Survival analysis:
Parametric and non-parametric survival

Mark Stevenson

Faculty of Veterinary and Agricultural Sciences

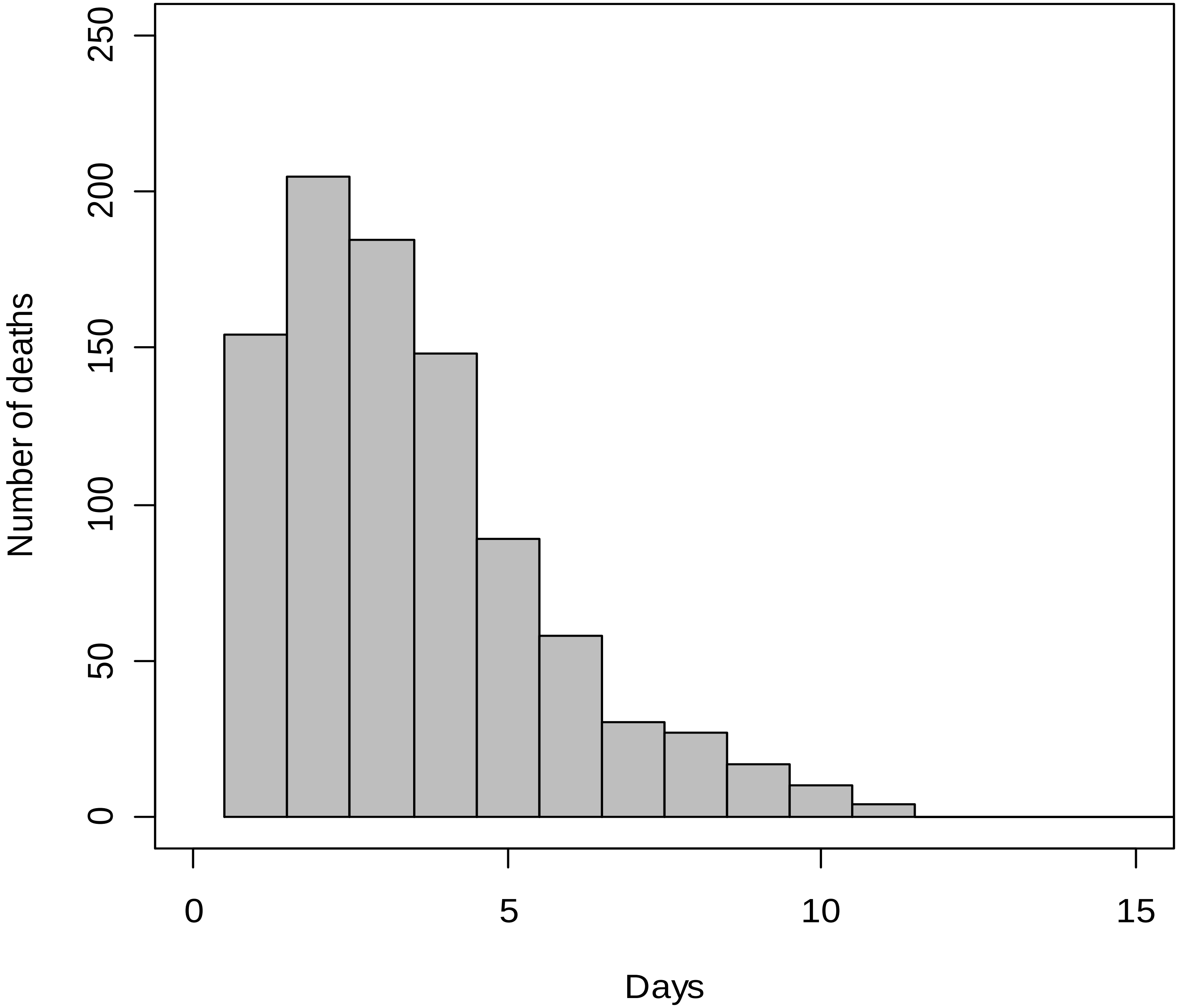
The University of Melbourne, Parkville Victoria 3010 Australia

mark.stevenson1@unimelb.edu.au

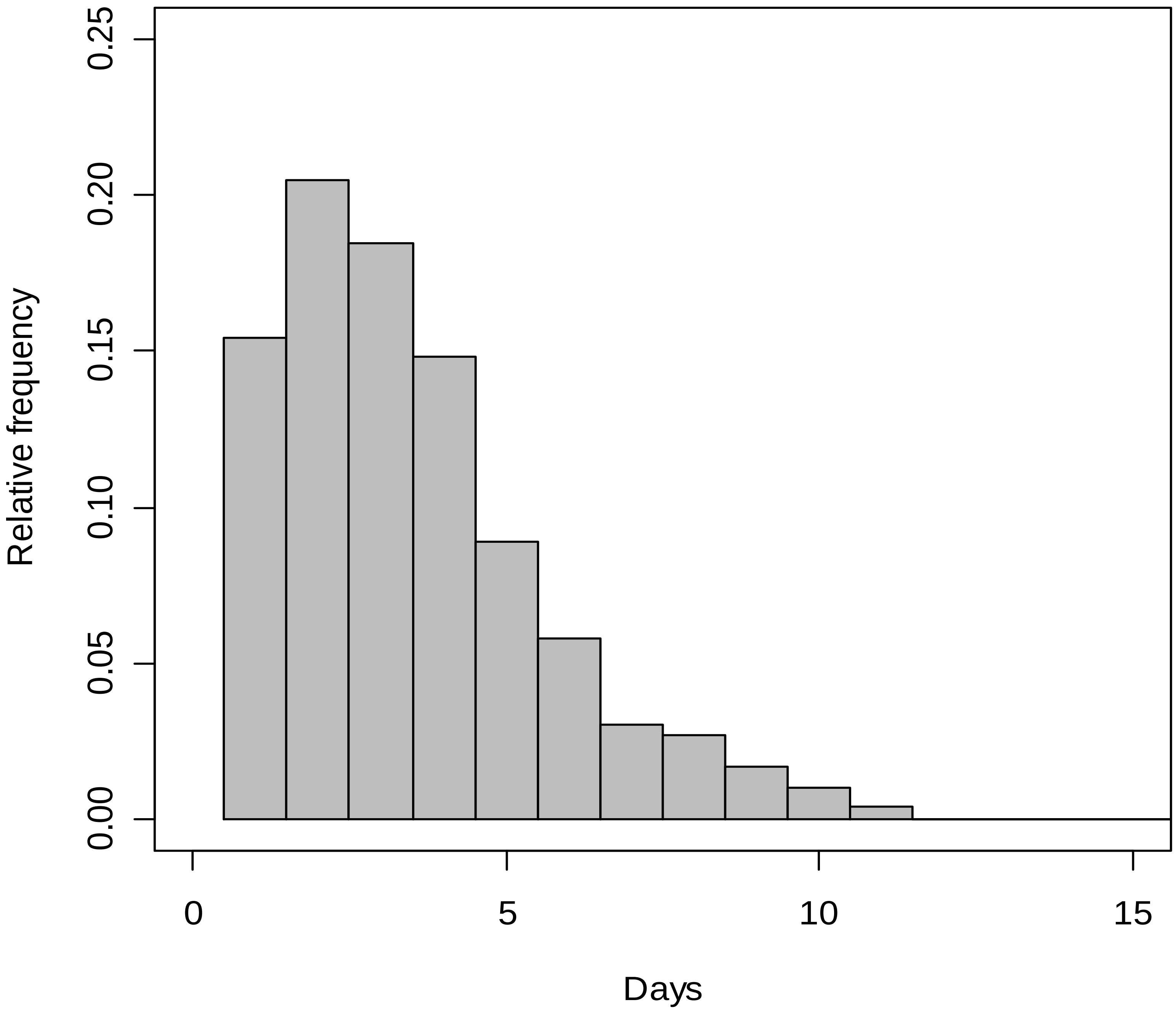
Roadmap

- Survival and hazard
- Parametric survival functions
  - Exponential
  - Weibull
- Non-parametric survival functions
- Presentation

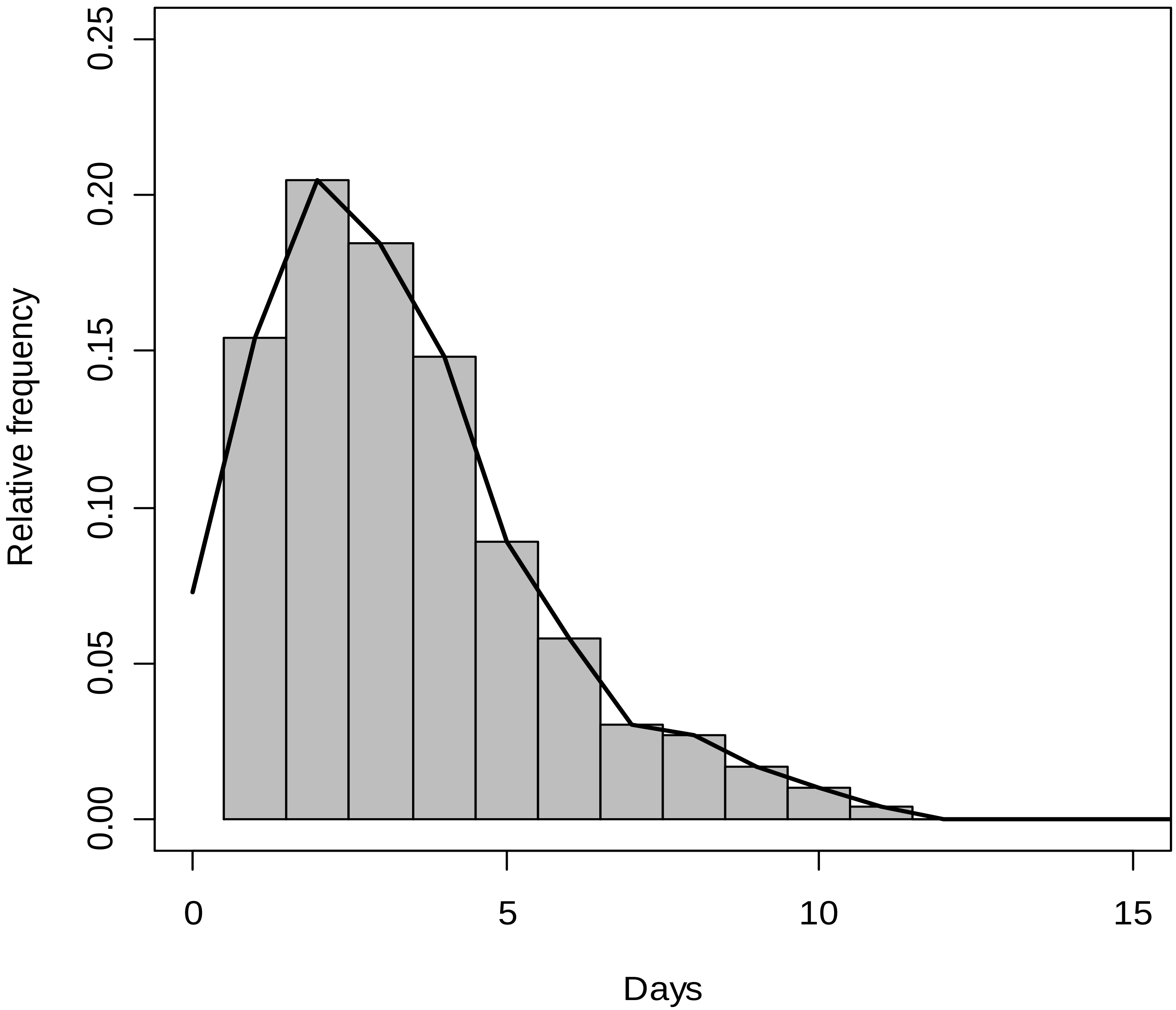
Frequency histogram showing the number of deaths as a function of time.



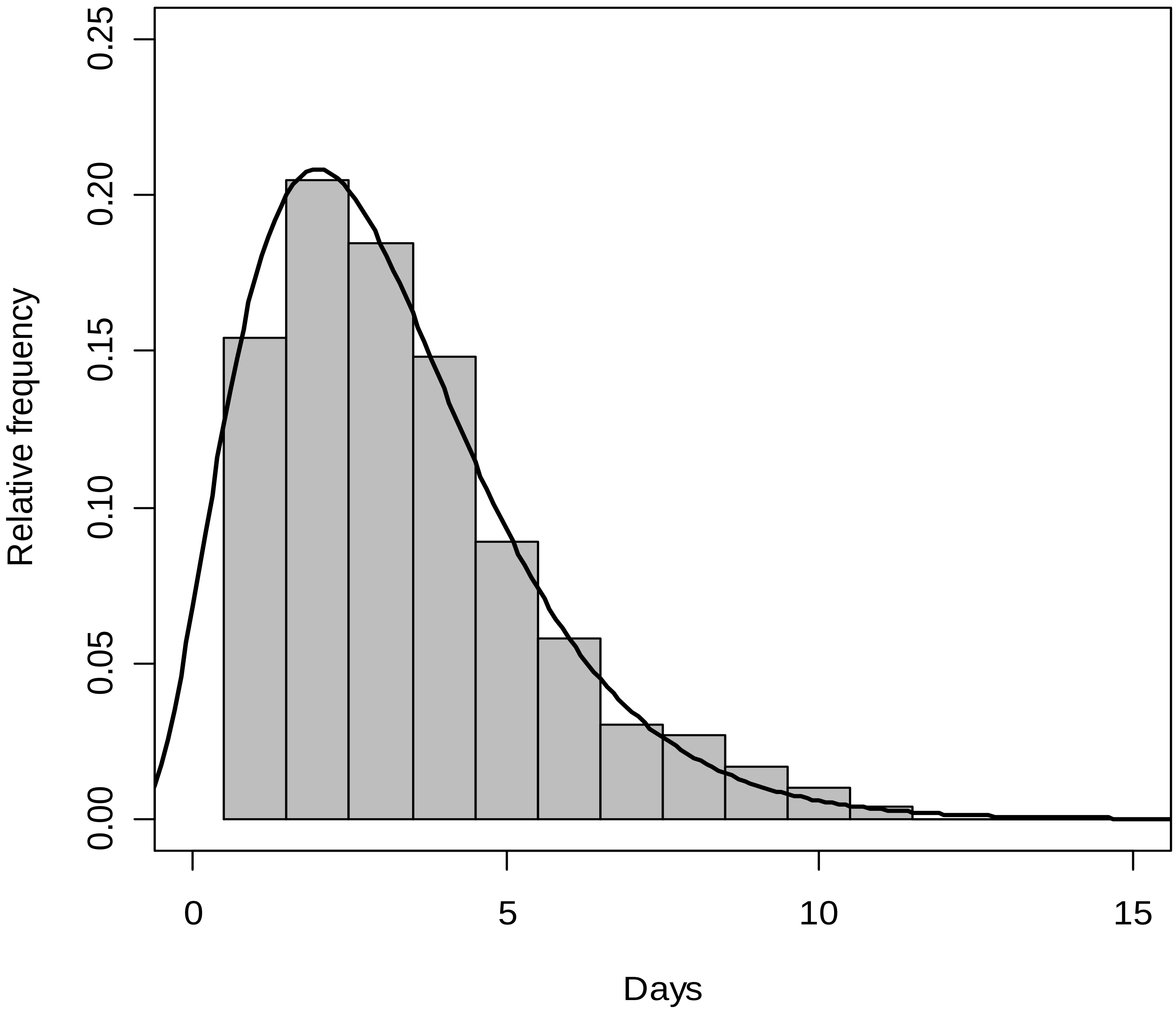
Frequency histogram showing the relative frequency of deaths as a function of time.



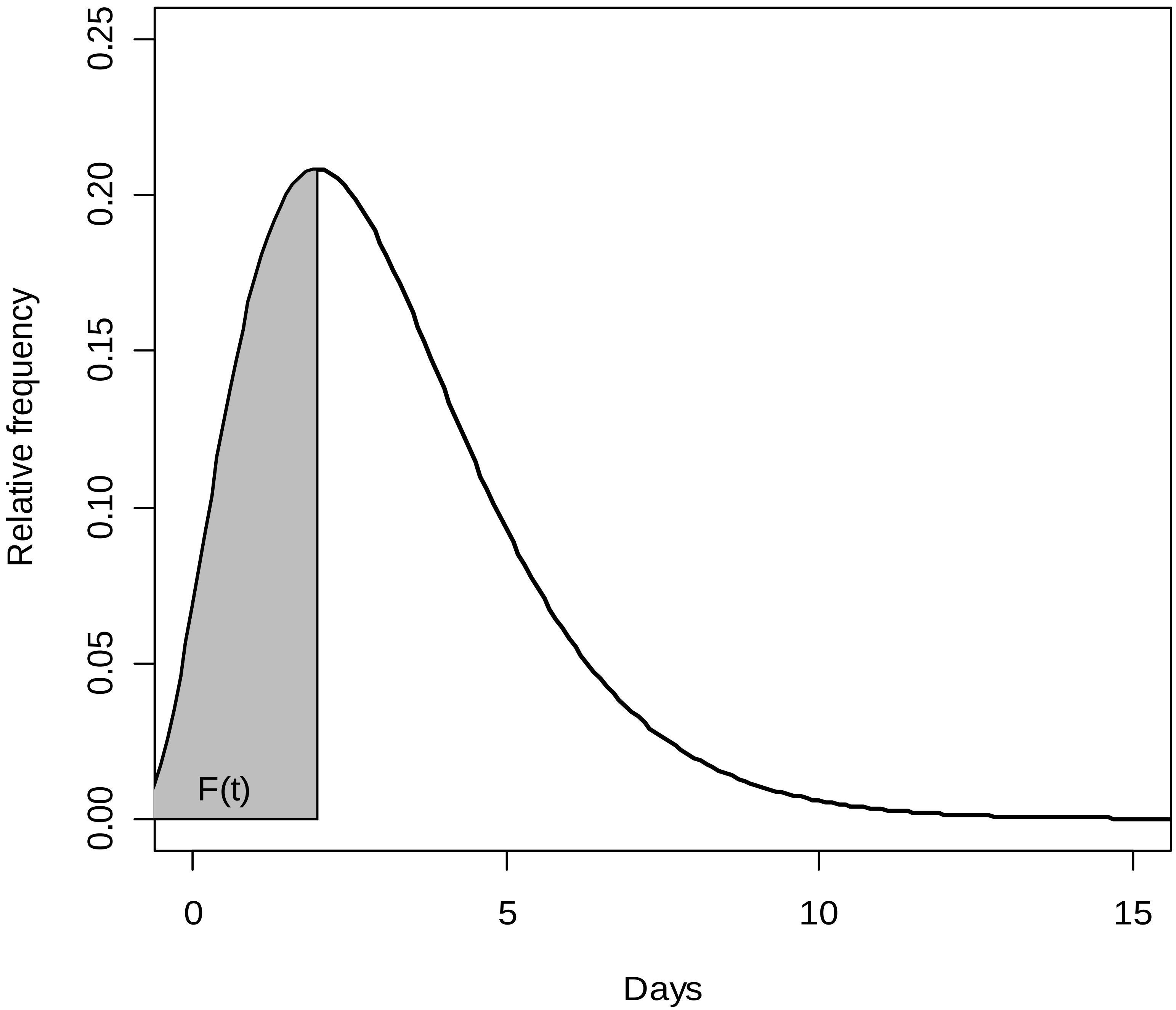
$f(t)$ = death density function.



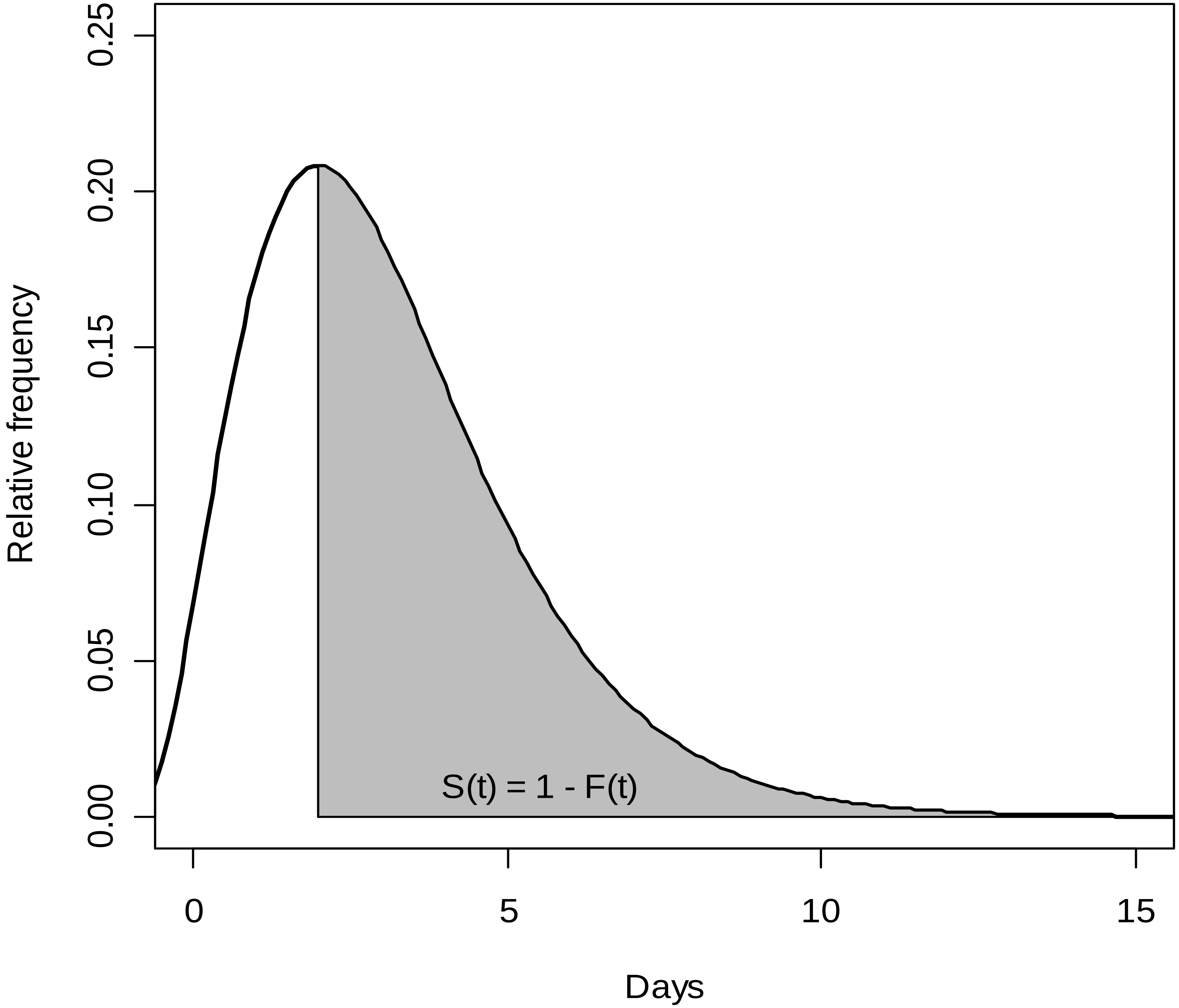
f(t) = death density function.



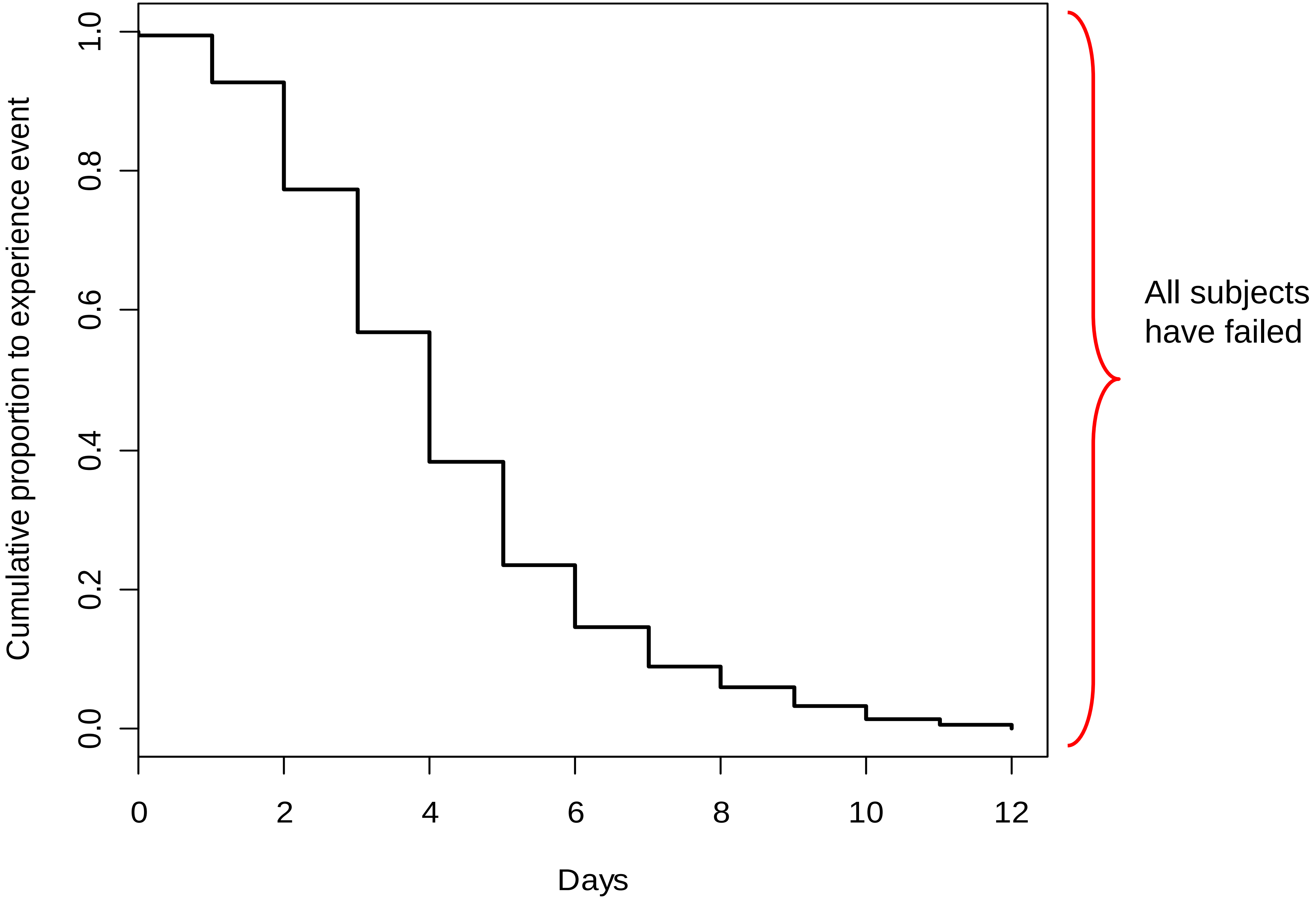
F(t) = the failure function = the proportion not surviving past time t.



S(t) = survival function = the proportion of the group that survive up to time t = 1 - F(t).



S(t) = survival function = the proportion of the group that survive up to time t = 1 - F(t).



Failurefunction: $F(t) = 1 - S(t)$ Probability of not surviving past time *t*

Instantaneous failure rate: $f(t) = \dfrac{dF(t)}{dt}$ The slope of function of F(t)

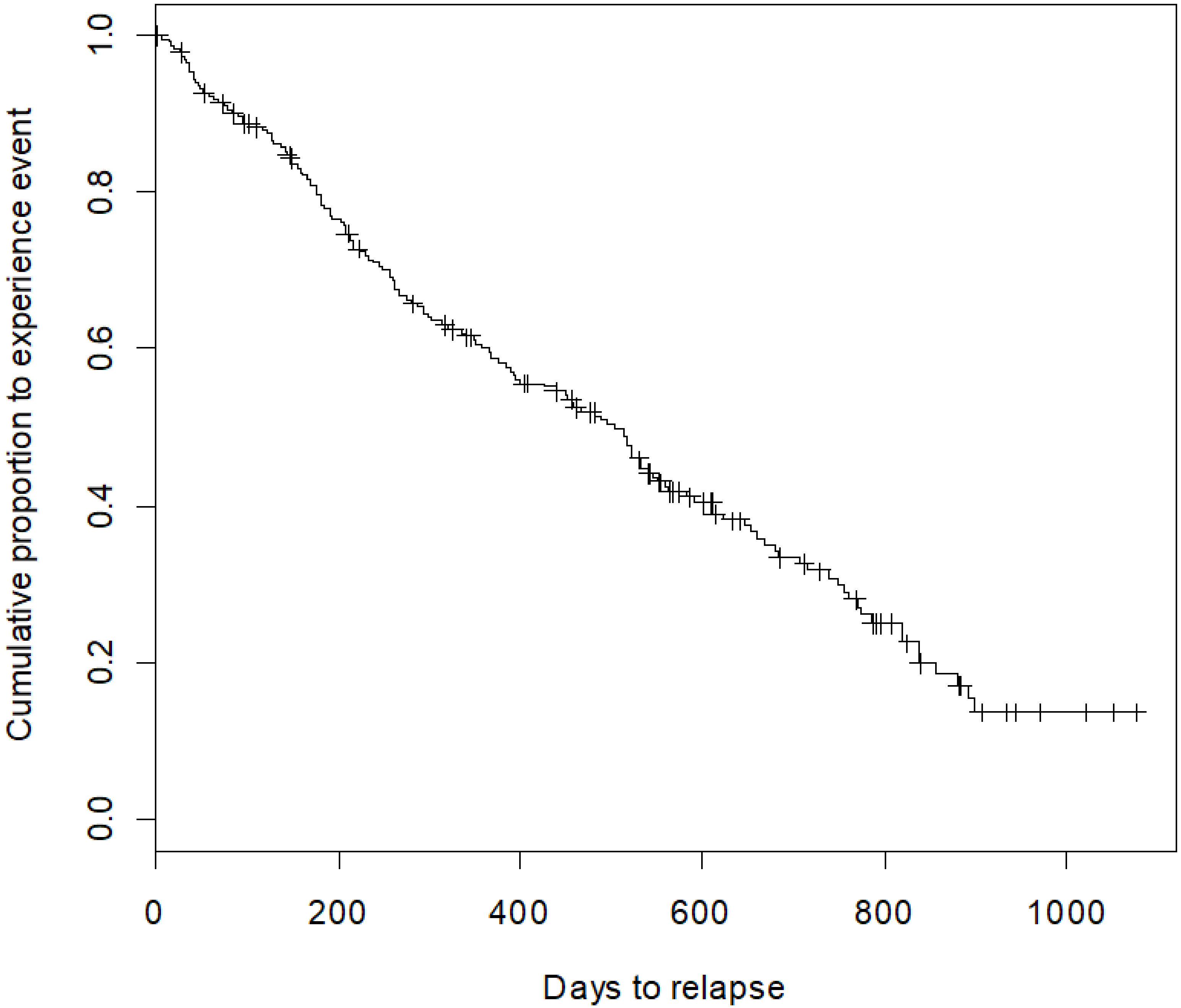
Instantaneous hazard function: $h(t) = \dfrac{f(t)}{S(t)}$ Probability of failure at time *t*

Cumulative hazard function: $H(t) = -\ln S\,(t)$ The accumulated hazard over time

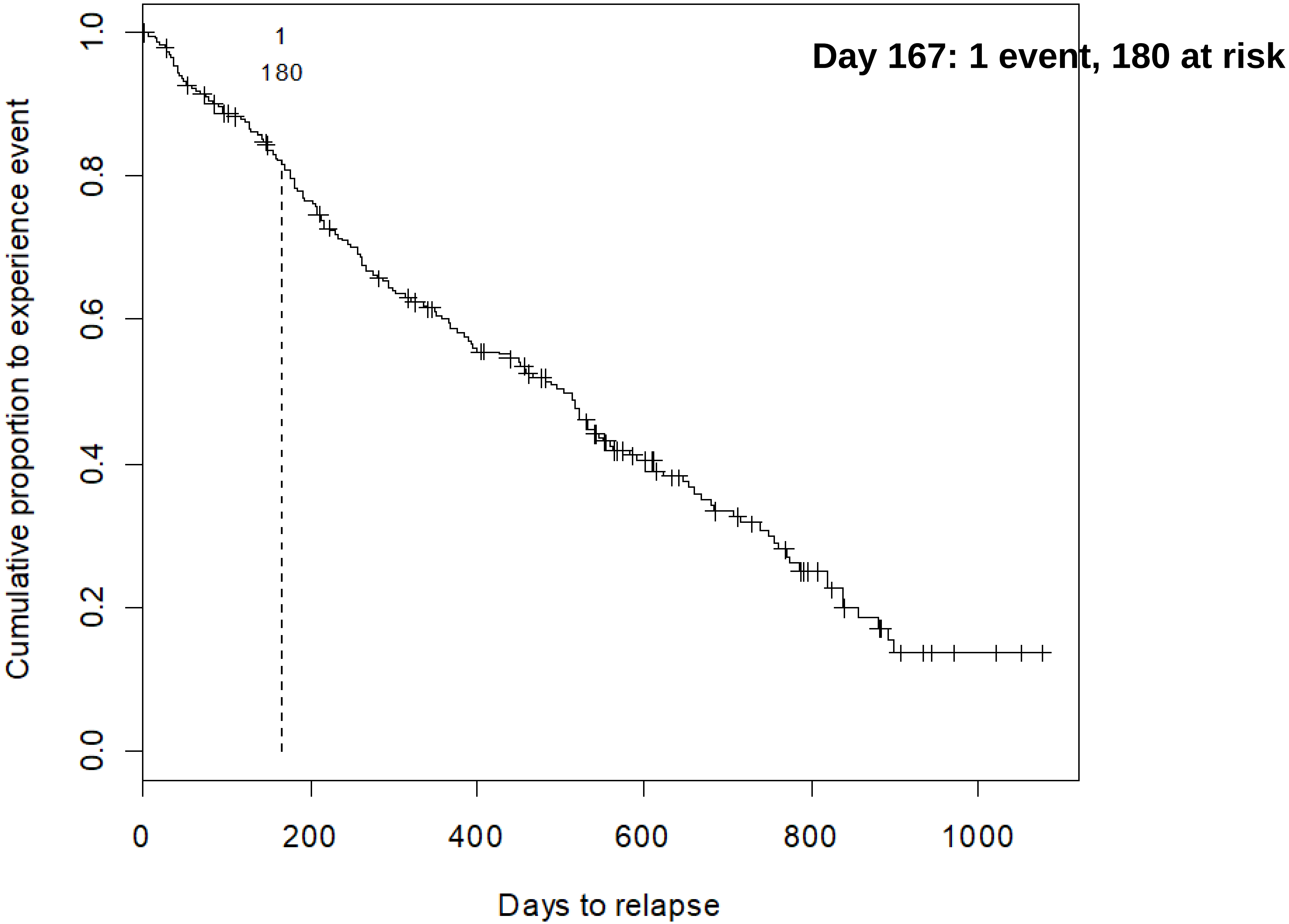
Survival and hazard

- Instantaneous hazard h(t)
  - ≡ 'hazard' (Dohoo and Martin)
  - the probability of an event occurring at time t, given that it hasn't occurred already

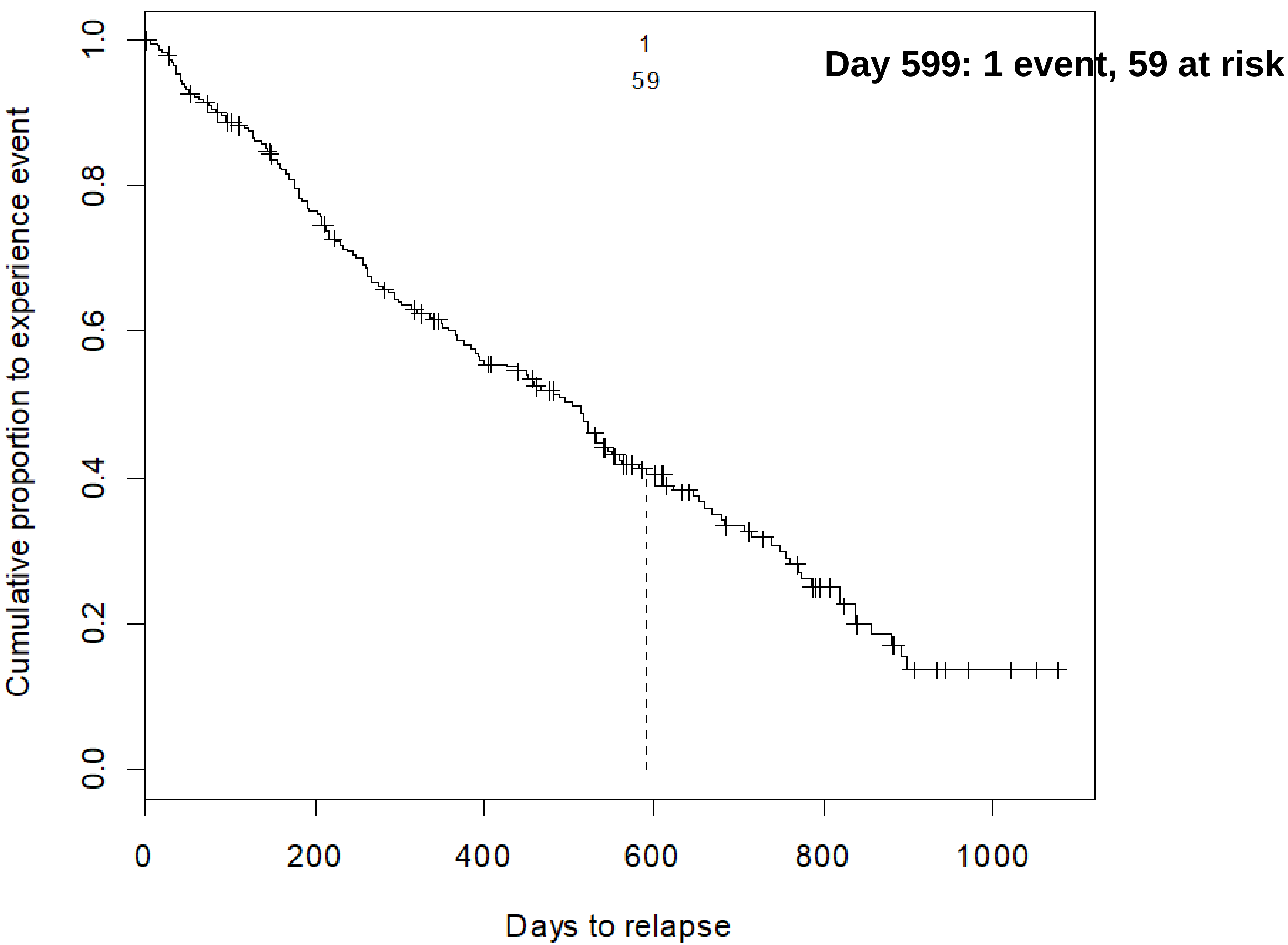
Kaplan-Meier survival curve showing the cumulative proportion of drug addict population that had not relapsed following discharge from treatment event as a function of calendar date.



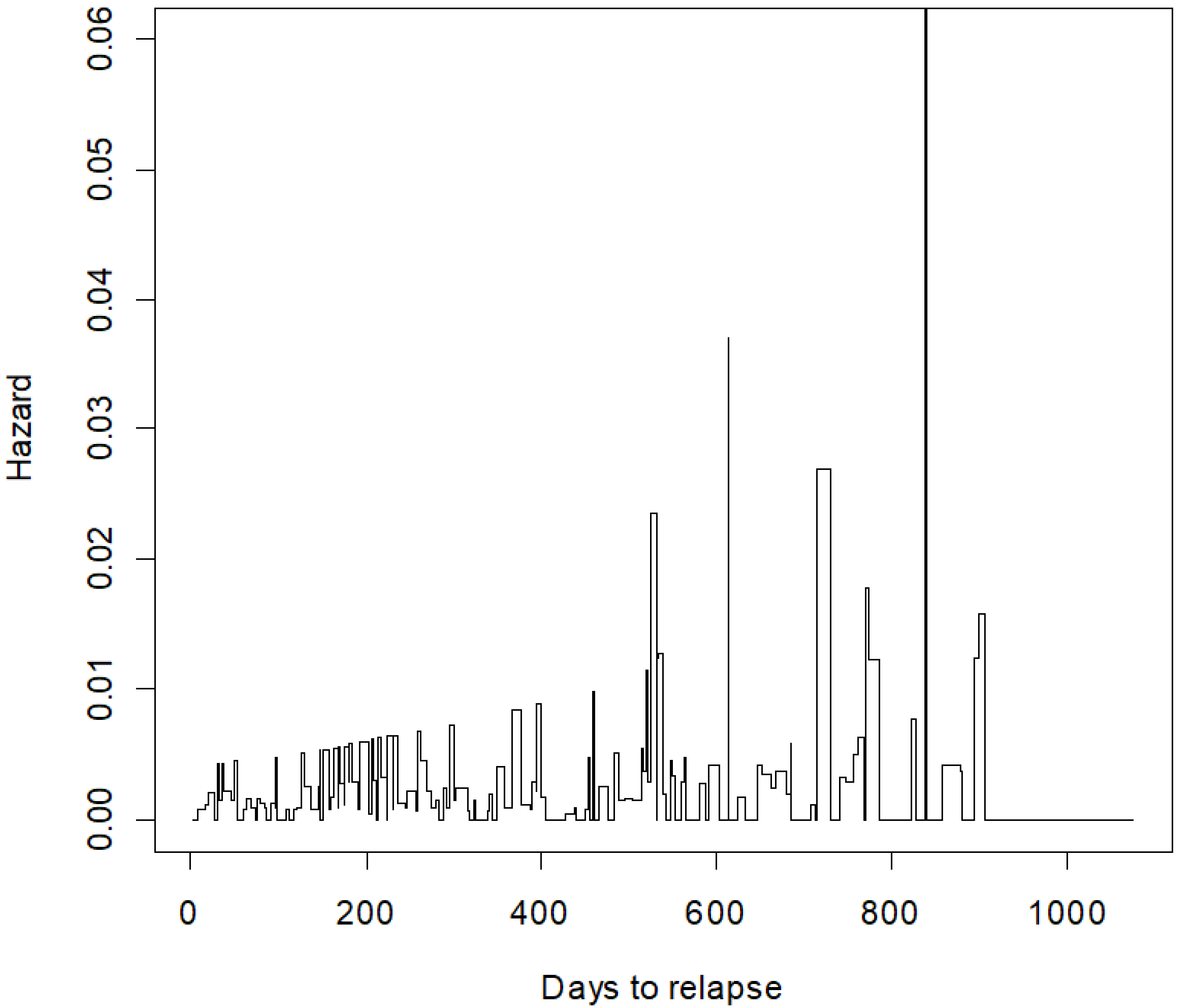
Kaplan-Meier survival curve showing the cumulative proportion of drug addict population that had not relapsed following discharge from treatment event as a function of calendar date.



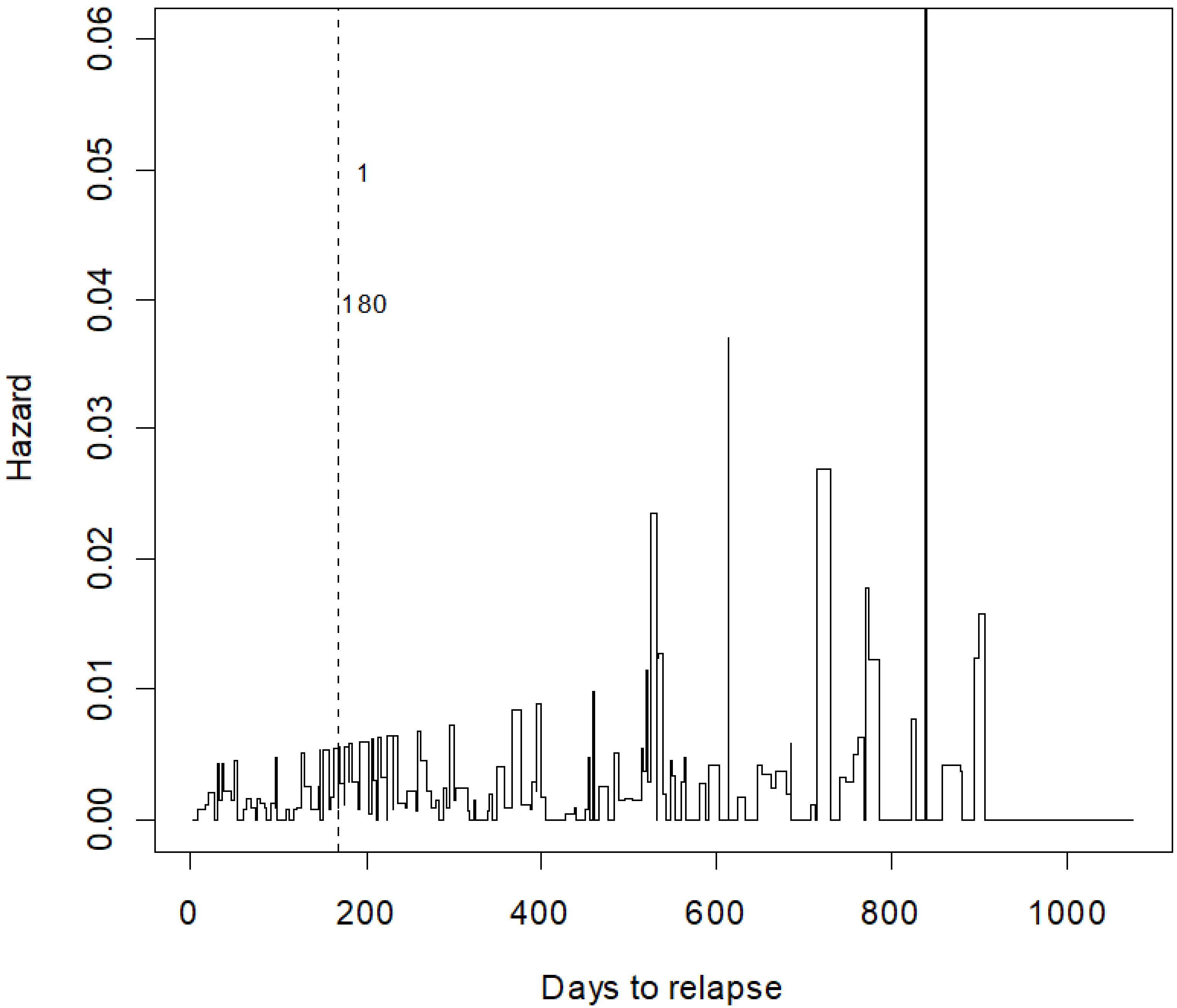
Kaplan-Meier survival curve showing the cumulative proportion of drug addict population that had not relapsed following discharge from treatment event as a function of calendar date.



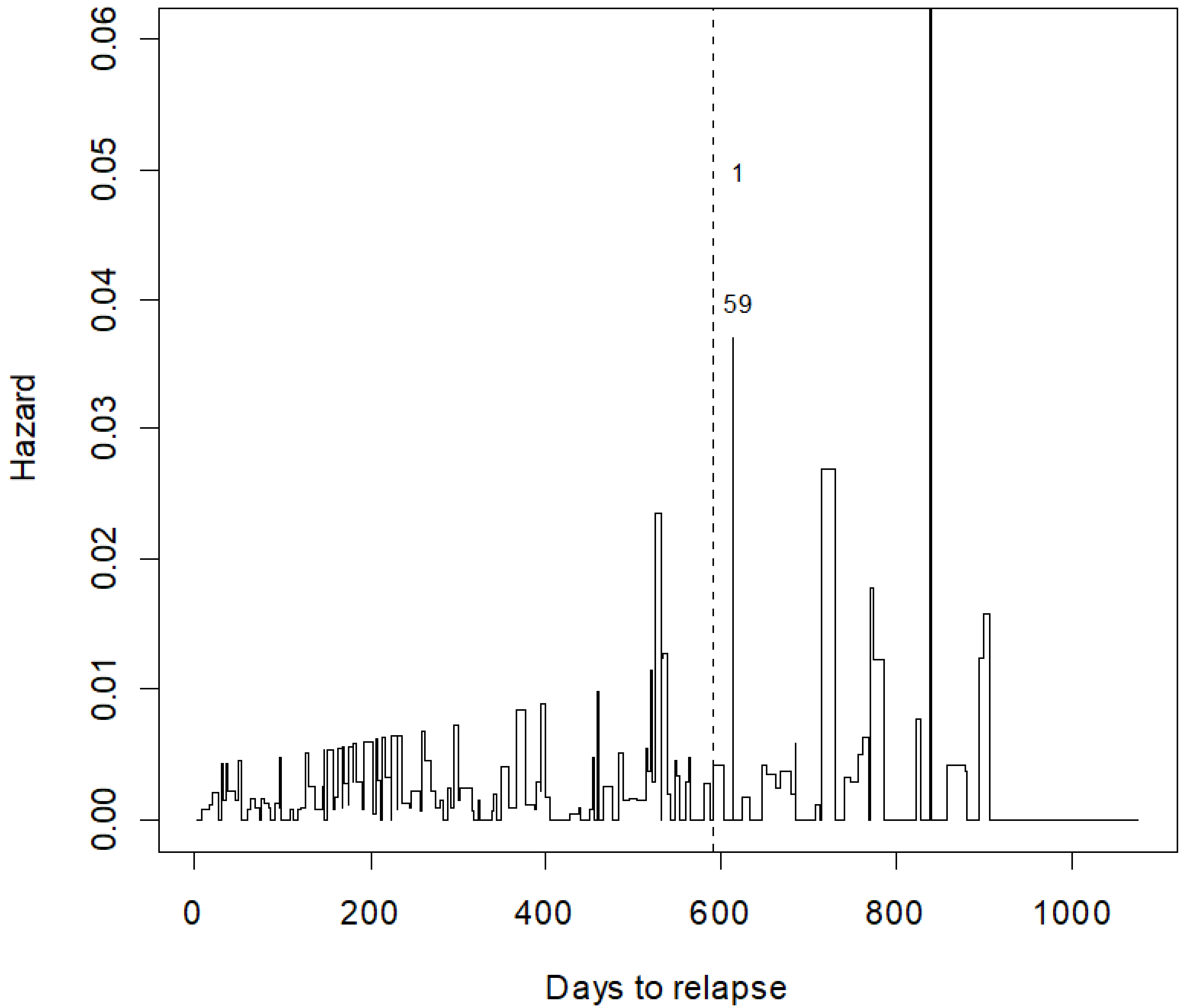
Instantaneous hazard of relapse in drug addicts following discharge from rehabilitation.



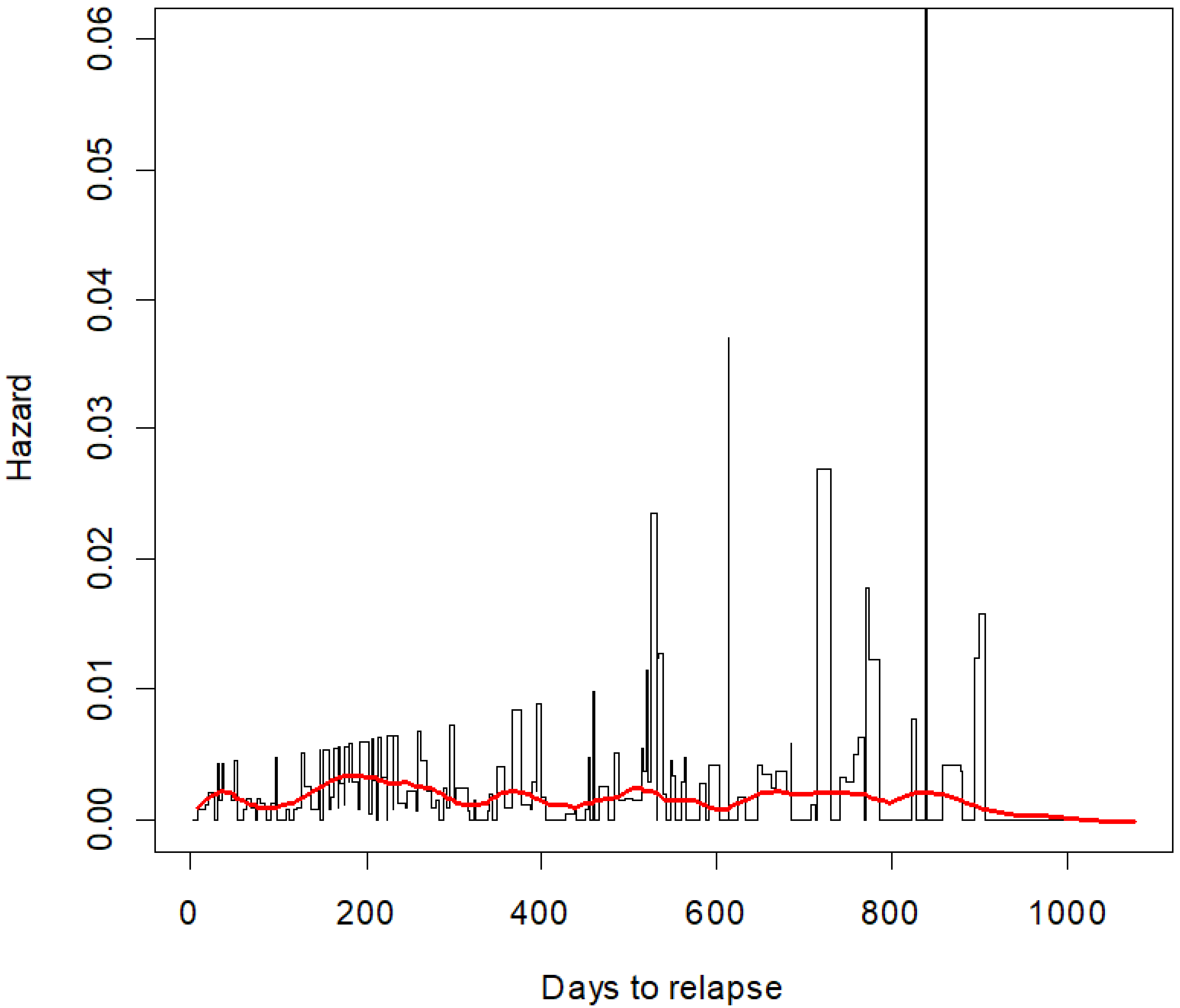
Instantaneous hazard of relapse in drug addicts following discharge from rehabilitation.



Instantaneous hazard of relapse in drug addicts following discharge from rehabilitation.



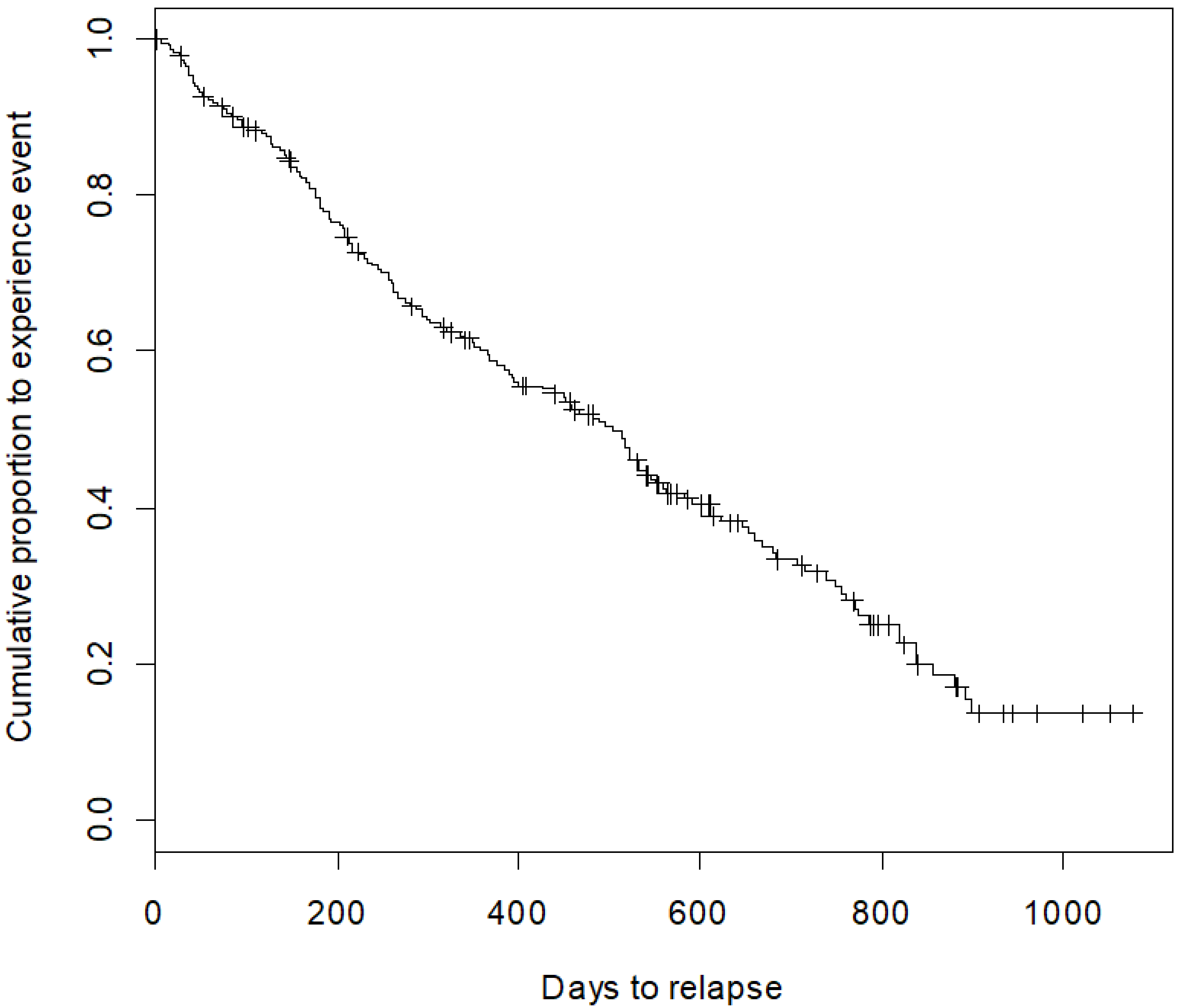
Instantaneous hazard of relapse in drug addicts following discharge from rehabilitation.



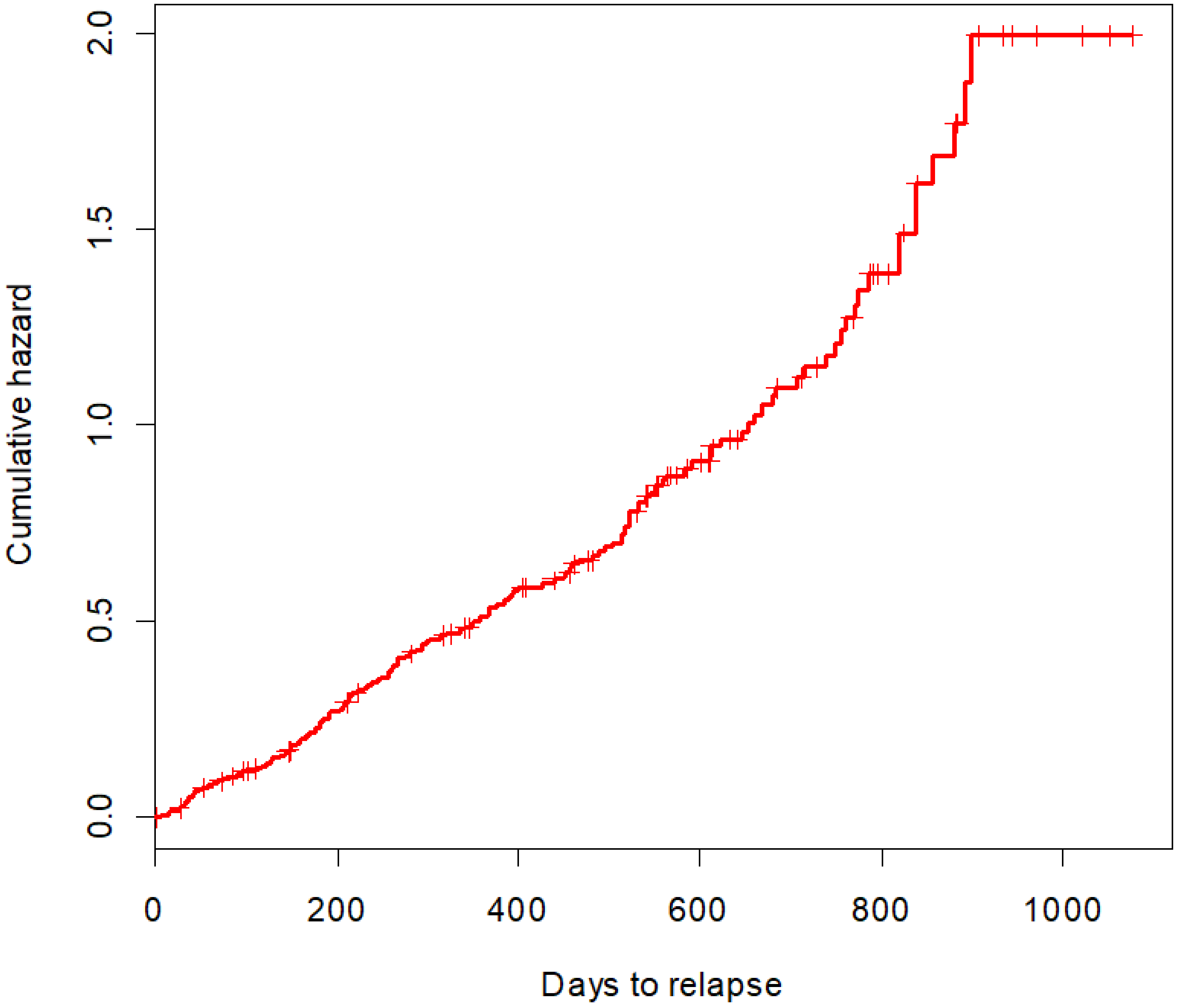
Survival and hazard

- Cumulative hazard H(t)
  - also referred to as ‘integrated hazard’
  - equals the total amount of accumulated risk that an individual has encountered from the beginning of the observation period
  - represents the expected number of events that would have occurred by time *t*

Kaplan-Meier survival curve showing the cumulative proportion of drug addict population that had not relapsed following discharge from treatment event as a function of calendar date.



Cumulative hazard of relapse in drug addicts following discharge from rehabilitation.



Roadmap

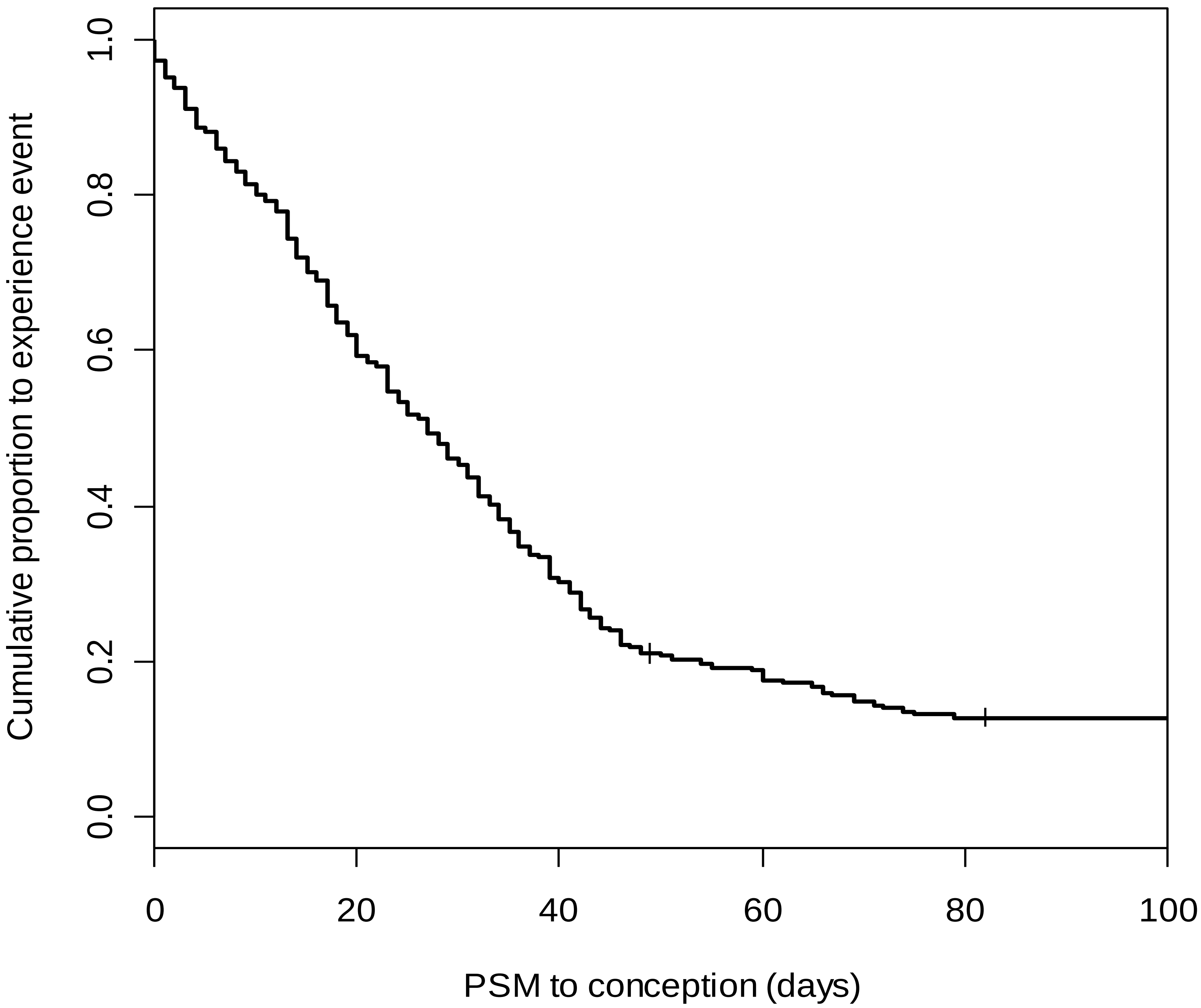
- Survival and hazard
- <u>Parametric survival functions</u>
  - Exponential
  - Weibull
- Non-parametric survival functions
- Presentation

Parametric survival distributions

- In the last lecture we talked about ways to describe survivorship
  - x axis: time
  - y axis: proportion of group that had not experienced event

- Some survival functions ‘predictable’ …

Days to conception as a function of days after Planned Start of Mating date in dairy cows.

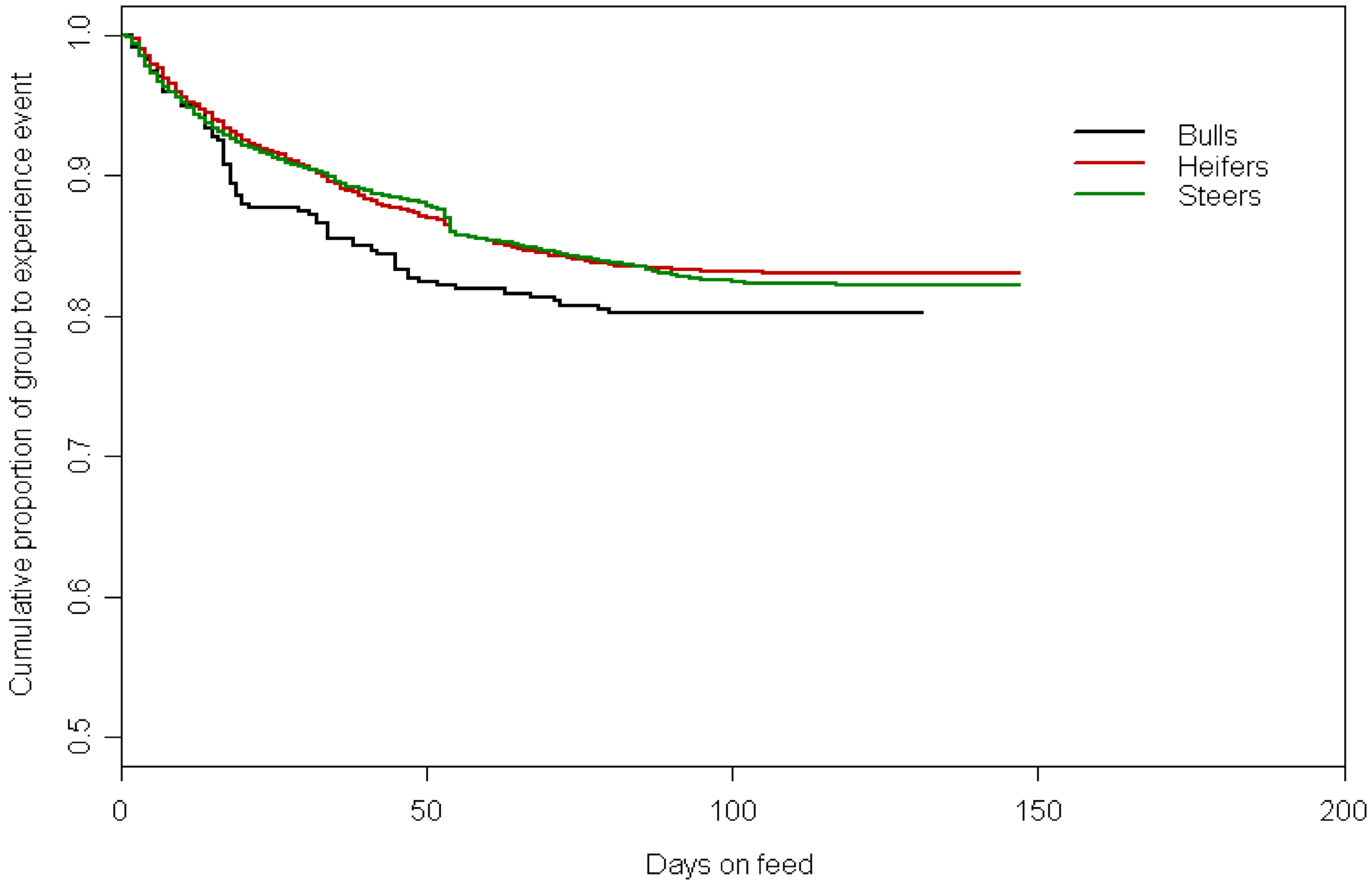


Parametric survival distributions

- Some survival functions ‘unpredictable’ …

Indonesian feedlot.

Kaplan-Meier survival curve showing cumulative proportion of animals pulled for all reasons as a function of calendar date, by stock class.



Parametric survival distributions

- Those distributions follow a nice, well-defined pattern can be described by <u>parametric</u> distributions

- Those distributions that are irregular in shape can only be described by <u>non-parametric</u> distributions

Parametric survival distributions

- The advantage of the parametric approach is that survival can be predicted (because we know how the curve will behave beyond the last observation)

- Non-parametric models are not suitable for prediction because there is uncertainty about the nature of the distribution beyond the last observation

Parametric survival functions

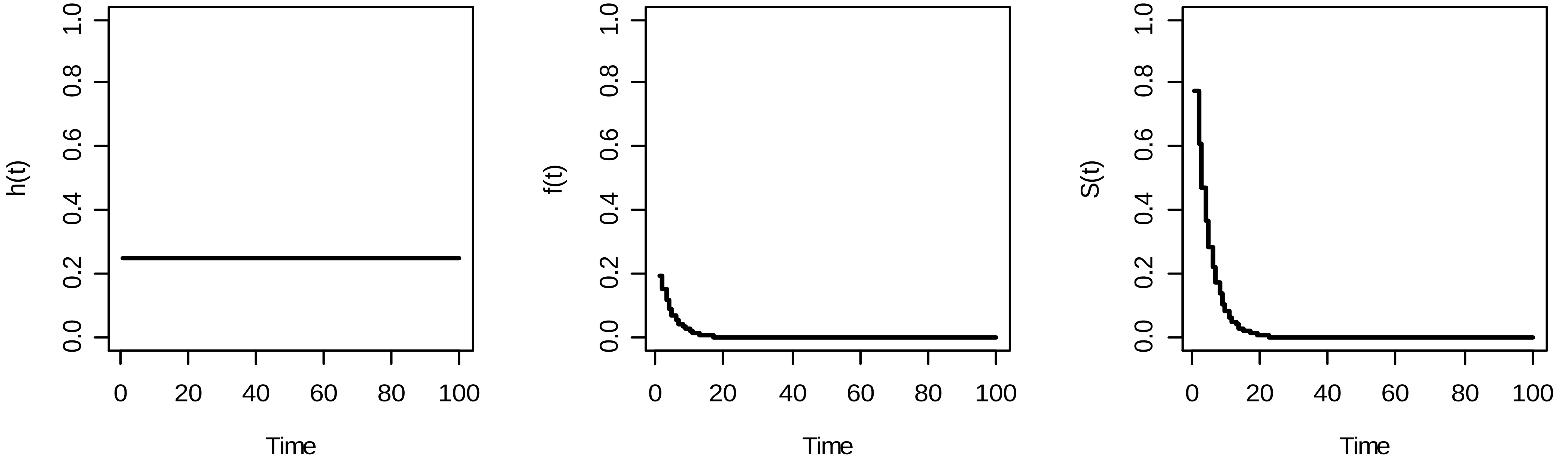
- Several parametric distributions are available, each characterised by different hazard functions:

Distribution	**h(t)**	**f(t)**	**S(t)**
Exponential	λ	$\lambda \exp[-\lambda t]$	$\exp[-\lambda t]$
Weibull	$\lambda p t^{p-1}$	$\lambda p t^{p-1} \exp[-\lambda t^p]$	$\exp[-\lambda t^p]$
Log-logistic	$[ab(at)^{b-1}] / [1 + (at)^b]$	$[ab(at)^{b-1}] / [1 + (at)^b]^2$	$[1 + (at)^b]^{-1}$

$$f(t) = h(t) \times S(t)$$

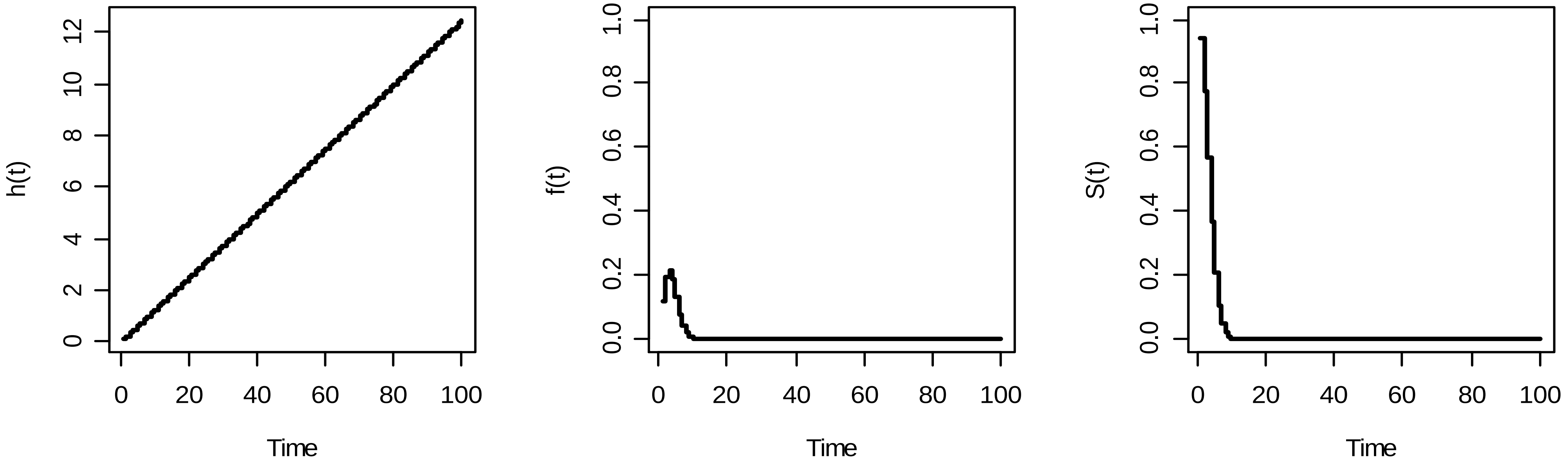
THE UNIVERSITY OF MELBOURNE

Exponential distribution = hazard constant over time.




```
t <- seq(from = 1, to = 100, by = 1)
lambda = 0.25
ht <- lambda
ft <- lambda * exp(-lambda * t)
St <- exp(-lambda * t)
```

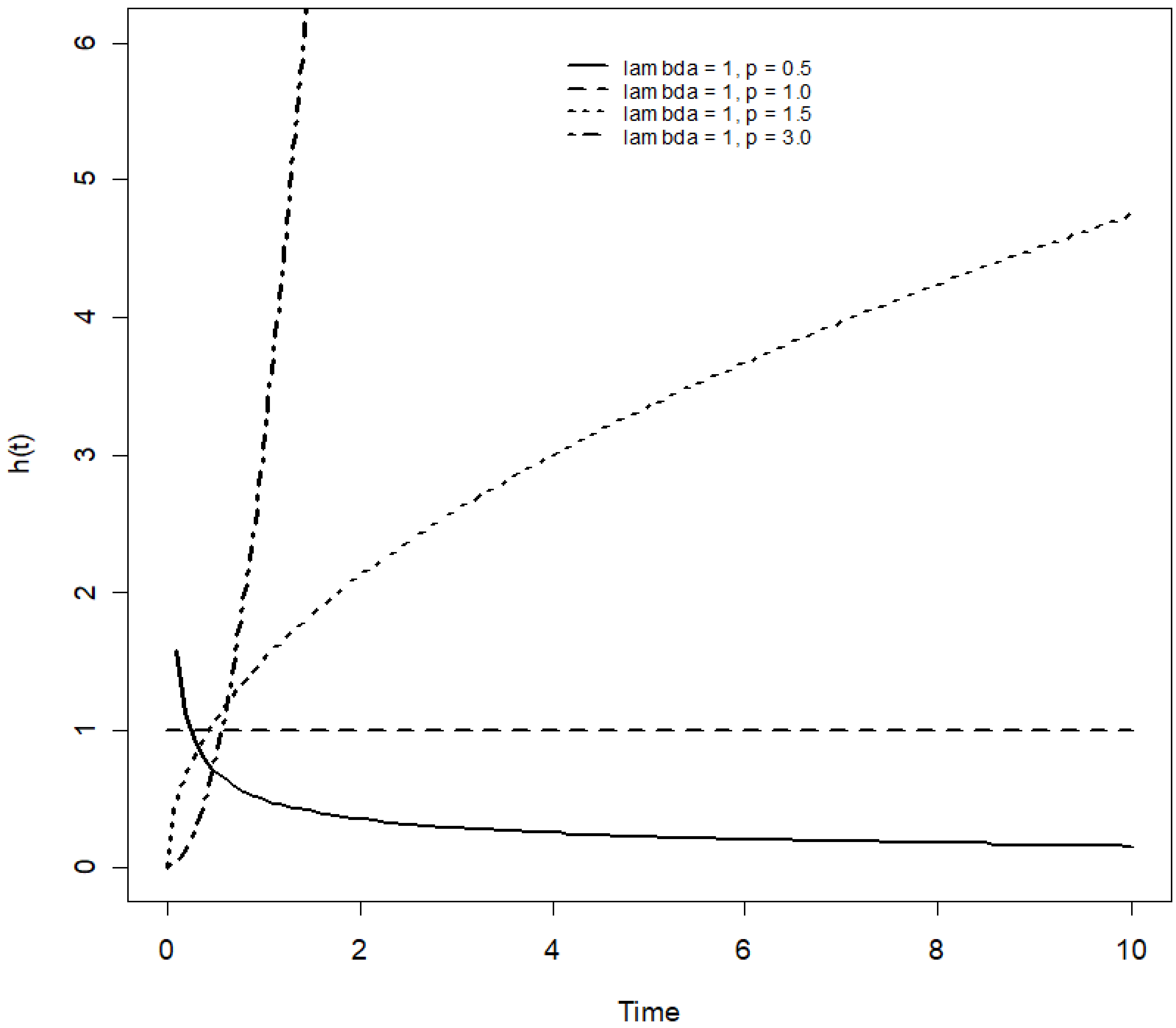
Weibull distribution = hazard varies according to lambda and p.




```
t <- seq(from = 1, to = 100, by = 1)
lambda = 0.25; p = 2
ht <- (lambda * p) * (lambda * t)^(p - 1)
ft <- (lambda * p) * (lambda * t)^(p - 1) * exp(-(lambda * t)^p)
St <- exp(-(lambda * t)^p)
```

Parametric survival functions

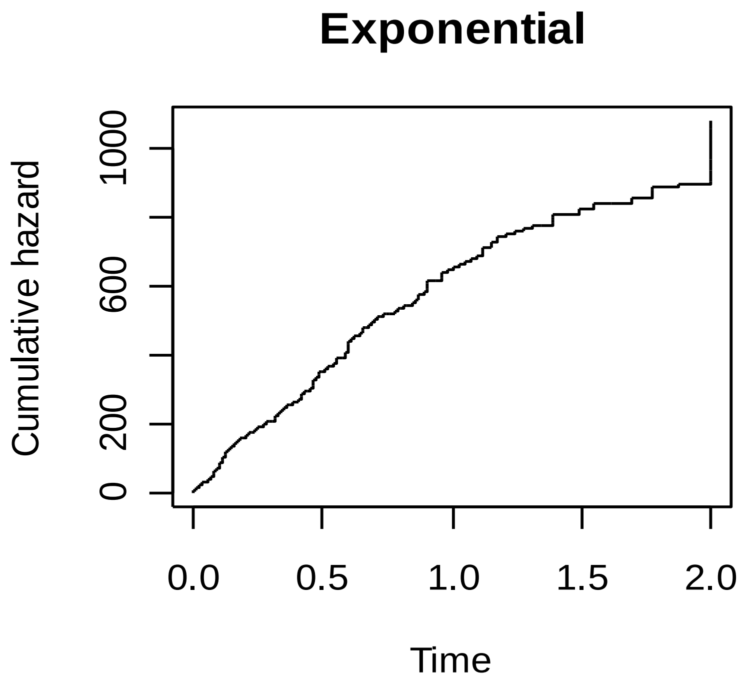
- Exercise
  - for the Weibull distribution, evaluate the effect of changing λ (the scale parameter) and p (the shape parameter)
  - what happens to hazard when
    - $p < 1$?
    - $p = 1$?
    - $p > 1$?

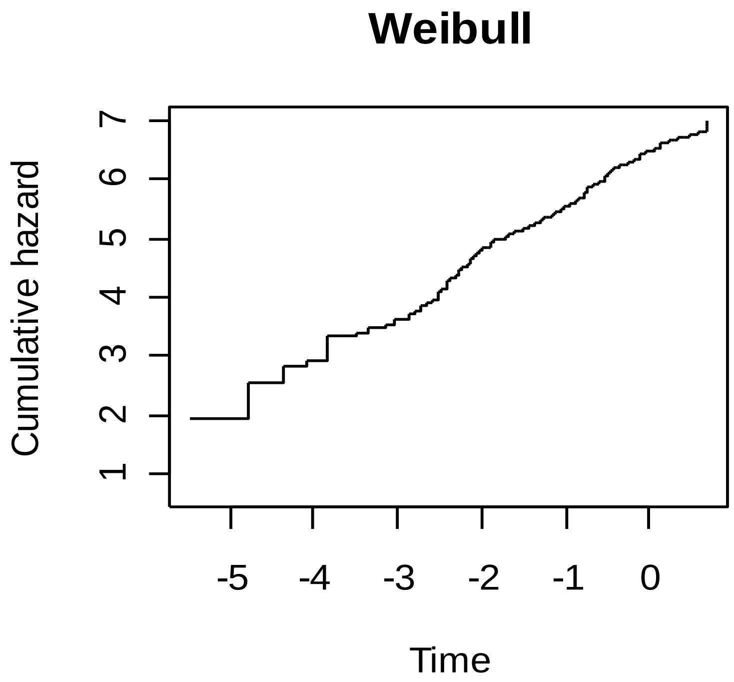


lam bda = 1, p = 0.5
lam bda = 1, p = 1.0
lam bda = 1, p = 1.5
lam bda = 1, p = 3.0
h(t)
0
1
2
3
4
5
6
0
2
4
6
8
10
Time

Parametric survival functions

- Is a parametric distribution appropriate for our data?
  - exponential distribution
    - plot the <u>cumulative hazard</u> as a function of t
    - if distribution appropriate, should be a straight line
  - Weibull distribution
    - plot the <u>log cumulative hazard</u> as a function of log t
    - if distribution appropriate, should be a straight line






```
library(survival); setwd("D:\\TEMP")
dat <- read.table("addict.csv", header = TRUE, sep = ",")
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none",
type = "kaplan-meier", data = dat)
Ht <- -log(addict.km$surv)
t <- addict.km$time

par(mfrow = c(2,2))
plot(Ht, t, type = "s", xlab = "Time", ylab = "Cumulative hazard",
main = "Exponential")
plot(log(Ht), log(t), type = "s", xlab = "Time", ylab = "Cumulative
hazard", main = "Weibull")
```

Parametric survival functions

- Other distributions
  - gamma
  - log-normal
  - log-logistic

  - see Dohoo and Martin page 427

Roadmap

- Survival and hazard
- Parametric survival functions
  - Exponential
  - Weibull
- <u>Non-parametric survival functions</u>
- Presentation

Non-parametric survival functions

- Non-parametric methods are used when no theoretical distribution adequately fits the data
  - less efficient than parametric methods
  - frequently (always?) used in epidemiology
  - two non-parametric methods for estimating the way survival changes over time:
    - the Kaplan-Meier method
    - life table methods

Non-parametric survival functions

- Kaplan-Meier method
  - also known as the Product Limit method
  - based on individual survival times and assumes that censoring is independent of survival time (that is, the reason an observation is censored is unrelated to the cause of failure)

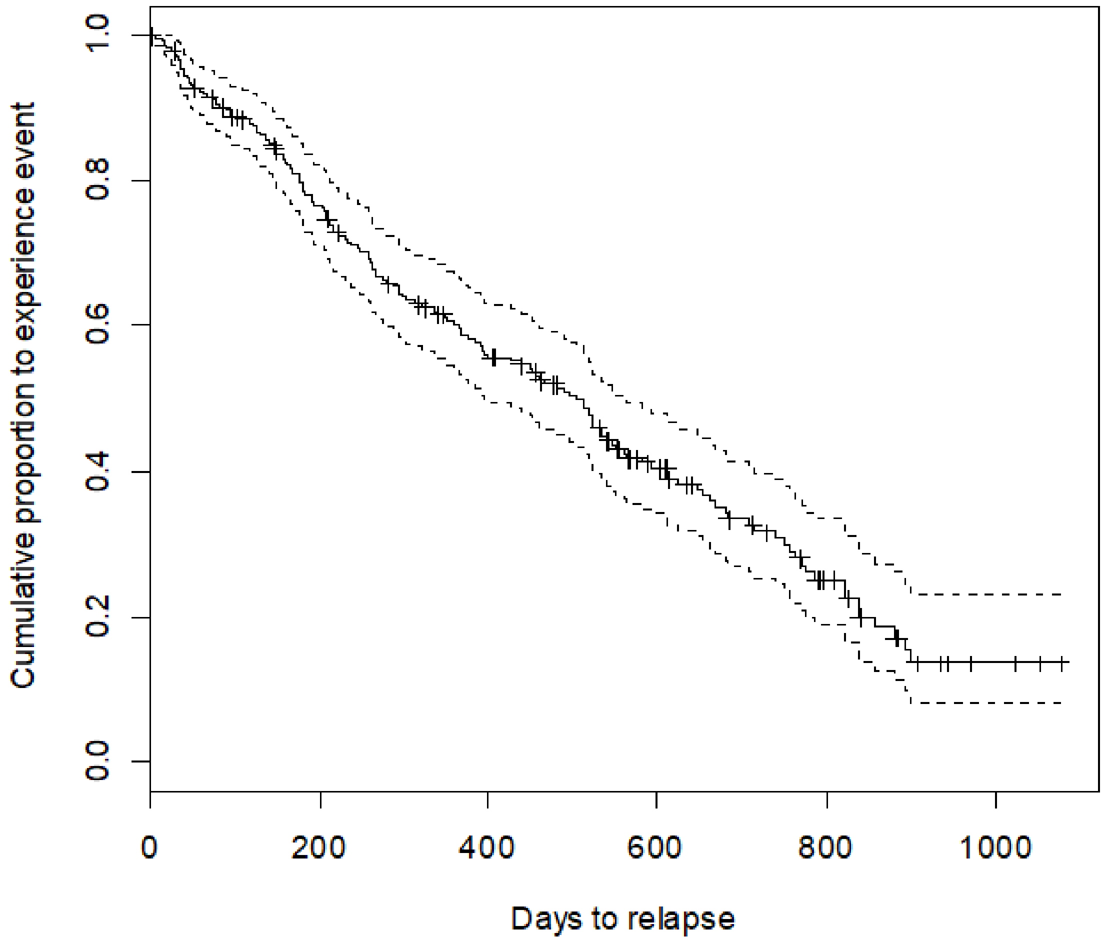
Non-parametric survival functions

- Kaplan-Meier method:

Time	Start n_i	Fail d_i	Censor w_i	At risk $r_i = (n_i - w_i)$	Surv prob $p_i = 1 - [d_i \div (n_i - w_i)]$	Cum surv
0	31	2	3	31-3 = 28	0.93	0.93 × 1.00 = 0.93
1	26	1	2	26-2 = 24	0.96	0.96 × 0.93 = 0.89
2	23	1	2	23-2 = 21	0.95	0.95 × 0.89 = 0.85
3	20	1	2	20-2 = 18	0.94	0.94 × 0.85 = 0.80
etc						

Non-parametric survival functions

- Confidence intervals
  - standard errors of survival estimate can be calculated at each point in time
  - standard errors can be used to calculate confidence intervals
  - standard errors can be used to test for differences in survival among strata




```
library(survival); setwd(("D:\\temp")
dat <- read.table("addict.csv", header = TRUE, sep = ",");

addict.km <- survfit(Surv(stop, status) ~ 1, type = "kaplan-meier",
data = dat);
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative
proportion to experience event", conf.int = TRUE);
```

Non-parametric survival functions

- Life-table (Cutler-Ederer) method

Time	Start ni	Fail di	Censor wi
0 to 1	31	3	4
2 to 3	24	2	4
etc			

Non-parametric survival functions

- Life-table (Cutler-Ederer) method

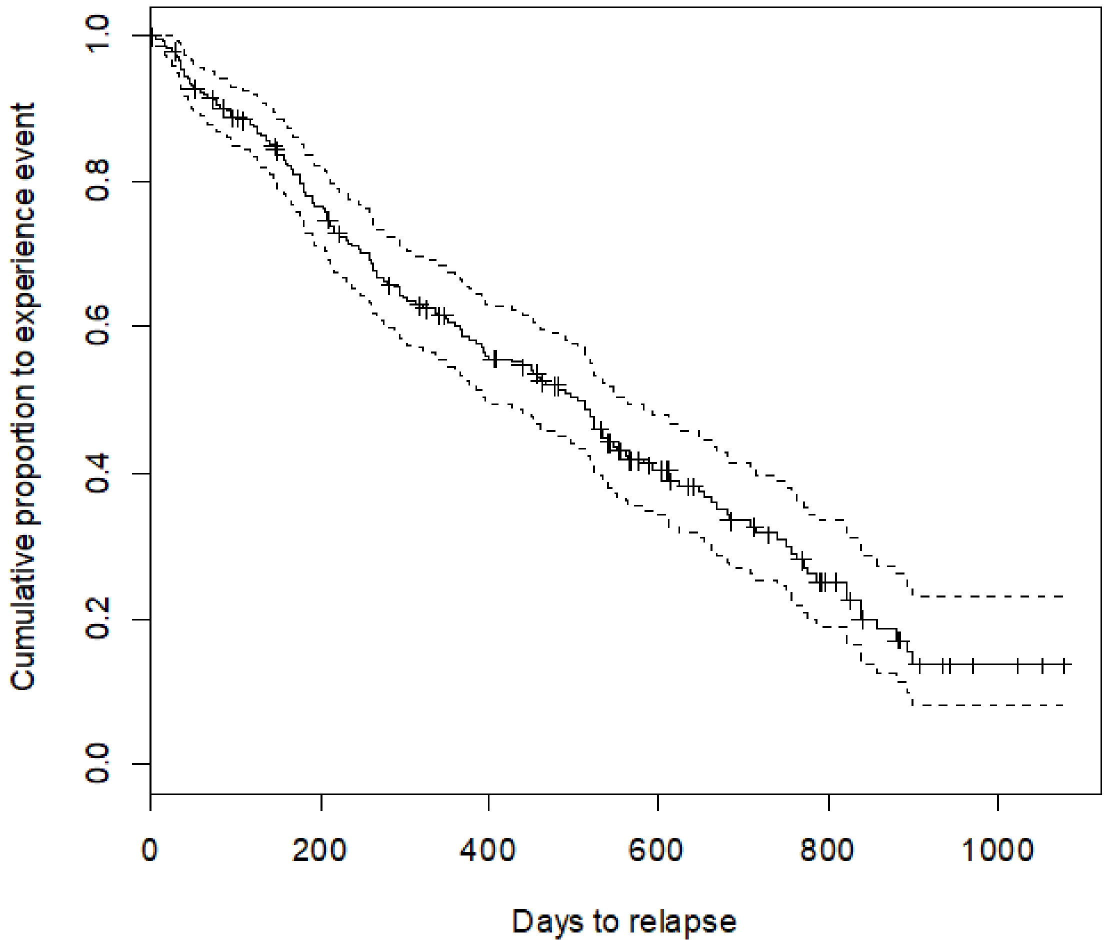
Time	Fail prob $q_i = d_i / [n_i - (w_i/2)]$	Surv prob $p_i = 1 - q_i$	Cum surv
0 to 1	0.10	0.90	0.90
2 to 3	0.09	0.91	$0.91 \times 0.90 = 0.92$
etc			

Non-parametric survival functions

- Life-table (Cutler-Ederer) method
  - assumes that subjects are withdrawn randomly throughout each interval – therefore, on average, they are withdrawn half way through the interval
  - this is not an important issue when the time intervals are short, but bias may introduced when time intervals are long
  - often used to produce life tables from large scale population surveys (e.g. death registers)
  - not often used these days

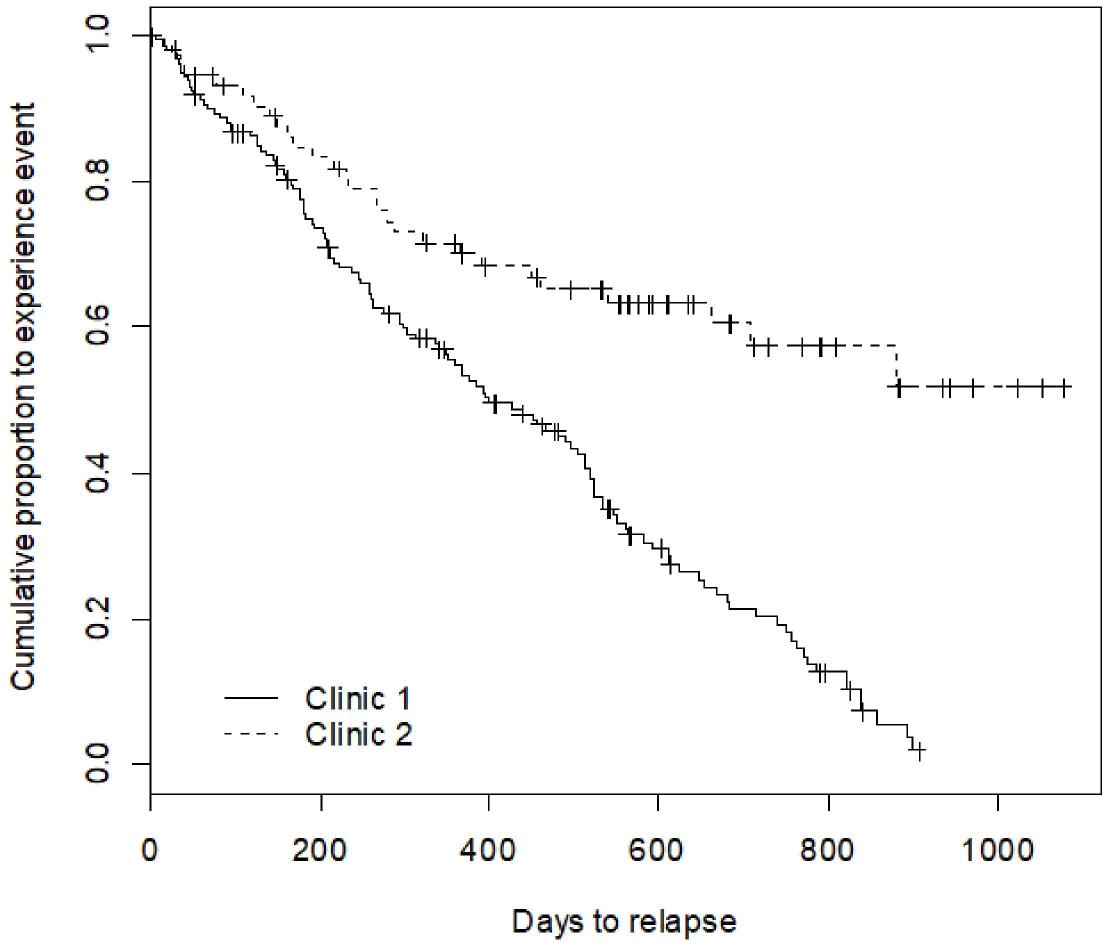
Non-parametric survival functions

- Caplehorn and others (1991)
  - compared retention in two methadone treatment clinics for heroin addicts
  - a patient's ‘survival’ time was determined as the time in days until the patient dropped out of the clinic or was censored at the end of the study
  - the two clinics differed according to their overall treatment policies
  - reference: Caplehorn J et. al. (1991). Methadone dosage and retention of patients in maintenance treatment. Medical Journal of Australia, 154:195 - 199.




```
library(survival); setwd(("D:\\temp")
dat <- read.table("addict.csv", header = TRUE, sep = ",");

addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none",
type = "kaplan-meier", data = dat);
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative
proportion to experience event");
```




```
library(survival); setwd(("D:\\temp")
dat <- read.table("addict.csv", header = TRUE, sep = ",")

addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-
meier", data = dat)
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative
proportion to experience event", lty = c(1,2), legend.text =
c("Clinic 1","Clinic 2"), legend.pos = 0, legend.bty = "n")
```

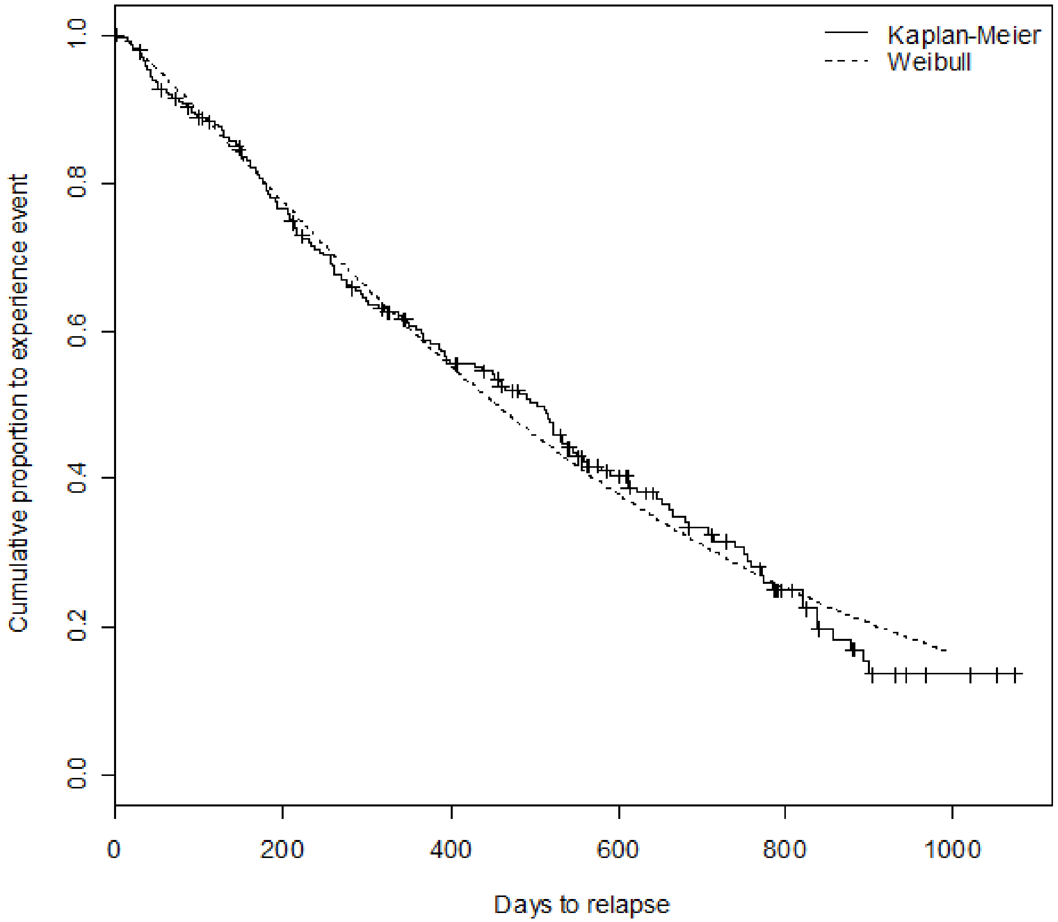
Comparison of Kaplan-Meier and Weibull estimates of survival:

```
library(survival); setwd(("D:\\temp")
dat <- read.table("addict.csv", header = TRUE, sep = ",")

addict.we <- survreg(Surv(stop, status) ~ 1, dist = "weib", data =
dat)
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none",
type = "kaplan-meier", data = dat)
```

Using the Weibull distribution μ (the intercept) = -log(λ) and σ (scale) = 1 / p. Thus the scale parameter λ = exp(-μ) and p = 1 / σ. See Tableman and Kim p 78.

```
p <- 1 / addict.we$scale
lambda <- exp(-addict.we$coeff[1])
t <- 1:1000
St <- exp(-(lambda * t)^p)
addict.we <- as.data.frame(cbind(t = t, St = St))
```




```
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative
proportion to experience event")
lines(addict.we$t, addict.we$St, lty = 2)
legend(x = "topright", legend = c("Kaplan-Meier", "Weibull"), lty =
c(1,2), bty = "n")
```

Non-parametric survival functions

- We conclude that the Weibull distribution provides an adequate fit to the observed data up to day 500, then it tends to underestimate survivorship

Roadmap

- Survival and hazard
- Parametric survival functions
  - Exponential
  - Weibull
- Non-parametric survival functions
- <u>Presentation</u>

Presentation

- Presentation of survival curves:
  - should be vertical step
  - good idea to mark censoring points
  - curves often become unstable at end of study period
  - good idea in figure title to quote number of events and number of censored observations
  - y-axis: "*Cumulative proportion to experience ...*"
  - x-axis: *"PSM to conception (days)"*

Roadmap

- Survival and hazard
- Parametric survival functions
  - Exponential
  - Weibull
- Non-parametric survival functions
- Presentation