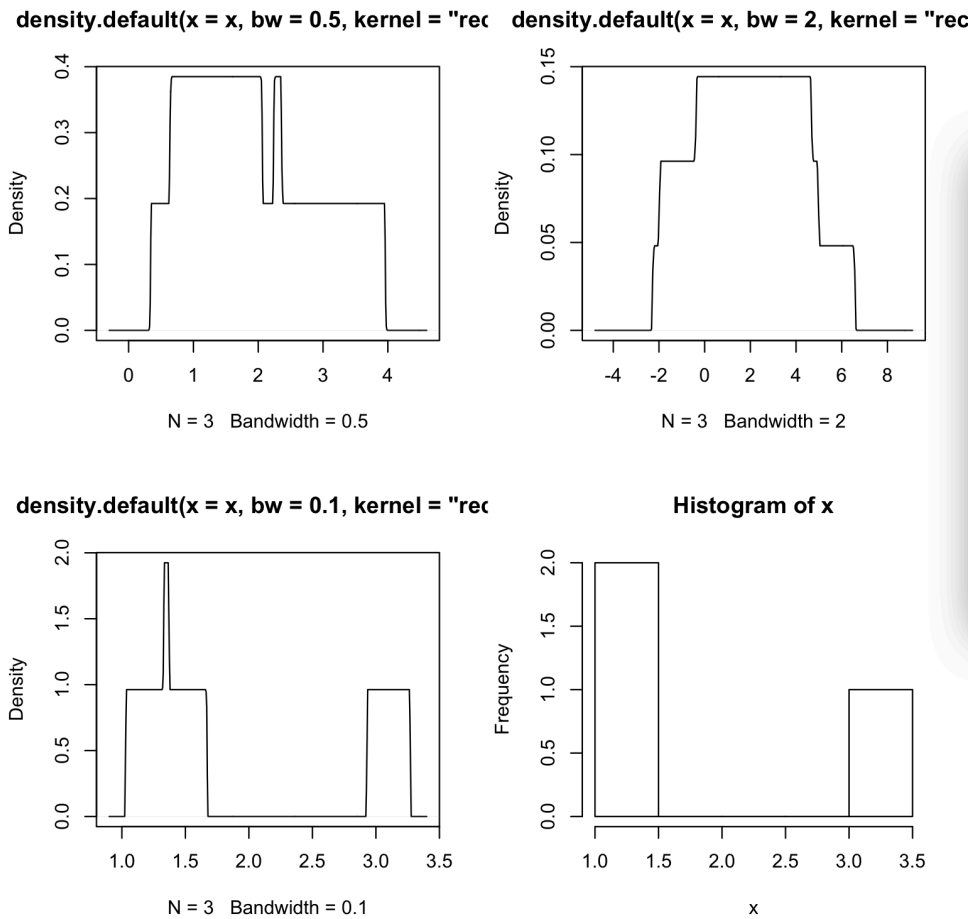
Kernel density estimation in R

Kernel density estimation can be done in R using the `density()` function in R. The default is a Guassian kernel, but others are possible also. It uses it's own algorithm to determine the bin width, but you can override and choose your own. If you rely on the `density()` function, you are limited to the built-in kernels. If you want to try a different one, you have to write the code yourself.

Kernel density estimation in R: effect of bandwidth for rectangular kernel

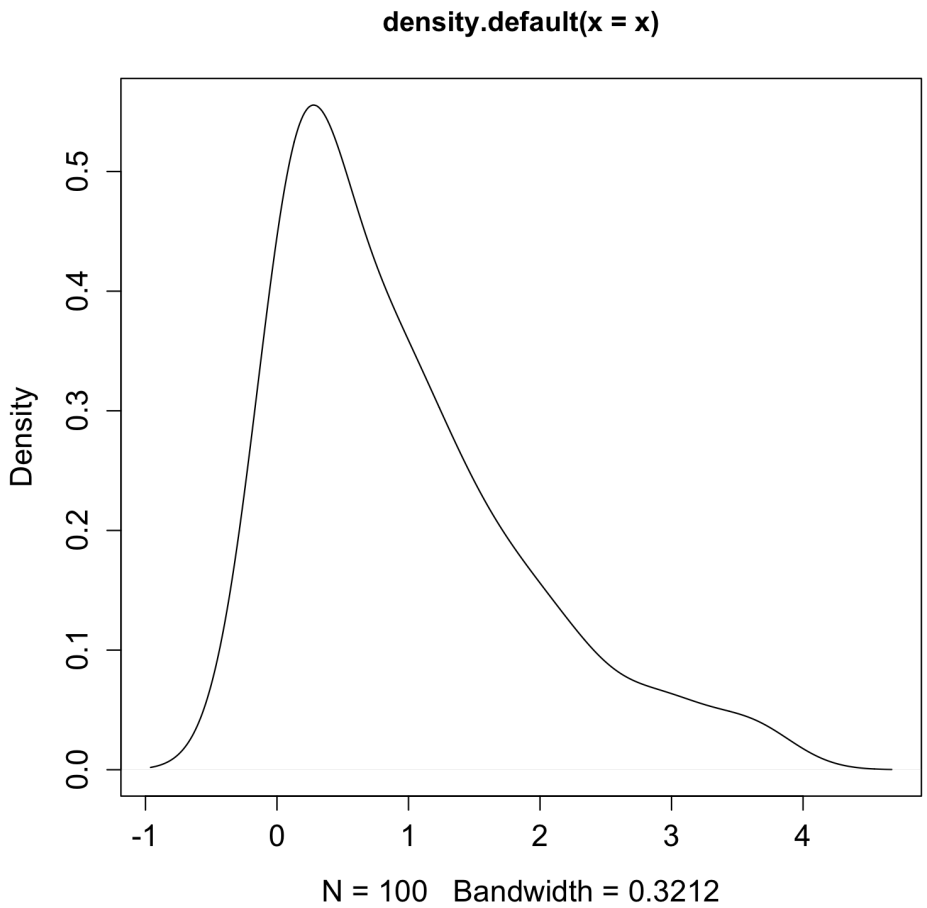


Kernel density estimation in R

Note that exponential densities are a bit tricky to estimate to using kernel methods. Here is the default behavior estimating the density for exponential data.

```
> x <- rexp(100)
> plot(density(x))
```

Kernel density estimation in R: exponential data with Gaussian kernel

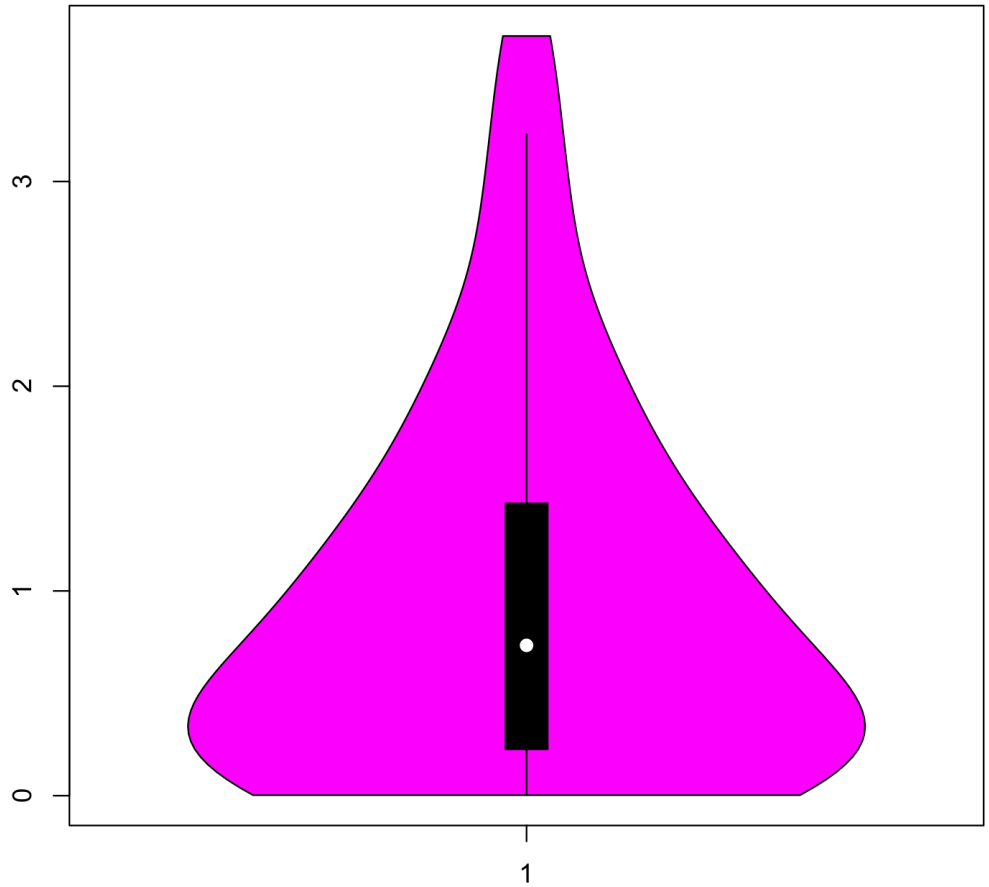


Violin plots: a nice application of kernel density estimation

Violin plots are an alternative to boxplots that show nonparametric density estimates of the distribution in addition to the median and interquartile range. The densities are rotated sideways to have a similar orientation as a box plot.

```
> x <- rexp(100)
> install.packages("vioplot")
> library(vioplot)
> x <- vioplot(x)
```

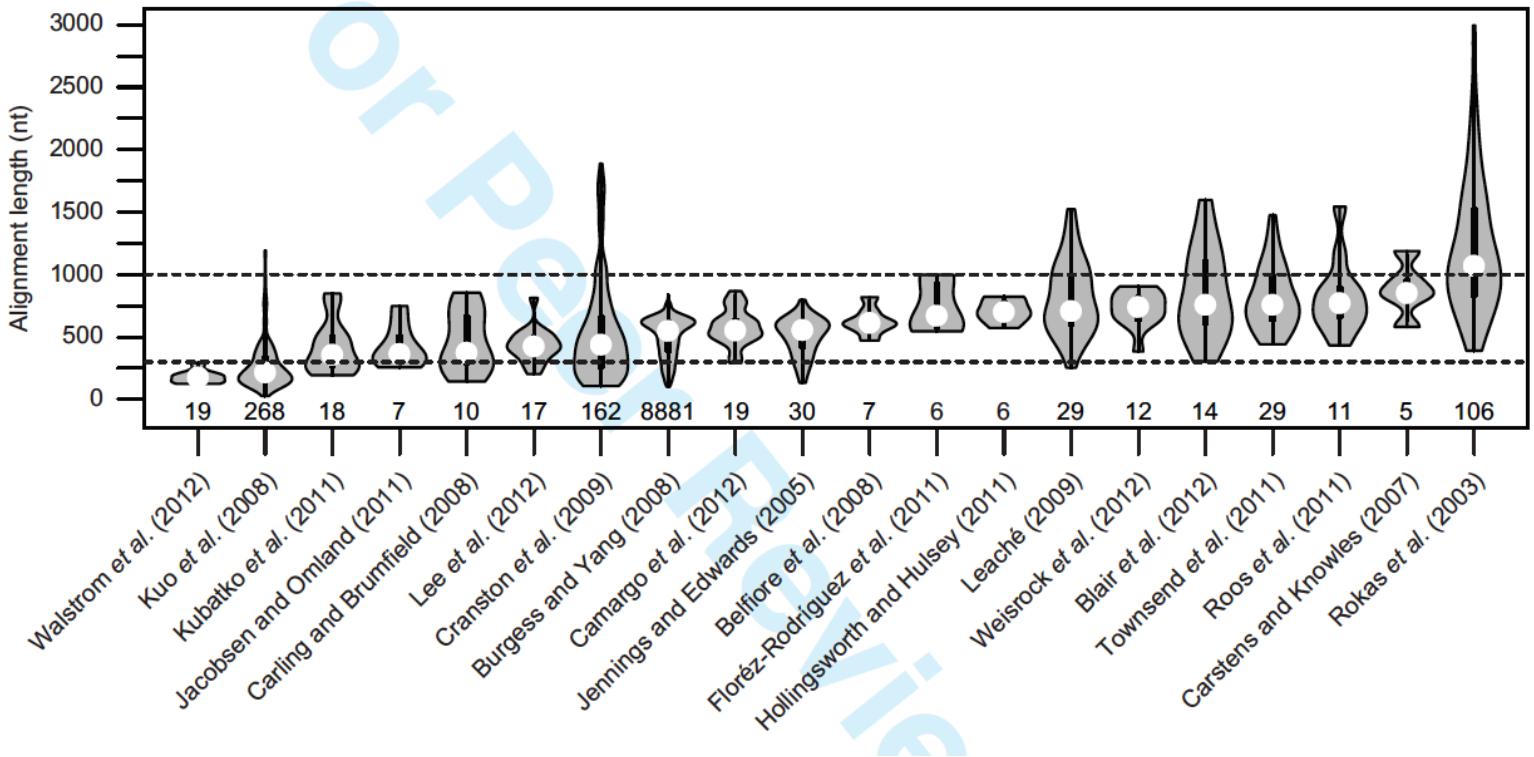
Kernel density estimation in R: violin plot



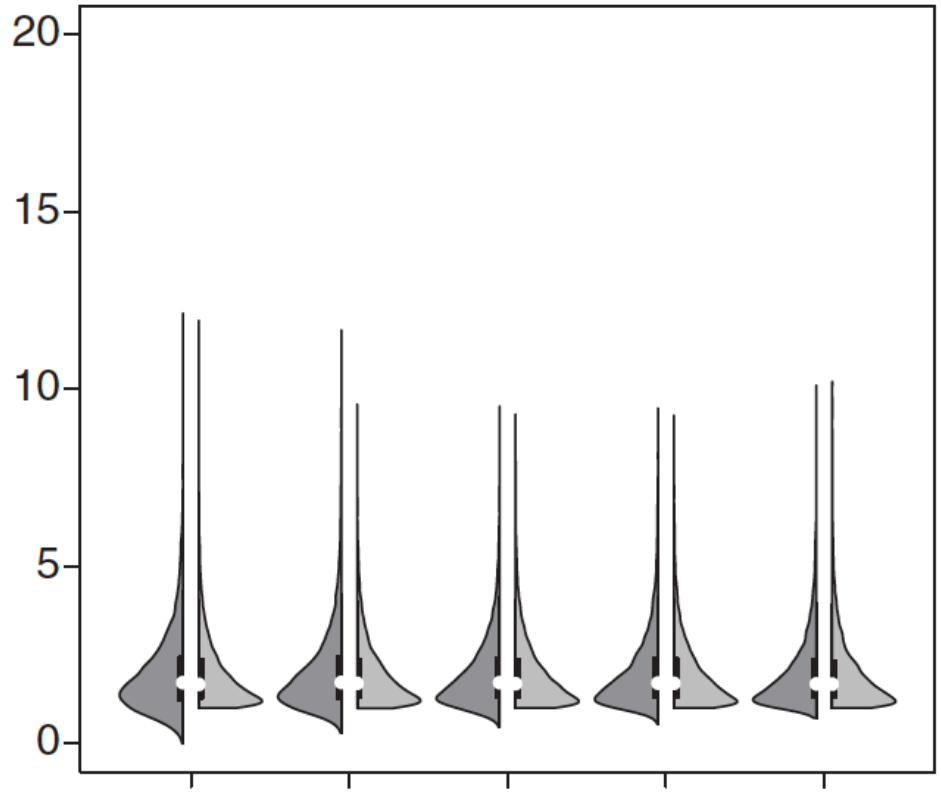
Kernel density estimation R: violin plot

The violin plot uses the function `sm.density()` rather than `density()` for the nonparametric density estimate, and this leads to smoother density estimates. If you want to modify the behavior of the violin plot, you can copy the original code to your own function and change how the nonparametric density estimate is done (e.g., replacing `sm.density` with `density`, or changing the kernel used).

Kernel density estimation in R: violin plot



Kernel density estimation in R: violin plot



Kernel density estimation in R: violin plot

```
> vioplot
function (x, ..., range = 1.5, h = NULL, ylim = NULL, names = NULL,
    horizontal = FALSE, col = "magenta", border = "black", lty = 1,
    lwd = 1, rectCol = "black", colMed = "white", pchMed = 19,
    at, add = FALSE, wex = 1, drawRect = TRUE)
{
    datas <- list(x, ...)
    n <- length(datas)
    if (missing(at))
        at <- 1:n
    upper <- vector(mode = "numeric", length = n)
    lower <- vector(mode = "numeric", length = n)
    q1 <- vector(mode = "numeric", length = n)
    q3 <- vector(mode = "numeric", length = n)
    ...
    args <- list(display = "none")
    if (!(is.null(h)))
        args <- c(args, h = h)
    for (i in 1:n) {
        ...
        smout <- do.call("sm.density", c(list(data, xlim = est.xlim),
            args))
```

Kernel density estimation

There are lots of popular Kernel density estimates, and statisticians have put a lot of work into establishing their properties, showing when some Kernels work better than others (for example, using mean integrated square error as a criterion), determining how to choose bandwidths, and so on.

In addition to the Guassian, common choices for the hazard function include

- Uniform, $K(u) = 1/2\ I(-1 \leq u \leq 1)$
- Epanechnikov, $K(u) = .75(1 - u^2)I(-1 \leq u \leq 1)$
- biweight, $K(u) = \frac{15}{16}(1 - x^2)^2 I(-1 \leq u \leq 1)$

Kernel-smoothed hazard estimation

To estimate a smoothed version of the hazard function using a kernal method, first pick a kernel, then use

$$\widehat{h} = \frac{1}{b} \sum_{i=1}^{D} K\left(\frac{t - t_i}{b}\right) \Delta \widetilde{H}(t_i)$$

where D is the number of death times and b is the babdwidth (instead of h). A common notation for bandwidth is h, but we use b because h is used for the hazard function. Also $\widehat{H}(t)$ is the Nelson-Aalen estimator of the cumulative hazard function:

$$\widetilde{H}(t) = \begin{cases} 0, & \text{if } t \leq t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i}, & \text{if } t > t_1 \end{cases}$$

Kernel-smoothed hazard estimation

The variance of the smoothed hazard is

$$\sigma^2[\widehat{h}(t)] = b^{-2} \sum_{i=1}^{D} \left[K\left(\frac{t - t_i}{b} \right) \right]^2 \Delta \widehat{V}[\widetilde{H}(t)]$$

Asymmetric kernels

A difficulty that we saw with the exponential also can occur here, the estimated hazard can give negative values. Consequently, you can use an asymmetric kernel instead for small t.

For $t < b$, let $q = t/b$. A similar approach can be used for large t, when $t_D - b < t < t_D$. In this case, you can use $q = (t_D - t)/b$ and replace x with $-x$ in the kernal density estimate for these larger times.

Asymmetric kernels

modified kernels, for the uniform kernel (6.2.1), are expressed by

$$K_q(x) = \frac{4(1+q^3)}{(1+q)^4} + \frac{6(1-q)}{(1+q)^3}x, \quad \text{for } -1 \le x \le q, \tag{6.2.6}$$

for the Epanechnikov kernel (6.2.2),

$$K_q(x) = K(x)(\alpha_E + \beta_E x), \quad \text{for } -1 \le x \le q, \tag{6.2.7}$$

where

$$\alpha_E = \frac{64(2-4q+6q^2-3q^3)}{(1+q)^4(19-18q+3q^2)}$$

and

$$\beta_E = \frac{240(1-q)^2}{(1+q)^4(19-18q+3q^2)},$$

Asymmetric kernels

and for the biweight kernel (6.2.3),

$$K_q(x) = K(x)(\alpha_{BW} + \beta_{BW}x), \quad \text{for } -1 \le x \le q, \tag{6.2.8}$$

where

$$\alpha_{BW} = \frac{64(8 - 24q + 48q^2 - 45q^3 + 15q^4)}{(1+q)^5(81 - 168q + 126q^2 - 40q^3 + 5q^4)}$$

and

$$\beta_{BW} = \frac{1120(1-q)^3}{(1+q)^5(81 - 168q + 126q^2 - 40q^3 + 5q^4)}.$$

Confidence intervals

A pointwise confidence interval can be obtained with lower and upper limits

$$\left(\widehat{h}(t)\exp\left[-\frac{Z_{1-\alpha/2}\sigma(\widehat{h}(t))}{\widehat{h}(t)}\right], \widehat{h}(t)\exp\left[\frac{Z_{1-\alpha/2}\sigma(\widehat{h}(t))}{\widehat{h}(t)}\right]\right)$$

Note that the confidence interfal is really a confidence interval for the smoothed hazard function and not a confidence interval for the actual hazard function, making it difficult to interpret. In particular, the confidence interval will depend on the both the kernel and the bandwidth. Coverage probabilities for smoothed hazard estimates (the proportion of times the confidence interval includes the true hazard rate) appears to have ongoing research.

Asymmetric kernels

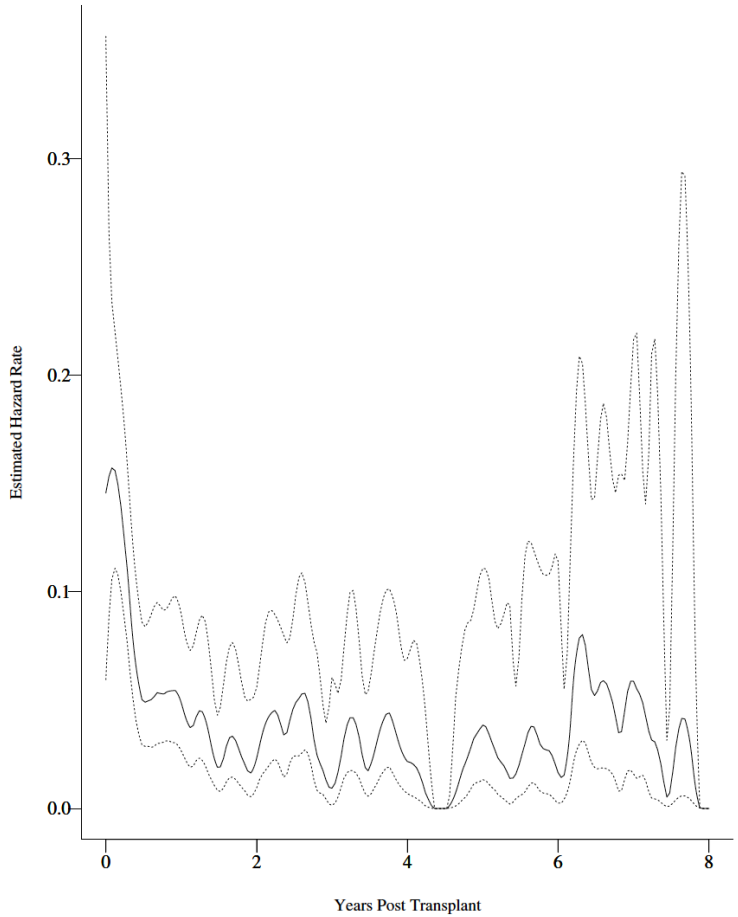


Figure 6.6 *Smoothed estimate of the hazard rate (———) and 95% confidence interval (------) for the time to death following a kidney transplant based on the biweight kernel and the best bandwidth.*

Effect of bandwidth

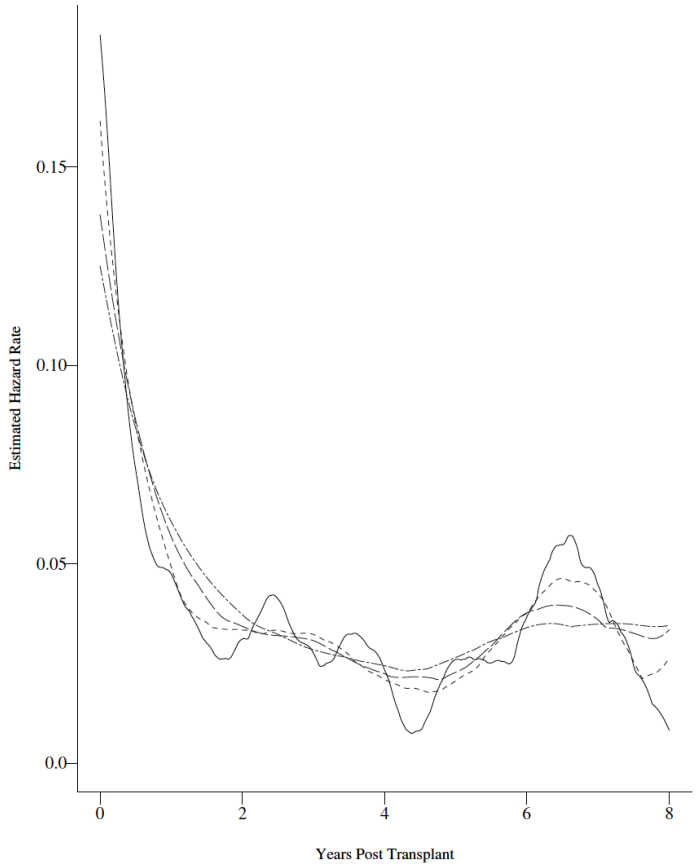


Figure 6.4 *Effects of changing the bandwidth on the smoothed hazard rate estimates for kidney transplant patients using the Epanechnikov kernel. bandwidth = 0.5 years (———) bandwidth = 1.0 years (-------) bandwidth = 1.5 years (— — —) bandwidth = 2.0 years (– · – · –)*

Effect of bandwidth

Because the bandwidth has a big impact, we somehow want to pick the optimal bandwidth. An idea is to minimize the squared area between the true hazard function and estimated hazard function. This squared area between the two functions is called the Mean Integrated Squared Error (MISE):

$$\begin{aligned} MISE(b) &= E \int_{\tau_L}^{\tau_U} [\widehat{h}(u) - h(u)]^2 \, du \\ &= E \int_{\tau_L}^{\tau_U} \widehat{h}^2(u) \, du - 2E \int_{\tau_L}^{\tau_U} \widehat{h}(u) h(u) \, du + f(h(u)) \end{aligned}$$

The last term doesn't depend on b so it is sufficient to minimize the function ignoring the last term. The first term can be estimated by $\int_{\tau_L}^{\tau_U} \widehat{h}^2(u) \, du$, which can be estimated using the trapezoid rule from calculus.

Effect of bandwidth

The second term can be approximated by

$$\frac{1}{b}\sum_{i\neq j} K\left(\frac{t-t_i}{b}\right)\Delta\widetilde{H}(t_i)\Delta\widetilde{H}(t_j)$$

summing over event times between τ_L and τ_U. Minimizing MISE is can be done approximately by minimizing

$$g(b) = \sum_i \left(\frac{u_{i+1}-u_i}{2}\right)[\widehat{h}^2(u_i) - \widehat{h}^2(u_{i+1})] - \frac{2}{b}\sum_{i\neq j} K\left(\frac{t-t_i}{b}\right)\Delta\widetilde{H}(t_i)\Delta\widetilde{H}(t_j)$$

The minimization can be done numerically by plugging in different values of b and evaluating.

Effect of bandwidth

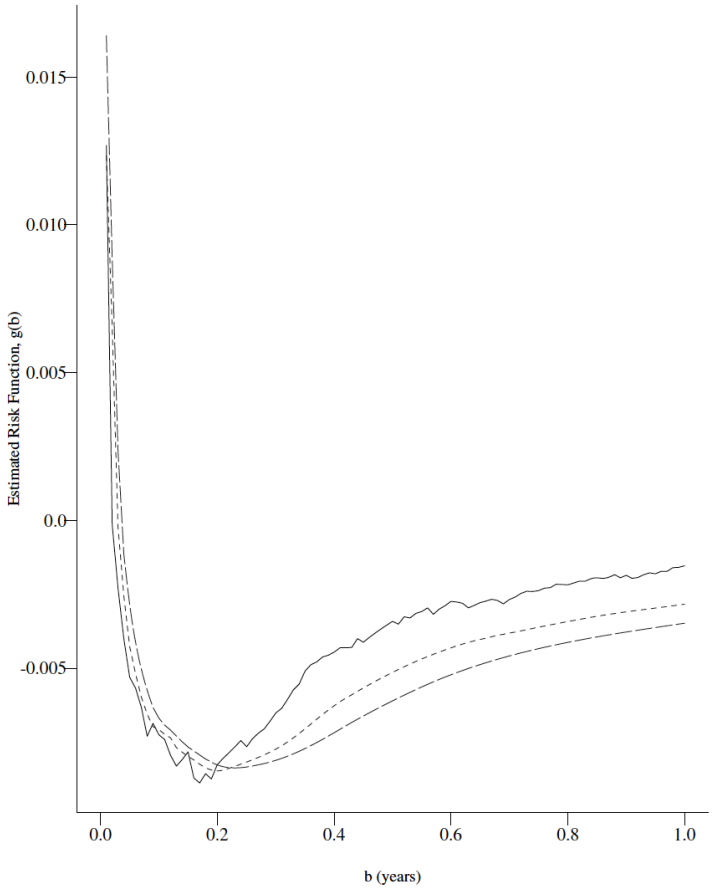


Figure 6.5 *Estimated risk function, g(b), for use in determination of the best bandwidth for the kidney transplant data. Uniform kernel (———); Epanechnikov kernel (------) Biweight kernel (— — —).*

Effect of bandwidth

For this example, the minimum occurs around $b = 0.17$ to $b = 0.23$ depending on the kernel.

Generally, there is a trade-off with smaller bandwidths having smaller bias but higher variance, and larger bandwidths (more smoothing) having less variance but greater bias. Measuring the quality of bandwidths and kernels using MISE is standard in kernel density estimation (not just survival analysis). Bias here means that $E[\widehat{h}(t)] \neq h(t)$.

Section 6.3: Estimation of Excess Mortality

The idea for this topic is to compare the survival curve or hazard rate for one group against a reference group, particular if the non-reference group is thought to have higher risk. The reference group might come from a much larger sample, so that its survival curve can be considered to be known.

An example is to compare the mortality for psychiatric patients against the general population. You could use census data to get the lifetable for the general population, and determine the excess mortality for the psychiatric patients. Two approaches are: a multiplicative model, and an additive model. In the multiplicative model, belonging to a particular group multiplies the hazard rate by a factor. In the additive model, belonging to a particular group adds a factor to the hazard rate.

Excess mortality

For the multiplicative model, if there is a reference hazard rate of $\theta_j(t)$ for the jth individual in a study (based on sex, age, ethnicity, etc.), then due to other risk factors, the hazard rate for the jth individual is

$$h_j(t) = \beta(t)\theta_j(t)$$

where $\beta(t) \geq 1$ implies that the hazard rate is higher than the reference hazard. We define

$$B(t) = \int_0^t \beta(u)\, du$$

as the **cumulative relative excess mortality**.

Excess mortality

Note that $\frac{d}{dt} = \beta(t)$. To estimate $B(t)$, let $Y_j(t) = 1$ if the jth individual is at risk at time t. Otherwise, let $Y_j(t) = 0$. Here $Y_j(t)$ is defined for left-truncated and right-censored data. Let

$$Q(t) = \sum_{j=1}^{n} \theta_j(t) Y_j(t)$$

where n is the sample size. Then we estimate $B(t)$ by

$$\widehat{B}(t) = \sum_{t_i \leq t} \frac{d_i}{Q(t_i)}$$

This value is comparing the actual number of deaths that have occurred by time t_i with the expected number of deaths based on the hazard rate and number of patients available to have died.

Excess mortality

The variance is estimated by

$$\widehat{V}[\widehat{B}(t)] = \sum_{t_i \leq t} \frac{d_i}{Q(t_i)^2}$$

$\beta(t)$ can be estimated by slope of $\widehat{B}(t)$, which can be improved by using kernel-smoothing methods on $\widehat{B}(t)$.

Excess mortality

For the additive model, the hazard is

$$h_j(t) = \alpha(t) + \theta_j(t)$$

Similarly to the multiplicative model, we estimate the cumulative excess mortality

$$A(t) = \int_0^t \alpha(u)\, du$$

In this case the expected cumulative hazard rate is

$$\Theta(t) = \sum_{j=1}^{n} \int_0^t \theta_j(u) \frac{Y_j(u)}{Y(u)}\, du$$

where

$$Y(u) = \sum_{j=1}^{n} Y_j(u)$$

is the number at risk at time u.

Excess mortality

The estimated excess mortality is

$$\widehat{A}(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i} - \Theta(t)$$

where the first term is the Nelson-Aalen estimator of the cumulative hazard. The variance is

$$\widehat{V}[\widehat{A}(t)] = \sum_{t_i \leq t} \frac{d_i}{Y(t)^2}$$

Excess mortality

For a lifetable where times are every year, you can also compute

$$\Theta(t) = \Theta(t-1) + \sum_t \frac{\lambda(a_j + t - 1)}{Y(t)}$$

where a_j is the age at the beginning of the study for patient j and λ is the reference hazard. Note that $\Theta(t)$ is a smooth function of t while $\widehat{A}(t)$ is has jumps.

Excess mortality

A more general model is to combine multiplicative and additive components, using

$$h_j(t) = \beta(t)\theta_j(t) + \alpha(t)$$

which is done in chapter 10.

Example: Iowa psychiatric patients

As an example, starting with the multiplicative model, consider 26 psychiatric patients from Iowa, where we compare to census data.

Iowa psychiatric patients

TABLE 1.7
Survival data for psychiatric inpatients

Gender	*Age at Admission*	*Time of Follow-up*
Female	51	1
Female	58	1
Female	55	2
Female	28	22
Male	21	30+
Male	19	28
Female	25	32
Female	48	11
Female	47	14
Female	25	36+
Female	31	31+
Male	24	33+
Male	25	33+
Female	30	37+
Female	33	35+
Male	36	25
Male	30	31+
Male	41	22
Female	43	26
Female	45	24
Female	35	35+
Male	29	34+
Male	35	30+
Male	32	35
Female	36	40
Male	32	39+

+Censored observation

Census data for Iowa

TABLE 6.2
1960 Iowa Standard Mortality

Males					
Age	*Survival Function*	*Hazard Rate*	*Age*	*Survival Function*	*Hazard Rate*
18–19	0.96394	0.00154	48–49	0.89596	0.00694
19–20	0.96246	0.00164	49–50	0.88976	0.00751
20–21	0.96088	0.00176	50–51	0.88310	0.00810
21–22	0.95919	0.00188	51–52	0.87598	0.00877
22–23	0.95739	0.00190	52–53	0.86833	0.00956
23–24	0.95557	0.00185	53–54	0.86007	0.01052
24–25	0.95380	0.00173	54–55	0.85107	0.01159
25–26	0.95215	0.00158	55–56	0.84126	0.01278
26–27	0.95065	0.00145	56–57	0.83058	0.01402
27–28	0.94927	0.00137	57–58	0.81902	0.01536
28–29	0.94797	0.00134	58–59	0.80654	0.01683
29–30	0.94670	0.00136	59–60	0.79308	0.01844
30–31	0.94541	0.00141	60–61	0.77859	0.02013
31–32	0.94408	0.00146	61–62	0.76307	0.02195
32–33	0.94270	0.00153	62–63	0.74650	0.02386
33–34	0.94126	0.00159	63–64	0.72890	0.02586
34–35	0.93976	0.00170	64–65	0.71029	0.02795

Excess mortality for Iowa psychiatric patients

TABLE 6.3

Computation of Cumulative Relative Mortality for 26 Psychiatric Patients

t_i	d_i	$Q(t_i)$	$\hat{B}(t)$	$\hat{V}[\hat{B}(t)]$	$\sqrt{\hat{V}[\hat{B}(t)]}$
1	2	0.05932	33.72	568.44	23.84
2	1	0.04964	53.86	974.20	31.21
11	1	0.08524	65.59	1111.84	33.34
14	1	0.10278	75.32	1206.51	34.73
22	2	0.19232	85.72	1260.58	35.50
24	1	0.19571	90.83	1286.69	35.87
25	1	0.18990	96.10	1314.42	36.25
26	1	0.18447	101.52	1343.81	36.66
28	1	0.19428	106.67	1370.30	37.02
32	1	0.18562	112.05	1399.32	37.41
35	1	0.16755	118.02	1434.94	37.88
40	1	0.04902	138.42	1851.16	43.03

Excess mortality for Iowa psychiatric patients

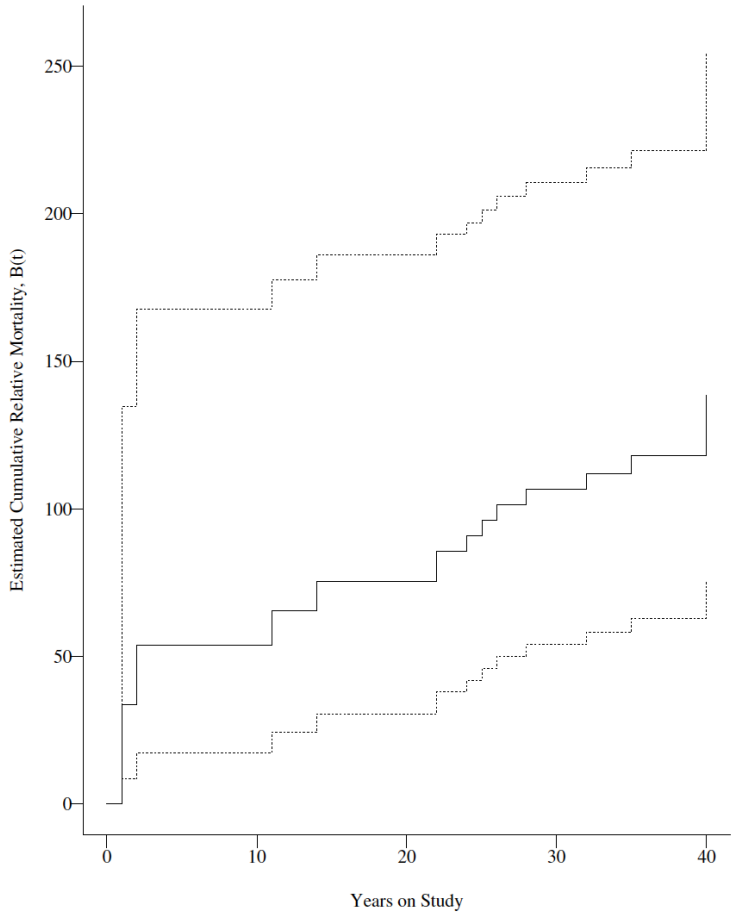


Figure 6.8 *Estimated cumulative relative mortality (solid line) and 95% pointwise confidence interval (dashed line) for Iowa psychiatric patients*

Excess mortality

The cumulative excess mortality is difficult to interpret. The slope of the curve is more meaningful. The curve is relatively linear. If we consider age 10 to age 30, the curve goes from roughly 50 to 100, suggesting a slope of $(100 - 50)/(30 - 10) = 2.5$, so that patients aged 10 to 30 had a roughly 2.5 times higher chance of dying.

This is a fairly low-risk age group, for which suicide is high risk factor. Note that the census data might include psychiatric patients who have committed suicide, so we might be comparing psychiatric patients to the general population which includes psychiatric patients, as opposed to psychiatric patients compared to people who have not been psychiatric patients, so this might bias results.

Survival curves

You can use the reference distribution to inform the survival curve instead of just relying on the data. This results in an adjusted or corrected survival curve. Let $S^*(t) = \exp[-\Theta(t)]$ (or use the cumulative hazard based on multiplying the reference hazard by the excess harzard) and let $\widehat{S}(t)$ be the standard Kaplan-Meier survival curve (using only the data, not the reference survival data). Then $S^c(t) = \widehat{S}(t)/S^*(t)$ is the corrected survival function. The estimate can be greater than 1, in which case the estimate can be set to 1.

Typically, $S^*(t)$ is less than 1, so that dividing by this quantity increases the estimated survival probabilities. This is somewhat similar in Bayesian statististics to the use of the prior, using the reference survival times as a prior for what the psychiatric patients are likely to experience.

Consequently, the adjusted survival curve is in between the kaplan-meier (data only) estimate, and the reference survival times.

Survival curves

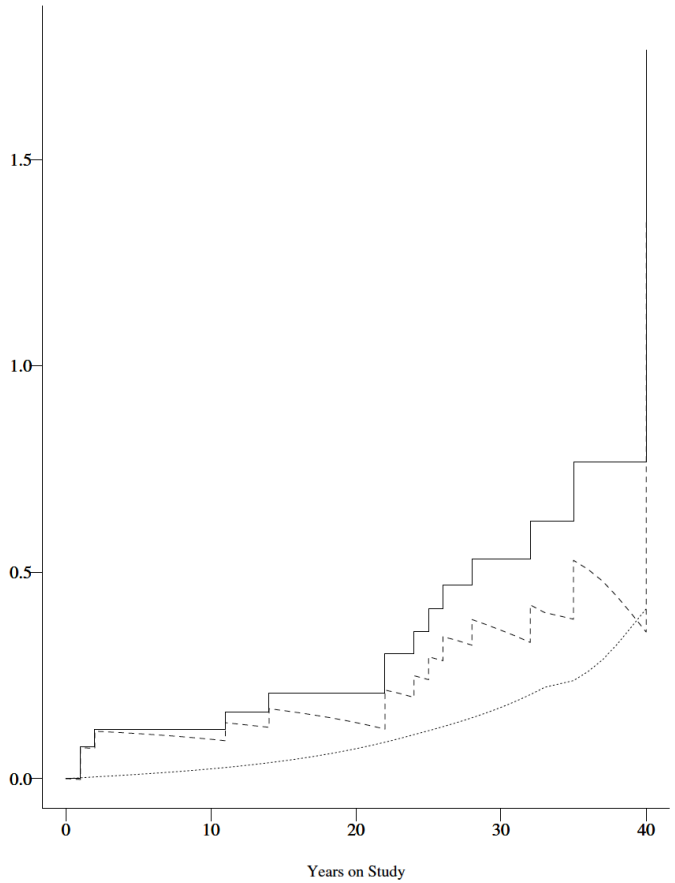


Figure 6.9 *Estimated cumulative excess mortality for Iowa psychiatric patients. Nelson–Aalen estimator (———) Expected cumulative hazard (------) Cumulative excess mortality (— — —)*

Survival curves

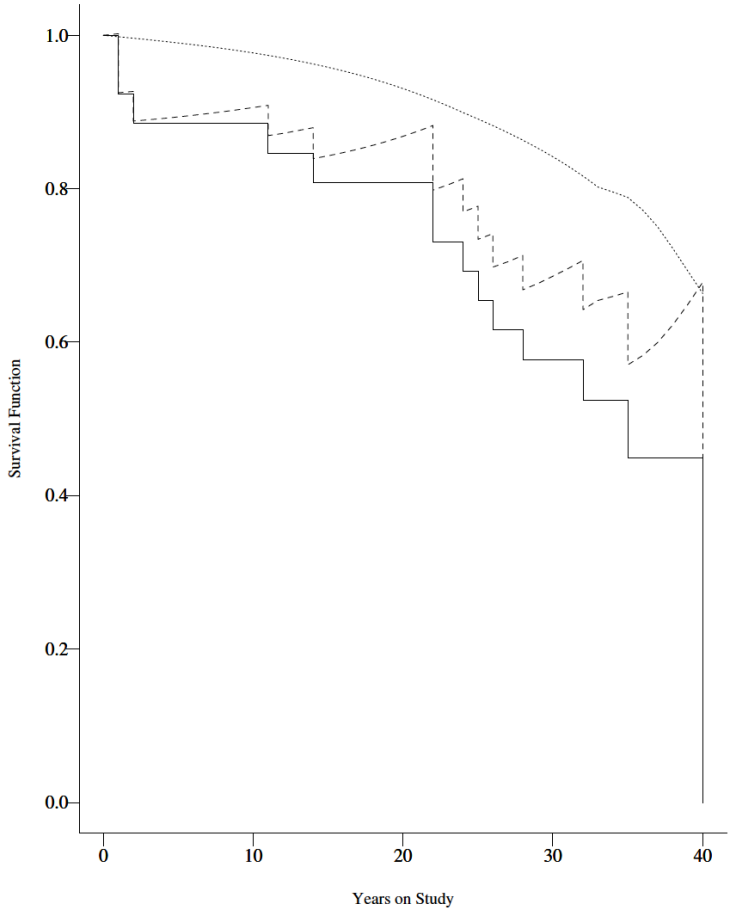


Figure 6.10 *Adjusted survival curves for Iowa psychiatric patients. Observed survival (———) Expected survival (------) Corrected survival (— — —)*

Bayesian nonparametric survival analysis

The previous example leads naturally to Bayesian nonparametric survival analysis. Here we have prior information (or prior beliefs) about the shape of the survival curve (such as based on a reference survival function). The survival curve based on this previous information is combined with the likelihood of the survival data to produce a posterior estimate of the survival function.

Reasons for using a prior are: (1) to take advantage of prior information or expertise of someone familiar with the type of data, (2) to get a reasonable estimate when the sample size is small.

Bayesian survival analysis

In frequentist statistical methods, parameters are treated as fixed, but unknown, and an estimator is chosen to estimate the parameters based on the data and a model (including model assumptions). Parameters are unknown, but are treated as not being random.

Philosophically, the Bayesian approach is to try to model all uncertainty using random variables. Uncertainty exists both in the form of the data that would arise from a probability model as well as the parameters of the model itself, so both observations and parameters are treated as random. Typically, the observations have a distribution that depends on the parameters, and the parameters themselves come from some other distribution. Bayesian models are therefore often hierarchical, often with multiple levels in the hierarchy.

Bayesian survival analysis

For survival analysis, we think of the (unknown) survival curve as the parameter. From a frequentist point of view, survival probabilities determine the probabilities of observing different death times, but there are no probabilities of the survival function itself.

From a Bayesian point of view, you can imagine that there was some stochastic process generating survival curves according to some distribution on the space of survival curves. One of the survival curves happened to occur for the population we are studying. Once that survival function was chosen, event times could occur according to that survival curve.

Bayesian survival analysis

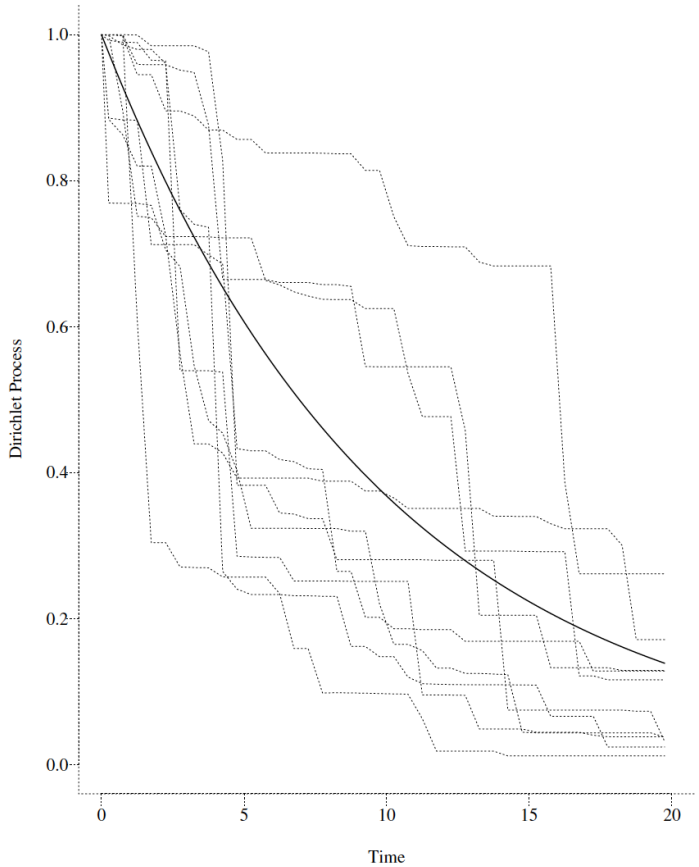


Figure 6.11 *Sample of ten sample paths (dashed lines) and their mean (solid line) for samples from a Dirichlet prior with* $S_0(t) = \exp(-0.1t)$ *and* $c = 5$.

Bayesian survival analysis

We imagine that there is a true survival curve $S(t)$, and an estimated survival curve, $\widehat{S}(t)$. We define a loss function as

$$L(S, \widehat{S}) = \int_0^\infty [\widehat{S}(t) - S(t)]^2 \, dt$$

The function $\widehat{S}$ that minimizes the expected value of the loss function is called the posterior mean, which is used to estimate the survival function.

A prior for survival curves

A typical way to assign a prior on the survival function is to use a Dirichlet process prior. For a Dirichlet process, we partition the real line into intervals $A_1, \ldots, A_k$, so that $P(X \in A_i) = W_i$. The numbers $(W_1, \ldots, W_k)$ have a k-dimension Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$. For this to be a Dirichlet distribution, we must have Z_i, $i = 1, \ldots, k$ are independent gamma random variables with shape parameter α_i and $W_i = \frac{Z_i}{\sum_{i=1}^{k} Z_i}$. By construction, the random numbers $W_i's$ are between 0 and 1 and sum to 1, so when interpreted as probabilities, they form a discrete probability distribution.

Essentially, we can think of a Dirichlet distribution as a distribution on unfair dice with k sides. We want to make a die that has k sides, and we want the probabilities of each side to be randomly determined. How fair or unfair the die is partly depends on the α parameters and partly depends on chance itself.

A prior for survival curves

We can also think of the Dirichlet distrbibution as generalizing the beta distribution. A beta random variable is a number between 0 and 1. This number partitions the interval [0,1] into two pieces, $[0, x)$ and $[x, 1]$. A Dirichlet random variable partitions the interval into k regions, using $k-1$ values between 0 and 1. The joint density for these $k-1$ values is

$$f(w_1, \ldots, w_{k-1}) = \frac{\Gamma[\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \left[\prod_{i=1}^{k-1} w_i^{\alpha_i - 1} \right] \left[1 - \sum_{i=1}^{k-1} w_i \right]^{\alpha_k - 1}$$

which reduces to a beta density with parameters (α_1, α_2) when $k = 2$.

Assigning a prior

To assign a prior on the space of survival curves, first assume an average survival function, $S_0(t)$. The Dirichlet prior determines when the jumps occur, and the exponential curve gives the decay of the curve between jumps. Simulated survival curves when $S_0(t) = e^{-0.1t}$ and $\alpha = 5S_0(t)$ are given below.

Bayesian survival analysis

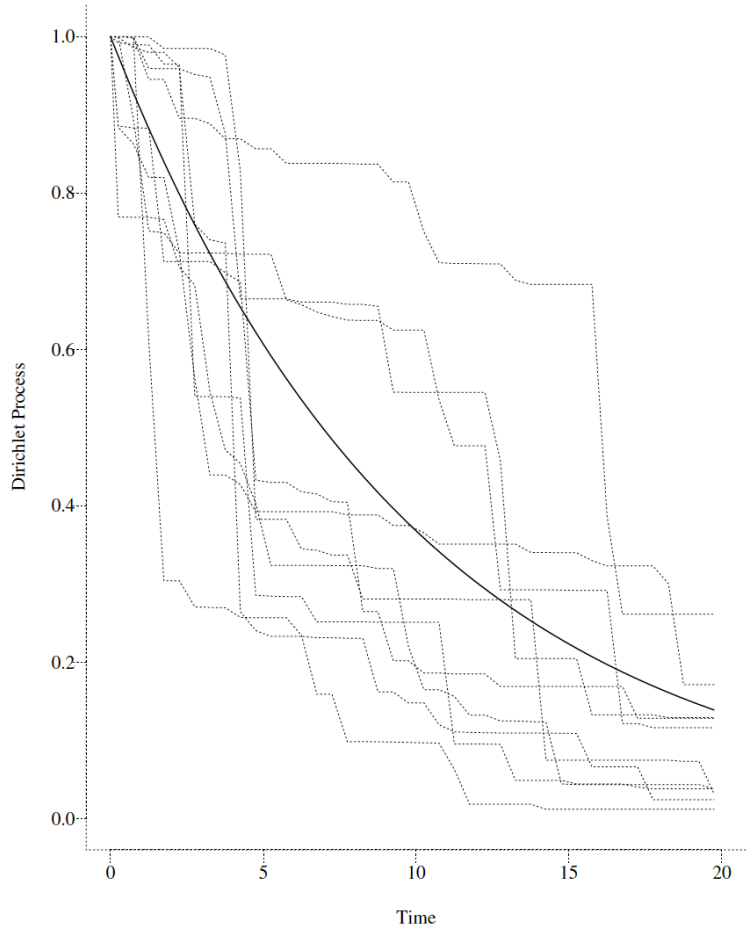


Figure 6.11 *Sample of ten sample paths (dashed lines) and their mean (solid line) for samples from a Dirichlet prior with* $S_0(t) = \exp(-0.1t)$ *and* $c = 5$.

Bayesian survival analysis

Other approaches are to have a prior for the cumulative hazard function and to use Gibb's sampling or Markov chain Monte Carlo. These topics would be more appropriate to cover after a class in Bayes methods.