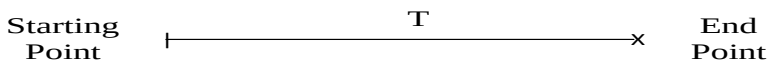
Survival Analysis

Biswabrata Pradhan
SQC & OR Unit
Indian Statistical Institute, Kolkata

bis@isical.ac.in

Introduction

- Survival analysis deals with time to event data.
- Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.
- Event: death, disease incidence, disease remission, relapse from remission or any designated experience of interest that may happen to an individual.
- By time, we mean years, months, weeks or days from the beginning of follow-up of an individual until an event occurs. Alternatively, time can refer to the age of an individual when an event occurs.
- End point: The point where event of interest occurs.



Introduction

- In considering the failure time data, it is important to have unambiguous definition of the time origin from which survival is measured.
- The natural time origin may be the occurrence of some event, such as randomization or entry into a study or diagnosis of a particular disease.
- Example: Laboratory animals are subject to doses of the carcinogenic substances and then observed to see if they develop tumors. Variable of interest is the time to appearance of a tumor, measured from when the dose is administered.

  - Here time origin is when the dose is administered.

- Date of joining a service.
- Insurance starting date.
- Date of subscribing a service.

Lifetime/Survival time/Failure time

- Leukemia patients: the event of interest is "going out of remission". The survival time/lifetime is "time in weeks (say) until a person goes out of remission".
- Disease-free cohort of individuals over several years to see who develops heart disease.
- Post-surgery survival time of heart patients.
- Duration of marriage/friendship etc.
- Customer churn: Time to churn (Duration: a customer stay with a particular company, service provider etc.).

  If the customer churned, lifetime is the number of days (or weeks, months, whatever) between the day they subscribed and the day they unsubscribed.

Lifetime/Survival time/Failure time

- Credit Risk Modeling: Time to default
  - Calculate the probability of default at different time point .
  - Credit risk models are used by financial companies to evaluate in advance the insolvency risk caused by credits that enter into default.
  - Traditional credit risk models aim at determining the probability of default on loan repayment.
  - Survival models is used to etsimate the probability of default by a certain time.
- Example: A bank granted loans between January and February 2009 with a short loan term of 12 months. A customer was considered defaulter if his had a period of 90 days without loan repayment.
- Insurance Attrition/Retention: Time of termination.

Lifetime

- Survival time/Failure time/Lifetime: A non-negative random variable represents the length of time between a suitable starting point and an end point.
- T: Survival time. It is a nonnegative random variable.
- The basic quantity employed to describe time-to-event phenomena is the **survival function**, the probability of an individual surviving beyond time t. It is defined as
$$S(t) = P[T \geq t].$$
- $S(t+) = P[T > t]$. $S(t+) = S(t)$, if T is continuous.
- In some contexts involving systems or lifetimes of manufactured products, $S(t)$ is referred to as the reliability function. Normally, it is denoted by $R(t)$.

Continuous Models

- T is continuous lifetime random variable defined over the interval $[0, \infty)$.
- $f(t)$: pdf of T; $F(t)$: cdf of T; $S(t) = 1 - F(t)$; $S(t) = \int_t^\infty f(u)du$.
- $F(t) = P[T \leq t]$. $F(0) = 0$, $F(\infty) = 1$, F is non-dcreasing and right continuous.
- Properties of $S(t)$
  (i) $S(0) = 1$

  (ii) $\lim_{t \to \infty} S(t) = 0$

  (iii) $S(t)$ is non-increasing continuous function in t.

  (iv) $S(t)$ is left continuous.
- Sometimes we wish to allow $S(\infty) > 0$, to consider settings where some individuals never fail.

Continuous Models: Hazard Rate

- **Hazard rate or Failure rate**: The hazard rate is defined by

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t}.$$

- $\lambda(t) = \frac{f(t)}{S(t)}$
- Hazard rate ranges between 0 and ∞.
- The hazard rate ($\lambda(t)$) specifies the instantaneous rate of death or failure (or occurring any event) at time t, given that the individual survives up to time t.
- Note: $\lambda(t)\Delta t$ is the approximate probability of death (of occurring an event) in $[t, t + \Delta t]$, given survival up to t.
- This is also known as force of mortality in demography.

Hazard Rate

- Hazard rate indicates the way the risk of failure varies with age or time.
- Prior information about the shape of the hazard function can help guide model selection.
- Hazard rate may be increasing, decreasing, constant, bathtub-shaped or of some other characteristics which describes the failure mechanism.
- Constant hazard: It occurs in stable settings where failure or death is due to random phenomenon such as shocks or accidents, which are external to the individual.
- Increasing hazard rate:
  - Due to natural aging or wear.
  - Patients (say, leukemia) not responding to treatment, where event of interest is death.

Hazard Rate

- Decreasing hazard
  - Infant mortality (due to infant disease)
  - Post surgery failure rate- event is death in persons who are recovering from surgery, because the potential for dying after surgery ususally decreases as the time after surgery increases.
- First increasing and then decreasing:
  - Such graph is expected for tuberculosis patients, since there potential for dying increases early in the disease and decreses later.
  - In connection with the duration of marriage.
- Bathtub-shaped hazard:
  - It is appropriate in populations followed from birth. During an early period, death result primarily from infant disease, after which the death rate stabilizes, followed by an increasing hazard rate due to natural aging process.

Relationship of $S(t)$ and $\lambda(t)$

- The functions, $f(t)$, $F(t)$, $S(t)$ and $\lambda(t)$ are mathematically equivalent specifications of the distribution of T.
- $\lambda(t) = \frac{f(t)}{S(t)}$
- $S(t) = \exp\left[-\int_0^t \lambda(u)du\right] = \exp\left[-\Lambda(t)\right]$, where $\Lambda(t)$ is known as cumulative hazard.
- $f(t) = \lambda(t)\exp\left[-\int_0^t \lambda(u)du\right]$.
- Example: If $\lambda(t) = \lambda$, then $f(t) = \lambda e^{-\lambda t}$: Exponential distribution.

Continuous Lifetime Distribution

- Exponential: $f(t) = \lambda e^{-\lambda t},\ t > 0, \lambda > 0.$

  $S(t) = e^{-\lambda t}$, $\lambda(t) = \lambda$, independent of time t (Memoryless property).

- Weibull: $f(t) = \alpha\lambda^{\alpha} t^{\alpha-1} e^{-(\lambda t)^{\alpha}},\ t > 0, \alpha > 0, \lambda > 0.$
  - $S(t) = e^{-(\lambda t)^{\alpha}}$
  - $\lambda(t) = \alpha\lambda^{\alpha} t^{\alpha-1}$. This is increasing in t, if $\alpha > 1$. Decreasing in t, if $\alpha < 1$. It reduces to exponential model, if $\alpha = 1$.

- Gamma distribution: $\frac{\lambda^n}{\Gamma(n)} t^{n-1} e^{-\lambda t},\ t > 0, \lambda > 0.$

  - Hazard is increasing in t for $n > 1$; decreasing for $n < 1$. Reduces to exponential model for $n = 1$.

- Log-normal distribution: $f(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}$

  - Hazard rate first increasing, then decreasing.

Discrete Models

- T takes on values $t_1, t_2, \ldots$, with $0 \leq t_1 < t_2 < \cdots$. The corresponding masses $p_1, p_2, \ldots,$.

$$p_j = P[T = t_j], \quad j = 1, 2, \ldots$$

- The survival function is

$$S(t_j) = P[T \geq t_j] = \sum_{l:t_l \geq t_j} p_l.$$

In general $S(t) = \sum_{j:t_j \geq t} p_j$.

- $S(t)$ is a left-continuous, non-increasing step function, with $S(0) = 1$ and $S(\infty) = 1$.

Discrete Models

- The discrete time hazard at t_j is

$$\lambda_j = \lambda(t_j) = P[T = t_j | T \geq t_j] = \frac{p_j}{S(t_j)}$$

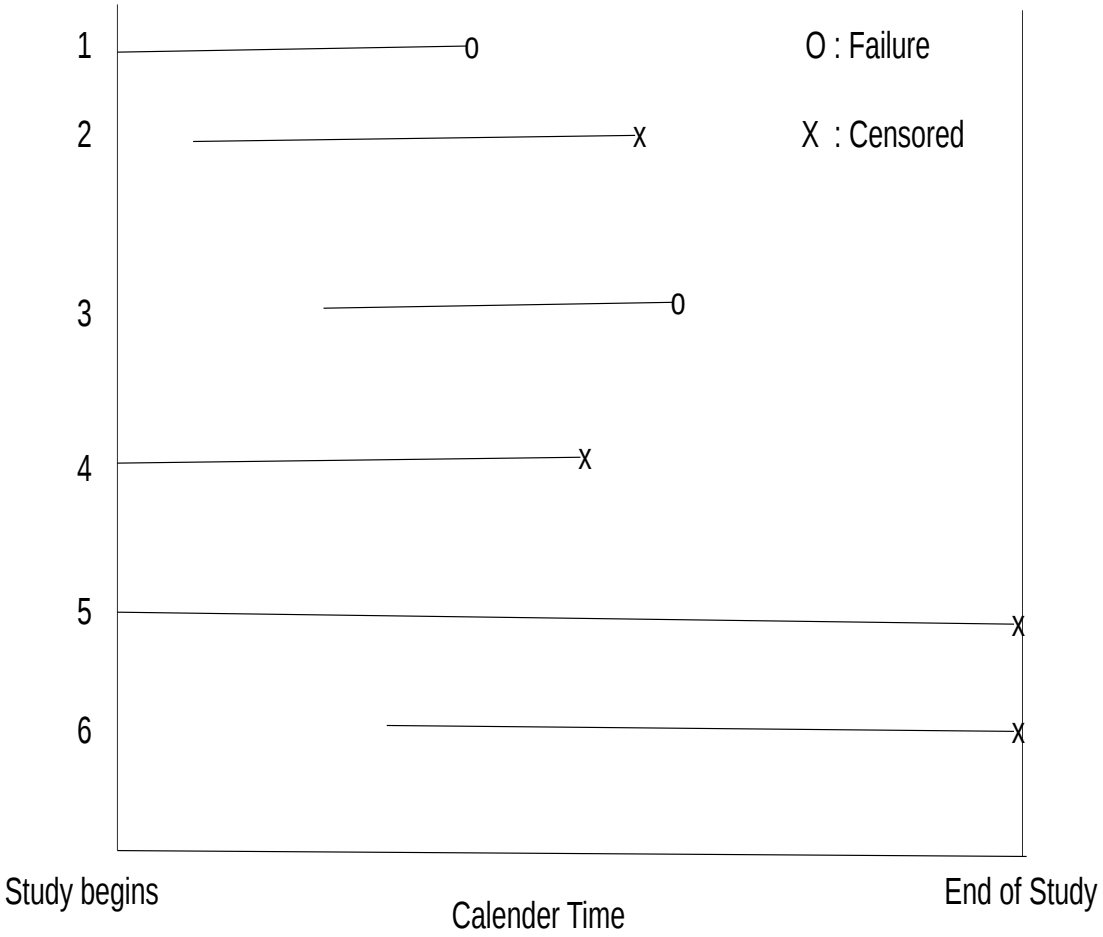
- Note that the hazard is zero at any time other than the mass point.

- In this case, the hazard is a conditional probability lying between 0 and 1, whereas it is like a conditional density in the continuous case.

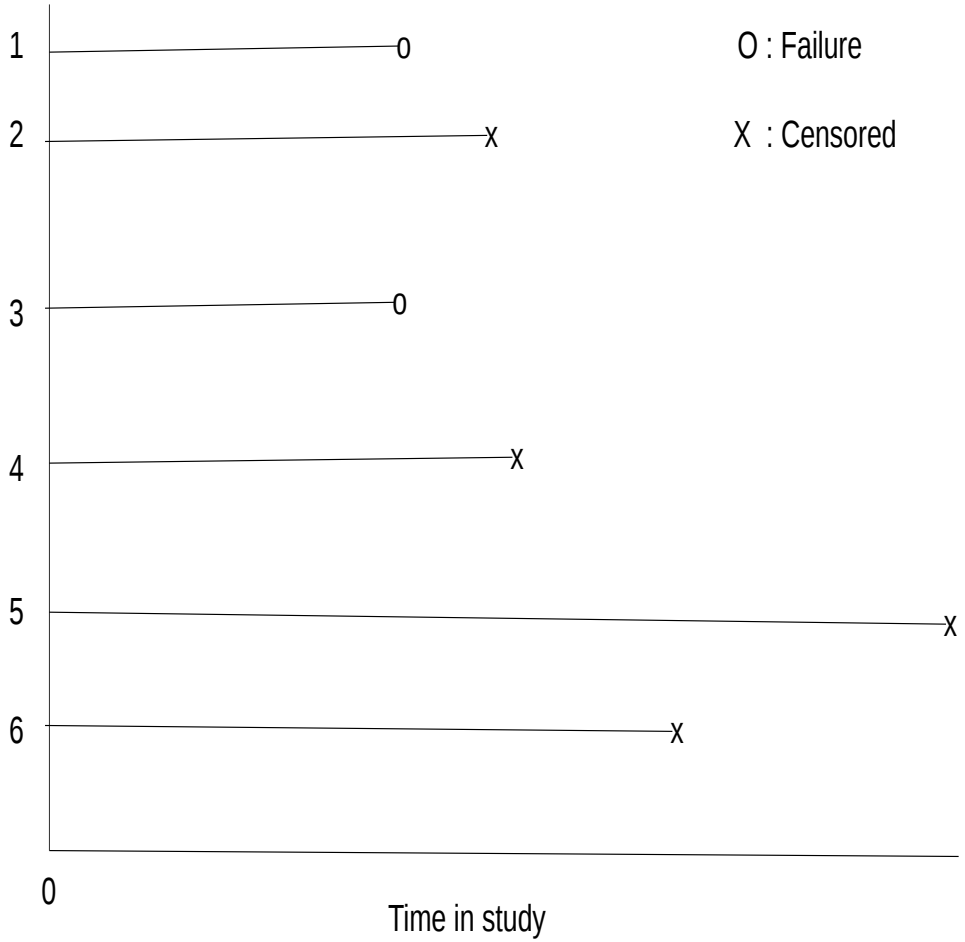
- As in the continuous case, the probability, survival functiona nd hazard functions give equivalent specifications of the distribution T.

  * $\lambda(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)}, \quad j = 1, 2, \ldots$
  * $S(t_j) = \prod_{l=1}^{j-1}(1 - \lambda_l)$
  * In general $S(t) = \prod_{j:t_j<t}(1 - \lambda_j)$ for all $t > 0$.

Survival Data



Survival Data



Features of Survival Data

- Features which are typically encountered in analysis of survival data:
- Individuals do not all enter the study at the same time. This is known as staggered entry.
- When the study ends, some individuals still haven't had the event yet
- Some individuals get lost in the middle of the study, and all we know about them is the last time they were still free of the event
- These features relate to "censoring" of the failure time events.
- Population may be heterogeneous.

Covariates/Explanatory variables

- Analysis of survival data in presence of heterogeneity in a population is an important issue.
- The use of explanatory variables or covariates in a regression analysis model is an important way to represent heterogeneity in a population.
- Covariates: Age, gender, socioeconomic status, dietary habits, smoking history, alcohol consumption, blood pressure, blood glucose level, hemoglobin level etc.
- The main object in many studies is to understand and exploit the relationship between lifetime and covariates.
- For lung cancer patients, age, type of tumor, smoking history etc. can be considered as covariates.
- In clinical trial, the treatment assigned to a patient may be considered as a covariate.

Objectives

- To investigate different characteristics of the lifetime (e.g. mean, median, quantiles, variance, survival function etc.)
- Two compare survival of two or more groups.
  Example: Treatment and Placebo groups.
- To investigate the effect of certain explanatory variables (known as covariates) on survival experience.

Data

- **Complete data**: Lifetimes are completely observed and recorded on n individuals.
- **Censored data**: A lifetime is censored if we do not observe it exactly but only know that it lies within a certain interval.
- Type of censoring: Right censoring, left censoring and interval censoring.
- **Right censoring**: Type-I censoring, Type-II censoring and Random Censoring
- **Type-I censoring**: n individuals are on a study upto a pre-specified time T_0. The event of interest is observed only if it occurs before T_0.
- **Type-II censoring**: The study continues until the failure of the first r individuals, where r is some predetermined integer ($r < n$).

Random Censoring

- Random censoring: Each individual is assumed to have a lifetime T and a censoring time C. One observe T or C whichever is earlier.
- Observation (X, δ), where $X = \min(T, C)$ and $\delta = I(T \leq C)$.
- $\delta = 1$ if lifetime is observed, that is the event of interest is observed.
- $\delta = 0$ if censored.
- For n individuals: $(X_1, \delta_1), \ldots, (X_n, \delta_n)$.
- Random censoring occurs due to: Loss to follow up, drop-out or study termination.
- Independent censoring: censoring should not convey any information about the future failure time.

  -T and C are independently distributed.

Left Censoring and Interval Censoring

- **Left-censoring**: It occurs when it is known that the event of interest occurred prior to a certain time t, but the exact time of occurrence is unknown.

  Example: We want to know the age at diagnosis in a follow-up study of diabetic retinopathy. At the time of examination, a 50-year old participant was found to have already developed retinopathy, but the exact age was not known. Thus the age at examination (i.e. 50) is a left-censored observation. So the age of diagnosis for this patient is at most 50 years.

- We observe $X_i = \max(T_i; U_i)$ and $\delta_i = 1$ if $T_i \geq U_i$; 0 if $T_i < U_i$

- **Interval-censoring**: This occurs when the event of interest is known to have occurred between times L and U. Observe (L_i, U_i) where $T_i \in (L_i, U_i)$.

  Example: Consider the retinopathy example. If medical records indicate that at age 45, the patient did not have retinopathy, his age at diagnosis is between 45 and 50 years.

Some Examples

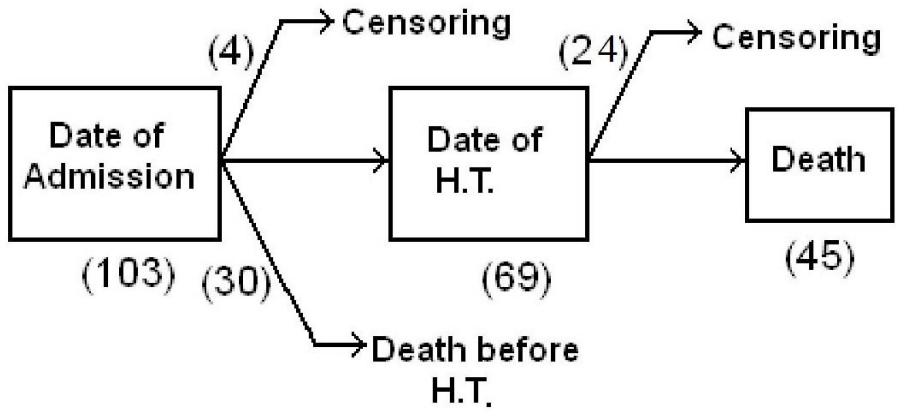
- Time to first use of Marijuana:

  In this study, 199 school boys were asked, "when did you first use marijuana?". The answers were the exact ages (uncensored observations); "I never used it", which are right-censored observations at the boys' current ages or "I have used it but cannot recall just when the first time was", which is a left-censored observations. Notice that a left-censored observation tells us only that the event has occurred prior to the boys' current age.

- In a study of age at which African children learn a task. Some already knew (left-censored), some learned during study (exact), some had not yet learned by end of study (right-censored).

Example: Stanford Heart Transplant data

- The patients with heart problem were admitted to the Stanford Program (Crowley and Hu, 1977, JASA). The observations started at that point of time.



- Covariates: age, previous history of surgery, waiting time for transplantation and mismatch score. One may be interested the effect of these covariates on post transplant survival time.

Truncated Data

- Truncation: A lifetime T is observed only when T belongs to a particular set $A = [a, b]$.

- An individual whose event time is not in this interval is not observed and no information is available to the investigator.

- When b is infinite, we have left truncation. We only observe those individuals whose event time $T > a$.

- Left truncation occurs when subjects enter the study at a particular age (not necessarily the origin for the event of interest).

  Consider a survival study of residents of a retirement center. Ages at death are recorded, as well as ages at which individuals entered the retirement community (truncated event). Since an individual must survive to a sufficient age to enter the retirement center, all individuals who died earlier will not enter the study.

Truncated Data Contd..

- Right truncation occur when $a = 0$. We observe the survival time only when $T \leq b$.

- Right truncation occurs when only individuals who have experienced the event of the interest are observable.

  A group of 258 patients (Lagakos et al., 1988) with AIDS had been exposed to HIV through blood transfusion on a known date (April 1, 1978). Patients who had not developed AIDS before the end of the study (June 30, 1986) are not included.

- In case of truncated data, we have to use conditional distribution for constructing the likelihood function.

- In censoring, there is at least partial information on each subject is available.

Remission Data of leukemia patients

- The remission times of 42 patients with acute leukemia were reported by Freiereich et al. (1963, Blood, 21(6)) in a clinical trail undertaken to assess the ability of 6-mercaptopurine (6-MP) to maintain Steriod-induced remission.

- Each patient was randomized to receive 6-MP or placebo

- The study was terminated after one year. The remission times

  6-MP (21 patients) : 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+.

  - **Placebo (21 patients)**: 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

- One may be interested to know the **relapsed rate** and the probability of having a remission time longer than 10 weeks in each group.

Estimation

- Parametrically: if we assume a distribution for T with pdf $f(t:\theta)$, then estimates of $\lambda(t)$, $S(t)$ follows from the estimate of θ.
- We find the maximum likelihood estimate of θ.
- Suppose we have random censored data $(x_1,\delta_1),\ldots,(x_n,\delta_n)$. Under the assumption of independent random censoring, the likelihood function becomes

$$L \propto \prod_{i=1}^{n} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}$$

- Example: Exponential distribution: $f(t;\lambda) = \lambda e^{-\lambda t}$.

  - Estimate of λ: $\hat{\lambda} = \frac{\sum \delta_i}{\sum x_i}$, $\sum \delta_i =$ number of failures.

  - Then estimate of $S(t)$ is $\hat{S}(t) = e^{-\hat{\lambda} t}$.

Estimation Contd.

- Consider the leukemia data. Assume that exponential distribution fits the data. Then

  6-MP: $\hat{\lambda} = \frac{9}{359} = 0.025$ per week. $\hat{\mu}$ (mean) = $\frac{1}{0.025}$ = 40 weeks.

  The probability of staying in remission for one year (or 52 weeks) or more is estimated by

$$\hat{S}(52) = \exp[-0.025 \times 52] = 0.273.$$

- Placebo: $\hat{\lambda} = \frac{21}{182} = 0.115$ per week. $\hat{\mu} = 8.7$ weeks.

$$\hat{S}(52) = \exp[-0.115 \times 52] = 0.003.$$

- What if the assumption is wrong?
- Non-parametric estimation means we do not specify a distribution for T.
- Non-parametric estimator for right-censored data: Kaplan-Meier estimator.

Kaplan-Meier Estimator

- Right censored data: $(x_1, \delta_1), \ldots, (x_n, \delta_n)$.
- Suppose $t_1 < t_2 < \ldots < t_k$ are ordered observed failure times
- d_i = number of failures at t_i.
- m_i = number of censoring in $[t_i, t_{i+1})$, $i = 0, 1, \ldots, k$, where $t_0 = 0$ and $t_{k+1} = \infty$.
- n_i = number of individuals at risk just prior to time t_i.
- Kaplan-Meier estimator of survival function $S(t)$ is given by

$$\hat{S}_{KM}(t) = \prod_{i:t_i<t} (1 - \frac{d_i}{n_i})$$

- $$Var(\hat{S}_{KM}(t)) = \{\hat{S}_{KM}(t)\}^2 \sum_{i:t_i<t} \frac{d_i}{n_i(n_i - d_i)}$$

Kaplan-Meier Estimate: Leukemia Data

6-MP (21 patients) : 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+.

R-Code

```
library(survival)
treatment<- c(6, 6, 6, 7, 10, 13, 16, 22, 23, 6, 9, 10, 11, 17, 19, 20, 25, 32, 32, 34, 35)
cens1<- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
surv.treatment <- survfit(Surv(treatment, cens1) 1, conf.type= "none")
summary(surv.treatment)
plot(surv.treatment, xlab="Time", ylab="Survival Probability")
```

Kaplan-Meier Estimate: Leukemia Data

6-MP (21 patients) : 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+.

R-Code

OUTPUT:

summary(surv.treatment)

Call: survfit(formula = Surv(treatment, cens1) 1, conf.type = "none")

time	n.risk	n.event	survival	std.err
6	21	3	0.857	0.0764
7	17	1	0.807	0.0869
10	15	1	0.753	0.0963
13	12	1	0.690	0.1068
16	11	1	0.627	0.1141
22	7	1	0.538	0.1282
23	6	1	0.448	0.1346

Note: R gives estimate of survival probabilty $S(t+)$.

Nonparametric Estimator of Cumulative Hazard

- Cumulative hazard: $\Lambda(t) = \int_0^t \lambda(u)du$.
- Nonparametric estimator of $\Lambda(t)$: $\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}$
- Plot $\hat{\Lambda}(t)$ give useful information about the shape of the hazard function. For example, $\Lambda(t)$ is linear if $\lambda(t)$ is constant.

Model Validation

- **Graphical Methods:**

  - PP Plot

  -QQ Plot

  - Hazard Plotting

- Goodness-of-fit Tests.

Model validation: Graphical Method

- $F(t;\theta)$: CDF
- P-P (Probability-Probability) Plot
- **Complete Data**
  - Lifetimes of n items: $x_1, \ldots, x_n$.
  - Corresponding ordered observations: $x_{(1)}, \ldots, x_{(n)}$.
  - Plot $F(x_{(i)};\hat{\theta})$ vs. $\frac{i-0.5}{n}$, $i = 1, 2, \ldots, n$. $\hat{\theta}$ = MLE of θ. $\frac{i-0.5}{n}$ is called plotting position.
  - If the parametric model is appropriate the points should lie around a straight line.
  - Exponential distribution: Plot $1 - e^{-\hat{\lambda}x_{(i)}}$ vs. $\frac{i-0.5}{n}$.
  - Weibull distribution: $1 - e^{-(\hat{\lambda}x_{(i)})^{\hat{k}}}$ vs. $\frac{i-0.5}{n}$.

P-P Plot: Right Censored Data

- Observations: $(x_1, \delta_1), \ldots, (x_n, \delta_n)$
- Let $t_1 < t_2 \cdots < t_k$ be the distinct failure times.
- Plot $F(t_i'; \hat{\theta})$ vs. $1 - \hat{S}_{\text{KM}}(t_i+)$, $i = 1, 2, \ldots, k$. $\hat{S}_{\text{KM}}(t)$ is the Kaplan-Meier estimator of $S(t)$.
- If the parametric model is appropriate the points should lie around a straight line.
- Exponential distribution: Plot $1 - e^{-\hat{\lambda} t_i}$ vs. $1 - \hat{S}_{\text{KM}}(t_i+)$
- Weibull distribution: $1 - e^{-(\hat{\lambda} t_i)^{\hat{k}}}$ vs. $1 - \hat{S}_{\text{KM}}(t_i+)$

Q-Q Plot: Complete Data

- Lifetimes of n items: $x_1, \ldots, x_n$.
- Corresponding ordered observations: $x_{(1)}, \ldots, x_{(n)}$.
- Plot $F^{-1}(\frac{i-0.5}{n}; \hat{\theta})$ vs. $x_{(i)}$, $i = 1, 2, \ldots, n$. $\hat{\theta}$ = MLE of θ. $\frac{i-0.5}{n}$ is called plotting position.
- If the parametric model is appropriate the points should lie around a straight line.
- Exponential distribution: Plot $-\frac{1}{\hat{\lambda}} \ln(1 - \frac{i-0.5}{n})$ vs. $x_{(i)}$.
- Weibull distribution:

Q-Q Plot: Censored Data

- Observations: $(x_1, \delta_1), \ldots, (x_n, \delta_n)$
- Let $t_1 < t_2 \cdots < t_k$ be the distinct failure times.
- Plot $F^{-1}(1 - \hat{S}_{\text{KM}}(t_i+); \hat{\theta})$ vs. t_i, $i = 1, 2, \ldots, n$.
- Exponential distribution: Plot $-\frac{1}{\hat{\lambda}} \ln(\hat{S}_{\text{KM}}(t_i+))$ vs. t_i.
- Weibull distribution:

Two-sample /K-sample problem

- The problem of comparing survival experience of two or more groups is an important issue in biomedical studies.
- For example

  - A diabetologist may wish to compare the retinopathy-free time of two groups of diabetic patients.

  - A clinical oncologist may be interested in comparing the ability of two or more treatments to prolong life or maintain health.

- These differences can be illustrated by drawing graphs of the estimated survivorship functions, but that gives only a rough idea of the difference between the distributions.
- It does not reveal whether the differences are significant or merely chance variations. So a statistical test is necessary.

Two-sample Problem: Log-Rank Test

- Test whether the survival functions of two groups (say treatments) are identical.
- To test whether two samples could have arisen from identical survival function. $H_0 : S_1(t) = S_2(t)$ against $H_0 : S_1(t) \neq S_2(t)$
- We observe the data as follows

  Group 1: $\{(x_{1j}, \delta_{1j}) |\ j = 1, 2, \cdots, n_1\}$

  Group 2: $\{(x_{2j}, \delta_{2j}) |\ j = 1, 2, \cdots, n_2\}$

- -t_j = Time of the jth failure time (across group)
  - d_{1j} = Number of failures for group 1 at time t_j
  - d_{2j} = Number of failures for group 2 at time t_j
  - n_{1j} = Number risk for group 1 prior to time t_j
  - n_{2j} = Number risk for group 2 prior to time t_j

Two-sample Problem: Log-Rank Test

- At j^{th} death time we have the following table.

Group	Number of Deaths	Number of Survivor	Total
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{1j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

- Construct similar table corresponding to other death/failure times.

Two-sample Problem: Log-Rank Test

- Calculate the following:
  $d_j = d_{1j} + d_{2j}$, $n_j = n_{1j} + n_{2j}$, $e_{1j} = \frac{n_{1j} d_{1j}}{n_j}$
  $O_1 - E_1 = \sum (d_{1j} - e_{1j})$
  $V_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$
  $V_1 = \sum_{j=1}^{k} V_{1j}$

- Test Statistic: $V_0 = (O_1 - E_1)^2 / V_1 \sim \chi_1^2$.

- The log-rank is sometimes called the Cox-Mantel test.

- Observed value of the log-rank statics for the leukemia example is

$$\chi^2_{logrank} = \frac{(10.251)^2}{6.257} = 16.793 \qquad p\text{-value} = 0.00004$$

Log-Rank Test: R-Code

```
library(survival)
treatment<- c(6, 6, 6, 7, 10, 13, 16, 22, 23, 6, 9, 10, 11, 17, 19, 20, 25, 32,
32, 34, 35)
> cens1<- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
> placebo<- c(1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23)
> cens2 <- rep(1, 21)
> time <- c(treatment, placebo)
> status <- c(cens1, cens2)
> combined <- c(rep(1, 21), rep(2, 21))
> fit <- survdiff(formula = Surv(time, status) ~ combined)
> fit Call: survdiff(formula = Surv(time, status) ~ combined)
Chisq= 16.8 on 1 degrees of freedom, p= 0.00004
```

Regression Models for Survival Data

- A problem frequently encountered in analyzing survival data is that of adjusting the survival function to account for covariates.
- Consider a failure time $T > 0$ and a vector of covariates $Z = (Z_1, \ldots\ldots, Z_p)$ associated with the failure time T.
- Z may include quantitative variables, such as blood pressure, temperature, age, weight etc. It can be qualitative, such as gender, race, treatment etc.
- Z may be time-dependent also: $Z(t) = (Z_1(t), \ldots, Z_p(t))'$. Example: Serial blood pressure.
- The objective is to establish the relationship between the failure time T and one more of the covariates.
- For example: One may be interested to compare the survival functions of two groups using a covariate Z. Define Z =0 for Group 1 and 1 for Group 2.

Regression Models Contd.

- Two approaches to the modeling of covariate effects on survival have become popular.
- The first approach is analogous to the classical linear regression approach. Here the natural logarithm of the survival time is modeled

$$Y = \ln(T) = \mu + \beta' Z + \sigma W,$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients and W is the error random variable.

- This model is called the accelerated failure-time model.
- It can be shown that

$$S(t|Z) = S_0(te^{-\beta' Z}),$$

where S_0 is the survival function of T without covariate effect.

Proportional hazard model

- $\lambda(t|Z) = \lambda_0(t)g(Z)$ with $g(Z) > 0$ and $g(0) = 1$.
- $\lambda_0(t)$ is called baseline hazard function.
- Why proportional?
  Consider two individuals with respective covariates Z and Z^*. Then

$$\frac{\lambda(t|Z)}{\lambda(t|Z^*)} = \frac{g(Z)}{g(Z^*)},$$

  which is independent of t. So the hazard rates are proportional.
- Normally $g(Z)$ is taken as

$$g(Z) = \exp(\beta_1 Z_1 + \cdots + \beta_p Z_p) = \exp(\beta^t Z).$$

- Then, $\lambda(t|Z) = \lambda_0(t)\exp(\beta^t Z)$

Proportional hazard model Contd.

- $\lambda(t|Z) = \lambda_0(t)\exp(\beta^t Z)$
- One can consider any parametric form of $\lambda_0(t)$.
- Example: $\lambda_0(t) = \lambda$, It is exponential regression model.
- Cox Proportional Hazard Model

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta^t Z),$$

where $\lambda_0(t)$ is arbitrary and unspecified.

- β is estimated by partial likelihood method.

- Data: $(x_i, \delta_i, \mathbf{z}_i)$

Proportional hazard model Contd.

- $\frac{\lambda(t|Z)}{\lambda(t|Z^*)}$ is called relative risk (RR) or hazard ratio.

  - Relative risk of an individual with risk factor Z having an event as compared to an individual with risk factor Z^*.

- Example: Suppose Z_1 indicates the treatment effect. $Z_1 = 1$ if treatment and $Z_1 = 0$ if placebo. All other covariates have the same value. Then

$$\frac{\lambda(t|Z)}{\lambda(t|Z^*)} = \exp(\beta_1).$$

- So $\exp(\beta_1)$ is the risk of having the event if the individual received the treatment relative to the risk of having the event should the individual have received the placebo.

Proportional Hazard Model: Example

- Consider the survival time data from 30 patients with AML (Acute myelogenous leukemia). Two covariates age and cellularity status are considered.

$$Z_1 = \begin{cases} 1 & \text{if patient is} \geq 50 \text{ years old} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if cellularity of marrow clot section is 100\%} \\ 0 & \text{otherwise} \end{cases}$$

- Sample Data:

Survival Time	Z_1	Z_2
18	0	0
9	0	1
28+	0	0
31	0	1
39+	0	1
.	.	.

Proportional Hazard Model: Example

- Consider the Cox proportional hazard model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2),$$

where $\lambda_0(t)$ is unspecified.

- Parameters (β's) are estimated by partial likelihood method.

Regression Analysis Result

Covariate	Regression	Standard Error	p Value	exp(coefficient)
Z_1	1.01	0.46	0.0013	2.75
Z_2	0.35	0.44	0.212	1.42

- The positive sign of regression coefficients indicate that the older patients ($\geq$ 50 years) and patients with 100% cellularity of the narrow clot section have a higher risk of dying

Proportional Hazard Model: Example

- Age is significantly related to survival after adjustment for cellularity.
- The estimated risk of dying for patients at least 50 years of age is 2.75 times higher than that for patients younger than 50.
- Patients with 100% cellularity have a 42% higher risk of dying than patients with less than 100 % cellularity.
- The relative risk for a patients who is over 50 years of age and whose cellularity is 100% compared to patients who are younger than 50 and whose cellularity is less than 100% = exp(1.01+0.35)= 3.90
- R-code: coxph(formula = Surv(time, status) $\sim Z_1 + Z_2$)

Cox PH Model for Leukemia data: R-Code

```
library(survival)
treatment<- c(6, 6, 6, 7, 10, 13, 16, 22, 23, 6, 9, 10, 11, 17, 19, 20, 25, 32,
32, 34, 35)
> cens1<- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
> placebo<- c(1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23)
> cens2 <- rep(1, 21)
> time <- c(treatment, placebo)
> status <- c(cens1, cens2)
group <- c(rep(1, 21), rep(0, 21))
> Regression <- coxph(formula = Surv(time, status) ~ group)
> Regression
Call: coxph(formula = Surv(time, status) ~ group)
```

Covariate	coef	exp(coef)	se(coef)	z	p
group	-1.57	0.208	0.412	-3.81	0.00014

Likelihood ratio test=16.4 on 1 df, p=5.26e-05 n= 42, number of events= 30

THANK YOU