MATH96048/MATH97075/MATH97183
Survival Models

Professor Axel Gandy

2021/22

These notes are based on earlier versions of the course, incorporting notes of Professor Nicholas Heard and Dr David Whitney.

1 Principles of Modelling Time to Event Data

parameters / entire distribution unknown.

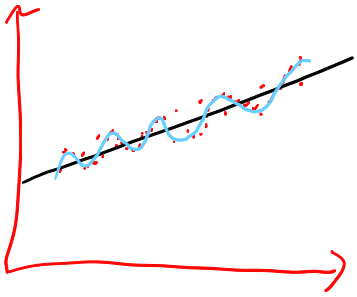
1.1 Statistical Modelling

A stochastic or statistical model of a system is a mathematical model which represents inherent uncertainties in the system as random variables. Any model which does not make such allowances for uncertainty, on the other hand, is said to be deterministic. In actuarial science, statistical models are especially useful for dealing with uncertainty in the time until a specific event will occur, such as predicting the number of premium payments that will be received before paying out on a life assurance contract. These statistical models for time to event data also have much wider applications in fields such as medicine, biostatistics and engineering.

Mathematical models are an imperfect representation of reality. Their utility comes from being able to approximately learn the consequences of hypothetically changing certain experimental inputs or actions. Statistical models extend this utility by capturing the uncertainty surrounding unknown future outcomes, providing the possibility of searching for an optimal decision under this uncertainty.

1.2 Model Choice

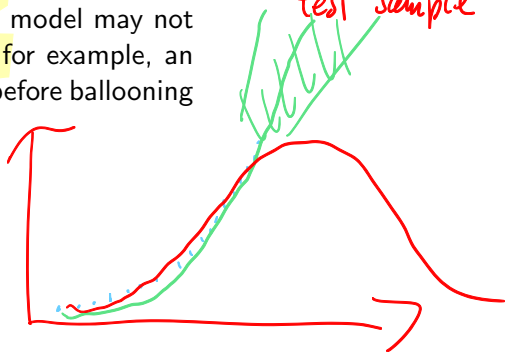
1.2.1 Complexity



The complexity of a good statistical model is often constrained for several reasons. The analyst might be restricted by the inputs for which data are readily available, or by computational or statistical limitations to the models which can be reliably fitted. Additionally, it is important that the chosen model satisfies the purposes for which it will be used; usually this means that the mathematical structure of the model need be easily interpretable for purposes of understanding and communication, and that the resulting fitted model will not overfit and understate uncertainty and risk. For these reasons, the statistical analyst will often seek to fit the most parsimonious model that the data will sensibly allow.

1.2.2 Sensitivity

residual plots
AIC, BIC, DIC
plot data
model predictions
does model fit make intuitive sense
training sample / test sample



Model choice and fitting should be viewed as an iterative procedure. Once a particular model has been fitted to the available data, the quality of the fit should be inspected; hypothesis tests such as goodness of fit tests can determine the suitability of the selected model. If the structure of the data is not well captured by the model, this may suggest that the model chosen is inadequate and the analyst should return to the model selection stage to consider other alternatives.

Note however that performance of the chosen model will also depend heavily on the quality of the data, with poor quality inputs likely to lead to unreliable output inference (*garbage in, garbage out*). If there are questions about the accuracy of the data being used, then it becomes particularly important to carry out a sensitivity analysis. This investigates the magnitude of change in the model outputs when small perturbations are applied to the model inputs.

Finally, it should be noted that the suitability of a well fitted model may not be comfortably relied upon outside the range of the data used; for example, an exponential relationship can appear fairly linear in the short term, before ballooning away from such a fit in the longer term.

2 Distributions of Event Times

2.1 Random Variable Modelling

Consider a homogeneous population of individuals, who each have an associated event time which is initially unknown and treated as a random variable. We will often refer to the period of time until the event occurs as the lifetime of the individual.

Throughout the course, let T denote the future lifetime of a new-born individual (aged 0). Unless otherwise stated, we shall assume T is a continuous random variable which takes values on the positive part of the real line $\mathbb{R}^+ = [0, \infty)$, with associated probability measure P.

More specifically, in this chapter we might assume the existence of a limiting age ω which T cannot exceed, so $T \in [0, \omega]$. For human life calculations, typically we take $\omega \approx 120$ years. $\omega = \infty$ possibility

We first define the cumulative distribution function, F, the survivor function S, the density f, the hazard h and the cumulative hazard H for a lifetime, and derive relationships between them.

This will provide us with a range of tools for specifying the distribution of a lifetime, and rules for moving between them.

2.2 Cumulative Distribution and Survivor Functions

Definition: The cumulative distribution function (CDF) of T is a function $F: \mathbb{R} \to [0,1]$,

$$F(t) = \mathrm{P}(T \le t),$$

the probability of death by age t.

(survival)

Definition: The survivor (or reliability) function of T is

$$S(t) = \mathrm{P}(T > t) \quad = \quad 1 - F(t),$$

the probability of surviving beyond age t.

Example: $T \sim$ Exponential (1)

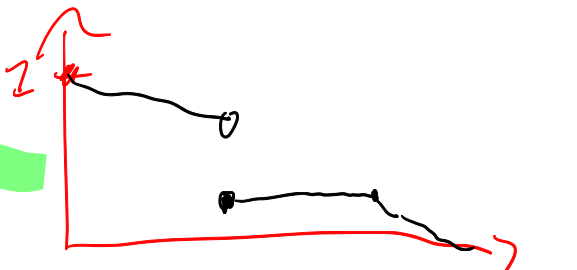
$F(t) = 1 - \exp(-t), \quad t \ge 0$

$S(t) = \exp(-t), \quad t \ge 0$

2.2.1 Criteria for a valid survivor function:

$S(t)$ is a valid survivor function iff

1. $S(0) = 1, \ \lim_{t \to \infty} S(t) = 0$
2. $0 \le S(t) \le 1, \forall t \in \mathbb{R}^+$;
3. Monotonicity: $\forall t_1, t_2 \in \mathbb{R}^+, t_1 < t_2 \implies S(t_1) \ge S(t_2)$;
4. S is càdlàg: $\forall t \in \mathbb{R}^+$,
   (a) $S(t^+) \equiv \lim_{u \downarrow t} S(u) = S(t)$ (S is right-continuous);
   (b) $S(t^-) \equiv \lim_{u \uparrow t} S(u)$ exists. (left limits exist for S).



can be constant for certain intervals

Most often: either continuous or only jumps

2.2.2 Future lifetime after age x

Definition: Let T_x be the future lifetime of an individual who has survived to age x, for $0 \le x \le \omega$, so $T_x \in [0, \omega - x]$. So clearly

- $T_0 \equiv T$;
- The distribution of T_x is the same as $T - x | T > x$, in other words: for all measurable sets A,

$$\mathrm{P}(T_x \in A) = \mathrm{P}(T - x \in A | T > x).$$

Example $T \sim$ Exponential (1)

$S_x(t) = P(T_x > t) = P(T - x > t \mid T > x)$

$= \frac{P(T - x > t, \ T > x)}{P(T > x)}$

$= \frac{P(T > x+t)}{P(T > x)} = \frac{\exp(-(x+t))}{\exp(-x)}$

$= \exp(-t)$ — independent of x

[memoryless property of exponential distribution]

Definition: The cumulative distribution and survivor functions of T_x are

$$\begin{aligned} F_x(t) &= \mathrm{P}(T_x \le t), \\ S_x(t) &= \mathrm{P}(T_x > t) = 1 - F_x(t). \end{aligned}$$

2.2.3 Relationship between T_x and T

For consistency with T, for the CDF we have

$$\begin{aligned} F_x(t) &= \mathrm{P}(T_x \le t) = \mathrm{P}(T \le x + t | T > x), \\ \Longrightarrow F_x(t) &= \frac{F(x+t) - F(x)}{S(x)} \end{aligned}$$

Def of cond distr

and for the survivor function,

$$S_x(t) = 1 - F_x(t) = \frac{S(x+t)}{S(x)}.$$

2.3 Density Function

Def: Let X be a r.v. and $F: \mathbb{R} \to \mathbb{R}$, $F(t) = P(X \le t)$. A function $f: \mathbb{R} \to [0, \infty)$ s.t. $\int_{-\infty}^{t} f(s)\,ds = F(t)$ $\forall t$ is called the pdf of X. [general case]

Definition: The probability density function (PDF) of the random variable T_x is

$$f_x(t) = \frac{d}{dt} F_x(t) = \lim_{h \downarrow 0} \frac{F_x(t+h) - F_x(t)}{h}.$$

Since we are assuming T_x is a continuous random variable, by definition this density exists.

For individuals who have currently survived to age x, $f_x(t)$ is the rate of death t further units of time into the future. That is, for such an individual, the probability of death within the interval $[t, t+h]$ for a small interval width h is approximately $f_x(t)h$.

2.4 Hazard Function / Force of Mortality

The hazard function plays a central role in survival analysis. We denote the hazard function (or force of mortality) at age x, $0 \le x \le \omega$, by $h(x)$.

Definition: The hazard function of T is defined as

$h: [0, \infty) \to [0, \infty)$, $h(x) = \lim_{h \downarrow 0} \frac{\mathrm{P}(T \le x + h | T > x)}{h}$.

(or hazard rate)

We will always assume this limit exists.

Example: $T \sim$ Exponential (1)

$h(x) = \lim_{h \downarrow 0} \frac{P(T \le x+h \mid T > x)}{h} = \lim_{h \downarrow 0} \frac{F_X(h)}{h} = \lim_{h \downarrow 0} \frac{1 - \exp(-h)}{h}$ (memoryless property)

Note

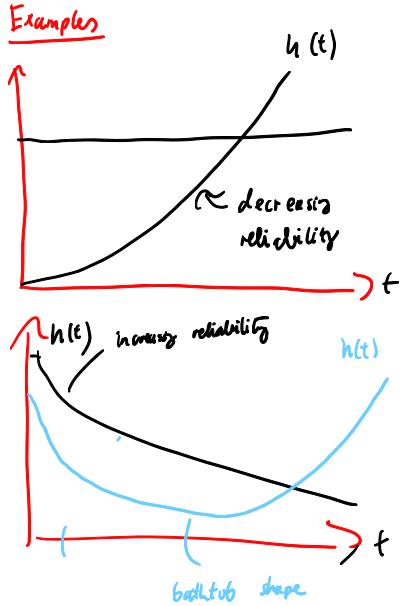
$$h(x) = \lim_{h \downarrow 0} \frac{F_x(h)}{h} = f_x(0).$$

The interpretation of $h(x)$ is important. It represents the instantaneous death rate for an individual who has survived to time x.

Or, approximately, for small h

$$\mathrm{P}(T \leq x + h | T > x) \approx h h(x). \tag{1}$$

Given an individual has reached age x, the probability of death in the next short period of time of length h is roughly proportional to h, the constant of proportionality being $h(x)$.



Important result

$$h(x+t) = \frac{f_x(t)}{S_x(t)},$$

or alternatively

$$f_x(t) = h(x+t) S_x(t).$$

So in particular, for $x = 0$ we have

$$f = hS.$$

Proof

$$\begin{aligned}
f_x(t) &= \lim_{h \downarrow 0} \frac{1}{h} \{\mathrm{P}(T_x \leq t+h) - \mathrm{P}(T_x \leq t)\} \\
&= \lim_{h \downarrow 0} \frac{1}{h} \{\mathrm{P}(T \leq x+t+h | T > x) - \mathrm{P}(T \leq x+t | T > x)\} \\
&= \lim_{h \downarrow 0} \frac{1}{h} \frac{\{F(x+t+h) - F(x)\} - \{F(x+t) - F(x)\}}{S(x)} \\
&= \lim_{h \downarrow 0} \frac{1}{h} \frac{F(x+t+h) - F(x+t)}{S(x)} \\
&= \frac{S(x+t)}{S(x)} \lim_{h \downarrow 0} \frac{1}{h} \frac{F(x+t+h) - F(x+t)}{S(x+t)} \\
&= \frac{S(x+t)}{S(x)} \lim_{h \downarrow 0} \frac{1}{h} \mathrm{P}(T \leq x+t+h | T > x+t) \\
&= \frac{S(x+t)}{S(x)} h(x+t) \\
&= S_x(t) h(x+t). \qquad \square
\end{aligned}$$

Summary

- T_x is a continuous random variable denoting the random future lifetime of an individual alive at age x.
- The distribution function $F_x(t)$is defined on $[0, \omega - x]$ with PDF $f_x(t) = F_x'(t)$.

CDF

- The survivor function and the PDF provide two methods of specifying the distribution of a continuous random variable.
- An additional way was provided via the hazard function

$$h(x+t) = \frac{f_x(t)}{S_x(t)} = \frac{-S_x'(t)}{S_x(t)}.$$

Why additionally consider the hazard rate when we have $F(t)$ and $f(t)$?

(i) It may be physically enlightening to consider the immediate risk.

(ii) Comparisons of groups of individuals are sometimes most incisively made via the hazard.

(iii) Hazard-based models are often convenient when there is censoring.

(iv) When fitting parametric models the form of the hazard function can be enlightening about the assumptions made by the model: e.g. Exponential $\implies$ constant hazard.

(v) The hazard rate does not need to satisfy as many conditions as pdf/CDF.

2.5 Cumulative Hazard Function

Definition: The cumulative (or integrated) hazard rate of T_x, denoted $H_x(t)$, is simply

$$H_x(t) = \int_0^t h(x+s)ds.$$

[$x=0$: $H(t) = \int_0^t h(s)\,ds$]

This leads to another important relationship,

$$S_x(t) = \exp\{-H_x(t)\}.$$

⇒ distribution can be defined via the hazard rate or the integrated hazard rate

Example $T \sim$ Exponential (1)
$S(t) = \exp(-t)$
$\therefore H(t) = t$

Proof

Recall,

$$h(x+t) = \frac{-S_x'(t)}{S_x(t)} \quad \text{pdf}$$

$$= -\frac{d}{dt}\log S_x(t).$$

$\therefore H(x,t) = -\log S_x(t) + C$

Furthermore we have the condition $S_x(0) = 1$ from which $S_x(t) = \exp\{-H_x(t)\}$ follows. If $C=0$

Note,

$$f_x(t) = h(x+t)\exp\{-H_x(t)\}.$$

2.6 Censoring

A defining feature of survival data, which renders many standard statistical analysis methods inappropriate, is that survival times are frequently censored.

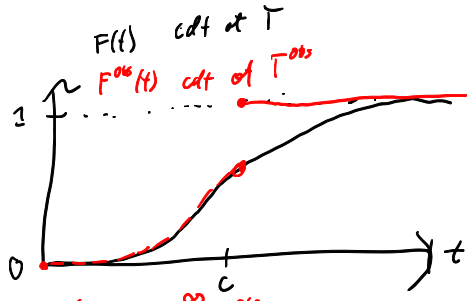
A survival time is said to be censored if the exact event time (e.g. time of death) has not been observed. Such observations provide only partial information about the value of T.

There are a number of common censoring mechanisms that prevent observation of some event times.

2.6.1 Right-censoring

An event time is right-censored if the censoring mechanism prematurely terminates observation of the individual, before the event has actually occurred. For example, if we lose track of an individual, or the study comes to an end. We only learn that $T > t$ for some $t \in (0, \omega)$.

If it is known beforehand that right-censoring is due to happen at time c for all events which have not yet occurred, then the time spent observing the individual $T_{\text{obs}} = \min(T, c)$ is a random variable of mixed type. That is, T_{obs} has continuous distribution over the interval $[0, c)$ and then a discrete atom of mass at time c.



$$E[T^{obs}] = \int_0^\infty S^{obs}(t)\,dt$$

2.6.2 Left-censoring

An event time is left-censored if we discover the event occurred before observation of the individual began; alternatively, it could be that we observe the event time but only have a lower bound on when the life of the subject began. In either case, all we can say is that the actual time to event T is less than some time t. For example, we begin a study of patients three months after an operation and find that some have died. All we know is that $T < 3$ months for those patients.

2.6.3 Interval-censoring

When all that is known is that the event occurred within an interval. For example, if we check the status of individuals every d days.

Although always undesirable, a censoring mechanism may inevitably be introduced through the design of an experiment or data collection process.

↑ "Standard" can adjust for these

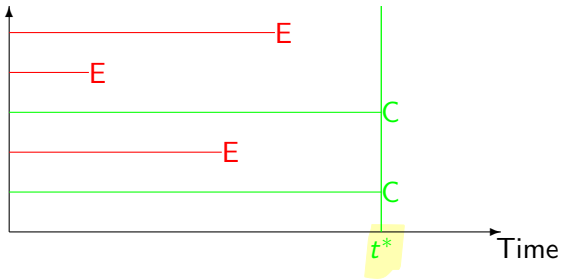
2.6.4 Truncation

Truncation happens when we wish to estimate the distribution of T but some of our data is sampled instead from a conditional distribution such as $T \mid a < T \leq b$. Truncation is somewhat similar to censoring but is not a type of censoring.

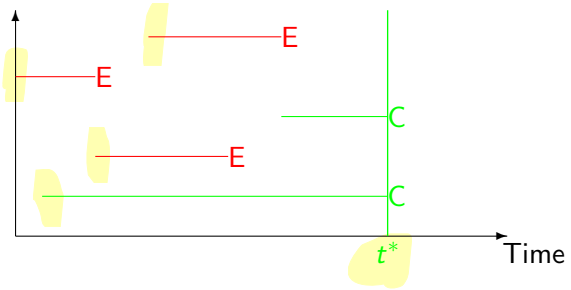
Cannot make statements about unconditional distribution → need additional data / assumptions

2.6.5 Type I censoring

Type I censoring occurs if we take n individuals and observe them for a pre-specified time t^*. Any non-events are right censored with censoring time t^*. Commonly found in medical studies, e.g. follow 15 patients for two years after operation.



- Generalised Type I censoring: As above, although now individuals join the study at different times ⇒ even censored observations have different durations on the study.



2.6.6 Type II censoring

Where we take n individuals and observe them until the d^{th} event occurs. The remaining non-events are right-censored, although (unlike Type I) this censoring time is not known in advance. Commonly found in reliability analysis, e.g. take 20 components and keep testing until half of them fail.

2.6.7 Competing risks

Suppose we are interested in the marginal distribution of failure time T_1 due to a particular cause, but the occurrence of a separate event, a competing risk, at random time T_2 would cause the individual to exit the study. Then we will only be able to observe $\min(T_1, T_2)$, along with the type of failure. Observations where $T_2 < T_1$ will constitute right-censored observations of T_1.

One important assumption made throughout this course is that there is independence between the censoring mechanism and the event time. That is, if C denotes a (possibly deterministic) censoring time, and T is the (perhaps unobserved) event time, for simplicity of inference we assume T and C are statistically independent random variables.

3 Parametric Distributions of Random Lifetimes T

3.1 Introduction

In Chapter 2 we defined various mathematical expressions that could be used to describe the probability distribution of a future random lifetime T. But until now, we have not considered what forms these quantities might take.

In this chapter we shall meet some standard parametric distributions that are commonly used in survival analysis.

Clearly, *any* distribution over the positive half-line $\mathbb{R}^+$ is a possible candidate. Moreover, distributions on $\mathbb{R}$ can be models for $\log T$. However, a number of standard distributions have emerged as being particularly appropriate for the task of analysing survival data.

We shall consider the three most popular models, these being the exponential, the Weibull and the Gompertz-Makeham distributions. Each will make a different assumption about the nature of the hazard rate.

The parametric models in this chapter do not admit a limiting age ω, all three have support over the whole of $\mathbb{R}^+$. If we were to truncate these distributions to lie on $[0, \omega]$, we should be mindful of the induced changes in the nature of the hazard functions.

3.2 Exponential Distribution

The exponential distribution is a natural starting point as a lifetime distribution. For $\lambda > 0$, if $T \sim \text{Exp}(\lambda)$ then

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

The constant hazard function reflects a property known as *lack of memory*; for the exponential distribution,

$$S_x(t) = S(t).$$

In practice however, the assumption of a constant hazard is often untenable and when fitting the exponential distribution to data it can be sensitive to outliers.

The other two distributions we now consider can be seen to generalise the exponential distribution; that is, each contains the exponential distribution within their family as a special case.

3.3 Weibull Distribution

The Weibull distribution can be written as

$$\begin{aligned} f(t) &= \eta \alpha^{-\eta} t^{\eta-1} \exp(-(t/\alpha)^{\eta}), \\ S(t) &= \exp(-(t/\alpha)^{\eta}), \\ h(t) &= \eta \alpha^{-\eta} t^{\eta-1}, \\ H(t) &= (t/\alpha)^{\eta}, \end{aligned}$$

where $\alpha > 0$ and $\eta > 0$ are known as the scale and shape parameters respectively.

The Weibull generalises the exponential distribution, which is recovered when $\eta = 1$.

Figures 1 and 2 show the hazard and density functions for different values of the shape parameter.

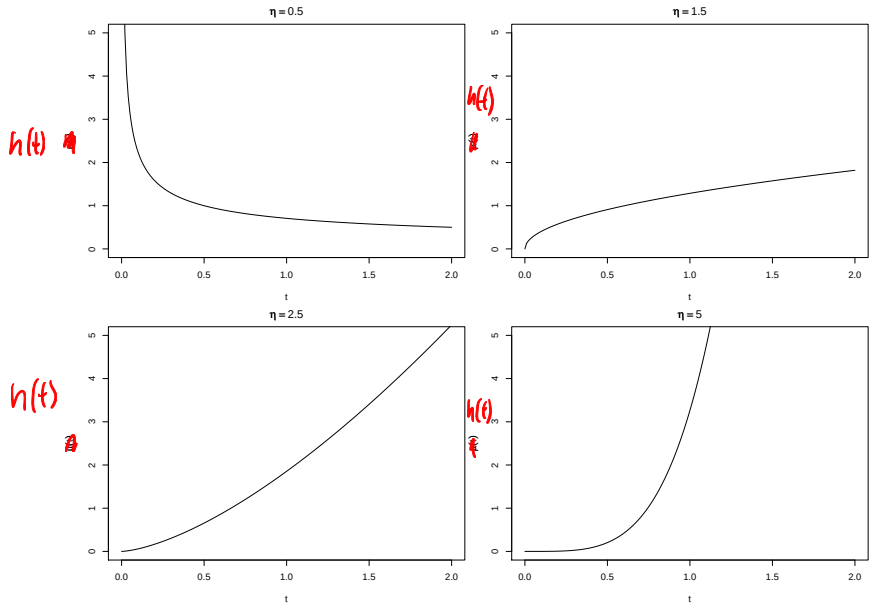


Figure 1: Hazard functions for Weibull with mean 1 and shape parameter η.

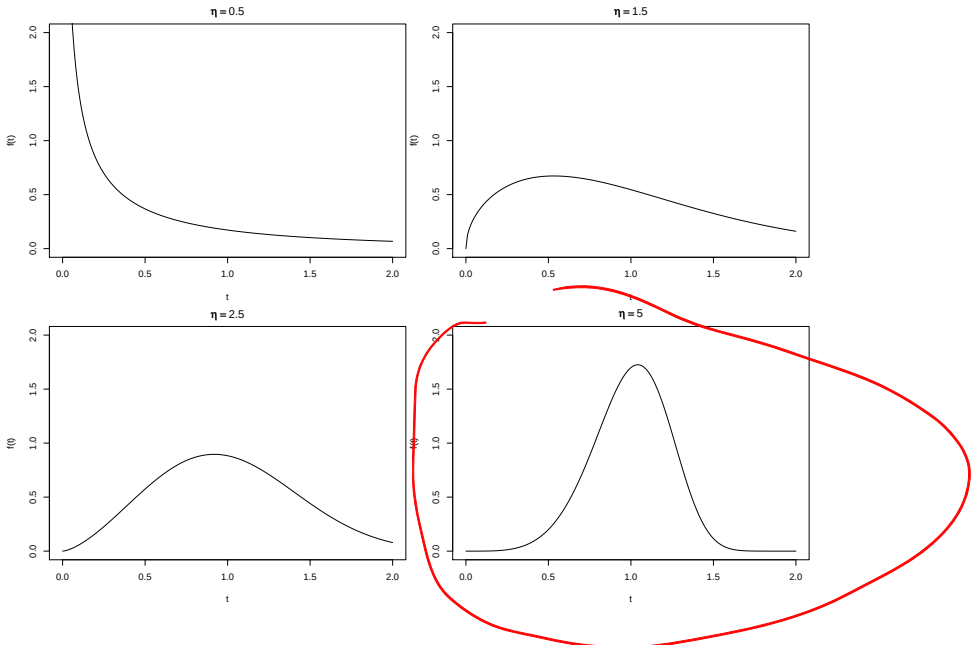


Figure 2: Density function for Weibull with mean 1 and shape parameter η.

The mean and variance of the Weibull density are

$$\alpha\Gamma(\eta^{-1}+1) \quad \text{and} \quad \alpha^2\{\Gamma(2\eta^{-1}+1) - [\Gamma(\eta^{-1}+1)^2]\}$$

respectively, where Γ is the gamma function $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$.

For $\eta < 1$ we have a monotonically decreasing (antitonic) hazard rate and for $\eta > 1$ we have a monotonically increasing (isotonic) hazard rate (see Fig. 1).

In particular for $1 < \eta < 2$ the hazard increases slower than linear in t; for $\eta = 2$ the increase is linear and for $\eta > 2$, faster than linear.

The Weibull distribution is probably the most widely used parametric distribution in survival analysis. Some possible reasons:

- Simplicity of the survivor and hazard functions.
- Covers a wide variety of distributional shapes (see Fig. 2).
- Empirically it has been found to be accurate in many contexts. It is an *extreme value distribution*.

3.4 Gompertz-Makeham Distribution

3.4.1 Gompertz and Makeham laws of mortality

The Gompertz law of mortality states that in a low mortality environment where external causes of death are rare, the force of mortality increases approximately exponentially with age. Empirical data gathered from observation of insects kept in laboratory conditions support this hypothesis.

Outside of such an environment, the Makeham law of mortality considers the risk of other, external causes of death to be approximately constant with age.

So together these laws suggest a hazard rate which is a sum of constant (Makeham) and exponential (Gompertz) terms. This motivates the Gompertz-Makeham distribution.

3.4.2 Gompertz-Makeham hazard function

The Gompertz-Makeham distribution is a three parameter distribution with hazard function

$$h(t) = \theta + \beta \exp(\gamma t)$$

for non-negative parameters θ, β, γ.

This distribution again generalises the exponential distribution. The survivor function and density function can be derived in the usual way (Exercise).

3.5 Choosing a Distribution

We have listed three common parametric distributions for lifetime random variables (there are many more: gamma, log-normal, log-logistic,...).

Given a particular data set of, say, n realisations of T, how do we decide which distribution is appropriate?

Each distribution makes particular assumptions about the form of the hazard. Knowledge of the 'true' hazard rate would enable us to decide which distribution was 'closest'. In practice we can compare to a non-parametric estimate of the integrated hazard which converges to the truth.

3.5.1 Empirical survivor function

Suppose we have gathered some survival time data for n individuals drawn from a population with common survivor function $S(t)$. Assume, for now, that there are no censored observations. Then the empirical survivor function

$$\hat{S}(t) = \frac{\text{number of observations} > t}{n}$$

provides an estimate of $S(t)$ derived purely from the data. Notice that $\hat{S}(t)$ is an antitonic step function with jumps at the death times. ($\hat{S}(t)$ is a *non-parametric*

CLT: Average or Sum of indep. r.v. behaves like a normal distr.

EVT Look at max (or min) of r.v.

If $\frac{\max(X_1,\dots,X_n) - c_n}{d_n} \xrightarrow{d} G$

then is either a Fréchet, Gumbel or a (reverse) Weibull distr.

cdf $F(x) = \begin{cases} e^{-(-x)^\alpha}, & x<0 \\ 1, & x \geq 0 \end{cases}$

[Weibull distr. flipped at 0]

[Fisher–Tippett–Gnedenko Theorem]

Weibull good model for min. of r.v.

estimator of $S(t)$. In Chapter 5 we shall see how to estimate S in the presence of censoring.)

Since $S(t) = \exp\{-H(t)\}$, the simple transformation

$$\hat{H}(t) = -\log \hat{S}(t)$$

then provides a similar, data-based estimate of the cumulative hazard.

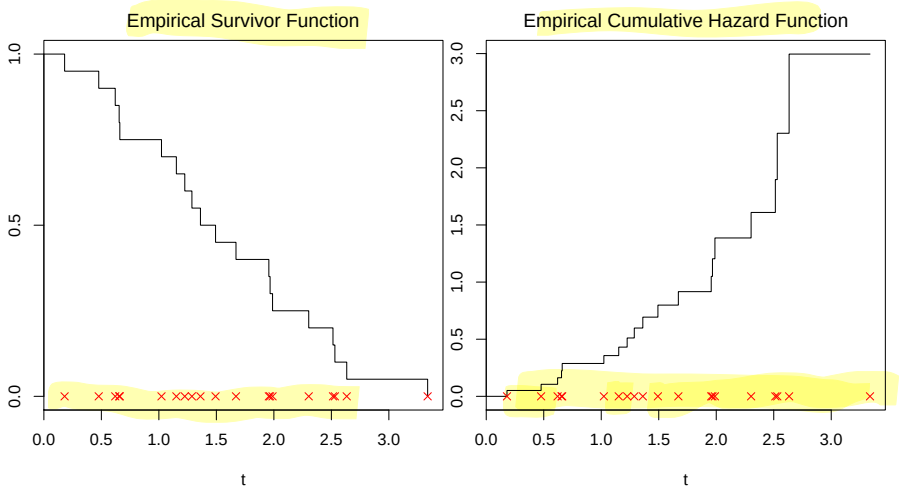


Figure 3: $\hat{S}(t)$ and $\hat{H}(t)$ for a sample of uncensored lifetime data.

We can use a plot of $\hat{H}(t)$ to see how the overall shape of the function compares with those dictated by some different distributions. In particular,

- $H(t)$ vs. t is linear for the exponential;
- $\log H(t)$ vs. $\log t$ is linear for the Weibull;
- $\log H(t)$ vs. t approaches linearity for large t for the Gompertz-Makeham.

After plotting $\hat{H}(t)$ in these ways we can decide which assumption is closest. This provides us with a crude but useful method of selecting a parametric model for a given set of data.

4 Fitting Parametric Distributions

In Chapter 3 we met some standard parametric families that might be appropriate for modelling T, and gave some simple criteria for choosing amongst them given a set of realised values.

We now assume that a distribution has been selected. What values should the parameters of the distribution take? This is a problem in **statistical inference**.

The most popular statistical approach to fitting parametric distributions to data is the method of **maximum likelihood (ML) estimation**.

In survival analysis the inference problem is often complicated by the presence of censored observations, and so we will need to consider ML estimation in this context.

4.1 Maximum Likelihood Estimation

Assume that the data $(t_1, \ldots, t_n)$ are n independent, possibly censored realisations from the distribution $F(t;\theta)$, where the form of F is known (e.g. Weibull) but the parameters θ are unknown.

Definition: The likelihood is the joint probability of the observed data, regarded as a function of the unknown parameters θ.

The **likelihood principle** states that all of the information about the parameters in a sample of data is contained in the likelihood function.

The **law of likelihood** extends this principle to state that the degree to which the data supports one parameter value over another is given by the ratio of their likelihoods.

The Maximum likelihood estimate (MLE) of θ is based solely upon the likelihood of the observed data. Maximum likelihood estimation thus provides a principled and general method of estimating parameters in parametric distributions using observed data.

As the name suggests the MLE is the value of the parameters that maximises the probability of the observed data,

$$\hat{\theta} = \arg\sup_{\theta \in \Theta} L(\theta),$$

$$L(\theta) = \prod_{i=1}^{n} \Pr(t_i | F(;\theta)),$$

density

where L is the likelihood function and $\Pr(t_i|F(;\theta))$ is the probability of observing the i^{th} observation given the distribution F with parameters θ.

If none of the observations are censored we find,

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta).$$

For each censored observation the contribution to the likelihood depends on the type of censoring. Recall, an observation is left (right) censored if we only have an upper (lower) bound for T, or interval censored if we only know that T lies in a specified interval.

- $\Pr(t_i|F(;\theta)) = S(t_i;\theta)$ for an observation right-censored at t_i;
- $\Pr(t_i|F(;\theta)) = F(t_i;\theta)$ for an observation left-censored at t_i;

- $\Pr(t_i|F(;\theta)) = F(t_i^{(u)};\theta) - F(t_i^{(l)};\theta)$ for an observation interval-censored within $[t_i^{(l)}, t_i^{(u)}]$.

From now on we shall only consider right-censoring, this being the most common.

We can then split the data into two disjoint sets relating to

$$U = \text{uncensored data}$$
$$C = \text{censored data.}$$

The likelihood function is then

$$L(\theta) = \prod_{i\in U} f(t_i;\theta) \prod_{i\in C} S(t_i;\theta).$$

It is almost always more convenient to work with the log-likelihood,

$$\begin{aligned}\ell(\theta) = \log\{L(\theta)\} &= \sum_{i\in U} \log f(t_i;\theta) + \sum_{i\in C} \log S(t_i;\theta) \\ &= \sum_{i\in U} \log h(t_i;\theta) + \sum_{i=1}^{n} \log S(t_i;\theta) \\ &= \sum_{i\in U} \log h(t_i;\theta) - \sum_{i=1}^{n} H(t_i;\theta).\end{aligned}$$

For an m-parameter distribution, the estimate $\hat{\theta}$ may be found by solving the likelihood equations

$$\frac{\partial}{\partial\theta_j}\ell(\theta) = 0 \quad (j = 1, \ldots, m).$$

Finding the solution often involves numerical techniques such as Newton and quasi-Newton methods.

4.1.1 Newton's Method

Newton's Method for optimising $\ell(\theta)$ begins by first making an initial guess $\theta = \theta_0 \in \mathbb{R}^m$, sufficiently close to the true optimum. Provided this initial guess θ_0 is incorrect, we would wish to add an increment $\delta \in \mathbb{R}^m$ to θ_0 s.t. $\ell(\theta_0 + \delta)$ is the optimum.

The (multivariate) Taylor expansion of ℓ about a value θ yields

$$\ell(\theta+\delta) \approx \ell(\theta) + \delta'\nabla\ell(\theta) + \frac{1}{2}\delta'\nabla^2\ell(\theta)\delta. \tag{2}$$

To solve the easier problem of maximising the Taylor expansion (2), we find the derivative of (2) wrt δ is equal to

$$0 = \nabla\ell(\theta) + \nabla^2\ell(\theta)\delta$$

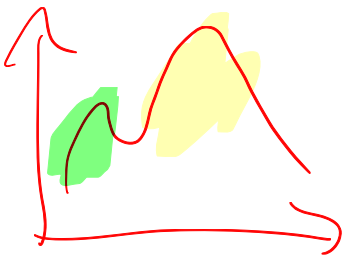
and hence get an approximate solution for δ of

$$\delta = -\{\nabla^2\ell(\theta)\}^{-1}\nabla\ell(\theta).$$

So the algorithm proceeds iteratively, setting

$$\theta_n = \theta_{n-1} - \{\nabla^2\ell(\theta_{n-1})\}^{-1}\nabla\ell(\theta_{n-1})$$

until sufficient convergence in the sequence $\theta_0, \theta_1, \theta_2, \ldots$ occurs.



$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \xrightarrow{d} N(0, I_1(\vartheta)^{-1})$ $(n \to \infty)$

$\hat{\vartheta}_n$: MLE based on n observations; ϑ: true value; $I_1(\vartheta)$: Fisher information matrix based on one observation. {under regularity conditions}

4.1.2 Asymptotic distribution of MLE

Maximum likelihood estimation not only provides a set of optimal parameter values for θ but also allows assessment of the variance in the estimates.

Consider the $m \times m$ information matrix $I(\theta) = -\nabla^2 \ell(\theta)$, so can be estimated via

$$I(\theta)_{ij} = \frac{-\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta), \quad i, j = 1, \ldots, m.$$

Then asymptotically for large samples, $\hat{\theta} \sim \mathrm{N}(\theta, I(\theta)^{-1})$.

Evaluating $I(\theta)$ by setting the unknown θ equal to $\hat{\theta}$, giving the observed information matrix, leads to an approximate covariance matrix for $\hat{\theta}$. That is, if $V = I(\hat{\theta})^{-1}$ then its ij^{th} entry v_{ij} is an estimate of the covariance between $\hat{\theta}_i$ and $\hat{\theta}_j$. ← based on sample size n

In particular the standard error of $\hat{\theta}_j$ is given by the square root of the j^{th} diagonal element of V,

$$\text{s.e.}(\hat{\theta}_j) \approx \sqrt{v_{jj}}.$$

4.1.3 Functional invariance of MLEs

Another key advantage of ML estimation is that the MLE of a function of the parameters $g(\theta)$ is simply

$$\widehat{g(\theta)} = g(\hat{\theta}).$$

For example, the exponential distribution has mean $\mathrm{E}(T) = 1/\lambda$, hence the MLE of $\mathrm{E}(T)$ is

$$\widehat{\mathrm{E}(T)} = \frac{1}{\hat{\lambda}} = \frac{1}{r}\sum_{i=1}^{n} t_i,$$

which is the total time on the study survived by the individuals divided by the number of deaths.

4.2 Examples of ML Estimation

We assume that we have observed lifetimes $t_1, \ldots, t_n$ with r uncensored observations and $n - r$ right-censored observations.

4.2.1 Exponential distribution

The log-likelihood is

$$\ell(\lambda) = r \log \lambda - \lambda \sum_{i=1}^{n} t_i.$$

$= \lambda$; $= \sum_{i \in U} \log(h(t_i; \lambda)) - \sum_{i=1}^{n} H(t_i; \lambda)$, $= \lambda t_i$

Hence

$$\frac{d\ell}{d\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i,$$

which may be set to zero and solved immediately to give

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i}.$$

check 2nd derivative

Also we have,

$$\frac{-d^2\ell}{d\lambda^2} = \frac{r}{\lambda^2} \implies \text{s.e.}(\hat{\lambda}) \approx \frac{\hat{\lambda}}{r^{1/2}}.$$

So we have approximately

$$\hat{\lambda} \sim \text{Normal}(\lambda, \lambda^2/r),$$

and then more approximately

$$\hat{\lambda} \sim \text{Normal}\left(\lambda, r \Big/ \left\{\sum_{i=1}^{n} t_i\right\}^2\right).$$

4.2.2 Weibull distribution

Recall the Weibull survivor function has form,

$$S(t) = \exp(-\lambda t^{\eta})$$

with scale λ and shape η. Note, we have changed the parameterisation slightly, writing λ for $\alpha^{-\eta}$.

The log-likelihood in the presence of right-censoring is

$$\ell(\lambda, \eta) = r \log(\lambda\eta) + (\eta - 1) \sum_{i \in U} \log t_i - \lambda \sum_{i=1}^{n} t_i^{\eta}.$$

We set

$$\frac{\partial \ell}{\partial \lambda} = 0 = \frac{r}{\hat{\lambda}} - \sum_{i=1}^{n} t_i^{\hat{\eta}} \tag{3}$$

and

$$\frac{\partial \ell}{\partial \eta} = 0 = \frac{r}{\hat{\eta}} + \sum_{i \in U} \log t_i - \hat{\lambda} \sum_{i=1}^{n} t_i^{\hat{\eta}} \log t_i. \tag{4}$$

Solving (3), we obtain

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\eta}}}.$$

Substituting into (4) gives

$$\frac{r}{\hat{\eta}} + \sum_{i \in U} \log t_i - \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\eta}}} \sum_{i=1}^{n} t_i^{\hat{\eta}} \log t_i = 0.$$

This is a non-linear equation in $\hat{\eta}$ which can only be solved using an iterative numerical procedure.

In practice it is common to maximise $\{\lambda, \eta\}$ simultaneously using Newton's method. An important by-product of which is an approximation of the covariance matrix from which standard errors on MLE can be obtained.

4.3 Hypothesis Testing for Nested Distributions

At the end of Chapter 3 we provided some heuristic methods for distinguishing between distributions. Here we consider asymptotic properties of the likelihood function to provide a more rigorous test that can be used for certain comparisons.

Suppose we are interested in making statements about a parameter subset $\theta^{(A)} \subseteq \theta \in \Theta$. Assume that θ is partitioned $\theta = (\theta^{(A)}, \theta^{(B)})$ and $\Theta = \Theta^{(A)} \times \Theta^{(B)}$.

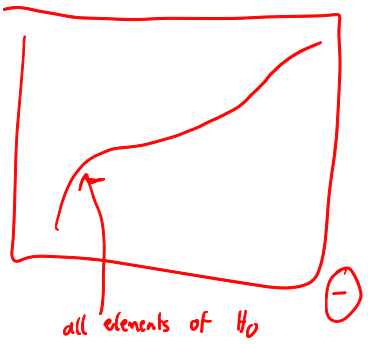
- For example, in the Weibull distribution we might take $\theta^{(A)} = \eta$ (shape), $\theta^{(B)} = \lambda$ (scale).

Let $\hat{\theta} = (\hat{\theta}^{(A)}, \hat{\theta}^{(B)})$ denote the MLE of $\theta = (\theta^{(A)}, \theta^{(B)})$.

We will consider two tests for the null hypothesis $H_0 : \theta^{(A)} = \theta_0^{(A)}$ vs. $H_1 : \theta^{(A)} \in \Theta^{(A)}$ which make use of the MLE.

- Note that for the Weibull distribution a test for $\eta = 1$ is a test of exponentiality.

The tests lead to the derivation of a confidence region for $\theta^{(A)}$ which is the collection of parameter values in the subset $\Theta^{(A)}$ not 'rejected' at a certain significance level.



4.3.1 Likelihood ratio statistic

Let

- $\hat{\theta}_{H_1} = \hat{\theta}$ be the unconstrained MLE over Θ.
- Let $\hat{\theta}_{H_0} = \left(\theta_0^{(A)}, \hat{\theta}^{(B)}_{\theta_0^{(A)}}\right)$, where $\hat{\theta}^{(B)}_{\theta_0^{(A)}}$ is the MLE estimate of $\theta^{(B)}$ over $\Theta^{(B)}$ with $\theta^{(A)}$ fixed at $\theta^{(A)} = \theta_0^{(A)}$.

Then the likelihood ratio test statistic is

$$W\left(\theta_0^{(A)}\right) = -2\left[\ell\left(\hat{\theta}_{H_0}\right) - \ell\left(\hat{\theta}_{H_1}\right)\right].$$

Under the null hypothesis $H_0 : \theta^{(A)} = \theta_0^{(A)}$, W is itself a random variable with an approximate chi-squared distribution with m_A degrees of freedom, where $m_A = \dim\left(\theta^{(A)}\right)$.

Large values of W relative to $\chi^2_{m_A}$ supply evidence against H_0.

The corresponding $1-\alpha$ confidence region for $\theta^{(A)}$ is

$$\left\{\theta^{(A)} : W\left(\theta^{(A)}\right) \leq \chi^2_{m_A,\alpha}\right\},$$

where $\chi^2_{m_A,\alpha} =$ is the upper 100α percentage point of $\chi^2_{m_A}$. Within this region we cannot reject the null hypothesis at the α significance level.

4.3.2 Wald statistic

Suppose $V = V\left(\hat{\theta}^{(A)}, \hat{\theta}^{(B)}\right)$ is the asymptotic covariance matrix for $\left(\hat{\theta}^{(A)}, \hat{\theta}^{(B)}\right)$ evaluated at the MLE.

Let V_A be the leading submatrix of V corresponding to $\theta^{(A)}$. Then,

$$W^*\left(\theta_0^{(A)}\right) = \left(\hat{\theta}^{(A)} - \theta_0^{(A)}\right)' V_A^{-1}\left(\hat{\theta}^{(A)} - \theta_0^{(A)}\right)$$

also has an approximate $\chi^2_{m_A}$ distribution under the null hypothesis $\theta^{(A)} = \theta_0^{(A)}$.

The corresponding $1-\alpha$ confidence region for $\theta^{(A)}$ is

$$\left\{\theta^{(A)} : W^*\left(\theta^{(A)}\right) \leq \chi^2_{m_A,\alpha}\right\}.$$

Both the Wald statistic and the likelihood ratio test statistic are based on large sample theory and both are asymptotically equivalent (and often give similar results in practice). However, large discrepancies are possible.

In such cases the likelihood ratio statistic is perhaps preferable as the results are invariant to reparameterisation.

Against this, the Wald statistic has the advantage that only one maximisation, over the unconstrained parameter space Θ, is required.

Note however, that the theory is for large samples and may give a poor approximation to small-sample results.