STAT40810 — Stochastic Models

Brendan Murphy

Week 5

Time to Event / Survival Data

Time To Event Data

- Time to event data arise when the time until some event occurs is recorded for a number of instances.
- It is also known as survival data (because of the historical origins of this area).
- Statistical methods of survival analysis are applied in a wide range of areas.

Applications

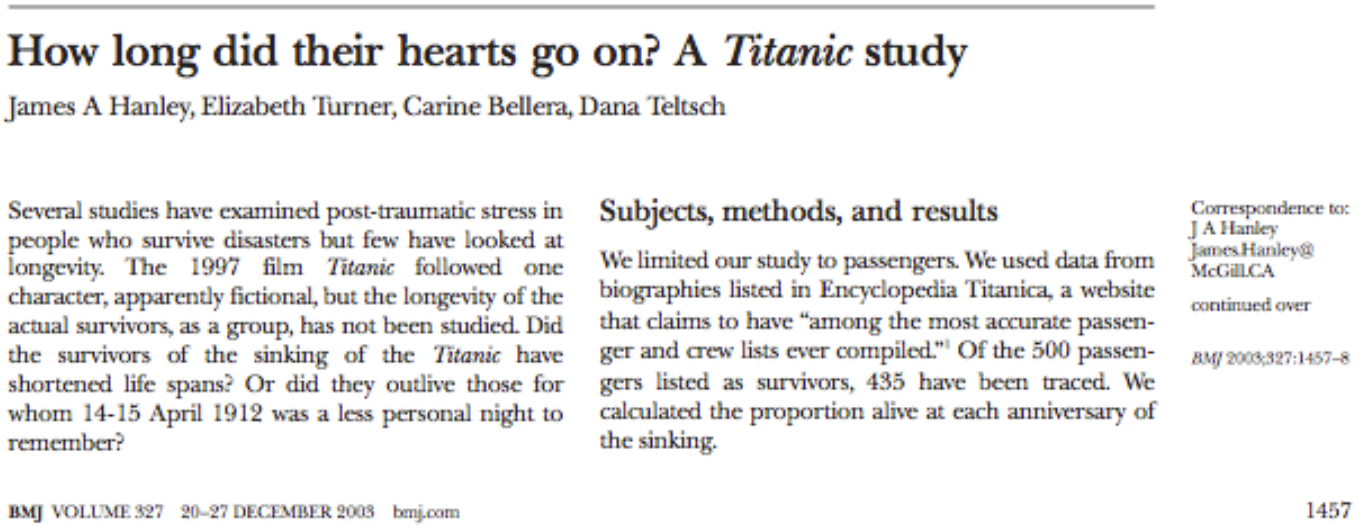
- **Survival analysis** has been applied in a wide range of areas:
  - laboratory studies of animals
  - clinical studies of humans
  - testing the failure of physical components
  - time to learn a particular skill
  - births, marriages, divorces, promotions and arrests.
  - finance
  - social sciences

  Survival times can be:
  - time from start of treatment to death
  - tumour free time
  - time from release from prison to next arrest, etc.
  - time between financial transactions
  - time between social interactions

Data Features

- Survival times are always **positive**.
- A second characteristic of survival data is the frequent occurrence of **covariate** information. For instance, a measure of immune function may be recorded for leukaemia patients who are in remission.
- Survival analysis involves **censoring**. Survival times are not always known exactly. For example, in clinical trials some subjects are still alive at the end of the study and some are 'lost to follow-up'. So all the survival times are not known and exact. These are called **censored observations**.

Titanic Study



How long did their hearts go on? A *Titanic* study

James A Hanley, Elizabeth Turner, Carine Bellera, Dana Teltsch

Several studies have examined post-traumatic stress in people who survive disasters but few have looked at longevity. The 1997 film *Titanic* followed one character, apparently fictional, but the longevity of the actual survivors, as a group, has not been studied. Did the survivors of the sinking of the *Titanic* have shortened life spans? Or did they outlive those for whom 14-15 April 1912 was a less personal night to remember?

Subjects, methods, and results

We limited our study to passengers. We used data from biographies listed in Encyclopedia Titanica, a website that claims to have "among the most accurate passenger and crew lists ever compiled."[1] Of the 500 passengers listed as survivors, 435 have been traced. We calculated the proportion alive at each anniversary of the sinking.

Correspondence to:
J A Hanley
James.Hanley@McGill.CA

continued over

BMJ 2003;327:1457–8

Figure : An interesting survival analysis study from the 2003 Christmas issue of the British Medical Journal.

Transplantation Study

Long-Term Renal Allograft Survival: Have we Made Significant Progress or is it Time to Rethink our Analytic and Therapeutic Strategies?

Herwig-Ulf Meier-Kriesche*, Jesse D. Schold and Bruce Kaplan

Department of Medicine, University of Florida, Gainesville, FL, USA
**Corresponding author: Herwig-Ulf Meier-Kriesche, meierhu@medicine.ufl.edu*

Impressive renal allograft survival improvement between 1988 and 1995 has been described using projections of half-lives based on limited actual follow up. We aimed, now with sufficient follow up available to calculate real half-lives.

Real half-lives calculated from Kaplan-Meier curves for the overall population as well as subsets of repeat transplants and African Americans recipients were examined.

Introduction

Survival analysis in renal transplantation have traditionally been made by Kaplan-Meier estimates, because of the necessity to evaluate data on outcomes before all patients reach a certain follow-up time (1). Each patient contributes to the survival estimate only up to the time they have available follow up, so that with increasing length of follow up progressively fewer patients contribute to the survival estimate. In this sense Kaplan-Meier survival estimates are only an estimation of the real survival outcome of the entire study population. In fact in order to make somewhat reliable deductions from Kaplan-Meier survival estimate comparisons, sufficient patients have to be at follow up at the point of analysis.

A calculated half-life is a projection of survival which has many assumptions in itself, most importantly that the slope

Figure : A survival analysis study from the American Journal of Transplantation.

Food Science

Determination of the End of Shelf-life for Milk using Weibull Hazard Method

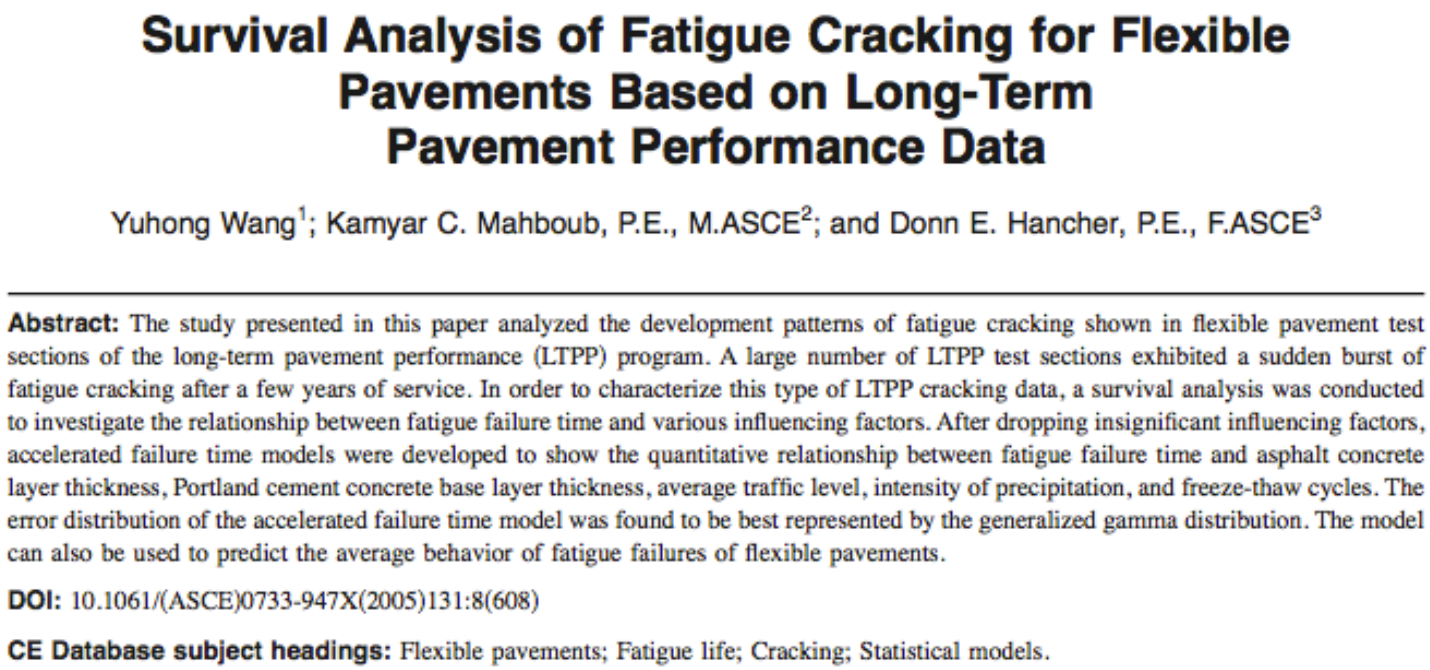
W. S. Duyvesteyn, E. Shimoni and T. P. Labuza*

W. S. Duyvesteyn: The Coca-Cola Company, Atlanta, GA 30301 (U.S.A.)
E. Shimoni, T. P. Labuza: Department of Food Science and Nutrition, University of Minnesota, St. Paul, MN 55108 (U.S.A.)

(Received October 6, 1999; accepted November 17, 2000)

The shelf life of pasteurized milk is traditionally estimated by the counts of both total and psychrotrophic microbial load. This study examins the relationship between the total and psychrotophic microbial growth in milk and its sensory shelf life as measured using the Weibull hazard method. Milk was stored at five constant temperatures (2, 5, 7, 12 and 15 °C) and both total and psychrotrophic microbial counts were used to obtain the lag time and the growth rate values. The lag time of the total and psychrotrophic growth responded to temperature following the Arrhenius equation. The loss of sensory quality of the milk followed a log shelf life vs. *temperature dependency. It was found that there was no correlation between the microbial count at the end of shelf life and the sensory quality of the milk. It is therefore suggested that microbial counts should not be used to determine the sensory shelf life of milk. The Weibull method gave end of shelf life values fairly similar to that of prior work using the American Dairy Science Association scoring method.*

Figure : A survival analysis study from LWT - Food Science and Technology.



Survival Analysis of Fatigue Cracking for Flexible Pavements Based on Long-Term Pavement Performance Data

Yuhong Wang[1]; Kamyar C. Mahboub, P.E., M.ASCE[2]; and Donn E. Hancher, P.E., F.ASCE[3]

Abstract: The study presented in this paper analyzed the development patterns of fatigue cracking shown in flexible pavement test sections of the long-term pavement performance (LTPP) program. A large number of LTPP test sections exhibited a sudden burst of fatigue cracking after a few years of service. In order to characterize this type of LTPP cracking data, a survival analysis was conducted to investigate the relationship between fatigue failure time and various influencing factors. After dropping insignificant influencing factors, accelerated failure time models were developed to show the quantitative relationship between fatigue failure time and asphalt concrete layer thickness, Portland cement concrete base layer thickness, average traffic level, intensity of precipitation, and freeze-thaw cycles. The error distribution of the accelerated failure time model was found to be best represented by the generalized gamma distribution. The model can also be used to predict the average behavior of fatigue failures of flexible pavements.

DOI: 10.1061/(ASCE)0733-947X(2005)131:8(608)

CE Database subject headings: Flexible pavements; Fatigue life; Cracking; Statistical models.

Figure : A survival analysis study from the Journal of Transportation Engineering.

Modelling employee withdrawal behaviour over time: A study of turnover using survival analysis

Mark John Somers*

School of Industrial Management, New Jersey Institute of Technology, 323 King Boulevard, Newark, NJ 07012, USA

and

PhD Program in Management, Rutgers University

Survival analysis techniques were used to test a model of turnover with a sample of 244 staff nurses. Estimates of survival and hazard functions indicated that withdrawal was not uniform over time, but rather occurred in distinct waves. Formal hypotheses were tested with a regression analogue of survival analysis, proportional hazards regression, and provided little support for a reasoned action model of turnover. Specifically, job satisfaction emerged as predictive of turnover while job search behaviour did not. Results from OLS and logistic regressions were consistent with prior research findings in that job search behaviour was a strong influence of employee turnover. Implications of these findings and directions for future research are discussed.

Figure : A survival analysis study from the Journal of Occupational and Organizational Psychology‘.

Survival function

- There are two important functions in survival analysis: the *survival function* and the *hazard function*.
- Let t be an observed value of the random variable $T \geq 0$ (where T denotes survival time) with density $f(t)$ and distribution function $F(t) = \mathbb{P}\{T \leq t\}$.
- The **survival function** $S(t)$ is the probability that the survival time is greater than t:

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= \mathbb{P}\{T > t\} \\ &= \mathbb{P}\{\text{an individual survives longer than t}\} \end{aligned}$$

Survival Function (cont.)

- So

$$\begin{aligned} S(t) &= 1 \text{ for } t = 0 \\ &= 0 \text{ as } t \to \infty. \end{aligned}$$

- Recall

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

- And

$$S(t) = \int_t^\infty f(s)ds.$$

Hazard Function

- The **hazard function** or **hazard rate** $h(t)$ is the risk of death at time t.
- It can be thought of as being related to the probability of an event at time t conditional on having survived (no event) up to time t.
- It can be defined as

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t).$$

- The hazard function has the interpretation

$$\begin{aligned} h(t)dt &\approx \mathbb{P}\{(t < T \leq t + dt) | T > t\} \\ &= \mathbb{P}\{\text{expiring in the interval } (t, t + dt] | \text{ survived past time } t\} \end{aligned}$$

Example: Exponential

- Suppose that $T \sim \text{Exponential}(\lambda)$.
- We have the following quantities.
  - Density:
$$f(t) = \lambda \exp(-\lambda t).$$
  - CDF:
$$F(t) = 1 - \exp(-\lambda t).$$
  - Survival:
$$S(t) = \exp(-\lambda t).$$
  - Hazard:
$$h(t) = \lambda.$$

Example: Gamma

- Suppose that $T \sim \text{Gamma}(\alpha, \beta)$ (Shape=α, Scale=β)
- We have the following quantities.
  - Density:
$$f(t) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} t^{\alpha-1} \exp\left(-\frac{t}{\beta}\right).$$
  - CDF:
$$F(t) = \text{Doesn't have a closed form expression}$$
  - Survival:
$$S(t) = \text{Doesn't have a closed form expression}$$
  - Hazard:
$$h(t) = \frac{f(t)}{1 - F(t)}$$

Example: Gamma

- The density can be computed using the `dgamma()` command.
- The CDF can be computed using the `pgamma()` command.
- Thus, we can compute the hazard too.
- The hazard function:
  - increases monotonically if $\alpha > 1$
  - decreases if $\alpha < 1$
  - $\lim_{t \to \infty} h(t) = 1/\beta$.

Example: Gamma Parameterization

- An alternative, but equivalent, formulation of the gamma distribution has density:

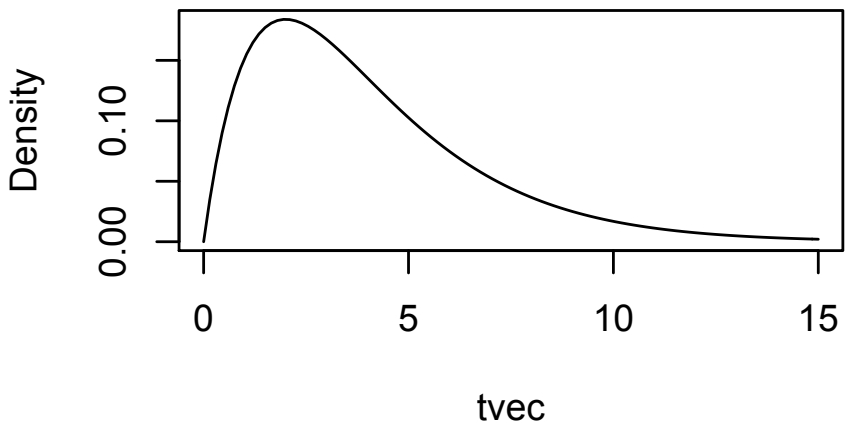
$$f(t) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\lambda t).$$

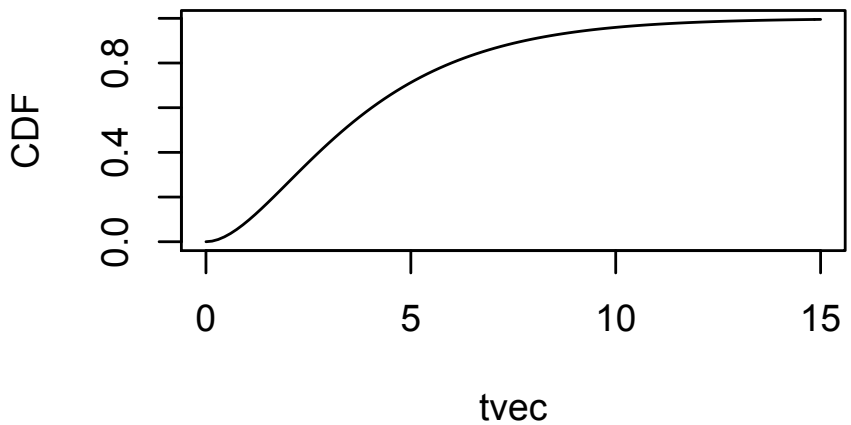
- In this case, we say that the shape$=\alpha$ and the rate$=\lambda$.
- So, rate is just the reciprocal of scale.

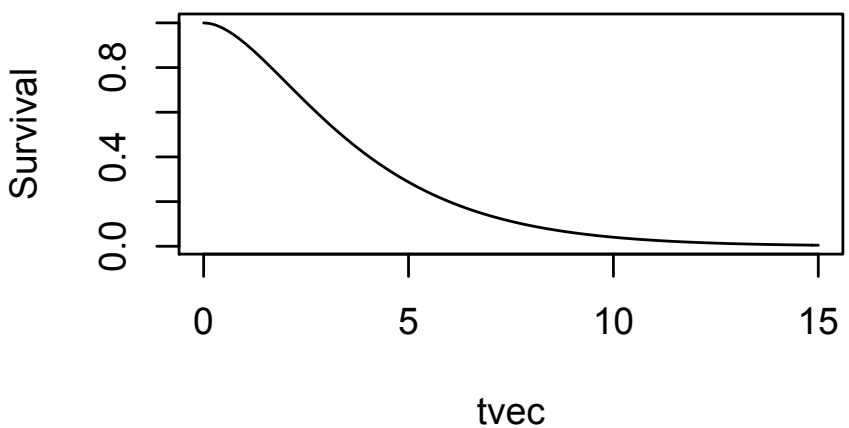
R Code

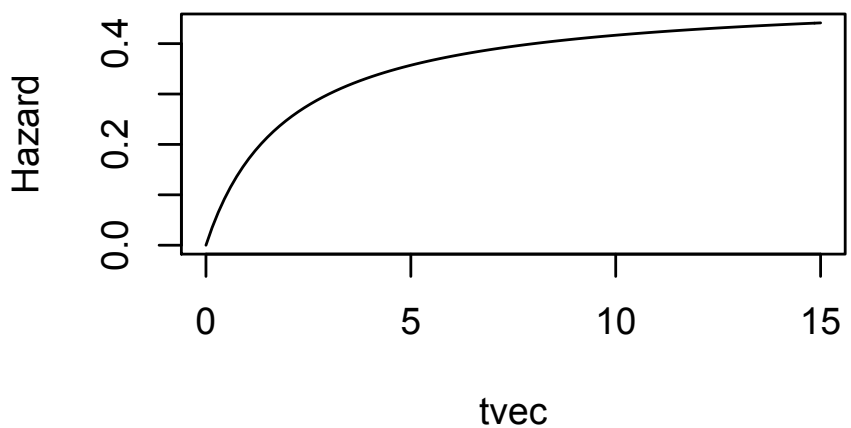
```
# Set up a grid of time values
tvec <- seq(0,15,length=101)

# Set the parameters of the distribution
alpha <- 2
beta <- 2

# Plot the density, CDF, Survival and hazard
par(mfrow=c(2,2))
# Density
plot(tvec,dgamma(tvec,shape=alpha,scale=beta),type="l",ylab="Density")

# CDF
plot(tvec,pgamma(tvec,shape=alpha,scale=beta),type="l",ylab="CDF")

# Survival
plot(tvec,1-pgamma(tvec,shape=alpha,scale=beta),type="l",ylab="Survival")

# Hazard
plot(tvec,dgamma(tvec,shape=alpha,scale=beta)/(1-pgamma(tvec,shape=alpha,scale=beta)),type="l"
  ,ylab="Hazard")

par(mfrow=c(1,1))
```

Plots: Gamma


Density
0.00
0.10
0
5
10
15
tvec


CDF
0.0
0.4
0.8
0
5
10
15
tvec


Survival
0.0
0.4
0.8
0
5
10
15
tvec


Hazard
0.0
0.2
0.4
0
5
10
15
tvec

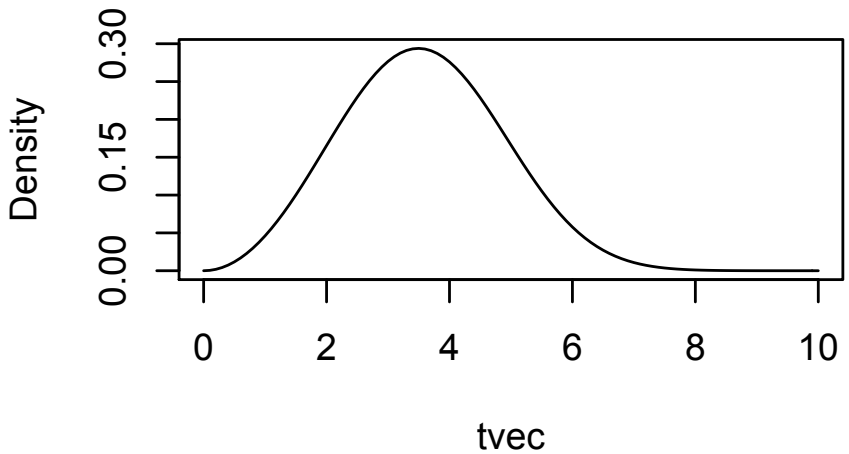
Example: Weibull

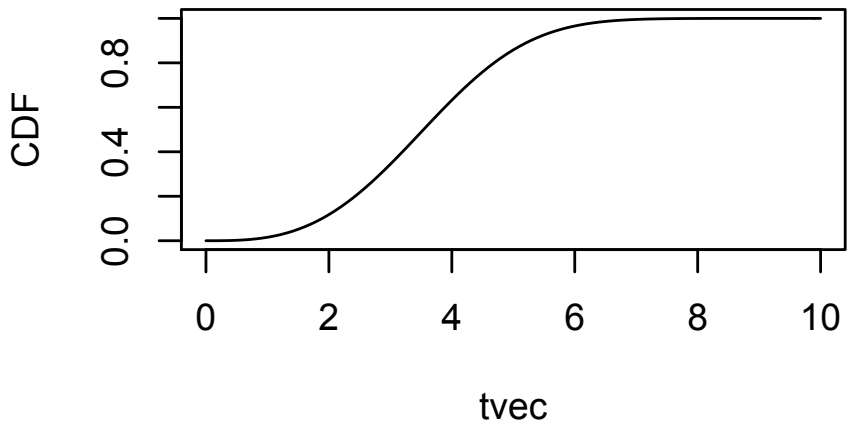
- Suppose that $T \sim \text{Weibull}(\alpha, \beta)$ (Shape=α, Scale=β)
- We have the following quantities.
  - Density:
$$f(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left[\left(-\frac{t}{\beta}\right)^{\alpha}\right].$$
  - CDF:
$$F(t) = 1 - \exp\left[\left(-\frac{t}{\beta}\right)^{\alpha}\right].$$
  - Survival:
$$S(t) = \exp\left[\left(-\frac{t}{\beta}\right)^{\alpha}\right]$$
  - Hazard:
$$h(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}$$

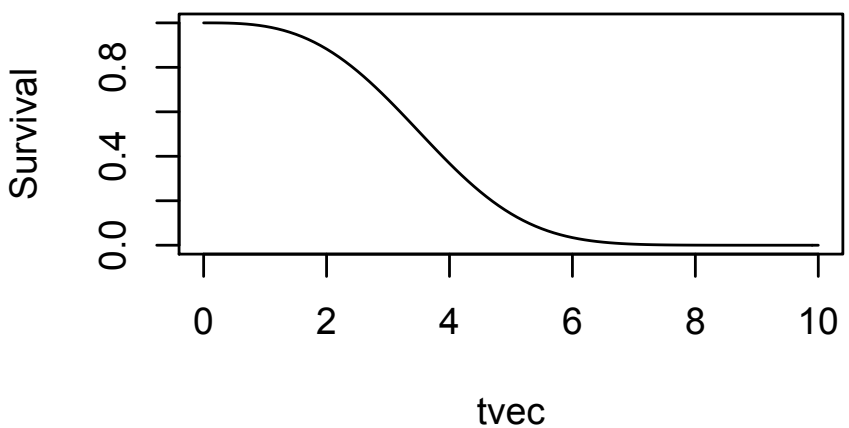
R Code

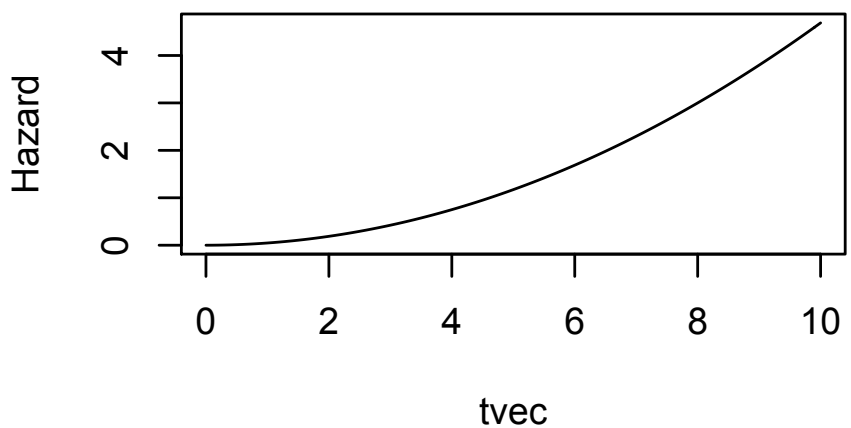
```
# Set up a grid of time values
tvec <- seq(0,10,length=101)

# Set the parameters of the distribution
alpha <- 3
beta <- 4

# Plot the density, CDF, Survival and hazard
par(mfrow=c(2,2))
# Density
plot(tvec,dweibull(tvec,shape=alpha,scale=beta),type="l",ylab="Density")

# CDF
plot(tvec,pweibull(tvec,shape=alpha,scale=beta),type="l",ylab="CDF")

# Survival
plot(tvec,1-pweibull(tvec,shape=alpha,scale=beta),type="l",ylab="Survival")

# Hazard
plot(tvec,dweibull(tvec,shape=alpha,scale=beta)/(1-pweibull(tvec,shape=alpha,scale=beta)),type="l"
   ,ylab="Hazard")

par(mfrow=c(1,1))
```

Plots: Weibull









Inference for Models

- In principle, we can fit any parametric model to survival data (if we have no censoring).
- We form the log-likelihood function and maximize it to find the MLE.
- We can also find confidence intervals for the parameters using the Hessian of the log-likelihood.

Example: Machine Failures

- The time (in days) between failures of a manufacturing machine were recorded.

  79 105 14 153 67 25 39 9 55 132 48 570

- We model the data as coming from an exponential distribution.
- We know that

$$\hat{\lambda} = \frac{1}{\overline{x}} = 0.00926$$

- We also know that

$$SE(\hat{\lambda}) \approx \frac{\hat{\lambda}}{\sqrt{n}} = 0.00267$$

- So, an approximate 95% confidence interval for λ is $(0.00392, 0.0146)$

R Code: Numerical Optimization

```
# Read in Failure Times
x <- scan()
79 105  14 153  67  25  39   9  55 132  48 570

# Form log-likelihood
logl <- function(lambda)
{
sum(dexp(x,lambda,log=TRUE))
}

# Initial parameter
lambda0 <- 1/mean(x)

#Optimization
fit <- optim(lambda0,logl,control=list(fnscale=-1),method="BFGS",hessian=TRUE)

#Parameters
lambdahat <- fit$par
SE <- sqrt(-1/fit$hess[1,1])

#95% CI
print(c(lambdahat-2*SE,lambdahat+2*SE))
```

Example: Machine Failures

- Suppose instead of observing the actual failure times, we only observe:

$$Y_i = \begin{cases} 1 & \text{if } T_i > 50 \\ 0 & \text{if } T_i \leq 50 \end{cases}$$

- That is, we observe

  `1 1 0 1 1 0 0 0 1 1 0 1`

- This is an extreme form of **censoring**.
- We want to model the **original** data as coming from an exponential distribution.
- Can we do it?

Approach 1: Indirectly estimating λ

- We know that
$$\mathbb{P}\{T_i > 50\} = \exp(-50\lambda) = p$$
.
- We know that
$$\mathbb{P}\{Y_i = 1\} = p \text{ and } \mathbb{P}\{Y_i = 0\} = (1 - p).$$

- So, we have a series of Bernoulli trials and the total number of trials is binomial(n, p)
- We can find the MLE for p and then solve for λ.
- So,
$$\hat{p} = \frac{7}{12} = 0.583$$

- And,
$$\hat{\lambda} = \frac{\log \hat{p}}{-50} = 0.0108$$
.

Approach 2: Directly estimating λ

- We know that

$$\mathbb{P}\{T_i > 50\} = \exp(-50\lambda)$$

and

$$\mathbb{P}\{T_i \leq 50\} = 1 - \exp(-50\lambda)$$

- So, the likelihood function for the data is

$$\prod_{i=1}^{n} [\exp(-50\lambda)]^{y_i} [1 - \exp(-50\lambda)]^{1-y_i}$$

- Maximizing this (or the log) with respect to λ gives

$$\hat{\lambda} = 0.0108 \text{ and } SE(\hat{\lambda}) = 0.0048$$

- So, an approximate 95% confidence interval for λ is $(0.0011, 0.0205)$
- Note, that the interval is wider because we lost information in the censoring.