Stat 486 / 886 Survival Analysis

Chapter 1. Introduction

Professor: Wenyu Jiang

Department of Mathematics and Statistics
Queen's University

Acknowledgement: Profs. Jerry Lawless, David Matthews (Waterloo),
Profs. Paul Peng, Dongsheng Tu (Queen's)

Sections of Chapter 1

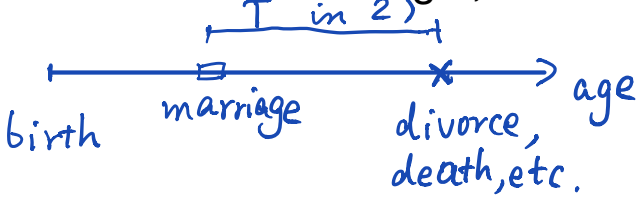
1. Scope of Survival Analysis
2. Introductory Examples
3. Survival Distributions
4. Censoring, Survival Data and Likelihood Function

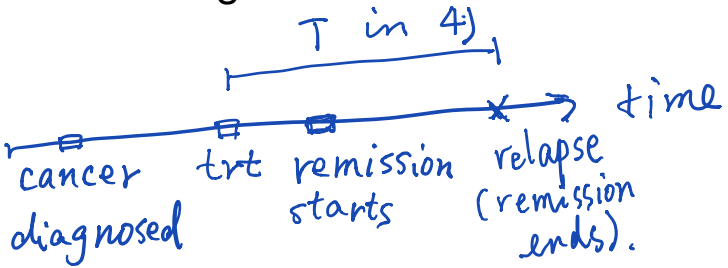
1.1 Scope of Survival Analysis

Response variable – time T to some event,
often called survival time, lifetime, failure time, or duration time.
For example:
1) T = length of life of a person = age at death;
2) T = duration of a marriage;
3) T = lifetime of a piece of equipment (eg. battery of a laptop);
4) T = time to relapse of cancer after chemotherapy.

Important issues

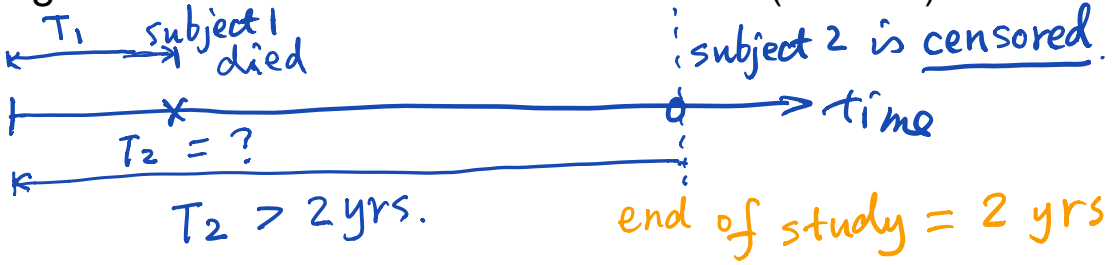
- Choices of time origin, unit and scale for defining time variables.





Important issues

- Length of time needed to observe events (failures) that define T;



  Note: Censoring is a special issue in survival analysis.

- Explanatory variables z for a given individual.
  z can be qualitative: gender, treatment, tumor type, etc.;
  can be quantitative: weight, waiting time before treatment;
  can change over time: white blood cell count.

  Regression setup:

  T – response (time); z – explanatory variables (covariates)

  A regression model describes how the distribution of T depends on the values of z.

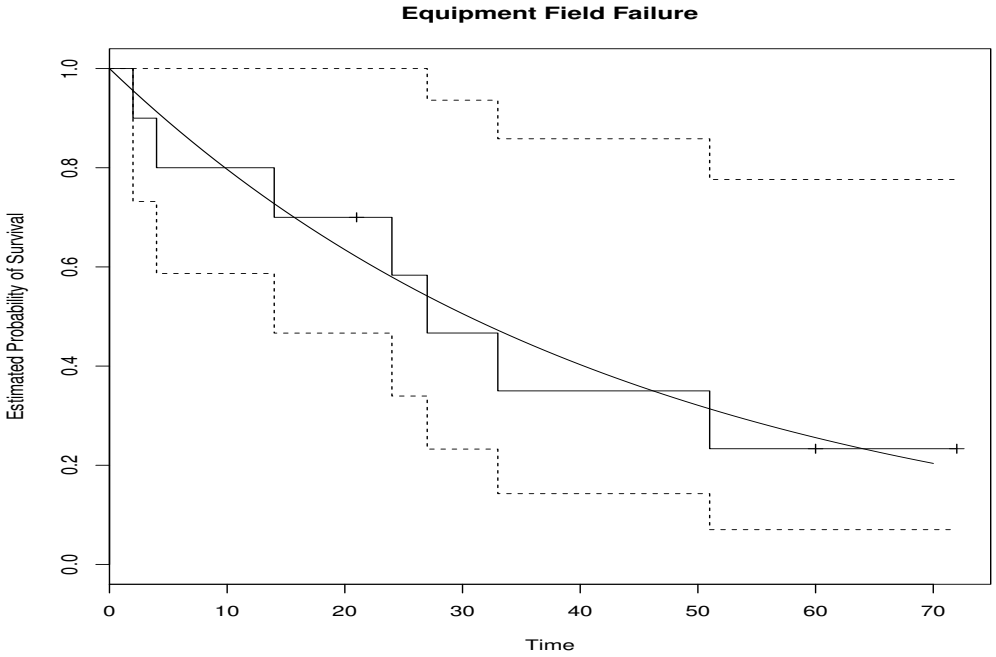
- The reference population, and choice of individuals for study.

1.2 Introductory Examples

Example 1.2.1 Equipment Field Failure

The time below show the number of days until the first malfunction ("failure") of 10 similar units of equipment. The units were installed at different times and when the data were assembled, three units had not yet failed; their number of days in service are shown with asterisks. (Reference: Lawless 2002)

2, 4, 14, 21*, 24, 27, 33, 51, 60*, 72*

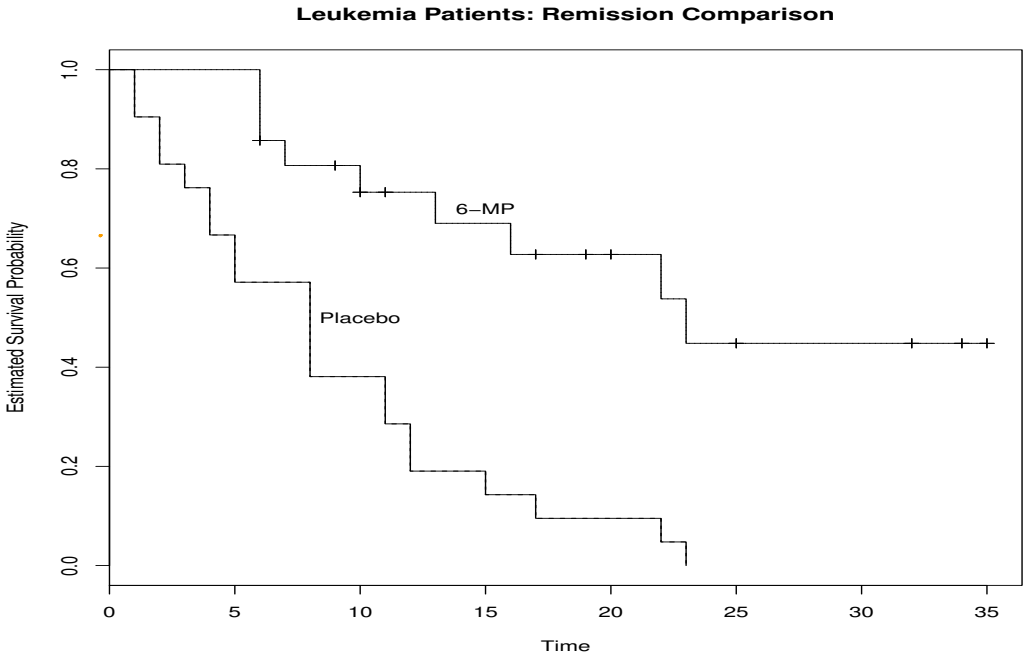

Equipment Field Failure
Estimated Probability of Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Time

Example 1.2.2 Leukemia Remission Comparison

The data below are remission times for patients with acute leukemia who were randomized in a clinical trial to receive either a drug (6-mercaptopurine, i.e. 6-MP) or a placebo. The intended purpose of the drug was to maintain remission in an individual. The times are in weeks and the asterisks denote censoring times. (Reference: Freireich et al. 1963; Klein and Moeschberger, 2003)

Control (Placebo) Group:
1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Drug 6-MP Group:
6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*


Leukemia Patients: Remission Comparison
Estimated Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
6–MP
Placebo
0
5
10
15
20
25
30
35
Time

Example 1.2.3 Bladder Cancer Recurrence

The data below are taken from a tumor recurrence study for patients with bladder cancer. Individuals had 0–4 recurrences during the follow-up time. (Reference: Wei, Lin and Weissfeld, JASA 1989)

Define the **response time** based on the follow-up time (futime) and first recurrence time (in months).

Covariates include
treatment group: 1=placebo, 2=drug thiotepa
size: size of the largest initial tumor (in centimeters)
number: number of tumors at initial diagnosis.

Data set:

Group	futime	number	size	recurrence	times		
...	...	...					
1	18	1	3	5			
1	18	1	1	12	16		
1	23	3	3				
...	...	...					
1	24	2	3	7	10	16	24
1	25	1	1	3	15	25	
...	...	...					

The examples above illustrate 3 questions of scientific interests.

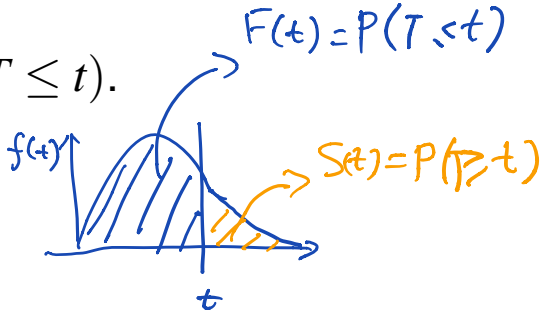
- Characterize the distribution of response time in a single study population.
  Analysis: univariate survival time models.
- Compare response time distributions in 2 or more study populations.
  Analysis: test/assess if survival time distributions are the same.
- Describe how response time in a study population depends on concurrently measured explanatory variables.
  Analysis: regression models.

1.3 Survival Distributions

1.3.1 Continuous time models

Failure time T: a continuous random variable, with
probability density function (pdf) $f(t)$, $t > 0$;
and cumulative distribution function (cdf) $F(t) = P(T \leq t)$.

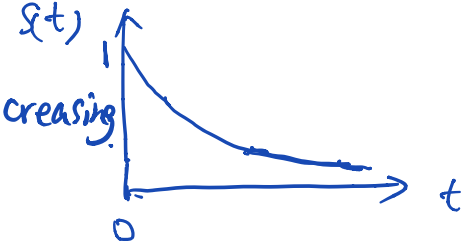
Recall that $f(t) = F'(t) = \frac{dF(t)}{dt}$ and $F(t) = \int_0^t f(x)dx$.



a. Survival function (sf)
The survival function of T is $S(t) = P(T \geq t)$,
the probability that the response time is at least t.

Properties:

$S(t)$ is monotone, non-increasing.



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

For cont r.v. T:

$$S(t) = 1 - F(t), \qquad f(t) = -S'(t).$$

$$S(t) = \int_t^\infty f(x)\,dx$$

b. Hazard function (hf)

The hazard function of T (also called intensity function) is defined as

$$h(t) = \lim_{\Delta t \to 0+} \frac{P(t \le T \le t+\Delta t | T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

Interpretation: the instantaneous failure rate at time t, for an individual who is alive at time t.

$$P(t \le T \le t+\Delta t \mid T \ge t) = \frac{P(t \le T \le t+\Delta t)}{P(T \ge t)} \approx \frac{f(t)\Delta t}{S(t)}$$

Also define cumulative hazard function for T,

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

Theorem 1.3.1 For a continuous failure time T, its survival function, hazard function and cumulative hazard function are related in the following way

$$S(t) = e^{-\int_0^t h(x)dx} = e^{-H(t)}.$$

Proof:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log S(t).$$

$$\int_0^t h(x)\,dx = -\log S(t) + C.$$

$\int_0^t h(x)\,dx = 0$ for $t=0$.

$S(0) = 1$

Initial condition: $S(0) = 1$

$$\Rightarrow C = 0.$$

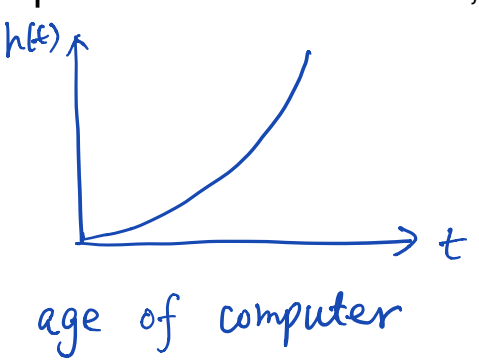
$$\Rightarrow \log S(t) = -\int_0^t h(x)\,dx.$$

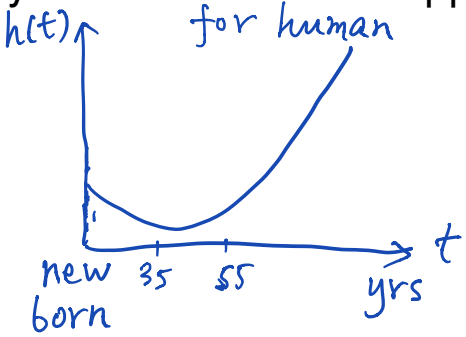
$$S(t) = e^{-\int_0^t h(x)dx} = e^{-H(t)}. \quad \square$$

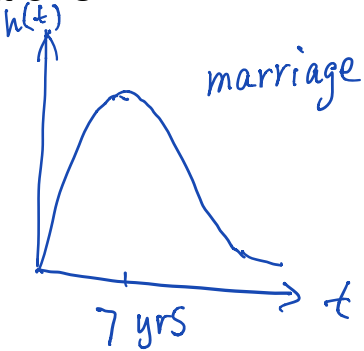
Notes:

- $h(t) \geq 0$, $\lim_{t\to\infty} S(t) = 0$, and $\lim_{t\to\infty} H(t) = \infty$.
- Each function, $S(t)$, $h(t)$ or $H(t)$, uniquely determines the distribution of T.
- In statistics (articles, application, packages), typically, $\log x$ is the notation for natural logarithm of x, i.e. $\log x = \log_e x$ or $\ln x$.

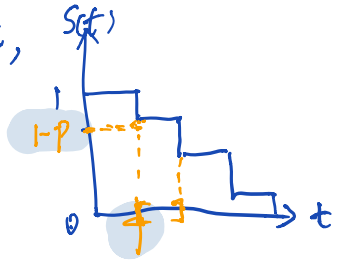
Shapes of hazard function, may be useful in some applications...







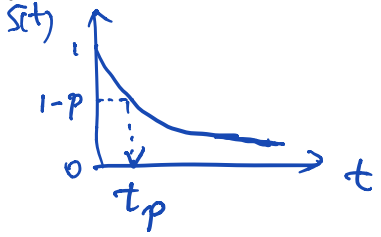
If T is not cont,



c. Some characteristics of the distribution of T

- Quantile (percentile): $t_p = \inf\{t : S(t) \le 1-p\}$,
  the pth quantile (or $100p$th percentile) of the distribution of T.

If T is a cont r.v.,
then t_p, $S(t_p) = 1-p$.
[$F(t_p) = p$].



$t_{0.5}$: median survival time.

- Mean $\mu = \mathrm{E}(T)$, and variance $\mathrm{Var}(T)$.
- Mean residual lifetime (mrl): $\mathrm{mrl}(t) = \mathrm{E}(T-t|T>t)$,
  expected remaining lifetime of a subject at time t.

1.3.2 Some Important Families of Distributions

a. Exponential distribution.

A failure time T has an exponential distribution if its survival function (sf) has the form

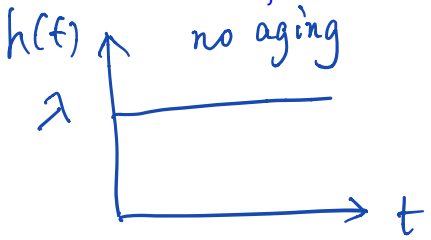
$$S(t) = e^{-\lambda t}, \text{ for } t > 0, \text{ with parameter } \lambda > 0.$$

It's a unique continuous distribution with a constant hazard function,

$$h(t) = \lambda, \text{ with } \lambda > 0.$$

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

$$= -S'(t)$$



The pdf is $f(t) = \lambda e^{-\lambda t}$, for $t > 0$.

memoriless property

$$\mathrm{E}(T) = \frac{1}{\lambda}.$$

An alternative parameterization: $\theta = 1/\lambda$.

mean

b. Weibull distribution.

A failure time T has a Weibull distribution if its sf has the form

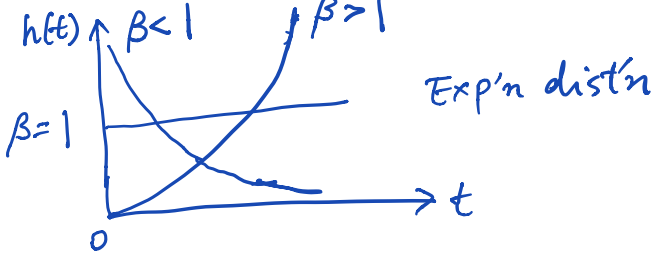
$$S(t) = e^{-\lambda t^{\beta}}, \text{ for } t > 0, \text{ with parameters } \lambda > 0, \beta > 0. \tag{1}$$

Note: If $\beta = 1$, it becomes the Exponential (λ) distribution.

Hazard function: $h(t) = \lambda \beta t^{\beta - 1}$.

$h(t) = -\frac{d}{dt} \log S(t) = -\frac{d}{dt}(-\lambda t^{\beta}) = \lambda \beta t^{\beta-1}$, $t > 0$.

[Recall $S(t) = e^{-\int_0^t h(x)\,dx}$]



Pdf: $f(t) = \lambda \beta t^{\beta - 1} e^{-\lambda t^{\beta}}$, for $t > 0$.

$= h(t) S(t)$.

Example 1.3.1 Connection between exponential and Weibull distributions.

Suppose that T has a Weibull distribution with parameters β and λ as described in (5). Let $Y = T^\beta$. Show that $T \sim$ Exponential(λ) (where λ is the hazard rate).

pf:

$$S_Y(y) = P(Y \geq y) = P(T^\beta \geq y)$$
$$= P(T \geq y^{\frac{1}{\beta}}) = S_T(y^{\frac{1}{\beta}}) = e^{-\lambda(y^{\frac{1}{\beta}})^\beta} = e^{-\lambda y}$$

for $y \geq 0$.

$\therefore$ $Y = T^\beta \sim Exp(\lambda)$. □

Characteristics of a Weibull random variable:

- $\mathrm{E}(T) = \Gamma(1+\frac{1}{\beta})/\lambda^{\frac{1}{\beta}}$ where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ is the gamma function.
- The pth quantile of T is $t_p = \{-[\log(1-p)]/\lambda\}^{\frac{1}{\beta}}$. Verify...

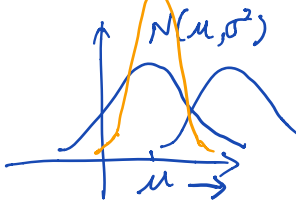
An alternative parameterization for Weibull distribution:
T is said to have a Weibull distribution if its survival function (sf) has the form

$$S(t) = e^{-(t/\alpha)^\beta}, \text{with parameters } \alpha > 0, \beta > 0. \tag{2}$$

This is also a common way to describe Weibull distribution.
Comparing (5) and (2), we see that $\lambda = \alpha^{-\beta}$.

c. Log normal distribution.

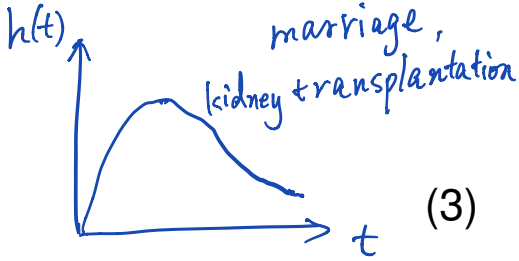
$T \geq 0$, $-\infty < \log T < \infty$



Let $Y = \log T$, and assume $Y \sim$ Normal (μ, σ^2).

Then T has a log normal distribution, denoted by LogNormal (μ, σ).

It's easy to see that the sf of T is



$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \tag{3}$$

where $\Phi(\cdot)$ is the cdf of a standard normal, N $(0, 1)$, distribution.

$$S_T(t) = P(T \geq t) = P(\log T \geq \log t)$$

$Y \sim N(\mu, \sigma^2)$ cdf of $N(0,1)$

$$= P\left(\frac{Y-\mu}{\sigma} \geq \frac{\log t - \mu}{\sigma}\right)$$

$$= P\left(Z \geq \frac{\log t - \mu}{\sigma}\right) = 1 - P\left(Z \leq \frac{\log t - \mu}{\sigma}\right)$$

$Z \sim N(0,1)$

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

d. Log location-scale models. (A family of distributions).

Let T be a failure time. Consider $Y = \log T$, i.e. log failure time.
Suppose there is a random variable W, with cumulative distribution function (cdf) $G(w)$, and probability density function (pdf) $g(w) = G'(w)$, $-\infty < w < \infty$, such that the cdf and pdf of Y can be expressed in the forms of

$$F_Y(y) = G\left(\frac{y-\mu}{\sigma}\right), \text{ and } f_Y(y) = \frac{1}{\sigma} g\left(\frac{y-\mu}{\sigma}\right), \tag{4}$$

then Y is said to have a location-scale distribution,
and T is said to have a log location-scale distribution.
The parameter μ is called the location parameter, $-\infty < \mu < \infty$,
the parameter σ is called the scale parameter, $\sigma > 0$.

In general, μ is not the mean, σ is not the standard deviation.

W is the standard distn in this loc-scale distn family.

Examples of Log Location-Scale Models

Example 1.3.2 Log normal distribution.

Let T be a failure time from LogNormal (μ, σ) distribution. Verify that T has a log location-scale distribution.

- Recall $W \sim N(0,1)$, its pdf is $g(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}$, $-\infty < w < \infty$, and its cdf is denoted by $\Phi(w)$.
- $Y \sim N(\mu, \sigma^2)$. Then $W = \frac{Y-\mu}{\sigma} \sim N(0,1)$, standard normal.

The cdf of $Y = \log T$ is $F_Y(y) = P(Y \le y) = P\left(\underbrace{\frac{Y-\mu}{\sigma}}_{W} \le \underbrace{\frac{y-\mu}{\sigma}}_{} \right) = \underbrace{\Phi}_{G}\left(\frac{y-\mu}{\sigma}\right)$.

$\therefore$ $Y \sim N(\mu, \sigma^2)$ is a loc-scale dist'n.

$\therefore$ $\Rightarrow$ T has a log loc-scale dist'n.

When $Y \sim N(\mu, \sigma^2)$.

Notes: $Y = \log T = \mu + \sigma W$ ⇐

$W \sim N(0, 1)$.

Allows for regression models. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$

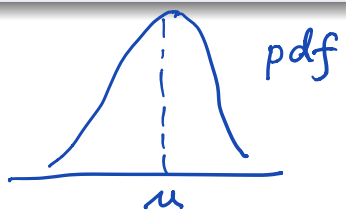
Definition: Extreme value distribution.

A random variable Y is said to have an extreme value distribution, EV (μ, σ), iff its cdf has the form

$$F_Y(y) = 1 - e^{-e^{\frac{y-\mu}{\sigma}}}, \text{ for } -\infty < y < \infty,$$

with parameters $-\infty < \mu < \infty$ and $\sigma > 0$.

Properties of the random variable $Y \sim$ EV (μ, σ).



- Its pdf is asymmetric, $f_Y(y) = \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma}} e^{-e^{\frac{y-\mu}{\sigma}}}$.
- $\mathrm{E}(Y) = \mu - 0.5772\sigma$, $\mathrm{Var}(Y) = \frac{\pi^2}{6}\sigma^2$, $\text{Median}(Y) = \mu - \sigma \log(\log 2)$.

Euler constant

EV $(\mu = 0, \sigma = 1)$ is called the standard EV distribution.

A random variable $W \sim$ EV $(0, 1)$ has cdf $G(w) = 1 - e^{-e^w}$, for $-\infty < w < \infty$.

Examples of Log Location-Scale Models

Consider $Y = \log T$ that has an EV (μ, σ) distribution.
Check for (4), it is easy to see Y has a location-scale distribution, and T has a log location-scale distribution.

$W \sim EV(0,1)$. cdf $G(w) \sim 1 - e^{-e^{w}}$, $-\infty < w < \infty$

$Y \sim EV(\mu, \sigma)$.

cdf of Y is $F_Y(y) = G\left(\frac{y-\mu}{\sigma}\right)$?

$F_Y(y) = P(Y \le y) = P\left(\frac{Y-\mu}{\sigma} \le \frac{y-\mu}{\sigma}\right) = P\left(W \le \frac{y-\mu}{\sigma}\right) = G\left(\frac{y-\mu}{\sigma}\right)$ □

$\frac{Y-\mu}{\sigma}$? $W \sim EV(0,1)$? ✓

For $Y \sim EV(\mu, \sigma)$, Let $W = \frac{Y-\mu}{\sigma}$. what's the dist'n of W ?

$F_W(w) = P(W \le w) = P\left(\frac{Y-\mu}{\sigma} \le w\right) = P(Y \le \mu + \sigma w)$

$= F_Y(\mu + \sigma w) = 1 - e^{-e^w}$, the cdf of $EV(0,1)$

$\therefore$ $Y \sim$ loc scale distn, $T \sim$ log loc-scale distn

Example 1.3.3 Connection between Weibull and EV distributions.

Let T be a failure time from Weibull distribution with cdf $F_T(t) = 1 - e^{-(t/\alpha)^\beta}$ (recall the parameterization (2)). Show that $Y = \log T$ has an EV distribution. (i.e. Weibull distribution for T is a log location-scale distribution.)

pf: For $Y = \log T$, the cdf

$F_Y(y) = P(Y \le y) = P(\log T \le y) = P(T \le e^y)$.

$= F_T(e^y) = 1 - e^{-\left(\frac{e^y}{\alpha}\right)^\beta} = 1 - e^{-\left(\frac{e^y}{e^\mu}\right)^\beta}$ Let $\alpha = e^\mu$

$= 1 - e^{-(e^{y-\mu})^\beta}$ let $\beta = \frac{1}{\sigma}$.

$= 1 - e^{-e^{\frac{y-\mu}{\sigma}}} = G\left(\frac{y-\mu}{\sigma}\right)$, where $G(\,)$ is the cdf of $W \sim EV(0,1)$.

$\therefore \quad Y \overset{=\log T}{\sim} EV(\mu, \sigma)$, is an loc-scale distn.

$\therefore \quad T \sim$ Weibull, is a log loc-scale distn.

Definition: Logistic distribution.

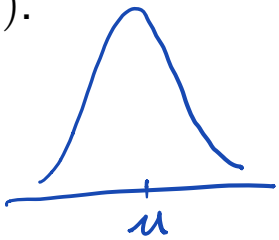
A random variable Y is said to have a logistic distribution, Logistic (μ, σ), iff its cdf has the form

$$F_Y(y) = \frac{e^{\frac{y-\mu}{\sigma}}}{1+e^{\frac{y-\mu}{\sigma}}}, \text{ for } -\infty < y < \infty,$$

with parameters $-\infty < \mu < \infty$ and $\sigma > 0$.

Properties of the random variable $Y \sim$ Logistic (μ, σ).

- Its pdf is symmetric about μ.
- $\mathrm{E}(Y) = \text{Median}(Y) = \mu$, $\mathrm{Var}(Y) = \frac{\pi^2}{3}\sigma^2$.



Logistic $(\mu = 0, \sigma = 1)$ is also called the standard logistic distribution.

A random variable $W \sim$ Logistic $(0, 1)$ has cdf $G(w) = \frac{e^y}{1+e^y}$, for $-\infty < w < \infty$.

Examples of Log Location-Scale Models

Consider $Y = \log T$ that has a Logistic (μ, σ) distribution.
Check for (4), it is easy to see Y has a location-scale distribution, and T has a log location-scale distribution.
T is said to have a log logistic distribution, LogLogistic (μ, σ).

Exercise 1.3.4 An alternative parameterization

Let T be a failure time from LogLogistic (μ, σ) distribution. Show that with some re-parameterization, the hazard function and sf of T can be expressed in the following forms

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}, \text{ and } S(t) = \frac{1}{1+\lambda t^{\alpha}}.$$

Examples of Log Location-Scale Models

If log failure time $Y = \log T$ has a location-scale distribution, the failure time T correspondingly has a log location-scale distribution.

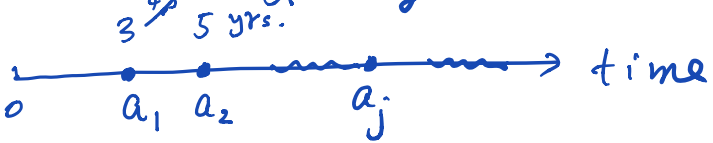
The connections are summarized for the following distributions:

T	$\log T$
log normal	normal
Weibull	EV
log logistic	logistic

In a survival model, we will either study T (and its distribution), or $\log T$ (and its distribution).

1.3.3 Discrete time distributions

Assume the failure time T is a discrete random variable, taking values $a_1, a_2, \ldots$, only. Typically $a_1 < a_2 < \cdots a_j < \cdots$



(This is a way to model failure time, useful for non-parametric analysis.)

For the discrete failure time, its distribution is described by the probability mass function (pmf) $f(t) = P(T = t)$,

$$= \begin{cases} 0, & \text{otherwise.} \\ P(T = a_j), & j = 1, 2, \cdots \end{cases}$$

and survival function $S(t) = P(T \geq t) = \sum_{k:a_k \geq t} f(a_k)$,

and the hazard function $h(t) = \dfrac{P(T=t)}{P(T \geq t)}$. $= P(T=t \mid T \geq t)$

At $t = a_j$,

$$h(t) = \frac{P(T=a_j)}{P(T \geq a_j)} = \frac{f(a_j)}{S(a_j)}$$

$= P(T=a_j \mid T \geq a_j)$.

At $t \neq a_1, a_2, \cdots$, $h(t) = 0$.

In fact, the hazard function

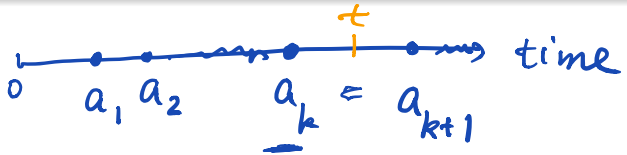
$$h(t) = \begin{cases} \frac{f(a_j)}{S(a_j)}, & \text{at } t = a_j,\ j = 1, 2, \ldots, \\ 0, & \text{if } t \neq a_1, a_2, \ldots \end{cases}$$

Theorem 1.3.2 For a discrete failure time T, show that its survival function can be expressed by

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

in terms of the hazard function.

Pf: $S(t) = P(T \geq t)$



$= P(T \geq a_1, T \geq a_2, \cdots, T \geq a_k, T \geq t)$ where $a_k \leq t < a_{k+1}$

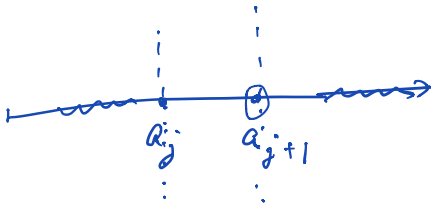
$= P(T \geq a_1) P(T \geq a_2 | T \geq a_1) P(T \geq a | T \geq a_2) \cdots P(T \geq a_k | T \geq a_{k-1}) P(T \geq t | T \geq a_k)$

$P(T \geq a_1, T \geq a_2)$

$P(T \geq a_2)$

$= 1 \cdot [1 - h(a_1)][1 - h(a_2)] \cdots [1 - h(a_{k-1})][1 - h(a_k)]$

$= \prod_{j: a_j < t} [1 - h(a_j)]$. □



[Note that

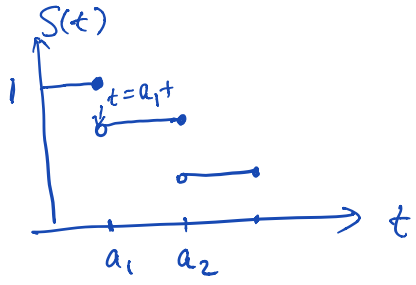
$$P(T \geq a_{j+1} \mid T \geq a_j).$$

$$= 1 - P(T = a_j \mid T \geq a_j)$$

$$= 1 - h(a_j) \quad].$$

Comment: What if $t = a_j$? Say $t = a_1$.

Convention: Assume $S(t)$ is left-continuous fctn.



$$S(\underbrace{a_1}_{t}) = 1, \qquad S(\underbrace{a_1 t}_{t}) = 1 - h(a_1)$$

1.4 Censoring, Survival Data and Likelihood Function

1.4.1 Censoring of survival time

Censoring is a special characteristic of survival (or time-to-event) data.

It refers to unobserved response times, that can occur in a study, for some subjects.

Common types of censoring:

a. Right censoring

- Due to loss to follow-up.
  Examples: Migration; end of study

- For these individuals, we only know that $T \geq t^*$.

b. Left censoring

- Only know that $T \leq t^*$.

Often occurs when recalling time of an event.
When did you drink the 1st time?
≤ 18 yrs.

c. Interval censoring

- Only know that $t_L \leq T \leq t_U$.
- Often occurs in studies with periodic follow-up.
- Example: In a study of a new drug for reducing cholesterol levels, $T =$ time that cholesterol level drops back to normal range, measured every 3 months.

1.4.2 Censoring and likelihood function

- Data with NO censoring

Denote the failure times (random variables) for subjects $1,\ldots,n$ by $T_1,\ldots,T_n$, and the observed data by $D=(t_1,\ldots,t_n)$.

Model assumption: $T_1,\ldots,T_n$ has joint pmf / pdf $f(t_1,\ldots,t_n;\theta)$.
Parameters of interest: θ (a vector in $\mathbb{R}^p$).

Likelihood function for observing D: $L(\theta)=f(t_1,\ldots,t_n;\theta)$,

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

if T_i's are i.i.d., each has a pmf / pdf $f(t_i;\theta)$, $i=1,\ldots,n$.

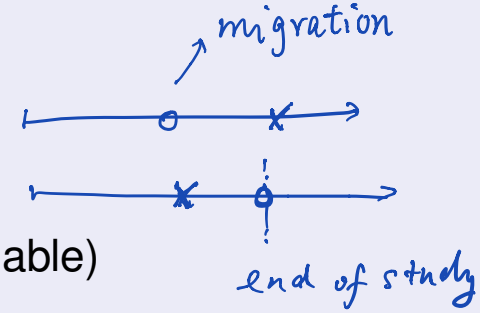
- Right-censored data
  - n individuals or subjects under study, $i = 1, \ldots, n$.
  - Each has a response time, some are exact failure time, some are censoring time.

Notation

T_i: (potential) failure time for subject i;
C_i: (potential) censoring time for subject i;

Denote the observed failure time of subject i by

$$X_i = \min(T_i, C_i). \quad \text{(a random variable)}$$



Define a censoring indicator for subject i

$$\Delta_i = I(T_i \le C_i) = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } T_i > C_i \end{cases}$$

The observed version of these random variables are t_i, c_i, x_i, δ_i.

The observed data is denoted by $D = \{(x_1, \delta_1), \ldots, (x_n, \delta_n)\}$

(X, Δ)

Data: $(10, 1)$: $T = 10$.

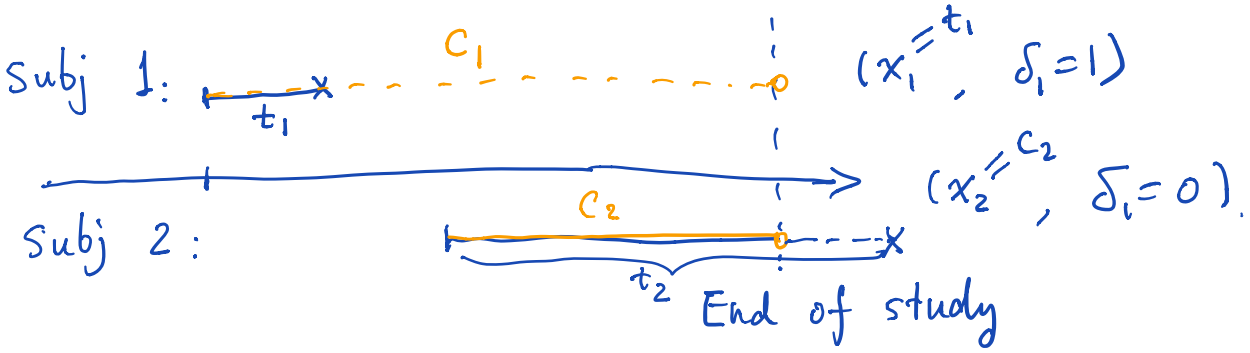
$(7, 0)$: $C = 7$, $T > 7$

Some censoring mechanisms in practical studies

i. Type I censoring:

Censoring times $C_1, \ldots, C_n$ are known constants.

Example: A study that has a pre-determined end-of-study time.



ii. Type II censoring:
Study continues until the first r failures are observed.

pre-specified.

Typical assumptions on censoring for survival models

1) Random censoring assumption:
$C_1, \ldots, C_n$ are independent random variables,
and independent of $T_1, \ldots, T_n$.

2) Independent censoring assumption:
T_i and C_i are independent given the covariates Z_i, $i = 1, \ldots, n$.

Note:
For 1) and 2), we must further assume that $C_1, \ldots, C_n$ are ancillary!
(no information about θ)

Remarks:

- Assumption 2) is more general than 1).
  See discussion on censoring mechanisms and assumptions in Kalbfleisch and Prentice, 2002, Section 1.3.
- Most models in survival analysis are based on Assumption 2).
- Many practical studies have Type I censoring mechanism.

Construction of likelihood for right-censored data

Denote the pdf and sf of T_i by $f(t;\theta)$ and $S(t;\theta)$.
Denote the pdf and sf of C_i by $g(c)$ and $U(c)$.

The likelihood contribution by subject i is

- $$P[T_i \in [x_i, x_i+\Delta x_i), C_i > x_i] = f(x_i;\theta)\Delta x_i \, U(x_i)$$

  if a failure time is observed $(\delta_i = 1)$;

- $$P[C_i \in [x_i, x_i+\Delta x_i), T_i > x_i] = g(c)\Delta x_i \, S(x_i;\theta)$$

  if a censoring time is observed $(\delta_i = 0)$.

  By previous assumptions on censoring, $g(x_i)$ and $U(x_i)$ do NOT depend on θ.

Likelihood function for data D (x_i, δ_i), $i = 1, \cdots, n$.

$$L(\theta) \propto \prod_{i:\text{ failure}} f(x_i;\theta) \prod_{i:\text{ censoring}} S(x_i;\theta)$$

or simply write

$$L(\theta) = \prod_{i=1}^{n} [f(x_i;\theta)]^{\delta_i} [S(x_i;\theta)]^{1-\delta_i}. \quad (5)$$

The likelihood (5) applies generally to right-censored data.

Example 1.4.1 Equipment Field Failure

See data given in Example 1.2.1.

Model the data by an exponential distribution, $T \sim \text{Exp}(\theta)$.
Pdf of T: $f(t;\theta) = \frac{1}{\theta} e^{-t/\theta}$.
Parameter: θ, which is the mean $\text{E}(T)$.

$S(t;\theta) = e^{-t/\theta}$.
$P(T \geq t) = \cdots$

$$L(\theta) = \prod_{i=1}^{n} [f(x_i;\theta)]^{\delta_i} [S(x_i;\theta)]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left[\frac{1}{\theta} e^{-x_i/\theta}\right]^{\delta_i} \left[e^{-\frac{x_i}{\theta}}\right]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left[\frac{1}{\theta}\right]^{\delta_i} e^{-\frac{x_i}{\theta}}$$
$$= \left[\frac{1}{\theta}\right]^{\sum \delta_i} e^{-\frac{\sum x_i}{\theta}}$$

$\prod_{i=1}^{n} \left(\frac{1}{\theta}\right)^{\delta_i}$
$\left[\frac{1}{\theta}\right]^{\delta_1+\delta_2+\cdots\delta_n}$.

log-likelihood: $\ell(\theta) = \log L(\theta)$

$= -\sum_{i=1}^{n} \delta_i \log \theta - \frac{\sum x_i}{\theta}$.

Score fct'n: $\ell'(\theta) = -\sum_{i=1}^{n} \delta_i \cdot \frac{1}{\theta} + \frac{\sum x_i}{\theta^2}$

Solve $\ell'(\theta) = 0$ (score equation) for θ.

$\sum \delta_i \frac{1}{\theta} = \frac{\sum x_i}{\theta^2} \Rightarrow \hat{\theta} = \frac{\sum x_i}{\sum \delta_i}$

MLE

Total # failures observed.

Hazard rate: $\lambda = \frac{1}{\theta}$. What's the mle of λ?

mle of λ is $\hat{\lambda} = \frac{1}{\hat{\theta}}$.

Invariance property of mle. Important and useful.

If $\hat{\theta}$ is the mle of θ (which can be a vector), and $\lambda = g(\theta)$ is a one-to-one transformation of θ, then the mle for λ is given by $\hat{\lambda} = g(\hat{\theta})$.