# 2


# Kaplan-Meier Survival Curves and the Log-Rank Test

The original version of this chapter was revised. A correction to this chapter can be found at https://doi.org/10.1007/978-1-4419-6646-9_11

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,* Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_2,

## Introduction

We begin with a brief review of the purposes of survival analysis, basic notation and terminology, and the basic data layout for the computer.

We then describe how to estimate and graph survival curves using the **Kaplan-Meier (KM)** method. The estimated survival probabilities are computed using a **product limit formula**.

Next, we describe how to compare two or more survival curves using the **log–rank test** of the null hypothesis of a common survival curve. For two groups, the log–rank statistic is based on the summed observed minus expected score for a given group and its variance estimate. For several groups, a computer should always be used because the log–rank formula is more complicated mathematically. The test statistic is approximately chi-square in large samples with $G - 1$ degrees of freedom, where $G$ denotes the number of groups being compared.

Several alternatives to the log–rank test will be briefly described. These tests are variations of the log rank test that weigh each observation differently. They are also large sample chi-square tests with $G - 1$ degrees of freedom.

Finally, we describe how to compute confidence intervals for the KM curve and for the median survival time.

## Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

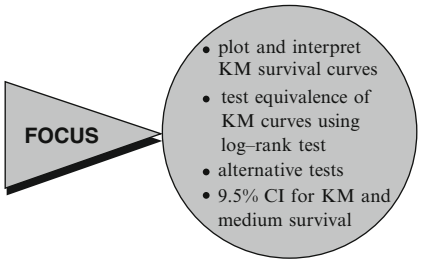

I. **Review** (pages 58–60)
II. **An example of Kaplan-Meier curves** (pages 61–65)
III. **General features of KM curves** (pages 66–67)
IV. **The log-rank test for two groups** (pages 67–71 )
V. **The log-rank test for several groups** (pages 71–73)
VI. **Alternatives to the log rank test** (pages 73–78)
VII. **Confidence intervals for KM curves** (pages 78–79)
VIII. **Confidence intervals for the median survival time** (page 80)
IX. **Summary** (page 81)

## Objectives

Upon completing the chapter, the learner should be able to:

1. Compute Kaplan-Meier (KM) probabilities of survival, given survival time and failure status information on a sample of subjects.
2. Interpret a graph of KM curves that compare two or more groups.
3. Draw conclusions as to whether or not two or more survival curves are the same based on computer results that provide a log–rank test and/or an alternative test.
4. Decide whether the log–rank test or one of the alternatives to this test is more appropriate for a given set of survival data.
5. Compute a 95% confidence interval for a KM survival probability.
6. Compute a 95% confidence interval for the median survival time obtained from a KM curve.

# Presentation




This presentation describes how to plot and interpret survival data using Kaplan-Meier (KM) survival curves and how to test whether or not two or more KM curves are equivalent using the log–rank test. We also describe alternative tests to the log–rank test. Furthermore, we provide formulae for computing 95% confidence intervals for a KM curve and for the median survival time.

## I. Review


We begin by reviewing the basics of survival analysis. Generally, survival analysis is a collection of statistical procedures for the analysis of data in which the outcome variable of interest is **time until an event occurs**. By **event**, we mean death, disease incidence, relapse from remission, or any designated experience of interest that may happen to an individual.

**Event:** death
disease
relapse

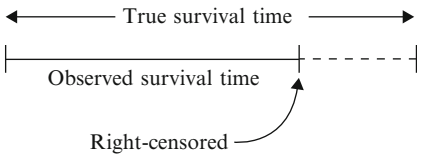
Time = survival time

Event = failure

When doing a survival analysis, we usually refer to the time variable as **survival time**. We also typically refer to the event as a **failure**.

**Censoring:** Don't know survival time exactly

Most survival analyses consider a key data analytical problem called **censoring**. In essence, censoring occurs when we have some information about individual survival time, but **we don't know the survival time exactly**.




Most survival time data is right-censored, because the true survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side of the observed time interval, giving us an observed survival time that is shorter than the true survival time. We want to use the observed survival time to draw implications about the true survival time.

***NOTATION***

$T$ = survival time
random variable
$t$ = specific value for $T$

As notation, we denote by a **capital *T*** the random variable for a person's survival time. Next, we denote by a **small letter *t*** any specific value of interest for the variable $T$.

$d$ = (0, 1) random variable

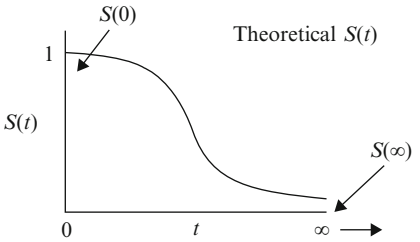
$= \begin{cases} 1 & \text{if failure} \\ 0 & \text{if censored} \end{cases}$

We let d denote a (0,1) random variable indicating either censorship or failure. A person who does not fail, that is, does not get the event during the study period, must have been censored either before or at the end of the study.

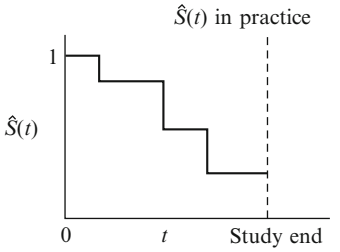
$S(t)$ = survivor function

$= \Pr(T > t)$

The survivor function, denoted by ***S(t)***, gives the probability that the random variable ***T*** exceeds the specified time $t$.




Theoretically, as $t$ ranges from 0 up to infinity, the survivor function is graphed as a decreasing smooth curve, which begins at $S(t) = 1$ at $t = 0$ and heads downward toward zero as $t$ increases toward infinity.

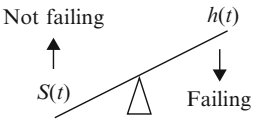



In practice, using data, we usually obtain estimated survivor curves that are **step functions**, as illustrated here, rather than smooth curves.

$h(t)$ = hazard function

= instantaneous potential given survival up to time $t$

The hazard function, denoted by ***h*(*t*)**, gives the **instantaneous potential** per unit time for the event to occur given that the individual has survived up to time $t$.




In contrast to the survivor function, which focuses on **not** failing, the hazard function focuses on failing; in other words, the higher the average hazard, the worse the impact on survival. The hazard is a **rate**, rather than a probability. Thus, the values of the hazard function range between zero and infinity.

$h(t)$ is a rate: 0 to $\infty$


Regardless of which function $S(t)$ or $h(t)$ one prefers, **there is a clearly defined relationship between the two**. In fact, if one knows the form of $S(t)$, one can derive the corresponding $h(t)$, and vice versa.

**General Data Layout:**

| Indiv. # | $t$ | $d$ | $X_1$ | $X_2 \ldots X_p$ |
|---|---|---|---|---|
| 1 | $t_1$ | $d_1$ | $X_{11}$ | $X_{12} \ldots X_{1p}$ |
| 2 | $t_2$ | $d_2$ | $X_{21}$ | $X_{22} \ldots X_{2p}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| $n$ | $t_n$ | $d_n$ | $X_{n1}$ | $X_{n2} \ldots X_{np}$ |

The general data layout for a survival analysis is given by the table shown here. The first column of the table identifies the study subjects. The second column gives the observed survival time information. The third column gives the information for $d$, the dichotomous variable that indicates censorship status. The remainder of the information in the table gives values for explanatory variables of interest.

**Alternative (ordered) data layout:**

| Ordered failure times, $t_{(f)}$ | # of failures $m_f$ | # censored in $[t_{(f)}, t_{(f+1)})$, $q_f$ | Risk set, $R(t_{(f)})$ |
|---|---|---|---|
| $t_{(0)} = 0$ | $m_0 = 0$ | $q_0$ | $R(t_{(0)})$ |
| $t_{(1)}$ | $m_1$ | $q_1$ | $R(t_{(1)})$ |
| $t_{(2)}$ | $m_2$ | $q_2$ | $R(t_{(2)})$ |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| $t_{(k)}$ | $m_k$ | $q_k$ | $R(t_{(k)})$ |

An alternative data layout is shown here. This layout is the basis upon which **Kaplan-Meier** survival curves are derived. The first column in the table gives ordered survival times from smallest to largest. The second column gives frequency counts of failures at each distinct failure time. The third column gives frequency counts, denoted by $q_f$, of those persons censored in the time interval starting with failure time $t_{(f)}$ up to but not including the next failure time, denoted by $t_{(f+1)}$. The last column gives the **risk set**, which denotes the collection of individuals who have survived at least to time $t_{(f)}$.

Table of ordered failures:

- Uses all information up to time of censorship;
- $S(t)$ is derived from $R(t)$.

To estimate the survival probability at a given time, we make use of the risk set at that time to include the information we have on a censored person up to the time of censorship, rather than simply throw away all the information on a censored person.

Survival probability:
Use **Kaplan-Meier (KM)** method.

The actual computation of such a survival probability can be carried out using the Kaplan-Meier (KM) method. We introduce the KM method in the next section by way of an example.

## II. An Example of Kaplan-Meier Curves

**EXAMPLE**

The data: remission times (weeks) for two groups of leukemia patients

| Group 1 ($n = 21$) treatment | Group 2 ($n = 21$) placebo |
|---|---|
| 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+, | 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23 |

*Note*: + denotes censored

| | # failed | # censored | Total |
|---|---|---|---|
| Group 1 | 9 | 12 | 21 |
| Group 2 | 21 | 0 | 21 |

Descriptive statistics:

$$\bar{T}_1(\text{ignoring} + \text{'s}) = 17.1,\ \bar{T}_2 = 8.6$$

$$\bar{h}_1 = .025,\ \bar{h}_2 = .115,\ \frac{\bar{h}_2}{\bar{h}_1} = 4.6$$

The data for this example derive from a study of the remission times in weeks for two groups of leukemia patients, with 21 patients in each group. Group 1 is the treatment group and group 2 is the placebo group. The basic question of interest concerns comparing the survival experience of the two groups.

Of the 21 persons in group 1, 9 failed during the study period and 12 were censored. In contrast, none of the data in group 2 are censored; that is, all 21 persons in the placebo group went out of remission during the study period.

In Chapter 1, we observed for this data set that group 1 appears to have better survival prognosis than group 2, suggesting that the treatment is effective. This conclusion was supported by descriptive statistics for the average survival time and average hazard rate shown. Note, however, that descriptive statistics provide overall comparisons but do not compare the two groups at different times of follow-up.

**EXAMPLE: (continued)**

Ordered failure times:

Group 1 (treatment)

| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ |
|---|---|---|---|
| 0 | 21 | 0 | 0 |
| 6 | 21 | 3 | 1 |
| 7 | 17 | 1 | 1 |
| 10 | 15 | 1 | 2 |
| 13 | 12 | 1 | 0 |
| 16 | 11 | 1 | 3 |
| 22 | 7 | 1 | 0 |
| 23 | 6 | 1 | 5 |
| >23 | — | — | — |

Group 2 (placebo)

| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ |
|---|---|---|---|
| 0 | 21 | 0 | 0 |
| 1 | 21 | 2 | 0 |
| 2 | 19 | 2 | 0 |
| 3 | 17 | 1 | 0 |
| 4 | 16 | 2 | 0 |
| 5 | 14 | 2 | 0 |
| 8 | 12 | 4 | 0 |
| 11 | 8 | 2 | 0 |
| 12 | 6 | 2 | 0 |
| 15 | 4 | 1 | 0 |
| 17 | 3 | 1 | 0 |
| 22 | 2 | 1 | 0 |
| 23 | 1 | 1 | 0 |

Group 2: no censored subjects
Group 2 (placebo)

| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $\hat{S}(t_{(f)})$ |
|---|---|---|---|---|
| 0 | 21 | 0 | 0 | 1 |
| 1 | 21 | 2 | 0 | $19/21 = .90$ |
| 2 | 19 | 2 | 0 | $17/21 = .81$ |
| 3 | 17 | 1 | 0 | $16/21 = .76$ |
| 4 | 16 | 2 | 0 | $14/21 = .67$ |
| 5 | 14 | 2 | 0 | $12/21 = .57$ |
| 8 | 12 | 4 | 0 | $8/21 = .38$ |
| 11 | 8 | 2 | 0 | $6/21 = .29$ |
| 12 | 6 | 2 | 0 | $4/21 = .19$ |
| 15 | 4 | 1 | 0 | $3/21 = .14$ |
| 17 | 3 | 1 | 0 | $2/21 = .10$ |
| 22 | 2 | 1 | 0 | $1/21 = .05$ |
| 23 | 1 | 1 | 0 | $0/21 = .00$ |

A table of ordered failure times is shown here for each group. These tables provide the basic information for the computation of KM curves.

Each table begins with a survival time of zero, even though no subject actually failed at the start of follow-up. The reason for the zero is to allow for the possibility that some subjects might have been censored before the earliest failure time.

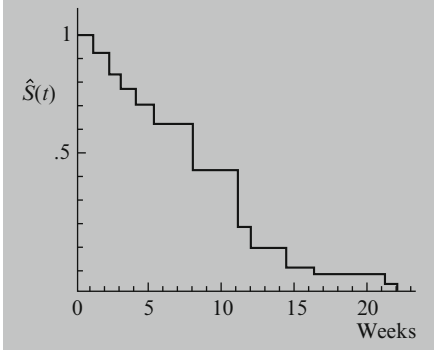
Also, each table contains a column denoted as $n_f$ that gives the number of subjects in the risk set at the start of the interval. Given that the risk set is defined as the collection of individuals who have survived at least to time $t_{(f)}$, it is assumed that $n_f$ includes those persons failing at time $t_{(f)}$. In other words, $n_f$ counts those subjects at risk for failing instantaneously prior to time $t_{(f)}$.

We now describe how to compute the KM curve for the table for group 2. The computations for group 2 are quite straightforward because there are no censored subjects for this group.

The table of ordered failure times for group 2 is presented here again with the addition of another column that contains survival probability estimates. These estimates are the KM survival probabilities for this group. We will discuss the computations of these probabilities shortly.

**EXAMPLE: (continued)**

**KM Curve for Group 2 (Placebo)**




$S(t) = \Pr(T > t)$

Group 2 (placebo)

| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $\check{S}(t_{(f)})$ |
|---|---|---|---|---|
| 0 | 21 | 0 | 0 | 1 |
| 1 | 21 | 2 | 0 | 19/21 = .90 |
| 2 | 19 | 2 | 0 | 17/21 = .81 |
| 3 | 17 | 1 | 0 | 16/21 = .76 |
| 4 | 16 | 2 | 0 | 14/21 = .67 |
| 5 | 14 | 2 | 0 | 12/21 = .57 |
| 8 | 12 | 4 | 0 | 8/21 = .38 |
| 11 | 8 | 2 | 0 | 6/21 = .29 |
| 12 | 6 | 2 | 0 | 4/21 = .19 |
| 15 | 4 | 1 | 0 | 3/21 = .14 |
| 17 | 3 | 1 | 0 | 2/21 = .10 |
| 22 | 2 | 1 | 0 | 1/21 = .05 |
| 23 | 1 | 1 | 0 | 0/21 = .00 |

$$\hat{S}(t_{(f)}) = \frac{\#\text{ surviving past } t_{(f)}}{21}$$

No censorship in group 2
Alternative formula: KM approach

A plot of the KM survival probabilities corresponding to each ordered failure time is shown here for group 2. Empirical plots such as this one are typically plotted as a step function that starts with a horizontal line at a survival probability of 1 and then steps down to the other survival probabilities as we move from one ordered failure time to another.

We now describe how the survival probabilities for the group 2 data are computed. Recall that a survival probability gives the probability that a study subject survives past a specified time.

Thus, considering the group 2 data, the probability of surviving past zero is unity, as it will always be for any data set.

Next, the probability of surviving past the first ordered failure time of 1 week is given by 19/21 or (.90) because 2 people failed at 1 week, so that 19 people from the original 21 remain as survivors past 1 week.

Similarly, the next probability concerns subjects surviving past 2 weeks, which is 17/21 (or. 81) because 2 subjects failed at 1 week and 2 subjects failed at 2 weeks leaving 17 out of the original 21 subjects surviving past 2 weeks.

The remaining survival probabilities in the table are computed in the same manner, that is, we count the number of subjects surviving past the specified time being considered and divide this number by 21, the number of subjects at the start of follow-up.

Recall that no subject in group 2 was censored, so the $q$ column for group 2 consists entirely of zeros. If some of the $q$'s had been nonzero, an alternative formula for computing survival probabilities would be needed. This alternative formula is called the Kaplan-Meier (KM) approach and can be illustrated using the group 2 data even though all values of $q$ are zero.

**EXAMPLE**

$$\hat{S}(4) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} = \frac{14}{21} = .67$$

$\Pr\ (T > t_{(f)} \ \ T \geq t_{(f)})$

$$\hat{S}(4) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} = \frac{14}{21} = .67$$

$\frac{19}{21} = \Pr(T > 1 | T \geq 1)$

$\frac{16}{17} = \Pr(T > 3 | T \geq 3)$

17 = # in risk set at week 3

$$\hat{S}(4) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16}$$

$$\hat{S}(8) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} \times \frac{12}{14} \times \frac{8}{12}$$

**KM** formula = product limit formula

Group 1 (treatment)

| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $\hat{S}\ (t_{(f)})$ |
|---|---|---|---|---|
| 0 | 21 | 0 | 0 | 1 |
| 6 | 21 | 3 | 1 | $1 \times \frac{18}{21}$ |
| | | . | | |
| | | . | | |
| | | . | | |

For example, an alternative way to calculate the survival probability of exceeding 4 weeks for the group 2 data can be written using the KM formula shown here. This formula involves the product of conditional probability terms. That is, each term in the product is the probability of exceeding a specific ordered failure time $t_{(f)}$ given that a subject survives up to that failure time.

Thus, in the KM formula for survival past 4 weeks, the term 19/21 gives the probability of surviving past the first ordered failure time, 1 week, given survival up to the first week. Note that all 21 persons in group 2 survived up to 1 week, but that 2 failed at 1 week, leaving 19 persons surviving past 1 week.

Similarly, the term 16/17 gives the probability of surviving past the third ordered failure time at week 3, given survival up to week 3. There were 17 persons who survived up to week 3 and 1 of these then failed, leaving 16 survivors past week 3. Note that the 17 persons in the denominator represents the number in the risk set at week 3.

Notice that the product terms in the KM formula for surviving past 4 weeks stop at the 4th week with the component 14/16. Similarly, the KM formula for surviving past 8 weeks stops at the eighth week.

More generally, any KM formula for a survival probability is limited to product terms up to the survival week being specified. That is why the KM formula is often referred to as a "product-limit" formula.

Next, we consider the KM formula for the data from group 1, where there are several censored observations.

The estimated survival probabilities obtained using the KM formula are shown here for group 1.

The first survival estimate on the list is $\hat{S}(0) - 1$, as it will always be, because this gives the probability of surviving past time zero.

**EXAMPLE: (continued)**

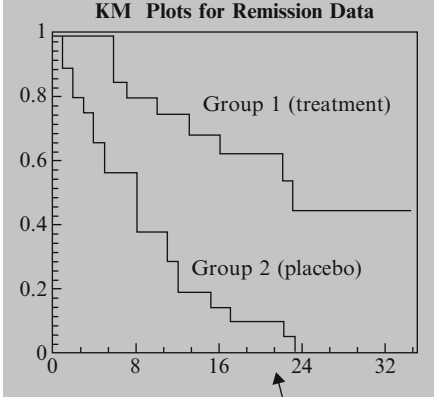
Group 1 (treatment)

| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $\hat{S}(t_{(f)})$ |
|---|---|---|---|---|
| 0 | 21 | 0 | 0 | (1) |
| 6 | 21 | 3 | 1 | $1 \times \left(\frac{18}{21}\right) = .8571$ |
| 7 | 17 | 1 | 1 | $.8571 \times \left(\frac{16}{17}\right) = .8067$ |
| 10 | 15 | 1 | 2 | $.8067 \times \frac{14}{15} = .7529$ |
| 13 | 12 | 1 | 0 | $.7529 \times \frac{11}{12} = .6902$ |
| 16 | 11 | 1 | 3 | $.6902 \times \frac{10}{11} = .6275$ |
| 22 | 7 | 1 | 0 | $.6275 \times \frac{6}{7} = .5378$ |
| 23 | 6 | 1 | 5 | $.5378 \times \frac{5}{6} = .4482$ |

Fraction at $t_{(f)}$: $\Pr(T > t_{(f)} \mid T \geq t_{(f)})$

Not available at $t_{(f)}$: failed prior to $t_{(f)}$ or censored prior to $t_{(f)}$

group 1 only




Obtain KM plots from computer package, e.g., SAS, Stata, SPSS R

The other survival estimates are calculated by multiplying the estimate for the immediately preceding failure time by a fraction. For example, the fraction is 18/21 for surviving past week 6, because 21 subjects remain up to week 6 and 3 of these subjects fail to survive past week 6. The fraction is 16/17 for surviving past week 7, because 17 people remain up to week 7 and 1 of these fails to survive past week 7. The other fractions are calculated similarly.

For a specified failure time $t_{(f)}$, the fraction may be generally expressed as the conditional probability of surviving past time $t_{(f)}$, given availability (i.e., in the risk set) at time $t_{(f)}$. This is exactly the same formula that we previously used to calculate each product term in the product limit formula used for the group 2 data.

Note that a subject might not be available at time $t_{(f)}$ for one of two reasons: (1) either the subject has failed prior to $t_{(f)}$, or (2) the subject has been censored prior to $t_{(f)}$. Group 1 has censored observations, whereas group 2 does not. Thus, for group 1, censored observations have to be taken into account when determining the number available at $t_{(f)}$.

Plots of the KM curves for groups 1 and 2 are shown here on the same graph. Notice that the KM curve for group 1 is consistently higher than the KM curve for group 2. These figures indicate that group 1, which is the treatment group, has better survival prognosis than group 2, the placebo group. Moreover, as the number of weeks increases, the two curves appear to get farther apart, suggesting that the beneficial effects of the treatment over the placebo are greater the longer one stays in remission.

The KM plots shown above can be easily obtained from most computer packages that perform survival analysis, including SAS, Stata, SPSS, and R. All the user needs to do is provide a KM computer program with the basic data layout and then provide appropriate commands to obtain plots.

## III. General Features of KM Curves

**General KM formula:**

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \hat{Pr}(T > t_{(f)} | T \geq t_{(f)})$$

The general formula for a KM survival probability at failure time $t_{(f)}$ is shown here. This formula gives the probability of surviving past the previous failure time $t_{(f-1)}$, multiplied by the conditional probability of surviving past time $t_{(f)}$, given survival to *at least* time $t_{(f)}$.

KM formula = product limit formula

$$\hat{S}(t_{(f-1)}) = \prod_{i=1}^{f-1} \hat{Pr}(T > t_{(i)} | T \geq t_{(i)})$$

The above KM formula can also be expressed as a product limit if we substitute for the survival probability $\hat{S}(t_{(f-1)})$, the product of all fractions that estimate the conditional probabilities for failure times $t_{(f-1)}$ and earlier.

**EXAMPLE**

$$\hat{S}(10) = .8067 \times \frac{14}{15} = .7529$$

$$= \boxed{\frac{18}{21} \times \frac{16}{17}} \times \frac{14}{15}$$

For example, the probability of surviving past 10 weeks is given in the table for group 1 (page 65) by .8067 times 14/15, which equals .7529. But the .8067 can be alternatively written as the product of the fractions 18/21 and 16/17. Thus, the product limit formula for surviving past 10 weeks is given by the triple product shown here.

$$\hat{S}(16) = .6902 \times \frac{10}{11}$$

$$= \boxed{\frac{18}{21} \times \frac{16}{17} \times \frac{14}{15} \times \frac{11}{12}} \times \frac{10}{11}$$

Similarly, the probability of surviving past 16 weeks can be written either as .6902 × 10/11, or equivalently as the five-way product of fractions shown here.

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \hat{Pr}[T > t_{(i)} | T \geq t_{(i)}]$$
$$= \hat{S}(t_{(f-1)}) \times \hat{Pr}(T > t_{(f)} | T \geq t_{(f)})$$

The general expression for the product limit formula for the KM survival estimate is shown here together with the general KM formula given earlier. Both expressions are equivalent.

Math proof:

Pr(A and B) = Pr(A) × Pr(B | A) always

A simple mathematical proof of the KM formula can be described in probability terms. One of the basic rules of probability is that the probability of a joint event, say A and B, is equal to the probability of one event, say A, times the conditional probability of the other event, B, given A.

$A = "T \geq t_{(f)}" \rightarrow A \text{ and } B = B$
$B = "T > t_{(f)}"$
$\Pr(A \text{ and } B) = \Pr(B) = S(t_{(f)})$

If we let A be the event that a subject survives to at least time $t_{(f)}$ and we let B be the event that a subject survives past time $t_{(f)}$, then the joint event A and B simplifies to the event B, which is inclusive of A. It follows that the probability of A and B equals the probability of surviving past time $t_{(f)}$.

No failures during $t_{(f-1)} < T < t_{(f)}$
$\Pr(A) = \Pr(T > t_{(f-1)}) = S(t_{(f-1)})$

Also, because $t_{(f)}$ is the next failure time after $t_{(f-1)}$, there can be no failures after time $t_{(f-1)}$ and before time $t_{(f)}$. Therefore, the probability of A is equivalent to the probability of surviving past the $(f-1)$th ordered failure time.

$\Pr(B|A) = \Pr(T > t_{(f)} | T \geq t_{(f)})$

Furthermore, the conditional probability of B given A is equivalent to the conditional probability in the KM formula.

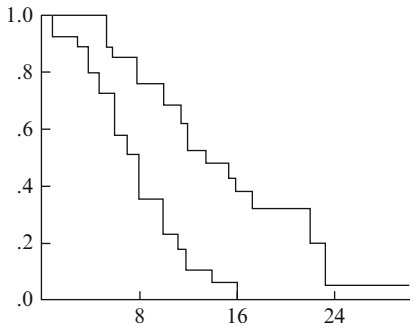
Thus, from Pr(A and B) formula,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B \mid A)$$
$$S(t_{(f)}) = S(t_{(f-1)}) \times \Pr(T > t_{(f)} | T \geq t_{(f)})$$

Thus, using the basic rules of probability, the KM formula can be derived.

## IV. The Log-Rank Test for Two Groups

Are KM curves statistically equivalent?




We now describe how to evaluate whether or not KM curves for two or more groups are statistically equivalent. In this section we consider two groups only. The most popular testing method is called the log–rank test.

When we state that two KM curves are "statistically equivalent," we mean that, based on a testing procedure that compares the two curves in some "overall sense," we do not have evidence to indicate that the true (population) survival curves are different.

- Chi-square test
- Overall comparison of KM curves
- Observed versus expected counts
- Categories defined by ordered failure times

The log–rank test is a large-sample chi-square test that uses as its test criterion a statistic that provides an overall comparison of the KM curves being compared. This (log–rank) statistic, like many other statistics used in other kinds of chi-square tests, makes use of observed versus expected cell counts over categories of outcomes. The categories for the log–rank statistic are defined by each of the ordered failure times for the entire set of data being analyzed.

**EXAMPLE**

Remission data: $n = 42$

| # failures | | | # in risk set | |
|---|---|---|---|---|
| $t_{(f)}$ | $m_{1f}$ | $m_{2f}$ | $n_{1f}$ | $n_{2f}$ |
| 1 | 0 | 2 | 21 | 21 |
| 2 | 0 | 2 | 21 | 19 |
| 3 | 0 | 1 | 21 | 17 |
| (4) | 0 | 2 | 21 | 16 |
| 5 | 0 | 2 | 21 | 14 |
| 6 | 3 | 0 | 21 | 12 |
| 7 | 1 | 0 | 17 | 12 |
| 8 | 0 | 4 | 16 | 12 |
| (10) | 1 | 0 | 15 | 8 |
| 11 | 0 | 2 | 13 | 8 |
| 12 | 0 | 2 | 12 | 6 |
| 13 | 1 | 0 | 12 | 4 |
| 15 | 0 | 1 | 11 | 4 |
| 16 | 1 | 0 | 11 | 3 |
| 17 | 0 | 1 | 10 | 3 |
| 22 | 1 | 1 | 7 | 2 |
| 23 | 1 | 1 | 6 | 1 |

As an example of the information required for the log–rank test, we again consider the comparison of the treatment (group 1) and placebo (group 2) subjects in the remission data on 42 leukemia patients.

Here, for each ordered failure time, $t_{(f)}$, in the entire set of data, we show the numbers of subjects ($m_{if}$) failing at that time, separately by group ($i$), followed by the numbers of subjects ($n_{if}$) in the risk set at that time, also separately by group.

Thus, for example, at week 4, no subjects failed in group 1, whereas two subjects failed in group 2. Also, at week 4, the risk set for group 1 contains 21 persons, whereas the risk set for group 2 contains 16 persons.

Similarly, at week 10, 1 subject failed in group 1, and no subjects failed at group 2; the risk sets for each group contain 15 and 8 subjects, respectively.

**Expected cell counts:**

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times \left(m_{1f} + m_{2f}\right)$$

↑ Proportion in risk set ↑ # of failures over both groups

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times \left(m_{1f} + m_{2f}\right)$$

We now expand the previous table to include expected cell counts and observed minus expected values for each group at each ordered failure time. The formula for the expected cell counts is shown here for each group. For group 1, this formula computes the expected number at time $f$ (i.e., $e_{1f}$) as the proportion of the total subjects in both groups who are at risk at time $f$, that is, $n_{1f}/(n_{1f} + n_{2f})$, multiplied by the total number of failures at that time for both groups (i.e., $m_{1f} + m_{2f}$). For group 2, $e_{2f}$ is computed similarly.

**EXAMPLE**

Expanded Table (Remission Data)

| | | # failures | | # in risk set | | # expected | | Observed-expected | |
|---|---|---|---|---|---|---|---|---|---|
| $f$ | $t_{(f)}$ | $m_{1f}$ | $m_{2f}$ | $n_{1f}$ | $n_{2f}$ | $e_{1f}$ | $e_{2f}$ | $m_{1f}-e_{1f}$ | $m_{2f}-e_{2f}$ |
| 1 | 1 | 0 | 2 | 21 | 21 | (21/42) × 2 | (21/42) × 2 | −1.00 | 1.00 |
| 2 | 2 | 0 | 2 | 21 | 19 | (21/40) × 2 | (19/40) × 2 | −1.05 | 1.05 |
| 3 | 3 | 0 | 1 | 21 | 17 | (21/38) × 1 | (17/38) × 1 | −0.55 | 0.55 |
| 4 | 4 | 0 | 2 | 21 | 16 | (21/37) × 2 | (16/37) × 2 | −1.14 | 1.14 |
| 5 | 5 | 0 | 2 | 21 | 14 | (21/35) × 2 | (14/35) × 2 | −1.20 | 1.20 |
| 6 | 6 | 3 | 0 | 21 | 12 | (21/33) × 3 | (12/33) × 3 | 1.09 | −1.09 |
| 7 | 7 | 1 | 0 | 17 | 12 | (17/29) × 1 | (12/29) × 1 | 0.41 | −0.41 |
| 8 | 8 | 0 | 4 | 16 | 12 | (16/28) × 4 | (12/28) × 4 | −2.29 | 2.29 |
| 9 | 10 | 1 | 0 | 15 | 8 | (15/23) × 1 | (8/23) × 1 | 0.35 | −0.35 |
| 10 | 11 | 0 | 2 | 13 | 8 | (13/21) × 2 | (8/21) × 2 | −1.24 | 1.24 |
| 11 | 12 | 0 | 2 | 12 | 6 | (12/18) × 2 | (6/18) × 2 | −1.33 | 1.33 |
| 12 | 13 | 1 | 0 | 12 | 4 | (12/16) × 1 | (4/16) × 1 | 0.25 | −0.25 |
| 13 | 15 | 0 | 1 | 11 | 4 | (11/15) × 1 | (4/15) × 1 | −0.73 | 0.73 |
| 14 | 16 | 1 | 0 | 11 | 3 | (11/14) × 1 | (3/14) × 1 | 0.21 | −0.21 |
| 15 | 17 | 0 | 1 | 10 | 3 | (10/13) × 1 | (3/13) × 1 | −0.77 | 0.77 |
| 16 | 22 | 1 | 1 | 7 | 2 | (7/9) × 2 | (2/9) × 2 | −0.56 | 0.56 |
| 17 | 23 | 1 | 1 | 6 | 1 | (6/7) × 2 | (1/7) × 2 | −0.71 | 0.71 |
| Totals | | 9 | (21) | | | 19.26 | (10.74) | −10.26 | (−10.26) |

# of failure times

$$O_i - E_i = \sum_{f=1}^{17} (m_{if} - e_{if}), \qquad i = 1, 2$$

When two groups are being compared, the log–rank test statistic is formed using the sum of the observed minus expected counts over all failure times for one of the two groups. In this example, this sum is −10.26 for group 1 and 10.26 for group 2. We will use the group 2 value to carry out the test, but as we can see, except for the minus sign, the difference is the same for the two groups.

**EXAMPLE**

$O_1 - E_1 = -10.26$
$O_2 - E_2 = 10.26$

**Two groups:**

$O_2 - E_2$ = summed observed minus expected score for group 2

$$\text{Log–rank statistic} = \frac{(O_2 - E_2)^2}{\text{Var}(O_2 - E_2)}$$

For the two-group case, the log–rank statistic, shown here at the left, is computed by dividing the square of the summed observed minus expected score for one of the groups — say, group 2 — by the variance of the summed observed minus expected score.

$$\mathrm{Var}(O_i - E_i) = \sum_j \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$
$$i = 1,\ 2$$

The expression for the estimated variance is shown here. For two groups, the variance formula is the same for each group. This variance formula involves the number in the risk set in each group ($n_{if}$) and the number of failures in each group ($m_{if}$) at time $f$. The summation is over all distinct failure times.

$H_0$: no difference between survival curves

The null hypothesis being tested is that there is no overall difference between the two survival curves. Under this null hypothesis, the log–-rank statistic is approximately chi-square with one degree of freedom. Thus, a P-value for the log–rank test is determined from tables of the chi-square distribution.

Log-rank statistic $\sim\chi^2$ with 1 df under $H_0$

Computer programs:
**Stata's "sts test":**

- descriptive statistics for KM curves
- log–rank statistic
- Alternative statistics to log–rank statistic

Several computer programs are available for calculating the log–rank statistic. For example the **Stata** package has a command called **"sts test"** that computes descriptive information about Kaplan-Meier curves, the log–rank statistic, and alternative statistics to the log–rank statistic, to be described later. Other packages, like **SAS** and **SPSS**, have procedures that provide results similar to those of **Stata**. A comparison of Stata, SAS, SPSS and R procedures and output is provided in the Computer Appendix at the back of this text.

**EXAMPLE**

Using Stata: Remission Data

| Group | Events observed | Events expected |
|---|---|---|
| 1 | 9 | 19.25 |
| 2 | 21 | 10.75 |
| Total | 30 | 30.00 |

Log rank = chi2(2) = 16.79
P-Value = Pr > chi2 = 0.000

For the remission data, the edited printout from using the **Stata** "sts test" procedure is shown here. The log–rank statistic is 16.79 and the corresponding P-value is zero to three decimal places. This P-value indicates that the null hypothesis should be rejected. We can therefore conclude that the treatment and placebo groups have significantly different KM survival curves.

**EXAMPLE**

$O_2 - E_2 = 10.26$
$\text{Var}(O_2 - E_2) = 6.2685$

$$\text{Log - rank statistic} = \frac{(O_2 - E_2)^2}{\widehat{\text{Var}}(O_2 - E_2)} = \frac{(10.26)^2}{6.2685} = 16.793$$

Although the use of a computer is the easiest way to calculate the log–rank statistic, we provide here some of the details of the calculation. We have already seen from earlier computations that the value of $O_2 - E_2$ is 10.26. The estimated variance of $O_2 - E_2$ is computed from the variance formula above to be 6.2685. The log–rank statistic then is obtained by squaring 10.26 and dividing by 6.285, which yields 16.793, as shown on the computer printout.

Approximate formula:

$$X^2 \approx \sum_{i}^{\#\text{ of groups}} \frac{(O_i - E_i)^2}{E_i}$$

An approximation to the log–rank statistic, shown here, can be calculated using observed and expected values for each group without having to compute the variance formula. The approximate formula is of the classic chi-square form that sums over each group being compared the square of the observed minus expected value divided by the expected value.

**EXAMPLE**

$$X^2 = \frac{(-10.26)^2}{19.26} + \frac{(10.26)^2}{10.74} = 15.276$$

Log-rank statistic = 16.793

The calculation of the approximate formula is shown here for the remission data. The expected values are 19.26 and 10.74 for groups 1 and 2, respectively. The chi-square value obtained is 15.276, which is slightly smaller than the log–rank statistic of 16.793.

## V. The Log-Rank Test for Several Groups

$H_0$: All survival curves are the same.

Log-rank statistics for $> 2$ groups involves variances and covariances of $O_i - E_i$.

Matrix formula: See Appendix at end of this chapter.

The log–rank test can also be used to compare three or more survival curves. The null hypothesis for this more general situation is that all survival curves are the same.

Although the same tabular layout can be used to carry out the calculations when there are more than two groups, the test statistic is more complicated mathematically, involving both variances and covariances of summed observed minus expected scores for each group. A convenient mathematical formula can be given in matrix terms. We present the matrix formula for the interested reader in an Appendix at the end of this chapter.

Use computer program for calculations.

We will not describe further details about the calculation of the log–rank statistic, because a computer program can easily carry out the computations from the basic data file. Instead, we illustrate the use of this test with data involving more than two groups.

$G$ ($\geq 2$) groups:
log–rank statistic $\sim \chi^2$ with $G - 1$ df

If the number of groups being compared is $G$ ($\geq 2$), then the log–rank statistic has approximately a large sample chi-square distribution with $G - 1$ degrees of freedom. Therefore, the decision about significance is made using chi-square tables with the appropriate degrees of freedom.

Approximation formula:

$$X^2 \approx \sum_{i}^{\#\text{ofgroups}} \frac{(O_i - E_i)^2}{E_i}$$

Not required because computer program calculates the exact log–-rank statistic

The approximate formula previously described involving only observed and expected values without variance or covariance calculations can also be used when there are more than two groups being compared. However, practically speaking, the use of this approximate formula is not required as long as a computer program is available to calculate the exact log–rank statistic.

We now provide an example to illustrate the use of the log–rank statistic to compare more than two groups.

**EXAMPLE**

**vets.dat:** survival time in days,

$n$ 137

Veteran's Administration Lung Cancer Trial

Column 1: Treatment (standard = 1, test = 2)
Column 2: Cell type 1 (large = 1, other = 0)
Column 3: Cell type 2 (adeno = 1, other = 0)
Column 4: Cell type 3 (small = 1, other = 0)
Column 5: Cell type 4 (squamous = 1, other = 0)
Column 6: Survival time (days)
Column 7: (Performance Status)
(0 = worst . . . 100 = best)
Column 8: Disease duration (months)
Column 9: Age
Column 10: Prior therapy (none = 0, some = 1)
Column 11: Status (0 = censored, 1 = died)

The data set "vets.dat" considers survival times in days for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their text (*The Statistical Analysis of Survival Time Data*, John Wiley, pp. 223–224, 1980). A complete list of the variables is shown here. Failure status is defined by the status variable (column 11).

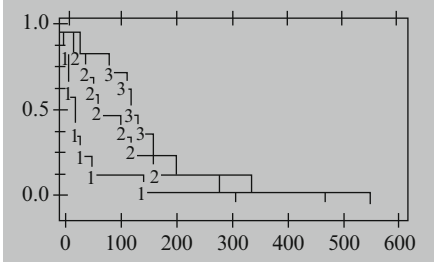
Among the variables listed, we now focus on the performance status variable (column 7). This variable is a continuous variable, so before we can obtain KM curves and the log–rank test, we need to categorize this variable.

**EXAMPLE: (continued)**

Performance Status Categories

| Group # | Categories | Size |
|---|---|---|
| 1 | 0–59 | 52 |
| 2 | 60–74 | 50 |
| 3 | 75–100 | 35 |

KM curves for performance status groups




| Group | Events observed | Events expected |
|---|---|---|
| 1 | 50 | 26.30 |
| 2 | 47 | 55.17 |
| 3 | 31 | 46.53 |
| Total | 128 | 128.00 |

Log-rank = chi2(2) = 29.18
P-value = Pr > chi2 = 0.0000
$G = 3$ groups; $df = G - 1 = 2$

Log-rank test is highly significant.

Conclude significant difference among three survival curves.

If, for the performance status variable, we choose the categories 0–59, 60–74, and 75–100, we obtain three groups of sizes 52, 50, and 35, respectively.

The KM curves for each of three groups are shown here. Notice that these curves appear to be quite different. A test of significance of this difference is provided by the log–rank statistic.

An edited printout of descriptive information about the three KM curves together with the log–rank test results are shown here. These results were obtained using the Stata package.

Because three groups are being compared here, $G = 3$ and the degrees of freedom for the log–rank test is thus $G - 1$, or 2. The log–rank statistic is computed to be 29.181, which has a P-value of zero to three decimal places. Thus, the conclusion from the log–rank test is that there is a highly significant difference among the three survival curves for the performance status groups.

## VI. Alternatives to the Log Rank Test

Alternative tests supported by Stata

Wilcoxen
Tarone-Ware
Peto
Flemington-Harrington

There are several alternatives to the log rank test offered by Stata, SAS, SPSS, and R designed to test the hypothesis that two or more survival curves are equivalent. In this section we describe the **Wilcoxon**, the **Tarone-Ware**, the **Peto**, and the **Flemington-Harrington** test. All of these tests are variations of the log rank test and are easily implemented in Stata.

Log rank uses

$$O_i - E_i = \sum_f \left(m_{if} - e_{if}\right)$$

i = group #
$f = f$th failure time

In describing the differences among these tests, recall that the log rank test uses the summed observed minus expected score $O - E$ in each group to form the test statistic. This simple sum gives the same weight – namely, unity – to each failure time when combining observed minus expected failures in each group.

Weighting the test statistic for two groups

Test statistic:

$$\frac{\left(\sum_f w(t_{(f)})\left(m_{if} - e_{if}\right)\right)^2}{\text{var}\left(\sum_j w(t_{(f)})\left(m_{if} - e_{if}\right)\right)}$$

i = 1, 2
$f = f$th failure time
$w(t_{(f)})$ = weight at $f$th failure time

The Wilcoxon, Tarone-Ware, Peto, and Flemington-Harrington test statistics are variations of the log rank test statistic and are derived by applying different weights at the f-th failure time (as shown on the left for two groups).

*Wilcoxon Test*

- $w(t_f) = n_f$ (number at risk)
- Earlier failures receive more weight
- Appropriate if treatment effect is strongest in earliest phases of administration

The Wilcoxon test (called the Breslow test in SPSS) weights the observed minus expected score at time $t_f$ by the number at risk $n_f$, over all groups at time $t_f$. Thus, the Wilcoxon test places more emphasis on the information at the beginning of the survival curve where the number at risk is large allowing early failures to receive more weight than later failures. This type of weighting may be used to assess whether the effect of a treatment on survival is strongest in the earlier phases of administration and tends to be less effective over time.

Weights Used for Various Test Statistics

| Test Statistic | $w(t_{(f)})$ |
|---|---|
| Log rank | 1 |
| Wilcoxon | $n_f$ |
| Tarone-Ware | $\sqrt{n_f}$ |
| Peto | $\tilde{s}(t_{(f)})$ |
| Flemington-Harrington | $\hat{S}(t_{(f-1)})^p \times[1 - \hat{S}(t_{(f-1)})]^q$ |

The Tarone-Ware test statistic also applies more weight to the early failure times by weighting the observed minus expected score at time $t_{(f)}$ by the square root of the number at risk $\sqrt{n_f}$. The Peto test weights the $f$-th failure time by the survival estimate $\tilde{s}(t_{(f)})$ calculated over all groups combined. This survival estimate $\tilde{s}(t_{(f)})$ is similar but not exactly equal to the Kaplan-Meier survival estimate. The Flemington-Harrington test uses the Kaplan-Meier survival estimate $\hat{S}(t)$ over all groups to calculate its weights for the f-th failure time, $\hat{S}(t_{(f-1)})^p[1 - \hat{S}(t_{(f-1)})]^q$. The weights for each of these test statistics are summarized on the left.

*Flemington-Harrington Test*

$$w(t) = \hat{S}\left(t_{(f-1)}\right)^{p}\left[1 - \hat{S}\left(t_{(f-1)}\right)\right]^{q}$$

if p = 1 and q = 0, $w(t) = \hat{S}(t_{(f-1)})$
if p = 0 and q = 1,
  $w(t) = 1 - \hat{S}(t_{(f-1)})$
if p = 0 and q = 0,
  w(t) = 1 (log rank test)

The Flemington-Harrington test allows the most flexibility in terms of the choice of weights because the user provides the values of p and q. For example, if p = 1 and q = 0 then $w(t) = \hat{S}(t_{(f-1)})$ which gives more weight for the earlier survival times when $\hat{S}(t_{(f-1)})$ is close to one. However, if p = 0 and q = 1 then $w(t) = 1 - \hat{S}(t_{(f-1)})$ in which case the later survival times receive more weight. If p = 0 and q = 0 then w(t) = 1, and the Flemington-Harrington test reduces to the log rank test.

Comparisons of Test Results: Remission Data, Testing Treatment (RX)

| Test | Chi-square (1 df) | P-value |
|---|---|---|
| Log rank | 16.79 | 0.0000 |
| Wilcoxon | 13.46 | 0.0002 |
| Tarone-Ware | 15.12 | 0.0001 |
| Peto | 14.08 | 0.0002 |
| FH (p = 3, q = 1) | 8.99 | 0.0027 |
| FH (p = 1, q = 3) | 12.26 | 0.005 |

On the left is a comparison of test results for the effect of treatment (vs. placebo) using the remission data. The log rank chi-square statistic (also displayed previously in this chapter) is the highest among these tests at 16.79. The Flemington–Harrington (FH) test with p = 3 and q = 1 yielded the lowest chi-square value at 8.99, although with this weighting it is not immediately obvious which part of the survival curve is getting the most weight. However, all the test results are highly significant yielding a similar conclusion to reject the null hypothesis.

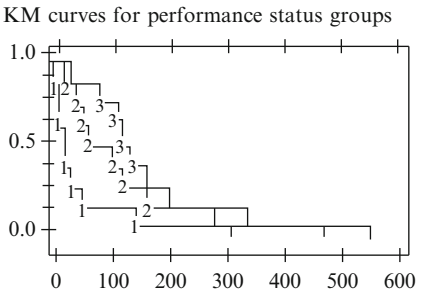
Vets Data, 3-Level Performance Status

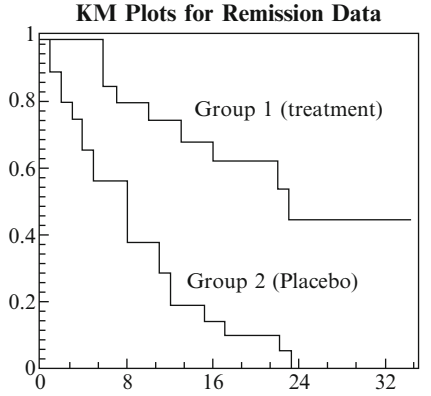
| Test | Chi-square (2 df) | P-value |
|---|---|---|
| Log rank | 29.18 | 0.0000 |
| Wilcoxon | 46.10 | 0.0000 |

Remission Data, 2-Level Treatment

| Test | Chi-square (1 df) | P-value |
|---|---|---|
| Log rank | 16.79 | 0.0000 |
| Wilcoxon | 13.46 | 0.0002 |

On the left are comparisons of the log rank and Wilcoxon tests for the 3-level performance status variable from the vets dataset discussed in the previous section. The Wilcoxon test yields a higher chi-square value (46.10) than the log rank test (29.18). In contrast, the log rank test for the effect of treatment (RX) from the remissions data yields a higher chi-square value (16.79) than the Wilcoxon test (13.46). However, both the Wilcoxon and log rank tests are highly significant for both performance status and for treatment variables.




A comparison of survival curves gives insight into why the Wilcoxon test yields a higher chi-square value than the log rank test for the 3-level performance status variable. The 3 curves being compared are farthest apart in the early part of followup before becoming closer later. By contrast, a comparison of the 2 curves for treatment shows the curves diverging over time.

*Choosing a Test*

- Results of different weightings usually lead to similar conclusions
- The best choice is test with most power
- Power depends on how the null is violated
- There may be a clinical reason to choose a particular weighting
- Choice of weighting should be a priori

In general, the various weightings should provide similar results and will usually lead to the same decision as to whether the null hypothesis is rejected. The choice of which weighting of the test statistic to use (e.g., log rank or Wilcoxon) depends on which test is believed to provide the greatest statistical power, which in turn depends on how it is believed the null hypothesis is violated.

If there is a clinical reason to believe the effect of an exposure is more pronounced toward the beginning (or end) of the survival function, then it makes sense to use a weighted test statistic. However, **one should make an a priori decision on which statistical test to use** rather than fish for a desired p-value. Fishing for a desired result may lead to bias.

Stratified log rank test

- $O - E$ scores calculated within strata
- $O - E$ scores then summed across strata
- Allows control of stratified variable

```
Stratified log-rank test
->1wbc3 = 1
           |  Events    Events
rx         | observed  expected
-----------+--------------------
0          |        0      2.91
1          |        4      1.09
-----------+--------------------
Total      |        4      4.00
           |
->1wbc3 = 2
           |  Events    Events
rx         | observed  expected
-----------+--------------------
0          |        5      7.36
1          |        5      2.64
-----------+--------------------
Total      |       10     10.00
->1wbc3 = 3
           |  Events    Events
rx         | observed  expected
-----------+--------------------
0          |        4      6.11
1          |       12      9.89
-----------+--------------------
Total      |       16     16.00
->Total
           |             Events
           |  Events    expected
rx         | observed       (*)
-----------+--------------------
0          |        9     16.38
1          |       21     13.62
-----------+--------------------
Total      |       30     30.00
```

The stratified log rank test is another variation of the log rank test. With this test the summed observed minus expected scores $O - E$ are calculated within strata of each group and then summed across strata. **The stratified log rank test provides a method of testing the equivalence of survival curves controlling for the stratified variable**. An example of the stratified log rank test is presented next using the remission data.

On the left is Stata output from performing a stratified log rank test for the effect of treatment (RX) stratified by a 3-level variable (LWBC3) indicating low, medium, or high log white blood cell count (coded 1, 2, and 3, respectively).

Within each stratum of LWBC3, the expected number of events is calculated for the treated group (RX = 0) and for the placebo group (RX = 1). The total expected number of events for the treated group is found by summing the expected number of events over the three strata: 2.91 + 7.36 + 6.11 = 16.38. Similarly the total expected number of events for the placebo group is calculated: 1.09 + 2.64 + 9.89 = 13.62. This compares to 9 observed cases from the treated group and 21 observed cases from the placebo group yielding a chi-square value of 10.14 with 1 degree of freedom (for 2 levels of treatment) and a corresponding p-value of 0.0014.

Recall that when we did not control for log white blood cell count, the log rank test for the effect of treatment yielded a chi-square value of 16.79 and a corresponding p-value rounded to 0.0000.

Log rank unstratified

$$O_i - E_i = \sum_j \left(m_{if} - e_{if}\right)$$

i = group #, f = fth failure time

Log rank stratified

$$O_i - E_i = \sum_s \sum_f \left(m_{ifs} - e_{ifs}\right)$$

i = group #, f = jth failure time, s = stratum #

Stratified or unstratified (G groups)
Under $H_0$:

log rank statistic $\sim\chi^2$ with G − 1 df

The only difference between the unstratified and stratified approaches is that for the unstratified approach, the observed minus expected number of events for each failure time is summed over all failure times for each group (i). With the stratified approach, the observed minus expected number of events is summed over all failure times for each group within each stratum and then summed over all strata. Either way, the null distribution is chi-square with G − 1 degrees of freedom, where G represents the number of groups being compared (not the number of strata).

Can stratify with other tests
Wilcoxon, Tarone-Ware, Peto, Flemington-Harrington

Limitation
Sample-size may be small within strata

The stratified approach can also be applied to any of the weighted variations of the log rank test (e.g., Wilcoxon). A limitation of the stratified approach is the reduced sample size within each stratum. This is particularly problematic with the remission dataset, which has a small sample size to begin with.

Alternatively
Test associations using modeling

- Can simultaneously control covariates
- Shown in next chapter

We have shown how the stratified log rank test can be used to test the effect of treatment while controlling for log white blood cell count. In the next chapter we show how modeling can be used to test an association of a predictor variable while simultaneously controlling for other covariates.

## VII. Confidence intervals for KM curves

95% CI for the KM curve:

$$\hat{S}_{KM}(t) \pm 1.96\sqrt{\hat{Var}[\hat{S}_{KM}(t)]}$$

where **Greenwood's formula** for $\text{Var}[\hat{S}_{KM}(t)]$ is given by

$$\hat{Var}[\hat{S}_{KM}(t)] = \left(\hat{S}_{KM}(t)\right)^2 \sum_{f:t_{(f)} \le t} \left[\frac{m_f}{n_f(n_f - m_f)}\right]$$

$t_{(f)}$ = f-ordered failure time
$m_f$ = number of failures at $t_{(f)}$,
$n_f$ = number in the risk set at $t_{(f)}$,

We now describe how to calculate (95%) confidence intervals (CIs) for the estimated Kaplan–Meier (KM) curve.

The 95% CI formula for estimated KM probability at any time point over follow-up has the general large sample form shown on the left, where $\hat{S}_{KM}(t)$denotes the KM survival estimate at time t and $\text{Var}[\hat{S}_{KM}(t)]$ denotes variance of $\hat{S}_{KM}(t)$. The most common approach used to calculate this variance uses the **Greenwood's formula**, also shown on the left.

$$\frac{m_f}{n_f(n_f - m_f)} = \underbrace{\frac{1}{n_f}}_{\text{weight}} \underbrace{\left(\frac{m_f}{n_f - m_f}\right)}_{\text{Conditional risk}}$$

$$= \frac{1}{n_f} P[T > t_{(f)} \mid T \geq t_{(f)}]$$

| t | n | m | $\frac{m}{n(n-m)}$ | $\sum \frac{m}{n(n-m)}$ | S(t) | Var[S(t)] |
|---|---|---|---|---|---|---|
| 0 | 21 | 0 | 0 | 0 | 1 | 0 |
| **6** | **21** | **3** | **0.0079** | **0.0079** | **0.857** | **0.0058** |
| 7 | 17 | 1 | 0.0037 | 0.0116 | 0.807 | 0.0076 |
| **10** | **15** | **1** | **0.0048** | **0.0164** | **0.753** | **0.0093** |
| 13 | 12 | 1 | 0.0076 | 0.0240 | 0.690 | 0.0114 |
| 16 | 11 | 1 | 0.0091 | 0.0330 | 0.628 | 0.0130 |
| 22 | 7 | 1 | 0.0238 | 0.0569 | 0.538 | 0.0164 |
| 23 | 6 | 1 | 0.0333 | 0.0902 | 0.448 | 0.0181 |

6 weeks: $0.857 \pm 1.96\sqrt{(.857)^2\,(0.0079)}$

$= 0.857 \pm 0.149$ **= (.708, 1.006)**

1.0

10 weeks: $\frac{m_3}{n_3(n_3 - m_3)} = \frac{1}{15(14)} = .0048$

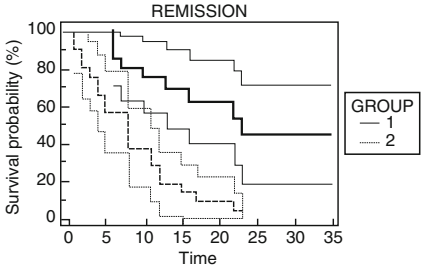
$$\sum_{f:t_{(f)} \leq t=10} \left[\frac{m_f}{n_f(n_f - m_f)}\right]$$

$= 0.0079 + 0.0037 + 0.0048 = 0.0164.$

$\hat{\text{Var}}[\hat{S}_{KM}(10)] = (0.753)^2(.0164) = 0.0093$

95% CI : $0.753 \pm 1.96\sqrt{0.0093} = (\mathbf{0.564}, \mathbf{0.942})$

Same CI at t =11 and 12, since no events occurred at those times.




The summation component of Greenwood's formula essentially gives at each failure time $t_{(f)}$, a weighted (by $1/n_f$) average of the conditional risk of failing at those failure times prior to $t_{(f)}$. Thus, the variance formula gives the square of the KM coordinate at each event time weighted by the cumulative estimate of the risk at time t.

We illustrate how Greenwood's variance is calculated for the treatment group (Group 1) of the remission times data described earlier. The layout for this computation is shown n the left.

At 6 weeks, the estimate of the survival function is 0.857. There were three events at 6 weeks and 21 patients at risk. Therefore, $m_f/n_f(n_f - m_f) = 3/(21\times18) = 0.0079$. As this is the only component of the sum, the variance is then $0.0079 \times 0.857^2 = 0.0058$. The corresponding 95% confidence interval is shown on the left, where the upper level should be modified to 1.

At 10 weeks, the estimate of the survival function is 0.753. There was 1 event at 10 months and 15 patients at risk. Therefore, $m_f/n_f(n_f\text{-}m_f) = 1/(15\times14) = 0.0048$.

There were two other risk components prior to this time, 0.0079 at time 6 and 0.0037 at time 7 and their sum is 0.0164. The variance at 10 weeks is then $0.0164\times0.753^2$, which equals 0.0093.

The 95% CI for the proportion of patients at 10 weeks is shown on the left to have the limits (0.564, 0.942). Note that since the variance is only defined at event times, the 95% confidence interval remains the same at 11 and 12 weeks also.

On the left we show the KM curves and their corresponding 95% confidence intervals for Group 1 (treatment) and Group 2 (placebo) for the remission time data.

## VIII. Confidence intervals for the median survival time

Remission data example:
Group 1 median = 8 weeks
95% CI ?

Formula for 95% CI derived from:

$$\frac{\left(\hat{S}_{\mathrm{KM}}(\mathrm{M}) - 0.5\right)^2}{\hat{\mathrm{Var}}[\hat{S}_{\mathrm{KM}}(\mathrm{M})]} \sim \chi^2_1 \text{ where}$$

M = true (unknown) median survival time, i.e., $S_{KM}(M) = 0.5$

$\hat{S}_{KM}(M)$ = estimated survival probability from KM curve at the true median survival time

$\hat{\mathrm{Var}}[\hat{S}_{\mathrm{KM}}(\mathrm{M})]$ uses Greenwood's formula.

**95% CI for median survival:**

$$\left(\hat{S}_{KM}(t) - 0.5\right)^2 < 3.84 \hat{Var}[\hat{S}_{KM}(t)]$$

| $t_{(i)}$ | $\hat{S}(t)$ | $(\hat{S}(t)-0.5)^2$ | 3.84 $\mathrm{Var}(\hat{S}(t))$ | Inequality satisfied? |
|---|---|---|---|---|
| 0 | 1 | 0.250 | - | |
| 1 | 0.90 | 0.160 | 0.016 | N |
| 2 | 0.81 | 0.096 | 0.028 | N |
| 3 | 0.76 | 0.068 | 0.033 | N |
| **4** | **0.67** | **0.029** | **0.041** | **Y** |
| **5** | **0.57** | **0.005** | **0.045** | **Y** |
| **8** | **0.38** | **0.014** | **0.044** | **Y** |
| 11 | 0.29 | 0.44 | 0.038 | N |
| 12 | 0.19 | 0.096 | 0.028 | N |
| 15 | 0.14 | 0.130 | 0.022 | N |
| 17 | 0.10 | 0.160 | 0.016 | N |
| 22 | 0.05 | 0.203 | 0.008 | N |
| 23 | 0.00 | 0.050 | 0 | N |

0.096 > 0.028 so the inequality is not satisfied
0.014 > 0.044 and the inequality is satisfied

Caution (ref B&C, 1982): upper limit should be adjusted to reflect censoring, e.g., SAS's LIFETEST adjusts above 95% CI from (**4,8**) to (**4, 11**).

Returning again to the remission time dataset, we now consider the calculation of the 95% CI for the median of the remission times for the placebo group. Recall that the median survival for this group is 8 weeks.

Brookmeyer and Crowley (1982) proposed a simple way to calculate the CI for the median survival time based on the fact that the square of a *standardized function* of the survival curve around the true (unknown) median value (M) is asymptotically $\chi^2$distributed. This relationship is shown mathematically on the left.

Using the above result about the standardized survival curve, a general formula for the 95% CI for the median survival is provided by the inequality expression shown on the left. **The times for which this inequality holds are plausible values of the true median, while the boundaries represent upper and lower times for the 95% CI for the median**. The lower boundary may be 0 and the upper boundary may not always exist.

For the remission time data, the calculation of the CI around the median of 8 weeks is given in the table shown on the left. Since the inequality in the CI formula is satisfied in the range t=4 weeks to 8 weeks, the resulting 95% CI is (**4,8**).

Brookmeyer and Crowley (B&C) caution that the upper limit for these estimates should be adjusted to reflect censoring. They recommend reporting semiopen intervals that extend one event beyond the event that satisfies the inequality; some packages (e.g., SAS, R) incorporate this recommendation.

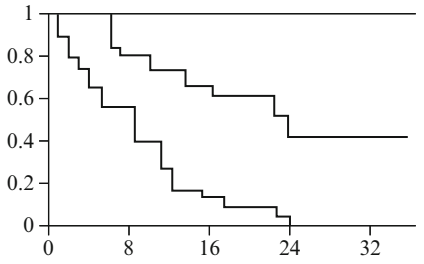
**In this example, the 95% CI for the median survival time obtained from SAS output is given by the interval (4, 11)**, with the upper limit now extended beyond 8 weeks but not including the event at month 11.

## IX. Summary

KM curves:

We now briefly summarize this presentation. First, we described how to estimate and graph survival curves using the Kaplan-Meier (KM) method.

KM curves:




$t_{(f)}$: $f$th ordered failure time

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \hat{P}r[T > t_{(i)} | T \geq t_{(i)}]$$
$$= \hat{S}(t_{(f-1)}) \times \hat{P}r(T > t_{(f)} | T \geq t_{(f)})$$

To compute KM curves, we must form a data layout that orders the failure times from smallest to largest. For each ordered failure time, the estimated survival probability is computed using the **product limit formula** shown here. Alternatively, this estimate can be computed as the product of the survival estimate for the previous failure time multiplied by the conditional probability of surviving past the current failure time.

Log-rank test:

$H_0$: common survival curve for all groups

$$\text{Log - rank statistic} = \frac{(O_2 - E_2)^2}{\text{Var}(O_2 - E_2)}$$

log–rank statistic $\sim \chi^2$ with $G - 1$ $df$ under $H_0$

$G$ = # of groups

When survival curves are being compared, the log–rank test gives a statistical test of the null hypothesis of a common survival curve. For two groups, the log–rank statistic is based on the summed observed minus expected scores for a given group and its variance estimate. For several groups, a computer should always be used since the log–rank formula is more complicated mathematically. The test statistic is approximately chi-square in large samples with $G - 1$ degrees of freedom, where $G$ denotes the number of groups being compared.

Greenwood's Variance formula:

$$\hat{Var}[\hat{S}_{KM}(t)] = (\hat{S}_{KM}(t))^2 \times \sum_{f:t_{(f)} \leq t} \left[ \frac{m_f}{n_f (n_f - m_f)} \right]$$

Large sample confidence intervals for KM curves can be computed based on Greenwood's formula, shown at the left, for the variance of an estimated KM survival probability.

95% CI for KM:

$$\hat{S}_{KM}(t) \pm 1.96\sqrt{\hat{Var}\left[\hat{S}_{KM}(t)\right]}$$

The expression for the 95% confidence interval is shown below Greenwood's formula.

95% CI for median survival:

$$\left(\hat{S}_{KM}(t) - 0.5\right)^2 3.84\,\hat{Var}\left[\hat{S}_{KM}(t)\right]$$

A large sample confidence interval formula for the median of the KM curve can also be computed using the inequality formula shown here on the left. The upper and lower boundaries of t for which this inequality holds provide the 95% confidence limits.

## Chapters

1. Introduction
✓2. Kaplan-Meier Survival Curves and the Log-Rank Test

This presentation is now complete. You can review this presentation using the detailed outline that follows and then try the practice exercises and test.

Next:

3. The Cox Proportional Hazards Model and Its Characteristics

Chapter 3 introduces the Cox proportional hazards (PH) model, which is the most popular mathematical modeling approach for estimating survival curves when considering several explanatory variables simultaneously.

## Detailed Outline

**I. Review** (pages 58–60)

A. The outcome variable is (survival) time until an event (failure) occurs.

B. Key problem: **censored data**, i.e., don't know survival time exactly.

C. Notation:
T = survival time random variable
t = specific value of T
d = (0, 1) variable for failure/censorship status
S(t) = survivor function
h(t) = hazard function

D. Properties of survivor function:
   i. theoretically, graph is smooth curve, decreasing from $S(t) = 1$ at time $t = 0$ to $S(t) = 0$ at $t = \infty$;
   ii. in practice, graph is step function.

E. Properties of $h(t)$:
   i. instantaneous potential for failing given survival up to time;
   ii. $h(t)$ is a rate; ranges from 0 to $\infty$.

F. Relationship of $S(t)$ to $h(t)$: if you know one you can determine the other.

G. Goals of survival analysis: estimation of survivor and hazard functions; comparisons and relationships of explanatory variables to survival.

H. Data layouts
   i. for the computer;
   ii. for understanding the analysis: involves **risk sets**.

**II. An Example of Kaplan-Meier Curves** (pages 61–65)

A. Data are from study of remission times in weeks for two groups of leukemia patients (21 in each group).

B. Group 1 (treatment group) has several censored observations, whereas group 2 has no censored observations.

C. Table of ordered failure times is provided for each group.

D. For group 2 (all noncensored), survival probabilities are estimated directly and plotted. Formula used is

$$\hat{S}(t_{(f)}) = \frac{\#\text{ surviving past } t_{(f)}}{21}.$$

E. Alternative approach for group 2 is given by a **product limit** formula.

F. For group 1, survival probabilities calculated by multiplying estimate for immediately preceding failure time by a conditional probability of surviving past current failure time, i.e.,

$$\hat{S}_{(f)} = \hat{S}_{(f-1)}\hat{P}r\left[T > t_{(f)} \middle| T \geq t_{(f)}\right].$$

**III. General Features of KM Curves** (pages 66–67)

A. Two alternative general formulae:

$$S_{(f)} = \prod_{i=1}^{f} \Pr\left[T > t_{(i)} \middle| T \geq t_{(i)}\right] \quad \begin{matrix}\text{(product limit}\\ \text{formula)}\end{matrix}$$

$$S_{(f)} = S_{(f-1)} \Pr\left[T > t_{(f)} \middle| T \geq t_{(f)}\right]$$

B. Second formula derived from probability rule:

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B|A)$$

**IV. The Log-Rank Test for Two Groups** (pages 67–71)

A. Large sample chi-square test; provides overall comparison of KM curves.

B. Uses observed versus expected counts over categories of outcomes, where categories are defined by ordered failure times for entire set of data.

C. Example provided using remission data involving two groups:

i. expanded table described to show how expected and observed minus expected cell counts are computed.

ii. for $i$th group at time $f$, where $i = 1$ or 2:
observed counts $= m_{if}$,
expected counts $= e_{if}$, where
expected counts = (proportion in risk set) × (# failures over both groups),

$$\text{i.e., } e_{if} = \left(\frac{n_{if}}{n_{1f} + n_{2f}}\right)(m_{1f} + m_{2f}).$$

D. Log-rank statistic for two groups:

$$\boxed{\frac{(O_i - E_i)^2}{\text{Var}(O_i - E_i)}},$$

where $i = 1.\ 2$,

$$O_i - E_i = \sum_f \left(m_{if} - e_{if}\right), \text{ and}$$

$$\text{Var}(O_i - E_i) = \sum_f \frac{n_{1f} n_{2f} \left(m_{1f} + m_{2f}\right)\left(n_{1f} + n_{2f} - m_{1f} - m_{2f}\right)}{\left(n_{1f} + n_{2f}\right)^2 \left(n_{1f} + n_{2f} - 1\right)}$$

$$i = 1,\ 2$$

E. $H_0$: no difference between survival curves.

F. Log-rank statistic $\sim \chi^2$ with 1 *df* under $H_0$.

G. Approximate formula:

$$X^2 = \sum_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i}, \text{ where } G = 2 = \#\text{ of groups}$$

H. Remission data example: Log-rank statistic = 16.793, whereas $X^2 = 15.276$.

**V. The Log-Rank Test for Several Groups** (pages 71–73)

A. Involves variances and covariances; matrix formula in Appendix.

B. Use computer for calculations.

C. Under $H_0$, log–rank statistic $\sim\chi^2$ with $G - 1$ *df*, where $G = \#$ of groups.

D. Example provided using vets.dat with interval variable "performance status"; this variable is categorized into $G = 3$ groups, so *df* for log–-rank test is $G - 1 = 2$, log–rank statistic is 29.181 ($P = 0.0$).

**VI. Alternatives to the Log-Rank Test** (pages 73–78)

A. Alternative tests supported by Stata: Wilcoxen, Tarone-Ware, Peto, and Flemington-Harrington.

B. Alternative tests differ by applying different weights at the j-th failure time.

C. The choice of alternative depends on the reason for the belief that the effect is more pronounced towards the beginning (or end) of the survival function.

D. The stratified-log–rank test is a variation of the log–rank test that controls for one or more stratified variables.

**VII. Confidence Intervals for KM Curves** (pages 78–79)

A. General form of 95% CI:

$$\hat{S}_{KM}(\mathrm{t}) \pm 1.96\sqrt{V\hat{a}r[\hat{S}_{KM}(\mathrm{t})]}$$

B. $\hat{\mathrm{V}}\mathrm{ar}[\hat{\mathrm{S}}_{\mathrm{KM}}(\mathrm{t})]$uses Greenwood's formula:

$$V\hat{a}r[\hat{S}_{KM}(\mathrm{t})] = \left(\hat{S}_{KM}(\mathrm{t})\right)^2 \sum_{f:t_{(f)}\le t}\left[\frac{m_f}{n_f(n_f - m_f)}\right]$$

C. Example using Remission Time Data

**VIII. Confidence Intervals for the Median Survival Time** (page 80)

A. General form of 95% CI where values of t satisfying the following inequality provide confidence limits:

$$\left(\hat{S}_{KM}(t) - 0.5\right)^2 < 3.84Var[\hat{S}_{KM}(t)]$$

B. Example using Remission Time Data

**IX. Summary** (page 81)

## Practice Exercises

1. The following data are a sample from the 1967-1980 Evans County study. Survival times (in years) are given for two study groups, each with 25 participants. Group 1 has no history of chronic disease (CHR = 0), and group 2 has a positive history of chronic disease (CHR = 1):

   Group 1 (CHR = 0): 12.3+, 5.4, 8.2, 12.2+, 11.7, 10.0, 5.7, 9.8, 2.6, 11.0, 9.2, 12.1+, 6.6, 2.2, 1.8, 10.2, 10.7, 11.1, 5.3, 3.5, 9.2, 2.5, 8.7, 3.8, 3.0

   Group 2 (CHR = 1): 5.8, 2.9, 8.4, 8.3, 9.1, 4.2, 4.1, 1.8, 3.1, 11.4, 2.4, 1.4, 5.9, 1.6, 2.8, 4.9, 3.5, 6.5, 9.9, 3.6, 5.2, 8.8, 7.8, 4.7, 3.9

   a. Fill in the missing information in the following table of ordered failure times for groups 1 and 2:

| Group 1 | | | | | Group 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $S(t_{(f)})$ | $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $S(t_{(f)})$ |
| 0.0 | 25 | 0 | 0 | 1.00 | 0.0 | 25 | 0 | 0 | 1.00 |
| 1.8 | 25 | 1 | 0 | .96 | 1.4 | 25 | 1 | 0 | .96 |
| 2.2 | 24 | 1 | 0 | .92 | 1.6 | 24 | 1 | 0 | .92 |
| 2.5 | 23 | 1 | 0 | .88 | 1.8 | 23 | 1 | 0 | .88 |
| 2.6 | 22 | 1 | 0 | .84 | 2.4 | 22 | 1 | 0 | .84 |
| 3.0 | 21 | 1 | 0 | .80 | 2.8 | 21 | 1 | 0 | .80 |
| 3.5 | 20 | | | | 2.9 | 20 | 1 | 0 | .76 |
| 3.8 | 19 | 1 | 0 | .72 | 3.1 | 19 | 1 | 0 | .72 |
| 5.3 | 18 | 1 | 0 | .68 | 3.5 | 18 | 1 | 0 | .68 |
| 5.4 | 17 | 1 | 0 | .64 | 3.6 | 17 | 1 | 0 | .64 |
| 5.7 | 16 | 1 | 0 | .60 | 3.9 | | | | |
| 6.6 | 15 | 1 | 0 | .56 | 4.1 | | | | |
| 8.2 | 14 | 1 | 0 | .52 | 4.2 | | | | |
| 8.7 | 13 | 1 | 0 | .48 | 4.7 | 13 | 1 | 0 | .48 |
| 9.2 | | | | | 4.9 | 12 | 1 | 0 | .44 |
| 9.8 | 10 | 1 | 0 | .36 | 5.2 | 11 | 1 | 0 | .40 |
| 10.0 | 9 | 1 | 0 | .32 | 5.8 | 10 | 1 | 0 | .36 |
| 10.2 | 8 | 1 | 0 | .28 | 5.9 | 9 | 1 | 0 | .32 |
| 10.7 | 7 | 1 | 0 | .24 | 6.5 | 8 | 1 | 0 | .28 |
| 11.0 | 6 | 1 | 0 | .20 | 7.8 | 7 | 1 | 0 | .24 |
| 11.1 | 5 | 1 | 0 | .16 | 8.3 | 6 | 1 | 0 | .20 |
| 11.7 | 4 | | | | 8.4 | 5 | 1 | 0 | .16 |
| | | | | | 8.8 | 4 | 1 | 0 | .12 |
| | | | | | 9.1 | | | | |
| | | | | | 9.9 | | | | |
| | | | | | 11.4 | 1 | 1 | 0 | .00 |

   b. Based on your results in part a, plot the KM curves for groups 1 and 2 on the same graph. Comment on how these curves compare with each other.

c. Fill in the following expanded table of ordered failure times to allow for the computation of expected and observed minus expected values at each ordered failure time. Note that your new table here should combine both groups of ordered failure times into one listing and should have the following format:

| $t_{(f)}$ | $m_{1f}$ | $m_{2f}$ | $n_{1f}$ | $n_{2f}$ | $e_{1f}$ | $e_{2f}$ | $m_{1f}-e1_f$ | $m_{2f}-e2_f$ |
|---|---|---|---|---|---|---|---|---|
| 1.4 | 0 | 1 | 25 | 25 | .500 | .500 | −.500 | .500 |
| 1.6 | 0 | 1 | 25 | 24 | .510 | .490 | −.510 | .510 |
| 1.8 | 1 | 1 | 25 | 23 | 1.042 | .958 | −.042 | .042 |
| 2.2 | 1 | 0 | 24 | 22 | .522 | .478 | .478 | −.478 |
| 2.4 | 0 | 1 | 23 | 22 | .511 | .489 | −.511 | .511 |
| 2.5 | 1 | 0 | 23 | 21 | .523 | .477 | .477 | −.477 |
| 2.6 | 1 | 0 | 22 | 21 | .516 | .484 | .484 | −.484 |
| 2.8 | 0 | 1 | 21 | 21 | .500 | .500 | −.500 | .500 |
| 2.9 | 0 | 1 | 21 | 20 | .512 | .488 | −.512 | .512 |
| 3.0 | 1 | 0 | 21 | 19 | .525 | .475 | .475 | −.475 |
| 3.1 | | | | | | | | |
| 3.5 | | | | | | | | |
| 3.6 | | | | | | | | |
| 3.8 | | | | | | | | |
| 3.9 | 0 | 1 | 18 | 16 | .529 | .471 | −.529 | .529 |
| 4.1 | 0 | 1 | 18 | 15 | .545 | .455 | −.545 | .545 |
| 4.2 | 0 | 1 | 18 | 14 | .563 | .437 | −.563 | .563 |
| 4.7 | 0 | 1 | 18 | 13 | .581 | .419 | −.581 | .581 |
| 4.9 | 0 | 1 | 18 | 12 | .600 | .400 | −.600 | .600 |
| 5.2 | 0 | 1 | 18 | 11 | .621 | .379 | −.621 | .621 |
| 5.3 | 1 | 0 | 18 | 10 | .643 | .357 | .357 | −.357 |
| 5.4 | 1 | 0 | 17 | 10 | .630 | .370 | .370 | −.370 |
| 5.7 | 1 | 0 | 16 | 10 | .615 | .385 | .385 | −.385 |
| 5.8 | 0 | 1 | 15 | 10 | .600 | .400 | −.600 | .600 |
| 5.9 | 0 | 1 | 15 | 9 | .625 | .375 | −.625 | .625 |
| 6.5 | 0 | 1 | 15 | 8 | .652 | .348 | −.652 | .652 |
| 6.6 | 1 | 0 | 15 | 7 | .682 | .318 | .318 | −.318 |
| 7.8 | 0 | 1 | 14 | 7 | .667 | .333 | −.667 | .667 |
| 8.2 | 1 | 0 | 14 | 6 | .700 | .300 | .300 | −.300 |
| 8.3 | 0 | 1 | 13 | 6 | .684 | .316 | −.684 | .684 |
| 8.4 | 0 | 1 | 13 | 5 | .722 | .278 | −.722 | .722 |
| 8.7 | 1 | 0 | 13 | 4 | .765 | .235 | .335 | −.335 |
| 8.8 | 0 | 1 | 12 | 4 | .750 | .250 | −.750 | .750 |
| 9.1 | 0 | 1 | 12 | 3 | .800 | .200 | −.800 | .800 |
| 9.2 | | | | | | | | |
| 9.8 | | | | | | | | |
| 9.9 | | | | | | | | |
| 10.0 | 1 | 0 | 9 | 1 | .900 | .100 | .100 | −.100 |
| 10.2 | 1 | 0 | 8 | 1 | .888 | .112 | .112 | −.112 |
| 10.7 | 1 | 0 | 7 | 1 | .875 | .125 | .125 | −.125 |
| 11.0 | 1 | 0 | 6 | 1 | .857 | .143 | .143 | −.143 |
| 11.1 | 1 | 0 | 5 | 1 | .833 | .167 | .167 | −.167 |
| 11.4 | 0 | 1 | 4 | 1 | .800 | .200 | −.800 | .800 |
| 11.7 | 1 | 0 | 4 | 0 | 1.000 | .000 | .000 | .000 |
| Totals | 22 | 25 | | | 30.79 | 16.21 | | |

d. Use the results in part c to compute the log–rank statistic. Use this statistic to carry out the log–rank test for these data. What is your null hypothesis and how is the test statistic distributed under this null hypothesis? What are your conclusions from the test?

2. The following data set called "anderson.dat" consists of remission survival times on 42 leukemia patients, half of whom get a certain new treatment therapy and the other half of whom get a standard treatment therapy. The exposure variable of interest is treatment status ($Rx = 0$ if new treatment, $Rx = 1$ if standard treatment). Two other variables for control as potential confounders are log white blood cell count (i.e., logwbc) and sex. Failure status is defined by the relapse variable (0 if censored, 1 if failure). The data set is listed as follows:

| Subj | Survt | Relapse | Sex | log WBC | *Rx* |
|---|---|---|---|---|---|
| 1 | 35 | 0 | 1 | 1.45 | 0 |
| 2 | 34 | 0 | 1 | 1.47 | 0 |
| 3 | 32 | 0 | 1 | 2.20 | 0 |
| 4 | 32 | 0 | 1 | 2.53 | 0 |
| 5 | 25 | 0 | 1 | 1.78 | 0 |
| 6 | 23 | 1 | 1 | 2.57 | 0 |
| 7 | 22 | 1 | 1 | 2.32 | 0 |
| 8 | 20 | 0 | 1 | 2.01 | 0 |
| 9 | 19 | 0 | 0 | 2.05 | 0 |
| 10 | 17 | 0 | 0 | 2.16 | 0 |
| 11 | 16 | 1 | 1 | 3.60 | 0 |
| 12 | 13 | 1 | 0 | 2.88 | 0 |
| 13 | 11 | 0 | 0 | 2.60 | 0 |
| 14 | 10 | 0 | 0 | 2.70 | 0 |
| 15 | 10 | 1 | 0 | 2.96 | 0 |
| 16 | 9 | 0 | 0 | 2.80 | 0 |
| 17 | 7 | 1 | 0 | 4.43 | 0 |
| 18 | 6 | 0 | 0 | 3.20 | 0 |
| 19 | 6 | 1 | 0 | 2.31 | 0 |
| 20 | 6 | 1 | 1 | 4.06 | 0 |
| 21 | 6 | 1 | 0 | 3.28 | 0 |
| 22 | 23 | 1 | 1 | 1.97 | 1 |
| 23 | 22 | 1 | 0 | 2.73 | 1 |
| 24 | 17 | 1 | 0 | 2.95 | 1 |
| 25 | 15 | 1 | 0 | 2.30 | 1 |
| 26 | 12 | 1 | 0 | 1.50 | 1 |
| 27 | 12 | 1 | 0 | 3.06 | 1 |
| 28 | 11 | 1 | 0 | 3.49 | 1 |

(*Continued on next page*)

| Subj | Survt | Relapse | Sex | log WBC | *Rx* |
|---|---|---|---|---|---|
| 29 | 11 | 1 | 0 | 2.12 | 1 |
| 30 | 8 | 1 | 0 | 3.52 | 1 |
| 31 | 8 | 1 | 0 | 3.05 | 1 |
| 32 | 8 | 1 | 0 | 2.32 | 1 |
| 33 | 8 | 1 | 1 | 3.26 | 1 |
| 34 | 5 | 1 | 1 | 3.49 | 1 |
| 35 | 5 | 1 | 0 | 3.97 | 1 |
| 36 | 4 | 1 | 1 | 4.36 | 1 |
| 37 | 4 | 1 | 1 | 2.42 | 1 |
| 38 | 3 | 1 | 1 | 4.01 | 1 |
| 39 | 2 | 1 | 1 | 4.91 | 1 |
| 40 | 2 | 1 | 1 | 4.48 | 1 |
| 41 | 1 | 1 | 1 | 2.80 | 1 |
| 42 | 1 | 1 | 1 | 5.00 | 1 |

a. Suppose we wish to describe KM curves for the variable logwbc. Because logwbc is continuous, we need to categorize this variable before we compute KM curves. Suppose we categorize logwbc into three categories - low, medium, and high - as follows:

   low (0–2.30), $n = 11$;
   medium (2.31–3.00), $n = 14$;
   high (>3.00), $n = 17$.

   Based on this categorization, compute and graph KM curves for each of the three categories of logwbc. (You may use a computer program to assist you or you can form three tables of ordered failure times and compute KM probabilities directly.)

b. Compare the three KM plots you obtained in part a. How are they different?

c. Below is an edited printout of the log–rank test comparing the three groups.

| Group | Events observed | Events expected |
|---|---|---|
| 1 | 4 | 13.06 |
| 2 | 10 | 10.72 |
| 3 | 16 | 6.21 |
| Total | 30 | 30.00 |

   Log-rank = chi2(2) = 26.39
   P-value = Pr > chi2 = 0.0000

   What do you conclude about whether or not the three survival curves are the same?

## Test

To answer the questions below, you will need to use a computer program (from SAS, Stata, SPSS, R or any other package you are familiar with) that computes and plots KM curves and computes the log–rank test. Freely downloadable files can be obtained from weblink http://www.sph.emory.edu/dkleinb/surv3.htm.

1. For the vets.dat data set described in the presentation:
   a. Obtain KM plots for the two categories of the variable cell type 1 (1 = large, 0 = other). Comment on how the two curves compare with each other. Carry out the log–rank, and draw conclusions from the test(s).
   b. Obtain KM plots for the four categories of cell type - large, adeno, small, and squamous. Note that you will need to recode the data to define a single variable which numerically distinguishes the four categories (e.g., 1 = large, 2 = adeno, etc.). As in part a, compare the four KM curves. Also, carry out the log–rank for the equality of the four curves and draw conclusions.
2. The following questions consider a data set from a study by Caplehorn et al. ("Methadone Dosage and Retention of Patients in Maintenance Treatment," *Med. J. Aust*., 1991). These data comprise the times in days spent by heroin addicts from entry to departure from one of two methadone clinics. There are two further covariates, namely, prison record and methadone dose, believed to affect the survival times. The data set name is addicts.dat. A listing of the variables is given below:

   Column 1: Subject ID

   Column 2: Clinic (1 or 2)

   Column 3: Survival status (0 = censored, 1 = departed from clinic)

   Column 4: Survival time in days

   Column 5: Prison record (0 = none, 1 = any)

   Column 6: Methadone dose (mg/day)

a. Compute and plot the KM plots for the two categories of the "clinic" variable and comment on the extent to which they differ.

b. A printout of the log–rank and Wilcoxon tests (using Stata) is provided below. What are your conclusions from this printout?

| Group | Events observed | Events expected |
|---|---|---|
| 1 | 122 | 90.91 |
| 2 | 28 | 59.09 |
| Total | 150 | 150.00 |

Log-rank = chi2(1) = 27.89
P-value = Pr > chi2 = 0.0000
Wilcoxon = chi2(1) = 11.63
P-value = Pr > chi2 = 0.0007

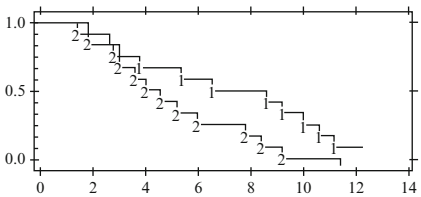
c. Compute and evaluate KM curves and the log–rank test for comparing suitably chosen categories of the variable "Methadone dose." Explain how you determined the categories for this variable.

## Answers to Practice Exercises

1. a.

| Group 1 | | | | | Group 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $S(t_{(f)})$ | $t_{(f)}$ | $n_f$ | $m_f$ | $q_f$ | $S(t_{(f)})$ |
| 0.0 | 25 | 0 | 0 | 1.00 | 0.0 | 25 | 0 | 0 | 1.00 |
| 1.8 | 25 | 1 | 0 | .96 | 1.4 | 25 | 1 | 0 | .96 |
| 2.2 | 24 | 1 | 0 | .92 | 1.6 | 24 | 1 | 0 | .92 |
| 2.5 | 23 | 1 | 0 | .88 | 1.8 | 23 | 1 | 0 | .88 |
| 2.6 | 22 | 1 | 0 | .84 | 2.4 | 22 | 1 | 0 | .84 |
| 3.0 | 21 | 1 | 0 | .80 | 2.8 | 21 | 1 | 0 | .80 |
| 3.5 | 20 | 1 | 0 | .76 | 2.9 | 20 | 1 | 0 | .76 |
| 3.8 | 19 | 1 | 0 | .72 | 3.1 | 19 | 1 | 0 | .72 |
| 5.3 | 18 | 1 | 0 | .68 | 3.5 | 18 | 1 | 0 | .68 |
| 5.4 | 17 | 1 | 0 | .64 | 3.6 | 17 | 1 | 0 | .64 |
| 5.7 | 16 | 1 | 0 | .60 | 3.9 | 16 | 1 | 0 | .60 |
| 6.6 | 15 | 1 | 0 | .56 | 4.1 | 15 | 1 | 0 | .56 |
| 8.2 | 14 | 1 | 0 | .52 | 4.2 | 14 | 1 | 0 | .52 |
| 8.7 | 13 | 1 | 0 | .48 | 4.7 | 13 | 1 | 0 | .48 |
| 9.2 | 12 | 2 | 0 | .40 | 4.9 | 12 | 1 | 0 | .44 |
| 9.8 | 10 | 1 | 0 | .36 | 5.2 | 11 | 1 | 0 | .40 |
| 10.0 | 9 | 1 | 0 | .32 | 5.8 | 10 | 1 | 0 | .36 |
| 10.2 | 8 | 1 | 0 | .28 | 5.9 | 9 | 1 | 0 | .32 |
| 10.7 | 7 | 1 | 0 | .24 | 6.5 | 8 | 1 | 0 | .28 |
| 11.0 | 6 | 1 | 0 | .20 | 7.8 | 7 | 1 | 0 | .24 |
| 11.1 | 5 | 1 | 0 | .16 | 8.3 | 6 | 1 | 0 | .20 |
| 11.7 | 4 | 1 | 3 | .12 | 8.4 | 5 | 1 | 0 | .16 |
| | | | | | 8.8 | 4 | 1 | 0 | .12 |
| | | | | | 9.1 | 3 | 1 | 0 | .08 |
| | | | | | 9.9 | 2 | 1 | 0 | .04 |
| | | | | | 11.4 | 1 | 1 | 0 | .00 |

b. KM curves for CHR data:




Group 1 appears to have consistently better survival prognosis than group 2. However, the KM curves are very close during the first 4 years, but are quite separate after 4 years, although they appear to come close again around 12 years.

c. Using the expanded table format, the following information is obtained:

| $t_{(f)}$ | $m_{1f}$ | $m_{2f}$ | $n_{1f}$ | $n_{2f}$ | $e_{1f}$ | $e_{2f}$ | $m_{1f}-e_{1f}$ | $m_{2f}-e_{2f}$ |
|---|---|---|---|---|---|---|---|---|
| 1.4 | 0 | 1 | 25 | 25 | .500 | .500 | −.500 | .500 |
| 1.6 | 0 | 1 | 25 | 24 | .510 | .490 | −.510 | .510 |
| 1.8 | 1 | 1 | 25 | 23 | 1.042 | .958 | −.042 | .042 |
| 2.2 | 1 | 0 | 24 | 22 | .522 | .478 | .478 | −.478 |
| 2.4. | 0 | 1 | 23 | 22 | .511 | .489 | −.511 | .511 |
| 2.5. | 1 | 0 | 23 | 21 | .523 | .477 | .477 | −.477 |
| 2.6 | 1 | 0 | 22 | 21 | .516 | .484 | .484 | −.484 |
| 2.8 | 0 | 1 | 21 | 21 | .500 | .500 | −.500 | .500 |
| 2.9 | 0 | 1 | 21 | 20 | .512 | .488 | −.512 | .512 |
| 3.0 | 1 | 0 | 21 | 19 | .525 | .475 | .475 | −.475 |
| 3.1 | 0 | 1 | 20 | 19 | .513 | .487 | −.513 | .513 |
| 3.5 | 1 | 1 | 20 | 18 | 1.053 | .947 | −.053 | .053 |
| 3.6 | 0 | 1 | 19 | 17 | .528 | .472 | −.528 | .528 |
| 3.8 | 1 | 0 | 19 | 16 | .543 | .457 | .457 | −.457 |
| 3.9 | 0 | 1 | 18 | 16 | .529 | .471 | −.529 | .529 |
| 4.1 | 0 | 1 | 18 | 15 | .545 | .455 | −.545 | .545 |
| 4.2 | 0 | 1 | 18 | 14 | .563 | .437 | −.563 | .563 |
| 4.7 | 0 | 1 | 18 | 13 | .581 | .419 | −.581 | .581 |
| 4.9 | 0 | 1 | 18 | 12 | .600 | .400 | −.600 | .600 |
| 5.2 | 0 | 1 | 18 | 11 | .621 | .379 | −.621 | .621 |
| 5.3 | 1 | 0 | 18 | 10 | .643 | .357 | .357 | −.357 |
| 5.4 | 1 | 0 | 17 | 10 | .630 | .370 | .370 | −.370 |
| 5.7 | 1 | 0 | 16 | 10 | .615 | .385 | .385 | −.385 |
| 5.8 | 0 | 1 | 15 | 10 | .600 | .400 | −.600 | .600 |
| 5.9 | 0 | 1 | 15 | 9 | .625 | .375 | −.625 | .625 |
| 6.5 | 0 | 1 | 15 | 8 | .652 | .348 | −.652 | .652 |
| 6.6 | 1 | 0 | 15 | 7 | .682 | .318 | .318 | −.318 |
| 7.8 | 0 | 1 | 14 | 7 | .667 | .333 | −.667 | .667 |
| 8.2 | 1 | 0 | 14 | 6 | .700 | .300 | .300 | −.300 |
| 8.3 | 0 | 1 | 13 | 6 | .684 | .316 | −.684 | .684 |
| 8.4 | 0 | 1 | 13 | 5 | .722 | .278 | −.722 | .722 |
| 8.7 | 1 | 0 | 13 | 4 | .765 | .235 | .335 | −.335 |
| 8.8 | 0 | 1 | 12 | 4 | .750 | .250 | −.750 | .750 |
| 9.1 | 0 | 1 | 12 | 3 | .800 | .200 | −.800 | .800 |
| 9.2 | 2 | 0 | 12 | 2 | 1.714 | .286 | .286 | −.286 |
| 9.8 | 1 | 0 | 10 | 2 | .833 | .167 | .167 | −.167 |
| 9.9 | 0 | 1 | 9 | 2 | .818 | .182 | −.818 | .818 |
| 10.0 | 1 | 0 | 9 | 1 | .900 | .100 | .100 | −.100 |
| 10.2 | 1 | 0 | 8 | 1 | .888 | .112 | .112 | −.112 |
| 10.7 | 1 | 0 | 7 | 1 | .875 | .125 | .125 | −.125 |
| 11.0 | 1 | 0 | 6 | 1 | .857 | .143 | .143 | −.143 |
| 11.1 | 1 | 0 | 5 | 1 | .833 | .167 | .167 | −.167 |
| 11.4 | 0 | 1 | 4 | 1 | .800 | .200 | −.800 | .800 |
| 11.7 | 1 | 0 | 4 | 0 | 1.000 | .000 | .000 | .000 |
| Totals | 22 | 25 | | | 30.79 | 16.21 | −8.790 | 8.790 |

d. The log–rank statistic can be computed from the totals of the expanded table using the formulae:

$$\text{log-rank statistic} = \frac{(O_i - E_i)^2}{\widehat{\text{Var}}(O_i - E_i)}$$

$$\text{Var}(O_i - E_i) = \sum_i \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$

The variance turns out to be 9.658, so that the log–-rank statistic is $(8.79)^2/9.658 = 7.993$.

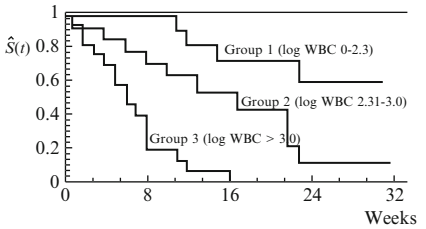
Using Stata, the results for the log–rank test are given as follows:

| Group | Events observed | Events expected |
|---|---|---|
| 1 | 22 | 30.79 |
| 2 | 25 | 16.21 |
| Total | 47 | 47.00 |

Log-rank = chi2(1) = 7.99
P-value = Pr > chi2 = 0.0047

The log–rank test gives highly significant results. This indicates that there is a significant difference in survival between the two groups.

2. a. For the Anderson dataset, the KM plots for the three categories of log WBC are shown below:




b. The KM curves are quite different with group 1 having consistently better survival prognosis than group 2, and group 2 having consistently better survival prognosis than group 3. Note also that the difference between group 1 and 2 is about the same over time, whereas group 2 appears to diverge from group 3 as time increases.

c. The log–rank statistic (26.391) is highly significant with P-values equal to zero to three decimal places. These results indicate that there is some overall difference between the three curves.

## Appendix: Matrix Formula for the Log-Rank Statistic for Several Groups

For $i = 1, 2, \ldots, G$ and $f = 1, 2, \ldots, k$, where $G =$ # of groups and $k =$ # of distinct failure times,

$n_{if} =$ # at risk in $i$th group at $f$th ordered failure time

$m_{if} =$ observed # of failures in $i$th group at $f$th ordered failure time

$e_{if} =$ expected # of failures in $i$th group at $f$th ordered failure time

$$= \left(\frac{n_{if}}{n_{1f} + n_{2f}}\right)\left(m_{1f} + m_{2f}\right)$$

$$n_f = \sum_{i=1}^{G} n_{if}$$

$$m_f = \sum_{i=1}^{G} m_{if}$$

$$O_i - E_i = \sum_{f=1}^{k} \left(m_{if} - e_{if}\right)$$

$$\mathrm{Var}(O_i - E_i) = \sum_{f=1}^{k} \left(\frac{n_{if}\left(n_f - n_{if}\right)m_{if}\left(n_f - m_f\right)}{n_f^2\left(n_f - 1\right)}\right)$$

$$\mathrm{Cov}(O_i - E_i,\ O_l - E_l) = \sum_{f=1}^{k} \left(\frac{-n_{if}\, n_{lf}\, m_f\left(n_f - m_f\right)}{n_f^2\left(n_f - 1\right)}\right)$$

$$\mathbf{d} = (O_1 - E_1,\ O_2 - E_2,\ \ldots,\ O_{G-1} - E_{G-1})'$$

$$\mathbf{V} = ((v_{il}))$$

where $v_{ii} = \mathrm{Var}\,(O_i - E_i)$ and $v_{il} = \mathrm{Cov}\,(O_i - E_i, O_l - E_l)$ for $i = 1, 2, \ldots, G - 1$; $l = 1, 2, \ldots, G - 1$.

Then, the log–rank statistic is given by the matrix product formula:

$$\boxed{\text{Log-rank statistic} \;=\; \mathbf{d}'\mathbf{V}^{-1}\mathbf{d}}$$

which has approximately a chi-square distribution with $G - 1$ degrees of freedom under the null hypothesis that all $G$ groups have a common survival curve.