MS-301
Survival analysis and Clinical Trials

Measurement of Survival Time (or Failure Time):

Following points should be kept in mind while measuring the survival time. The time origin should be precisely defined for each individual. The individuals under study should be as similar as possible at their time origin. The time origin need *not* be and usually is *not* the same calendar time for each individual. Most clinical trials have staggered entries, so that patients enter the study over a period of time. The survival time of a patient is measured from his/her own date of entry. Figure (1.1) and (1.2) show staggered entries and how these are aligned to have a common origin.

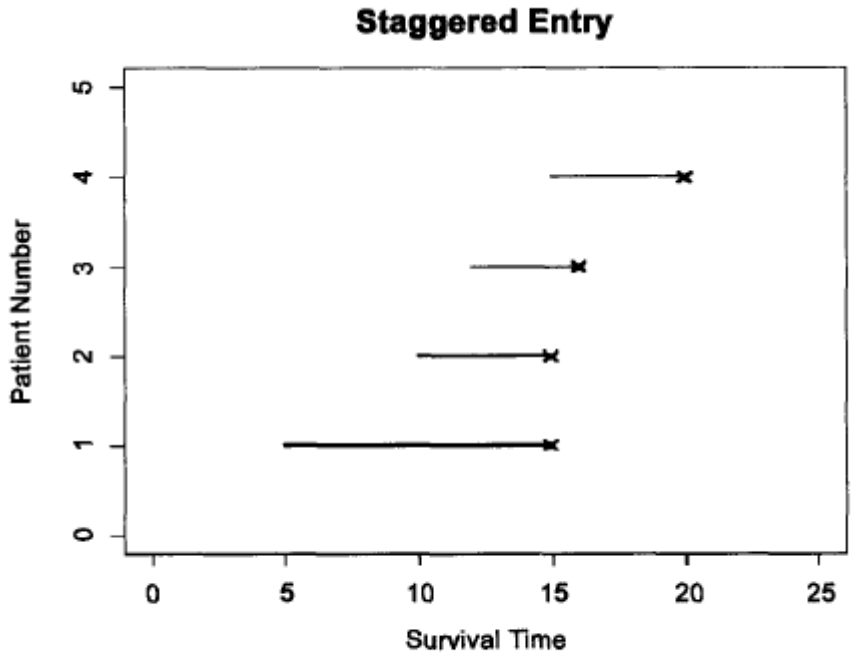


Figure 1.1

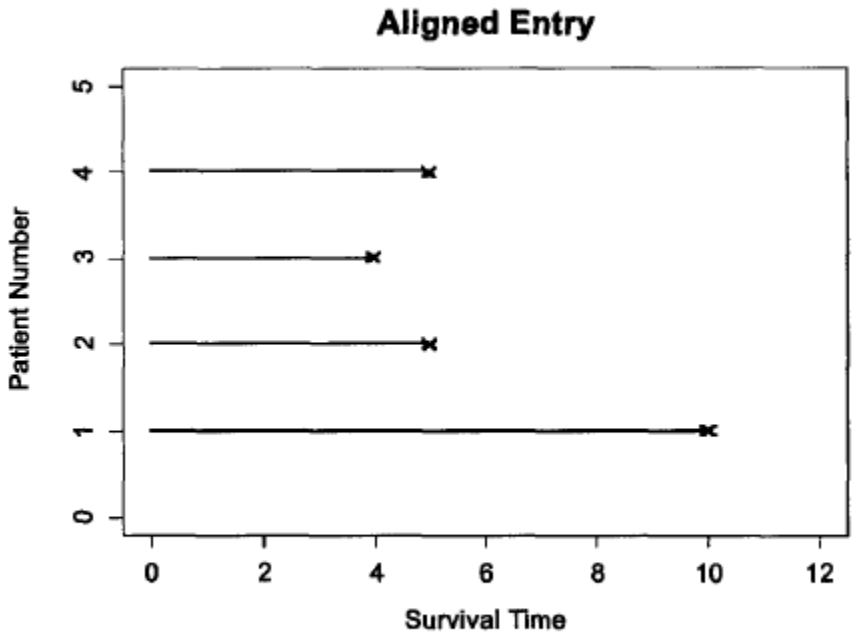


Figure 1.2

Censoring:

The techniques for reducing experimental time are known as censoring. In survival analysis the observations are lifetimes which can be indefinitely long. So quite often the experiment is so designed that the time required for collecting the data is reduced to manageable levels.

Two types of censoring:

Type-I censoring (Time Censoring):

A number (say n) of identical items are simultaneously put into operation. However, the study is discontinued at a predetermined time t_0. Suppose n_u items have failed by this time and the remaining $n_c = n - n_u$ items remain operative. These are called the censored items. Therefore the data consists of the lifetimes of the n_u failed items and the censoring time t_0 for the remaining n_c items, (see Figure 1.3).

Example

Power supplies are major units for most electronic products. Suppose a manufacturer conducts a reliability test in which 15 power supplies are operated over the same duration. The manufacturer decides to terminate the test after 80000

hrs. Suppose 10 power supplies fail during the fixed time interval. Then remaining five are type I censored.

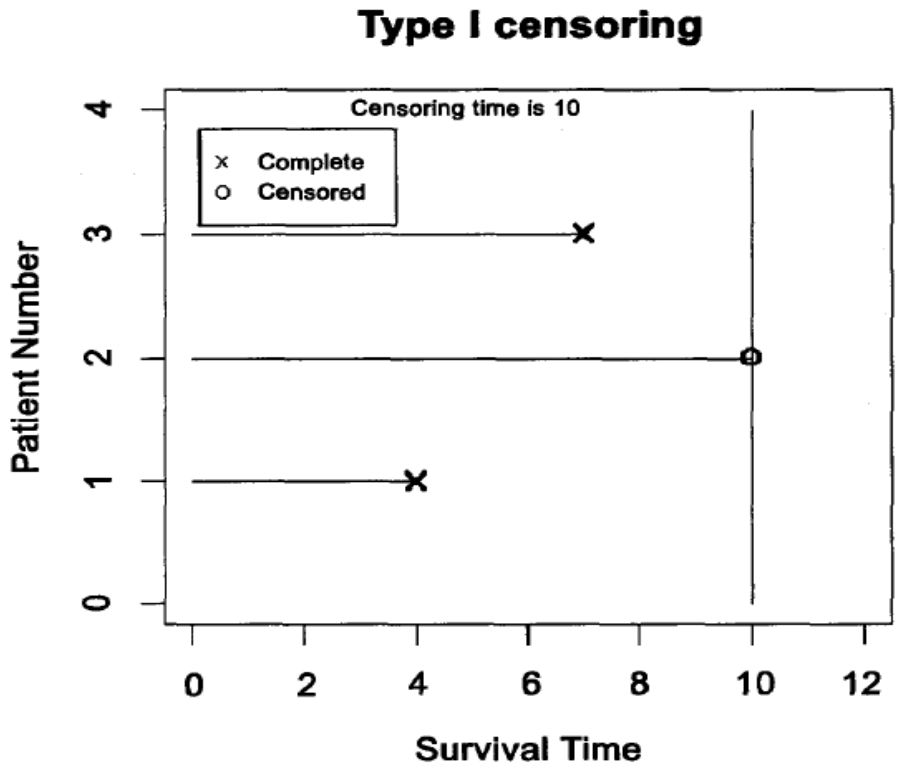


Figure 1.3

***Type II Censoring (Order Censoring)*:**
A number (say n) of identical components are simultaneously put into operation. The study is discontinued when a predetermined number k $(< n)$ of the items fail. Hence the failure times of the k failed items are available. These are the k smallest order statistics of the complete random sample. For the remaining items the censoring time $x_{(k)}$, which is the failure time of the item failing last, is available. (See Figure 1.4.)

Example:
Twelve ceramic capacitors are subjected to a life test. In order to reduce the test time, the test is terminated after eight capacitors fail. The remaining are type II censored.

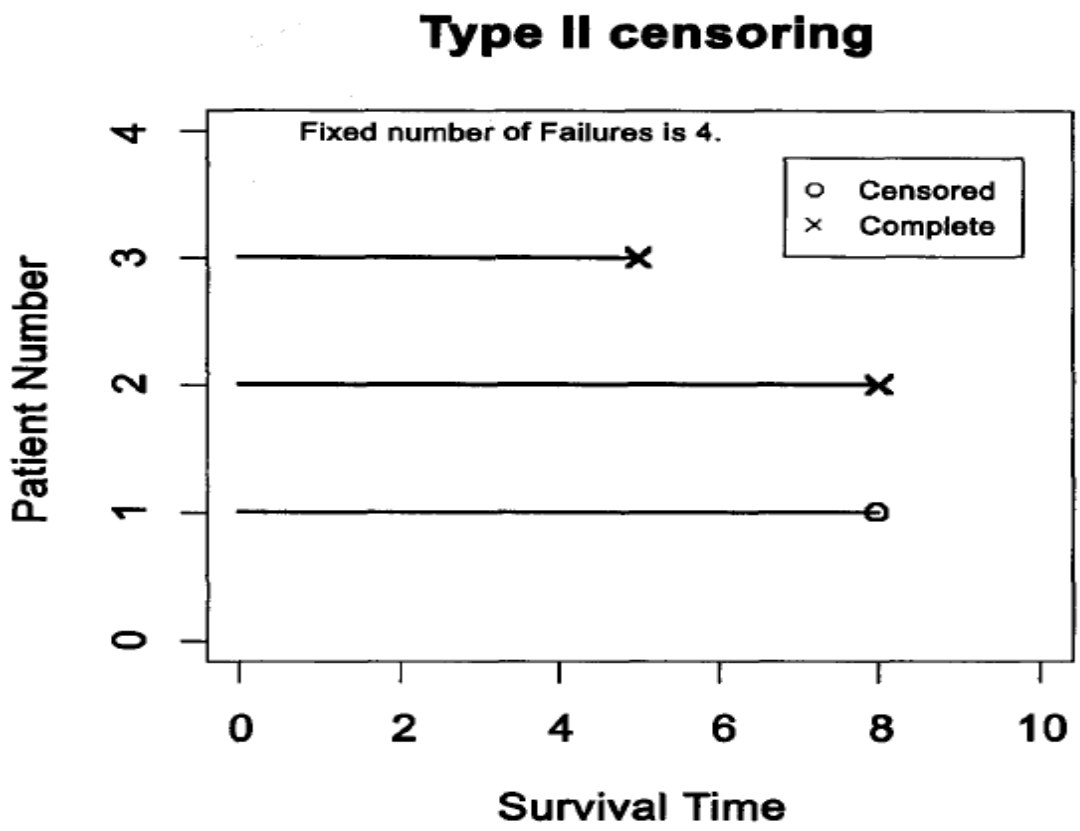


Figure 1.4

Ageing

The concept of ageing plays an important role in the choice of models for the lifetime distributions. It is particularly useful in engineering applications to model the lifetime distributions of units subjected to wear and tear or shocks.

Let T be a continuous, non-negative valued random variable representing the lifetime of a unit. This is the time for which an individual (or unit) carries out its appointed task satisfactorily and then passes into "failed" or "dead" state thereafter. The age of the working unit or living individual is the time for which it is already working satisfactorily without failure. No states besides "living" (operating) or "dead" (failed) are envisaged.

Functions Characterizing Life-time Random Variable

The probabilistic properties of the random variable are studied through its cumulative distribution function F or other equivalent functions defined below:

1. **Survival function or Reliability function**

$$\overline{F}(t) = 1 - F(t) = P[X > t], \quad t \geq 0.$$

2. **Probability density function**

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}\overline{F}(t),$$

3. **Hazard function or failure rate function**

$$\begin{aligned} r(t) &= \lim_{0<h\to 0} \frac{1}{h} P[t < T \leq t + h | T > t] \\ &= \lim_{0<h\to 0} \frac{\overline{F}(t) - \overline{F}(t+h)}{h\overline{F}(t)} \\ &= \frac{f(t)}{\overline{F}(t)}, \text{ provided } \quad F(t) < 1, \text{ and} f(t) \text{ exists.} \end{aligned}$$

Conversely,

$$\overline{F}(t) = \exp\{-\int_0^t r(u)du\}.$$

4. **Cumulative hazard function**

$$R(t) = \int_0^t r(u)du, \quad t \geq 0.$$

$$\overline{F}(t) = \exp[-R(t)].$$

5. **Mean Residual life function**

Let a unit be of age *t.* That is, it has survived without failure up to time *t.* Since the unit has *not* yet failed it has certain amount of residual life time. Let T_t be the residual life-time and $\overline{F}_t$ be its survival function

$$\overline{F}_t(x) = P[T_t > x] = P[T > t + x | T > t] = \frac{\overline{F}(t+x)}{\overline{F}(t)}$$

Then the mean residual life function is defined as

$$L_F(t) = E[T_t] = \int_0^\infty \overline{F}_t(u)du = \int_0^\infty \frac{\overline{F}(t+u)}{\overline{F}(t)}du, \quad t \geq 0.$$

This gives,

$$L_F(0) = E(T) = \mu.$$

and

$$r(t) = [1 + L'(t)]/L(t).$$

Increasing Failure rate (IFR) or Increasing Hazard rate (IHR):
The d.f. f(x) is said to be increasing failure rate or increasing hazard rate if failure rate $\gamma(t)$ increases as t increases.

Decreasing Failure rate (DFR) or Decreasing Hazard rate (DHR):
The d.f. f(x) is said to be decreasing failure rate or decreasing hazard rate if failure rate $\gamma(t)$ decreases as t increases.

Constant Failure rate (CFR) or Constant Hazard rate (CHR):
The d.f. f(x) is said to be constant failure rate or constant hazard rate if failure rate $\gamma(t)$ constant as t increases.

Some Parametric Families of Probability Distributions:

1. Weibull Family

The Weibull distribution is a generalization of the exponential distribution that is appropriate for modeling the lifetimes having constant, strictly increasing (and unbounded) or strictly decreasing hazard functions. It is given by the distribution function.

$$F(t) = 1 - e^{-\lambda t^\gamma}, \quad t > 0, \quad \lambda > 0, \gamma > 0$$

Where λ and γ are both positive valued parameters. It is clear that $\gamma = 1$ gives the exponential distribution with mean $\frac{1}{\lambda}$. Hence this may be viewed as a generalization of the exponential distribution. It is interesting to look at its failure rate.

$$f(t) = -\frac{d}{dt}\overline{F}(t) = \lambda\gamma t^{\gamma-1}e^{-\lambda t^{\gamma}}, \quad t > 0,$$

$$\text{and} \quad r(t) = \frac{f(t)}{\overline{F}(t)} = \lambda\gamma t^{\gamma-1}, \quad t > 0.$$

2. Gamma Family

The gamma distribution is another important generalization of the exponential distribution. The probability density function for the gamma distribution is

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

Where λ and γ are positive parameters denoting scale and shape respectively. Putting γ = 1 we get the exponential family. It is often difficult to differentiate between Weibull and gamma distributions based on their probability density functions, since shapes of these plots are similar. The differences between these two distributions become apparent when their hazard rates are compared. The behaviour of hazard rate for gamma distribution can only be indirectly investigated as *F(t)* and hence *r(t)* do not have closed form expressions. The reason that the gamma distribution is less popular in modelling than Weibull is partially attributed to this fact.

$$\begin{aligned} r(t) = \frac{f(t)}{\overline{F}(t)} &= \frac{\lambda^{\gamma} t^{\gamma-1} e^{-\lambda t}}{\Gamma(\gamma)} \left[\int_t^{\infty} \frac{\lambda^{\gamma} x^{\gamma-1} e^{-\lambda x} dx}{\Gamma(\gamma)}\right]^{-1} \\ &= \left[\int_t^{\infty} (\frac{x}{t})^{\gamma-1} e^{-\lambda(x-t)} dx\right]^{-1} \\ &= \left[\int_0^{\infty} \left(\frac{y+t}{t}\right)^{\gamma-1} e^{-\lambda y} dy\right]^{-1} \quad \text{for} \quad 0 < t < \infty, \end{aligned}$$

3. Log-normal Family

A random variable *T* is said to have lognormal distribution when *Y* = loge *T* is distributed as normal (Gaussian) with mean μ and variance σ^2. The p.d.f. and survival function of lognormal distribution, respectively are:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log_e t - \mu)^2\right], \quad t > 0, \sigma > 0$$

and

$$\overline{F}(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x}\left[-\frac{1}{2\sigma^2}(\log_e x - \mu)^2\right] dx.$$

The mean and the variance of the distribution are given by

$$E(T) = \exp[\mu + \sigma^2/2]$$

$$Var(T) = [e^{2\mu+\sigma^2}][e^{\sigma^2} - 1].$$

The density curve is positively skew and the skewness increases with σ^2. There are no closed-form expressions for survival and hazard function.
However, computing survival and hazard functions is not difficult. We can write

$$\overline{F}(t) = P\left[Z > \frac{\log_e t - \mu}{\sigma}\right],$$

where $Z \sim N(0,1)$. Thus

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

Where Φ (.) represents distribution function of standard normal variate. So using the table of the cumulative probability integral for *Z*, one can evaluate the survival function of T. Similarly, using the table of ordinates of standard normal distribution we can compute *f(t)* and thus we can get values for hazard function, *h(t)* = $\frac{f(t)}{\bar{F}(t)}$
The hazard function is non-monotonic; initially it increases, reaches a maximum and then decreases to zero as time approaches infinity.

4. Linear Failure Rate Family

It is given by

$$F(x) = 1 - \exp\{-(x + \frac{1}{2}\theta x^2)\}, x > 0, \theta \geq 0,$$

$$f(x) = (1 + \theta x)e^{-(x+\frac{1}{2}\theta x^2)} \text{ and}$$

$$r(x) = \frac{f(x)}{\overline{F}(x)} = (1 + \theta x).$$

5. Makeham Family

This is given by the distribution function

$$F(x) = 1 - \exp[-\{x + \theta(x + e^{-x} - 1)\}], \quad x > 0, \quad \theta \geq 0.$$

$\theta = 0$ again leads to the exponential distribution.

$$f(x) = exp\{-[x + \theta(x + e^{-x} - 1)]\}[1 + \theta(1 - e^{-x})] \text{ and}$$

$$r(x) = [1 + \theta(1 - e^{-x})]$$

6. Pareto Family

Simple (one parameter) form of this distribution is given by

$$F(x) = 1 - (1 + \theta x)^{-1/\theta}, \quad x > 0, \theta > 0$$
$$f(x) = (1 + \theta x)^{-(1/\theta + 1)}, \quad x > 0, \quad \theta > 0$$
$$r(x) = (1 + \theta x)^{-1}, \quad x > 0, \quad \theta > 0.$$

In the last chapter we have investigated several parametric distributions which are useful in modelling lifetime data. In this chapter we shall consider the techniques of analysis of lifetime data such as point and interval estimation of the unknown parameters and testing hypotheses regarding these parameters. In general, we shall use the method of maximum likelihood for estimation.

Method of Maximum Likelihood

Let T_1, T_1,..., T_n be a random sample from a life distribution having probability density f(x; $\underline{\theta}$) where $\underline{\theta}$) = $\{\theta_1, \theta_2, ..., \theta_p\}$ ε $\boldsymbol{\theta}$ is the vector of unknown parameters. Since the lifetimes are independent, the likelihood function L($\underline{t}$, $\underline{\theta}$), is the product of probability density functions evaluated at each sample point. Thus,

$$L(\underline{t}, \underline{\theta}) = \prod_{i=1}^{n} f(t_i, \underline{\theta}),$$

Where $\underline{t}$ = $(t_1,t_2,...,t_n)$ is the data point. The maximum likelihood estimator $\underline{\hat{\theta}}$ is the value of $\underline{\theta}$ which maximizes L($\underline{t}$, $\underline{\theta}$)for fixed $\underline{t}$. That is, $\underline{\hat{\theta}}$ is the maximum likelihood estimator of $\underline{\theta}$, if L($\underline{t}$, $\underline{\hat{\theta}}$) $\geq$ L(, $\underline{\check{\theta}}$) for any other estimator.

Parametric Analysis for survival Data:

A. The Exponential Distribution

Let $t_1, t_2,...,t_n$ be a random sample from an exponential distribution with parameter λ.

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

$$L(\underline{t};\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}.$$

The log likelihood function is

$$\log L(\underline{t},\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} t_i.$$

The score is

$$U(\lambda) = \frac{\delta}{\delta\lambda} \log L(\underline{t},\lambda)$$
$$= \frac{n}{\lambda} - \sum_{1}^{n} t_i.$$

$$\left[\frac{\delta}{\delta\lambda} \log L(\underline{t},\lambda)\right]_{\lambda=\hat{\lambda}} = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_1^n t_i}.$$

$$I(\lambda) = \frac{n}{\lambda^2}.$$

Sample information at $\hat{\lambda}$ is $\frac{n}{\hat{\lambda}^2}$ and var($\hat{\lambda}$) is $\frac{\hat{\lambda}^2}{n}$

Notice that the maximum likelihood estimator of λ is the ratio of total number of failures to the total lifetime of all the units.

The Gamma Distribution

Let $t_1, t_2, ..., t_n$ be a random sample from a gamma distribution with scale parameter λ and shape parameter γ

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

$$L(\underline{t};\lambda,\gamma) = \frac{\lambda^{n\gamma}}{[\Gamma(\gamma)]^n} e^{-\lambda \sum_1^n t_i} \prod_{i=1}^{n} (t_i)^{\gamma-1}.$$

$$\log L(\underline{t};\lambda,\gamma) = n\gamma \log \lambda - n \log \Gamma(\gamma) - \lambda \sum_{i=1}^{n} t_i + (\gamma - 1) \sum_{i=1}^{n} \log t_i.$$

The score vector has components;

(i) $\frac{\delta}{\delta\lambda} \log L(\lambda,\gamma) = \frac{n\gamma}{\lambda} - \sum_1^n t_i$ and

(ii) $\frac{\delta}{\delta\gamma} \log L(\lambda,\gamma) = n \log \lambda - \frac{n\Gamma'(\gamma)}{\Gamma(\gamma)} + \sum_1^n \log t_i$, where $\Gamma'(\gamma) = \frac{\delta}{\delta\gamma}\Gamma(\gamma)$.

The MLEs of λ and γ satisfy

$$\hat{\lambda} = \hat{\gamma}(\bar{t})^{-1}$$

and

$$n \log \hat{\lambda} + \sum_1^n \log t_i = \frac{n\Gamma'(\gamma)}{\Gamma(\gamma)}.$$

Substituting for $\hat{\lambda}$ in (4.3.3) from (4.3.2), we get

$$n \log(\frac{\hat{\gamma}}{\bar{t}}) + \sum_1^n \log t_i = \frac{n\Gamma'(\hat{\gamma})}{\Gamma(\hat{\gamma})}.$$

Or

$$\frac{\Gamma'(\hat{\gamma})}{\Gamma(\hat{\gamma})} - \log(\hat{\gamma}) = \log R,$$

where

$$R = \frac{(\prod_1^n t_i)^{1/n}}{\bar{t}}$$

= Ratio of the geometric mean and the arithmetic mean.

Some iterative numerical method such as Newton - Raphson procedure must be used for solving the equation.

Sample information matrix

$$\frac{\delta^2}{\delta\lambda^2}\log L(\lambda,\gamma) = -\frac{n\gamma}{\lambda^2}.$$

$$\frac{\delta^2}{\delta\lambda\delta\gamma}\log L(\lambda,\gamma) = \frac{n}{\lambda}.$$

$$\frac{\delta^2}{\delta\gamma^2}\log L(\lambda,\gamma) = -n\left[\frac{\Gamma''(\gamma)}{[\Gamma(\gamma)]} - \frac{[\Gamma'(\gamma)]^2}{[\Gamma(\gamma)]^2}\right].$$

$$i(\lambda,\gamma) = \begin{bmatrix} \frac{n\gamma}{\lambda^2} & -\frac{n}{\lambda} \\ -\frac{n}{\lambda} & n[\frac{\Gamma''(\gamma)}{(\Gamma(\gamma)]} - \frac{[\Gamma'(\gamma)]^2}{\Gamma(\gamma)^2}] \end{bmatrix}.$$

The Weibull Distribution

Let $t_1, t_2, ..., t_n$ be a random sample from the Weibull distribution with scale parameter λ and shape parameter γ.

$$f(t;\lambda,\gamma) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}), \quad t \geq 0, \quad \lambda, \quad \gamma > 0$$

$$\log L(\lambda,\gamma) = n\log\lambda + n\log\gamma + \sum_{i=1}^{n}(\gamma-1)\log t_i - \lambda\sum_{1}^{n} t_i^{\gamma}.$$

The elements of the score vector are
(i) $\frac{\delta}{\delta\lambda}\log L(\underline{t};\lambda,\gamma) = \frac{n}{\lambda} - \sum_1^n t_i^{\gamma}$ and
(ii) $\frac{\delta}{\delta\gamma}\log L(\underline{t};\lambda,\gamma) = \frac{n}{\gamma} + \sum_1^n \log t_i - \lambda\sum_1^n t_i^{\gamma}\log t_i$.
The MLEs of λ and γ satisfy the equations:

$$\frac{n}{\hat{\lambda}} - \sum_{1}^{n} t_i^{\hat{\gamma}} = 0.$$

$$\frac{n}{\hat{\gamma}} + \sum_{1}^{n}\log t_i - \hat{\lambda}\sum_{1}^{n} t_i^{\hat{\gamma}}\log t_i = 0.$$

$$\hat{\lambda} = n[\sum_{1}^{n} t_i^{\hat{\gamma}}]^{-1}.$$

$$\hat{\lambda} = [\frac{n}{\hat{\gamma}} + \sum_{1}^{n}\log t_i][\sum_{1}^{n} t_i^{\hat{\gamma}}\log t_i]^{-1}.$$

$$n[\sum_1^n t_i^{\hat{\gamma}}]^{-1} = [\frac{n}{\hat{\gamma}} + \sum_1^n \log t_i][\sum_1^n t_i^{\hat{\gamma}} \log t_i]^{-1}.$$

$$n[\sum_1^n t_i^{\hat{\gamma}}]^{-1}[\sum_1^n t_i^{\hat{\gamma}} \log t_i] - \frac{n}{\hat{\gamma}} - \sum_1^n \log t_i = 0.$$

$$h(\hat{\gamma}) = 0.$$

Graphical Procedure for estimating the parameters

The survival function of the Weibull distribution is

$$\overline{F}(t) = \exp(-\lambda t^{\gamma}).$$

Hence

$$\log\{[\overline{F}(t)]^{-1}\} = \lambda t^{\gamma}$$

and therefore

$$\log\log[\{\overline{F}(t)\}^{-1}] = \log\lambda + \gamma \log t.$$

Let $t_{(1)} < t_{(2)} ... < t_{(n)}$ be the order statistics from the random sample. Estimate $\overline{F}(t_{(i)})$ by $\hat{\overline{F}}(t_{(i)})$ where $\hat{\overline{F}}(t_{(i)})$ is empirical survival function. Plot $\log\log[\{\hat{\overline{F}}(t_{(i)})\}^{-1}]$ against $\log[t_{(i)}]$ for $i = 1, 2, ..., n$. If the underlying distribution is indeed Weibull, the graph will be approximately a straight line. A line could be fitted by the usual least squares techniques or just by inspection. The slope of the line will give the initial estimate of γ and the y - intercept will provide an initial estimate of λ.

Sample Information Matrix

$$i(\lambda, \gamma) = \begin{bmatrix} \frac{n}{\lambda^2} & \sum_1^n t_i^{\gamma} \log t_i \\ \frac{n}{\gamma^2} + \lambda \sum_1^n t_i^{\gamma} (\log t_i)^2 & \end{bmatrix}$$

Alternatively, $(\hat{\lambda}, \hat{\gamma})$ can be obtained by using Newton - Raphson method of scoring with the estimates obtained from the graphical method as the initial solution. Let $\underline{\theta} = (\lambda, \gamma)'$ and $\underline{\theta}^{(0)} = (\hat{\lambda}, \hat{\gamma})'$ as obtained from graphical

Example:

Following are the times (in minutes) to break down of an insulating fluid between electrodes recorded at voltage 36kv. Assume Weibull distribution and estimate the parameters of the distribution.

.35, .59, .96, .99, 1.69, 1.97, 2.07, 2.58, 2.71, 2.90, 3.67, 3.99, 5.35, 13.77

Solution :

Graphical Method.

Following figure (Figure 4.5) shows the graph of $\log[-\log F_n(t(i))]]$ vs $\log[t(i)]$.

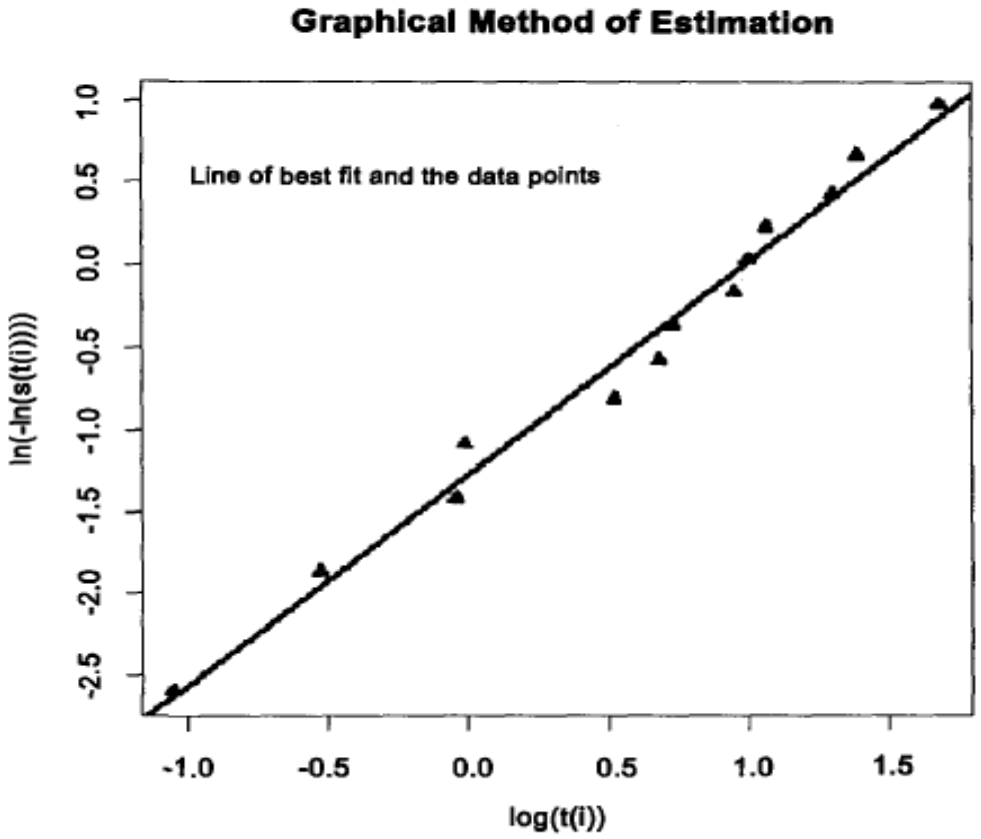


Figure 4.5

The points show strong linear trend. A line of best fit, obtained by the method of least squares, is also shown on the plot.

Correlation coefficient = 0.9847

Regresion coefficient = 1.29732.

The assumption of Weibull model is justifiable. From the graph, estimate of $\gamma = 1.29732$. In order to get refined estimate of γ we use approximate trial and error method. For this we compute $h(\hat{\gamma})$ for values of $\hat{\gamma}$ in the range $(1, 3)$. $h(\hat{\gamma}) < 0$ for $\hat{\gamma} = 1.15$ and $h(\hat{\gamma}) > 0$ for $\hat{\gamma} = 1.1$. Now we use simple bi-section method to get $\hat{\gamma} = 1.126458$ with $h(\hat{\gamma}) = -1.00156e^{-12}$. Using above value of $\hat{\gamma}$ and (4.3.8), we get $\hat{\lambda} = 0.2631458$.

Nonparametric Estimation of the Survival Function

In the previous chapters, we have considered parametric models for life distribution and methods of estimation of the unknown parameters involved in these models. These methods are discussed for complete as well as censored data. Such a parametric modelling is based on prior knowledge of the failure characteristics of the individual (or unit). However, in many practical situations such prior knowledge may not be available.

Uncensored (complete) Data

Let x_1, x_2, ..., x_n be a random sample of size n from a distribution F. A non-parametric estimator of F(x) is $F_n(x)$ where $F_n(x)$ is sample (or empirical) distribution function defined as

$$F_n(x) = \frac{1}{n} \text{ [The number of observations } \leq x]$$

The Glivenko - Cantelli theorem tells us that

$$\sup_x [|F_n(x) - F(x)|] \to 0 \text{ as } n \to \infty \quad \text{with probability 1.}$$

Hence, $F_n(x)$ as a function is a consistent estimator of $F(x)$. Then obviously, a non-parametric consistent estimator of survival function $[\overline{F}(x)]$ is $\overline{F}_n(x)$ where $\overline{F}_n(x)$ is defined by

$$\overline{F}_n(x) = \frac{\#(\text{observations} \quad > x)}{n}$$

It is seen that for a fixed x, $n\overline{F}_n(x)$ follows the binomial distribution with parameters n and $\overline{F}(x)$. Hence the standard asymptotic theory leads to the asymptotic normality of $\sqrt{n}(\overline{F}_n(x) - \overline{F}(x))$. Unbiasedness of this estimator is obvious. Further, using these results, confidence interval for $\overline{F}(x)$, for fixed x, can be constructed. Confidence bands for the entire $\overline{F}(x), 0 < x < \infty$ may be constructed using the distribution of $\sup_x |\overline{F}_n(x) - \overline{F}(x)|$.

Example: A complete data set of n = 23 ball bearing failure times to test endurance of deep groove ball bearings has been extensively studied. The ordered set of failure times measured in 10^6 revolutions is 17.88, 28.92, 33.0, 41.52, 42.12, 45.60, 48.48, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04,173.40.

Observe that the data set contain two observations tied at 68.64. An empirical survival function has a downward step of $\frac{1}{n}$ at each observed lifetime. It is also the survival function corresponding to a discrete random variable for which *n* mass values are equally likely. Ties are not difficult to adjust for since the formula remains the same and the function will take a downward step of *d/n* if there are *d* tied observations at a particular time point.

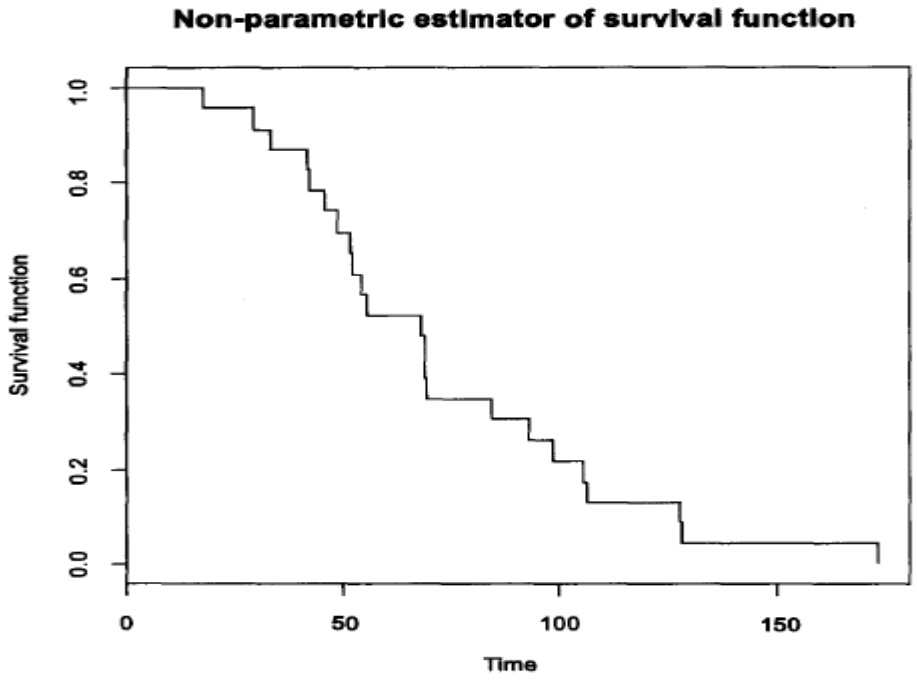
In Figure 5.1 the non-parametric estimator of $\overline{F}(t)$ is shown where the downward steps in $\overline{F}_n(t)$ are connected by vertical lines.

A point estimate $[S_n(t)]$ of the survivor function at t $(\overline{F}(t))$, for $t = 50$ (in the units of 10^6) is given by, $S_n(50) = \frac{16}{23} = 0.696$. Approximate 95% confidence interval for $S(t)$ is $S_n(50) \pm 1.96\sqrt{\frac{S_n(50)(1-S_n(50))}{23}}$ which reduces to

$$0.508 < S(50) < 0.884.$$

***Confidence bands*:** The confidence bands for the survival function can be obtained by using Kolmogorov - Smirnov statistic.

$$D_n = \sup_x |F_n(x) - F(x)|$$



Therefore

$$\sup_x |\overline{F}_n(x) - \overline{F}(x)| = D_n.$$

From the tables of D_n statistic, find $D_{n,(1-\alpha)}$ such that

$$P[D_n \leq D_{n,(1-\alpha)}] = 1 - \alpha.$$

That is, $\quad P[\sup_x |\overline{F}_n(x) - \overline{F}(x)| \leq D_{n,(1-\alpha)}] = 1 - \alpha.$

This gives

$$P[\overline{F}_n(x) - D_{n,(1-\alpha)} \leq \overline{F}(x) \leq \overline{F}_n(x) + D_{n,(1-\alpha)} \quad \forall x] = 1 - \alpha.$$

As $0 \leq \overline{F}(x) \leq 1$,

$$L_n(x) = \max[\overline{F}_n(x) - D_{n,(1-\alpha)}, 0]$$

and

$$U_n(x) = min[\overline{F}_n(x) + D_{n,(1-\alpha)}, 1]$$

gives the required confidence band.

The asymptotic probability distribution of D_n is an infnite sum, which may, in practice, be approximated by its first term $2e^{-2d^2}$.

That is, $\lim_{n\to\infty} P[D_n > d/\sqrt{n}] = 2e^{-2d^2}$. Setting this equal to α and solving we get

$$\lim_{n\to\infty} P[\hat{S}(t) - \frac{d_{(1-\alpha)}}{\sqrt{n}} < S(t) < \hat{S}(t) + \frac{d_{(1-\alpha)}}{\sqrt{n}}]$$
$$= 1 - \alpha.$$

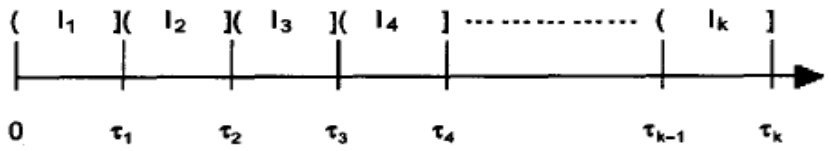
Then $\hat{S}_n(t) \pm \frac{d_{(1-\alpha)}}{\sqrt{n}}$ define an asymptotic $100(1-\alpha)\%$ non-parametric confidence band for $S(t)$ for $t \leq T_{(n)}$. In Dixon and Massey (1983), the tables for $d_{(1-\alpha)}/\sqrt{n}$ are available. Nair (1984) has shown that the asymptotic critical value of $d_{(1-\alpha)}/\sqrt{n}$ is valid for n as small as 25.

Censored Data

The general case of randomly right censored data will be considered here.

(A) *The Actuarial Method*: Life tables have historically been used by acturies for estimating the survival distribution of humans, but apply well to reliability and biostatistical situations for which grouped data (rather than raw data) which display the combined survival experience of a cohort of individuals who fall into natural groupings by age or calender time interval are available.

Notation : Suppose the time interval $(0, \tau]$ is under consideration. Let this be partitioned into a fixed sequence of intervals $I_1, ..., I_k$. These intervals are almost always, but not necessarily of equal lengths and for human populations the length of each interval is usually one year.



For a life table, let

n_i = # alive at the beginning of I_i,

d_i = # died during I_i,

ℓ_i = # lost to follow-up during I_i,

w_i = # withdrawn during I_i,

$P_i = P$ [surviving through I_i / alive at the beginning of I_i]

$Q_i = 1 - P_i$

Table 5.1 (Cutler and Ederer (1958), Miller, (1981)] is an example of a life table. In this table, Cutler and Ederer reviewed annual cohorts of Connecticut residents with localized kidney cancer diagnosed in the years 1946 to 1951. The study terminated on Dec. 31, 1951. Within each annual cohort, the patients were subdivided by years after diagnosis, commonly called as "time-on-study". For each such time interval, two groups of patients were defined : those who died during the interval and those who were lost to follow-up. (The later group might also include deaths from other causes). During the last time-on-study interval of each cohort, a third category was defined ; those known to be alive at the end date of study. The term used to describe these patients was "withdrawn alive". In our terminology, patients lost to follow-up or withdrawn alive are said to have censored survival times.

Life Table

Year after Diagnosis	Alive at the Beginning of Interval	Died during the Interval	Lost to followup during the Interval	Withdrawn Alive during the Interval
0 - 1	126	47	4	15
1 - 2	60	5	6	11
2 - 3	38	2	-	15
3 - 4	21	2	2	7
4 - 5	10	-	-	6

We break up the survival probability $S(\tau_k)$ into a product of conditional probabilities:

$$\begin{aligned} S(\tau_k) &= P[T > \tau_k] \\ &= P[T > \tau_1]P[T > \tau_2/T > \tau_1] \times P[T > \tau_k|T > \tau_{k-1}] \\ &= P_1.P_2...P_k, \end{aligned}$$

where

$$P_i = P[T > \tau_i | T > \tau_{i-1}].$$

The actuarial method estimates P_i seperately and then multiplies the estimates to get the estimate of $S(\tau_k)$.

For an estimate of P_i we would have used $(1 - \frac{d_i}{n_i})$ if the data were complete i.e. there were no losses and withdrawals in I_i. We assume that, on the average, those individuals who are lost to follow-up or withdrawn during I_i were at risk for half of the interval. Therefore, the effective sample size is defined as

$$n_i' = n_i - \frac{1}{2}(\ell_i + w_i)$$

and

$$\hat{Q}_i = q_i = \frac{d_i}{n_i'}, \quad \hat{P}_i = p_i = 1 - \hat{Q}_i.$$

Then the actuarial estimate of $S(\tau_k)$ is

$$\hat{S}(\tau_k) = \pi_{i=1}^{k} \hat{P}_i.$$

Example: Following table shows the annual failures and removals (censored) of a fleet of 200 single engine aircrafts. Removals resulted from aircrafts eliminated from the inventory for various reasons other than engine failure.

Year	No. of failures	No. of removals
1981	5	0
1982	10	1
1983	12	5
1984	8	2
1985	10	0
1986	15	6
1987	9	3
1988	8	1
1989	4	0
1990	3	1

Following table shows the life table for engine failure data

Year	Working at the beginning of interval n_i	Failed during the interval d_i	Censored during the interval $(\ell_i + w_i)$
0-1	200	5	0
1-2	195	10	1
2-3	184	12	5
3-4	167	8	2
4-5	157	10	0
5-6	147	15	6
6-7	126	9	3
7-8	114	8	1
8-9	105	4	0
9-10	101	3	1

Note : Year 1980 is taken as base year.

Acturial Estimate of Reliability Function

τ_i	n_i	$\hat{p}_i$	$\hat{S}(\tau_i)$: Reliability at τ_i
1	200	0.975	0.975
2	194.5	0.949	0.925
3	181.5	0.934	0.864
4	166	0.952	0.823
5	157	0.936	0.770
6	144	0.896	0.690
7	124.5	0.928	0.640
8	113.5	0.930	0.595
9	105	0.962	0.573
10	100.5	0.970	0.556

$$\hat{S}(\tau_k) = 0.556$$

(B) *Product - Limit (Kaplan - Meier) Estimator*

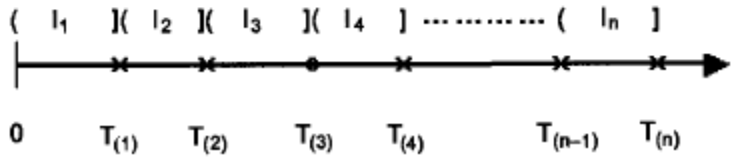
Let $X_1, X_2, ..., X_n$ be the lifetimes of the n individuals (units). With each X_i there is associated a random variable C_i, known as its censoring variable. What we observe is $T_i = min(X_i, C_i)$ and

$$\delta_i = \begin{cases} 1 \text{ if } X_i \leq C_i \\ 0 \text{ if } X_i > C_i \end{cases}$$

For the moment assume that there are no ties. Let $T_{(1)} < T_{(2)} ... < T_{(n)}$ be the order statistics corresponding to $T_1, T_2, ..., T_n$ and with a little abuse of notation, define $\delta_{(i)}$ to be the value associated with $T_{(i)}$. That is, $\delta_{(i)} = \delta_j$ if $T_{(i)} = T_j$. Let $R(t)$ denote the risk set at time t, which is the set of subjects still alive at time t^- (just prior to t) and $n_i = \#R(T_{(i)}) = \#$ alive at $T_{(i)}^-$, and $d_i = \#$ died at $T_{(i)}$.

It may be noted that for the data with no ties $d_i = 1$ or 0 depending on $\delta_{(i)}$ is 1 or 0.

The time interval of interest, in this case, is $(0, \tau]$ where $\tau = T_{(n)}$. We consider the subdivision of this interval into n subintervals I_i with end points $T_{(i)}$.



On the time axis, $\times$ denotes $\delta_{(i)} = 1$ i.e. an uncensored observation and 0 denotes $\delta_{(i)} = 0$ i.e. censored observation.

Let $P_i = P$ [surviving through I_i / alive at the beginning of I_i.]

$= P[T > T_{(i)} | T > T_{(i-1)}]$, and

$Q_i = 1 - P_i$. Its estimator is then $\hat{Q}_i = q_i = \frac{d_i}{n_i}$ and hence that of P_i is

$$\hat{P}_i = p_i = 1 - q_i = \begin{cases} 1 - \frac{1}{n_i} & \text{if } \delta_{(i)} = 1 \text{ (uncensored)} \\ 1 & \text{if } \delta_{(i)} = 0 \text{ (censored)} \end{cases}$$

The PL (product limit) estimator of the survival function when no ties are present is then

$$\begin{aligned} \hat{S}(t) &= \prod_{U_i T_{(i)} \le t} (1 - \frac{1}{n_i}) \\ &= \prod_{T_{(i)} \le t} (1 - \frac{1}{n_i})^{\delta_{(i)}} \\ &= \prod_{T_{(i)} \le t} (1 - \frac{1}{n - i + 1})^{\delta_{(i)}} \\ &= \prod_{T_{(i)} \le t} (\frac{n - i}{n - i + 1})^{\delta_{(i)}} \end{aligned}$$

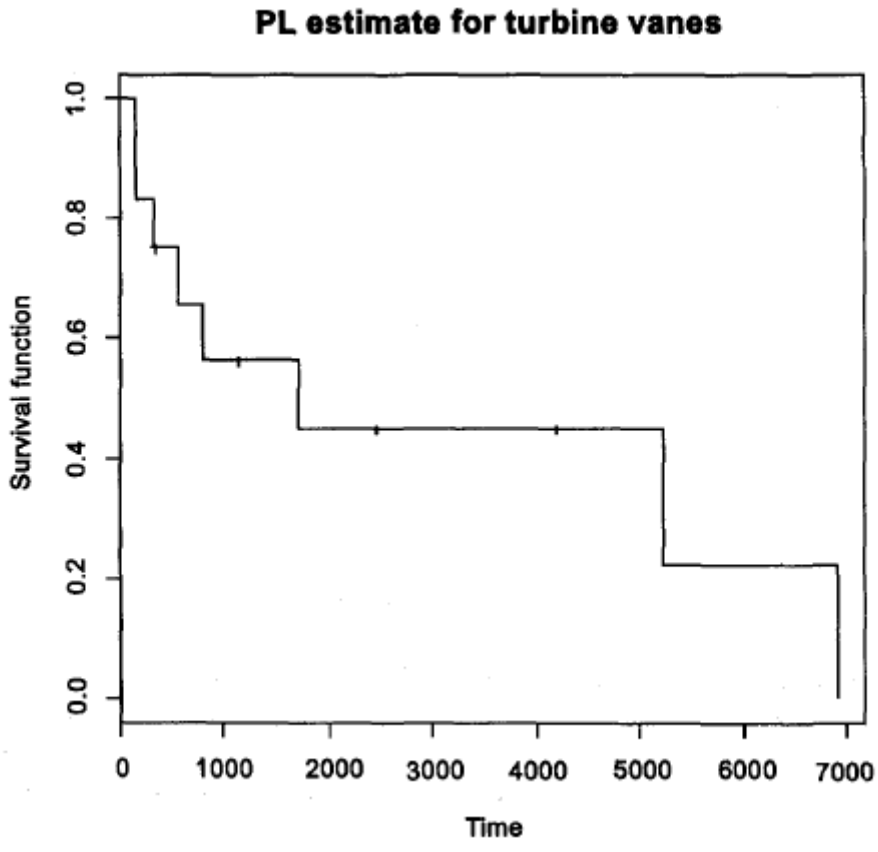
Example: The following failure and censor times (in operating hrs) were recorded on 12 turbine vanes : 142, 320, 345+, 560, 805, 1130+, 1720, 2480+, 4210+, 5280, 6890. (+ indicates censored observation). Censoring was a result of failure mode other than wearout. Plot PL estimate of survival function.

Solution:

Estimation of Survival Function for Turbine Vanes

j	$\tau_{(j)}$	d_j	n_j	$\hat{S}(\tau_{(j)})$
1	142	1	12	0.9167
2	149	1	11	0.8334
3	320	1	10	0.7500
4	345^+	0	–	**
5	560	1	8	0.6563
6	805	1	7	0.5625
7	1130^+	0	–	**
8	1720	1	5	0.4500
9	2480^+	0	–	**
10	4210^+	0	–	**
11	5230	1	2	0.2250
12	6890	1	1	0.0000

** These are censored observations, hence we have not computed survival probability for them. However, it may be noted that in the relevant intervals the survival probability remains constant.



Proportional Hazards Model (PH Model) : This is the most cited and used model which is introduced by Cox (1972) in his path breaking paper. The simplest form of proportional hazards model is :

$$h(t, \underline{Z}) = h_0(t)\psi(\underline{Z}) \tag{8.1.1}$$

where $h_0(t)$ is the baseline failure rate and $\psi(\underline{Z})$ is the link function bringing in the covariates. It satisfies $\psi(\underline{0}) = 1$ and $\psi(\underline{Z}) \geq 0$ for all $\underline{Z}$. Note that the failure rate function $h(t,)$ is a function of time t as well as the covariate values $\underline{z}$.

The following two parametric link functions are commonly used:

(i) $\psi(\underline{Z}; \underline{\beta}) = \exp(\underline{\beta}'\underline{Z})$: log linear form.

(ii) $\psi(\underline{Z}; \underline{\beta}) = 1 + \underline{\beta}'\underline{Z}$: linear form.

We shall consider the base line hazard rate, $h_0(t)$, as completely unknown and the covariates as fixed quantities, thus leading to the semi-parametric PH model.

Clinical Trials

Introduction

A clinical trial is a research study conducted to assess the utility of an intervention in volunteers. Interventions may be diagnostic, preventative or treatment in nature and may include drugs, biologics, medical devices or methods of screening. Interventions may also include procedures whose aim is to improve quality of life or to better understand how the intervention works in volunteers.

Quality clinical research must be well planned, closely and carefully monitored and conducted, and appropriately analyzed and reported. Greater attentiveness to detail at the design stage argues for greater efficiency at the analysis and reporting stages. This is important on a per protocol basis as well as across the entire clinical development plan to support regulatory filing.

Phases of Clinical Trials and Objectives

It is well known that evidence to support regulatory approval of a new drug derives from clinical trials. Such trials are categorized as Phase I, Phase II or Phase III. Although these categories may not be mutually exclusive (nor in some cases mutually exhaustive), there is general agreement as to what types of clinical studies comprise the bulk of the trials within each phase.

Phase I Trials

Phase I trials may consist of “early Phase I" trials, early dose ranging trials, bioavailability or pharmacokinetic trials, or mechanism of action studies. Early Phase I trials represent the initial introduction of the drug in humans, in order to characterize the acute pharmacological effect. For most classes of drugs, healthy subjects are enrolled, in an attempt to reduce the risk of serious toxicity and to avoid confounding pharmacological and disease effects. The idea is to introduce the drug to humans without inducing acute toxicity.

Early dose ranging trials, often called dose tolerance or dose titration trials, are also most often conducted in healthy subjects. Both the effects of single dosing and multiple dosing schemes are studied. The objective of these trials is to determine a ‘tolerable’ dose range, such that as long as future dosing remains in this range, no intolerable side effects of toxicities would be expected to be seen.

Early Phase I trials and early dose ranging trials don't establish nor quantitate efficacy characteristics of a drug. These studies have to be conducted first, so that acute pharmacological effects may be described, and a range of tolerable doses determined, which guide clinical use of the drug for later studies.

The primary objectives of Phase I bioavailability and pharmacokinetic trials are to characterize what happens to the drug once it's injected into the human body.

That is, properties such as absorption, distribution, metabolism, elimination, clearance, and half-life need to be described. These trials also usually enroll healthy subjects and are often called "blood level trials."
Mechanism of action trials attempt to identify how the drug induces its effects. An example is the class of H2-receptor antagonists, such as cimetidine, ranitidine, famotidine and zanatidine, which by blocking the H2-receptor reduce the secretion of gastrin which in turn leads to a reduction of gastric acid production. Another example is the H1-receptor antagonist, seldane, which by blocking the H1-receptor reduces histamine release. Other examples are the ACE (angiotensin-converting-enzyme) inhibitors (e.g. captopril, enalapri, quinapril) which are competitive inhibitors of ACE. ACE inhibitors block the formation of the chemical angiotensin II (AT-II) which causes muscles surrounding blood vessels to contract. Blocking the formation of AT-II leads to reduced vasoconstriction, increased vasodilation, and reduced blood pressure.
Bioavailability or pharmacokinetic studies and mechanism of action studies provide additional information so that the drug may be clinically used more effectively and safer in future studies.

Phase II Trials
Phase II trials represent the earliest trials of a drug in patients. Patients should have the disease under investigation. Patients who enter such trials represent a relatively restricted yet homogeneous population. In some areas of drug development - such as oncology, Phase II trials are categorized as Phase IIA and Phase IIB.
Phase IIA trials may include clinical pharmacology studies in patients, and more extensive or detailed pharmacokinetic and pharmacodynamic studies in patients. Phase IIB trials are controlled and represent the initial demonstration of efficacy and safety of a drug at the doses from the clinical pharmacology studies. Also of interest is to estimate the effective dose range, to characterize the dose response curve, and to estimate the minimally effective dose. Often it is difficult to distinguish between Phase IIB trials and Phase III trials, particularly in terms of objectives. The primary differences are the inclusion/exclusion criteria and the sample size.

Phase III Trials
Phase III trials may be viewed as extensions of Phase IIB trials. They are larger and the inclusion/exclusion criteria may be less restrictive than those of Phase IIB trials. For a drug to proceed to the Phase III portion of the development program, it must be deemed effective from the Phase IIB program. At this stage, effectiveness has been indicated, but not confirmed.

The primary objectives of the Phase III program are to confirm the effectiveness of the drug in a more heterogeneous population, and to collect more and longer term safety data. Information from Phase IIB, provides pilot data for the purpose of sample size determination in Phase III. For the purpose of obtaining more safety data under conditions which better approximate the anticipated clinical use of the drug, relatively large, uncontrolled, non-comparative trials may also be conducted in Phase III. Since if the drug is given approval to be marketed, it may be used in the elderly, in the renally impaired, etc., and since such patients are usually excluded from other trials, studies in special populations may also be conducted in Phase III.

The Clinical Development Plan

The clinical development plan for a new drug includes Phase I, Phase II, and Phase III trials. In viewing the types of trials within each phase of clinical development, it is obvious that the objectives of the trials describe characteristics of a drug which should be known before proceeding sequentially with subsequent clinical use. Further upon the successful completion of the trials through Phase III, sufficient information should exist for the drug to be approved to be marketed.

The drug sponsor may wish to include other trials in the clinical development plan particularly to provide a marketing `hook' for launch. Prior to finalizing the clinical development plan the drug sponsor should formulate draft labeling. The draft labeling should accommodate what is required to be said and what is desired to be said about the compound in the package insert of the marketed product. The clinical development plan then serves as a blueprint for labeling.

Basically, the labeling should communicate characteristics of the drug and give instructions for its use. Usually, the objectives of the trials described in Phase I, Phase II, and Phase III, if met in carrying out the attendant investigations, provide sufficient information to communicate the characteristics of the drug. However, since the population studied pre-market approval is likely to be more homogeneous than the user population post-market approval, and since inferences are based upon group averages, there may be insufficient information from the usual Phase I, Phase II, and Phase III program as to optimal clinical use of the drug, particularly in individual patients.

Therefore, drug sponsors may consider implementing a `Phase III 1/2' program directed more toward clinical use than toward establishing efficacy as a characteristic of the drug, which in our mind is what the typical pivotal proof of efficacy trials in Phase III do. Such a targeted program may be unnecessary if more efficient and more optimal designs and methods, such as response surface methodology, and evolutionary operations procedures are incorporated into the clinical development program as early as Phase II. In addition, being proactive in

developing an integrated data base consisting of all data collected on a compound, so that meta-analysis and other techniques may be used, should enable the drug sponsor to do a better job at labeling.

Bio statistical Aspects of a Protocol

A protocol has to be developed for each clinical trial. An important responsibility of the statistician or biostatistician assigned to the protocol is to provide its statistical content. This includes: ensuring that the objectives are clear; recommending the most appropriate design (experimental design and determination of sample size) for the condition being studied; assessing the adequacy of endpoints to address study objectives; assigning participants to protocol interventions to minimize bias; and developing the statistical analysis section. In addition it is imperative that the biostatistician provides a review of the protocol for completeness and consistency.

Background or Rationale

Sufficient information should be given in this section to set the stage for the clinical trial for which the protocol is being developed. This requires integrating the results (with references) of previous studies that have bearing on the current protocol. The section should end with a paragraph explaining why the current protocol is needed or why it is being developed.

Objective

The objective or research question of the protocol should be defined so that it is unambiguous. For example, in an investigation about the antihypertensive efficacy of drug D in some defined population, the statement: “The objective of this investigation is to assess the efficacy of drug D" is ambiguous.
It provides only general information as to the question (“Is D efficacious?").
The statement: “The objective of this investigation is to assess whether drug D is superior to placebo P in the treatment of hypertensive patients with diastolic blood pressure (DBP) between 90 and 105 mm HG for six months" is better as the hypertensive population to be treated and what is meant by efficacious in a comparative sense are specified.

However, the data or endpoint(s) upon which antihypertensive efficacy will be based is (are) not specified. DBP is stated, but how will it be measured? Using a sphygmomanometer or a digital monitor? Will DBP be measured in the sitting, standing or supine position? Further, what function of the DBP will be used? The change from baseline to the end of the treatment period? Or whether the patient

achieves a therapeutic goal of normotension (DBP $\leq$ 80 mm HG) by the end of the treatment period?
If there is more than one question or objective, one should identify which is primary versus which is secondary.

Plan of Study
The plan of study entails all that is to be done in order to enroll and treat patients, monitor the study, ensure patient safety, and collect valid data. The study population has to be specified. Design aspects of the study, including all procedures to be used in the diagnoses, treatment or management of patients must be delineated.

Study Population
Characteristics of the population of patients to be entered into the protocol must be specified. This is typically accomplished by specifying inclusion and exclusion criteria appropriate for the disease and drug under study. In specifying these criteria, one has to be cognizant that patients entered must have the disease under study, that the condition of the patient must not compromise patient safety by participating in the protocol, and that the patients entered should enable the efficacy of the drug under study to be determined (absence of masking or confounding factors).

The inclusion criteria specify the demography of the patient population, their disease characteristics, acceptable vital signs ranges, acceptable clinical laboratory tests ranges, etc. Exclusion criteria are generally the complement of the inclusion criteria, with delineation of a subset that specifically excludes patients from entry. For example, non-menopausal females who are pregnant or who do not agree to practice an acceptable form of birth control during the intervention period are excluded; as are patients who do not agree to abstain from using concomitant medications that may mask or interfere with the activity of the drug under study. Inclusion/exclusion criteria essentially define the population to be studied. Therefore they provide general descriptors of the population to which inferences from analyses of the data collected pertain.