Chapter 1

(AST405) Lifetime data analysis

Md Rasel Biswas

Lecture Outline

1. 1. Basic Concepts and Models
   - 1.1 Introduction
   - 1.2 Lifetime Distributions
   - 1.3 Some important failure time models
   - 1.4 Regression models

Section 1

1. Basic Concepts and Models

Subsection 1

1.1 Introduction

Introduction

- Lifetime data have important use in many research areas, including health sciences, engineering, social sciences, etc.
- Applications of lifetime distribution methodology range from investigation of the durability of manufactured items to studies of human disease and their treatment
- Lifetime data are also referred as “survival time” or “failure time data”

Regression models and lifetables

- Some methods dealing with lifetime data are quite old, but the field expanded rapidly after 1970, specially after publishing of Sir David Cox's famous paper (Cox 1972)

Regression Models and Life-Tables

By D. R. Cox

Imperial College, London

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. Healy in the Chair]

- Number of citations: 62649 (Google Scholar, September 25, 2024)
- Software packages for lifetime data analysis are widely available since 1980

Regression models and lifetables

Figure 1: Sir David Cox (1924 – 2022)

Example 1.1.1

- Manufactured items with mechanical or electronic components are often subjected to life tests in order to obtain information on their durability.
- This involves putting items in operation, often in a laboratory setting, and observing them until they fail.
- It is common here to refer to the lifetimes as "failure times," since when an item ceases operating satisfactorily, it is said to have "failed."

Example 1.1.2

- Demographers and social scientists are interested in the duration of certain life "states" for humans.
- Consider, for example, marriage and, in particular, the marriages formed during the year 1980 in a particular country.
- Then the lifetime of a marriage would be its duration; a marriage may end due to annulment, divorce, or death.

Example 1.1.3

- In medical studies dealing with potentially fatal diseases one is interested in the survival time of individuals with the disease, measured from the date of diagnosis or some other starting point.
- For example, it is common to compare treatments for a disease in terms of the survival time distributions for patients receiving the different treatments.

Example 1.1.4

- A standard experiment in the investigation of carcinogenic substances is one in which laboratory animals are subjected to doses of the substance and then observed to see if they develop tumors.
- A main variable of interest is the time to appearance of a tumor, measured from when the dose is administered.

Time scale and time origin

- The definition of lifetime includes a "time scale" and "time origin", and also the specification of the event (e.g. failure or death) that determines the lifetime
- Time scale is not always real or chronological time, e.g.
  - miles driven can be used as a time scale with motor vehicles
  - number of pages for a computer printer or a photocopier, etc.

Censoring

- The chronological time needed to observe the lifetimes of all individuals in a study may be large enough that practical constraints prevent full observations
- If an individual's lifetime is only known to be exceed a certain value, then it is known as "censored" observation and the process is known as censoring
- For example, if a life test is terminated after 28 days and one item had not failed by then, then we would only know that its lifetime is greater than 28 days and it is referred as "censoring time"
- There are different types of censoring, e.g. right, left and interval censoring, which will be discussed in detail in the next chapter

Censoring

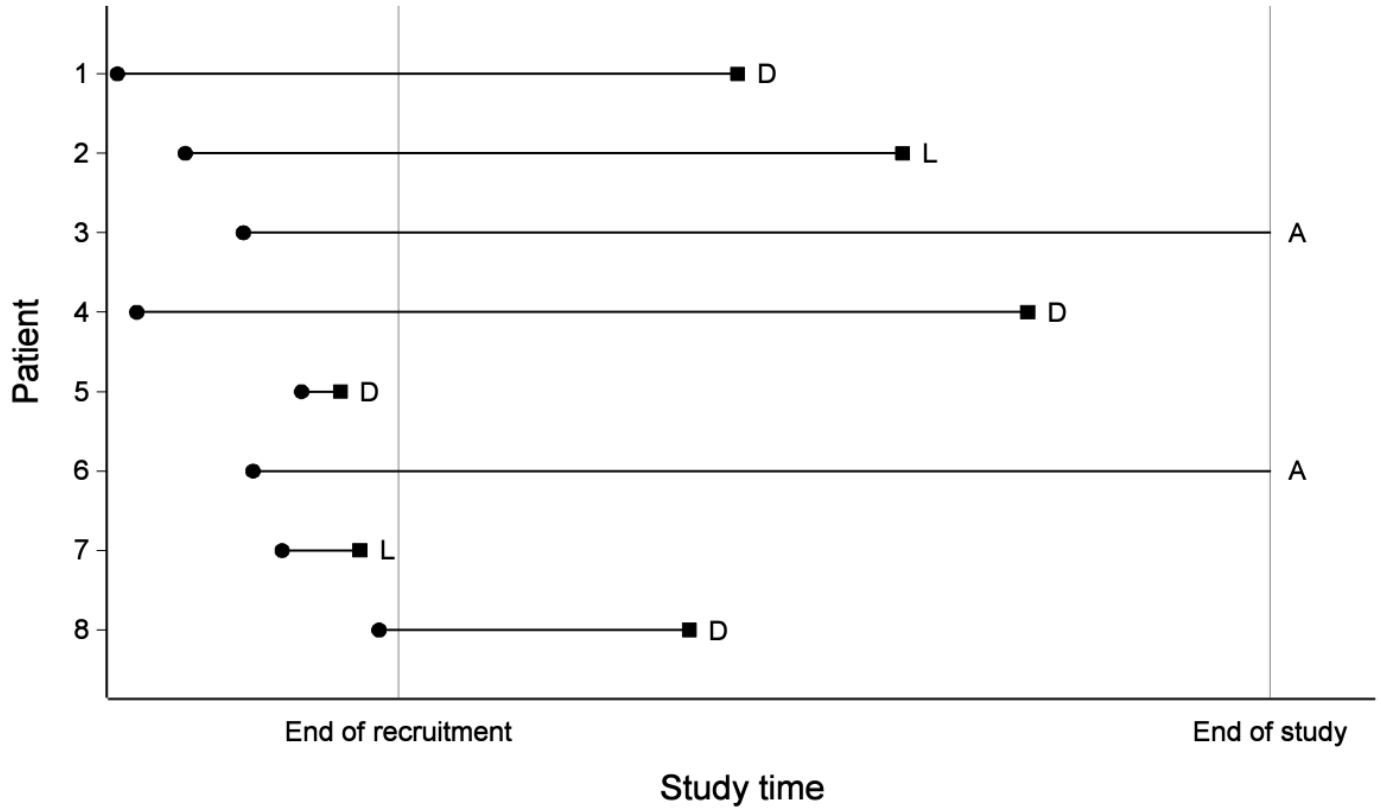


Figure 2: Lifetime of eight subjects in a survival study

Example 1.1.5

- Nelson (1972) described the results of a life test experiment in which specimens of a type of electrical insulating fluid were subjected to a constant voltage stress.
- The length of time until each specimen failed, or "broke down," was observed.
- The table in the next slide gives results for seven groups of specimens, tested at voltages ranging from 26 to 38 kilo-volts (kV).

Example 1.1.5

Table 1.1. Times to Breakdown (in minutes) at Each of Seven Voltage Levels

Voltage Level (kV)	n_i	Breakdown Times
26	3	5.79, 1579.52, 2323.7
28	5	68.85, 426.07, 110.29, 108.29, 1067.6
30	11	17.05, 22.66, 21.02, 175.88, 139.07, 144.12, 20.46, 43.40, 194.90, 47.30, 7.74
32	15	0.40, 82.85, 9.88, 89.29, 215.10, 2.75, 0.79, 15.93, 3.91, 0.27, 0.69, 100.58, 27.80, 13.95, 53.24
34	19	0.96, 4.15, 0.19, 0.78, 8.01, 31.75, 7.35, 6.50, 8.27, 33.91, 32.52, 3.16, 4.85, 2.78, 4.67, 1.31, 12.06, 36.71, 72.89
36	15	1.97, 0.59, 2.58, 1.69, 2.71, 25.50, 0.35, 0.99, 3.99, 3.67, 2.07, 0.96, 5.35, 2.90, 13.77
38	8	0.47, 0.73, 1.40, 0.74, 0.39, 1.13, 0.09, 2.38

Example 1.1.5

- The main purpose of the experiment was to investigate the distribution of "time to breakdown" for the insulating fluid and to relate this to the voltage level
  - breakdown times tend to decrease as the voltage increases.
- The experiment was run long enough to observe the failure of all the insulation specimens tested.
- If a decision had been made in the preceding experiment to terminate testing after 180 minutes had elapsed, then two of the observations in the 26- and 28-kV sample and one each in the 30- and 32-kV samples would have been censored.
  - In each case, we would not know the exact failure time of the item, but only that it exceeded 180 minutes.

Example 1.1.7

- Gehan (1965) have discussed the results of a clinical trial, in which the drug *6-mercaptopurine* (6-MP) was compared to a placebo with respect to the ability to maintain remission in acute leukemia patients.
- Remission times for two groups of 21 patients each, one group given the placebo and the other the drug 6-MP are available.

Table 1.3. Lengths of Remission (in weeks) for Two Groups of Patients[a]

6-MP	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

[a] Stars denote censored observations.

Subsection 2

1.2 Lifetime Distributions

Distribution function

- Let T be a nonnegative random variable representing lifetimes of individuals in some population
- Let $f(t)$ be the probability density function (*pdf*) of T and the cumulative density function (*cdf*) of T can be defined as

$$F(t) = \Pr(T \leq t) = \int_0^t f(x)\, dx \tag{1}$$

Survivor function

- The probability that an individual survives to time t is given by the survivor function

$$S(t) = \Pr(T \geq t) = \int_t^{\infty} f(x)\, dx \tag{2}$$

- In some context involving lifetimes of manufactured items, $S(t)$ is referred to as the *reliability function*
- The survivor function $S(t)$ is a monotone decreasing function with

$$S(0) = 1 \text{ and } S(\infty) = \lim_{t \to \infty} S(t) = 0.$$

Survivor function

- A useful relationship

$$E(T) = \int_0^\infty S(x)\, dx \tag{3}$$

  - The mean survival time is the area under the survivor curve

Quantiles

- The p^{th} quantile of the distribution of T is the value t_p such that

$$\Pr(T \leq t_p) = F(t_p) = p \quad \Rightarrow \quad t_p = F^{-1}(p) = S^{-1}(1-p)$$

  - The p^{th} quantile t_p is also known as $100p^{\text{th}}$ percentile.
  - The 0.5 quantile $t_{.5}$ is called the median of the distribution.

Mortality rate

- In life table, the mortality rate at time t is the proportion of population who die between times t and $(t+1)$ among individuals alive at time t

$$q_{\mathrm{t}} = \mathsf{Pr}(t \leq T < t+1 \,|\, T \geq t) \tag{4}$$

  - Mortality rate is probability and it lies between 0 and 1

- Calculating the mortality rate for ever smaller intervals of time results in the hazard function (also called hazard rate), $h(t)$.

Hazard function

- Hazard function $h(t)$ is an important concept for lifetime distributions and it can be defined as the limit of the mortality rate

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{5}$$

- Hazard function $h(t)$ specifies instantaneous rate of failure at time t given that the individual survives up to time t.

Hazard function

- Hazard function is conditional failure rate not probability (so can take any positive value, i.e. between 0 to ∞, unlike the mortality rate which is bounded by one)
- For a mortality rate

$$P = \Pr(t \leq T < t + \Delta t \,|\, T \geq t) = 1/4,$$

the corresponding hazard rate depends on the length of the interval Δt

P	Δt	$(P/\Delta t) =$ rate
$\frac{1}{4}$	$\frac{1}{3}$day	$\frac{1/4}{1/3} = 0.75$/day
$\frac{1}{4}$	$\frac{1}{21}$week	$\frac{1/4}{1/21} = 5.25$/week

Hazard function

- The probability that an individual fails in $[t, t + \Delta t)$ given that the individual survives up to time t is approximated by

$$h(t)\Delta t \approx \Pr(t \leq T < t + \Delta t \mid T \geq t) \tag{6}$$

- The $h(t)$ can also be regarded as the expected number of events experienced by an individual in unit time, given that the event has not occurred before then, and assuming that the hazard is constant over that time period.
- Hazard functions are sometimes given other names, such as conditional failure rate, hazard rate, force of mortality, etc.

Relationship between different functions

- The functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ give mathematically equivalent specifications of the distribution of T
- From a given expression of one function, say hazard function, expressions of other functions (e.g. density function) can be derived

Relationship between different functions

Expressing $S(t)$ in terms of $h(t)$

Relationship between different functions

Expressing $S(t)$ in terms of $h(t)$

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

Relationship between different functions

Expressing $S(t)$ in terms of $h(t)$

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

$$\int_0^t h(x)\,dx = \int_0^t \Big[-\frac{d}{dx}\log S(x)\Big]dx$$

Relationship between different functions

Expressing $S(t)$ in terms of $h(t)$

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

$$\int_0^t h(x)\,dx = \int_0^t \Big[-\frac{d}{dx}\log S(x)\Big]dx$$

$$-\int_0^t h(x)\,dx = \log S(x)\Big|_0^t$$

$$-\int_0^t h(x)\,dx = \log S(t) - \log S(0)$$

Relationship between different functions

Expressing $S(t)$ in terms of $h(t)$

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

$$\int_0^t h(x)\,dx = \int_0^t \Big[-\frac{d}{dx}\log S(x)\Big]dx$$

$$-\int_0^t h(x)\,dx = \log S(x)\Big|_0^t$$

$$-\int_0^t h(x)\,dx = \log S(t) - \log S(0)$$

Note $S(\infty) = 0$ and $S(0) = 1$

Relationship between different functions

Expressing $S(t)$ in terms of $h(t)$

$$-\int_0^t h(x)\,dx = \log S(t) \;\Rightarrow\; S(t) = \exp\Big(-\int_0^t h(x)\,dx\Big)$$

It is useful to define the *cumulative hazard function* as

$$H(t) = \int_0^t h(x)\,dx \tag{7}$$

Cumulative hazard function

- Relationship between $S(t)$ and $H(t)$

$$S(t) = \exp\Big(-H(t)\Big) \Rightarrow H(t) = -\log S(t)$$

  - $S(\infty) = 0 \Rightarrow H(\infty) = \infty$

- For a given time t, the greater the risk, the smaller $S(t)$, and hence the shorter mean survival time $E(T)$, and vice verse

Cumulative hazard function

- It is possible for the cumulative hazard function to exceed unity

$$H(t) \geq 1 \;\Rightarrow\; -\log S(t) \geq 1 \;\Rightarrow\; S(t) \leq e^{-1} = 0.368$$

- The cumulative hazard is then greater than unity when the probability of an event occurring after time t is less than 0.37

Relationship between different functions

Expressing $f(t)$ in terms of $h(t)$

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

$$f(t) = h(t)S(t) = h(t)\ \exp\Big(-\int_0^t h(x)\,dx\Big)$$

Example 1.2.1

- Suppose T has p.d.f.

$$f(t) = \beta t^{\beta-1} \exp(-t^{\beta}),\ t > 0$$

  - Obtain survivor function and hazard function of T

Discrete models

- Sometimes, lifetimes are grouped or measured as a number of cycles of some sort
- In such situations, T may be treated as a discrete random variable
- Let T can take on values $t_1, t_2, \ldots,$ with

$$0 = t_0 \leq t_1 < t_2 < t_3 < \cdots$$

Discrete models

- The probability mass function

$$f(t_j) = \mathsf{Pr}(T = t_j), \quad j = 1, 2, ...$$

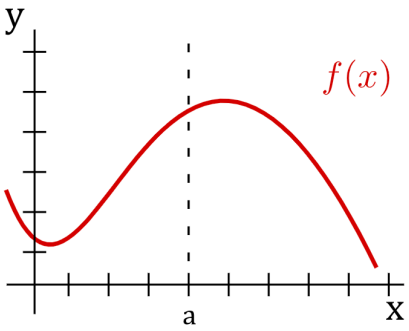
- The survivor function

$$S(t) = \mathsf{Pr}(T \geq t) = \sum_{j:t_j \geq t} f(t_j) \tag{8}$$

- When considered as a function for all $t \geq 0$, $S(t)$ is left continuous, nonincreasing step function, with

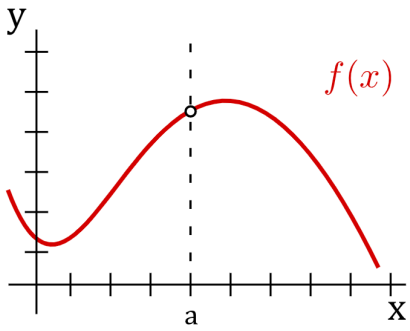
$$S(0) = 1 \text{ and } S(\infty) = 0$$

Continuity

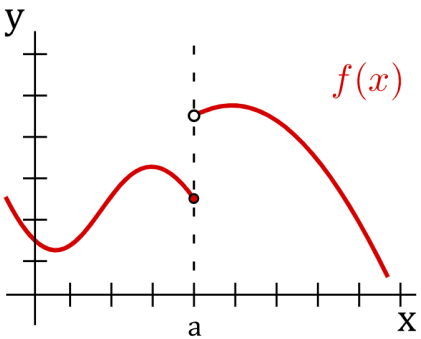


continuous at $x = a$

$\left(\lim_{x \to a} f(x) = f(a)\right)$

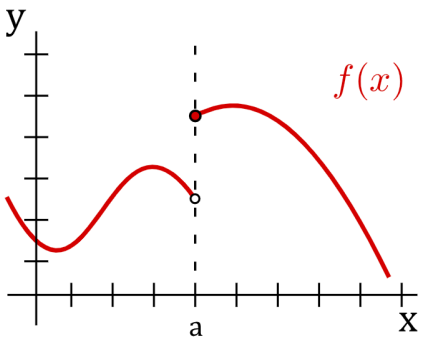


f(a) not defined



$\lim_{x \to a^-} f(x) = f(a)$

left continuous at $x = a$



$\lim_{x \to a^+} f(x) = f(a)$

right continuous at $x = a$

Discrete models

- The discrete time hazard function is defined as

$$\begin{aligned} h(t_j) &= \Pr(T = t_j \mid T \geq t_j) \\ &= \frac{f(t_j)}{S(t_j)}, \quad j = 1, 2, ... \end{aligned}$$

Discrete models

- Using the relationship

$$f(t_j) = S(t_j) - S(t_{j+1}),$$

We can show

$$h(t_j) = \frac{f(t_j)}{S(t_j)} = 1 - \frac{S(t_{j+1})}{S(t_j)}, \quad j = 1, 2, ...$$

- As in the continuous case, discrete hazard function uniquely determines the distribution of the survival time T

Discrete models

Expressing $S(t)$ in terms of $h(t)$

$$S(t) = \prod_{j:t_j<t} [1 - h(t_j)] \tag{9}$$

Expressing $S(t)$ in terms of $h(t)$

$$f(t_j) = h(t_j) \prod_{i=1}^{j-1} [(1 - h(t_i))] \tag{10}$$

Discrete models

- We have

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)} \quad \Rightarrow \quad 1 - h(t_j) = \frac{S(t_{j+1})}{S(t_j)}$$

Discrete models

- Assume t lies between t_j and t_{j+1}, then

$$S(t) = S(t_{j+1}) = S(t_j) \times \frac{S(t_{j+1})}{S(t_j)}$$

Discrete models

- Then

$$S(t) = S(t_j) \times \frac{S(t_{j+1})}{S(t_j)} = S(t_j)\Big[1 - h(t_j)\Big]$$

- That means

$$S(t) = S(t_j)\Big[1 - h(t_j)\Big]$$

Discrete models

- We can write

$$S(t) = S(t_j)\Big[1 - h(t_j)\Big]$$
$$= S(t_{j-1}) \times \frac{S(t_j)}{S(t_{j-1})}\Big[1 - h(t_j)\Big]$$

- That means

$$S(t) = S(t_{j-1})\big[1 - h(t_{j-1})\big]\big[1 - h(t_j)\big]$$

Discrete models

- In general

$$\begin{aligned} S(t) &= S(t_0)[1-h(t_0)][1-h(t_1)]\cdots[1-h(t_{j-1})][1-h(t_j)] \\ &= \prod_{j:t_j<t}[1-h(t_j)] \end{aligned}$$

Discrete models

Expressing $f(t)$ in terms of $h(t)$

$$
\begin{aligned}
f(t_j) &= S(t_j) - S(t_{j+1}) \\
&= \prod_{x:x<t_j} [1 - h(x)] - \prod_{x:x<t_{j+1}} [1 - h(x)] \\
&= \prod_{i=1}^{j-1} [1 - h(t_i)] - \prod_{i=1}^{j} [1 - h(t_i)]
\end{aligned}
$$

Discrete models

Expressing $f(t)$ in terms of $h(t)$

$$
\begin{aligned}
f(t_j) &= S(t_j) - S(t_{j+1}) \\
&= \prod_{i=1}^{j-1}[1 - h(t_i)] - \prod_{i=1}^{j}[1 - h(t_i)] \\
&= \prod_{i=1}^{j-1}[1 - h(t_i)][1 - 1 + h(t_j)] \\
&= h(t_j) \prod_{i=1}^{j-1}[1 - h(t_i)]
\end{aligned}
$$

Discrete models

- The probability that an individual fails at time t_j

$$f(t_j) = h(t_j) \prod_{i=1}^{j-1} [1 - h(t_i)]$$

- The individual survives the preceding discrete failure times $t_1, \ldots, t_{j-1}$ with corresponding (conditional) probabilities $[1 - h(t_1)], \ldots, [1 - h(t_{j-1})]$
- Having survived just before t_j, the individual fails at t_j with probability $h(t_j)$

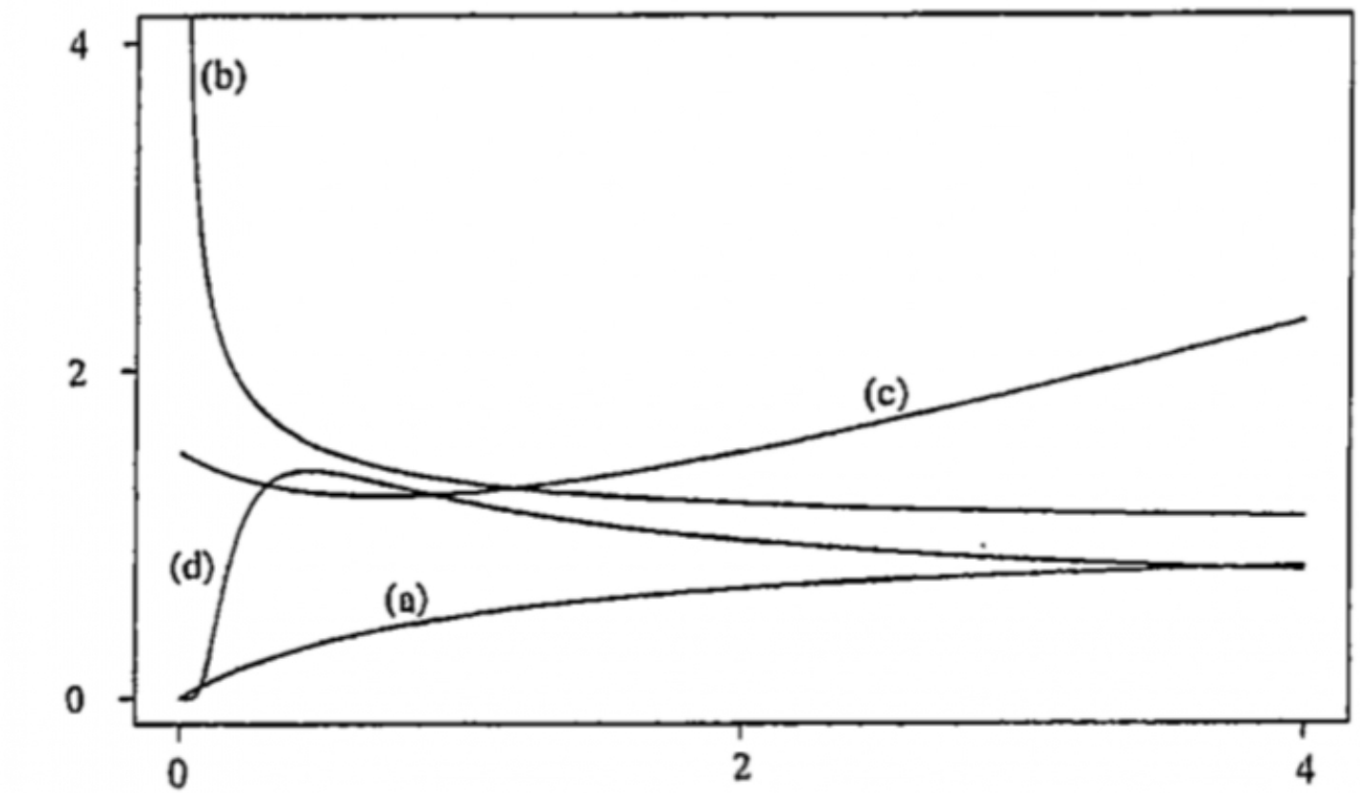
Some remarks on hazard functions

- The hazard function is an important characteristic of a lifetime distribution that indicates the way the risk of failure varies with age or time, and this is of interest in most applications.
- In many instances, information is available on how failure rates change with time and such prior information about the shape of the hazard function can help guide model selection.
- The model/information for hazard function can easily be translated for survivor and density functions using the formulas derived earlier

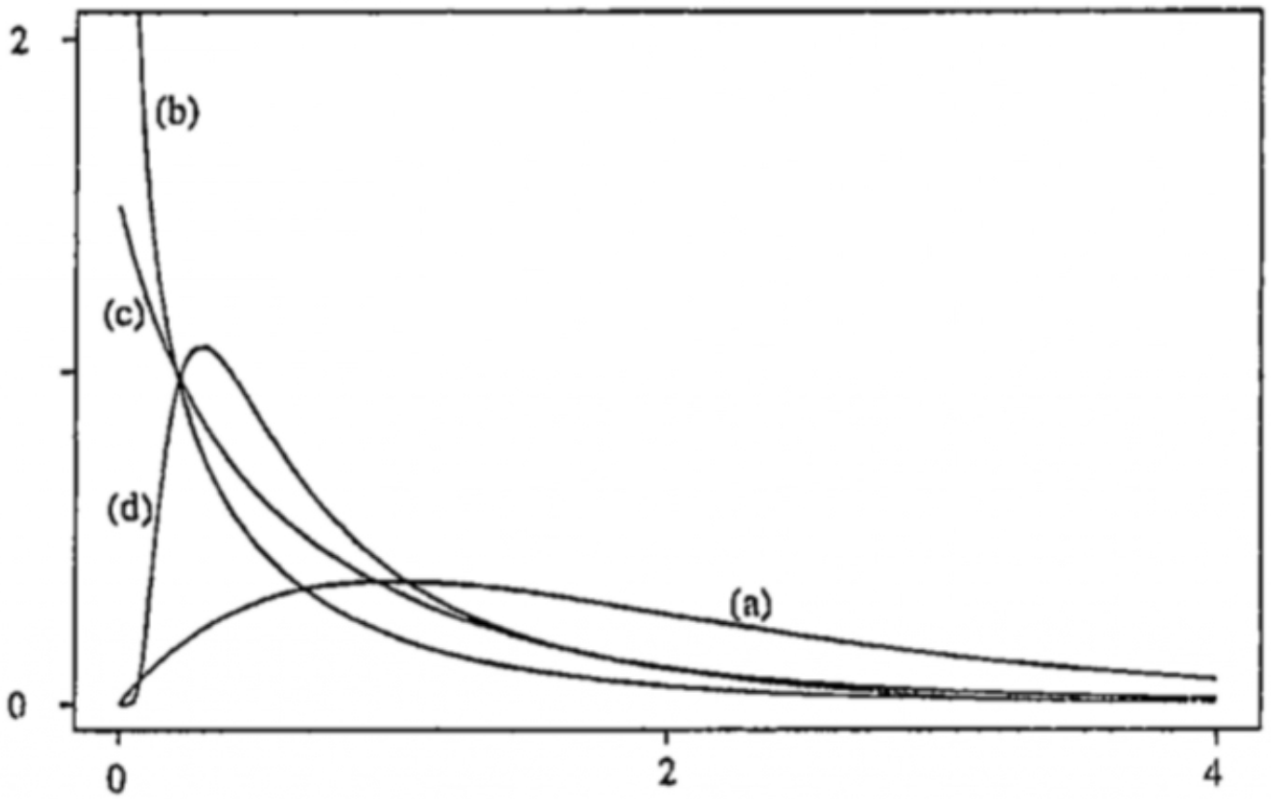
Different shapes of hazard functions

- The shapes of hazard functions could be different, such as
  - monotone increasing (e.g. positive aging) (a)
  - monotone decreasing (e.g. negative aging) (b)
  - bathtub-shaped or U-shaped (e.g. age at death of human populations, lifetime of manufactured items, etc.) (c)
  - inverse bathtub-shaped (e.g. survival after treatment for cancer, duration of marriage, etc.) (d)

Different shapes of hazard functions



Different shapes of density functions



Some remarks on hazard functions

- Shapes of density function could be different corresponding to the shapes of hazard functions
- Although different survivor functions can have the same basic shape, their hazard functions can differ dramatically
- The hazard function is usually more informative about the underlying mechanism of failure than the survivor function.
- Modelling the hazard function is an important method for summarizing survival data

Subsection 3

1.3 Some important failure time models

Introduction

- Various parametric families of models are used in the analysis of lifetime data and a few distributions have the usefulness in a wide-range of situations
- The most commonly used univariate distributions for failure time data
  - exponential, Weibull, log-normal, and log-logistic
- Notations
  - $T \rightarrow$ lifetime, takes only nonnegative values, i.e. from 0 to ∞
  - $Y = \log T \rightarrow$ log-lifetime, takes any value on the real line, i.e. from $-\infty$ to ∞

The exponential distribution

- The exponential distribution is characterized by a constant **hazard function**

$$h(t) = \lambda,\ t \geq 0$$

  - $\lambda > 0$

- The **cumulative hazard function**

$$H(t) = \int_0^t h(x)\,dx = \int_0^t \lambda\,dx = \lambda t$$

- The **survivor function**

$$\begin{aligned} S(t) &= \exp(-H(t)) \\ &= \exp(-\lambda t) \end{aligned}$$

The exponential distribution

- The **probability distribution function**

$$f(t) = h(t)\, S(t) = \lambda \exp(-\lambda t)$$

- Reprametrization $\theta = \lambda^{-1}$ Then, $T \sim \text{Exp}(\text{scale} = \theta)$, where

$$f(t) = (1/\theta)\, \exp(-t/\theta), \;\; t \geq 0$$

The exponential distribution

- Properties
  - $E(T) = \theta$
  - $V(T) = \theta^2$
- **Quantiles**, the pth quantile

$$F(t_p) = p \Rightarrow 1 - \exp(-t/\theta) = p$$
$$\Rightarrow t_p = -\theta \log(1-p)$$

  - The median, $.5th$ quantile

$$t_{.5} = -\theta \log(.5)$$

The exponential distribution

- The exponential distribution with $\theta = 1$ is known as standard exponential distribution
- If $T \sim \text{Exp}(\theta)$ then

$$(T/\theta) \sim \text{Exp}(1)$$

  - The mean and variance of $\text{Exp}(1)$ is 1
  - The median of the $\text{Exp}(1)$ is $-\log(.5) = 0.6931$
  - The density function of $\text{Exp}(1)$ is positively skewed

The exponential distribution

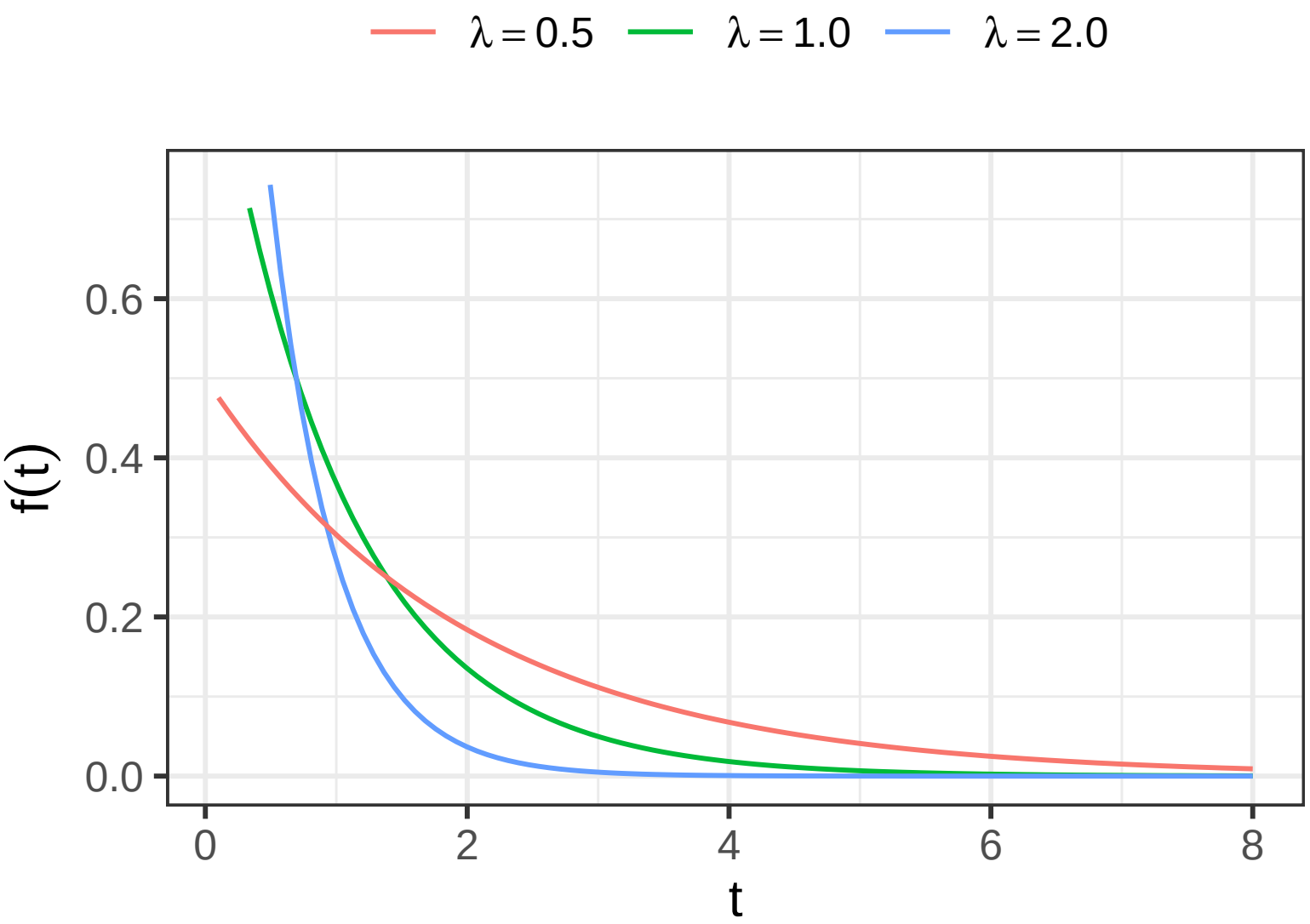


Figure 3: Density function of exponential distribution

The exponential distribution

- Historically, the exponential was the first widely discussed lifetime distribution model
  - This was in part because of the availability of simple statistical methods for it
- The assumption of a constant hazard function is very restrictive, so the model's applicability is fairly limited

The Weibull distribution

- The Weibull distribution is the most widely used lifetime distribution model.
- It has applications to the lifetimes or durability of manufactured
  - It is used as a model with diverse types of items, such as ball bearings, automobile components, and electrical insulation.
- It is also used in biological and medical applications, for example, in studies on the time to the occurrence of tumors in human populations or in laboratory animals.

The Weibull distribution

- The hazard function of Weibull distribution

$$h(t) = \lambda\beta(\lambda t)^{\beta-1},\ \lambda > 0,\ \beta > 0.$$

- Show that $h(t)$ is
  1. monotone increasing for $\beta > 1$
  2. monotone deccreasing for $\beta < 1$
  3. constant for $\beta = 1$
- Exponential distribution is a special case
  - For $\beta = 1$, Weibull distribution reduces to exponential distribution with $h(t) = \lambda$

The Weibull distribution

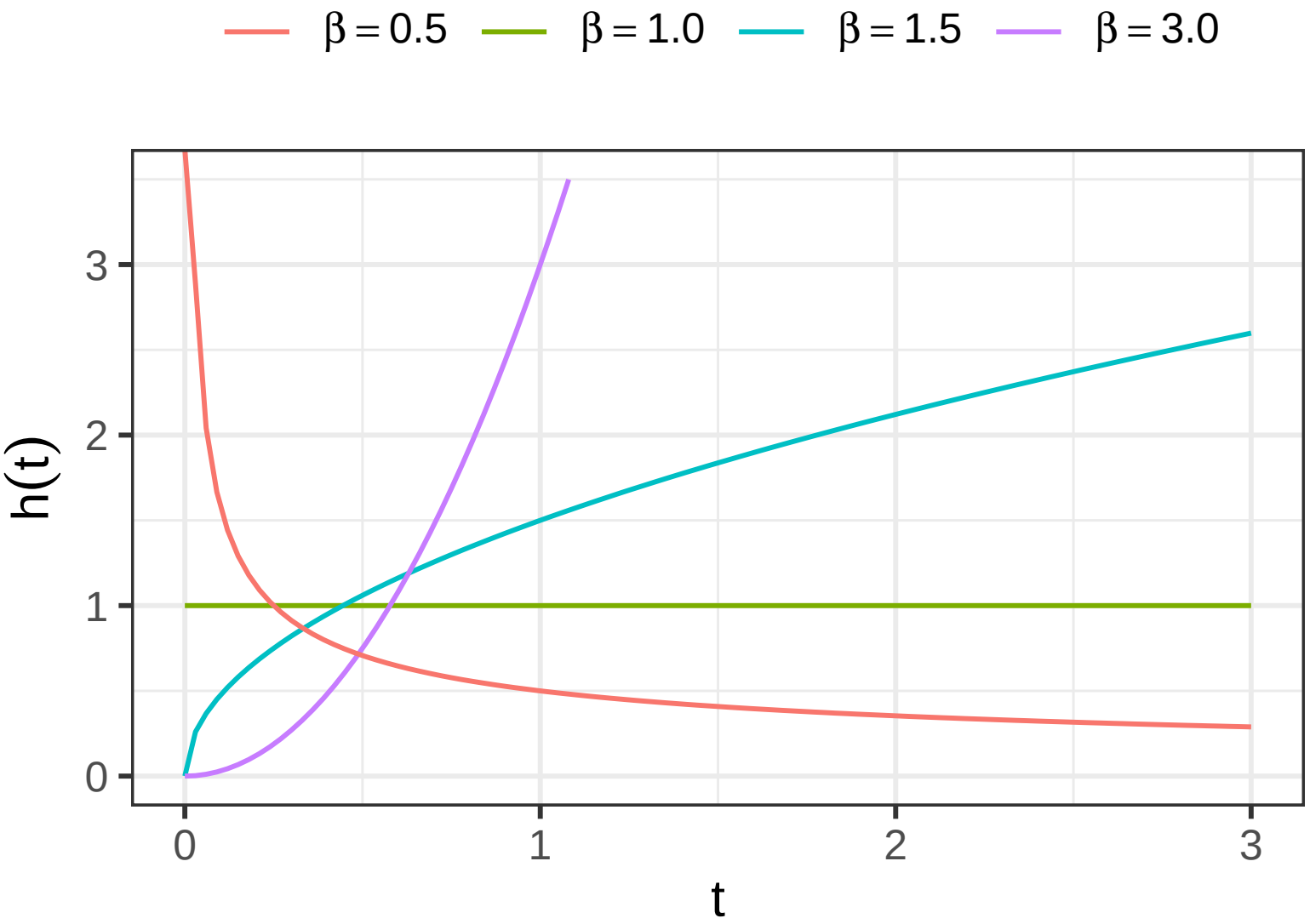


Figure 4: Hazard function of Weibull distribution $(\lambda = 1.0)$

The Weibull distribution

- The cumulative hazard function

$$H(t) = \int_0^t h(x)\,dx = \int_0^t \lambda\beta\,(\lambda x)^{\beta-1}\,dx = (\lambda\,t)^{\beta}$$

- The survivor function

$$S(t) = \exp[-H(t)] = \exp[-(\lambda t)^{\beta}]$$

- The density function

$$f(t) = h(t)\,S(t) = \lambda\beta\,(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}]$$

The Weibull distribution

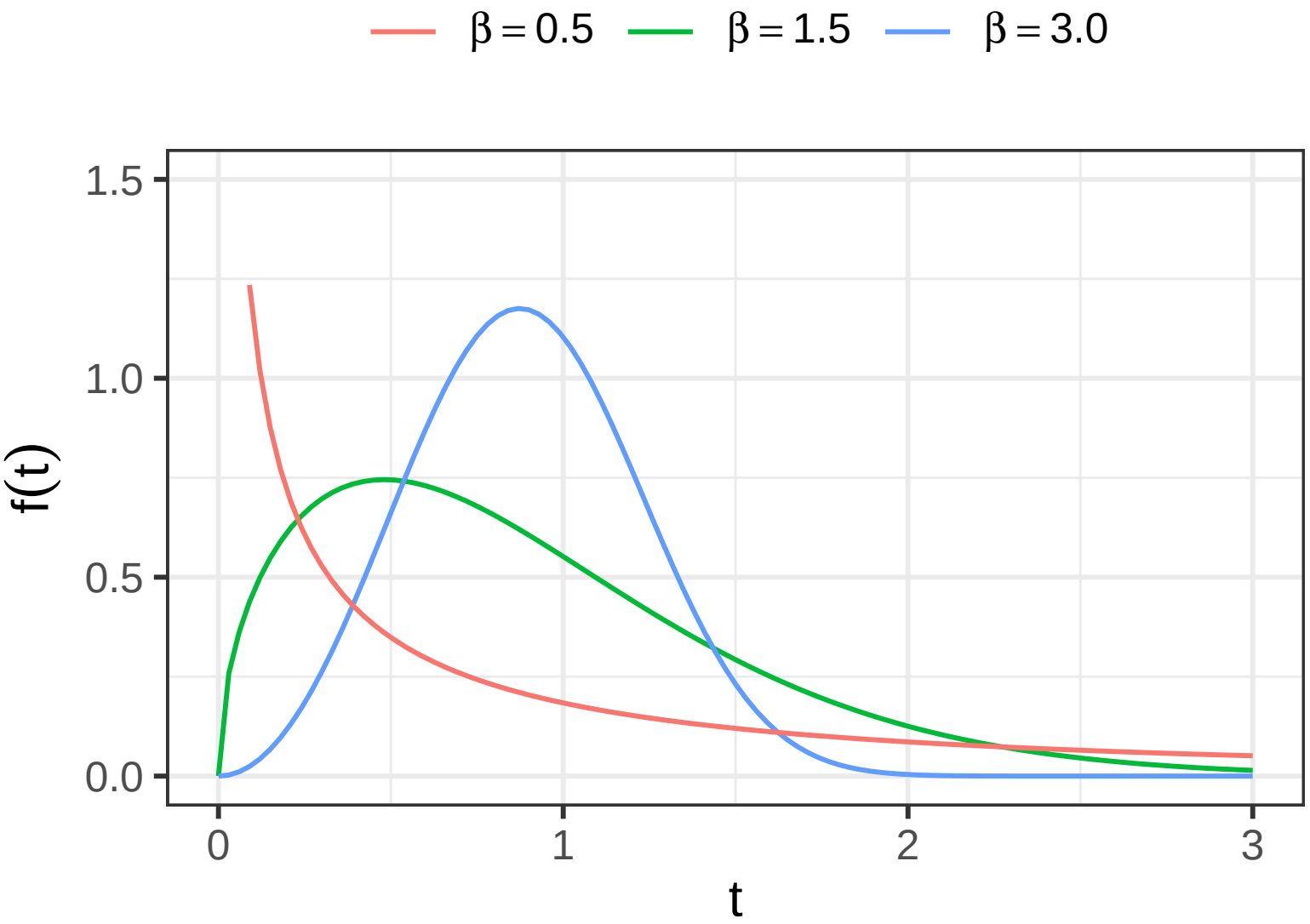


Figure 5: Density function of Weibull distribution $(\lambda = 1.0)$

The Weibull distribution

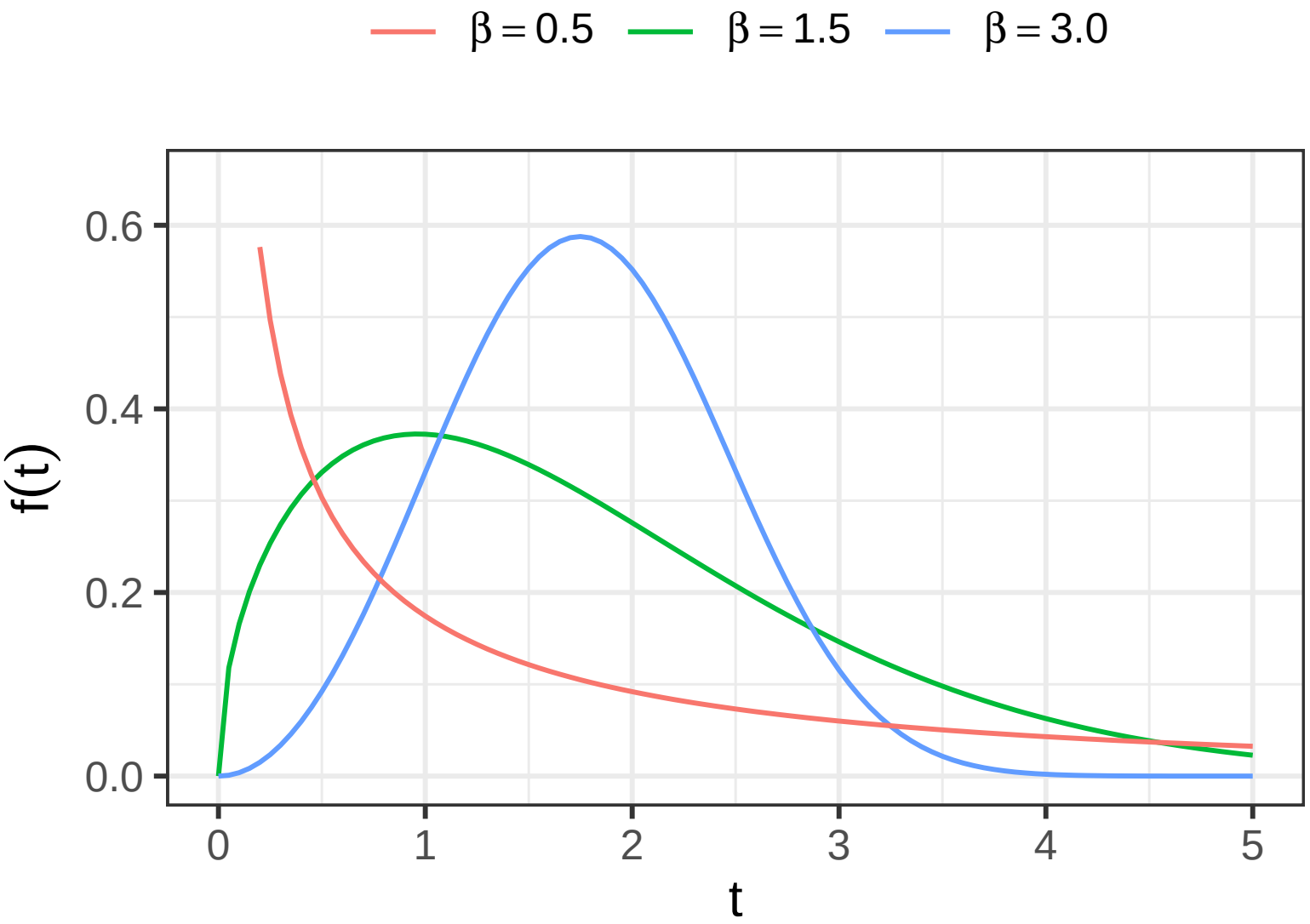


Figure 6: Density function of Weibull distribution $(\lambda = 2.0)$

The Weibull distribution

- Show that the rth moment of Weibull distribution

$$E(T^r) = \lambda^{-r}\Gamma(1 + r/\beta)$$

  - Obtain the expressions of $E(T)$ and $V(T)$

The Weibull distribution

- The p^{th} quantile can be obtained as

$$F(t) = 1 - e^{-(\lambda t_p)^{\beta}} = p \quad \Rightarrow \quad t_p = \alpha[-\log(1-p)]^{1/\beta}$$

  - $\alpha = 1/\lambda$ is known as the scale parameter of the distribution
  - The shape of the distribution depends on β, which is known as the shape parameter
- It can be shown that α is the .632 quantile of the distribution irrespective of the value of β
  - i.e. α is greater than the median of the distribution!

The extreme value distribution

- Let T follows a Weibull distribution

$$T \sim \text{Weib}(\alpha, \beta) \text{ with } \alpha = 1/\lambda$$

- Extreme value distribution (also known as Gumbel distribution) is closely related to Weibull distribution
- If lifetime T follows a Weibull distribution then log-lifetime $Y = \log T$ follows an extreme value distribution
- Extreme value distribution has two parameters, which have one-to-one connection with the Weibull distribution parameters!

The extreme value distribution

- $T \sim \text{Weib}(\alpha, \beta) \quad \Leftrightarrow \quad Y = \log T \sim \text{EV}(u, b)$
  - $u = \log \alpha$ and
  - $b = (1/\beta)$
- The pdf of Y

$$f(y) = (1/b) \exp\left[\frac{y-u}{b} - \exp\left(\frac{y-u}{b}\right)\right] \quad -\infty < y < \infty$$

  - $-\infty < u < \infty$ and $b > 0$

The extreme value distribution

- *Exercise*: obtain the pdf of $Y = \log T$, where $T \sim \text{Weib}(\alpha, \beta)$
  - *Hints.* $J = \frac{dt}{dy} = e^y$ and

$$f_Y(y) = f_T(e^y) \mid J \mid$$

The extreme value distribution

- The survivor function

$$\begin{aligned} S(y) &= \int_{y}^{\infty} f(x\,dx) \\ &= \int_{y}^{-\infty} (1/b) \exp\left[\frac{x-u}{b} - \exp\left(\frac{x-u}{b}\right)\right] dx \\ &= \exp\left[-\exp\left(\frac{y-u}{b}\right)\right] \end{aligned}$$

- The cumulative hazard function

$$H(y) = \exp\left(\frac{y-u}{b}\right)$$

- The hazard function

$$h(y) = \frac{d\,H(y)}{dy} = (1/b) \exp\left(\frac{y-u}{b}\right)$$

The extreme value distribution

Standard extreme value distribution

- If $Y \sim \text{EV}(u, b)$, then

$$\frac{Y - u}{b} \sim \text{EV}(0, 1),$$

the *standard extreme value distribution*.

The extreme value distribution

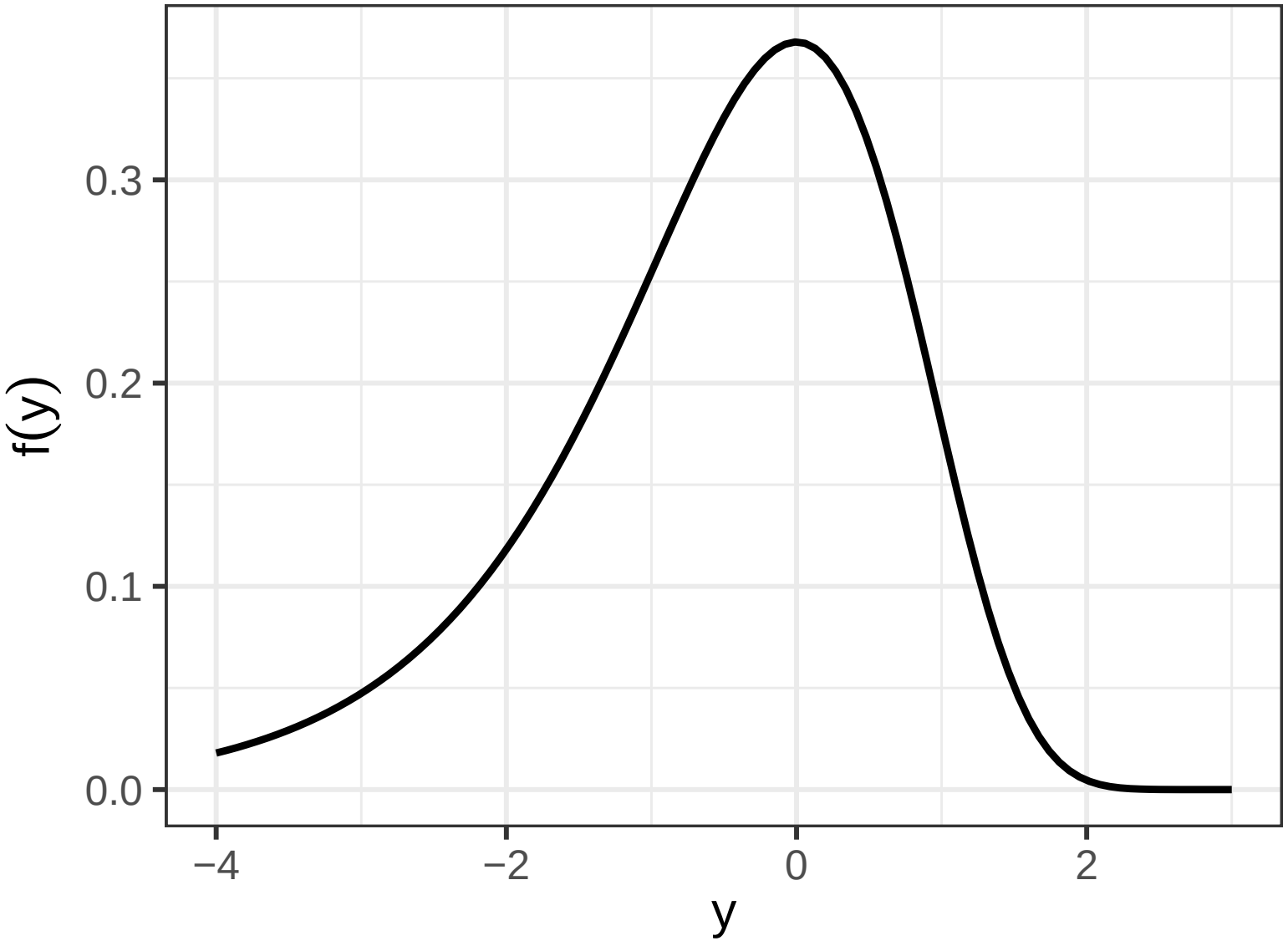


Figure 7: The density function of extreme value distribution with $u = 0$ and $b = 1$

The extreme value distribution

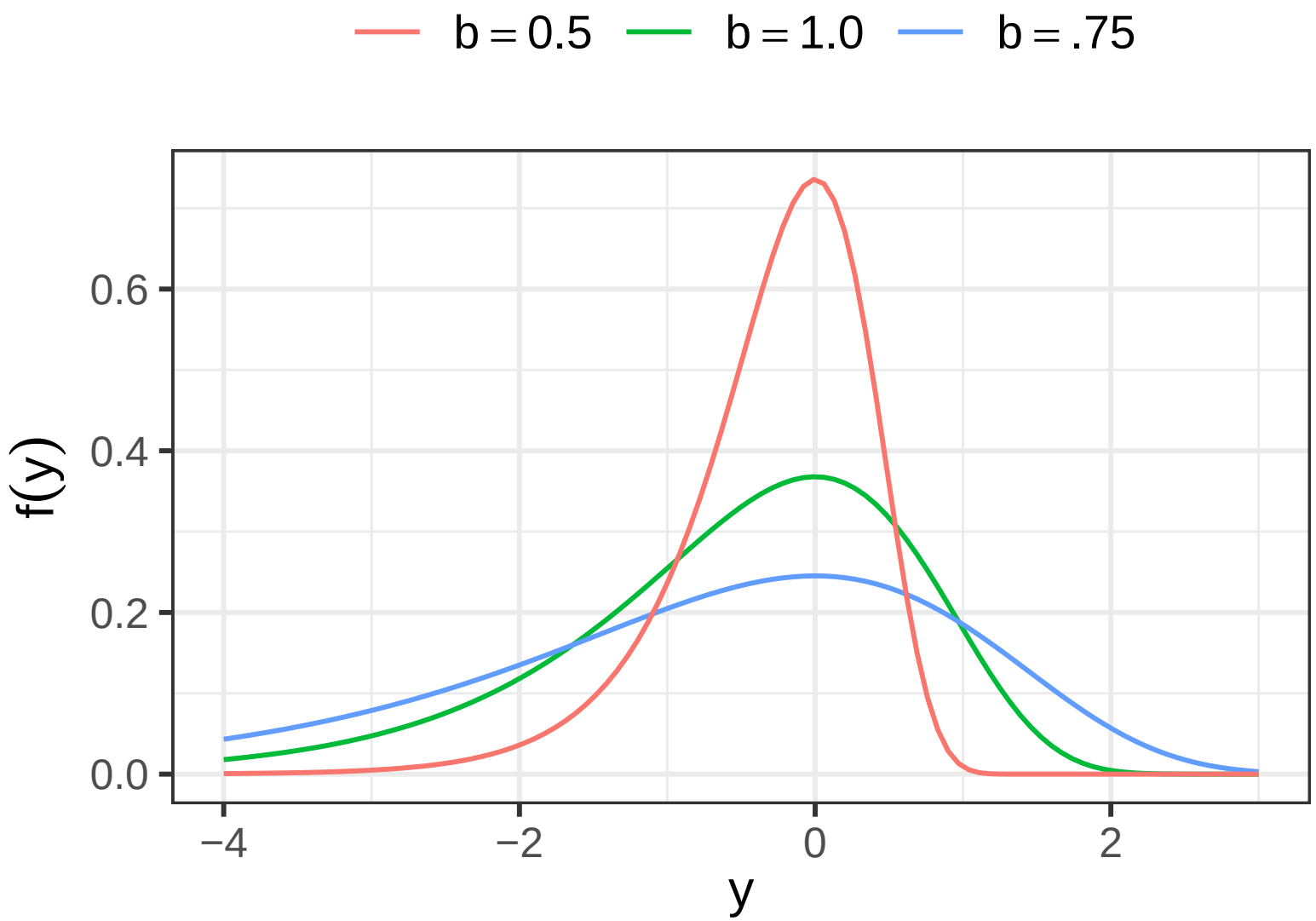


Figure 8: The density function of extreme value distribution with $u = 0$

The extreme value distribution

- The moment generating function of $EV(u, b)$

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta y} f(y)\, dy$$

$$= \int_{-\infty}^{\infty} e^{\theta y} (1/b) \exp\left[\frac{y-u}{b} - \exp\left(\frac{y-u}{b}\right)\right] dy$$

  - Let $(y-u)/b = z$

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta(u+bz)} \exp\left[z - \exp(z)\right] dz$$

  - Let $e^z = x$

$$M(\theta) = \int_{0}^{\infty} e^{\theta u}\, x^{\theta b}\, e^{-x}\, dx = e^{\theta u}\, \Gamma(\theta b + 1)$$

The extreme value distribution

- If $Y \sim EV(u, b)$

$$M(\theta) = e^{\theta u}\,\Gamma(\theta b + 1)$$

- If $Y \sim EV(0, 1)$

$$M(\theta) = \Gamma(\theta + 1)$$

The extreme value distribution

- Moments of standard extreme value distribution $Z \sim EV(0,1)$

$$E(Z) = \frac{d}{d\theta} M(\theta)\big|_{\theta=0} = \Gamma'(1) = -\gamma \quad \text{(Euler's constant)}$$
$$V(Z) = \Gamma''(1) - \gamma^2 = \pi^2/6$$

- For $Y \sim EV(u,b)$, show that

$$E(Y) = u - \gamma b \quad \text{and} \quad V(Y) = b^2(\pi^2/6)$$

The extreme value distribution

The pth quantile of extreme value distribution

$$F(y_p) = p$$
$$S(y_p) = 1 - p$$
$$y_p = u + b \log\left[-\log(1-p)\right]$$

- Show that the location parameter u is the .632 quantile of $Y \sim EV(u, b)$

The log-normal distribution

- The lifetime T is said to be log-normally distributed if log-lifetime $Y = \log T$ is normally distributed.
- The parameters of normal distribution μ and σ are also considered as the parameters of log-normal distribution

$$Y = \log T \sim N(\mu, \sigma^2) \;\Rightarrow\; T = \exp(Y) \sim \log N(\mu, \sigma^2)$$

The log-normal distribution

- Let $Y = \log T \sim N(\mu, \sigma^2)$, show that the density function of $T = \exp(Y)$

$$f_T(t) = f_Y(\log T)\left|\frac{dy}{dt}\right|$$

$$= \frac{1}{\sigma t\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right]$$

  - $t > 0$, $\sigma > 0$, and $-\infty < \mu < \infty$

The log-normal distribution

- The survivor function of $T = \exp(Y)$

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

  - $\Phi(\cdot) \rightarrow$ distribution function of $N(0, 1)$

- The hazard function is defined as $f(t)/S(t)$, which takes the value 0 at $t = 0$, increases to a maximum and then decrease, approaching 0 as $t \rightarrow \infty$.

The log-normal distribution

- It can be shown

$$E(T) = \exp(\mu + \sigma^2/2)$$
$$V(T) = [\exp(\sigma^2) - 1][\exp(2\mu + \sigma^2)]$$

- For log-normal distribution
  - $\exp(\mu) \rightarrow$ the scale parameter
  - $1/\sigma \rightarrow$ the shape parameter
- Show that for $T \sim \log N(\mu, \sigma^2)$

$$t_{.5} = \exp(\mu)$$

The log-normal distribution

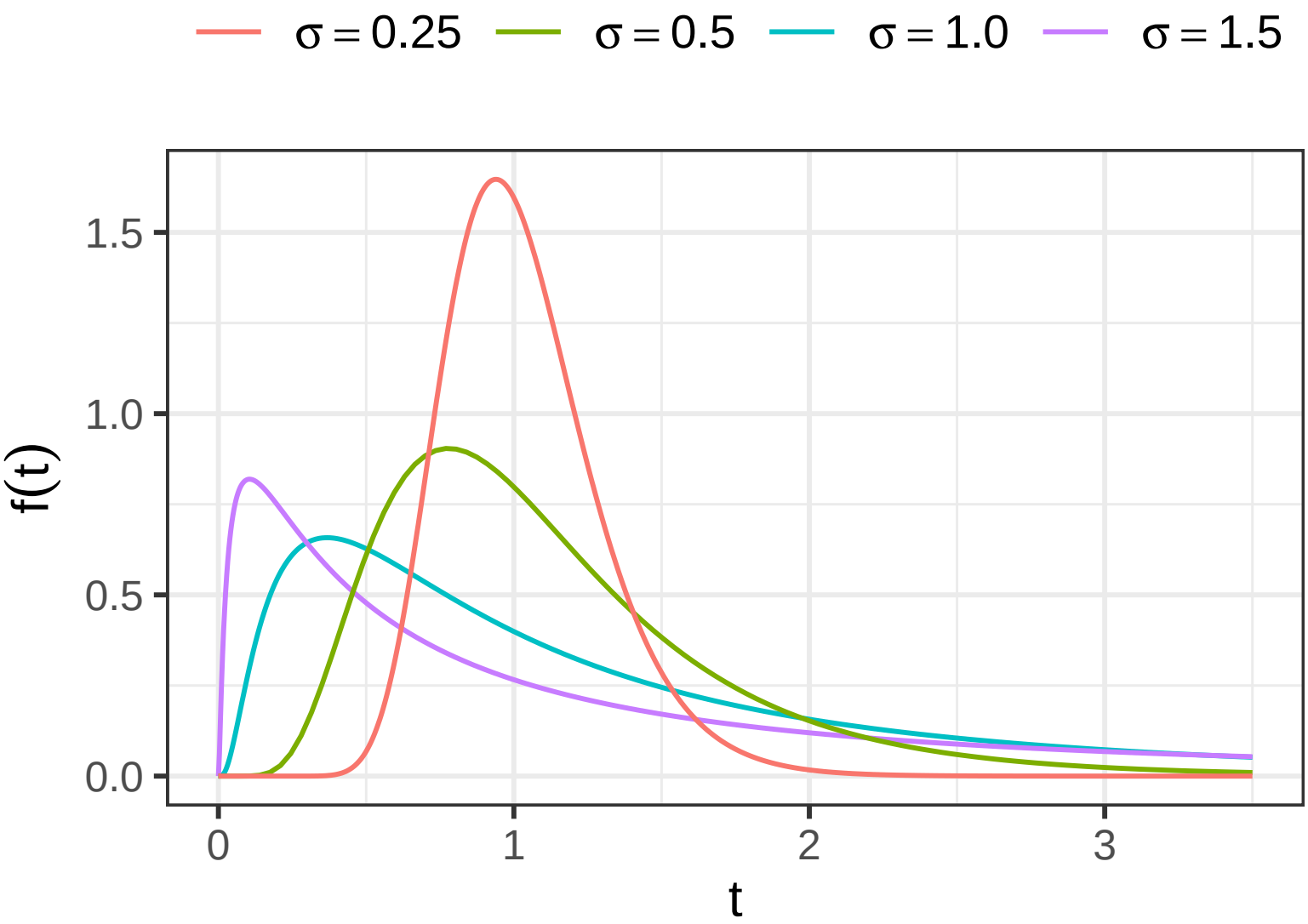


Figure 9: Density function of log-normal distribution with $\mu = 0$

The log-normal distribution

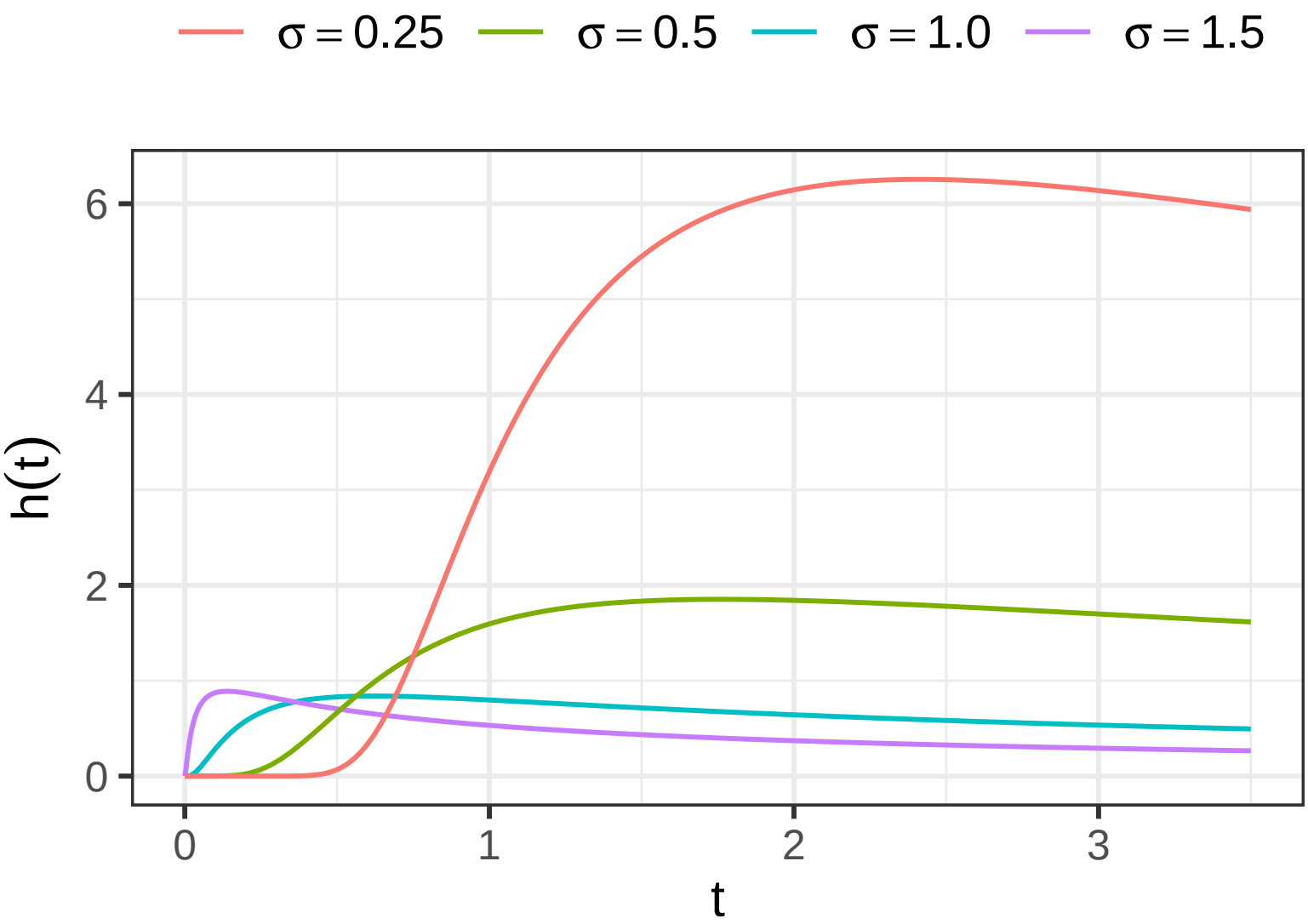


Figure 10: Hazard function of log-normal distribution with $\mu = 0$

The Log-logistic distribution

- If $Y = \log T$ follows a logistic distribution then T follows a log-logistic distribution
- The p.d.f. of a logistic distribution with parameters u and b

$$f(y) = \frac{(1/b)\exp\left[(y-u)/b\right]}{\left\{1+\exp\left[(y-u)/b\right]\right\}^2}$$

  - $-\infty < y < \infty$, $-\infty < u < \infty$, $b > 0$

The Log-logistic distribution

- The survivor function of a logistic distribution

$$S(y) = \frac{1}{\{1 + \exp\left[(y-u)/b\right]\}}$$

- The hazard function of logistic distribution

$$h(y) = \frac{(1/b)\exp\left[(y-u)/b\right]}{\{1 + \exp\left[(y-u)/b\right]\}}$$

The Log-logistic distribution

- The p.d.f. of log-logistic distribution

$$f_T(t) = f_Y(\log T)\left|\frac{dy}{dt}\right|$$
$$= \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{\left[1+(t/\alpha)^{\beta}\right]^2}$$

  - $\alpha = \exp(u) > 0$ and $\beta = 1/b > 0$

The Log-logistic distribution

- The survivor function of $T \sim \text{LLogis}(\alpha, \beta)$

$$S(t) = \int_t^{\infty} \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{\left[1 + (x/\alpha)^{\beta}\right]^2} dx$$

  - Let $(x/\alpha)^{\beta} = y$

$$S(t) = \int_{(t/\alpha)^{\beta}}^{\infty} \frac{1}{(1+y)^2} dy$$

$$= \frac{-1}{1+y}\Bigg|_{(t/\alpha)^{\beta}}^{\infty}$$

$$= \left[1 + (t/\alpha)^{\beta}\right]^{-1}$$

The Log-logistic distribution

- The p.d.f.

$$f(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{\left[1 + (t/\alpha)^{\beta}\right]^2}$$

- The survivor function

$$S(t) = \left[1 + (t/\alpha)^{\beta}\right]^{-1}$$

- The hazard function

$$h(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{\left[1 + (t/\alpha)^{\beta}\right]}$$

The Log-logistic distribution

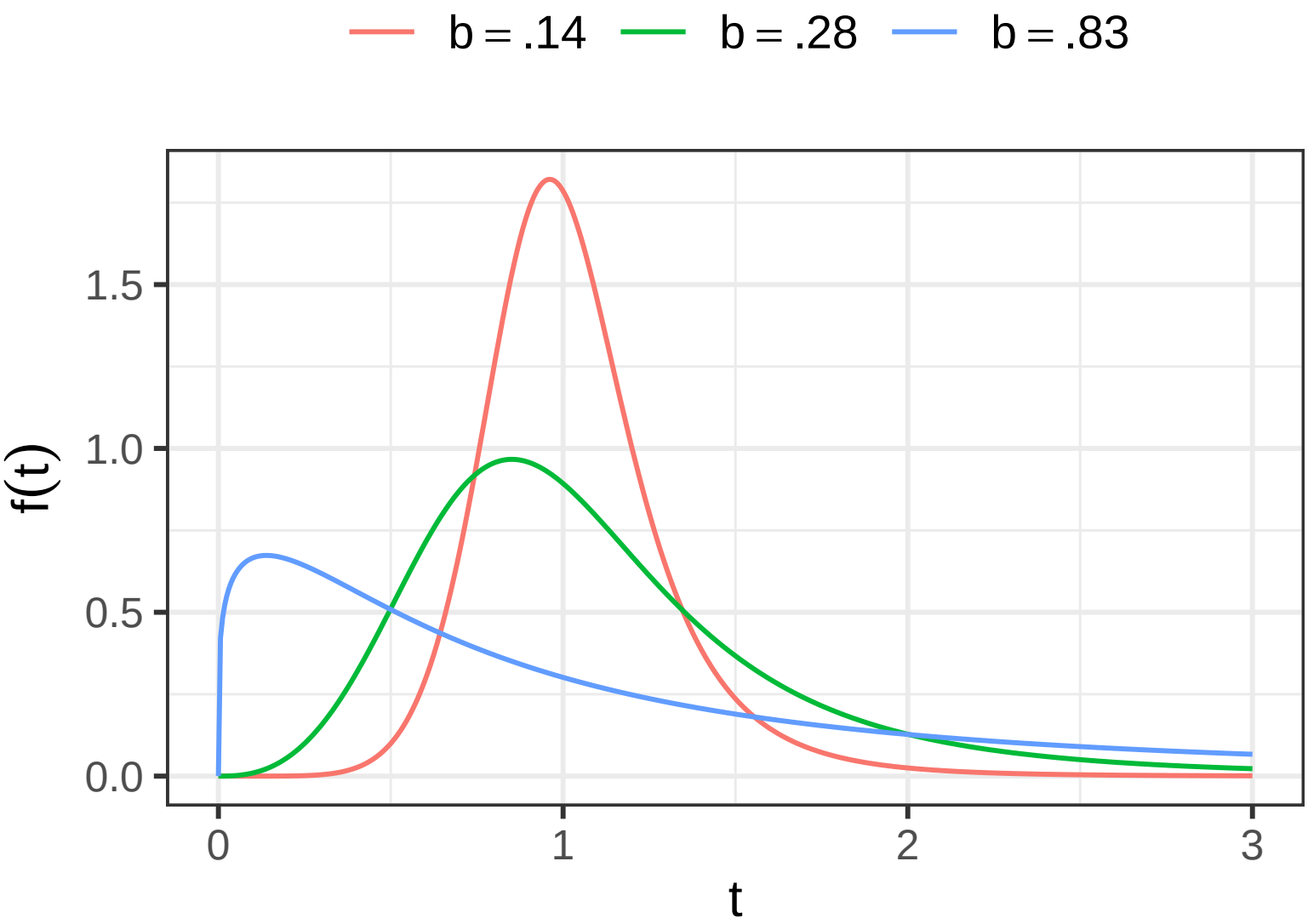


Figure 11: The density function of log-logistic distribution with $u = 0$

The Log-logistic distribution

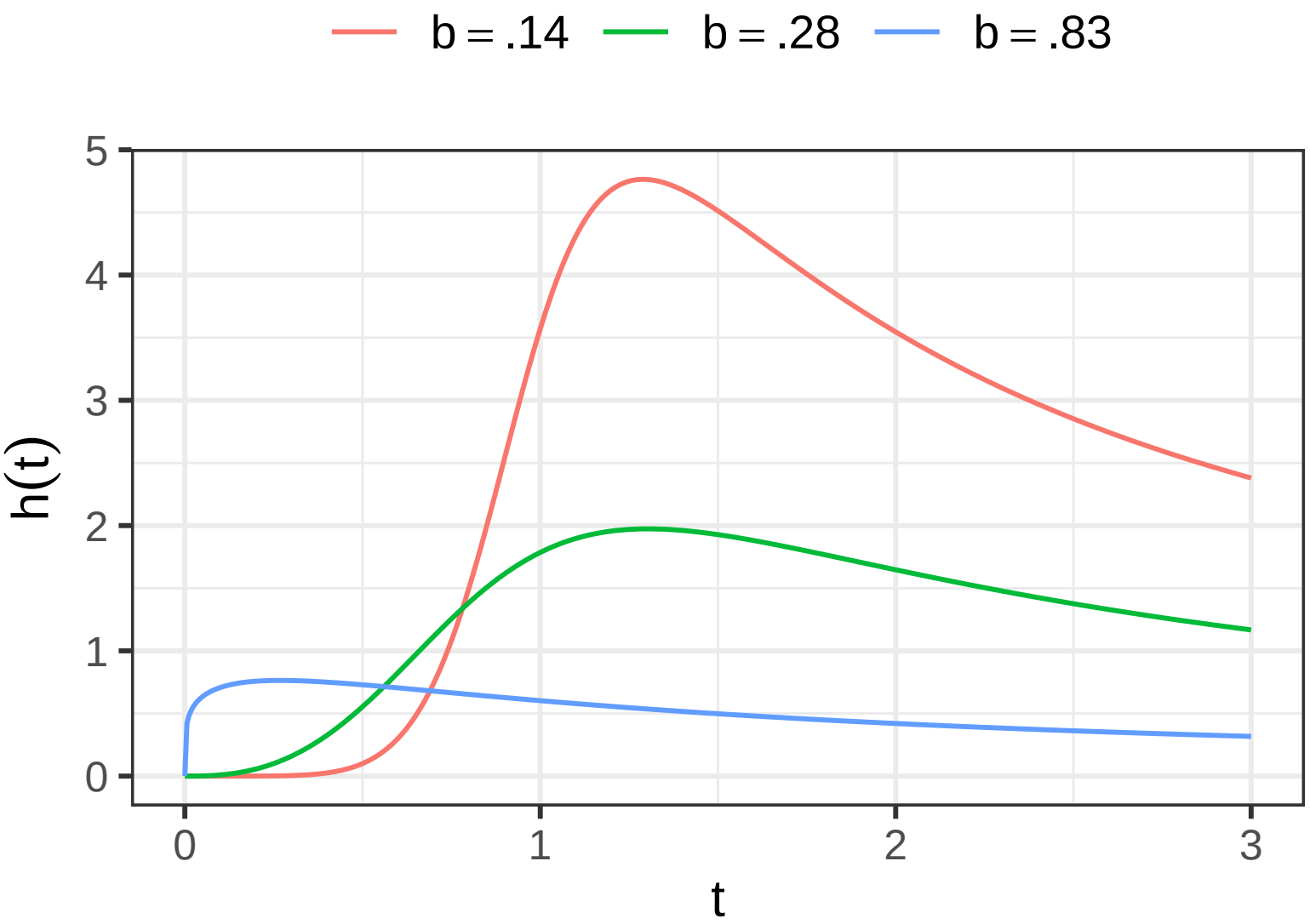


Figure 12: The hazard function of log-logistic distribution with $u = 0$

The Log-logistic distribution

- Show that for $T \sim \text{LLogis}(\alpha, \beta)$, provided $\beta > r$

$$E(T^r) = \alpha^r \, \Gamma(r/\beta + 1)\Gamma(1 - r/\beta)$$

- Beta distribution of the first kind

$$f(x; \alpha, \beta) = \frac{1}{\text{Beta}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \qquad 0 < x < 1$$

- Beta distribution of the second kind

$$f(x; \alpha, \beta) = \frac{1}{\text{Beta}(\alpha, \beta)} \, \frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}} \qquad x > 0$$

The Log-logistic distribution

- The logistic and normal distribution have similar shapes
- For $\beta > 1$, the hazard function of log-logistic distribution has the same characteristic as that of log-normal distribution, i.e. $h(0) = 0$, increases to maximum and then approaches 0 monotonically as $t \to \infty$.
- For $\beta < 1$, the hazard function is monotone decreasing

The gamma distribution

- The gamma distribution has a *pdf* of the form

$$f(t) = \frac{\lambda(\lambda t)^{k-1}\, e^{-\lambda t}}{\Gamma(k)} \quad t > 0$$

  - $k > 0$ and $\lambda > 0$
  - $\lambda^{-1} \rightarrow$ scale parameter
  - $k \rightarrow$ shape parameter
- For $k = 1$, gamma distribution reduces to exponential distribution

The gamma distribution

- Incomplete gamma function

$$I(k,x) = \frac{1}{\Gamma(k)} \int_0^x u^{k-1}\, e^{-u}\, du$$

The gamma distribution

- Survivor function

$$S(t) = \int_t^{\infty} \frac{\lambda(\lambda x)^{k-1} e^{-\lambda x}}{\Gamma(k)} dx$$

  - Let $\lambda x = t$

$$S(t) = \frac{1}{\Gamma(k)} \int_{\lambda t}^{\infty} y^{k-1} e^{-k} dy = 1 - I(k, \lambda t)$$

The gamma distribution

- The hazard function

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

  - For $k > 1$, with $h(0) = 0$ and $\lim_{t \to 0} h(t) = \lambda$
  - $0 < k < 1$, $h(t)$ is monotone decreasing, with

$$\lim_{t \to 0} h(t) = \infty \; and \; \lim_{t \to \infty} h(t) = \lambda$$

The gamma distribution

- The distribution with $\lambda = 1$ is called one-parameter gamma distribution, denoted by $Ga(k)$, and has p.d.f.

$$f(t) = \frac{t^{k-1}\, e^{-t}}{\Gamma(k)} \quad t > 0$$

- If T follows a gamma distribution with scale parameter λ^{-1} and shape parameter k, then show that $\lambda T \sim Ga(k)$
  - *Hints.* $Y = \lambda T$ and $f_Y(y) = f_T(y/\lambda)\,|dt/dy|$

The gamma distribution

- If $Y \sim Ga(k)$ then $2Y \sim \chi^2_{(2k)}$
- Let $T_1, \ldots, T_n$ are independent and identical and exponentially distributed with parameter λ
  - $\sum_{i=1}^{n} T_i$ follows a gamma distribution with parameters λ and n
- The moment generating function of $Y \sim Ga(k)$ is $M(\theta) = (1-\theta)^{-k}$

The gamma distribution

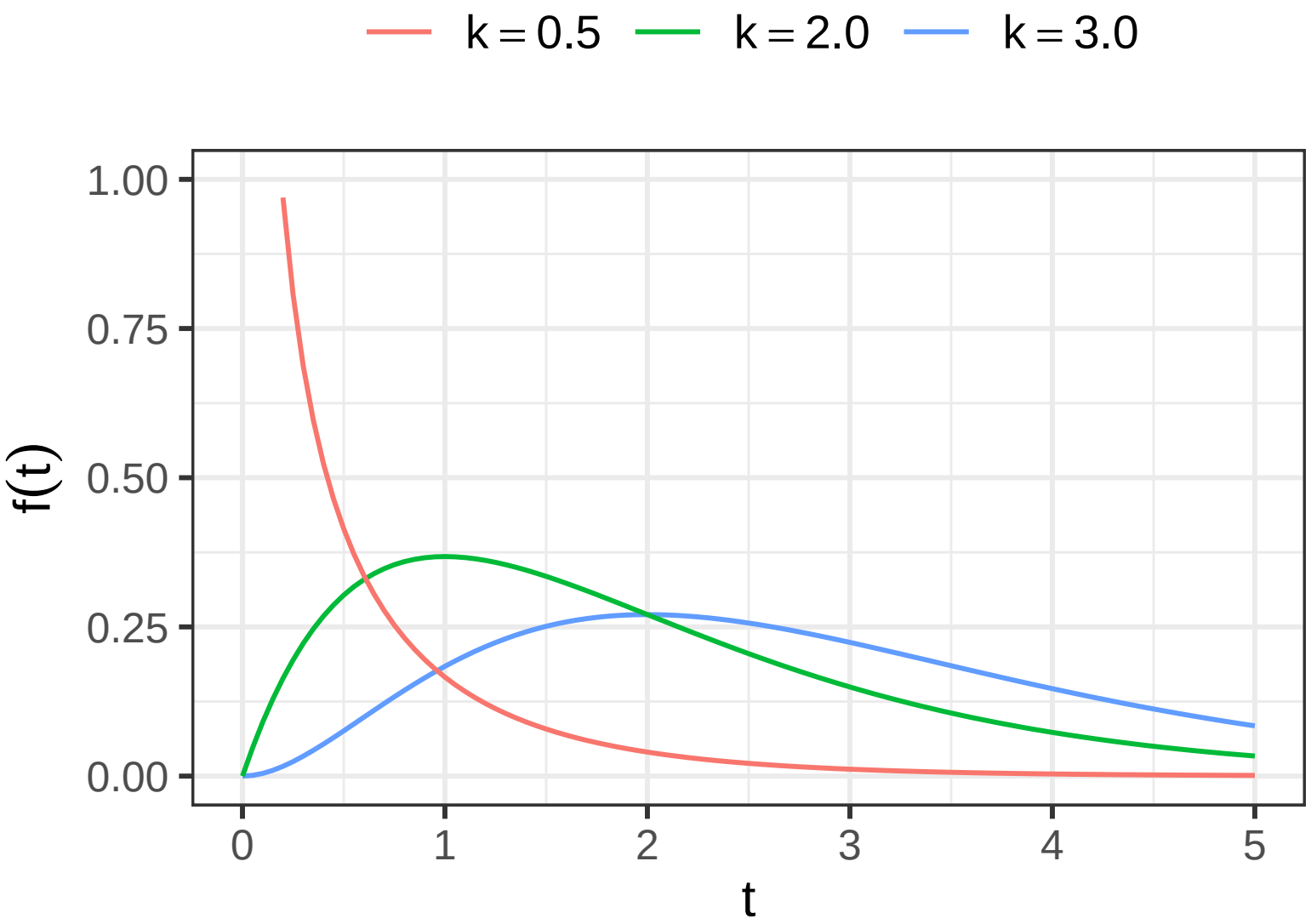


Figure 13: Density function of standard gamma distribution

The gamma distribution

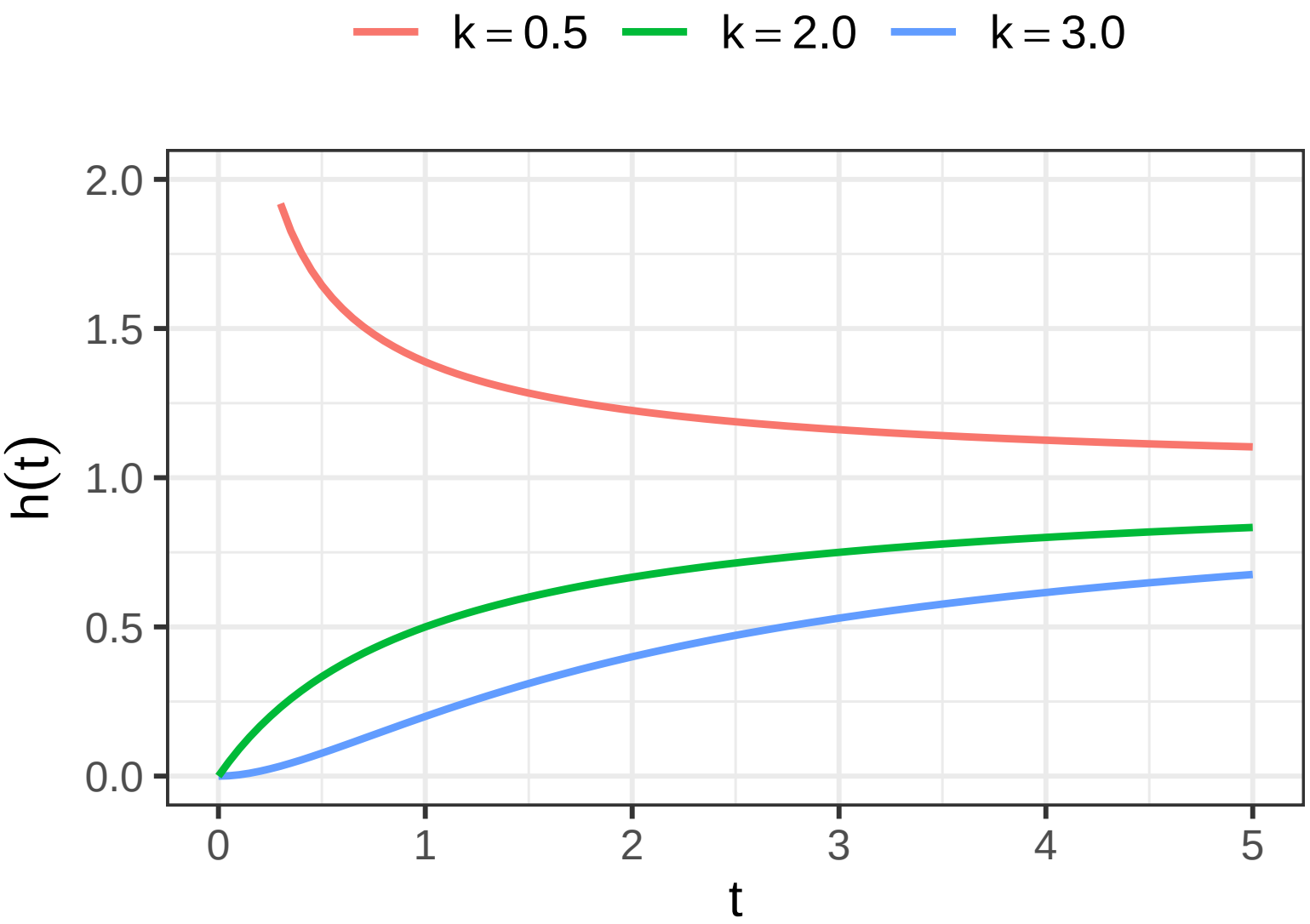


Figure 14: Hazard function of standard gamma distribution

Log-Location-scale models

- A parametric location-scale model for a random variable Y on $(-\infty, \infty)$ is a distribution with p.d.f. of the form

$$f(y) = (1/b)\, f_0\left(\frac{y-u}{b}\right) \quad -\infty < y < \infty$$

  - $-\infty < u < \infty$, location parameter
  - $b > 0$, scale parameter
  - $f_0(z)$ is a specified p.d.f. on $(-\infty, \infty)$

Log-Location-scale models

- The cumulative density function of Y

$$\begin{aligned} F(y) &= \int_{-\infty}^{y} (1/b)\, f_0\left(\frac{x-u}{b}\right) dx \\ &= \int_{-\infty}^{(y-u)/b} f_0(z)\, dz \\ &= F_0\left(\frac{y-u}{b}\right) \end{aligned}$$

- Similarly, the survivor function of Y

$$S(y) = 1 - F_0\left(\frac{y-u}{b}\right) = S_0\left(\frac{y-u}{b}\right)$$

Log-Location-scale models

- The distribution of the standardized variable $Z = (Y - u)/b$
  - Probability density function of Z

$$f_Z(z) = f_Y\left(\frac{y-u}{b}\right)\left|\frac{dy}{dz}\right| = (1/b)\, f_0(z)\,(b) = f_0(z)$$

  - Survivor function of Z

$$S_Z(z) = \int_z^{\infty} f_0(x)\, dx = S_0(z)$$

  - Cumulative density function of Z

$$F_Z(z) = F_0(z)$$

Log-Location-scale models

- There is an one-to-one correspondence between some lifetime and log-lifetime distributions

Lifetime (T)		log-Lifetime (Y)
Weibull	$\longleftrightarrow$	extreme value
log-logistic	$\longleftrightarrow$	logistic
log-normal	$\longleftrightarrow$	normal

- Parameters of *lifetime* distributions

$$\text{scale } (\alpha) \text{ and shape } (\beta)$$

- Parameters of *log-lifetime* distributions

$$\text{location } (u = \log \alpha) \text{ and scale } (b = 1/\beta)$$

Log-Location-scale models

- For the standardized log-lifetimes $Z = (Y - u)/b$
- The density, cumulative density, and survivor functions can be expressed in terms of $f_0(\cdot)$, $F_0(\cdot)$, and $S_0(\cdot)$, respectively
- For example, the survivor functions of log-lifetimes are defined as

$$S_0(z) = \exp(-e^z) \longrightarrow \text{extreme value}$$
$$S_0(z) = 1 - \Phi(z) \longrightarrow \text{normal}$$
$$S_0(z) = (1 + e^z)^{-1} \longrightarrow \text{logistic}$$

Log-Location-scale models

- Using the transformation $T = \exp(Y)$, lifetime distributions can be obtained from each of the distributions of location-scale family

$$\begin{aligned} S_T(t) &= P(T \geq t) \\ &= P(\log T \geq \log t) \\ &= S_0\left(\frac{\log t - u}{b}\right) \\ &= S_0^{\star}\left(\left(\frac{t}{\alpha}\right)^{\beta}\right) \end{aligned}$$

  - $S_0^{\star}(x) = S_0(\log x)$

Log-Location-scale models

- Obtain the survivor function of $T \sim \text{Weib}(\alpha, \beta)$ from $Y \sim EV(u, b)$

$$\begin{aligned} S(t) &= S_0^{\star}\Big((t/\alpha)^{\beta} \Big) \\ &= S_0\Big(\log (t/\alpha)^{\beta} \Big) \\ &= \exp\Big(- e^{\log (t/\alpha)^{\beta}} \Big) \\ &= \exp\Big(- (t/\alpha)^{\beta} \Big) \end{aligned}$$

- Similarly, obtain the expressions of survivor function of log-logistic and log-normal distribution using the relationship $S(\cdot) = S_0^{\star}(\cdot)$

Subsection 4

1.4 Regression models

Regression models

- Regression models are used to understand the relationship between lifetime and a set of covariates (e.g. age, gender, disease status, values of bio-markers, etc.), some of which may depend on time
- Regression models considered for lifetimes can be divided into two broad categories
  - parametric models
  - semiparametric models

Parametric regression models

- Parametric models discussed in this chapter (e.g. Weibull, log-logistic, etc.) can be considered for modeling lifetime
- In parametric regression model, one of the parameters of the assumed lifetime distribution is expressed as a function of available covariates

Parametric regression models

- Let T be the lifetime and $\mathbf{x} = (x_1, \ldots, x_p)'$ be the available p covariates
- Assume $T \sim \text{Exp}(\theta)$ and since $\theta > 0$, a reasonable model for θ would be

$$\theta(\mathbf{x}) = \exp\left(\beta'\mathbf{x}\right), \; where \; \beta = (\beta_1, \ldots, \beta_p)'$$

- The model specification $\theta(\mathbf{x}) = \exp\left(\beta'\mathbf{x}\right)$ ensures $\theta(\mathbf{x}) \geq 0$ for any set of values of β and $\mathbf{x}$
- For the given set of covariates $\mathbf{x}$, the survivor function is defined as

$$S(t \,|\, \mathbf{x}) = \exp(-t/\theta(\mathbf{x}))$$

Parametric regression models

- If $Y = \log T$ follows a distribution of location-scale family, the model $u(\mathbf{x}) = \beta' \mathbf{x}$ would be useful, $-\infty < u(\mathbf{x}) < \infty$
- The corresponding survivor function has the form

$$S_Y(y \,|\, \mathbf{x}) = P(Y \geq y \,|\, \mathbf{x}) = S_0\left(\frac{y - u(\mathbf{x})}{b}\right)$$

  - For example, if $S_0(\cdot)$ is the survivor function of standard normal distribution, then the model $u(\mathbf{x}) = \beta' \mathbf{x}$ represents the multiple linear regression model!

Semiparametric regression models

- In semiparametric regression model, the dependence of Y or T on $\mathbf{x}$ is specified by a parametric function without making any distributional assumption regarding Y or T
- For lifetime data, the most famous semiparametric regression mode is Cox's proportional hazards model (Cox 1972)
- Cox's model cosiders the hazard function of T given $\mathbf{x}$ of the form

$$h(t \,|\, \mathbf{x}) = h_0(t)\, \exp(\beta' \mathbf{x})$$

  - $h_0(t) \longrightarrow$ arbitrary "baseline" hazard function
  - Time-dependent covariates can be included in Cox's proportional hazards model

Exercises

1. Obtain graphs of probability density, survivor, and cumulative hazard functions of the following distributions using R codes.

(a) Weibull distribution with (i) scale parameter 10, and shape parameter 1.5 and (ii) scale parameter 10, and shape parameter 0.95

(b) Logistic distribution with (i) location parameter 10 and scale parameter 1.5 and (ii) location parameter 10 and scale parameter 0.75

Acknowledgements

This lecture is adapted from materials created by Mahbub Latif

References

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187–202.

Gehan, Edmund A. 1965. "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples." *Biometrika* 52 (1-2): 203–24.

Nelson, Wayne. 1972. "Graphical Analysis of Accelerated Life Test Data with the Inverse Power Law Model." *IEEE Transactions on Reliability* 21 (1): 2–11.