Survival Analysis

Lecture 1

Marta Fiocco & Hein Putter

Department of Medical Statistics and Bioinformatics
Leiden University Medical Center

Outline

Introduction

Structure of the course

What is survival analysis?

Examples

Time-to-event variables

Basic concepts describing survival

Survival function

Hazard function

Mean Residual Life Function and Median Life

Parametric models

General

Exponential distribution

Weibull distribution

Other distributions

To do for next week

Outline

Introduction

- Structure of the course
- What is survival analysis?
- Examples
- Time-to-event variables

Basic concepts describing survival

- Survival function
- Hazard function
- Mean Residual Life Function and Median Life

Parametric models

- General
- Exponential distribution
- Weibull distribution
- Other distributions

To do for next week

Structure of the course

- Teachers:
  - Marta Fiocco (LUMC)
  - Hein Putter (LUMC)
- Marta Fiocco: `m.fiocco@lumc.nl`
- Hein Putter: `h.putter@lumc.nl`
- Format:
  - 2 × 45 min. class in the morning, based on Klein&Moeschberger, *Survival Analysis*, also other material will be used
  - 2 × 45 min. practical session in the afternoon, exercises from KM, R, other exercises

Structure of the course

- The final test consists on two parts:
    1. Report on data analysis (included: one or more theoretical questions, a data set to analyze, presentation of the analysis at the end of the course).
    2. Written exam with problems.
- The final grade will be based on both parts (score on both parts must be at least 6 in order to pass the exam).

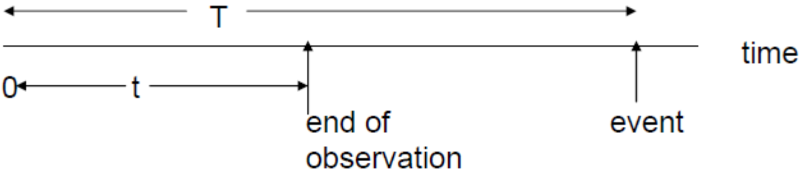
What is survival analysis?

Course description:

- Survival analysis is the study of the distribution of life times,
- i.e. the times from an initiating event (birth, diagnosis, start of treatment) to some terminal event (relapse, death).
- It is most prominently (but not only) used in the biomedical sciences.
- Special feature of survival data: need time to observe the event of interest.
- As a result for a number of subjects the event is not observed, it is only known that the event has not taken place yet. This phenomenon is called **censoring** and it requires special statistical methods.

Censoring

- Censoring means: observation ends before occurrence of event



- Censoring: if we only observe $T > t$, but T unknown

Examples

- Survival time (in general): measured from birth to death for an individual. This is the survival time we need to investigate in a life expectancy study.
- Survival time of a treatment for a population with certain disease: measured from the time of treatment initiation until death

Examples

- Time between stem cell transplant and treatment failure (relapse or death)
- Reliability of light bulbs
- The time of interest may be time to something *good* happening
- For example, we may be interested in how long it takes to eradicate an infection after treatment with antibiotics.

More complicated examples

- Recurrent events: distribution of new skin tumors over time
- Time to relapse in the presence of competing event death
- Time between implantation of a hip prothesis and its failure
- Time to death given/since relapse
- Survival time due to heart disease: (the event is death from heart disease): measured from birth (or other time point such as treatment initiation for heart disease patients) to death caused by heart disease
  - This may be a bit tricky if individuals die from other causes: this is competing risk problem; i.e other risks are competing with heart disease to produce an event death

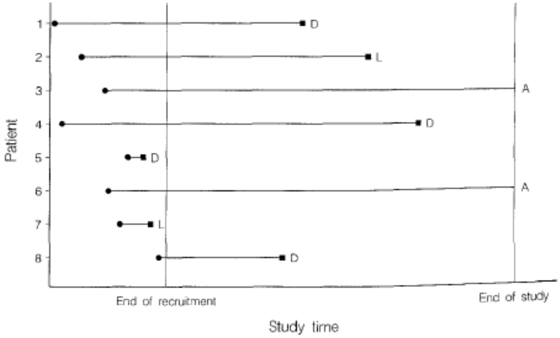
Cohort studies

- A group of individuals is followed in time for the occurrence of some event of interest
- Covariates (explanatory variables) are observed at entry in the study and possibly updated during follow-up
- Interest is in the relation between the outcome event and the covariates
- The statistical methodology for this type of data is indicated by different terminology:
  - Survival analysis
  - Event history analysis
  - Analysis of time-to-event variables
  - Analysis of censored data

Patient time and study time

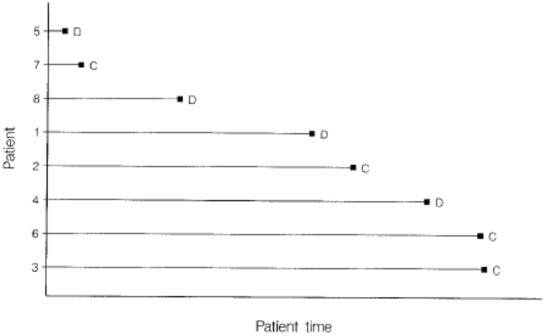
- In a typical study, patients are not all recruited at exactly the same time but accrue over a period of months or even years
- After recruitment, patient are followed up until they die, or until the end of the study, when the data are analyzed
- The actual survival times will be observed for a number of patients, after recruitment some patients may be lost to follow-up, while others will still be alive at the end of the study
- **Study time**: calendar time period in which an individual is in the study
- **Patient time**: period of time that a patient spends in the study, measured from that patient's time origin

Study time for eight patients in a survival study



- ●: time of entry to the study;
- individuals 1, 4, 5 and 8 die (**D**) during the course of the study; 2 and 7 are lost to follow-up (**L**);
- individuals 3 and 6 are still alive (**A**) at the end of the observation period.

Patients time for eight patients in a survival study



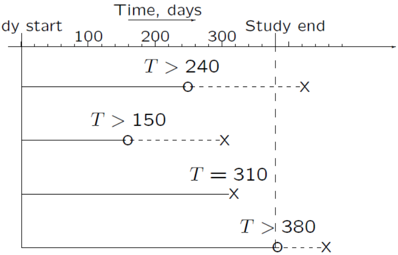
- for individuals 1, 4, 5 and 8: recorded the survival time (**D**)
- the survival times of the remaining individuals are right-censored (**C**).

Examples

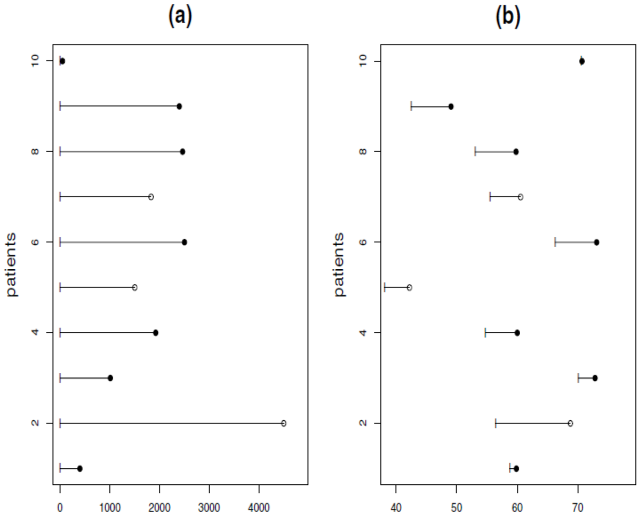
Some individuals may not be observed for the full time to failure, e.g. because of:

- Loss to follow-up
- Drop out
- Termination of the study

(type of censoring: right censoring)



Survival times on two different times-scales



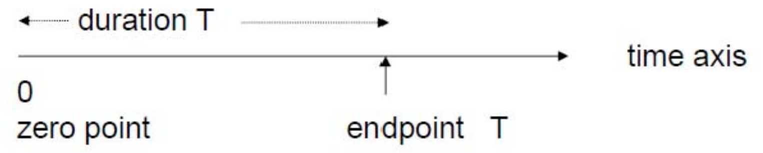
Examples

- **Left figure:** follow-up time-scale; this is the time-scale of interest in the study since we want to study the lifetimes after inclusion in the study.
- Zero is the starting point for all observations on the follow-up time-scale
- Censored observations are indicated with an open ball, whereas those that are observed to die during the study are marked with a filled ball

- **Right figure:** survival times on the age time-scale.
- If age was used as the time-scale for the analysis, the data would have delayed entry because the patient entered the study at different ages and are alive at inclusion, and this must be dealt with in a subsequent analysis.
- Otherwise there is a risk that long lifetimes will be over-represented (long lifetimes have a higher probability of being sampled), which is referred to as length-bias.
- Working on the age time-scale, such data are also called *left-truncated* since the subjects are sampled to being alive (later in the course)

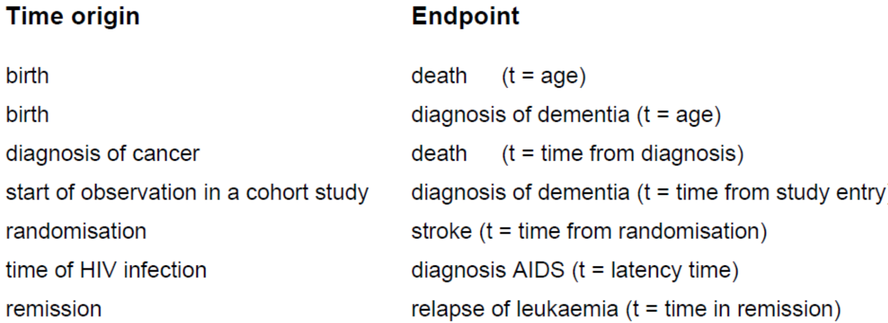
Time-to-event variables

- Individuals are followed in time for the occurrence of a certain event ("endpoint").



- We are interested in T, the time between a well defined starting point (origin) and a well defined endpoint (event). T is often called the "survival time".

Examples of time-to-event variables



Time origin	Endpoint
birth	death (t = age)
birth	diagnosis of dementia (t = age)
diagnosis of cancer	death (t = time from diagnosis)
start of observation in a cohort study	diagnosis of dementia (t = time from study entry
randomisation	stroke (t = time from randomisation)
time of HIV infection	diagnosis AIDS (t = latency time)
remission	relapse of leukaemia (t = time in remission)

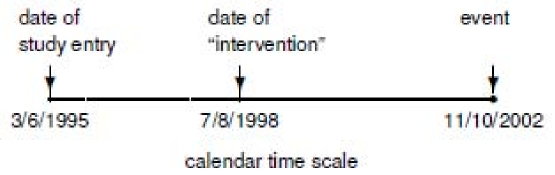
Questions of interest

- What is the distribution of T?
- What is the probability that the event occurs within $(0, t)$ for arbitrary t?
- Which factors do influence the occurrence of the event?

The Observation Process: Time Scales

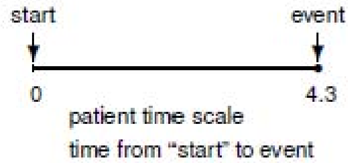
TIME: Several time scales may be involved

- Calendar (study) time scale



The Observation Process: Time Scales

- Personal or patient time scale
  - Time since infection/intervention/start risk behaviour
  - Time since birth (age)
  - Time since intermediary event (AIDS, previous infection, relapse)



Outline

Introduction

Structure of the course

What is survival analysis?

Examples

Time-to-event variables

Basic concepts describing survival

Survival function

Hazard function

Mean Residual Life Function and Median Life

Parametric models

General

Exponential distribution

Weibull distribution

Other distributions

To do for next week

Basic quantities:

Let X be the time until event. Four functions characterize the distribution of X:

- Survival function $S(x)$: probability of an individual surviving to time x
- Hazard rate $h(x)$: chance an individual of age x experiences the event in the next instant of time
- Probability density function $f(x)$: unconditional probability of the event's occurring at x
- Mean residual life function $mrl(x)$: mean time to the event of interest, given it has not occurred at x

Characterizations of survival time distributions

- Survival function $S(x) = Pr(X > x)$
- $S(x)$ continuous, strictly decreasing function
- $S(0) = 1$, $S(\infty) = 0$
- $S(x)$ is the complement of the cumulative distribution function: $S(x) = 1 - F(x)$, where $F(x) = Pr(X \leq x)$
- $S(x) = \int_x^\infty f(t)\mathrm{d}t$
- $f(x) = -\frac{\mathrm{d}S(x)}{\mathrm{d}x}$

Survival function

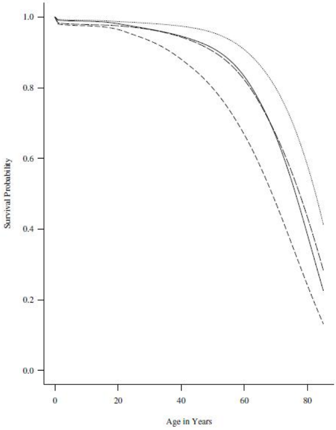


Figure 2.2 *Survival Functions for all cause mortality for the US population in 1989. White males (————); white females (······); black males (------); black females (—— ——).*

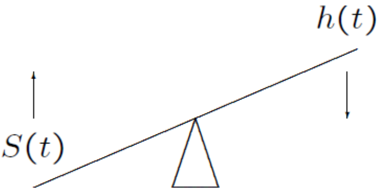
Hazard function

- $h(x) = lim_{\Delta x \to 0} \frac{Pr[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$
- $h(x) \geq 0$
- Cumulative hazard function $H(x)$: $\underline{H(x) = \int_0^x h(u)\mathrm{d}u}$
- $h(x)\Delta x$ approximate probability of an individual of age x of experiencing the event in the next instant

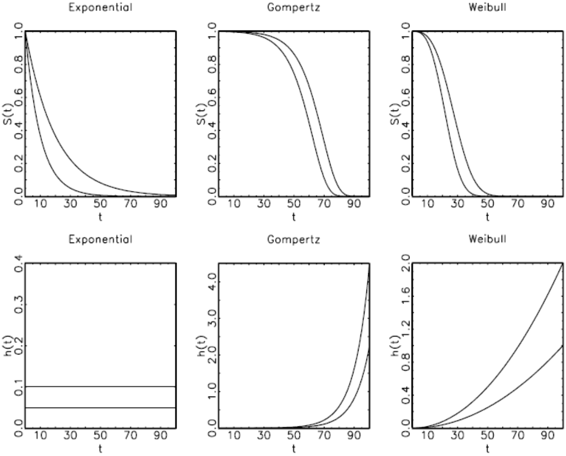
Relationship between functions

- $S(x) = 1 - F(x) = \int_x^\infty f(u)\mathrm{d}u$
- $-\frac{\mathrm{d}S(x)}{\mathrm{d}x} = f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$
- $H(x) = \int_0^x h(u)\mathrm{d}u = -ln[S(x)]$
- $S(x) = \exp[-H(x)] = \exp[-\int_0^x h(u)\mathrm{d}u]$
- $f(x) = h(x)\exp[-\int_0^x h(u)\mathrm{d}u] = h(x)\exp[-H(x)]$
- if X continuous: $\underline{h(x)} = \frac{f(x)}{S(x)} = \underline{-\frac{\mathrm{d}ln[S(x)]}{\mathrm{d}x}}$

High hazard rate= Low survival



Hazard function



Hazard function

Alternative names for $h(x)$

- failure rate
- instantaneous death rate
- risk function
- incidence/mortality rate
- incidence density
- event rate

Hazard rate $h(t)$

- It is particularly useful in determining the appropriate failure distributions utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time
- Many events if both survival probability and hazard are high

Different types of hazard rates

- Models with increasing hazard rates may arise when there is natural aging
- Decreasing hazard functions are much less common but find occasional use when there is a very early likelihood of failure, such as in certain types of electronic devices or in patients experiencing certain types of transplants
- Bathtub-shaped hazard for populations since ’birth’: followed by a constant hazard rate and later increase
- hazard rate increasing early and then declining: survival after successful surgery where there is an initial increase in risk due to complications just after the procedure, followed by a steady decline in risk as the patient recovers

Hazard function

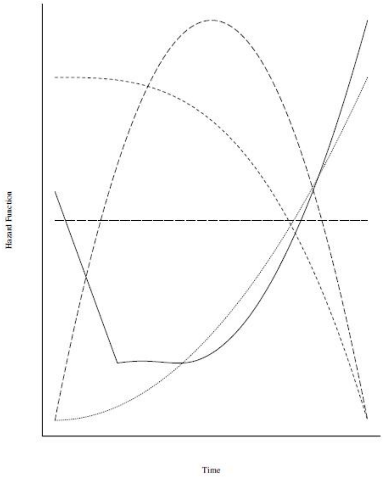


Figure 2.4 *Shapes of hazard functions. Constant hazard (——— ———); increasing hazard (------); decreasing hazard (- - - - - -); bathtub shaped (————); bumpshaped (———).*

Mean and mean residual life function

- $\mu = E(X) = \int_0^\infty t f(t) \mathrm{d}t = \int_0^\infty S(t) \mathrm{d}t$
- $mrl(x) = E(X - x | X > x) = \frac{\int_x^\infty (t-x) f(t) \mathrm{d}t}{S(x)} = \frac{\int_x^\infty S(t) \mathrm{d}t}{S(x)}$
- $\sigma^2 = E(X^2) - [E(X)]^2 = \int_0^\infty t^2 f(t) \mathrm{d}t - \mu^2$
- use integration in parts $[\int u \mathrm{d}v = uv - \int v \mathrm{d}u + C]$ to derive the expression for $Var(X)$
- $Var(X) = 2 \int_0^\infty t S(t) \mathrm{d}t - [\int_0^\infty S(t) \mathrm{d}t]^2$

Quantiles and Median

- pth quantile of the distribution of X is the smallest x_p such that $S(x_p) \leq 1 - p$
- for continuous X, the median lifetime $x_{0.5}$ can be calculated by $S(x_{0.5}) = 0.5$

Outline

Introduction

- Structure of the course
- What is survival analysis?
- Examples
- Time-to-event variables

Basic concepts describing survival

- Survival function
- Hazard function
- Mean Residual Life Function and Median Life

Parametric models

- General
- Exponential distribution
- Weibull distribution
- Other distributions

To do for next week

General

- In practice and in this course, more emphasis on non- and semi-parametric models
- But parametric models useful for gaining insight in mechanism (simulation)
- Flexibility by choice of parameters and possibly by piecewise hazard/survival functions
- Remember: $f(x), S(x), h(x), mrl(x)$ all give the same information

Exponential distribution

- constant hazard function $h(x) = \lambda$ (time-independent)
- $S(x) = exp[-\int_0^x h(u)\mathrm{d}u] = exp[-\lambda x]$
- $f(x) = -\frac{\mathrm{d}S(x)}{\mathrm{d}x} = exp[-\lambda x]$
- memoryless property

$$P(X > s + t|X > s) = P(X > t)$$

- If *X* is a lifetime of a component the memoryless property asserts that the probability that the component will last more than *s+t* units given that it has lasted more than *s* units is the same as that of a new component lasting more than *t* units

Exponential distribution

- "no-aging" property:
$$mrl(x) = \frac{\int_x^\infty S(t)\mathrm{d}t}{S(x)} = \frac{\int_x^\infty exp[-\lambda t]\mathrm{d}t}{exp[-\lambda x]} = \frac{(-1/\lambda)exp[-\lambda x]}{exp[-\lambda x]} = \frac{1}{\lambda}$$
- Example
  - A number of km that a car run before its battery wears out is exponentially distributed with an average equal to 10,000 km
  - We want to travel for 5000 km; compute the probability to complete the trip without having to replace the car battery
  - What can be said when the distribution is not exponential? (see notebook)

The constant hazard rate appears too restrictive in both health and industrial applications

Weibull distribution

- the exponential distribution arises when the life distribution has constant hazard rate function: $h(x) = \lambda$
- but there are situations in which is more realistic to assume that $h(x)$ either increases or decreases over time
- example of such a hazard function

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

- note that $h(t)$ increases when $\alpha > 1$, decreases when $\alpha < 1$ and is constant (reducing to the exponential) when $\alpha = 1$

Weibull distribution

- the distribution whose hazard is given by the definition above is called the *Weibull* distribution with parameters (α, λ)
- $\lambda > 0$: scale parameter; $\alpha > 0$: shape parameter
- Special case: $\alpha = 1$ (exponential)
- $S(x) = \exp[-\int_0^x h(u)du] = \exp[-\int_0^x \lambda\alpha u^{\alpha-1} du] = \exp[-\lambda u^\alpha|_0^x] = \exp[-\lambda x^\alpha]$
- $h(x) = -\frac{\mathrm{d}ln[S(x)]}{\mathrm{d}x} = \lambda\alpha x^{\alpha-1}$
- $f(x) = h(x)S(x)$

Weibull distribution

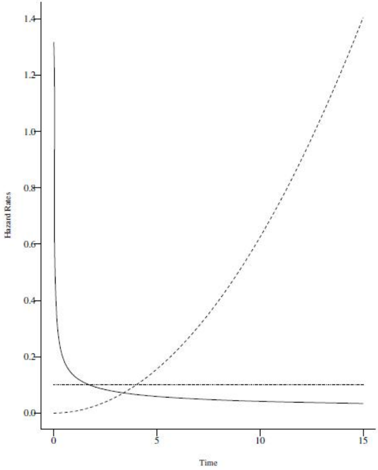


Figure 2.5 *Weibull hazard functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (————); $\alpha = 1.0$, $\lambda = 0.1$ (------); $\alpha = 3.0$, $\lambda = 0.00208$ (———).

Weibull distribution

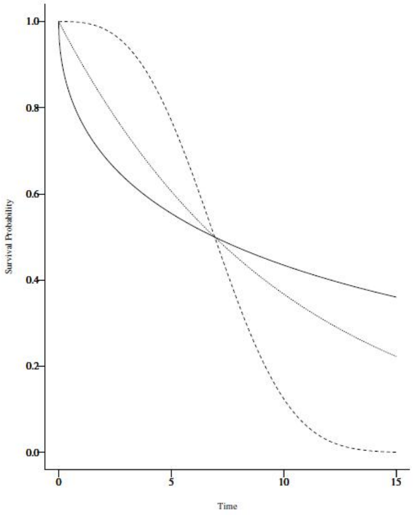
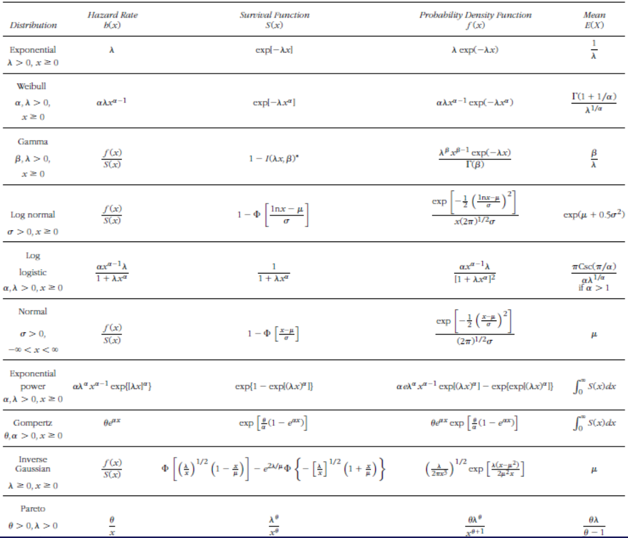


Figure 2.1 *Weibull Survival functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (————); $\alpha = 1.0$, $\lambda = 0.1$ (······); $\alpha = 3.0$, $\lambda = 0.00208$ (------).

General comments

- The Weibull distribution is flexible enough to accommodate increasing ($\gamma > 1$), decreasing ($\gamma < 1$), or constant hazard rates ($\gamma = 1$).
- This fact together with the model's relatively simple survival, hazard, and probability density functions, have made it a very popular parametric model.
- The shape of the Weibull distribution depends upon the value of γ), thus, the reason for referring to this parameter as the *shape* parameter.

Other distributions



Distribution	*Hazard Rate* $b(x)$	*Survival Function* $S(x)$	*Probability Density Function* $f(x)$	*Mean* $E(X)$
Exponential $\lambda > 0, x \geq 0$	λ	$\exp[-\lambda x]$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$
Weibull $\alpha, \lambda > 0,$ $x \geq 0$	$\alpha\lambda x^{\alpha-1}$	$\exp[-\lambda x^{\alpha}]$	$\alpha\lambda x^{\alpha-1}\exp(-\lambda x^{\alpha})$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Gamma $\beta, \lambda > 0,$ $x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - I(\lambda x, \beta)^{*}$	$\frac{\lambda^{\beta}x^{\beta-1}\exp(-\lambda x)}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$
Log normal $\sigma > 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - \Phi\left[\frac{1\text{n}x - \mu}{\sigma}\right]$	$\frac{\exp\left[-\frac{1}{2}\left(\frac{1\text{n}x-\mu}{\sigma}\right)^2\right]}{x(2\pi)^{1/2}\sigma}$	$\exp(\mu + 0.5\sigma^2)$
Log logistic $\alpha, \lambda > 0, x \geq 0$	$\frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$	$\frac{1}{1+\lambda x^{\alpha}}$	$\frac{\alpha x^{\alpha-1}\lambda}{[1+\lambda x^{\alpha}]^2}$	$\frac{\pi \text{Csc}(\pi/\alpha)}{\alpha\lambda^{1/\alpha}}$ if $\alpha > 1$
Normal $\sigma > 0,$ $-\infty < x < \infty$	$\frac{f(x)}{S(x)}$	$1 - \Phi\left[\frac{x-\mu}{\sigma}\right]$	$\frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{(2\pi)^{1/2}\sigma}$	μ
Exponential power $\alpha, \lambda > 0, x \geq 0$	$\alpha\lambda^{\alpha}x^{\alpha-1}\exp\{[\lambda x]^{\alpha}\}$	$\exp\{1 - \exp[(\lambda x)^{\alpha}]\}$	$\alpha e \lambda^{\alpha}x^{\alpha-1}\exp[(\lambda x)^{\alpha}] - \exp\{\exp[(\lambda x)^{\alpha}]\}$	$\int_0^{\infty} S(x)dx$
Gompertz $\theta, \alpha > 0, x \geq 0$	$\theta e^{\alpha x}$	$\exp\left[\frac{\theta}{\alpha}(1 - e^{\alpha x})\right]$	$\theta e^{\alpha x}\exp\left[\frac{\theta}{\alpha}(1 - e^{\alpha x})\right]$	$\int_0^{\infty} S(x)dx$
Inverse Gaussian $\lambda \geq 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$\Phi\left[\left(\frac{\lambda}{x}\right)^{1/2}\left(1 - \frac{x}{\mu}\right)\right] - e^{2\lambda/\mu}\Phi\left\{-\left[\frac{\lambda}{x}\right]^{1/2}\left(1 + \frac{x}{\mu}\right)\right\}$	$\left(\frac{\lambda}{2\pi x^3}\right)^{1/2}\exp\left[\frac{\lambda(x-\mu^2)}{2\mu^2 x}\right]$	μ
Pareto $\theta > 0, \lambda > 0$	$\frac{\theta}{x}$	$\frac{\lambda^{\theta}}{x^{\theta}}$	$\frac{\theta\lambda^{\theta}}{x^{\theta+1}}$	$\frac{\theta\lambda}{\theta - 1}$

Outline

Introduction

- Structure of the course
- What is survival analysis?
- Examples
- Time-to-event variables

Basic concepts describing survival

- Survival function
- Hazard function
- Mean Residual Life Function and Median Life

Parametric models

- General
- Exponential distribution
- Weibull distribution
- Other distributions

To do for next week

To do for next week

- Study Sections 2.1–2.5 of Klein&Moeschberger
- Make assignments of this afternoon