Chapter 1

(AST305) Lifetime Data Analysis I

Md Rasel Biswas

Lecture Outline

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

Lifetime Data

- **Lifetime data** have important use in many research areas, including health sciences, engineering, social sciences, etc.
- Applications range from investigating the **durability of manufactured products** to studying **human diseases and their treatments**.
- Lifetime data are also known as "**survival time data**" or "**failure time data**"

Example 1.1.1 – Engineering

- Manufactured items (mechanical or electronic) are often tested for durability.
- Items are operated under controlled conditions and observed until they fail.
- Here, lifetimes are referred to as **“failure times.”**

Example 1.1.2 – Social Sciences

- Demographers study the duration of human life “states.”
- Example: The **duration of marriage** formed in 1980 in a country.
- Marriage ends with **annulment, divorce, or death.**

Example 1.1.3 – Medical Studies

- For patients with potentially fatal diseases, the key outcome is **survival time**, usually measured from diagnosis or treatment.
- Treatments are often compared using the **distribution of survival times.**

Example 1.1.4 – Laboratory Experiments

- In carcinogenic studies, laboratory animals are exposed to substances and followed until tumors appear.
- The primary outcome is **time to tumor appearance**.

Time Origin and Time Scale

Time Origin

- The **zero point** from which survival time is measured for each subject.
- Marks the **start of follow-up**.
- **Examples:** In a medical study, time origin can be the *date of diagnosis*; in a clinical trial, the *date of randomization*; and in an employment study, the *date of recruitment*, all marking when the follow-up clock starts.
- **Key Points**
  - **Must be clearly defined**
  - **Can differ between individuals** (rolling recruitment)

Time Origin and Time Scale

Time Scale

- The **metric used to measure time** from the time origin.
- Determines the unit and meaning of “time” in the model.
- **Common Time Scales:**
    - *Time since study entry* (follow-up duration from study start, in hours, minutes, seconds etc.)
    - *Attained age* (participant’s actual age at each point in follow-up).
- Time need not always be chronological:
    - **Miles driven** for a vehicle until breakdown.
    - **Number of pages printed** for a printer or photocopier.

Censoring

- In practice, full observation of lifetimes may not be feasible.
- If we only know a subject's lifetime **exceeds a certain value**, it is a **censored observation**.
- **Example:** A life test stops after 28 days. If an item has not failed, its lifetime is known only as **"greater than 28 days."**
- **Types of censoring:** right, left, and interval censoring (details in next chapter).

Censoring

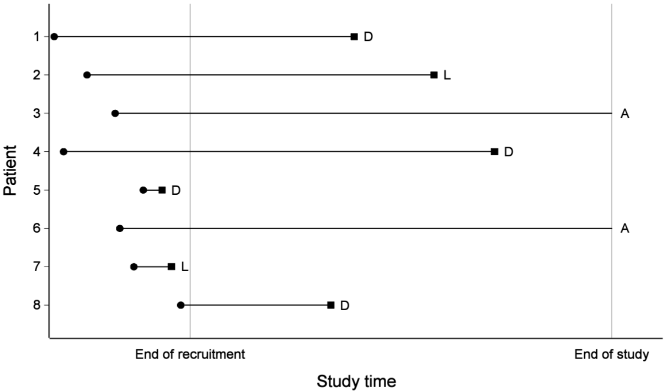


Figure 1: Lifetime of eight subjects in a survival study. D=Death, L=Lost, A=Alive

Example 1.1.5 – Electrical Insulating Fluid

- Nelson (1972) studied electrical insulating fluid under constant voltage stress.
- Failure time (time to breakdown) was recorded.
- Specimens were tested at different voltages (26–38 kV).

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

- Key result: **Higher voltage $\rightarrow$ shorter breakdown times.**
- If testing had stopped at 180 minutes, some observations would have been **censored** (true failure times unknown but >180 minutes).

Example 1.1.7 – Clinical Trial (Leukemia)

- Gehan (1965) compared **6-mercaptopurine (6-MP)** with placebo in acute leukemia patients.
- Outcome: **remission time** – the duration patients remained in remission (a state where symptoms reduce or disappear, but the disease may return later).
- Two groups of 21 patients each (placebo vs. 6-MP).

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

Cumulative Distribution Function (CDF)

- Let T be a **nonnegative random variable** representing lifetimes of individuals in a population.
- The probability density function (*pdf*) of T is denoted by $f(t)$, and the **cumulative distribution function** (*cdf*) of T is defined as

$$F(t) = \Pr(T \leq t) = \int_0^t f(x)\, dx, \quad t \geq 0$$

Survivor Function

- The probability that an individual survives **beyond** time t is

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

- In reliability engineering (e.g., lifetimes of machines), $S(t)$ is also called the **reliability function**.
- Properties of $S(t)$:
  - $S(0) = 1$ (everyone is alive at the start).
  - $S(t)$ is monotone decreasing.
  - $\lim_{t\to\infty} S(t) = 0$.

Survivor Function

- A useful relationship between mean survival and the survivor function:

$$E(T) = \int_0^\infty S(x)\,dx$$

This shows that the **mean survival time** equals the area under the survivor curve.

Quantiles

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

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

- Special cases:
  - t_p is also the $100p^{\text{th}}$ percentile.
  - The 0.5 quantile ($t_{0.5}$) is the **median lifetime**.

Mortality Rate

- In life tables, the **mortality rate** at time t is the probability of dying between t and $t+1$ among those alive at time t:

$$q_t = \Pr(t \leq T < t+1 \,|\, T \geq t)$$

- Properties:
  - Mortality rate is a probability, so $0 \leq q_t \leq 1$.
  - As the interval shrinks, mortality rate leads to the **hazard function**.

Hazard Function

- Defined as the **instantaneous failure rate** at time t, conditional on survival up to t:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}$$

- Key points:
  - Hazard is a **rate, not a probability** $\rightarrow$ it can take values from 0 to ∞.
  - It represents the **instantaneous risk of failure** at time t.

Hazard Function

- Approximation:

$$\Pr(t \leq T < t + \Delta t \mid T \geq t) \approx h(t)\Delta t$$

- Alternative interpretation: for small Δt, $h(t)\,\Delta t$ approximates the **conditional failure probability** over $[t, t + \Delta t)$.
- Other names: *force of mortality*, *conditional failure rate*.

Hazard Function

- Suppose we are told that for a certain interval of length Δt, the probability of failure is

$$P = \Pr(t \leq T < t + \Delta t \mid T \geq t) = \tfrac{1}{4}.$$

- To get the corresponding **hazard rate**, divide by Δt:

$$h(t) \approx \frac{P}{\Delta t}$$

Then the hazard depends on the interval length:

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

Relationship Between Different Functions

- The four functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ are mathematically equivalent ways of describing 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 $h(t)$

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

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

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

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

Since $S(0) = 1$,

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

Cumulative Hazard Function

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

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

- Relationship with survivor function:

$$S(t) = \exp[-H(t)] \quad \Rightarrow \quad H(t) = -\log S(t)$$

- Notes:
  - $S(\infty) = 0 \Rightarrow H(\infty) = \infty$.
  - It is possible for $H(t) > 1$ (since it is not a probability).

Cumulative Hazard Function

- 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
- 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 $h(t)$

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

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

Example 1.2.1

- Suppose T has pdf

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

  - Obtain survivor function and hazard function of T

Discrete Models: Motivation

- Sometimes lifetimes are measured in **cycles or intervals** (e.g., weeks, months, visits).
- Then T is a **discrete random variable**.
- Possible values: $t_1, t_2, ...$, with $0 = t_0 < t_1 < t_2 < \cdots$.

Discrete Probability Functions

- **Probability mass function (pmf):**

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

- **Survivor function:**

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

- Properties:
    - Step function (non-increasing).
    - Left-continuous.
    - $S(0) = 1, \ S(\infty) = 0.$

Note: In continuous-time settings we often write $S(t) = \Pr(T > t)$ (right-continuous); here, with discrete times, $S(t) = \Pr(T \geq t)$ is left-continuous.

Discrete Hazard Function

- Defined as
$$h(t_j) = \Pr(T = t_j \mid T \geq t_j) = \frac{f(t_j)}{S(t_j)}.$$
- Interpretation: probability of failing **at time** t_j given survival up to that time.
- Range: $0 \leq h(t_j) \leq 1$.
- Unlike continuous hazard, this is a **probability, not rate**.

Discrete Hazard Function

- Connection to survivor:

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

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

- Survivor in terms of hazards:

$$S(t) = \prod_{t_j < t} [1 - h(t_j)].$$

Intuition

To survive until t_j: must "not fail" at earlier times $[1 - h(t_1)], [1 - h(t_2)], \dots, [1 - h(t_{j-1})]$.

Discrete Hazard Function

- pmf in terms of hazards:

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

Intuition

To fail exactly at t_j: multiply survival so far by hazard at t_j.

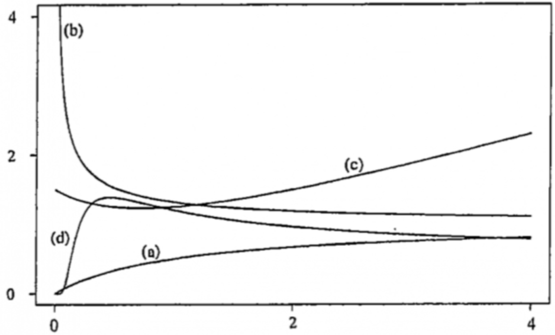
Some Remarks on Hazard Functions

- 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

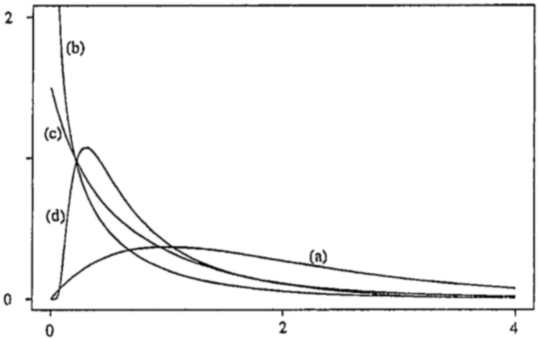
Different Shapes of Hazard Functions

- 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

- In survival analysis, **parametric models** are often used to describe lifetime data.
- Only a few distributions have proven useful across many applications.
- The most common univariate models are:
  - Exponential
  - Weibull
  - Log-normal
  - 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 density function**

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

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

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

The Exponential Distribution

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

$$
\begin{aligned}
F(t_p) = p \;\Rightarrow\; & 1 - \exp(-t_p/\theta) = p \\
\Rightarrow\; & t_p = -\theta \log(1 - p)
\end{aligned}
$$

  - 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 $T \sim$ Exp(θ) then

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

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

The Exponential Distribution

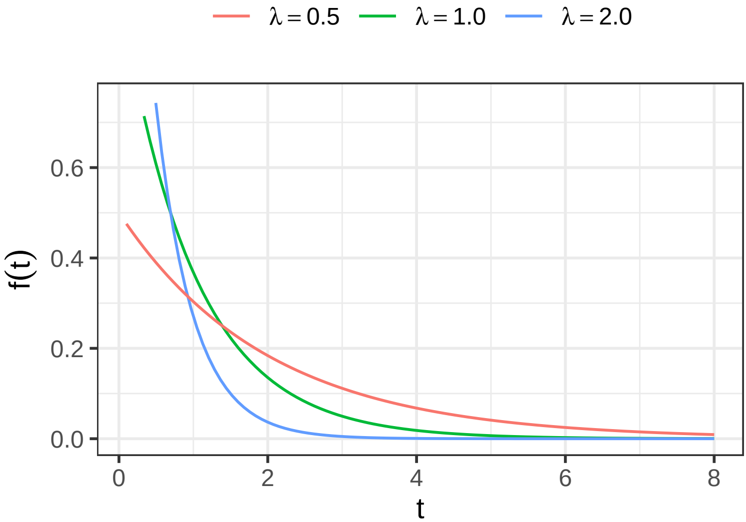


Figure 2: Density function of exponential distribution

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 decreasing 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

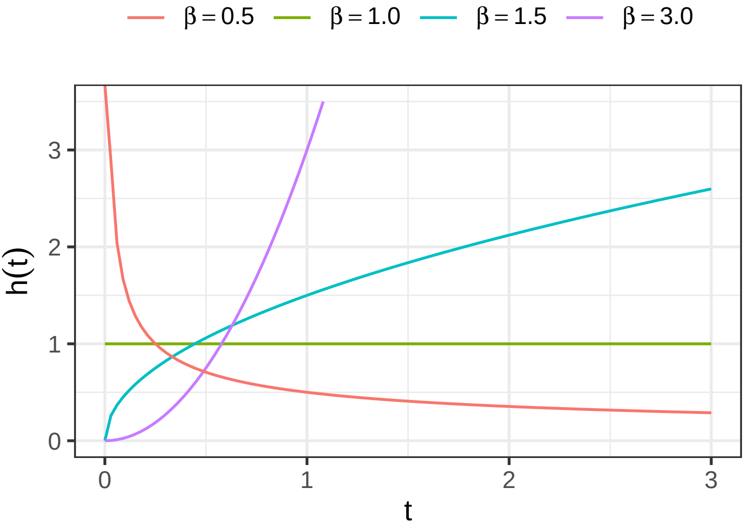


Figure 3: Hazard function of Weibull distribution $(\lambda = 1.0)$

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

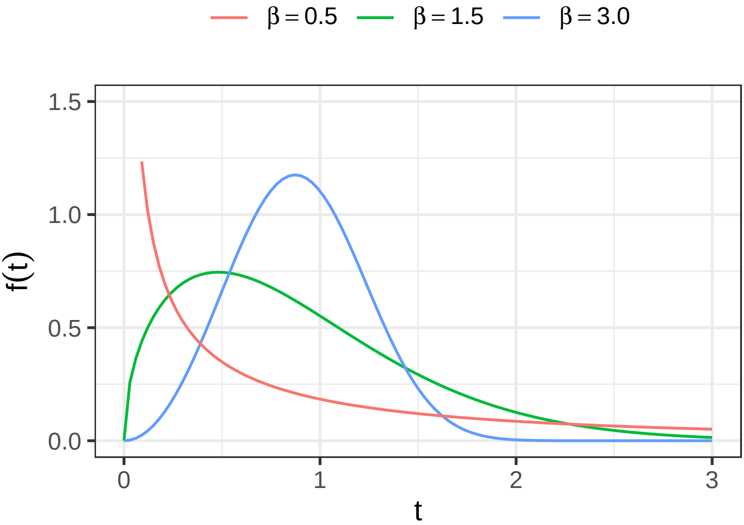


Figure 4: Density function of Weibull distribution $(\lambda = 1.0)$

The Weibull Distribution

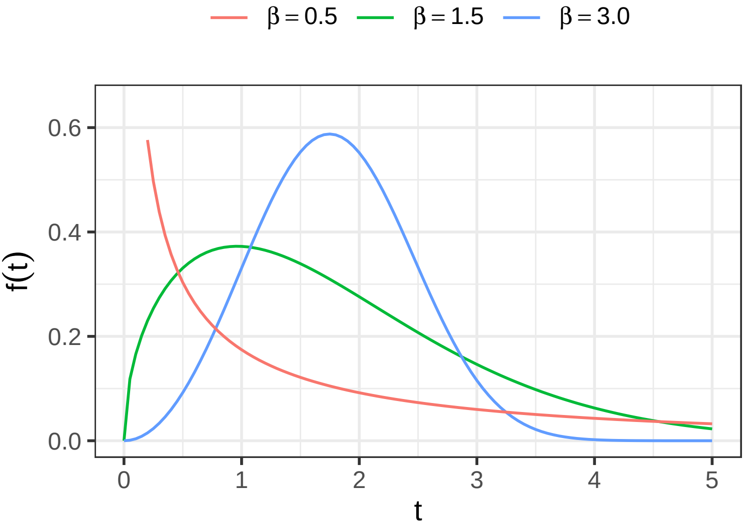


Figure 5: Density function of Weibull distribution $(\lambda = 2.0)$

The Weibull Distribution

- Show that the r^{th} 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

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

  - $\alpha = 1/\lambda$ 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

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

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

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

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

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

the *standard extreme value distribution*.

The Extreme Value Distribution

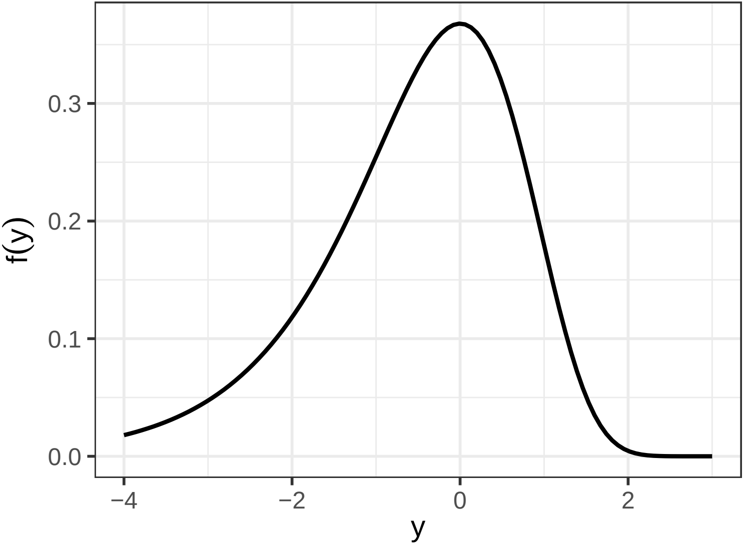


Figure 6: The density function of extreme value distribution with $u = 0$ and $b = 1$

The Extreme Value Distribution

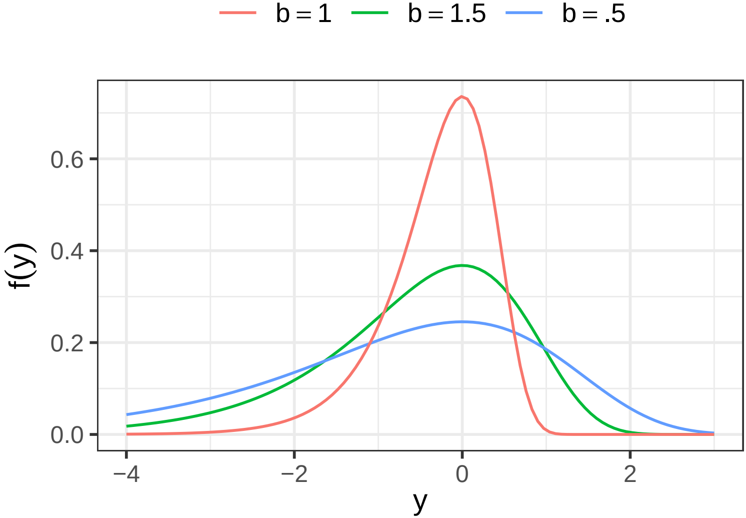


Figure 7: The density function of extreme value distribution with $u = 0$

The Extreme Value Distribution

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

$$M_Y(\theta) = E[e^{\theta Y}] = \int_{-\infty}^{\infty} e^{\theta y} \, f(y) \, dy$$

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

$$= \int_{-\infty}^{\infty} e^{\theta(u+bz)} \exp\left[z - \exp(z)\right] dz \qquad \left[\text{let } \frac{y-u}{b} = z\right]$$

$$= \int_{0}^{\infty} e^{\theta u} \, x^{\theta b} \, e^{-x} \, dx = e^{\theta u} \, \Gamma(\theta b + 1), \qquad [\text{let } e^z = x]$$

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 p^{th} quantile of extreme value distribution

$$
\begin{aligned}
F(y_p) &= p \\
S(y_p) &= 1 - p \\
\exp\left[-\exp\left(\frac{y_p - u}{b}\right)\right] &= (1 - p) \\
y_p &= u + b\log\left[-\log(1 - p)\right]
\end{aligned}
$$

- 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

$$\begin{aligned} Y &= \log T \sim N(\mu, \sigma^2) \\ \Rightarrow\ T &= \exp(Y) \sim \log N(\mu, \sigma^2) \end{aligned}$$

The Log-normal Distribution

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

$$f_T(t) = f_Y(\log T)\left|\frac{dy}{dt}\right|$$
$$= \frac{1}{\sigma t\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right]$$

  - $t > 0$, $\sigma > 0$, 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 decreases, 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

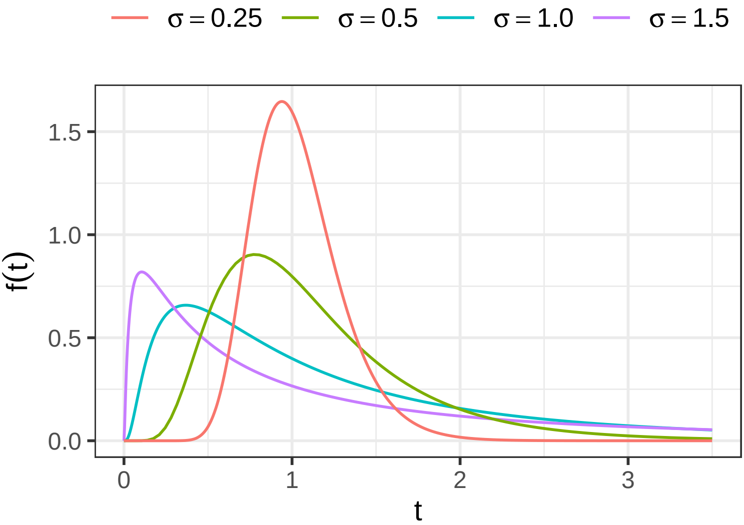


Figure 8: Density function of log-normal distribution with $\mu = 0$

The Log-normal Distribution

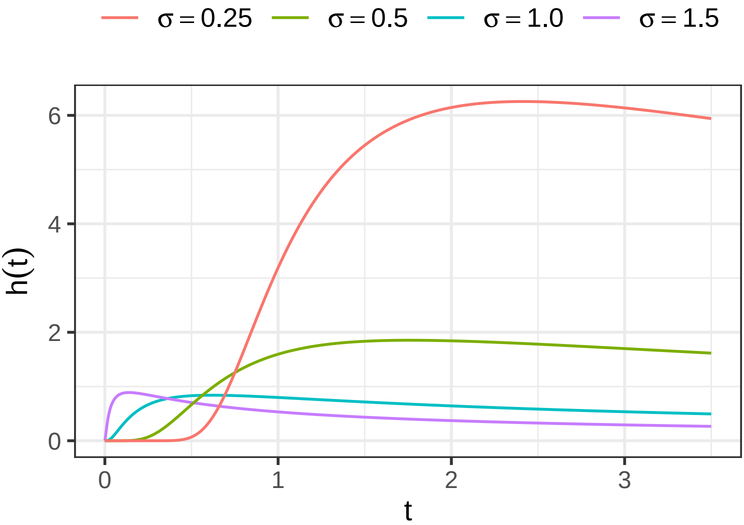


Figure 9: Hazard function of log-normal distribution with $\mu = 0$

The Log-logistic distribution

- If $Y = \log T$ follows a logistic distribution then T follows a log-logistic distribution
- The probability density function of a logistic distribution is a unimodal, bell-shaped curve symmetric around its mean, and for location parameter u and scale parameter b it is given by

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

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

The Log-logistic distribution

- The survivor function of a logistic distribution is

$$
\begin{aligned}
S(y) = \int_y^\infty f(x)\,dx &= \int_y^\infty \frac{e^{(x-u)/b}}{b\left[1+e^{(x-u)/b}\right]^2}\,dx. \\
&= \cdots \\
&= \frac{1}{\left[1+e^{(y-u)/b}\right]}
\end{aligned}
$$

- The hazard function of logistic distribution is

$$
h(y) = \frac{f(y)}{S(y)} = \frac{e^{(y-u)/b}}{b\left[1+e^{(y-u)/b}\right]}.
$$

The Log-logistic distribution

- The pdf of log-logistic distribution

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

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

The Log-logistic distribution

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

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

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

$$\begin{aligned} S(t) &= \int_{(t/\alpha)^{\beta}}^{\infty} \frac{1}{(1+y)^2} \, dy \\ &= \left. \frac{-1}{1+y} \right|_{(t/\alpha)^{\beta}}^{\infty} \\ &= \left[1 + (t/\alpha)^{\beta}\right]^{-1} \end{aligned}$$

The Log-logistic distribution

- The pdf

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

- The survivor function

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

- The hazard function

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

The Log-logistic distribution

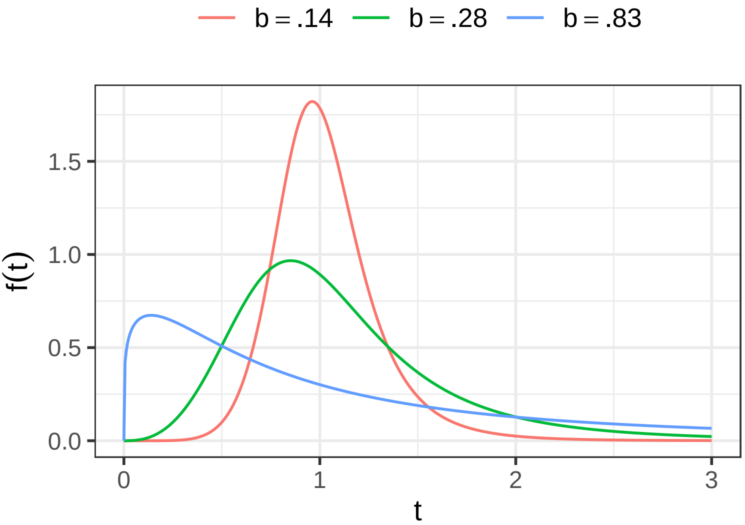


Figure 10: The density function of log-logistic distribution with $u = 0$

The Log-logistic distribution

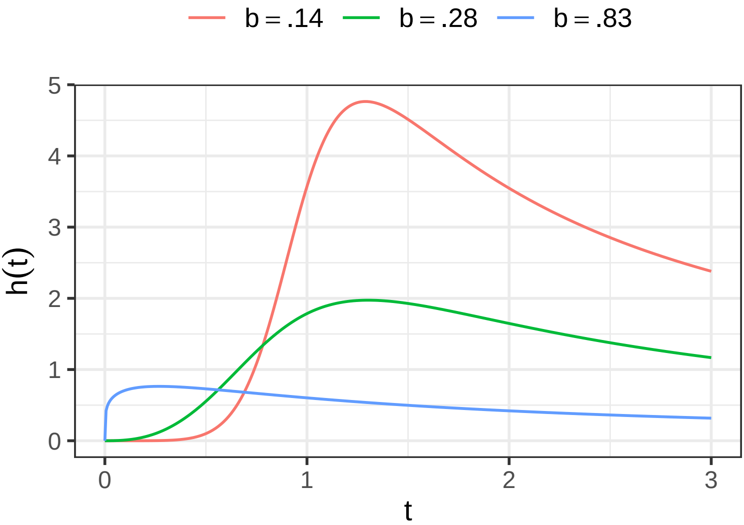


Figure 11: The hazard function of log-logistic distribution with $u = 0$

The Log-logistic distribution

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

$$E(T^r) = \int_0^\infty t^r \, \frac{\beta}{\alpha} \, \frac{(t/\alpha)^{\beta-1}}{\left(1 + (t/\alpha)^\beta\right)^2} \, dt.$$
$$= \cdots$$
$$= \alpha^r \, \Gamma\left(\frac{r}{\beta} + 1\right) \Gamma\left(1 - \frac{r}{\beta}\right)$$

You might have to recognize and use a Beta integral.

- Beta (first kind): for $0 < x < 1$,

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx = \mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

- Beta function (second kind / Beta prime):

$$\int_0^\infty \frac{x^{a-1}}{(1+x)^{a+b}} \, dx = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}, \quad a > 0, \; b > 0.$$

The Log-logistic distribution

Log-logistic hazard shape

- For $\beta > 1$: $h(0) = 0$, increases to a maximum, then decreases to 0 as $t \to \infty$ (unimodal hazard).
- For $\beta = 1$: $h(t) = 1/(\alpha + t)$, strictly decreasing.
- For $0 < \beta < 1$: $h(t)$ is monotone decreasing (starts at ∞).

The Gamma Distribution

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

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

  - shape $k > 0$
  - rate $\lambda > 0$ (scale $1/\lambda$)
- For $k = 1$, gamma distribution reduces to exponential distribution

The Gamma Distribution

- *Regularized lower 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 $y = \lambda x$

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

The Gamma Distribution

Gamma hazard shape

The hazard function

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

- For $k > 1$: $h(0) = 0$, $h(t)$ increases, $\lim_{t\to\infty} h(t) = \lambda$.
- For $k = 1$: $h(t) \equiv \lambda$.
- For $0 < k < 1$: $h(t)$ decreases, $\lim_{t\to 0} h(t) = \infty$, $\lim_{t\to\infty} h(t) = \lambda$.

In short: Gamma has increasing hazard rate for $k > 1$, constant for $k = 1$, and decreasing hazard rate for $0 < k < 1$

The Gamma Distribution

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

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

- 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 $T_1, \dots, T_n$ are iid 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 $Y \sim \text{Ga}(k)$ (rate 1) is

$$M_Y(\theta) = (1-\theta)^{-k}, \quad \theta < 1.$$

The Gamma Distribution

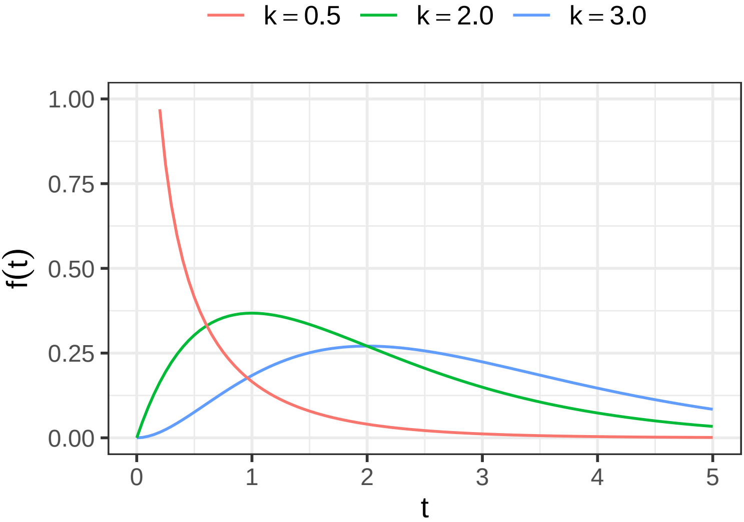


Figure 12: Density function of standard gamma distribution

The Gamma Distribution

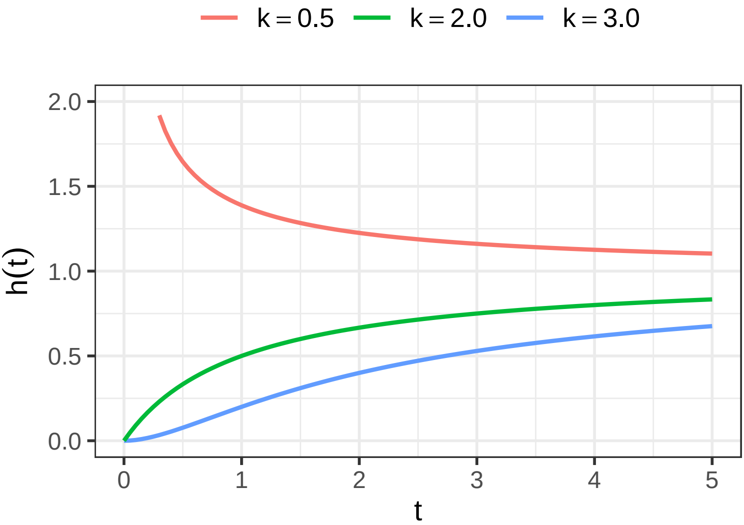


Figure 13: Hazard function of standard gamma distribution

Log-Location-scale Models

Definition: Location Scale Family A location-scale family is a family of distributions formed by *translation* and *rescaling* of a standard family member.

- A parametric location-scale model for a random variable Y is a distribution with pdf of the form

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

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

Log-Location-scale Models

- The cumulative distribution 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

scale (α) and shape (β)

- Parameters of *log-lifetime* distributions

location $(u = \log \alpha)$ 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

1.4 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
$$\theta(\mathbf{x}) = \exp\left(\beta' \mathbf{x}\right), \qquad \text{where } \beta = (\beta_1, \ldots, \beta_p)'$$
- 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
$$S(t \mid \mathbf{x}) = \exp(-t/\theta(\mathbf{x}))$$

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 model is Cox's proportional hazards model (Cox 1972)
- Cox's model considers 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 Dr. 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.