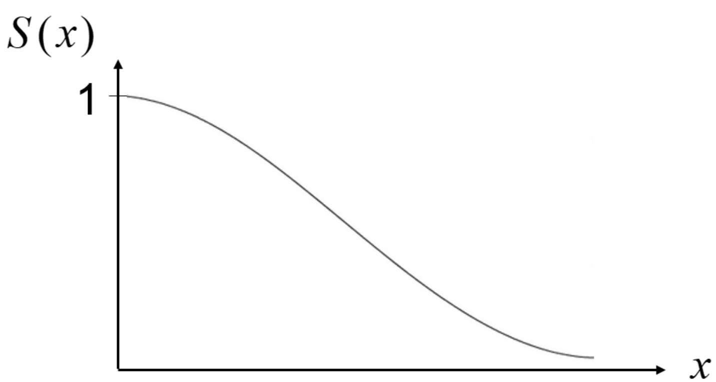
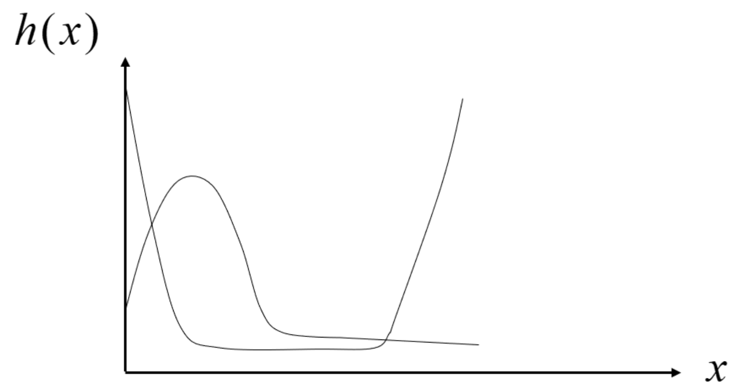
Analysis of Survival Data

WELCOME!





Lecture 1: Introduction

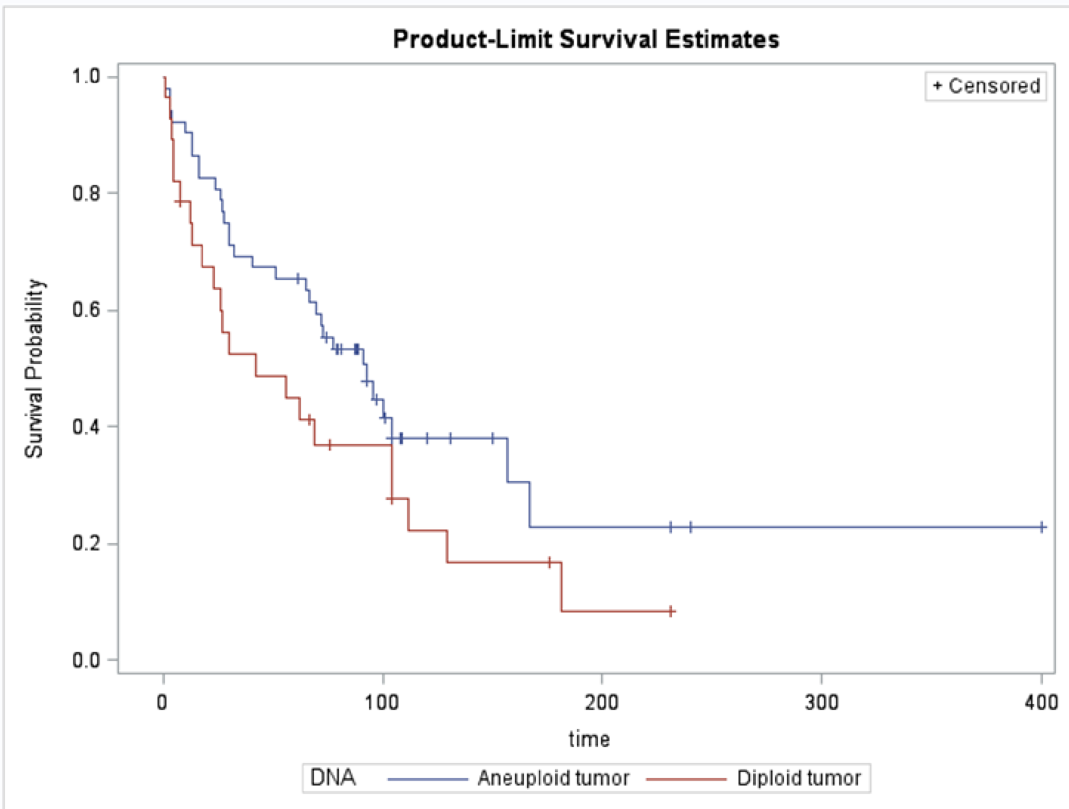
Inger Persson

Program L1

- **Survival analysis - introduction**
  - What is this course about?
  - The “flipped classroom”
- **Basic quantities and models**
  - Survival data
  - Basic concepts in Survival analysis
  - Common parametric models for survival data
  - Regression models for survival data

What is this course about?

- **Estimating survival probabilities**



UPPSALA UNIVERSITET

What is this course about?

- **Proportional hazards regression**

$$h(t \mid \mathbf{Z}) = h_0(t)c(\boldsymbol{\beta}^t\mathbf{Z})$$

UPPSALA UNIVERSITET

Learning outcomes

After this course you should:

- have basic knowledge of the most common methods for the analysis of survival data
- be able to apply these methods on data from the medical field and other subject areas
- be able to assimilate the content of scientific articles based on survival data
- have the ability to both orally and in written form account for results of analyses based on methods for survival data

Why SAS?

We'll be using the SAS software, and there are two main resons for doing so:

- Future employers in biostatistics use SAS (as it is the only "validated" statistical software and as such fulfils requirements of e.g. medical agencies)
- Some of the functions we'll be using are not (yet) implemented in R

UPPSALA UNIVERSITET

Examination

- **Home assignments (using SAS)**
  1) Estimation of basic quantities and hypothesis testing
  2) Cox proportional hazards regression
  3) Article review
  4) Oral presentation

- **Final exam (pen and paper)**

  See Ladok for date and registration

UPPSALA UNIVERSITET

Grades

You will get credits for the home assignment part and the in-class examination separately, but the course will be marked as complete only if you pass all parts.

To **pass** with distinction (VG), you need to score 90% on the exam (60% for G), including any bonus points from home assignments and attendance.

Bonus points from home assignments:
If you pass on the first try you get 2 bonus points (a total of 4/6* bonus points if you pass all assignments on the first try)

Any bonus points will be added to your exam score.

* Math students: 2 home assignments (4 possible bonus points). One less task on the exam.
* Stat students: 3 home assignments (6 possible bonus points).

Flipping the classroom

	At home	In-class
Traditional lectures	Solve problems Active Alone	Listen to lecture Passive With teacher
Flipped classroom	Listen to lecture Passive Alone	Solve problems Active With teacher
Flipped classroom + technology	Interactive lecture Active Alone	Solve problems Active With peers With teacher

Presentation borrowed from David Black-Schaffer, IT dept. UU

”Interactive lecture”

Short lecture segments + self-assessment quizzes

Max 5-10 minutes video, then a quiz

Online lectures posted as Assignments at Studium
(most web browsers except Safari should work)

UPPSALA UNIVERSITET

The ”flipped classroom”

This only works if you prepare, and show up in class!

Bonus points for attendance:

- Watch all online lectures on time: 2 bonus points
- Attend 6/7 of the in-class meetings: 2 bonus points

UPPSALA UNIVERSITET

Why?

How much do we remember?

Lectures

Remarkable lectures

Video lectures (stop, rewind, repeat)

Video lectures with built-in quizzes

Discuss the content with someone else

Explain the content to someone else

UPPSALA UNIVERSITET

In-class sessions

- **Online lecture follow-up**
  Any parts from the online lectures that you had trouble understanding
- **Review questions**
  Based on the content from online lecture and book
  IPA (individually, peers, all)
- **Exercises**
  Pen and paper or SAS

UPPSALA UNIVERSITET

Basic rules

- **Be on time!**

  I will start on time, and I expect you to be here when we start.

  Our time is limited and valuable, it is very disturbing to me and to your fellow class mates if you drop in late.

- **Keep quiet (when I talk)**

  When you have questions, raise them out loud.

- **Interrupt at any time**

  Questions are always welcome!

UPPSALA UNIVERSITET

Survival data

Data involving time to the occurrence of a certain event.

Examples of events:

- Death (which accounts for the name)
- The appearance of a tumor
- The development of some disease
- Conception (pregnancy)
- Cessation of smoking (quitting)

UPPSALA UNIVERSITET

Survival data

Applications of statistical methods for survival data analysis have been extended beyond biomedical research to other fields, e.g.:

- Lifetime of electronic devices (reliability engineering)
- Felon’s time to parole (criminology)
- Duration of first marriage (sociology)
- Length of app subscription (marketing)
- Worker’s compensation claims (insurance)
- etc…

Survival analysis

Survival analysis encompasses a wide variety of methods for analyzing the timing of events.

Terminology varies:

- **Survival analysis** in biostatistics, which has the richest tradition in this area
- **Failure-time analysis** (or reliability analysis) in engineering
- **Event-history analysis** in sociology
- **Time-to-event analysis** more generally

UPPSALA UNIVERSITET

Survival analysis

X = time to some specified event

$X \geq 0$

Different ways of describing X:

1) Survival function $S(x)$
2) Probability density function $f(x)$
3) Hazard function (rate) $h(x)$
4) Cumulative Hazard function $H(x)$
5) Mean Residual Life and Median lifetime

1) Survival function

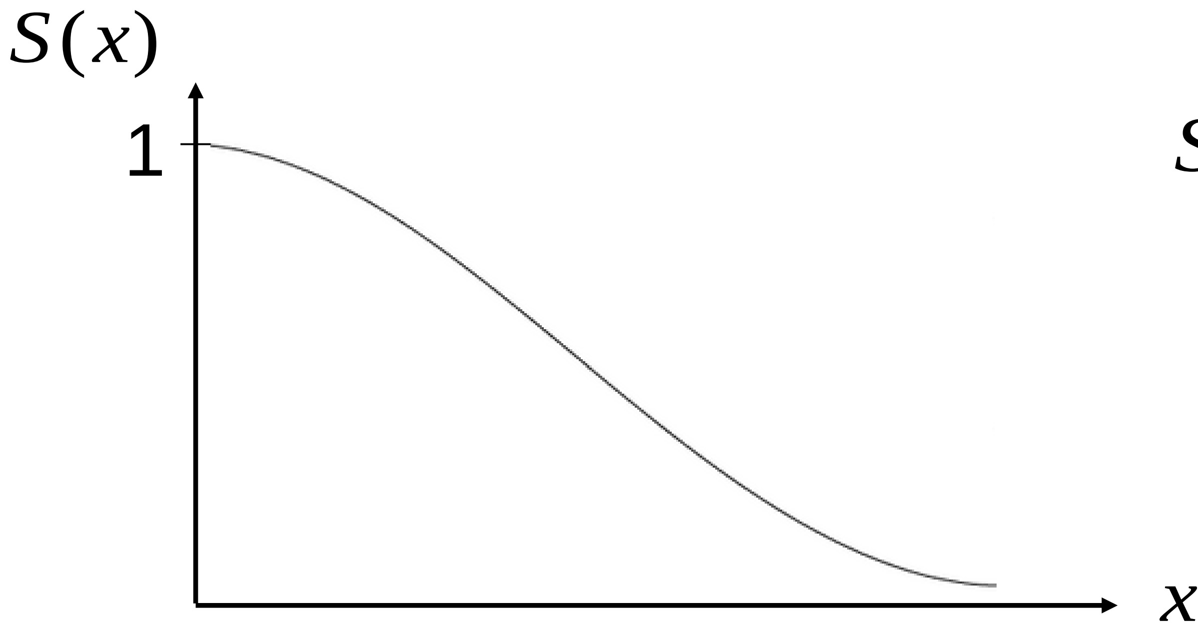
$$S(x) = \Pr(X > x) = 1 - F(x)$$

$$F(x) = \Pr(X \le x)$$

The probability to experience the event beyond a certain time, *x*.

E.g. the probability to survive beyond time *x*.

1) Survival function



$$S(x) = \int_x^\infty f(t)dt$$

Probability density function

A non-increasing function, equal to 1 at time 0 and equal to 0 as time approaches infinity.

UPPSALA UNIVERSITET

2) Probability density function

$$f(x) = -\frac{dS(x)}{d(x)}$$

A nonnegative function, with the area under *f(x)* equal to 1.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$S(x) = \int_{x}^{\infty} f(t)dt$$

Example: All cause mortality in the US

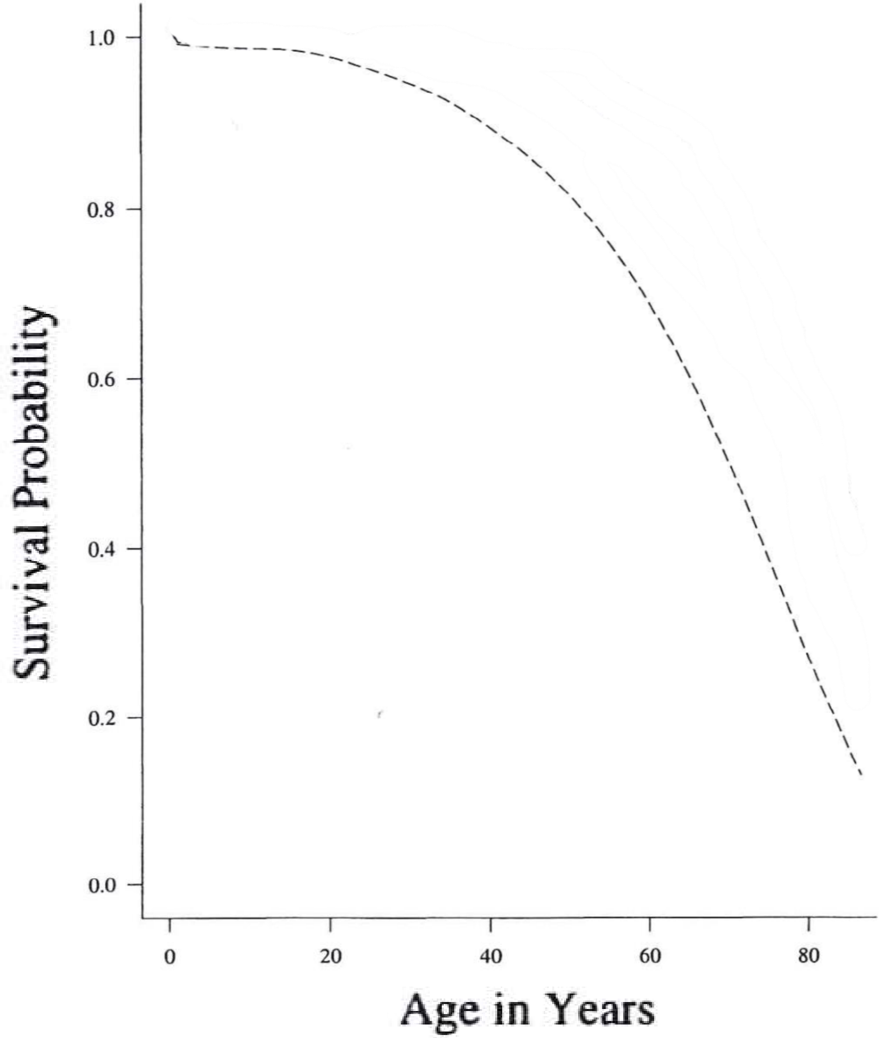


Figure 2.2 *Survival Functions for all cause mortality for the US population in 1989. White males (————); white females (······) black males (------) females (—— ——).*

Survival function for discrete X

The previous definitions hold when X is a continuous random variable.

When X is a discrete random variable:

$$S(x) = \Pr(X > x) = \sum_{x_j > x} p(x_j)$$

Probability mass function

$$p(x_j) = \Pr(X = x_j)$$

3) Hazard function (intensity function)

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X < x + \Delta x \mid X \ge x)}{\Delta x}$$

Conditional on the fact that you still haven't experienced the event at time x.

$h(x)\,\Delta x$ = "approximate" probability of an individual to experience the event in the next instant ($x + \Delta x$).

Describes how the probability of experiencing the event changes over time.

3) Hazard function (intensity function)

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X < x + \Delta x \mid X \ge x)}{\Delta x}$$

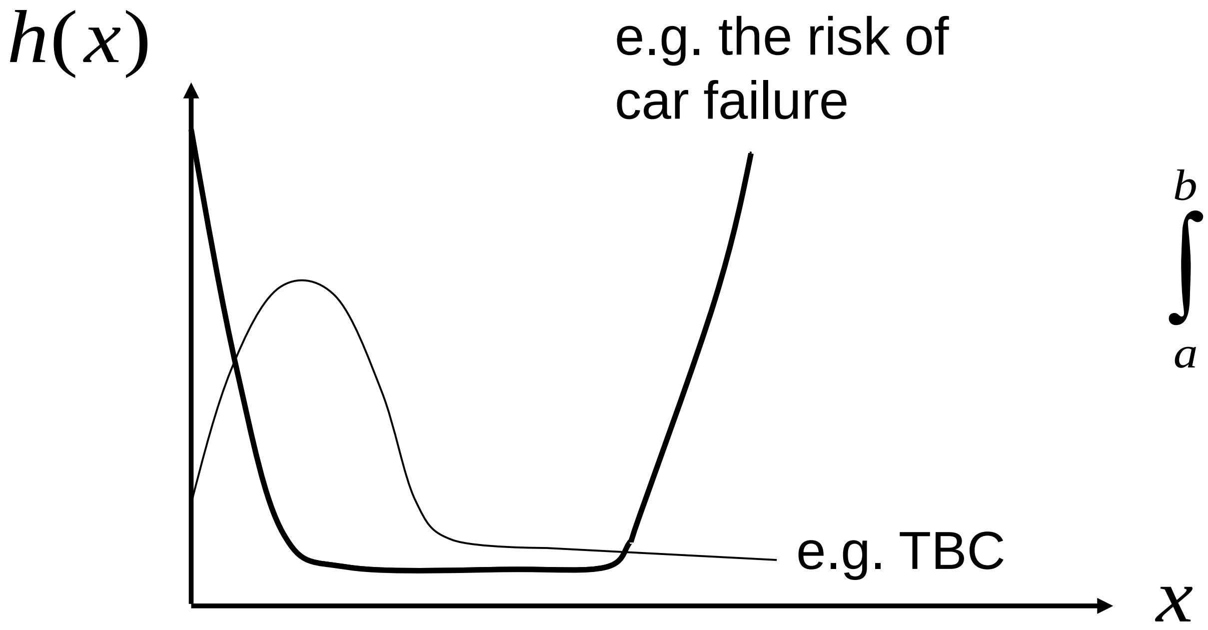
Conditional on the fact that you still haven't experienced the event at time x.

Compare:

$$f(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X < x + \Delta x)}{\Delta x}$$

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d \ln(S(x))}{dx}$$

3) Hazard function (intensity function)



$$\int_a^b h(x)dx$$ can be > 1 (!)

Only one restriction: $h(x) \geq 0$

4) Cumulative Hazard function

$$H(x) = \int_0^x h(t)dt \quad = -\ln\big(S(x)\big)$$

$$S(x) = e^{-H(x)} \quad = e^{-\int_0^x h(t)dt}$$ (for continuous lifetimes)

Denotes the total amount of risk accumulated up to a certain time point, i.e. the number of times we would expect to observe events over a given time period, if events were repeatable.

Example: Survival vs. Hazard function

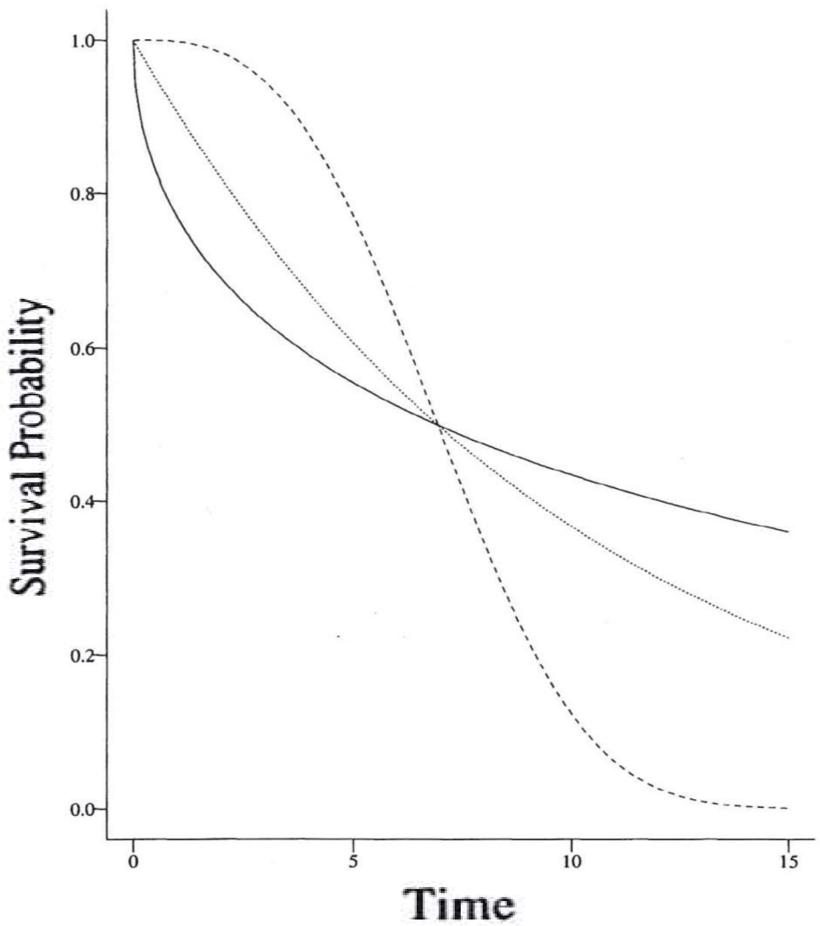


Figure 2.1 *Weibull Survival functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (———); $\alpha = 1.0$, $\lambda = 0.1$ (······); $\alpha = 3.0$, $\lambda = 0.00208$ (------).

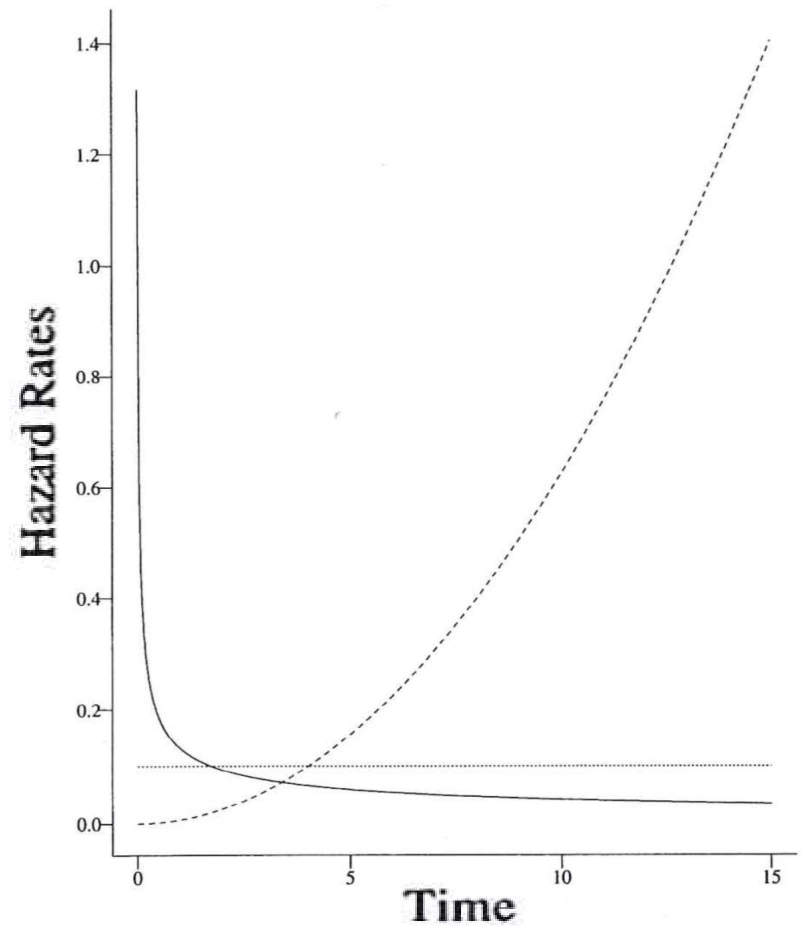


Figure 2.5 *Weibull hazard functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (———); $\alpha = 1.0$, $\lambda = 0.1$ (------); $\alpha = 3.0$, $\lambda = 0.00208$ (— — —).

5) Mean Residual Life

$$\mathrm{mrl}(x) = E(X - x \mid X > x)$$

The expected remaining lifetime at time x.

$\mathrm{mrl}(0) = \mu$ = average lifetime (from birth)

Quantiles of the survival data distribution

The pth quantile (or the 100pth percentile) of the distribution of X is the smallest x_p that fulfills

$$S(x_p) \leq 1 - p$$

The time where 100p% have experienced the event.

Example: All cause mortality in the US

$$S(x_{0.2}) \leq 0.8$$

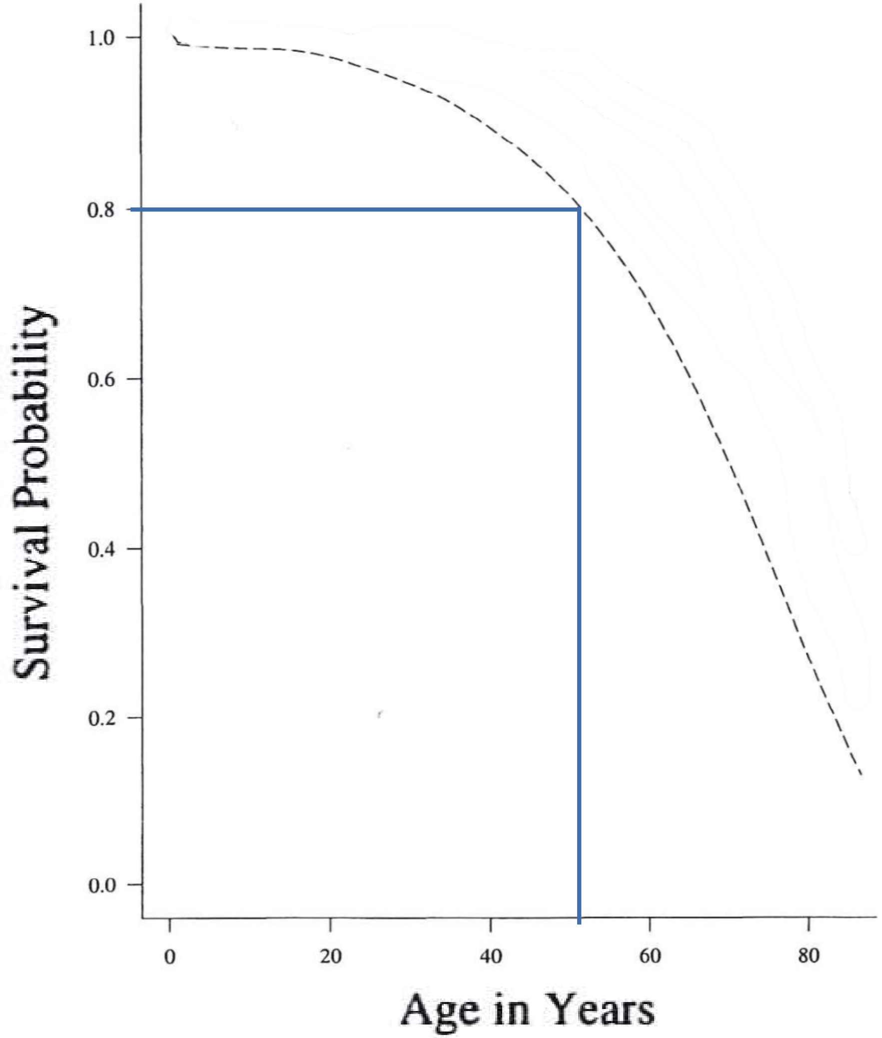


Figure 2.2 *Survival Functions for all cause mortality for the US population in 1989. White males (————); white females (······) black males (------) females (——-——).*

The 20th percentile of the survival time distribution is approximately 50 years

Median Lifetime

The **median lifetime** is the time at which half of the individuals have experienced the event.

The median lifetime is the 50th percentile of the distribution of X.

In other words, the median lifetime is the value $x_{0.5}$ that fulfills

$$S(x_{0.5}) \leq 0.5$$

Example: All cause mortality in the US

$$S(x_{0.5}) \leq 0.5$$

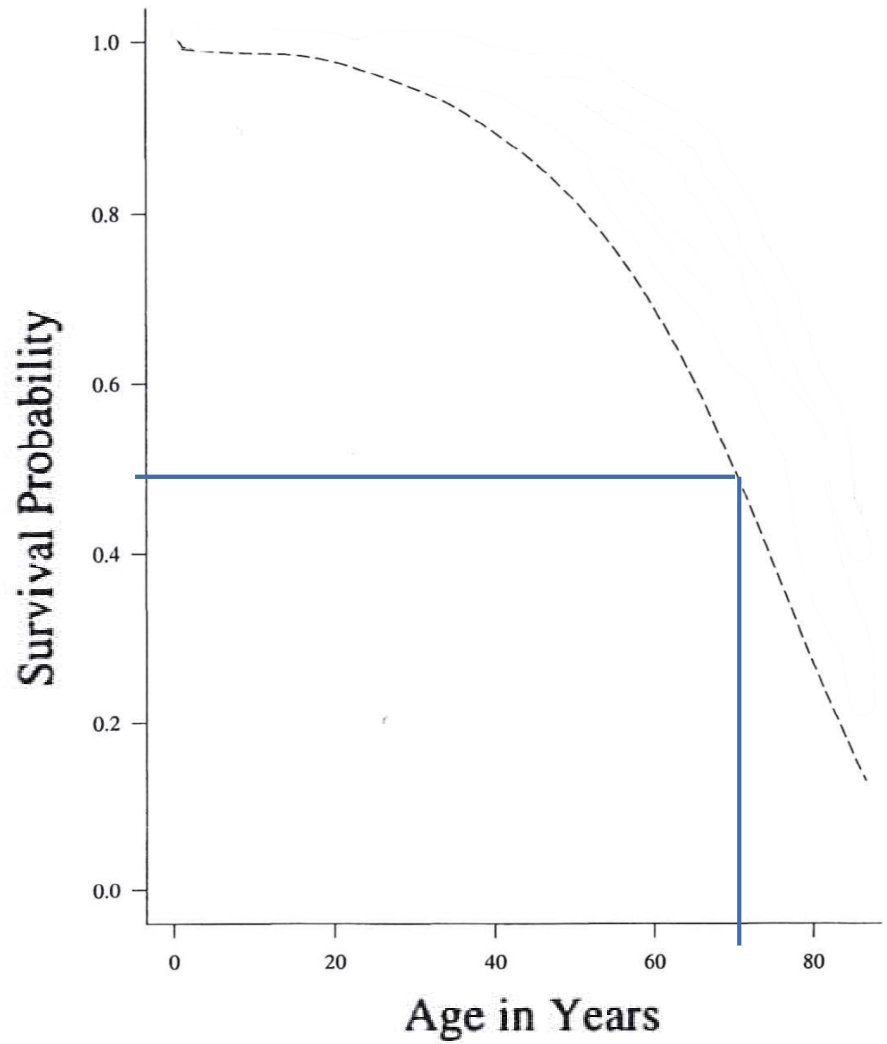


Figure 2.2 *Survival Functions for all cause mortality for the US population in 1989. White males (———); white females (⋯⋯) black males (------) females (—— ——).*

The median lifetime is approximately 70 years

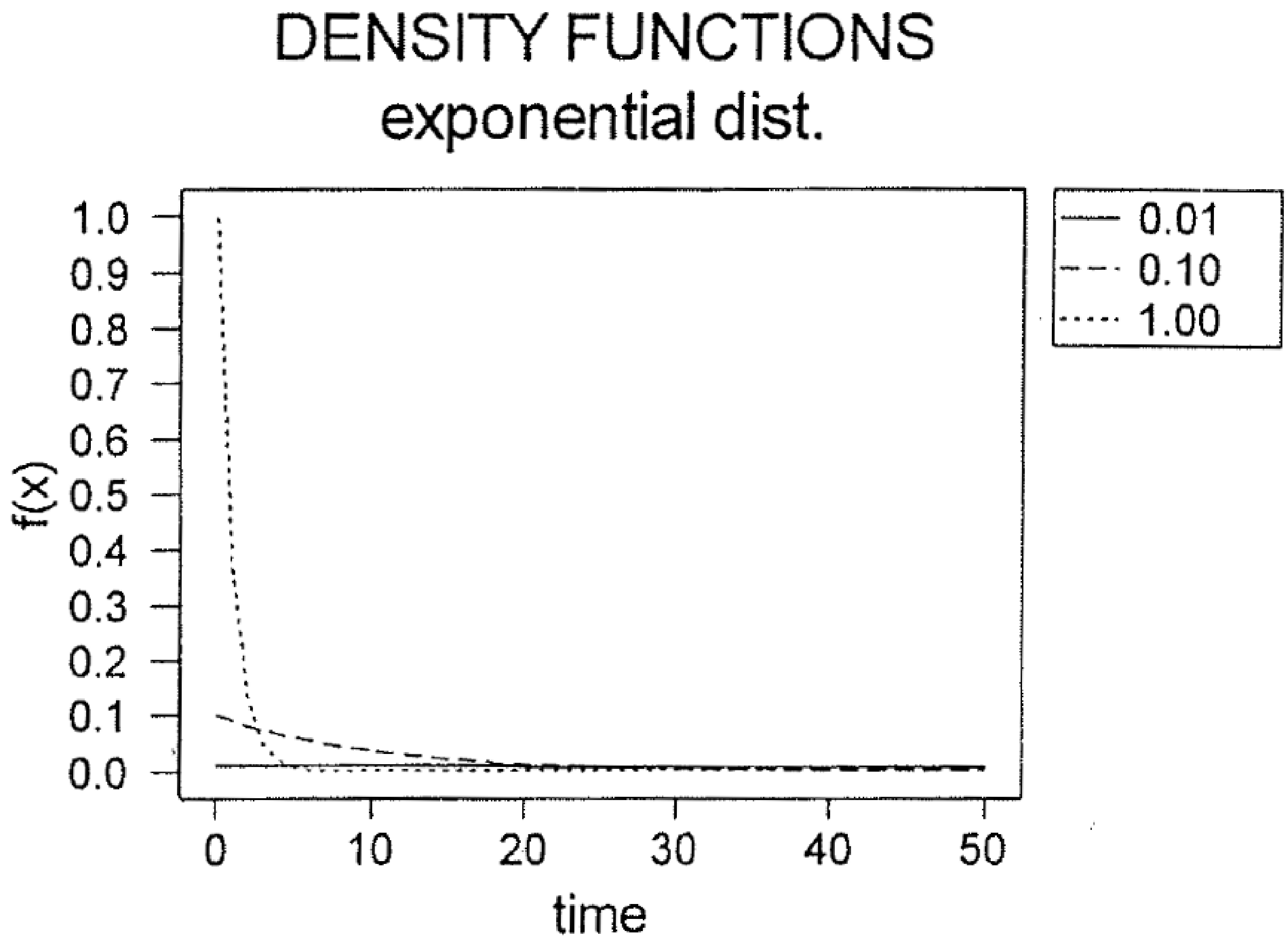
Common parametric models

Survival data can follow a number of parametric distributions, e.g.

- Exponential
- Weibull
- Gamma
- Log normal
- Log logistic
- etc..

Exponential distribution

$$f(x) = \lambda e^{(-\lambda x)}$$

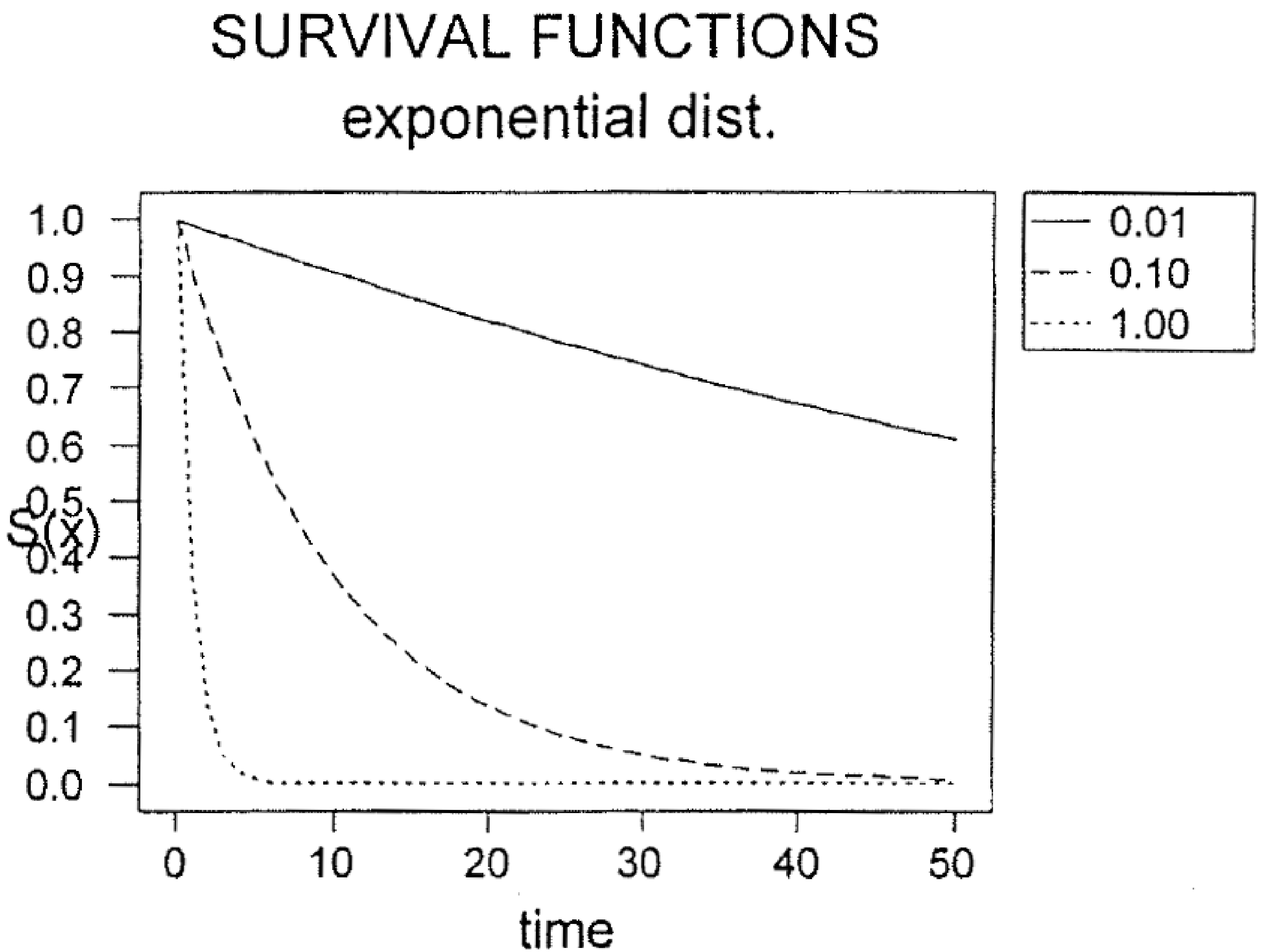


Exponential distribution

$$S(x) = \int_x^\infty f(t)dt$$

$$= \int_x^\infty \lambda e^{(-\lambda t)} dt$$

$$\boxed{S(x) = e^{(-\lambda x)}}$$



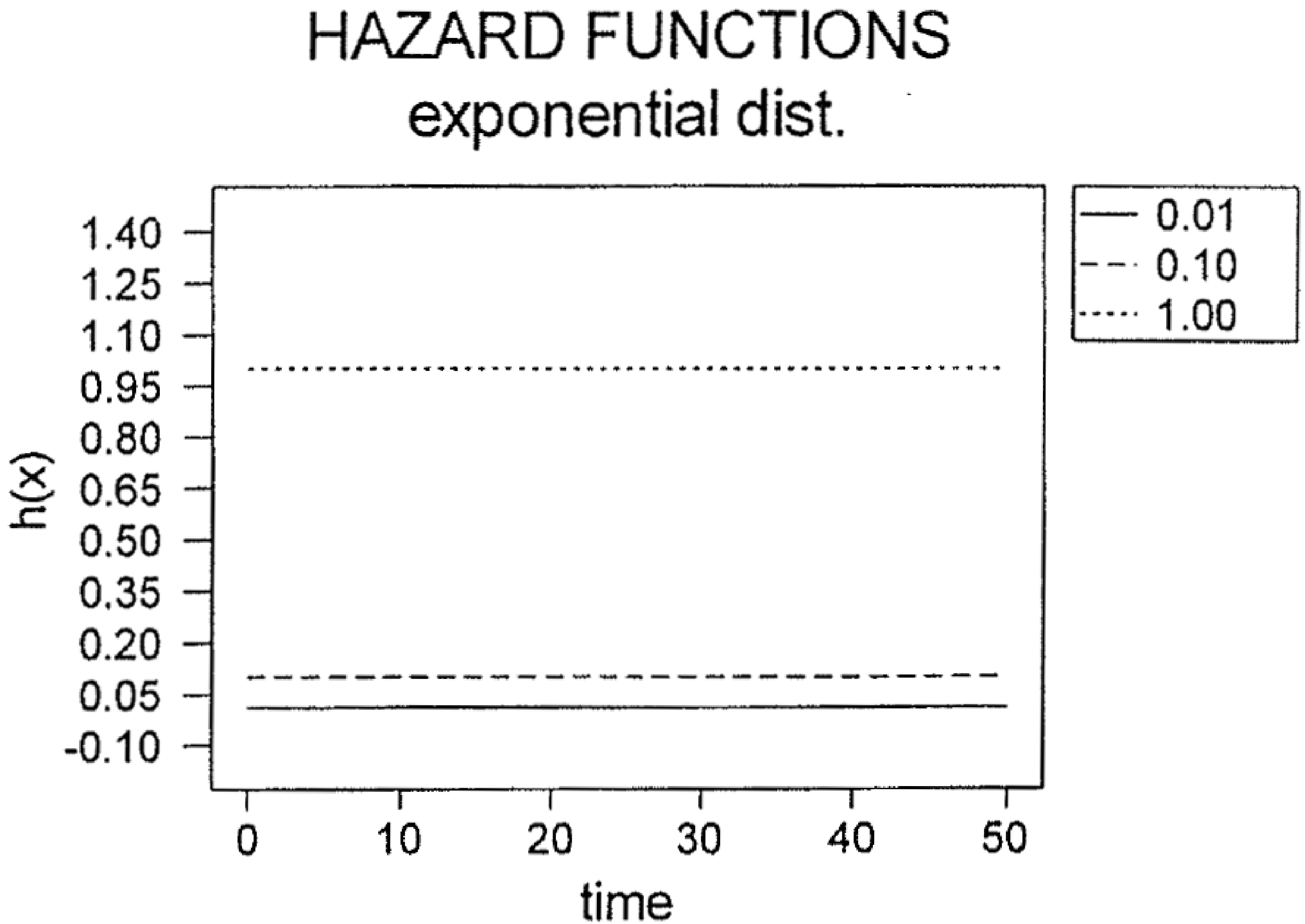
Exponential distribution

$$h(x) = \frac{f(x)}{S(x)}$$

$$= \frac{\lambda e^{(-\lambda x)}}{e^{(-\lambda x)}}$$

$$\boxed{h(x) = \lambda}$$

Constant hazard function



Exponential distribution

A constant hazard rate means that the probability to experience the event at any time t, does not depend on t.

The constant hazard rate also means that the mean residual life (the expected remaining lifetime) is constant.

Weibull distribution

$$\boxed{f(x) = \alpha\lambda x^{\alpha-1} e^{(-\lambda x^{\alpha})}}$$

$$S(x) = \int_x^{\infty} f(t)dt = \int_x^{\infty} \alpha\lambda t^{\alpha-1} e^{(-\lambda t^{\alpha})} dt$$

$$\boxed{S(x) = e^{(-\lambda x^{\alpha})}}$$

$$h(x) = \frac{f(x)}{S(x)} = \frac{\alpha\lambda x^{\alpha-1}\cancel{e^{(-\lambda x^{\alpha})}}}{\cancel{e^{(-\lambda x^{\alpha})}}}$$

$$\boxed{h(x) = \alpha\lambda x^{\alpha-1}}$$

Weibull distribution

Weibull hazard rates can be increasing ($\alpha > 1$), decreasing ($\alpha < 1$), or constant ($\alpha = 1$),

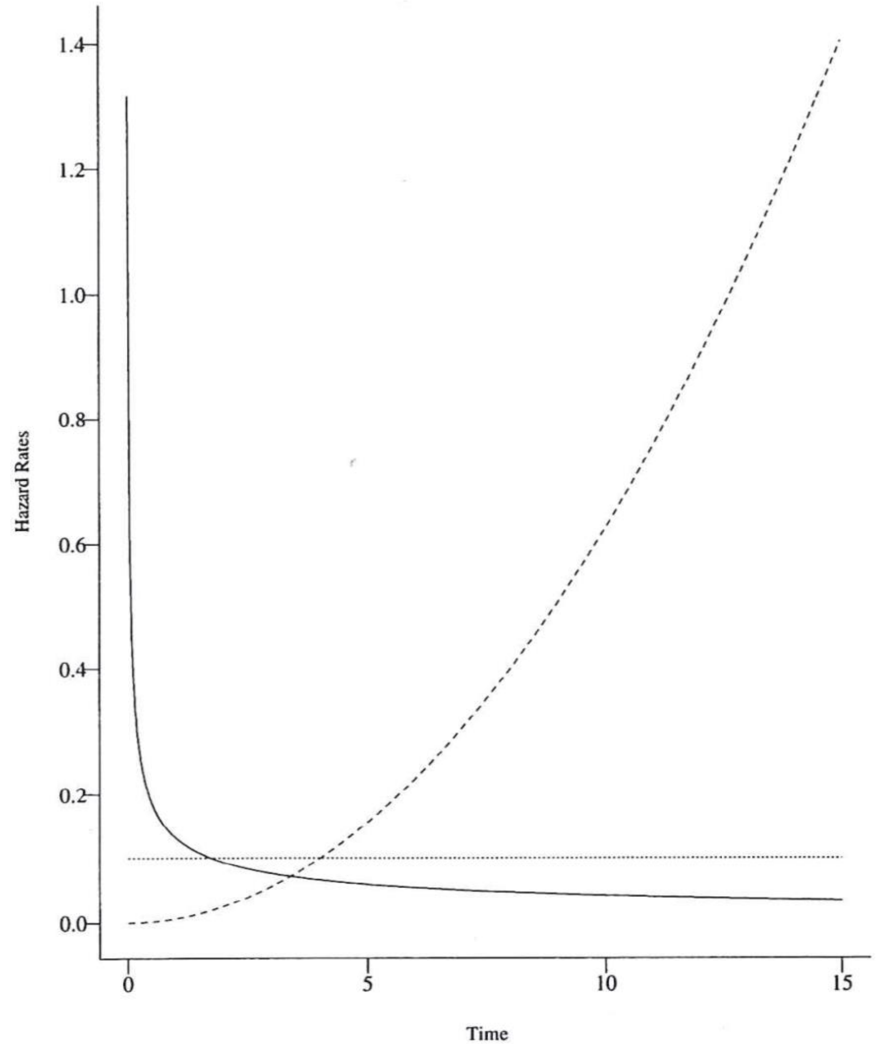


Figure 2.5 *Weibull hazard functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (———); $\alpha = 1.0$, $\lambda = 0.1$ (------); $\alpha = 3.0$, $\lambda = 0.00208$ (— — —).

UPPSALA UNIVERSITET

Log normal distribution

$$f(x) = \frac{e^{\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right)}}{x(2\pi)^{1/2}\sigma} = \phi\left(\frac{\ln x - \mu}{\sigma}\right) \Big/ x$$

Density function of a standard Normal variable.

$$S(x) = 1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

Cumulative distribution function of a standard Normal variable.

$$h(x) = \frac{f(x)}{S(x)}$$

Log normal distribution

Log normal hazard rates are hump-shaped

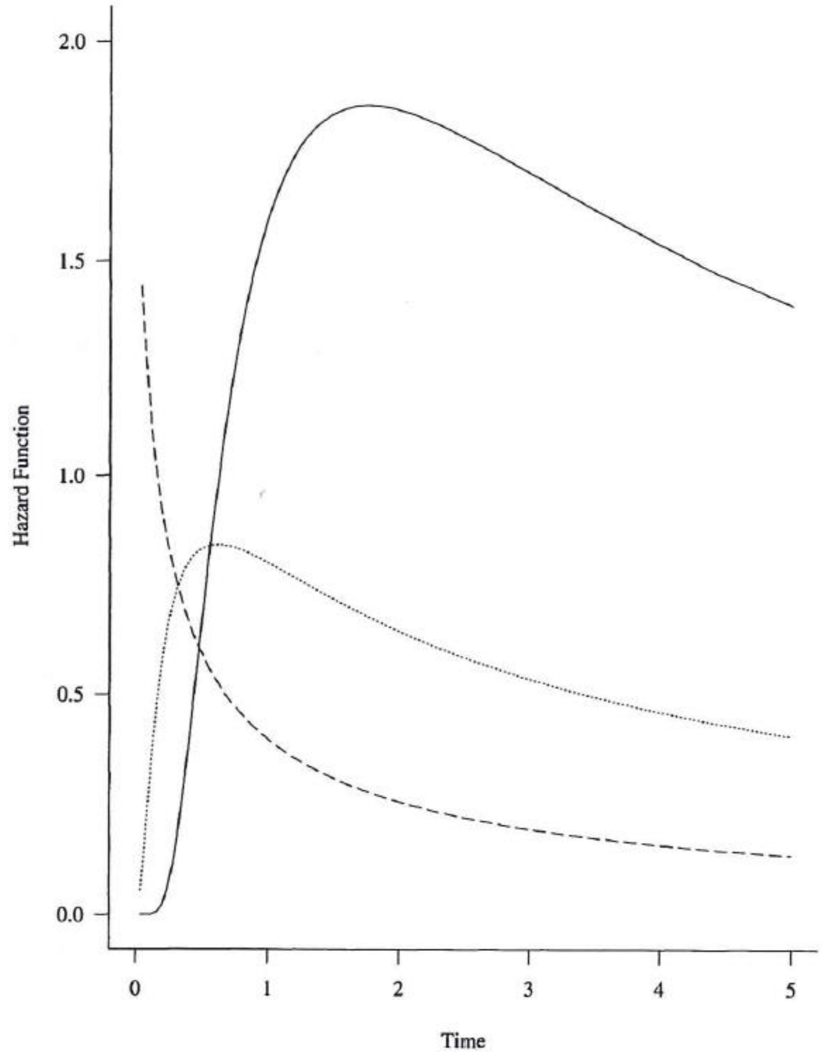


Figure 2.8 *Log normal hazard rates.* $\mu = 0, \sigma = 0.5$ (———); $\mu = 0, \sigma = 0.1$ (------); $\mu = 0, \sigma = 2.0$ (— — —)

Log logistic distribution

Log logistic hazard rates can be monotone decreasing ($1/\sigma \leq 1$), or hump-shaped ($1/\sigma > 1$)

Hazard rates similar to the log normal, but with simpler hazard and survival functions.

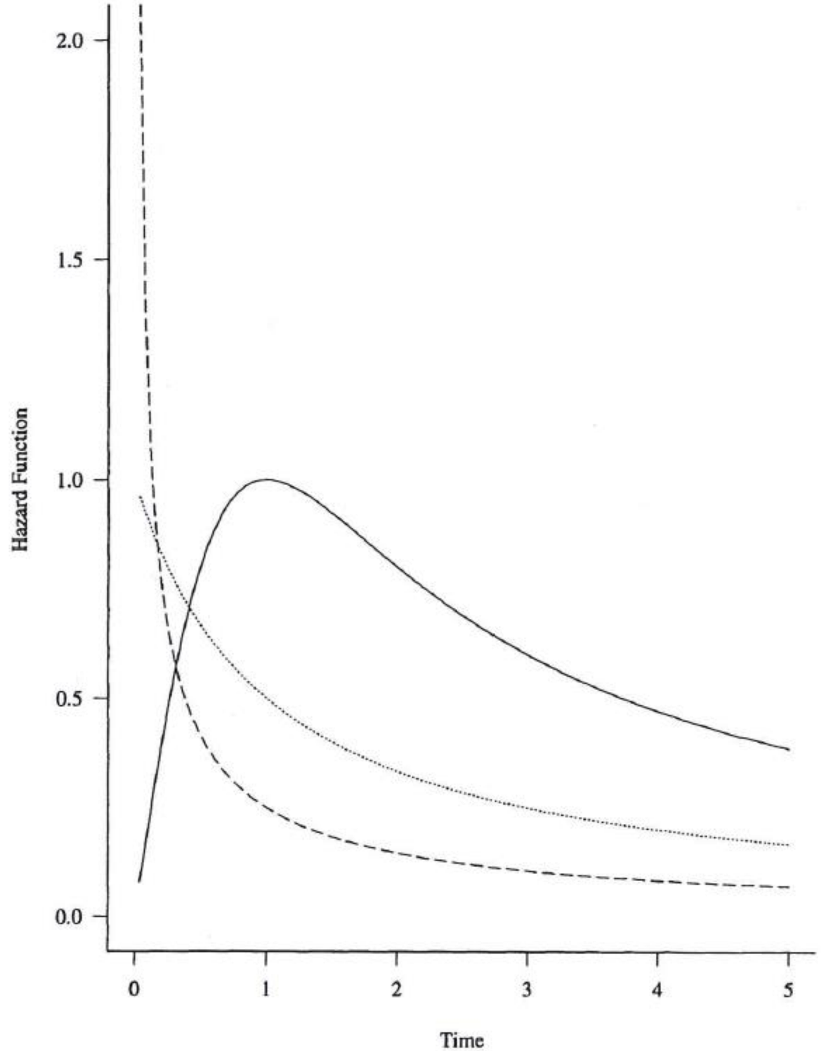


Figure 2.9 *Log logistic hazard rates.* $\lambda = 1, \sigma = 0.5$ (———); $\lambda = 1, \sigma = 1.0$ (······); $\lambda = 1, \sigma = 2.0$ (------)

Gamma distribution

Gamma hazard rates can be monotone increasing (β > 1), monotone decreasing (β < 1), or constant (β = 1)

Properties similar to the Weibull distribution but mathematically more complicated.

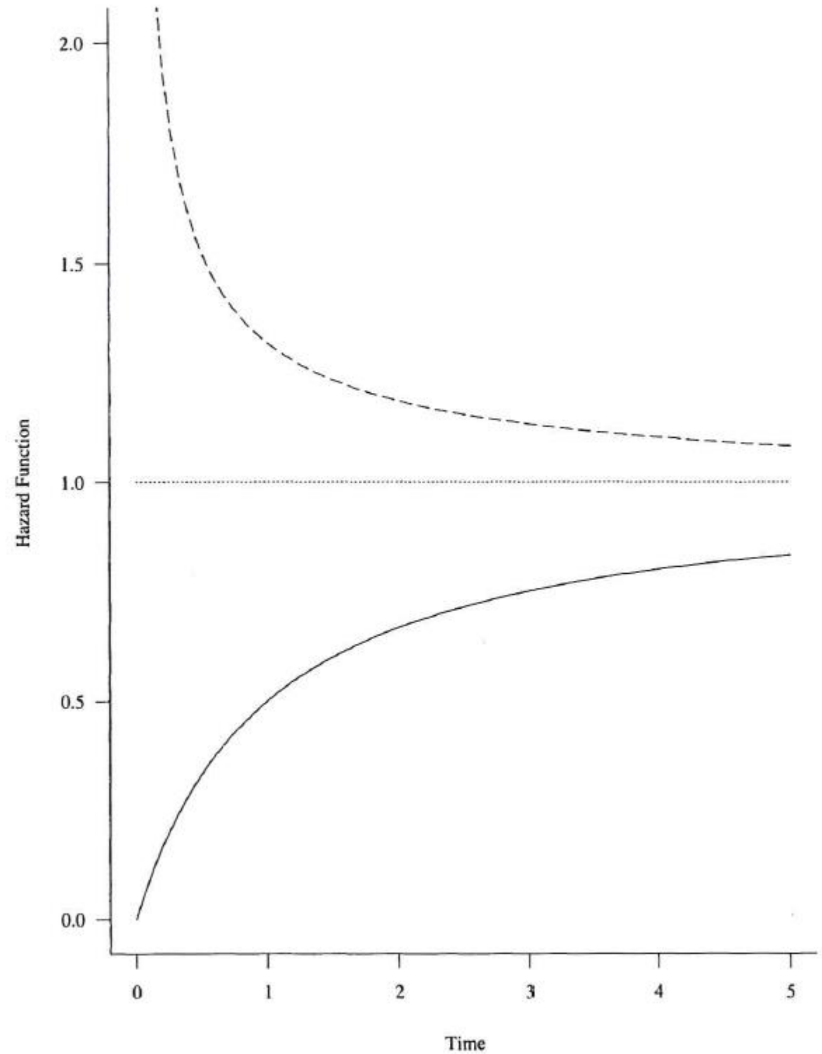


Figure 2.10 *Gamma hazard rates.* $\lambda = 1, \beta = 2.0$ (———); $\lambda = 1, \beta = 1.0$ (------); $\lambda = 1, \beta = 0.5$ (— — —)

Estimating parametric model parameters

Once a parametric distribution is known the model parameters (λ, α, σ, etc) can be estimated by maximum likelihood methods.

There are also methods of estimating the fit of a certain distribution (e.g. Kolmogorov-Smirnov test).

Regression models for survival data

As for other types of data, you might wish to investigate the relationship between the survival time X and one or more explanatory variables.

X = survival time (time to event)

$\mathbf{Z}^t = (Z_1, \ldots, Z_p)$ = vector of explanatory variables associated with X.

Z_i can be quantitative, qualitative, and/or time-dependent.

Time-dependent explanatory variables

Variables that depend on the survival time are said to be **time-dependent (**the effect is not constant over time).

A time-dependent variable can e.g. be:

- a variable that denotes if some intermediate event has occurred by time x (e.g. development or recovery of some other disease, cessation of smoking, a transplant, change in blood pressure, etc.)
- The amount of time passed since some intermediate event

Time-dependent explanatory variables

If $\mathbf{Z}^t$ includes time-dependent variables:

$$\mathbf{Z}^t(x) = [Z_1(x), \; \ldots , Z_p(x)].$$

Modelling of covariate effects on survival

Two popular approaches:

1) Accelerated failure-time model, analogous to the classical linear regression approach.
2) Modelling the conditional hazard rate as a function of the covariates (which is the major approach).

Accelerated failure-time model

$Y = \ln(X)$ is modelled.

A linear model is assumed for Y:

$$\boxed{Y = \mu + \boldsymbol{\gamma}^t \mathbf{Z} + \sigma W}$$

$\boldsymbol{\gamma}^t = (\gamma_1, \dots, \gamma_p)$ = vector of regression coefficients

W = error distribution.

The regression coefficients are estimated using maximum likelihood methods.

Accelerated failure-time model

$X = e^{Y}$

The effect of the explanatory variables in the original time scale is accelerated (or degraded) by a constant factor, thereof the name of the model (further explanation in the book).

The use of the accelerated failure-time model is restricted by the error distributions one can assume.

Modelling the conditional hazard rate

Two classes of models:

- Multiplicative hazard rate models
- Additive hazard rate models

Multiplicative hazard rate models

$$\boxed{h(x|\mathbf{z}) = h_0(x)c(\boldsymbol{\beta}^t\mathbf{z})}$$

Baseline hazard rate

Function of covariates

$h_0(x)$ is the underlying hazard function, describing how the risk of event per time unit changes over time at baseline levels of covariates.

$h_0(x)$ can have a specified parametric form, or be an arbitrary nonnegative function.

Example: Exponential distribution

$$\boxed{h(x|\mathbf{z}) = h_0(x)c(\boldsymbol{\beta}^{\mathrm{t}}\mathbf{z})}$$

$$h_0(x) = \lambda \qquad c(\boldsymbol{\beta}^t\mathbf{z}) = \exp(\boldsymbol{\beta}^t\mathbf{z})$$

Commonly used function

Risk of experiencing the event at time x for covariate values $\mathbf{z}$ is then

$$h(x\,|\,\mathbf{z}) = \lambda\exp(\boldsymbol{\beta}^t\mathbf{z})$$

Estimates of $\boldsymbol{\beta}$ can be obtained by maximum likelihood estimation.

Arbitrary baseline hazard function

If the interest lies in comparing hazard rates (or survival rates) between different groups, e.g. individuals with different covariate values, the baseline hazard is of less importance and doesn't have to be specified.

Multiplicative hazard rate models
Proportional hazard rates

Consider two individuals with covariate values $\mathbf{z}_1$ and $\mathbf{z}_2$.

The relationship between the hazard rates of these individuals:

$$\frac{h(x \mid \mathbf{z}_1)}{h(x \mid \mathbf{z}_2)} = \frac{\cancel{h_0(x)}\,c(\boldsymbol{\beta}^t \mathbf{z}_1)}{\cancel{h_0(x)}\,c(\boldsymbol{\beta}^t \mathbf{z}_2)} = \frac{c(\boldsymbol{\beta}^t \mathbf{z}_1)}{c(\boldsymbol{\beta}^t \mathbf{z}_2)}$$

The ratio is constant, independent of time.

This means that the hazard rates are proportional.

Additive hazard rate models

$$h(x|\mathbf{z}) = h_0(x) + \sum_{j=1}^{p} z_j(x)\beta_j(x)$$

Baseline
hazard rate

Additive models are usually estimated by nonparametric (weighted) least-squares methods.

We'll focus on the multiplicative models in this course.