Detailed Outline

I. Review (pages 48–50)

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
δ = (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 ∞.

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 51–56)

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

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

E. Alternative formula 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.,

$$\hat{S}_{(j)} = \hat{S}_{(j-1)}\,\hat{\Pr}\left[T > t_{(j)} \,\middle|\, T \geq t_{(j)}\right].$$

III. General Features of KM Curves (pages 56–58)

A. Two alternative general formulae:

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

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

B. Second formula derived from probability rule:
$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B \mid A)$

IV. The Log–Rank Test for Two Groups (pages 58–62)

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 ith group at time j, where $i = 1$ or 2:
observed counts $= m_{ij}$,
expected counts $= e_{ij}$, where
expected counts = (proportion in risk set) × (# failures over both groups),

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

D. Log–rank statistic for two groups:

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

where $i = 1, 2$,

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

$$\widehat{\text{Var}}\left(O_i - E_i\right) = \sum_j \frac{n_{1j}n_{2j}\left(m_{1j} + m_{2j}\right)\left(n_{1j} + n_{2j} - m_{1j} - m_{2j}\right)}{\left(n_{1j} + n_{2j}\right)^2\left(n_{1j} + n_{2j} - 1\right)}, \ i = 1, 2$$

E. H_0: no difference between survival curves.

F. Log–rank statistic ~χ^2 with 1 *df* under H_0.

G. Approximate formula:

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

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

V. The Log–Rank Test for Several Groups (pages 62–64)

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

B. Use computer for calculations.

C. Under H_0, log–rank statistic ~χ^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. The Peto Test (pages 65–66)

A. Peto test weights observed minus expected score at time j by number at risk, n_j, whereas log–rank test uses equal weights.

B. As with log–rank statistic, Peto statistic ~χ^2 with $G - 1$ *df*, where G = # of groups.

C. Use computer for calculations.

D. Peto test emphasizes beginning of survival curve; early failures receive larger weights; in contrast, log–rank test emphasizes tail of survival curve.

E. In practice, Peto and log–rank test give similar, but not necessarily equal, results.

VII. Summary (page 67)

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				
$t_{(j)}$	n_j	m_j	q_j	$S(t_{(j)})$	$t_{(j)}$	n_j	m_j	q_j	$S(t_{(j)})$
0.0	25	0	0	1.00	0.0	25	1	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

(Continued on next page)

Group 1					Group 2				
$t_{(j)}$	n_j	m_j	q_j	$S(t_{(j)})$	$t_{(j)}$	n_j	m_j	q_j	$S(t_{(j)})$
					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:

$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}	e_{1j}	e_{2j}	$m_{1j}-e_{1j}$	$m_{2j}-e_{2j}$
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

(Continued on next page)

$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}	e_{1j}	e_{2j}	$m_{1j}-e_{1j}$	$m_{2j}-e_{2j}$
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

(Continued on next page)