Introduction to Survival Analysis

CF Jeff Lin, MD., PhD.

December 27, 2005

Introduction to Survival Analysis

Introduction to Survival Analysis

1. Survival Function, hazard Function
2. Censoring
3. Life Table
4. Kaplan-Meier Method
5. Log Rank Test

Introduction to Survival Analysis

Conversion of the student, the teacher, and the statistician.

1. Student: Tell me about Life and Death.
2. Teacher: The answer depends on what you want to know about it.
3. Student: How do I choose the right question?
4. Teacher: It depends ...
5. Statistician: I can tell you if you just tell me how you collected your data.

A Simple Question

1. 100 patients were admitted to hospital on Sep 7, 2000,
2. 99 patients were discharged on Sep 11, 2000,
3. 1 patient died on Sep 12, 2000,
4. What's the death rate on Sep 12, 2000?
5. $\frac{1}{100} = 1\%$?
6. $\frac{1}{1}$=100%?

The answer really depend on how you collect your data!

Example: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990

T

1. he following Figure 1 are long-term follow-up ersults of 7 subjects with lung cancer from January 1, 1980 to December 31, 1990.
2. $\times$ denote death and $\bigcirc$ dennote alive at the last visit.
3. What is t he 5-eayr survival rate?

Follow-up 7 Subjects from 1980 to 1990

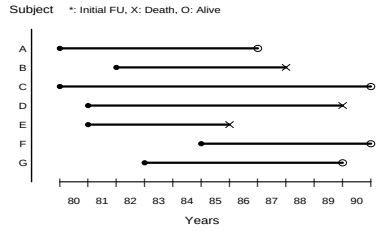

Figure 1: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990

Example: Survival Time

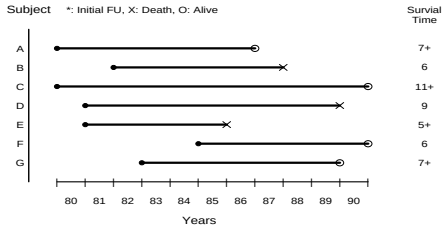

Figure 2: Follow-up Time: Pearson-Year

Example: Period Prevalence and Incidence Rate

$$\text{Period Prevalence} = \frac{3}{7} = 0.428 \quad (1)$$

$$\text{Incidence Rate} = \frac{3}{\sum(7+6+\cdots+7)} = \frac{3}{51} = 0.058 \quad (2)$$

Example: Survival Time

Figure 3: Survival Time from Time Zero

Background

1. **Survival analysis** are used to analyze the length of time between a **starting event** (entry into follow-up) and an **outcome event** (such as death).
2. Such data are often **censored**; that is, not all subjects who enter the study are followed long enough to observe the time of the outcome event.
3. In addition, such data often have highly skewed distributions.
4. Special statistical methods are required in order to analyze such data.

Background

5. More precisely, **survival analysis** is the study of the distribution of life times, and is a loosely defined statistical term that encompasses variety of statistical technique for analyzing **positive-valued random variables**.

6. Typically, the value of the random variable is the times from an **initiating event time point** to some **terminal event time point**, i.e. from time of birth (start of treatment) to death(relapse).

Background

7. Examples of this time-to-event data arise in diverse field
   (a) survival rate in medicine
   (b) Mortality in public health
   (c) Life table in epidemiology
   (d) Vital statistics in actuarial science and demography
   (e) Reliability in engineering
   (f) Event history analysis in social science
   (g) Queue process in business, unemployment in economics.

Complete Observations: One Year Study

1. In a **1 year** study of 50 animals, all survived for 1 year and 20 developed skin cancer.

2. Estimate the 1 year skin cancer incidence proportion.

3. The proportion developing skin cancer during the first year is estimated to be $\hat{p} \times 100 = 40$ percent with $s.e.(\hat{p}) = \sqrt{\hat{p}\hat{q}/n} \times 100 = \pm 6.9\%$.

Complete Observations: One Year Study

4. An approximate confidence interval can be computed based on the normal distribution.

5. An exact confidence interval can be computed using the binomial distribution.

6. Note that the estimation methods would be different if half the animals died cancer-free during the year.

Complete Observations: Life Time Study

1. In a **life-time** study of 50 animals, 20 developed skin cancer. Estimate the life-time cancer incidence proportion.

2. This is almost the same as the example above, for a 1 year study.

3. Simple proportions, as used here, are only appropriate if all subjects are followed for an equal interval of time.

4. In this case the interval of time is defined as a lifetime.

Complete Observations: Life Time Study

5. An **"equal follow-up interval"** is usually defined by a fixed time interval, such as 1 year, but can be measured in any units of time, such as lifetimes or generations.

6. Note that the average time to cancer and the median time to cancer are not meaningful for the entire population, since not all animals developed cancer during follow-up.

Complete Observations: Life Time Study

7. When all subjects are followed for an equal time, as in this lifetime example, it is sometimes useful to summarize the average and median times to event among those observed to have an event during follow-up.

8. The average time to cancer among those developing cancer during a lifetime is an interpretable statistic.

Complete Observations: Life Time Study

9. The average time to cancer among those developing cancer during a one year follow-up is less interpretable.

10. The average time to cancer among those developing cancer during a follow-up that varied between 1 and 5 years is very difficult to interpret.

Complete Observations

1. Let us consider the complete observations (all deaths observed) of an 20-animal study in the following Table 1:

Table 1: Time (months) to death data: complete observations

3.1	5.6	7.1	9.6	6.4	34.3	18.5	51.2	14.1	17.3
5.2	7.8	46.3	25.0	8.8	29.1	23.7	33.9	4.7	43.9

Complete Observations: R

```
> time.comp<-c(3.1, 5.6,  7.1,  9.6, 6.4, 34.3, 18.5,
               51.2, 14.1, 17.3, 5.2, 7.8, 46.3, 25.0,
               8.8, 29.1, 23.7, 33.9,  4.7, 43.9)
> summary(time.comp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.100   6.925  15.700  19.780  30.300  51.200
> sd(time.comp)
[1] 15.35344
```


Complete Observations: R

```
> stem(time.comp)

  The decimal point is 1 digit(s) to the right of the
  0 | 35566789
  1 | 0479
  2 | 459
  3 | 44
  4 | 46
  5 | 1

> plot.density(density(time.comp))
```


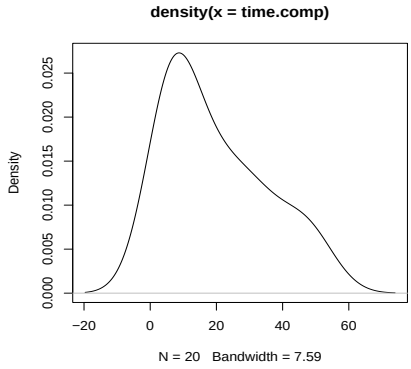

Figure 4: Density Plot: Complete Observations of Survival Time

Complete Observations

2. The average time from entry to death is 19.7 months (S.D.= 15.3).
3. The time range from 3.1 to 51.2 months.
4. Half of the subjects were dead below 15 months. The distribution of time is not normal distributed and is **highly skew**.

Censored Data

Censored Data

1. **Censored data** arise from losses to follow-up and from varying follow-up intervals.
2. Censored data make it more difficult to compute interpretable summaries.
3. How would you compute the 5 year death fraction based on the following outcomes from a 5 year study of 50 subjects in the following Table 2 ?

Censored Data

Table 2: Time (months) to death data with censored observations

Number	Observed outcome
10	Drop-out alive before 5 years
5	die during study
35	alive at 5 years

Censored Data

4. What do we know about the 10 subjects who were lost to follow-up would have died within 5 years, had they didn't dropped out.
5. If all study subjects are followed for a fixed equal period of time, then an event proportion (risk) can be estimated for that interval with no ambiguity.
6. When subjects are followed for differing lengths of follow-up, it is often more appropriate to estimate an event rate.

Constant Event Rate: Person-Year

1. Suppose 100 subjects (initially cancer-free) were followed until lung cancer or loss to follow-up, whichever came first.
2. Suppose that the average follow-up interval was 4.675 years and 3 incident primary cancers were observed.
3. The total person-years of follow-up was 467.5 person-years.
4. The incidence rate is estimated to be $3/467.5$, or 6.417 per 1000 person-years.

Constant Event Rate: Person-Year

5. A confidence interval for the rate can be computed using the **Poisson distribution**.
6. In this example, we have assumed that the **incidence rate is constant** throughout the study period (and is estimated to be 6.417 per 1000 person-years).
7. The assumption of a **constant event rate** may not be plausible or may be inconsistent with the data in many studies.

Survival Functions and Hazard Function

Basic Survival Functions

1. It is often useful to summarize the survival experience of a study group.
2. The summary is especially useful if the study group is representative of a larger population.
3. The survival experience of the study group is an estimate of the survival experience of the wider population.
4. If the study group is a random sample from the target population, then probabilistic measures of the accuracy of the estimate can be computed.

Basic Survival Functions

Survival analysis methods are tailored to work well with the specific characteristics of the data and the specific objectives that arise in survival studies.

Basic Survival Functions

1. Often, survival data are distinguished from other types of data because they are censored.
   - Censored data (those observations whose times to event we do not get to observe completely) prevent the use of standard methods of statistical summarization and inference.
   - In particular, right censored data are reported as lower bounds for the actual unobserved event times.

Basic Survival Functions

2. We are often interested in the whole distribution of survival times.
   - Survival times often have a distribution in the population that is very different from a Gaussian (Normal) distribution.
   - Many standard approximate statistical methods are not accurate for such data.
   - Many standard statistical methods are instead oriented towards inference for the mean survival time, μ and standard deviation σ.

Basic Survival Functions

3. Considerations for survival distribution
   - The extremes of the distribution of times to event (extreme quantiles) are often of interest in survival analysis.
   - For example, many people hope that they will live to the 95^{th} percentile, rather than the 50^{th} percentile.
4. The rate of occurrence of events per unit time is often of interest in survival analysis.

Survival Functions

1. Let X be the time from a well-defined time point zero to a well-defined time point when some specified event occurs.
2. We deal with a single nonnegative random variable, X.
3. Let $X \geq 0$ and $f(X)$ be the probability density (mass) function.
4. **Probability Density Function (p.d.f) of** X is

$$f(X = x) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x)}{\Delta x} = \frac{dF(x)}{dx} \tag{3}$$

with $\int_0^\infty f(x)dx = 1$.

Survival Functions

5. The range of X is $[0, \infty]$, and this should be understood as the domain of definition for function of x.
6. **Survival Function** is the probability of an individual surviving beyond time x (experiencing the event after time x).

$$S(x) = Pr(X > x) = \int_x^\infty f(t)\, dt \tag{4}$$

Survival Functions

7. In the context of equipment item failures, $S(x)$ is referred to as the reliability function.
8. $S(x_1) - S(x_2)$ is the fraction of the population that dies between ages x_1 and x_2 for $x_1 < x_2$.
9. Survival functions are monotone, decreasing (nonincreasing) functions equal to one at zero and zero at the time approaches infinity.
10. $S(0) = 1$, if the every member of the population eventually has an event, then $S(\infty) = 0$.

Survival Functions

11. If some member of the population never have the event, then it is possible that the survival curve does not approach 0 as time increase.
12. The notation dealing with this is not standardized, but one practical implication is that a survival curve estimate need not reach 0 by the end of follow-up.
13. When X is a continuous random variable, the survival function is the complement of the cumulative distribution function.

Failure Functions

1. **Failure Function** is the cumulative distribution.

$$F(x) = Pr(X \leq x) = 1 - S(x) \tag{5}$$

$$\begin{aligned} f(x) &= \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x)}{\Delta x} = \frac{dF(x)}{dx} && (6) \\ &= -\frac{dS(x)}{dx} = \lim_{\Delta x \to 0} \frac{S(x) - S(x + \Delta x)}{\Delta x} && (7) \end{aligned}$$

Failure Functions

2. $f(x)dx \approx$ fraction who die between age x and $x + \Delta x$ when Δx is a short interval of time.
3. The density is positive.
4. $\int_0^\infty f(t)dt = 1$.

Discrete Survival Functions

1. When X is a discrete random variable, different techniques are required.
2. Suppose that X take on values $x_j, j = 1, 2, \ldots, n$, with probability mass function (p.m.f) $p(x_j) = Pr(X = x_j)$, where $x_1 < x_2 < \ldots < x_n$.
3. **Survival Function 2** for a discrete random variable X is

$$S(x) = Pr(X > x) = \sum_{x_j > x} p(x_j) \tag{8}$$

where $S(0) = 1$ and $p(x_j) = S(x_{j-1}) - S(x_j)$.

Hazard Function

A fundamental in survival analysis is the **hazard function**.

Hazard Function

This function is also known as

1. The **hazard rate** in survival analysis
2. The **conditional failure rate** in reliability
3. The **force mortality** in demography
4. The **intensity function** in stochastic processes
5. The **age-specific failure rate** in epidemiology
6. The **inverse of Mill's ratio** in economics

Hazard Function (Hazard Rate)

1. Let X is a continuous random variable.
2. **Hazard Function, (Hazard Rate)** is conditional probability that specifies the instantaneous rate of failure at $X = x$ conditional upon survival to time x, and is defined as

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x | X \geq x)}{\Delta x} && (9) \\ &= \frac{f(x)}{S(x)} && (10) \\ &= -\frac{d}{dx} \ln[S(x)] && (11) \\ f(x) &= h(x)S(x) && (12) \end{aligned}$$

Hazard Function (Hazard Rate)

3. Note that **death rates** are generally reported among those still surviving, and are the same as the hazard function.
4. The concept of the hazard function has been discovered in many field has many names.
5. This function is known as conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic process, the age-specific failure rate in epidemiology. The inverse of the Mill's ratio in economics, or simply as the hazard rate.

Cumulative Hazard Function

1. **Cumulative Hazard Function** is defined

$$H(x) = \int_0^x h(u)\, du = -\ln[S(x)] \tag{13}$$

2. Thus, for continuous survival time,

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)\, du\right] \tag{14}$$

Hazard Function

1. Hazard function is particularly useful in determining the appropriate failure distribution utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time.
2. There are many general shapes for the hazard rate.

Hazard Function

3. The only restriction on $h(x)$ is that it be nonnegative, i.e.,

$$h(x) \geq 0. \tag{15}$$

4. One may believe that the hazard rate for the occurrence of a particular event is increasing, decreasing, constant, bathtub-shaped, hump-shaped or possessing some other characteristic which describes the failure mechanism.

Hazard Function

5. $H(x)$ is the expect number of events when following a single person to time x, with replacement at death.
6. It is easy to estimate $S(x)$.
7. This makes it easy to examine the shape of $H(x)$ graphically, which tells us about the hazard function as the slope of $H(x)$.

Discrete Hazard Function

When X is a discrete random variable, **Hazard Function 2**, for discrete hazard function is defined as

$$h(x) = Pr(X = x_j | X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})} \quad j = 1, 2, \ldots \tag{16}$$

$$h(x_j) = \frac{S(x_{j-1}) - S(x_j)}{S(x_{j-1})} = 1 - \frac{S(x_j)}{S(x_{j-1})} \tag{17}$$

$$S(x_{j-1}) \times h(x_j) = S(x_{j-1}) - S(x_j) \tag{18}$$

$$S(x_j) = S(x_{j-1})[1 - h(x_j)] \tag{19}$$

Discrete Hazard Function

1. Thus, for discrete survival time, the survival function is the product of conditional survival probability as

$$S(x) = \sum_{x_j > x} p(x_j) = \prod_{x_j \le x} \left[1 - h(x_j)\right] = \prod_{x_j \le x} \frac{S(x_j)}{S(x_{j-1})} \qquad (20)$$

Discrete Hazard Function

2. And the cumulative hazard function for discrete random variable is

$$H(x) = \sum_{x_j \le x} \ln[1 - h(x_j)] \qquad (21)$$

$$\cong \sum_{x_j \le x} h(x_j); \quad \text{if } h(x_j) \text{ is small for } j = 1, 2, \ldots \qquad (22)$$

Discrete Hazard Function

3. The equation (21) is based on the relationship for continuous lifetimes $S(x) = \exp[-H(x)]$ will be preserved for discrete lifetimes.

4. The equation (22) is directly estimable from a sample of censored or truncated lifetimes and the estimator has a very desirable statistical properties, however, the relationship $S(x) = \exp[-H(x)]$ for the equation (22) no longer holds true.

Continuous Hazard Function

1. For a continuous lifetimes, the failure distribution is said to have an increasing failure rate (IFR) property, if the hazard function $h(x)$ is nondecreasing for $x \ge 0$, and an increasing failure rate on the average (IFRA), if the ratio of the cumulative hazard function to time $H(x)/x$ is nondecreasing for $x > 0$.

2. For a continuous lifetimes, the failure distribution is said to have a decreasing failure rate (DFR) property, if the hazard function $h(x)$ is nonincreasing for $x \ge 0$.

Hazard Function

- We will work with both continuous and discrete survival functions.
- In practice the distinction between continuous and discrete survival function is not very important.
- The distinctions require very different notations.

Censoring

1. The three basic requirements for measuring failure time are time origin, scale for measuring the passage of time and meaning of the point event.

2. The time origin should be precisely defined.

3. The time origin need not be and usually is not at the same calendar time.

Censoring

4. Most randomized clinical trials have staggered entry, so time origin is usually his own date of entry.
5. The scale of measuring time is often clock time (real time), although other possibility certainly arise, such as operating time of a system, mileage of a car.
6. The meaning of point event of failure must be defined precisely such as death.

Censoring

7. The tools of survival analysis are designed to yield inferences about the distribution of the times to event, X, (lifetime) in a population.
8. A special source of difficulty in survival analysis is that some individuals may not be observed for the full time to failure.

Censoring

9. Some lifetimes are known to have occurred only within certain interval,.
10. Such **incomplete observation** of the failure time is called **censoring**.
11. Censoring is a point event and that the period of observation for censored individuals must be recorded.
12. In practice, we often do not observe X for a random sample, but only known that X lies in an observed interval (L, R) (**interval censoring**) or might only observe a subject conditional on certain conditions (truncation).

Censoring

13. Formally, an observation is said to be **right censored** at time R if the exact value of the observation is not known but only that it is greater than or equal R.
14. Similarly, an observation is said to be **left censored** at time L if it is known only that the observation is less than or equal to L.
15. Right censoring is very common.
16. We use different notation for the observed data to clarify that it is different from the measure, X, that we are interested in.

Life Table Method (Actuarial Method) Kaplan-Meier Method (product-Limit Method

Life Table Method (Actuarial Method)

1. The classical method of estimating $S(t)$ in epidemiology and actuarial science is the actuarial method or life table method.
2. Early methods for estimating survival functions were developed by
   - Berkson, J. and Gage, R.P.: Calculation of survival rates for cancer. Proc. Staff Meet. Mayo Clin. 25: 270-286 (1950).
   - Cutler, S.J. and Ederer, F.: Maximum utilization of the life table method in analyzing survival. J. Chronic Dis. 8: 699-712 (1958).

Life Table Method (Actuarial Method)

3. The same concepts are used in modern calculation methods.
4. The resulting survival curves are often referred to as actuarial curves because they are analogous to those used by actuaries.

Life Table Method (Actuarial Method)

5. However, there is an important difference between survival curves and actuarial curves.
6. Survival curves are based on data from a longitudinal study of subjects through time and are used to summarize what happened to the study group through their lifetimes.
7. Actuarial curves are more typically based on age-specific death rates observed during a short calendar interval.
8. Actuarial methods combine current age-specific death rates from several age-cohorts of subjects to forecast lifetime experience for new cohorts.

Life Table Method (Actuarial Method)

9. The early methods of Berkson and Gage were developed to be used when the time axis is grouped into intervals and the numbers of subjects dying or lost to follow-up in each interval are recorded.
10. Time intervals of length 1 to 5 years were commonly used.
11. The methods based on grouping were useful for hand calculation and for illustration, but are less widely used now that computers can calculate estimates based on the recorded survival times.

Life Table Method (Actuarial Method)

12. Now, survival times are commonly computed based on dates of entry and death.
13. Sometimes, only the month is available in the original data source and the day of an event is not recorded for some subjects.
14. If the time to event typically takes months, rather than days, then dates can be imputed for these missing values with little effect on the resulting estimates.

Life Table Method (Actuarial Method)

15. The basic construction of life table introduces notation of hazard function, density function, and survival function.
16. A cohort is a group of individual who have some common origin from which the event time will be calculated.
17. They are followed over time and their event time or censoring time is recorded to fall in one of $s+1$ adjacent, nonoverlapping intervals.
18. A traditional cohort life table presents the actual mortality experience of the cohort from the birth of each individual to the death of the last surviving member of the cohort.

Life Table Method (Actuarial Method)

19. Let time be partitioned into a fixed sequence of $s+1$ intervals, $I_1, I_2, \ldots, I_s, I_{s+1}$. These intervals are adjacent, nonoverlapping.
20. These intervals are almost always, but not necessarily, of equal lengths, and for human populations the length of each interval is usually one year. Deaths, losses, and withdraws are counted for each time interval.
21. We use the notation introduced by the notation introduced by Gehan, E.A. Estimating survival functions from the life table, Journal of Chronic Disease, 21: 629-644 (1969).

Life Table Method (Actuarial Method)

The basic construction of the cohort life table is described below:

1. $I_i = [t_{i-1}, t_i)$ denotes the i'th interval of follow-up time that includes all time t satisfying $T_{i-1} < t \leq t_i$. Define $T_0 = 0$ and $T_{s+1} = \infty$.
2. There are $s+1$ intervals including the last one of infinity length.
3. $t_{mi} = (t_i + t_{i-1})/2$ is the midpoint of i'th interval $I_i = [t_{i-1}, t_i)$. These times are used for plotting hazard and density functions.
4. $z_i = (t_i - t_{i-1})$ is the length (width) of the i'th interval.

Life Table Method (Actuarial Method)

5. $n_i^{'}$ is the number of individuals being followed at the beginning of the i'th interval, that is the effective sample size.
6. d_i is the number of individuals dying during the i'th interval.
7. l_i is the number of individuals loss to follow-up during the i'th interval.
8. For example,individuals who move away during the interval and whose mortality status cannot subsequently be ascertained are lost to follow-up.

Life Table Method (Actuarial Method)

9. w_i is the number of individuals who are withdrawn alive during the i'th interval because of the close of the study.
10. These counts enter into the computations exactly like l_i counts.
11. $n_i = n_i^{'} - [(l_i + w_i)/2]$ is the number of individuals expected to be at risk for death, on average, during the i'th interval (at its midpoint).
12. Note: most of the time, $l_i + w_i$ are censored, assuming that censoring times are uniformly distributed over the interval.
13. $n_i^{'} = n_{i-1}^{'} - (d_{i-1} + l_{i-1} + w_{i-1})$.

Life Table Method (Actuarial Method)

We can break up the survival probability $S(t_i)$ into a product of probabilities as

$$\begin{aligned} S(t_i) &= Pr[T > t_i] && (23)\\ &= Pr[T > t_1]\, Pr[T > t_2 \mid T > t_1] \cdots Pr[T > i_i \mid T > t_{i-1}] \\ &= p_1 \cdot p_2 \cdots p_i && (24) \end{aligned}$$

$$\text{where } p_i = Pr[T > t_i \mid T > t_{i-1}] \quad (25)$$

Life Table Method (Actuarial Method)

The basic notations in construction of life table are described as below:

1. conditional death function is estimated as

$$\hat{q}_i = \frac{d_i}{n_i} = Pr[\text{ dying during } I_i \mid \text{surviving beyond } I_{i-1}] \quad (26)$$

2. That is the estimated condition probability of dying during the i'th interval, given survival beyond $(i-1)$'th interval.

Life Table Method (Actuarial Method)

3. Conditional survival function is estimated as

$$\hat{p}_i = 1 - \hat{q}_i = Pr[\text{ surviving beyond } I_i \mid \text{surviving beyond } I_{i-1}] \quad (27)$$

4. That is the estimated condition probability of surviving through the i'th interval, given survival beyond $(i-1)$'th interval.
5. $\hat{P}_i$ is the cumulative proportion surviving to the beginning of the i'th interval, t_{i-1}, the estimated survival function for the individuals who survive beyond time t_{i-1}.

Life Table Method (Actuarial Method)

6. $\hat{P}_i$ is often denoted as the survival function at time t_{i-1} as

$$\hat{S}_{i-1} = \hat{P}_i = \hat{p}_{i-1} \times \hat{P}_{i-1} \tag{28}$$

$$\hat{S}_i = \hat{S}_{i-1} \times p_i \tag{29}$$

Life Table Method (Actuarial Method)

7. Density function is estimated as

$$\hat{f}_i = \hat{f}(t_{mi}) = \frac{\hat{P}_i - \hat{P}_{i+1}}{t_i - t_{i-1}} = \frac{\hat{P}_i\,\hat{q}_i}{z_i} = \frac{\hat{S}_{i-1} - \hat{S}_i}{z_i} \tag{30}$$

8. The density is the probability of dying during an interval per unit time.

Life Table Method (Actuarial Method)

9. Hazard function is estimated as

$$\hat{h}_i = \hat{h}(t_{mi}) = \frac{\hat{f}(t_{mi})}{\hat{P}(t_{mi})} \tag{31}$$

$$\text{where } \hat{P}(t_{mi}) = \frac{\hat{P}_{i+1} + \hat{P}_i}{2} = \frac{\hat{P}_i(1+\hat{p}_i)}{2} \tag{32}$$

$$\text{so } \hat{h}(t_{mi}) = \frac{2\,\hat{f}(t_{mi})}{\hat{P}_{i+1} + \hat{P}_i} = \frac{2\,\hat{q}_i}{(t_i - t_{i-1})(1+\hat{p}_i)} \tag{33}$$

Life Table Method (Actuarial Method)

1. The actuarial method gives an estimate for each p_i separately and then multiplies the estimates together to estimate $S(t_k)$.

2. The actuarial estimate is

$$\hat{S}(t_k) = \prod_{i=1}^{i=k} \hat{p}_i \tag{34}$$

Table 3: Survival Analysis: Life Table

Interval	Num Mid Point	Width	Num Enter Int	Num Lost Follow	Num With-Drawn	Num Exp Risk	Num Die	Cond Prop Die	Cond Prop Surv	Cum Prop Surv	$f(t_{mi})$	$\hat{\lambda}(t_{mi})$
$[t_0, t_1)$	t_{m1}	h_1	n_1'	l_1	w_1	n_1	d_1	$\hat{q}_1$	$\hat{p}_1$	$\hat{P}_1 = 1.0$	$f(t_{m1})$	$\hat{\lambda}(t_{m1})$
$[t_0, t_2)$	t_{m2}	h_2	n_2'	l_2	w_2	n_2	d_2	$\hat{q}_2$	$\hat{p}_2$	$\hat{P}_2$	$f(t_{m2})$	$\hat{\lambda}(t_{m2})$
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
$[t_0, t_s)$	t_{ms}	h_s	n_s'	l_s	w_s	n_s	d_s	$\hat{q}_s$	$\hat{p}_s$	$\hat{P}_s$	$f(t_{ms})$	$\hat{\lambda}(t_{ms})$
$[t_s, t_\infty)$												
					Example Data							
$[0,1)$	0.5	1	913	19	77	865.0	312		0.639	1.000	0.361	0.441
$[1,2)$	1.5	1	505	3	71	468.0	96		0.795	0.639	0.131	0.228
$[2,3)$	2.5	1	335	4	58	304.0	45		0.852	0.508	0.075	0.160
$[3,4)$	3.5	1	228	3	27	213.0	29		0.864	0.433	0.059	0.146
$[4,5)$	4.5	1	169	5	35	149.0	7		0.953	0.374	0.018	0.048
$[5,6)$	5.5	1	122	1	36	103.5	9		0.913	0.356	0.031	0.091
$[6,7)$	6.5	1	76	0	17	67.5	3		0.956	0.325	0.014	0.045
$[7,8)$	7.5	1	56	2	10	50.0	1		0.980	0.311	0.006	0.020
$[8,9)$	8.5	1	43	0	8	39.0	3		0.923	0.305	0.024	0.080
$[9,\infty)$	-	-	32	-	-	32.0	32		0.000	0.281	-	-

Life Table Method (Actuarial Method)

3. To estimate variance of $S(t_i)$, we use Greenwood's Formula.

$$\mathbf{Var}[\hat{S}(t_k)] = \hat{S}^2(t_k) \sum_{i=1}^{i=k} \frac{d_i}{n_i\,(n_i - d_i)} \tag{35}$$

$$(1-\alpha)100\%\ \text{C.I.} \qquad \hat{S}(t_k) \pm z_{1-\alpha/2}\ \text{s.e}\ \{S(t_k)\} \tag{36}$$

4. One difficulty with this procedure arises from the fact that the confidence intervals are symmetric.

5. When the estimated survival function is close to zero or unity.

6. The survival function that lie outside the interval $(0,1)$.

Life Table Method (Actuarial Method)

7. Another better transformation is

$$\mathbf{Var}[\log(-\log \hat{S}(t_k))] \cong \frac{1}{[\log \hat{S}(t_k)]^2} \sum_{i=1}^{i=k} \frac{d_i}{n_i\,(n_i - d_i)} \tag{37}$$

8. $(1-\alpha)100\%$ C.I. of is

$$\hat{S}(t_k)\ [\hat{S}(t_k)]^{\exp[\pm z_{1-\alpha/2}\text{s.e. }(\log(-\log \hat{S}(t_k)))]} \tag{38}$$

Life Table Method (Actuarial Method)

9. The standard error of $\hat{f}_i = \hat{f}(t_{mi})$ is

$$\text{s.e. } (f_i) \approx \frac{\hat{S}_{i-1}\hat{q}_i}{(t_i - t_{i-1})} \sqrt{\left[\sum_{j=1}^{i-1} \frac{\hat{q}_i}{n_i \hat{p}_i}\right] + \left[\frac{\hat{p}_i}{n_i \hat{q}_i}\right]} \tag{39}$$

Life Table Method (Actuarial Method)

10. Broadly speaking, the behavior of the life table estimates is acceptable under random independent censorship provided that censoring is fairly evenly distributed across individuals and not too heavy, intervals are not too wide, and sample sizes are not too small.

11. It is nevertheless wise to remember that properties depend on the censoring and lifetime distributions at hand, that estimates of survival probabilities will be slightly biased, and that the adequacy of the variance estimate is not fully known unless censoring is very light.

Life Table Method (Actuarial Method)

1. A **cohort life table** is a group of people who are followed through-out the course of the study. Clinical life table follows the same population time.

2. The people at risk at the beginning of the interval I_i are those people who survived (not dead, lost, or withdraw) the previous interval I_{i-1}.

Life Table Method (Actuarial Method)

3. Another type of life table is the current life table.

4. **Current life table** starts with large number people, time is age, follows up short period, for example one year.

5. In a current life table a group of people with age t_{i-1} are considered to be at risk at the beginning of the interval $I_i = (t_{i-1}, t_i]$, and this group of people is completely different from those at risk in the previous interval I_{i-1}.

Life Table Method (Actuarial Method)

6. Typically, different age groups in the population are followed at time same time.

7. Inference bases on current population at short calendar time.

8. For example, $\hat{S}(\text{age} = 40)$, estimated survival rate at age 40 years old based current life table is quite different for individuals who is age zero (just birth) in current life table. Those age zero individuals are expected surviving longer.

Kaplan-Meier (Product-Limit) Method

Kaplan-Meier (Product-Limit) Method

1. Consider the following example with 10 subjects, we have survival time (censoring time) of

$$1^+, 3, 4^+, 5, 5, 6^+, 7, 7, 7^+, 8^+ \tag{40}$$

2. The "+" sign represent censored time of the observations.

Kaplan-Meier (Product-Limit) Method

1. With right censored data, we know which observations are still being followed and we can observe which of them have an event.
2. Now, we are concerned with how many have an event (for estimating the survival curve) rather than which ones have an event.
3. What we know is the number of subjects being followed and the number with an event at each moment in the follow-up.

Kaplan-Meier (Product-Limit) Method

4. Denote a data set with n of observations including right censored observation with follow-up times and status indicators by (x_i, δ_i), for $i = 1, 2, \ldots, n$.
5. This information can be organized several ways.
6. For example in life table method, the information could be grouped according to a division of time axis into disjoint subintervals.

Kaplan-Meier (Product-Limit) Method

7. In life table method, the grouping of data into time intervals does not retain all of the information in the origin data set.
8. All the information is kept by recording the information at each time point (as function of time, rather than in a table).

Kaplan-Meier (Product-Limit) Method

9. The product-limit (PL) estimator, proposed by Kaplan and Meier (1958), is similar to the actuarial estimator except the lengths of the intervals I_i, be the i'th ordered censored or uncensored observation.
10. The product-limit estimator has intervals determined by the data.
11. Intervals can be though of as very short, or as each containing just one type of data observation.

Kaplan-Meier (Product-Limit) Method

1. Let $X_1, X_2, X_3, \ldots, X_n$ be independently identical distributed (i.i.d.) each with density function F, and survival function S.
2. However, censoring time are often effectively random.
3. Sometimes, individuals will experience some other competing event of interest which causes them to be removed from the study.
4. Some events which cause the individual to be randomly censored, with respect to event of interest, are accident death, migration of human population.

Kaplan-Meier (Product-Limit) Method

5. Let $C_1, C_2, \ldots, C_n$ be i.i.d. each with distribution function G.
6. C_i is the censoring time associated with T_i. We can only observe $(T_1, \delta_1), (T_2, \delta_i), \ldots, (T_n, \delta_n)$ where

$$T_i = \min(X_i, C_i) = X_i \wedge C_i \tag{41}$$

$$\delta_i = \mathsf{I}(X_i \leq C_i) \begin{cases} 1 & \text{if } X_i \leq C_i, \text{ that is, } T_i \text{ is not censored,} \\ 0 & \text{if } X_i > C_i, \text{ that is, } T_i \text{ is censored.} \end{cases} \tag{42}$$

Kaplan-Meier (Product-Limit) Method

(a) Consider the following example with 10 subjects, we have survival time (censoring time) of

$$1^+, 3, 4^+, 5, 5, 6^+, 7, 7, 7^+, 8^+ \tag{43}$$

(b) The "+" sign represent censored time of the observations.

$$\begin{array}{rcccccccccc} T_i & = & 1 & 3 & 4 & 5 & 5 & 6 & 7 & 7 & 7 & 8 \\ C_i & = & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \end{array} \tag{44}$$

Kaplan-Meier (Product-Limit) Method

7. We observe the pairs of data as $(t_1, \delta_1), (t_2, \delta_2), (t_i, \delta_i), \ldots, (t_n, \delta_n)$, for $i = 1, 2, \ldots, n$.
8. Let $t_{(1)} < t_{(2)}, \ldots, t_{(n)}$ be the order statistics of $t_1, t_2, \ldots, t_n$.
9. Define $\delta_{(i)}$ to be the value of δ associated with $t_{(i)}$.

Kaplan-Meier (Product-Limit) Method

(a) Consider the following example with 10 subjects, we have survival time (censoring time) of

$$1^+, 3, 4^+, 5, 5, 6^+, 7, 7, 7^+, 8^+ \tag{45}$$

(b) The "+" sign represent censored time of the observations.

$$\begin{array}{rcccccccccc} T_i & = & 1 & 3 & 4 & 5 & 5 & 6 & 7 & 7 & 7 & 8 \\ C_i & = & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \end{array} \tag{46}$$

Kaplan-Meier (Product-Limit) Method

10. We can consider the interval $t_{(i)} - t_{(i-1)} \to 0$.
11. If there are ties in the observed t_i values, then order the observations with respect to δ_i as well.
12. That is $(t, 0) > (t, 1)$.
13. If there are no tied values of t_i, then for short enough intervals, there will be at most one t_i in any interval in the limit.

Kaplan-Meier (Product-Limit) Method

14. The traditional approach of Kaplan-Meier estimator (product-limit estimator) is based on order statistics.
15. We first define risk set at time t, $\mathbb{R}(t)$, which is the set of subjects still alive at time $t-$ (just before time t).
16. That is, the indices of the subjects still alive and uncensored (still in the study) at time t.

Kaplan-Meier (Product-Limit) Method

The Kaplan-Meier estimate for the survival curve

$$
\begin{aligned}
n_i &= \text{in } \mathbb{R}(t) = \text{ alive at time } t- && (47)\\
d_i &= \text{ died at time } t_{(i)} && (48)\\
p_i &= Pr[\text{ surviving through } I_i \mid \text{alive at the beginning of } I_i \\
&= Pr[T > t_{(i)} | T > t_{(i-1)}] && (49)\\
q_i &= 1 - p_i && (50)\\
\hat{q}_i &= \frac{d_i}{n_i} && (51)\\
\hat{S}(t) &= \prod_{t_{(i)} \le t} \hat{p}_i = \prod_{t_{(i)} \le t} (1 - \hat{q}_i). && (52)
\end{aligned}
$$

Kaplan-Meier (Product-Limit) Method

17. It is also called the product-limit estimator.
18. Note that it is the same as the life table estimator when the intervals of time are taken to be arbitrarily (thus the term limit above).
19. The product-limit estimator is a step function with jumps at the observed event times.
20. The size of these jumps depends not only on the number of events observed at each event time $t_{(i)}$, but also on the pattern of the censored observations prior to time $t_{(i)}$.

Kaplan-Meier (Product-Limit) Method

21. The variance of the product-limit estimator is commonly estimated by **Greenwood's formula**.
22. The Greenwood's variance estimator of $\hat{S}(t)$ is (originally based on $\widehat{\text{Var}}[\log(\hat{S}(t))]$) as

$$
\begin{aligned}
\widehat{\text{Var}}[\log(\hat{S}(t))] &\approx \sum_{t'_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} && (53)\\
\widehat{\text{Var}}[\hat{S}(t)] &\approx \hat{S}^2(t) \sum_{t'_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} && (54)
\end{aligned}
$$

Kaplan-Meier (Product-Limit) Method

23. An approximate $(1-\alpha)100\%$ confidence interval for $\hat{S}(t)$ is

$$(1-\alpha)100\% \text{ C.I. of } S(t): \quad \hat{S}(t) \pm z_{1-\alpha/2} \text{ s.e. } [\hat{S}(t)] \qquad (55)$$

where the s.e. is the square root of the Greenwood variance formula.

Kaplan-Meier (Product-Limit) Method

24. A better confidence interval is based on approximate variance $[v(t)]$ of $\log(-\log \hat{S}(t))$

$$
\begin{aligned}
&[v(t)] = \textbf{Var}(\log(-\log(\hat{S}(t)))) \\
&\approx \frac{1}{[\sum_{t'_{(i)} \le t} \log(\frac{n_i - d_i}{n_i})]^2} \sum_{t'_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} && (56)
\end{aligned}
$$

25. $(1-\alpha) \times 100\%$ C.I.

$$[\hat{S}(t)]^{\exp(+\Delta)} < S(t) < [\hat{S}(t)]^{\exp(-\Delta)}] \qquad (57)$$

where $\Delta = z_{1-\alpha/2} \sqrt{(\hat{v}(t))}$.

Kaplan-Meier (Product-Limit) Method

26. The justification for these formulas is not as clear as in the case of life tables because the number of terms in the product is random and there is more dependence between terms.

27. However, they can be justified as approximations to the asymptotic variance of $\hat{S}(t)$.

Example: Kaplan-Meier (Product-Limit) Method

1. Consider the following example with 10 subjects, we have survival time (censoring time) of

$$1^+, 3, 4^+, 5, 5, 6^+, 7, 7, 7^+, 8^+ \tag{58}$$

2. The "+" sign represent censored time of the observations.

$$\begin{array}{rcccccccccc} T_i & = & 1 & 3 & 4 & 5 & 5 & 6 & 7 & 7 & 7 & 8 \\ C_i & = & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \end{array} \tag{59}$$

Kaplan-Meier (Product-Limit) Method

t_i	observed survival time
d_i	the number of events observed at time t_i
n_i	the number of individuals still under observation just before time t_i
q_i	the fraction of the n individuals who do have an event at time t_i, i.e. d_i/n_i
p_i	the fraction of the n individuals who do not have an event at time t_i, i.e. $(n_i - d_i)/n_i$
$S(t_i)$	the KM estimate of the survival function at time t_i
$s.e.$	the approximate standard error of $S(t_i)$

Example: Kaplan-Meier (Product-Limit) Method

Table 4: Survival Analysis: Kaplan-Meier Method

t_i	d	n	q_i	p_i	$S(t) = Pr(T > t)$	$s.e.$
0	0	10	0	1.0	1.0	-
3	1	9	1/9	$8/9 \approx 0.89$	0.889	0.104
5	2	7	2/7	$5/7 \approx 0.71$	0.634	0.169
7	2	4	2/4	$2/4 = 0.5$	0.317	0.179

Example: Kaplan-Meier (Product-Limit) Method

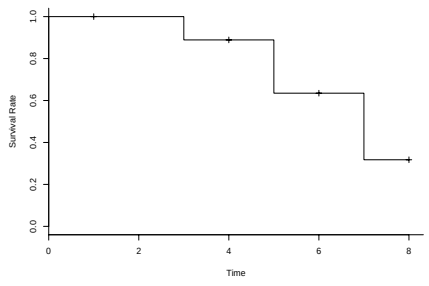

Figure 5: Kaplan-Meier Survival Curve

Example: Kaplan-Meier (Product-Limit) Method

```
> exkm.time <- c(1,3,4,5,5,6,7,7,7,8)
> exkm.censor<-c(0,1,0,1,1,0,1,1,0,0)
> data.frame(exkm.time,exkm.censor)
   exkm.time exkm.censor
1          1           0
2          3           1
3          4           0
4          5           1
5          5           1
6          6           0
7          7           1
8          7           1
9          7           0
10         8           0
```

Example: Kaplan-Meier (Product-Limit) Method

```
> survfit(Surv(exkm.time,exkm.censor))
Call: survfit(formula = Surv(exkm.time, exkm.censor))
      n  events  median 0.95LCL 0.95UCL
     10       5       7       5     Inf

> summary(survfit(Surv(exkm.time,exkm.censor),
   type=c("kaplan-meier"),error=c("greenwood"),
   conf.type=c("plain")))
Call: survfit(formula = Surv(exkm.time, exkm.censor),
                      type = c("kaplan-meier"),
    error = c("greenwood"), conf.type = c("plain"))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3      9       1    0.889   0.105        0.684        1.000
    5      7       2    0.635   0.169        0.303        0.967
    7      4       2    0.317   0.180        0.000        0.670
```


Example: Kaplan-Meier (Product-Limit) Method

```
plot(survfit(Surv(exkm.time,exkm.censor),  conf.type="none"),
     bty="l", xlab="Time", ylab="Survival Rate")
```


Example: Kaplan-Meier (Product-Limit) Method

```
data ex;
input time censor @@;
cards;
1  0   3  1   4  0   5  1
5  1   6  0   7  1   7  1
7  0   8  0

run;
proc lifetest  method=km;
    time time*censor(0);
    run;
```


Kaplan-Meier (Product-Limit) Method

We now consider another example of the remission duration for acute leukemia that 21 children were acute leukemia and were treated with 6-MP, the data are as following

$$10, 7, 32^+, 23, 22, 6, 16, 34^+, 32^+, 25^+, 11^+, 20^+,$$
$$19^+, 6, 17^+, 35^+, 6^+, 13, 9^+, 6^+, 10^+ \quad (60)$$

Example: Kaplan-Meier (Product-Limit) Method

```
> AML01.time<- c(10,7,32,23,22,6,16,34,32,25,11,20,
        19,6,17,35,6,13,9,6,10)
> AML01.censor<-c(1,1, 0, 1, 1,1, 1, 0, 0, 0, 0, 0,
         0,1, 0, 0,0, 1,0,0, 0)
> data.frame(AML01.time,AML01.censor)

> survfit(Surv(AML01.time,AML01.censor))
Call: survfit(formula = Surv(AML01.time, AML01.censor))

      n  events  median 0.95LCL 0.95UCL
     21       8      23      16     Inf
```


Example: Kaplan-Meier (Product-Limit) Method

```
> summary(survfit(Surv(AML01.time,AML01.censor),
  type=c("kaplan-meier"),
  error=c("greenwood"),conf.type=c("plain")))
Call: survfit(formula = Surv(AML01.time, AML01.censor),
    type = c("kaplan-meier"),
    error = c("greenwood"), conf.type = c("plain"))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       2    0.905  0.0641        0.779        1.000
    7     17       1    0.852  0.0794        0.696        1.000
   10     15       1    0.795  0.0922        0.614        0.975
   13     12       1    0.729  0.1056        0.521        0.936
   16     11       1    0.662  0.1149        0.437        0.888
   22      7       1    0.568  0.1318        0.309        0.826
   23      6       1    0.473  0.1397        0.199        0.747
```

Example: Kaplan-Meier (Product-Limit) Method

```
> plot(survfit(Surv(AML01.time,AML01.censor),
      conf.type="none"), bty="l",
      xlab="Time", ylab="Survival Rate")
```


Example: Kaplan-Meier (Product-Limit) Method

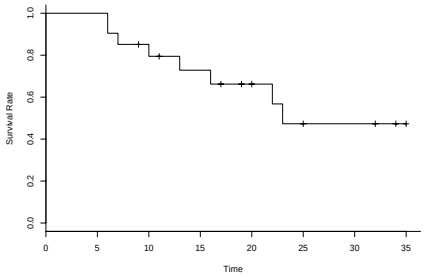

Figure 6: Acute Leukemia: Kaplan-Meier Survival Curve

Example: Kaplan-Meier (Product-Limit) Method

```
> plot(survfit(Surv(AML01.time,AML01.censor)))
```


Example: Kaplan-Meier (Product-Limit) Method

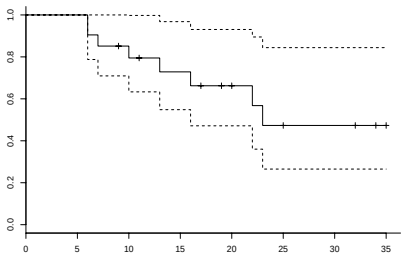

Figure 7: Acute Leukemia: KM Survival Curve with C.I.

Kaplan-Meier (Product-Limit) Method

1. Note that the survival curve is plotted as a series of horizontal lines based on the computed values of the KM estimate.
2. The Kaplan-Meier survival curve, which is just the graphical representation of the survival function, is easily interpreted.
3. The vertical axis represents the estimated fraction (or percent) of the population that has not died.
4. The horizontal axis represents the time since entry into the study.

Kaplan-Meier (Product-Limit) Method

5. For example, based on the survival curve in above example, we estimate that over 75% of similar children survived for at 12 months while less than 50
6. Note that survival curves never increase.

Kaplan-Meier (Product-Limit) Method

7. The percent surviving shown in the survival curve is relative to the total number entering the study.
8. Thus, the relevant population consists of those who satisfy the entry criteria.
9. As in this example, if all subjects enter the study at the same time and all subjects are followed to the end of the study then simple proportions can be used to estimate the fraction alive at any time during the study.

Kaplan-Meier (Product-Limit) Method

10. However, note that the mean survival time cannot be calculated, because the time to death for the surviving subjects is not known.
11. In this example, the median survival time can be estimated, but the 25 percentile cannot be estimated.
12. Note that if the subject with the longest follow-up has an event, then the Kaplan-Meier survival curve drops to 0 at the time of that event.
13. If the subject with the longest follow-up is censored, then the Kaplan-Meier estimate is undefined after that time.

Kaplan-Meier (Product-Limit) Method

14. The horizontal time axis measures time relative to entry into the study.
15. The time origin could be, and often is, a different calendar date for each subject in the study.
16. The time axis usually measures time from a well-defined event which defines the beginning of follow-up for each subject, such as birth.
17. The time axis can also start on a particular date, such as $1/1/89$.

Kaplan-Meier (Product-Limit) Method

18. The event, or outcome, of interest is often an event other than death.
19. For example, time to relapse, time to progression, and time to diagnosis are all appropriately analyzed with survival analysis methods.

Kaplan-Meier (Product-Limit) Method

20. Generally, the vertical axis measures the fraction of the population that is event-free.
21. The variability of the survival curve is usually larger for longer times because there are fewer subjects with longer follow-up, due to censoring.

Kaplan-Meier (Product-Limit) Method

22. In particular, a long at segment often appears at the right end of the KM estimate and should not, generally, be interpreted as representing a fraction of the population that is unlikely to die because they are "cured".
23. Instead, it may be due to imprecision of the estimate (based on a few long-term survivors).

Kaplan-Meier (Product-Limit) Method

24. The survival function (or curve) can be used to compute an estimate for the median survival time, $t_{0.5}$.

25. The time at which the survival function jumps from above to below 0.5 is the most commonly used estimate of $t_{0.5}$.

26. If there is an interval of times (t_L, t_U) for which $S(t) = Pr(T > t) = 0.5$ for $t_L \leq t < t_U$, then any time in the interval can be used to estimate the median, but the average of the endpoints, $\hat{t}_{0.5} = (t_L + t_U)/2$ is commonly used.

Kaplan-Meier (Product-Limit) Method

27. The KM estimate is only appropriate when the causes of censoring are independent of (unrelated to) subsequent mortality.

28. For example, if subjects are likely to be censored just before they die then the KM estimate can be severely biased.

Kaplan-Meier (Product-Limit) Method

29. We consider the enrollment and follow-up experience for subjects in a study.

30. Subjects were enrolled when they were diagnosed with a particular disease at the study center.

31. There are more subjects under observation (at risk) at the beginning of the study than there are at the end of the study because of deaths and losses.

32. The methods of survival analysis allow the data from subjects lost to follow-up to be used until the time at which they are lost.

Kaplan-Meier (Product-Limit) Method

33. For the purposes of survival analysis, we often assume that death rates vary with the time since entry into the study, but not with respect to calendar time of entry during the study period. With this assumption of homogeneity of death rates, it is appropriate to measure survival from the time of enrollment.

34. The assumption of homogeneity of death rates can be avoided with the use of regression methods,

35. which allow analysis to be adjusted for patient characteristics such as age and date of enrollment.

Summary of Basic Survival Analysis

1. Analysis of time from one event to another event. Some examples and counter-examples are:

(a) Example: Time from admission to discharge among burn patients.

(b) Examples with cancer:
- Time from remission to relapse.
- Time from diagnosis to remission.
- Time from diagnosis to death.

Summary of Basic Survival Analysis

(c) Examples:
- Time from first treatment to death in ESRD patients.
- Time from first treatment to transplant in ESRD patients.

Summary of Basic Survival Analysis

(d) counter-examples
- NOT: time to cancer among those getting cancer.
- NOT: dichotomous death outcome for hospital discharge (use dichotomous response methods: logistic regression, chi-square, discriminant).
- NOT: insurance actuarial tables, based on cross-section

Summary of Basic Survival Analysis

2. Clinical or personal versus statistical experience.

3. We remember exceptional events rather than the norm.

4. Out of 500 patients treated, one might remember the exceptional cases.

5. In contrast, many statistical summaries are oriented towards summarizing the norm. Statistical tools help summarize the norm as well as to identify distinguished cases.

Summary of Basic Survival Analysis

6. Statistical summaries and non-Gaussian distributions:

(a) means, probability density function, and histograms (for complete Gaussian data).
(b) medians, hazard function, and survival curve (for censored or non-Gaussian data).

7. Statistical significance versus importance.

Summary of Basic Survival Analysis

(a) Statistical significance is hard to achieve with small sample sizes. Consider two groups, A and B, with $n = 2$ in group A and $n = 3$ in group B.
(b) No matter how big the difference is between the two sets of numbers, the random chances (probability) that the difference between the two groups is as large or larger than is observed is at least $0.1 = 1/(\frac{5}{2})$ since there are $(\frac{5}{3}) = (\frac{5}{2})$ ways to distribute 5 numbers between these two groups.
(c) Thus, by chance, the 2 largest values would end up in group A and the smallest in the other with probability 0.1.
(d) So, the p-value is at least 0.1.

Summary of Basic Survival Analysis

(e) If there is any difference at all between the two populations, statistical significance will occur if the sample size is large enough, even if the difference is unimportant.
(f) Consider two large samples from populations that differ slightly.
(g) The difference is often significant (small p-value) because

$$t \approx \lim_{n\to\infty} \frac{\mu_1 - \mu_2}{\sigma\sqrt{2/n}} \to \infty \tag{61}$$

Summary of Basic Survival Analysis

8. When to use survival analysis versus the fraction with event.

(a) Use survival analysis when the events are spread out over a long period of time.
(b) e.g. Survival analysis curve for time to death for diabetes would be useful.
(c) Use a fraction when the events are clustered near the entry time.
(d) e.g. Hospital death rate = fraction of burn admissions discharged dead.
(e) Time to death is of secondary importance to fraction dead.

Summary of Basic Survival Analysis

9. How to describe a single sample of survival data:

10. Survival curve, median or other percentiles (s.e. of estimates).

11. Crude death rates=Total number of events divided by total follow-up. e.g. 156 patients followed for total of 2431 months with 15 deaths while on transplant yields death rate of $6.17 = 1000 \times 15/2431$ per 1,000 person months or $7.4 = 6.17 \times 12/10$ deaths per 100 person years.

Comaprison Survival Rates for Two Samples

Freich et al. (1963) and Gehan (1965) report the results of a clinical trial of 6-mercaptopurine (6-MP) versus placebo in 42 children, 21 children in eah group. Treatment allocation was randomized. Patients were followed until their leukemia return (relapse).

- Is there any difference between two survival rates?

Table 5: A clinical trial of 6-mercaptopurine (6-MP) versus placebo.

Placebo				6-MP			
Time	Censor	Time	Censor	Time	Censor	Time	Censor
1	1	5	1	10	1	20	0
22	1	4	1	7	1	19	0
3	1	15	1	32	0	6	1
12	1	8	1	23	1	17	0
8	1	23	1	22	1	35	0
17	1	5	1	6	1	6	1
2	1	11	1	16	1	13	1
11	1	4	1	34	0	9	0
8	1	1	1	32	0	6	0
12	1	8	1	25	0	10	0
2	1			11	0		

Log-Rank Test

Log-Rank Test

1. Often, one of the main objectives of statistical analysis is to compare two or more samples to each other.

2. In survival applications, the comparison can be directed towards a variety of parameters.

3. The comparison can be directed towards contrasting the death rates, the survival curves, the mean lifetimes, or the median lifetimes.

Log-Rank Test

4. The methods for comparison and the results of the comparison do not usually differ with the parameterization chosen because lower death rates, a higher survival curve, and longer lifetimes all tend to correspond to each other.

5. The objective is to determine whether the survival times in one group tend to be longer than the times in the other group.

6. If one survival curve is higher than the other (on the vertical axis) at a particular time, then a larger proportion of that sample has survived to that time.

Log-Rank Test

7. If one survival curve is higher than the other at all times, then the survival in that group tends to be longer than the survival in the other group, for example, evaluating mortality after exposure to a risk factor.

8. Compare the survival curves of the exposed and unexposed groups analyzing the time from entry to death as the time of the outcome event and the time of loss to follow-up as a censored observation.

Log-Rank Test

9. Several analysis can be useful in making such comparisons.
   (a) Plot the estimated survival curves on the same axes for comparison.
   (b) Interpret the two curves, if possible.

Log-Rank Test

(c) If one curve is consistently above the other, then the comparison of the two survival patterns is clear.
(d) If the curves cross once, then the comparison is harder to summarize; one group has lower event rates at the beginning while the other group has lower event rates at later times.
(e) If the curves overlap or cross many times, then a reasonable summary may be that the survival distributions are similar to each other.

Log-Rank Test

(f) Compute relevant summary proportions with or without the event (e.g., at 1 year and 5 years).
(g) The survival curve estimates the fraction that are event-free at each time.
(h) Each "curve" is usually plotted as a "staircase" function of time.
(i) Test for differences with the log rank test. (alternatively, the Peto-Wilcoxon or Prentice-Wilcoxon, but not the Breslow-Wilcoxon or Gehan test).

Log-Rank Test

(j) Compute and report the crude event rates (total number of events divided by the total time of follow-up) in each group for descriptive purposes (this assumes a constant event rate).
(k) Estimate the event rates during a series of time intervals and plot them as a function of time.
(l) More generally, compare several curves. Caution: ordinal groups (dose) are handled differently.

Log-Rank Test

10. The resulting plot can be quite informative.

11. We could compare survival at specific time points, or we are more interested in comparing two survival curves.

12. However, real differences can only be revealed by application of statistical tests of significance.

Log-Rank Test

13. When there are no censored observations, standard nonparametric tests can be used to compare survival distributions; for example, the Wilcoxon or
14. Mann-Whitney for the comparison of two samples, and the Kruskal-Wallis test for the comparison of several groups.
15. A family of nonparametric tests for samples with censoring will be considered in this chapter.

Log-Rank Test

1. Considering first sample of two samples, let $(X_{11}, X_{12}, \cdots, X_{1n_1})$ be i.i.d. each with survival time cumulative density function F_1, (survival function S_1), and $C_{11}, C_{12}, \cdots, C_{1n_1}$ be i.i.d. each with censoring time cumulative density function G_1, $C_{1i}, i = 1, 2, \cdots, n_1$ is the censoring time associated with T_{1i}.
2. We can observe $(T_{1i}, \delta_{1i}), i = 1, 2, \cdots, n_1$, where $T_{1i} = T_{1i} \wedge C_{1i},\ \delta_{1i} = I(T_{1i} \leq C_{1i})$.

Log-Rank Test

3. For the second sample, let $X_{21}, X_{22}, \cdots, X_{2n_2}$ be i.i.d. each with survival time cumulative density function F_2 (survival function S_2), and $C_{21}, C_{22}, \cdots, C_{2n_2}$ be i.i.d. each with censoring time density function G_2, C_{2i} is the censoring time associated with T_{2i}.
4. We can observe $(T_{2i}, \delta_{2i}), i = 1, 2, \cdots, n_2$, where $T_{2i} = T_{2i} \wedge C_{2i},\ \delta_{2i} = I(T_{2i} \leq C_{2i})$.

Log-Rank Test

1. The usual two-sample problem is to test

$$H_0 : F_1 = F_2 \tag{62}$$

2. In terms of medical research, we are interested in

$$H_0^\star : S_1(t) = S_2(t) \tag{63}$$
$$\text{v.s. } H_A^\star : S_1(t) \leq S_2(t) \tag{64}$$

3. H_A is an one-side alternative hypothesis with strict inequality at any time t.

Log-Rank Test

4. Let another testing hypothesis be

$$H_0^\dagger : \lambda_1(t) = \lambda_2(t) \tag{65}$$
$$\text{v.s. } H_A^\dagger : \lambda_1(t) \leq \lambda_2(t) \tag{66}$$

5. These two hypotheses in not exact the same such that

$$H_0^\dagger \Leftrightarrow H_0^\star \tag{67}$$
$$H_A^\dagger \Rightarrow H_A^\star \tag{68}$$
$$H_A^\dagger \not\Leftarrow H_A^\star \tag{69}$$

Log-Rank Test

$$(1)\ H_0^\star : S_1(t) = S_2(t) \tag{70}$$
$$\text{v.s. } H_A^\star : S_1(t) \leq S_2(t) \tag{71}$$
$$(2)\ H_0^\dagger : \lambda_1(t) = \lambda_2(t) \tag{72}$$
$$\text{v.s. } H_A^\dagger : \lambda_1(t) \leq \lambda_2(t) \tag{73}$$

$$(3)\ H_0^\dagger \Leftrightarrow H_0^\star \tag{74}$$
$$H_A^\dagger \Rightarrow H_A^\star \tag{75}$$
$$H_A^\dagger \not\Leftarrow H_A^\star \tag{76}$$

Log-Rank Test

1. Consider two hazard function and two survival functions as

$$\lambda_1 > \lambda_2 \tag{77}$$
$$\int_0^t \lambda_1(u)\,du > \int_0^t \lambda_2(u)\,du \tag{78}$$
$$\exp[-\int_0^t \lambda_1(u)\,du] < \exp[-\int_0^t \lambda_2(u)\,du] \tag{79}$$
$$S_1(t) < S_2(t) \tag{80}$$

2. We should plot two or more survival curves and (cumulative) hazard curves to see any cross over survival curves before we test any hypothesis.

Log-Rank Test

1. Formally, we test the hypothesis that the population survival distributions are equal.

2. The null and alternative hypotheses are

$$H_0 : S_1(t) = S_2(t), \quad \text{for all } T.0. \tag{81}$$
$$H_A : S_1(t) \neq S_2(t), \quad \text{for some } t > 0. \tag{82}$$

3. The log rank test is most useful for detecting consistent differences between survival curves. Difference methods should be used to document crossing survival curves.

Log Rank Test: Single 2×2 Table

1. Suppose we have two populations, for example, one population receive new treatment and another population receive standard treatment.

2. Suppose we have data include two groups from two population, the patients in either group may either die within a year or survival beyond a year.

3. The data may be summarized in a 2×2 Table as Table 6.

Log Rank Test: Single 2×2 Table

Table 6: Log Rank Test: Single 2×2 Table

	Dead	Alive	Total
Group (Population) 1	d_1	$n_1 - d_1$	n_1
Group (Population) 2	d_2	$n_2 - d_2$	n_2
Total	$d_1 + d_2 = d_.$	$n - d_.$	n

Log Rank Test: Single 2×2 Table

1. Denote

$$p_1 = Pr[\text{ Dead } | \text{ population 1 }] \tag{83}$$
$$p_2 = Pr[\text{ Dead } | \text{ population 1 }] \tag{84}$$

2. To test

$$H_0 : p_1 = p_2, \tag{85}$$

is the same to test **Risk Difference**

$$H_0 : p_1 - p_2 = 0. \tag{86}$$

Log Rank Test: Single 2×2 Table

$$\hat{p}_1 = \frac{d_1}{n_1} \tag{87}$$
$$\hat{p}_2 = \frac{d_2}{n_2} \tag{88}$$
$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{d_.}{n} \tag{89}$$
$$\hat{q} = 1 - \hat{p} \tag{90}$$
$$\text{Risk Difference} = \hat{p}_1 - \hat{p}_2 \tag{91}$$
$$\mathbf{Var}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} \tag{92}$$

Log Rank Test: Single 2×2 Table

3. Test statistic

$$\mathsf{z} = \frac{\mid \hat{p}_1 - \hat{p}_2 \mid}{\sqrt{\hat{p}(1-\hat{q})(\frac{1}{n_1}+\frac{1}{n_2})}} \tag{93}$$

4. Include the continuity correction

$$\mathsf{z}_c = \frac{\mid \hat{p}_1 - \hat{p}_2 \mid - \frac{n}{2}}{\sqrt{\hat{p}(1-\hat{q})(\frac{1}{n_1}+\frac{1}{n_2})}} \tag{94}$$

5. To test $H_0 : p_1 = p_2$;

$$p - \text{value} = 2\left[1 - \Phi(\mathsf{z})\right] \tag{95}$$

Log Rank Test: Single 2×2 Table

6. Actually, test Risk Difference is the same as using Pearson's chi-square test as

$$X^2 = \left[\frac{\mid \hat{p}_1 - \hat{p}_2 \mid}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}\right]^2 \tag{96}$$

$$= \frac{n\Big(|d_1(n_2-d_2)-(n_1-d_1)d_2|\Big)^2}{\Big[\, n_1\, n_2\, (d_\cdot)\, (n-d_\cdot)\,\Big]} \tag{97}$$

$$= \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{98}$$

Log Rank Test: Single 2×2 Table

7. Include the continuity correction,

$$X_c^2 = \frac{n\Big(|d_1(n_2-d_2)-(n_1-d_1)d_2| - \frac{n}{2}\Big)^2}{\Big[\, n_1\, n_2\, (d_\cdot)\, (n-d_\cdot)\,\Big]} \tag{99}$$

$$\text{p-value} = Pr[\chi_1^2 > X^2] \tag{100}$$

8. Note: Pearson's chi-square test, as the equations: 97 and 98, is an approximation to the exact discrete conditional distribution under H_0.

Log Rank Test: Single 2×2 Table

9. Given that four margins $n_1, n_2, d_\cdot, n - d_\cdot$ are fixed, the random variable D_1, which is the entry in the $(1,1)$ cell of the 2×2 table, has a hypergeometric distribution

$$Pr[D_1 = d_1] = \frac{\binom{n_1}{d_1}\binom{n_2}{d_2}}{\binom{n}{d_\cdot}} \tag{101}$$

$$\mathbf{E}(D_1) = \frac{n_1\, d_\cdot}{n} \tag{102}$$

$$\mathbf{Var}(D_1) = \frac{n_1\, n_2\, d_\cdot\, (n-d_\cdot)}{n^2\, (n-1)} \tag{103}$$

Log Rank Test: Single 2×2 Table

10. Consequently,

$$n_1\, (n_2 - d_2) - (n_1 - d_1)\, d_2 = n(\, d_1 - \mathbf{E}(D_1)\,) \tag{104}$$

$$n_1\, n_2\, (d_\cdot)\, (n-d_\cdot) = n^2\, (n-1)\, \mathbf{Var}(D_1) \tag{105}$$

$$X^2 = \frac{n\Big(|d_1(n_2-d_2)-(n_1-d_1)d_2|\Big)^2}{\Big[\, n_1\, n_2\, (d_\cdot)\, (n-d_\cdot)\,\Big]} \tag{106}$$

$$= \frac{n}{n-1}\left[\frac{d_1 - \mathbf{E}(D_1)}{\sqrt{\mathbf{Var}(D_1)}}\right]^2 \tag{107}$$

Log Rank Test: Sequence of 2×2 Table

1. Now suppose we have a k-sequence of 2×2 tables.
2. For example, we might have k strata of 2 groups that receive 2 different treatments.
3. Because there may be differences among k strata, we do not want to combined all k tables into a single 2×2 table.

Log Rank Test: Sequence of 2×2 Table

4. We want to test

$$H_0 \ : \ p_{11} = p_{21} \ , \ \cdots \ , \ p_{1k} = p_{2k}, \text{ simultaneous statement} \quad (108)$$

$$H_a \ : \ p_{1i} > p_{2i}, \text{ in any one stratum} \quad (109)$$

5. Where

$$p_{1i} \ = \ Pr[\text{ Dead } | \text{ Treatment 1, strata i }]$$

$$p_{2i} \ = \ Pr[\text{ Dead } | \text{ Treatment 2, strata i }]$$

Log Rank Test: Sequence of 2×2 Table

6. Consider stratum 1 as

Table 7: Log Rank Test with Sequence 2×2 Table: Stratum 1

	Dead	Alive	Total
Group (Population) 1	d_{11}	$n_{11} - d_{11}$	n_{11}
Group (Population) 2	d_{21}	$n_{21} - d_{21}$	n_{21}
Total	$d_{11} + d_{21} = d_{.1}$	$n_{.1} - d_{.1}$	$n_{.1}$

Log Rank Test: Sequence of 2×2 Table

7. till stratum k as

Table 8: Log Rank Test with Sequence 2×2 Table: Stratum k

	Dead	Alive	Total
Group (Population) 1	d_{1k}	$n_{1k} - d_{1k}$	n_{1k}
Group (Population) 2	d_{2k}	$n_{2k} - d_{2k}$	n_{2k}
Total	$d_{1k} + d_{2k} = d_{.k}$	$n_{.k} - d_{.k}$	$n_{.k}$

Log Rank Test: Sequence of 2×2 Table

8. We can use Mantel-Haenszel statistic to test association of a sequence 2×2 table

$$\theta_{MH} \ = \ \frac{\sum_1^k \Big(d_{1i} - \mathbf{E}(D_{1i}) \Big)}{\sum_1^k \sqrt{\mathbf{Var}(D_1 i)}} \quad (110)$$

$$\mathbf{E}(D_{1i}) \ = \ \frac{n_{1i}\, d_{.i}}{n_{.i}} \quad (111)$$

$$\mathbf{Var}(D_{1i}) \ = \ \frac{n_{1i}\, n_{2i}\ d_{.i}\, (n_{.i} - d_{.i})}{n_{.i}^2\, (n_{.i} - 1)} \quad (112)$$

where, for $i = 1, 2, \cdots, k$.

Log Rank Test: Sequence of 2×2 Table

9. Including the continuity correction, the Mantel-Haenszel statistic is

$$\theta_{MHc} \ = \ \frac{\left| \sum_1^k \Big(d_{1i} - \mathbf{E}(D_{1i}) \Big) \right| - \frac{1}{2}}{\sum_1^k \sqrt{\mathbf{Var}(D_{1i})}} \quad (113)$$

Log Rank Test: Sequence of 2×2 Table

10. When the tables are independent, then under H_0,

$$\theta_{MH} \sim \text{ asym } \ N(0,1) \quad (114)$$

either when k is fixed and $n_i \to \infty$ or $k \to \infty$ and the tables are also identically distributed.

11. Note: $\theta_{MH}^2 \ \sim \ \text{asym} \ \chi_1^2$ distribution.

Log Rank Test (Nonparametric Methods): Comparison of Two Samples Hazard Rate

Comparison of Two Samples Hazard Rate

$\Rightarrow$ Comparison of Two Samples Survival Rate

Log Rank Test: Comparison of Two Samples Hazard Rate

1. When there are no censored observation, standard nonparametric tests can be used to compare survival distributions; for example,
   - the Wilcoxon or Mann-Whitney for the comparison of two samples, and
   - the Kruskal-Wallis test for the comparison for two samples.

Log Rank Test: Comparison of Two Samples Hazard Rate

2. After construction a 2×2 table for each uncensored time, the evidence against the null hypothesis can be summarized in the following statistic:

$$\mathbf{U} = \sum_{i=1}^{i=k} w_i \left[d_{1i} - \mathbf{E}(D_{1i}) \right] \tag{115}$$

$$\mathbf{Var}(\mathbf{U}) = \sum_{i=1}^{i=k} w_i^2 \, \mathbf{Var}(D_{1i}) \tag{116}$$

$$Z = \frac{\sum_{i=1}^{i=k} w_i \left[d_{1i} - \mathbf{E}(D_{1i}) \right]}{\sqrt{\sum_{i=1}^{i=k} w_i^2 \, \mathbf{Var}(D_{1i})}} \tag{117}$$

Log Rank Test: Comparison of Two Samples Hazard Rate

3. Under H_0,

$$\mathbf{E}(\mathbf{U}) = 0 \tag{118}$$

$$Z \sim \text{ asym } N(0,1), \text{ under } H_0 \tag{119}$$

$$p - \text{value} = \Phi(|Z| \geq z) \tag{120}$$

Log Rank Test: Comparison of Two Samples Hazard Rate

4. Under H_0:

$$Z_{\text{LR}}^2 = X_{\text{LR}}^2 = \frac{\mathbf{U}^2}{\mathbf{Var}(\mathbf{U})} \tag{121}$$

$$X_{\text{LR}}^2 \sim \text{ asym } \chi_1^2, \text{ under } H_0 \tag{122}$$

5. $X_{\text{LR}}^2 = Z^2$ is chi-square distribution at 1 degree of freedom.

6. Log Rank Test: Comparison of Two Samples Hazard Rate

7. w_i is weight associated with the 2×2 table at time t_i, and

8. w_i can be any function of previous history $n_{1m}, n_{2m}, d_{1m}, d_{2m}, m \leq i$ except d_{1i}, d_{2i}.

Log Rank Test: Comparison of Two Samples Hazard Rate

9. Note: these sequence of tables are **not independent**, however, we still sum over the variance because of conditionally uncorrelated.

Log Rank Test: Comparison of Two Samples Hazard Rate

10. The choice

$$w_i = 1 \tag{123}$$

gives the **log-rank test (also called Cox-Mantel Test, Mantel-Cox Test, Mantel-Haenszel Test, Peto-Mantel-Haenszel Test, Generalized Mantel-Haenszel Test)**.

11. Log-rank test put equal weight on each observation and therefore, by default, is more sensitive to exposures with a constant relative risk, i.e., proportional hazard effect.

Log Rank Test: Comparison of Two Samples Hazard Rate

12. The choice

$$w_i = n_i \tag{124}$$

gives the **(Generalized Gehan) Wilcoxon Test, (Gehan-Breslow Test, Gehan Test, Generalized Mann-Whitney Test, Generalized Breslow Test)**.

13. It reduced to the Wilcoxon test in the absence of censoring.

Log Rank Test: Comparison of Two Samples Hazard Rate

14. The generalized Wilcoxon test put more weight on the beginning observations and because of that its use is more powerful in detecting the effects of short term risks.

15. The generalized Wilcoxon test is less sensitive than the log-rank test to differences between groups that occur at later points in time.

16. To put in another way, although both statistics test the same null hypothesis, they differ in their sensitivity to various kinds of departures from that hypothesis.

Log Rank Test: Comparison of Two Samples Hazard Rate

17. Log-rank test is more suitable when the alternative to the null hypothesis of no difference between two groups of survival times is that the hazard of death at any given time for an individual in one group is proportional to the hazard at that time for a similar individual the other group.

18. This is the assumption of proportional hazards, which underlines a number of methods for analyzing survival data.

Log Rank Test: Comparison of Two Samples Hazard Rate

19. Some statistician suggest that the Wilcoxon test is more appropriate than the log-rank test for comparing the two survival functions for other types of departure from the null hypothesis.

20. Wilcoxon test is more powerful in situations where event times have log-normal distributions with a common variance but with different means in the two groups.

21. Neither test is particularly good at detecting differences when survival curves cross.

Log Rank Test: Comparison of Two Samples Hazard Rate

22. Other statisticians suggest that Gehan-Wilcoxon test uses weight $n_{\cdot i}$ which carries the information of mortality and censoring.

23. Actually, Generalized Wilcoxon test tries to test two hypothesis

$$(1)\ H_{T_0} \ : \ \lambda_{T_1}(t) = \lambda_{T_2}(t) \tag{125}$$

$$(2)\ H_{C_0} \ : \ \lambda_{C_1}(t) = \lambda_{C_2}(t) \tag{126}$$

24. Unfortunately, this statistic can reject the null hypothesis when (1) is true and (2) is false because we are only interested (1) hypothesis.

Log Rank Test: Comparison of Two Samples Hazard Rate

25. Peto-Wilcoxon test uses $W_i = \hat{S}_{\text{combined}}(T_i)$ that really have to do with mortality.

26. So Peto-Wilcoxon statistic only tests $(1) H_{T_0} : \lambda_{T_1}(t) = \lambda_{T_2}(t)$ hypothesis.

Log Rank Test: Comparison of Two Samples Hazard Rate

27. Note: In SAS, Wilcoxon test has two situation:
    (a) Wilcoxon Test uses weight n_i in testing "STRATA" variable(s).
    (b) Wilcoxon Test is really similar to Peto-Wilcoxon test in "TEST" covariate(s).

28. Compare several groups simultaneously can be generalized from previous section. We will not discuss in details, most survival analysis software handles Comparison of several groups simultaneously.

Comaprison Survival Rates for Two Samples

Freich et al. (1963) and Gehan (1965) report the results of a clinical trial of 6-mercaptopurine (6-MP) versus placebo in 42 children, 21 children in eah group. Treatment allocation was randomized. Patients were followed until their leukemia return (relapse).

- Is there any difference between two survival rates?

Table 9: A clinical trial of 6-mercaptopurine (6-MP) versus placebo.

Placebo				6-MP			
Time	Censor	Time	Censor	Time	Censor	Time	Censor
1	1	5	1	10	1	20	0
22	1	4	1	7	1	19	0
3	1	15	1	32	0	6	1
12	1	8	1	23	1	17	0
8	1	23	1	22	1	35	0
17	1	5	1	6	1	6	1
2	1	11	1	16	1	13	1
11	1	4	1	34	0	9	0
8	1	1	1	32	0	6	0
12	1	8	1	25	0	10	0
2	1			11	0		

Comaprison Survival Rates for Two Samples

```
> setwd("C://temp//Rdata")
> AML<-read.csv("GehanAML.csv",header = TRUE,sep = ",",dec=".")
> attach(AML)
> Surv(time,censor)
 [1]  1  22   3  12   8  17   2  11   8  12   2   5   4  15
[15]  8  23   5  11   4   1   8  10   7  32+ 23  22   6  16
[29] 34+ 32+ 25+ 11+ 20+ 19+  6  17+ 35+  6  13   9+  6+ 10+
```

Comaprison Survival Rates for Two Samples

```
> gehan.surv<-survfit(Surv(time, censor)~group,
     type=c("kaplan-meier"),
    error=c("greenwood"), conf.type=c("log"))
> summary(gehan.surv)
Call: survfit(formula = Surv(time, censor) ~ group,
    type = c("kaplan-meier"),
    error = c("greenwood"), conf.type = c("log"))
```


Comaprison Survival Rates for Two Samples

group=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	21	2	0.9048	0.0641	0.78754	1.000
2	19	2	0.8095	0.0857	0.65785	0.996
3	17	1	0.7619	0.0929	0.59988	0.968
4	16	2	0.6667	0.1029	0.49268	0.902
5	14	2	0.5714	0.1080	0.39455	0.828
8	12	4	0.3810	0.1060	0.22085	0.657
11	8	2	0.2857	0.0986	0.14529	0.562
12	6	2	0.1905	0.0857	0.07887	0.460
15	4	1	0.1429	0.0764	0.05011	0.407
17	3	1	0.0952	0.0641	0.02549	0.356
22	2	1	0.0476	0.0465	0.00703	0.322
23	1	1	0.0000	NA	NA	NA

Comaprison Survival Rates for Two Samples

group=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	21	3	0.857	0.0764	0.720	1.000
7	17	1	0.807	0.0869	0.653	0.996
10	15	1	0.753	0.0963	0.586	0.968
13	12	1	0.690	0.1068	0.510	0.935
16	11	1	0.627	0.1141	0.439	0.896
22	7	1	0.538	0.1282	0.337	0.858
23	6	1	0.448	0.1346	0.249	0.807

Comaprison Survival Rates for Two Samples

```
> plot(gehan.surv,bty="l", conf.int=F, lty=1:2, lwd=2,
       xlab="time to remission (weeks)", ylab="survival")
> lines(gehan.surv, conf.int=T, lty=1:2, lwd=1.0,cex=2)
> legend(25,0.9,c("Control","6-MP"), lty=1:2,lwd=2)
```


Comaprison Survival Rates for Two Samples

```
> survdiff(Surv(time, censor)~group)
Call:
survdiff(formula = Surv(time, censor) ~ group)

         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 21       21     10.7      9.77      16.8
group=2 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```


Comaprison Survival Rates for Two Samples

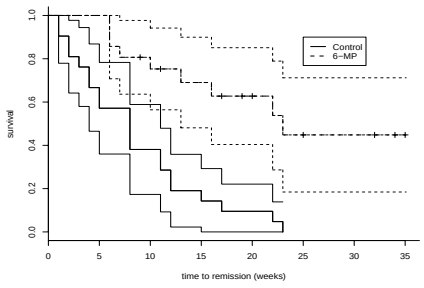

Figure 8: Comparison for Two Survival Curves