficients, via something along these lines:

```
library("survival")
heroin <- matrix(scan("heroin2-data",skip=7),byrow=T,ncol=6)
```

```
#
tt <- heroin[ ,2]    # time, in years
delta <- heroin[ ,3] # 63% observed, 37% censored
x1 <- heroin[ ,4]    # methadone dose
x2 <- heroin[ ,5]    # prison or not
x3 <- heroin[ ,6]    # clinic
nn <- length(tt)     # 238
# here we go:
showme <- coxph(Surv(tt,delta) ~ x1 + x2 + x3)
summary(showme)
```

Give an interpretation of what the `coxph` machinery delivers here. Which coefficients are significantly present, and what does it mean, for Australian drug users?

(b) Try to duplicate the essence of this table, by programming the log-partial-likelihood function yourself,

$$\ell_n(\beta) = \sum_{i=1}^n \{x_i^{\rm t}\beta - \log S_n(t_i,\beta)\}\,\delta_i = \sum_{i=1}^n \int_0^\tau \{x_i^{\rm t}\beta - \log S_n(s,\beta)\}\,{\rm d}N_i(s),$$

with

$$S_n(s,\beta) = \sum_{\text{risk set}} \exp(x_j^{\rm t}\beta) = \sum_{j=1}^n Y_j(s)\exp(x_j^{\rm t}\beta).$$

When having made such a `logPL` programme, check that it works, by asking for something like `logPL(c(-2.22,0.44,-1.11))`. Then throw it to an optimiser, to find *both* the Cox estimators $\widehat\beta_j$ *and* their standard errors (estimated standard deviations):

```
nils <- nlm(minuslogPL,c(0,0,0),hessian=T)
cox <- nils$estimate
Jhat <- nils$hessian
se <- sqrt(diag(solve(Jhat)))
show <- cbind(cox,se,cox/se)
print(round(show,4))
```

That the standard errors actually can be computed (rather simply) in this fashion has to do with the large-sample approximation

$$\operatorname{Var}\widehat\beta \doteq \widehat J^{-1},$$

with $\widehat J = -\partial^2\ell_n(\widehat\theta)/\partial\beta\partial\beta^{\rm t}$ the Hessian matrix of second order derivatives at the maximising point (the anglophied world is winning, apparently, regarding this term; it really ought to be 'Hesse matrix', just ask Ludwig Otto Hesse, 1811–1874, but then people won't understand you). See ABG Section 2.1 for this, along with a few Nils Exercises below.

(c) In addition to issues essentially related to the β_j, one needs to estimate and assess also cumulative hazard rates $A(t\,|\,x) = A_0(t)\exp(x_i^{\rm t}\beta)$ and survival curves

$$S(t\,|\,x) = \Pr\{T \ge t\,|\,x\} = \exp\{-A(t\,|\,x)\} = \exp\{-A_0(t)r(x)\},$$

with $r(x) = \exp(x^{\rm t}\beta)$. These tasks involve estimation of the cumulative baseline hazard function $A_0(t) = \int_0^t \alpha_0(s)\,{\rm d}s$, for which we use the so-called Breslow–Aalen estimator,

$$\widehat{A}_0(t) = \int_0^t \frac{\sum_{i=1}^n {\rm d}N_i(s)}{\sum_{i=1}^n Y_i(s)\exp(x_i^{\rm t}\widehat{\beta})}.$$

Give a motivation for this $\widehat{A}_0$, and compute it.

(d) Choose a given type of Australian drug user, with specific covariates $x_0 = (x_{1,0}, x_{2,0}, x_{3,0})$, of your choice. Compute and plot and behold the estimated cumulative hazard rate $\widehat{A}(t\,|\,x_0)$ and survival function $\widehat{S}(t\,|\,x_0)$. Explain how the survival function ought to be interpreted here.

(e) Try to duplicate a version of Figure 0.4.

(f) Check for possible interaction effects, for methadone dose and clinic.

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

This exercise is meant to be a brief going-through of what happens with maximum likelihood (ML) estimators in the classical terrain of regular models without censoring. There are basically four wondrously important and now easy-to-use results about such estimators, say $\widehat{\theta}$ for an underlying parameter vector $\theta = (\theta_1, \ldots, \theta_p)$:

(i) the distribution of $\widehat{\theta}$ is approximately multinormal;

(ii) the ML is approximately unbiased;

(iii) its variance matrix can be estimated (relatively easily) as part of the process;

(iv) results (i)-(ii)-(iii) carry over to any focus parameter, say using $\widehat{\phi} = \phi(\widehat{\theta})$ for estimating $\phi = \phi(\theta)$, via the delta method.

These methods and results (with variations and extensions, e.g. to regression models) are in constant use, also in a long list of R and software packages (e.g. for standard errors and Wald ratios and p-values for all generalised linear models, etc.).

The importance and supreme usefulness of these results relate also to the fact that they hold for *any* parametric model, all the usual ones plus those you might feel a need to invent tomorrow.

Going through the points below is meant to give you ‘the basics’ about these points, and why they hold. The point will then be that a good understanding of these issues and methods, for the easier classical terrain, will help you when it comes to extensions, generalisations, modifications, for the partly more complicated world of models and estimators in this course, e.g. ML in parametric survival analysis models (ABG Ch. 5) and Cox estimators in the Cox regression model (ABG Ch. 4).

Consider now data points $y_1, \ldots, y_n$, i.i.d. from a given parametric model, with density $f(y, \theta)$, smooth in its parameters. We use θ_0 to indicate the true parameter value. The log-likelihood function is

$$\ell_n(\theta) = \sum_{i=1}^n \log f(y_i, \theta),$$

since there is no censoring or other complications. We shall need for the 1st and 2nd derivatives,

$$\begin{aligned} U_n(\theta) &= \sum_{i=1}^n \frac{\partial \log f(y_i,\theta)}{\partial\theta} = \sum_{i=1}^n u(y_i,\theta), \\ I_n(\theta) &= \sum_{i=1}^n \frac{\partial^2 \log f(y_i,\theta)}{\partial\theta\partial\theta^{\mathrm{t}}} = \sum_{i=1}^n u(y_i,\theta). \end{aligned}$$

Note that $U_n(\theta)$ is a p-dimensional vector function and that $I_n(\theta)$ is a $p\times p$-matrix function.

(a) Using the Central Limit Theorem, show that

$$(1/\sqrt{n})U_n(\theta_0) \to_d U \sim \mathrm{N}_p(0,J_0),$$

with

$$J_0 = \mathrm{Var}\, u(y,\theta_0) = -\mathrm{E}\,\frac{\partial^2 \log f(y,\theta_0)}{\partial\theta\partial\theta^{\mathrm{t}}}$$

the so-called Fisher information matrix, computed at the true parameter value. That these two expressions defining J_0 are equal is part of the game, and is called the Bartlett identity (show it, that part is not hard). The $u(y,\theta)$ is called the score function of the model, and must assume here that it has a finite variance matrix.

(b) Using the Law of Large Numbers, show that

$$J_n = -(1/n)I_n(\theta_0) \to_{\mathrm{pr}} J_0.$$

(c) From the defining properties of the ML, argue via Taylor that

$$0 = U_n(\widehat\theta) = U_n(\theta_0) + I_n(\theta_0)(\widehat\theta - \theta_0) + \delta_n,$$

where δ_n ought to be small in size.

(d) From this, deduce that

$$\sqrt{n}(\widehat\theta - \theta_0) = J_n^{-1}U_n(\theta_0) + \delta_n',$$

where δ_n' is small. Argue that

$$\sqrt{n}(\widehat\theta - \theta_0) \to_d J_0^{-1}U \sim J_0^{-1}\mathrm{N}_p(0,J_0) \sim \mathrm{N}_p(0,J_0^{-1}).$$

This is the 'mathematical version' of the basic result about ML estimation in regular models.

(e) By carrying $\sqrt{n}$ over, we reach the 'practical version' of the same result, namely that

$$\widehat\theta \approx_d \mathrm{N}_p(\theta_0,\widehat\Sigma),$$

with estimated variance matrix

$$\widehat\Sigma = (1/n)\widehat J_0^{-1} = \widehat J_{\mathrm{obs}}^{-1},$$

writing now

$$\widehat J_{\mathrm{obs}} = -\frac{\partial^2 \ell_n(\widehat\theta)}{\partial\theta\partial\theta^{\mathrm{t}}} = n\widehat J_0$$

for the so-called Fisher's observed information matrix.

(f) Note that $\widehat{J}_{\rm obs}$ increases with sample size, since $\widehat{J}_{\rm obs}/n$ tends to J_0. Explain that when a dataset is doubled, from size n to $2n$, the lengths of the confidence intervals go down with a factor of around $\sqrt{2}$. How many more data points would you need, compared to the n you already have access to, if you need your confidence intervals do have (approximately) half-as-long widths?

(g) A basic fact about the multinormal distribution is that if $X \sim {\rm N}_p(\xi, \Sigma)$, then the quadratic form $(X-\xi)^{\rm t}\Sigma^{-1}(X-\xi)$ has a χ^2_p distribution. Argue from this that

$$Q_n = n(\widehat\theta-\theta_0)^{\rm t}\widehat{J}_0(\widehat\theta-\theta_0) = (\widehat\theta-\theta_0)^{\rm t}\widehat{J}_{\rm obs}(\widehat\theta-\theta_0) \to_d \chi^2_p.$$

Explain how this may be used to test $H_0\colon \theta = \theta_{\rm fix}$, for any given $\theta_{\rm fix}$, and also how a 90% confidence region can be constructed for the unknown θ_0.

(h) Then the extension of the above results to focus parameters: Suppose a parameter $\phi = \phi(\theta) = \phi(\theta_1,\ldots,\theta_p)$ is of importance, a smooth function of the model parameters. The true parameter value is $\phi_0 = \phi(\theta_0)$. Show that the ML of ϕ is simply $\widehat\phi = \phi(\widehat\theta)$. Show also that $\widehat\phi$ is approximately normal, approximately unbiased, and with an easy estimate for its standard deviation, via the delta method of Exercise 17. Specifically,

$$\sqrt{n}(\widehat\phi-\phi_0) \to_d {\rm N}(0,\kappa^2), \quad \text{with} \quad \kappa_0^2 = c^{\rm t}J_0^{-1}c,$$

where $c = \partial\phi(\theta_0)/\partial\theta$. The practical translation of this statement, directly useful for a long list of applications, is that

$$\widehat\phi \approx_d {\rm N}(\phi_0, \widehat\kappa^2),$$

with

$$\widehat\kappa^2 = (1/n)\widehat\kappa_0^2 = (1/n)\widehat{c}^{\rm t}\widehat{J}_0^{-1}\widehat{c} = \widehat{c}^{\rm t}\widehat{J}_{\rm obs}^{-1}\widehat{c},$$

and $\widehat{c} = \partial\phi(\widehat\theta)/\partial\theta$ the gradient vector of partial derivatives calculated at the ML position.

(i) [xx briefly: outside model conditions, the sandwich matrix, etc. xx]

(j) I list as a separate point the following couple of comments. First, what is shown in the course of the points above usually takes *much more splace and time* in regular textbooks (with lots of details having to do with regularity conditions implying remainder terms tending to zero etc.); I think Lehmann's classic 1983 book needs some fifteen pages to go reach the most crucial result (d), for example. So my proof, which is really 'a good indication of a proof', is considerably shorter. For more details, also regarding the important extensions of the above to regression models (as for Poisson and logistic regression, etc.), see Claeskens and Hjort (2008, Ch. 2) or Schweder and Hjort (2016, Appendix). Clean, concise proofs are available, with all details taken proper care of, in the case of log-concave densities; see Hjort and Pollard (1993).

31. The likelihood theory works also for censored data

Exercise 30 deals with the classic terrain of independent and fully observed data points, and establishes the basic results for maximum likelihood (ML) and a few related quantities and tools. This has been a veritable success story, in probability theory and statistics, since about 1922, when Sir Ronald A. Fisher invented the ML, to the present. It was not so easy to lift these machines to the world of survival analysis models, however, since the censoring business makes the models, the estimators, and their behaviour, more complicated. Among the first papers dealing with these issues, sorting out limiting normality under good conditions, etc., are Borgan (1984), then Hjort (1985a, 1986, 1992); cf. also the broad treatment in ABG (Ch. 5).

Consider survival data of the usual form (t_i, δ_i) for individuals $i = 1, \ldots, n$, and assume that the hazard rate is of a suitable parametric form $\alpha(s, \theta)$. Examples include the exponential, the Weibull, the gamma, the Gompertz, the log-normal, etc. Below, let θ_0 denote the true parameter value.

(a) Check that you are able to deduce several (equivalent) expressions for the log-likelihood function, met also in several earlier exercises:

$$\begin{aligned}\ell_n(\theta) &= \sum_{\delta_i=1} \log f(t_i, \theta) + \sum_{\delta_i=0} \log S(t_i, \theta) \\ &= \sum_{i=1}^n \{\delta_i \log \alpha(t_i, \theta) - A(t_i, \theta)\} \\ &= \int_0^\tau \{\log \alpha(s, \theta)\, \mathrm{d}N(s) - Y(s)\alpha(s, \theta)\, \mathrm{d}s\}.\end{aligned}$$

Here τ is an upper bound for the time window under consideration.

(b) We attempt to follow the path of Exercise 30, and start with the first derivative. Consider therefore the random function

$$U_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} = \int_0^\tau \psi(s, \theta)\, \{\mathrm{d}N(s) - Y(s)\alpha(s, \theta)\, \mathrm{d}s\},$$

with $\psi(s, \theta) = \partial \log \alpha(s, \theta)/\partial \theta$.

(c) At the true parameter value, show that $U_n(\theta_0) = \int_0^\tau \psi(s, \theta_0)\, \mathrm{d}M(s)$, involving the martingale $M(t) = N(t) - \int_0^t Y(s)\alpha(s, \theta_0)\, \mathrm{d}s$. Hence $U_n(\theta_0)$ is very conveniently a martingale (perhaps Borgan, 1984, was the first to realise this key point, in decent generality), evaluted at the end-point τ.

(d) Show that

$$\begin{aligned}\langle (1/\sqrt{n})U_n(\theta_0), (1/\sqrt{n})U_n(\theta_0)\rangle &= (1/n) \int_0^\tau \psi(s, \theta_0)\psi(s, \theta_0)^{\mathrm{t}}\, \mathrm{d}\langle M, M\rangle(s) \\ &= (1/n) \int_0^\tau \psi(s, \theta_0)\psi(s, \theta_0)^{\mathrm{t}} Y(s)\alpha(s, \theta_0)\, \mathrm{d}s,\end{aligned}$$

and that this converges in probability to the matrix

$$J_0 = \int_0^\tau \psi(s, \theta_0)\psi(s, \theta_0)^{\mathrm{t}} y(s)\alpha(s, \theta_0)\, \mathrm{d}s.$$

It is assumed that $Y(s)/n$ tends uniformly in probability to a limit function $y(s)$. Conclude from martingale limit theorems, cf. previous exercises, that

$$(1/\sqrt{n})U_n(\theta_0) \to_d U \sim \mathrm{N}_p(0, J_0).$$

So we're in good shape, with the required result for the first derivative random function, though it tooks more work and more advanced limit theory than for the i.i.d. case of Exercise 30.

(e) Next up is the second derivative $p \times p$ matrix function. Show that

$$\begin{aligned} I_n(\theta) &= \frac{\partial^2 \ell_n(\theta)}{\partial\theta\partial\theta^{\rm t}} \\ &= \int_0^\tau \big[\psi^*(s,\theta)\{{\rm d}N(s) - Y(s)\alpha(s,\theta)\,{\rm d}s\} - \psi(s,\theta)\psi(s,\theta)^{\rm t} Y(s)\alpha(s,\theta)\,{\rm d}s\big], \end{aligned}$$

where $\psi^*(s,\theta) = \partial^2 \log\alpha(s,\theta)/\partial\theta\partial\theta^{\rm t}$.

(e) Try to show, whether this means 'prove' or 'make very plausible', that

$$-(1/n)I_n(\theta_0) \to_{\rm pr} J_0.$$

(f) So we're in good shape, and have essentially been able to modify and extend the simpler arguments of Exercise 30 for the present more complicated world of survival models, partly thanks to the martingale theory. Conclude that for the ML estimator $\widehat\theta$, we have

$$\sqrt{n}(\widehat\theta - \theta_0) \to_d J_0^{-1}U \sim {\rm N}_p(0, J_0^{-1}).$$

This is the clear extension of the classical ML result for i.i.d. data to the world of survival models, involving now a new definition of the J_0 matrix.

(g) For the special case of no censoring, and with time window the full half-line, corresponding to $\tau = \infty$, show indeed that J_0 becomes equal to the Fisher information matrix of Exercise 30.

(h) We need a good estimator of J_0. There are several options, actually. The perhaps simplest to work with, since it comes out of computer optimisation programmes, as the Hessian matrix, is

$$\widehat J_0 = -\frac{1}{n}\frac{\partial^2(\widehat\theta)}{\partial\theta\partial\theta^{\rm t}}.$$

Try to prove that $\widehat J_0 \to_{\rm pr} J_0$, i.e. that it is consistent. Another estimator, which also works, based on the form of J_0, is

$$\widetilde J_0 = (1/n)\int_0^\tau Y(s)\psi(s,\widehat\theta)\psi(s,\widehat\theta)^{\rm t}\alpha(s,\widehat\theta)\,{\rm d}s.$$

(h) For a simple illustration, consider the exponential model with a constant hazard rate, $\alpha(s,\theta) = \theta$. Find an expression for the ML estimator $\widehat\theta$, and check each step of the arguments above. Put up the limit distribution for $\sqrt{n}(\widehat\theta - \theta_0)$. How much is lost in precision, if the censoring distribution is another exponential, with rate γ?

32. Parametric hazard rate regression models

Importantly, the full machinery above extends nicely to the case of regression models, and this is 76% of ABG's Chapter 5, regarding both the basic methods and the ensuing type of results. The setting is that of survival data (t_i, δ_i, x_i), with a covariate vector x_i of length p. We start from

$$\alpha_i(s) = \alpha(s,\theta)\exp(x_i^{\rm t}\beta),$$

for $i = 1,\ldots,n$. Show that the log-likelihood function may be expressed as

$$\ell_n(\theta,\beta) = \sum_{i=1}^n \int_0^\tau \big[\{\log\alpha(s,\theta) + x_i^{\rm t}\beta\}\,{\rm d}N_i(s) - Y_i(s)\alpha(s,\theta)\exp(x_i^{\rm t}\beta)\,{\rm d}s\big].$$

Now go back Down Under, to the $n = 238$ Australian drug users in the two clinics, cf. Exercise 29. There we applied Cox's semiparametric model, with an unspecified baseline hazard function. Now try the parametric machinery, with two models: (i) that of the constant baseline hazard, $\alpha_0(s) = \theta$; and (ii) that of a Weibull type baseline hazard, $\alpha_0(s) = \theta \exp(\gamma s)$. Plot the Breslow–Aalen estimator $\widehat{A}_0(t)$ alongside the parametrically fitted versions, and comment on your findings.

[xx a bit more her. confidence interval for γ. xx]

33. [xx more on this, for the Cox regression model. xx]

[xx well: to be written, very soon. for survival regression data (t_i, δ_i, x_i), Cox's log-partial-likelihood,

$$\ell_n(\beta) = \sum_{i=1}^n \{x_i^{\rm t}\beta - \log S_n(t_i,\beta)\}\,\delta_i = \sum_{i=1}^n \int_0^\tau \{x_i^{\rm t}\beta - \log S_n(s,\beta)\}\,{\rm d}N_i(s),$$

with

$$S_n(s,\beta) = \sum_{\rm risk\ set} \exp(x_j^{\rm t}\beta) = \sum_{j=1}^n Y_j(s)\exp(x_j^{\rm t}\beta).$$

programme is: work with 1st and 2nd derivatives, and aim for a parallel story to those of Exercises 30 and 31 (even if the going gets tougher). point to Andersen and Gill (1982) and Hjort and Pollard (1993). xx]

References

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

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.

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.

Cunen, C. and Hjort, N.L. (2018). Survival and event history models and methods via Gamma processes. [Manuscript in progress.]

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. (1980). *Limit Theorems in Probability Theory and Statistics* [well, this 'Hjorts lille grønne' 150-page compendium is in Norwegian; it was used as basic curriculum at the start of the høyere grad statistics studies for almost twenty years, I think].

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. (1992). On inference in parametric survival data models. *International Statistical Review* **xx**, 355–387.

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).

Hjort, N.L. and Pollard, D.B. (1993). Asymptotics for minimisers of convex processes. Statistical Research Report, Department of Mathematics, University of Oslo.

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

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, 1983.

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

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* **14**, 945–966.

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.

Stoltenberg, E.Aa. and Hjort, N.L. (2018). Models and inference for on-off data via clipped Ornstein–Uhlenbeck processes. Submitted for publication.

Stoltenberg, E.Aa. and Hjort, N.L. (2018). Modelling and analysing the Beta- and Gamma Oslo Police Tweetery data. [Manuscript, in progress.]

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 Gjessing, H.K. (2001). Understanding the shape of the hazard rate: a process point of view. *Statistical Science* **16**, 1–22.

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