USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/∼das01/StatsShortCourse/

17th March 2004

Module 4 : 17th March

Survival Analysis

- Discrete & Continuous Time Models
- Non-parametric analysis Kaplan-Meier plots
- Proportional Hazards
- Accelerated Life
- Parametric Modelling
- Multivariate Survival & Competing Risks
- Multi-State Models

SECTION 1.

SURVIVAL ANALYSIS

In many medical studies an outcome of interest is the time to an event

- death from identified cause
- recurrence of a tumour
- conception during fertility treatment
- or discharge from hospital

The distinguishing feature of such data is that at the end of the follow up period the event may not have be observed, and thus the survival time is **censored**. In this circumstance, we do not know when (or, indeed, whether) the patient will experience the event, only that he or she has not done so by the end of the observation period.

Thus censoring may occur as

- truncation of study period
- loss to follow up from non-specific cause
- loss to a "competing" event unrelated to the condition being studied (eg: patients being studied after a heart transplant may die from some other disease or in an accident)

Typically, patients are recruited/introduced into the study over a period and followed up to a fixed date beyond the end of recruitment, and the last patients recruited will thus be studied for a shorter period than those recruited first and will be less likely to experience the event. Also, it is assumed that patients lost to follow up have the same prognosis as those remaining in the study.

Mathematical Notation:

Survival (or lifetime, or time-to-event) analysis is a special type of regression modelling that explains the observed variability in a **response** variable Y via consideration of **predictors** $X = (X_1, ..., X_K)$. The principal difference between survival analysis and conventional regression is that account is taken of potential **censoring** in the response variable

- we may observe some actual responses (survival, failure) times,
- censored responses where we do not observe an actual failure but rather only that the failure occurs after a **censoring time** (the end of study) – this is called **right-censoring**
- occasionally, we observe **left-censoring** or **interval-censoring**

- the response data is thus bivariate (Y, Z) where Y is the time at which the response is measured, and

$$Z = \begin{cases} 1 & \text{Failure is observed} \\ 0 & \text{Censored} \end{cases}$$

- The potential presence of censoring fundamentally changes how we view the modelling process - previously we have looked at probability densities and expected responses etc.
- we have previously only dealt with data y for which we need to specify $P[Y = y]$; we now need to think about
  - $P[Y > y]$ for right censoring
  - $P[Y \leq y]$ for left censoring
- We now take an alternative view, and examine **survivor** and **hazard** functions.

1.1 SURVIVAL IN DISCRETE TIME

The probability **mass function** for response variable Y is f_Y,

$$f_Y(y) = P[Y = y] \qquad y = 0, 1, 2, ...$$

The **distribution function** F_Y is

$$\begin{aligned} F_Y(y) &= P[Y \leq y] = \sum_{t=0}^{y} f_Y(t) \\ &= P[Y = 0] + P[Y = 1] + ... + P[Y = y] \end{aligned}$$

that is a **cumulative probability function**. Note that the function $F_Y(y)$ is a **non-decreasing** function.

As noted above, the major difference between conventional statistical modelling and survival/reliability modelling is the presence of **censoring**. In conventional regression modelling, the probability contribution for data point i with response y_i is $f_Y(y_i)$.

For right-censored data with censoring at y_i, however, we only observe the event

$$Y > y_i$$

that is, death/failure has **not occurred** before y_i time units. This event has probability

$$P[Y > y_i] = 1 - F_Y(y_i)$$

This motivates consideration of the **survivor (reliability) function**

$$S_Y(y) = 1 - F_Y(y)$$

Note that $S_Y(y)$ is a **non-increasing** function.

The likelihood function (via which inference and testing will be done) is thus

$$\left\{ \prod_{i:Z_i=1} f_Y(y_i) \right\} \times \left\{ \prod_{i:Z_i=0} S_Y(y_i) \right\}$$

that is

LIKELIHOOD FOR UNCENSORED DATA

$\times$

LIKELIHOOD FOR CENSORED DATA

and the role of the predictors can be introduced via the parameters of f_Y and F_Y.

Let

$$f_Y(y) = P[Y = y] \qquad y = 0, 1, 2, ...$$

define a discrete failure distribution. Then

$$S_Y(y) = P[Y > y] = 1 - F_Y(y) = \sum_{j=y+1}^{\infty} f_Y(j)$$

For example, we might have (for some probability π)

$$f_Y(y) = (1-\pi)^y \pi \qquad y = 0, 1, 2, ...$$

and

$$S_Y(y) = (1-\pi)^{y+1} \qquad y = 0, 1, 2, ...$$

1.2 THE DISCRETE HAZARD FUNCTION

As an alternative method of specification, we consider thediscrete **hazard function**

$$h_Y(y) = P\left[\text{Failure at } y | \text{Survival } \geq y\right] = \frac{f_Y(y)}{S_Y(y)}$$

and the **integrated hazard**

$$H_Y(y) = \sum_{t=0}^{y} h_Y(t)$$

and it can be shown that

$$S_Y(y) = \exp\left\{-H_Y(y)\right\}$$

so that

$$f_Y(y) = \left\{ \prod_{j=0}^{y-1} (1 - h_Y(j)) \right\} \times h_Y(y)$$

and

$$S_Y(y) = \prod_{j=0}^{y} (1 - h_Y(j))$$

If

$$f_Y(y) = (1-\pi)^y \pi \qquad y = 0, 1, 2, ...$$

then

$$h_Y(y) = \frac{(1-\pi)^y \pi}{(1-\pi)^{y+1}} = \frac{\pi}{1-\pi}$$

that is, a constant, independent of y.

1.3 THE CONTINUOUS TIME MODEL

The probability **density function** for continuous response variable Y is f_Y, and the expectation, likelihood function and so on that are required for regression modelling are formed from f_Y. The **distribution function** F_Y is

$$F_Y(y) = P[Y \leq y] = \int_0^y f_Y(t)\, dt$$

In conventional regression modelling, the likelihood contribution for data point i with response y_i is $f_Y(y_i)$. For right-censored data with censoring at y_i, we have again the survivor function

$$S_Y(y) = 1 - F_Y(y)$$

1.4 CONTINUOUS HAZARDS

As a further alternative method of specification, we consider the **continuous hazard function**

$$\begin{aligned} h_Y(y) &= P\left[\text{Failure at } y | \text{Survival } \geq y\right] \\ &= \frac{f_Y(y)}{S_Y(y)} \end{aligned}$$

and the **integrated hazard**

$$H_Y(y) = \int_0^y h_Y(t)\, dt$$

and it can be shown that

$$S_Y(y) = \exp\left\{-H_Y(y)\right\}$$

1.5 THE KAPLAN-MEIER CURVE

The **Kaplan-Meier curve** (or **product-limit estimate)** is a non-parametric estimate of the **survivor** function; it takes into account the censored data and produces a decreasing "step-function" curve, where the downward steps take place at the times of the failures, giving the estimated survival function at the jth failure/censoring time as

$$\widehat{S}_j = \prod_{i=1}^{j} \left(1 - \frac{z_i}{n - i + 1}\right) \tag{1}$$

This curve can be used to report an estimated survival probability at a given time (1 year, 5 years etc.).

Standard errors for these estimated survival probabilities are also available.

Construction: Let

- sample size n comprise observed and censored failure times
- $0 < y_{(1)} < y_{(2)} < ... < y_{(m)}$, be the distinct failure times, sorted into ascending order
- d_j be the number of number of failures observed at time $y_{(j)}$
  - usually $d_j = 1$
  - certainly $d_j \geq 1$ ($d_j > 1$ implies tied failure times)
- n_j be the number of patients "at risk" of failure at time $t_{(j)}$, that is, the number of patients who have failure/censoring time greater than or equal to $t_{(j)}$.

Then the observed probability of surviving beyond $t_{(j)}$ (conditional on having survived that long) is

$$\widehat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \widehat{q}_j$$

say, where $q_j = d_j/n_j$ is the estimated conditional probability of failure at time $t_{(j)}$. Using the chain rule for probabilities, the estimated probability of surviving at least until time t is

$$\widehat{P}(t) = \prod_{j=1}^{J_t} \widehat{p}_j = \prod_{j=1}^{J_t} \left(1 - \frac{d_j}{n_j}\right) \tag{2}$$

where $J_t = \max\left\{j : t_{(j)} \leq t\right\}$. $\widehat{P}$ is identical to the S_j function from formula (1),and thus we have $\widehat{S}_{KM}(t) = \widehat{P}(t)$.

STANDARD ERRORS: The estimate in (2) should be reported with an associated standard errror. A number of possibilities have been suggested. Let $P_j = P\left(t_{(j)}\right)$. Then

- **Greenwood's Formula**

$$s.e.\left(\widehat{P}_j\right) = \widehat{P}_j \sqrt{\sum_{i=0}^{j-1} \frac{d_i}{n_i - d_i}}$$

- **Peto's Formula**

$$s.e.\left(\widehat{P}_j\right) = \widehat{P}_j \sqrt{\frac{1 - \widehat{P}_j}{n'_j}}$$

where n'_j is an "adjusted" or "effective" sample size, the number of survivors at the beginning of the interval $\left(t_j, t_{(j+1)}\right)$.

1.6 THE NELSON-AALEN CURVE

The **Nelson-Aalen estimate** is a non-parametric estimate of the **cumulative hazard** function; it takes the form

$$\widehat{H}(t) = \sum_{j=1}^{J_t} \left(\frac{d_j}{n_j} \right) \tag{3}$$

where $J_t = \max\left\{ j : t_{(j)} \leq t \right\}$. From this, we can construct another estimate of the survivor function

$$\widehat{S}_{FH}(t) = \exp\left\{ -\widehat{H}(t) \right\}$$

this is the **Fleming-Harrington estimate** of the survivor function.

STANDARD ERRORS: If $\widehat{H}_j = \widehat{H}(t_{(j)})$, can use

- **Greenwood**

$$s.e.\left(\widehat{H}_j\right) = \sqrt{\sum_{i=0}^{j} \frac{d_i}{n_i\left(n_i - d_i\right)}}$$

- **Tsiatis**

$$s.e.\left(\widehat{H}_j\right) = \sqrt{\sum_{i=0}^{j} \frac{d_i}{n_i^2}}$$

- **Klein**

$$s.e.\left(\widehat{H}_j\right) = \sqrt{\sum_{i=0}^{j} \frac{d_i\left(n_i - d_i\right)}{n_i^3}}$$

1.7 THE COX REGRESSION MODEL

The **Cox** (or **Proportional Hazards**) model provides a simple way of introducing the influence of predictors into the survival model. The basic components are a **baseline hazard** function, h_0 and a linear predictor and (positive) link function g (similar to the GLM modelling of previous chapters). Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the hazard function takes the form

$$h_Y(y; x) = g(x^T \beta) h_0(y)$$

that is, the hazard is modified in a multiplicative fashion by the linked-linear predictor.

Typically, g is the exponential function.

From the previously established relationships,

$$S_Y(y;x) = \exp\left\{-\int_0^y h_Y(t)\ dt\right\} = \exp\left\{-\int_0^y g(x^T\beta)h_0(y)\ dt\right\}$$

If a coefficient β_k is positive, the hazard is **increased**, and the expected failure time **decreased.**

The relevance/significance of a particular predictor is assessed using a **Wald** test based on the magnitude of

$$t = \frac{\widehat{\beta}}{s.e.\left(\widehat{\beta}\right)}$$

If $|t| > 2$, then the hypothesis that $\beta = 0$ can be rejected.

1.8 THE ACCELERATED LIFE MODEL

The **Accelerated Life** model provides another way of introducing the influence of predictors into the survival model. The basic components now are a **baseline survivor** function, S_0 and a linear predictor and (positive) link function g as above. Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the survivor function takes the form

$$S_Y(y; x) = S_0(g(x^T\beta)y)$$

that is, the time scale is modified in a multiplicative fashion by the linked-linear predictor.

Again, typically, g is the exponential function. This model allows direct modelling of the influence of predictors on survival.

1.9 FRAILTY MODELLING

The idea of frailty modelling is to introduce **random effects** terms into the linear predictor that appears in the proportional hazards and accelerated life models. For example, we extend

$$x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{iD} + ... + \beta_D x_{iD}$$

(which is a deterministic function of the parameter $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_D)$ and predictors $x = (x_1, x_2, ..., x_D)$) to include a random component that is specific to the individual patient concerned, that is, we have

$$x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_D x_{iD} + L_i$$

where L_i is some (usually zero mean) random variable.

1.10 THE LOG-RANK TEST

The **log-rank** test is a standard test for significant differences between two (or more) survivor functions that differ because of the influence of the different levels of a discrete predictor.

$$\begin{aligned} H_0 &: S_1 = S_2 \\ H_1 &: S_1 \neq S_2 \end{aligned}$$

It is a non-parametric test based on ranks of samples for the two or more subgroups.

Asymptotic or exact versions of the test can be carried out; SPSS and other packages give further alternatives.

1.11 PARAMETRIC MODELLING

It is possible to fit and compare **parametric** survival models to such data. Parametric densities, survivor functions, hazards etc. can be readily used in the formation of a likelihood, potentially within the proportional hazards/accelerated life framework.

Typical models used are

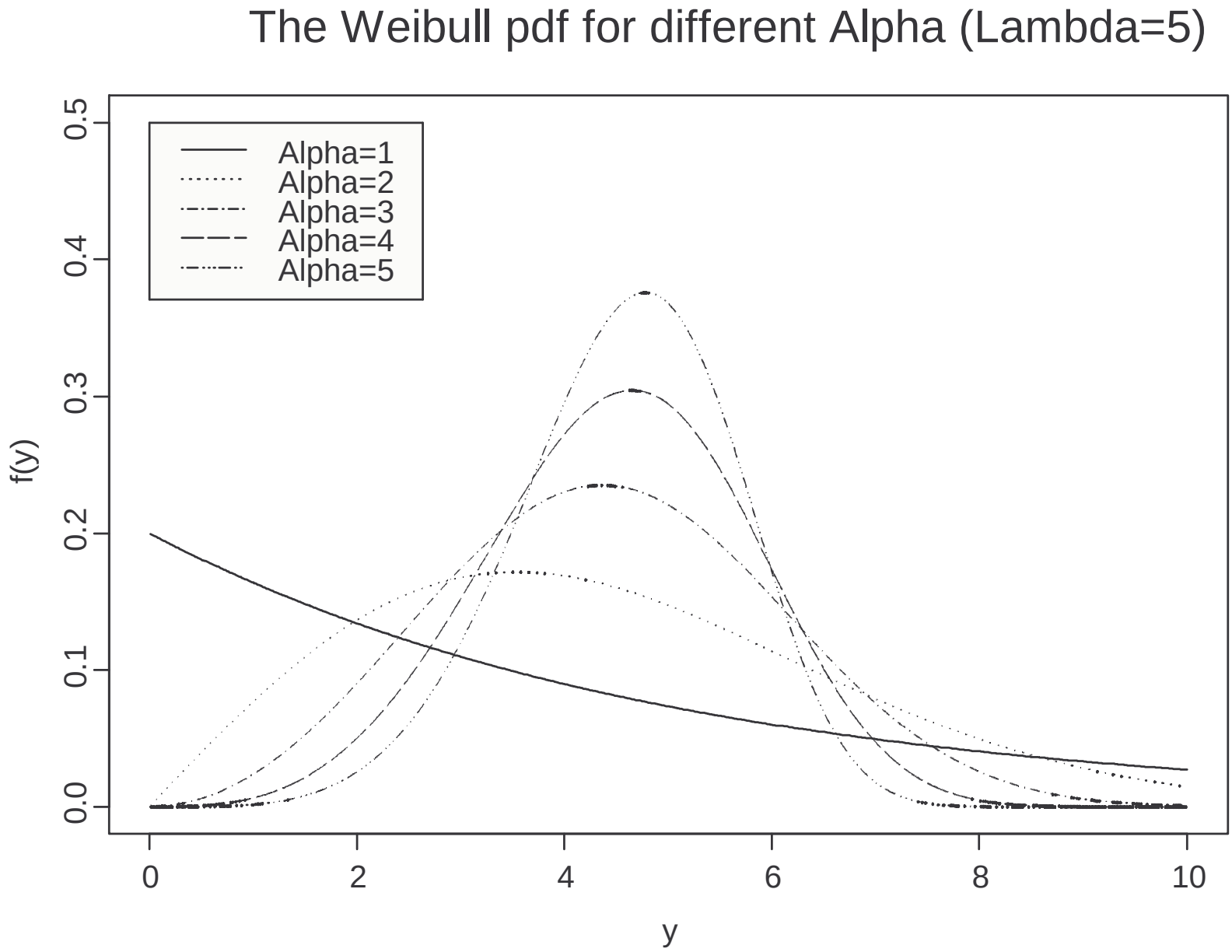
- Weibull
- Gamma
- Log-Logistic
- Log-Normal
- Pareto

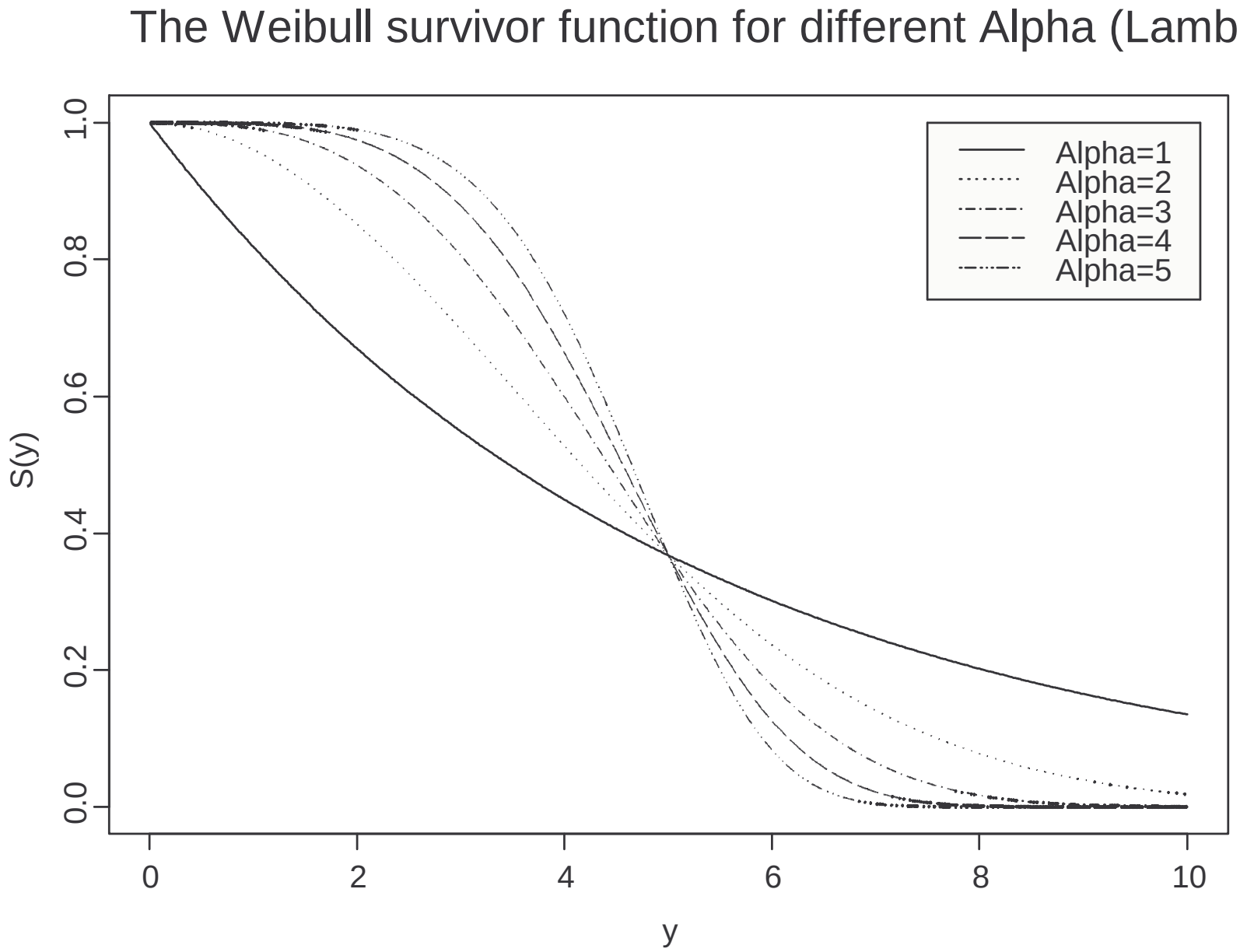
1.11.1 WEIBULL MODEL

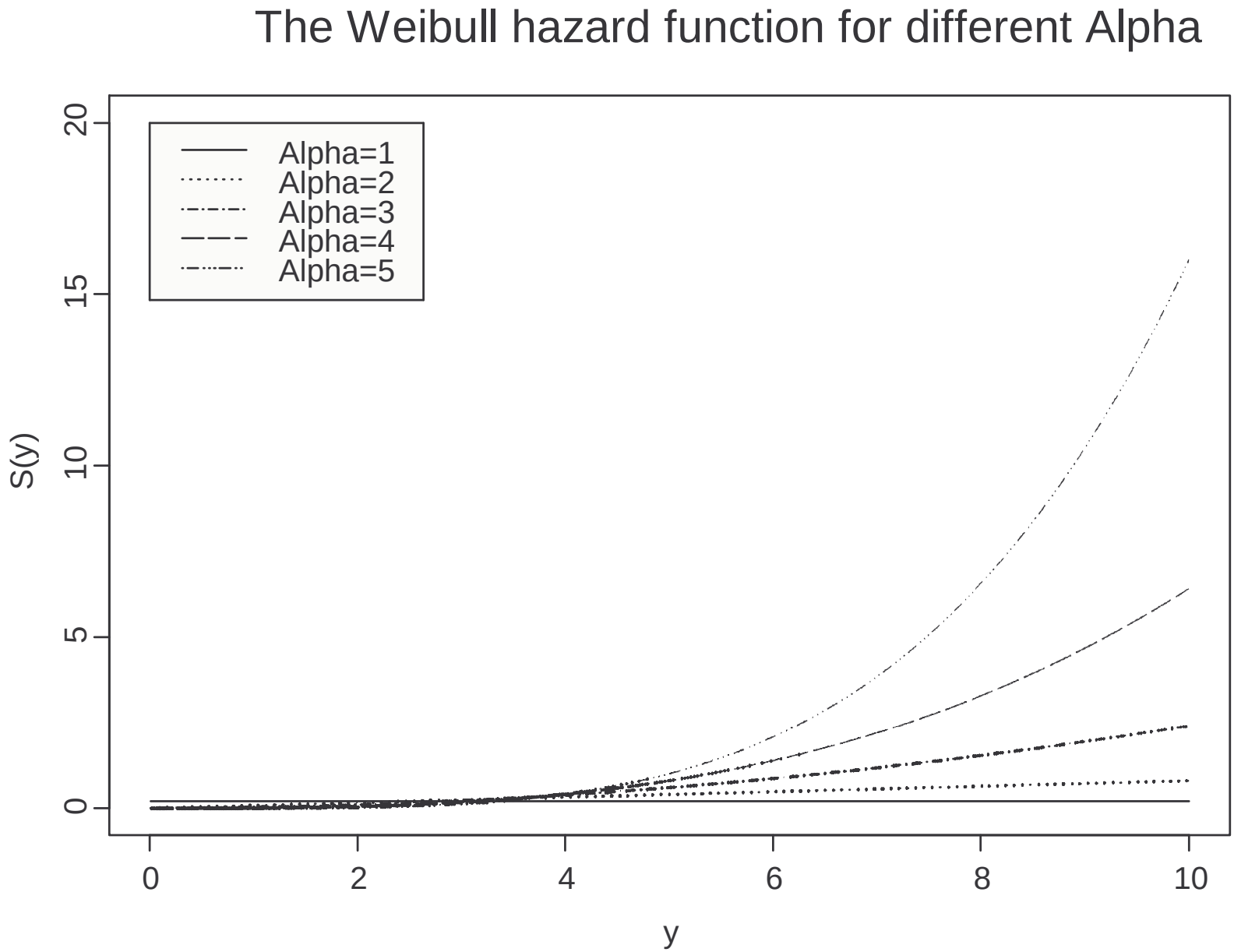
The Weibull distribution is a two-parameter probability model that is the most commonly used in reliability modelling. For $y > 0$,

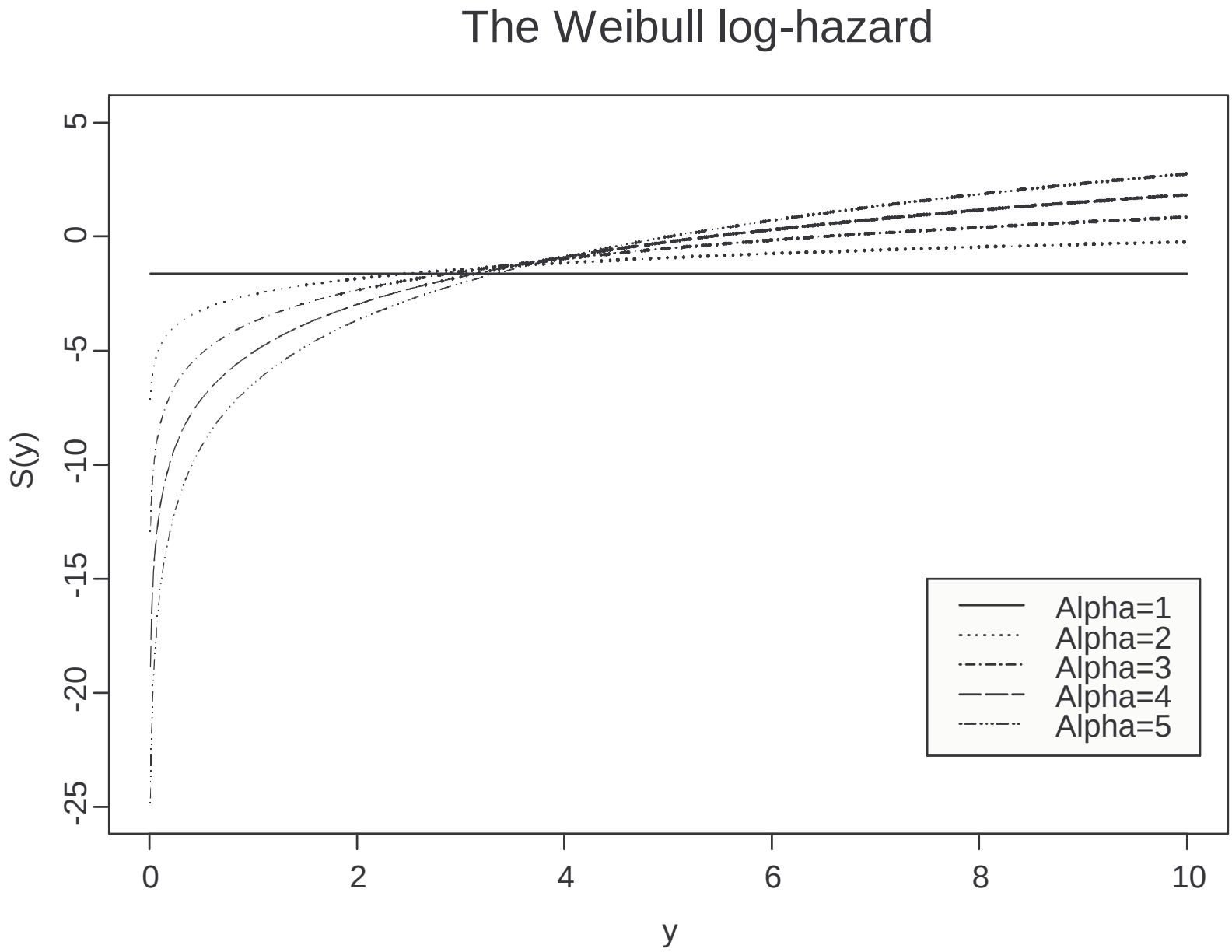
$$\begin{aligned} f(y) &= \frac{\alpha}{\lambda^\alpha} y^{\alpha-1} \exp\left\{-\left(\frac{y}{\lambda}\right)^\alpha\right\} \\ F(y) &= 1 - \exp\left\{-\left(\frac{y}{\lambda}\right)^\alpha\right\} \qquad \Longrightarrow \qquad S(y) = \exp\left\{-\left(\frac{y}{\lambda}\right)^\alpha\right\} \\ h(y) &= \frac{\alpha}{\lambda^\alpha} y^{\alpha-1} \\ H(y) &= \left(\frac{y}{\lambda}\right)^\alpha \end{aligned}$$

for parameters $\alpha, \lambda > 0$ (the *shape* and *scale* parameters respectively).


The Weibull pdf for different Alpha (Lambda=5)
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
f(y)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
y


The Weibull survivor function for different Alpha (Lambda=
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
S(y)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
y


The Weibull hazard function for different Alpha
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
S(y)
0
5
10
15
20
0
2
4
6
8
10
y


The Weibull log-hazard
S(y)
5
0
-5
-10
-15
-20
-25
0
2
4
6
8
10
y
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5

1.11.2 GAMMA MODEL

The Gamma distribution is a two-parameter probability model. The functions interest for the Weibull distribution are, for $y > 0$,

$$f(y) = \frac{1}{\lambda^{\alpha}\Gamma(\alpha)} y^{\alpha-1} \exp\left\{-\left(\frac{y}{\lambda}\right)\right\}$$

for parameters $\alpha, \lambda > 0$ (the *shape* and *scale* parameters respectively), and $\Gamma(\alpha)$ is the **Gamma Function**, a special function defined by

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$$

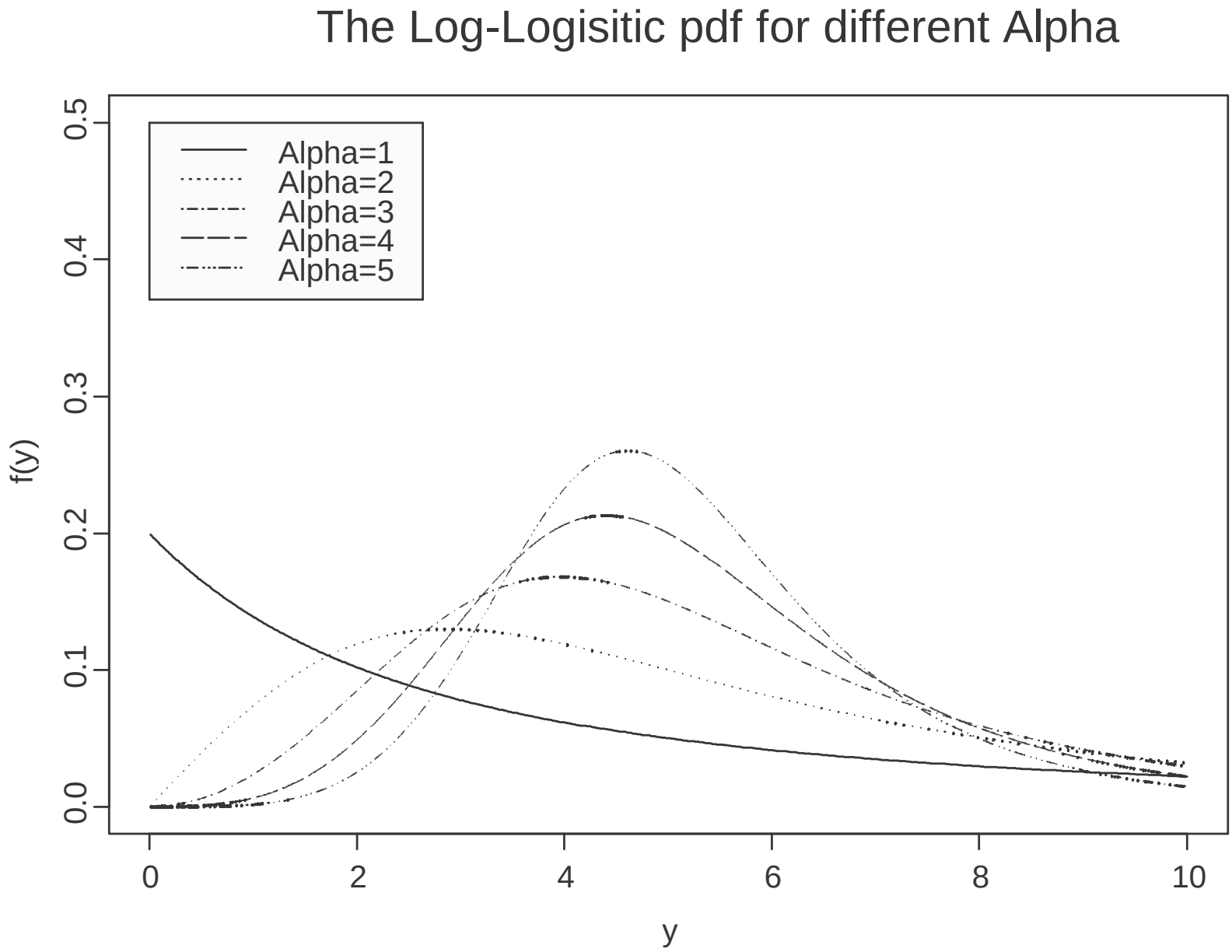
None of the other functions are available in a straightforward way, by can be computed numerically.

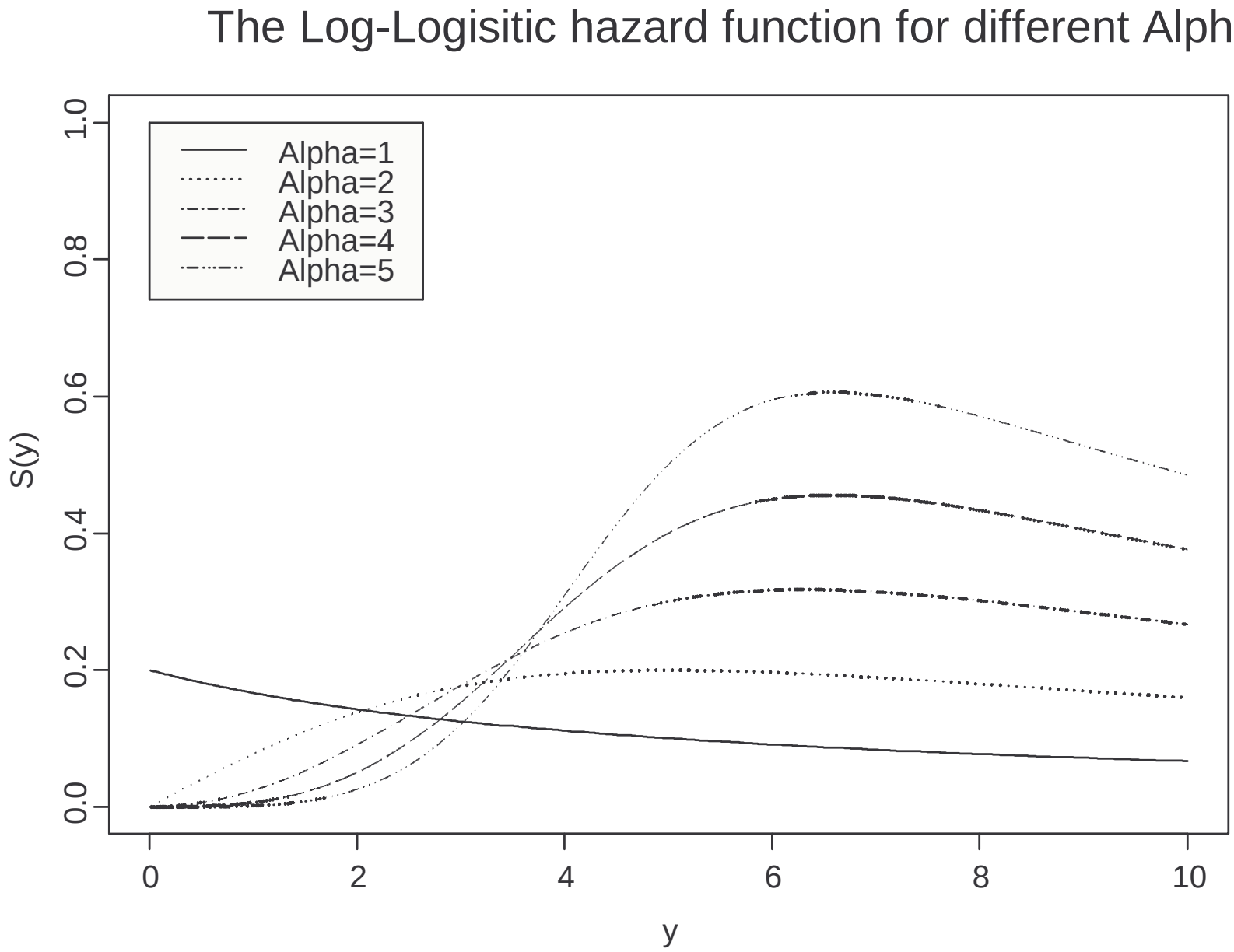
1.11.3 LOG-LOGISTIC

The **log-logistic** distribution is a two-parameter probability model for which the functions interest for are, for $y > 0$,

$$\begin{aligned} f(y) &= \frac{\alpha\lambda^{\alpha}y^{\alpha-1}}{(\lambda^{\alpha}+y^{\alpha})^2} \\ F(y) &= \frac{y^{\alpha}}{\lambda^{\alpha}+y^{\alpha}} \qquad \Longrightarrow \qquad S(y) = \frac{\lambda^{\alpha}}{\lambda^{\alpha}+y^{\alpha}} \\ h(y) &= \frac{\alpha y^{\alpha-1}}{(\lambda^{\alpha}+y^{\alpha})} \\ H(y) &= \log(\lambda^{\alpha}+y^{\alpha}) \end{aligned}$$

for parameters $\alpha, \lambda > 0$.


The Log-Logisitic pdf for different Alpha
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
f(y)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
y


The Log-Logisitic hazard function for different Alpha
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
S(y)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
y

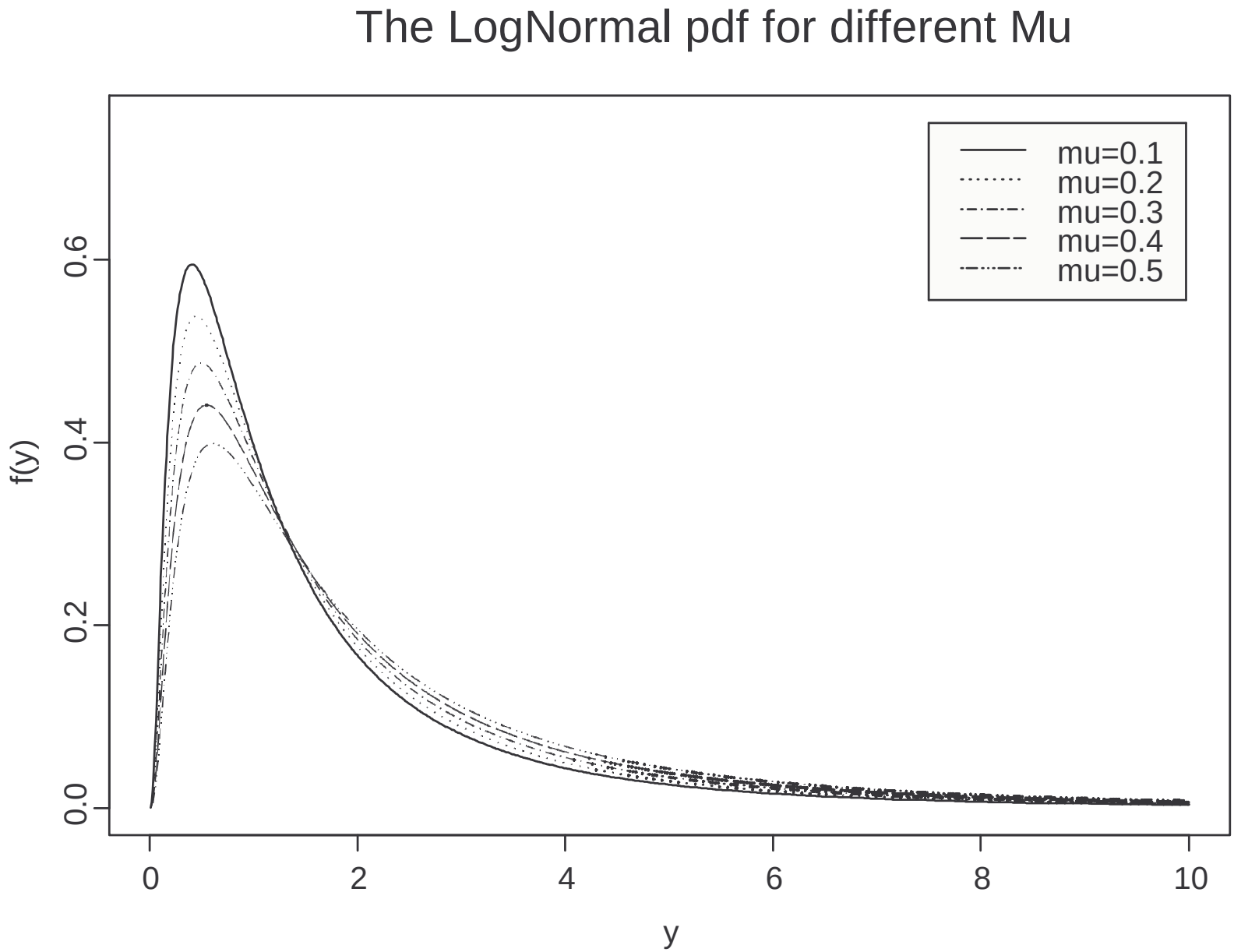
1.11.4 LOGNORMAL MODEL

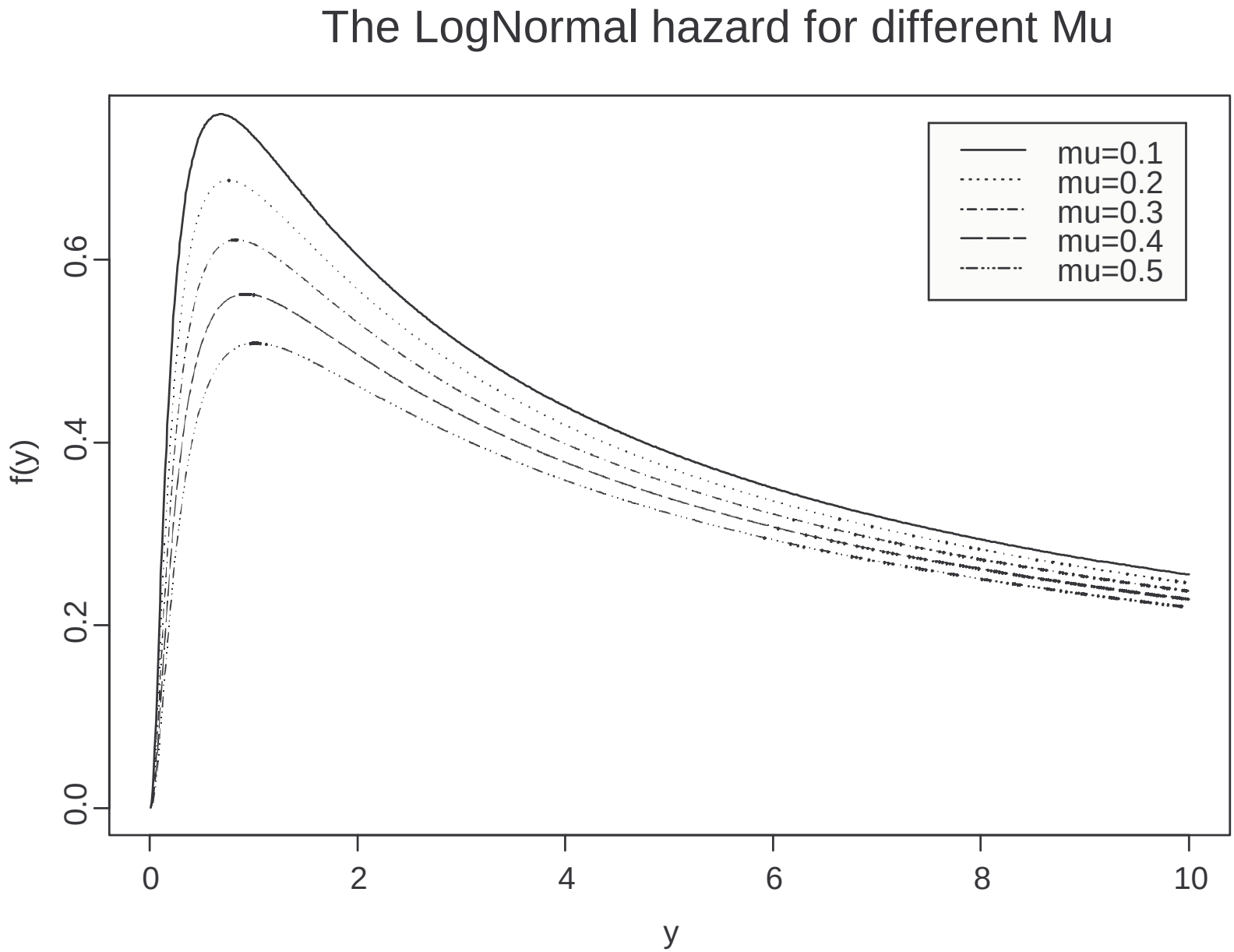
The lognormal distribution is a two-parameter probability model for which the functions interest for are, for $y > 0$,

$$\begin{aligned} f(y) &= \left(\frac{1}{2\pi y^2\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(\log y - \mu)^2\right\} \\ F(y) &= \Phi\left(\frac{(\log y - \mu)}{\sigma}\right) \qquad \Longrightarrow \qquad S(y) = 1 - \Phi\left(\frac{(\log y - \mu)}{\sigma}\right) \\ H(y) &= -\log\left(1 - \Phi\left(\frac{(\log y - \mu)}{\sigma}\right)\right) \end{aligned}$$

for parameters μ, and $\sigma > 0$, where Φ is the standard normal distribution function. This model presumes that the log survival time is normally distributed

$$\log Y \sim Normal\left(\mu, \sigma^2\right)$$


The LogNormal pdf for different Mu
mu=0.1
mu=0.2
mu=0.3
mu=0.4
mu=0.5
f(y)
0.0
0.2
0.4
0.6
0
2
4
6
8
10
y

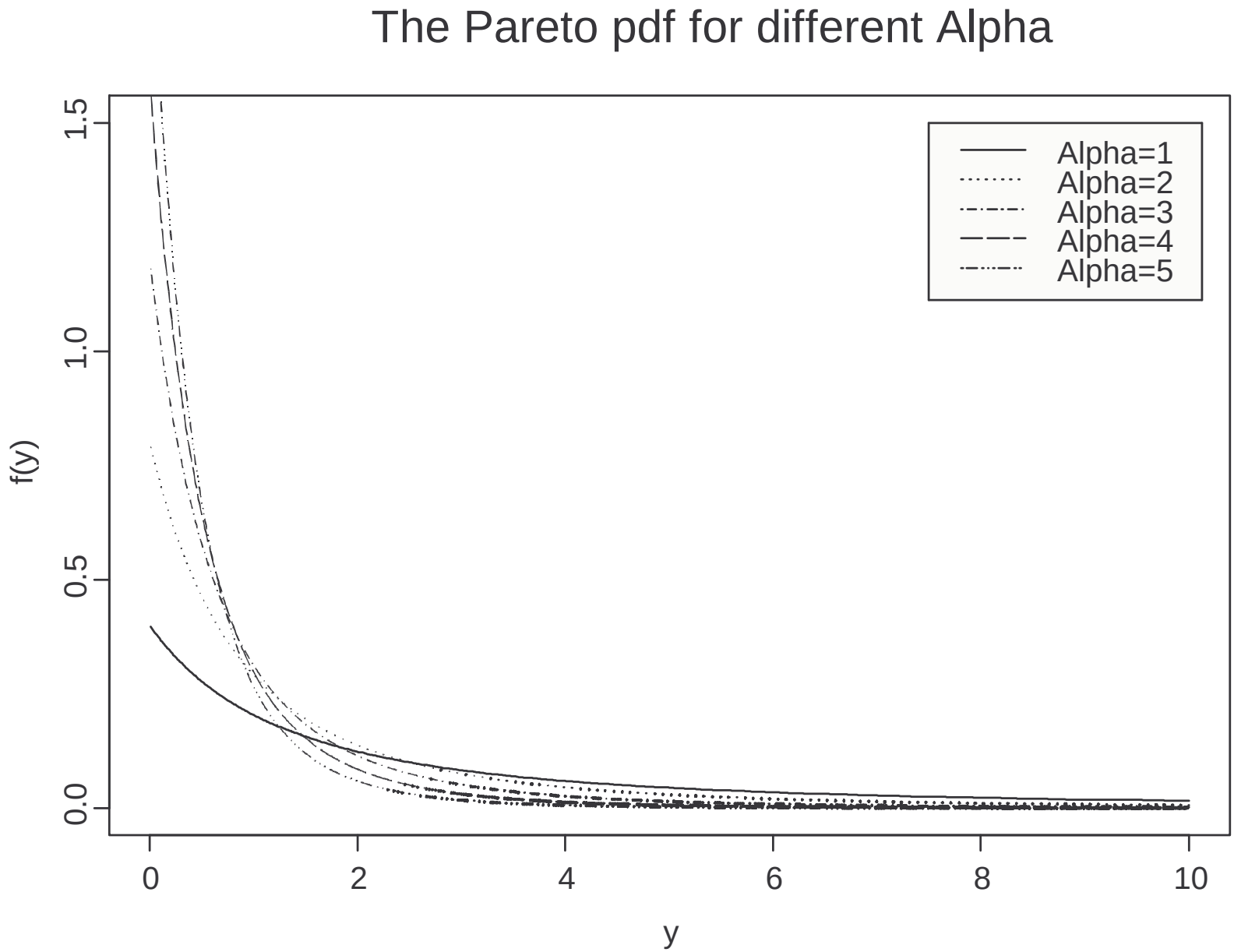

The LogNormal hazard for different Mu
mu=0.1
mu=0.2
mu=0.3
mu=0.4
mu=0.5
f(y)
0.0
0.2
0.4
0.6
0
2
4
6
8
10
y

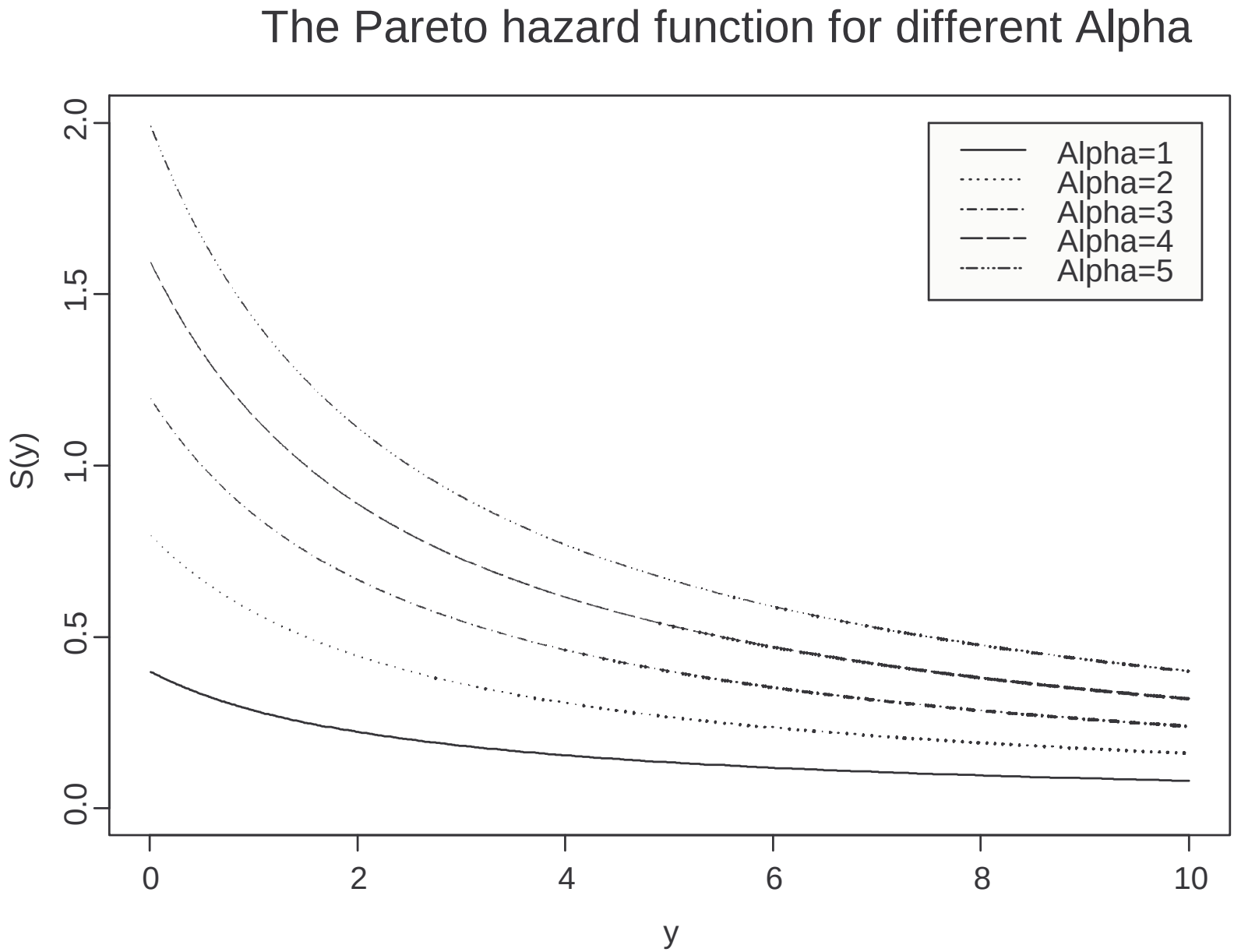
1.11.5 PARETO MODEL

The **Pareto** distribution is a two-parameter probability model for which the functions interest for are, for $y > 0$,

$$\begin{aligned} f(y) &= \frac{\alpha\theta^{\alpha}}{(\theta+y)^{\alpha+1}} \\ F(y) &= 1-\left(\frac{\theta}{\theta+y}\right)^{\alpha} \qquad \Longrightarrow \qquad S(y)=\left(\frac{\theta}{\theta+y}\right)^{\alpha} \\ h(y) &= \frac{\alpha}{(\theta+y)} \\ H(y) &= \alpha\log(\theta+y) \end{aligned}$$

for parameters $\alpha, \theta > 0$.


The Pareto pdf for different Alpha
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
f(y)
0.0
0.5
1.0
1.5
0
2
4
6
8
10
y


The Pareto hazard function for different Alpha
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
S(y)
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8
10
y

1.12 MULTIVARIATE ANALYSIS & COMPETING RISKS

An important generalization of conventional survival analysis extends survival time Y to be a **vector** quantity, that is, we have say K different aspects of failure, with random variable $Y = (Y_1, ...Y_K)$ requiring a joint probability model. Typically, such models are difficult to construct.

A more common experimental situation is one of **competing risks**; that is, there are K potential causes of failure, but **at most one** is observed for each individual in the study. Then the failure time, T, is defined by

$$T = \min \{Y_1, ..., Y_K\}$$

If the cause of failure, C, is recorded as $C = k$, we observe

$$Y_1 > t, ..., Y_{k-1} > t, Y_k = t, Y_{k+1} > t, ..., Y_K > t$$

whereas if the observation is censored, we observe

$$Y_1 > t, ..., Y_{k-1} > t, Y_k > t, Y_{k+1} > t, ..., Y_K > t$$

A joint model is again often difficult to construct, and in addition there are issues to do with identifiability of the "marginal" failure processes for the components of Y.

i.e. without sufficient data, there are problems in estimating the models for $Y_1, ..., Y_K$ considered on theor own

1.13 MULTI-STATE MODELLING

In **multi-state** modelling, rather than just having the standard failed/not failed (dead/alive) dichotomy, with

$$Z = \begin{cases} 0 & \text{Censored} \\ 1 & \text{Failure is observed} \end{cases}$$

we have an extension to polytomy, where

$$Z(t) = \begin{cases} 0 & \text{Censored at time } t \\ 1 & \text{State 1 at time } t \\ \vdots & \vdots \\ M & \text{State } M \text{ at time } t \end{cases}$$

This kind of modelling is very useful for modelling disease progression; the different states could correspond to different stages of the disease.

In such a model, we attempt to estimate the probability

$$\pi_{ij}(t_i, t_j) = P[\text{In state } j \text{ at time } t_j | \text{In state } i \text{ at time } t_i]$$

for $t_i < t_j$, or rate, λ_{ij}, of transition from one state to another.

In a discrete time framework, homogeneous **Markov Models** are typically used, characterized by a **transition matrix** P, with $(i, j)^{th}$ entry π_{ij}, independent of t, with

$$\sum_{j=0}^{M} \pi_{ij} = 1$$

A **multi-state** process is a random process $\{Z(t)\}_{t \geq 0}$ describing the state within which the individual lies at time t.