Lecture Three

Dr. Mutua Kilai

Department of Pure and Applied Sciences

Jan-April 2024

Innovative technology for a dynamic world

Kirinyaga University

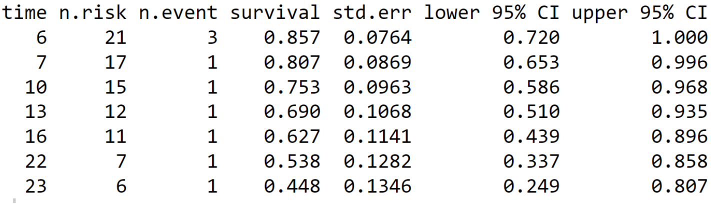
Kaplan Meier Example 2

The data: remission times (weeks) for two groups of leukemia patients.
Group 1 ($n = 21$) the treatment group and Group II ($n = 21$) the control group. Construct and plot the KM curves.

Group 1										
6	6	6	7	10	13	16	22	23	6+	9+
10+	11+	17+	19+	20+	25+	32+	32+	34+	35+	
Group II										
1	1	2	2	3	4	4	5	5	8	8
8	8	11	11	12	12	15	17	22	23	

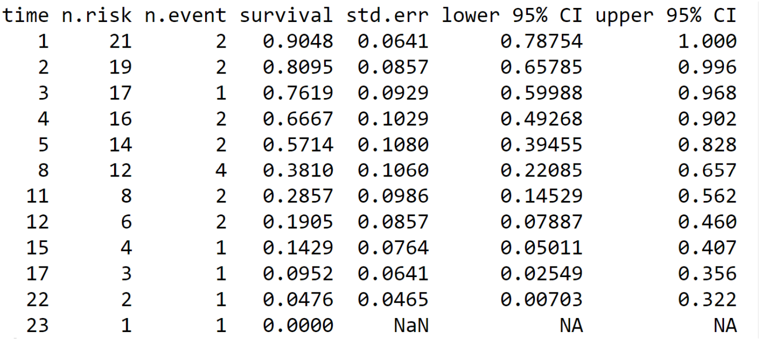
Solution

For group 1



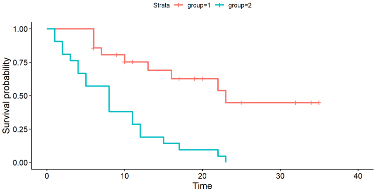
```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   6     21       3    0.857  0.0764        0.720        1.000
   7     17       1    0.807  0.0869        0.653        0.996
  10     15       1    0.753  0.0963        0.586        0.968
  13     12       1    0.690  0.1068        0.510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

Solution Cont'd



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

Survival Curves



- 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.

R Codes

```
# reading in the data
times <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,3
           1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23
status <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,
            1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
group <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
# data frame
data <- data.frame(times, status,group)
# fitting the model
fit <- survfit(Surv(times, status) ~ group, data = data)
# plotting
ggsurvplot(fit, data = data)
```

Likelihood Function

- Suppose that we have n units with lifetime governed by a survivor function $S(t, \theta)$ with associated density $f(t, \theta)$ and hazard function $h(t, \theta)$.
- Suppose unit i is observed for a time t_i. If the unit died at t_i its contribution to the likelihood function is the density at that duration.
- If the unit is still alive at t_i all we know is that the lifetime exceeds t_i. The probability of the event is $L_i = S(t_i)$ which becomes the contribution of a censored observation to the likelihood

Cont'd

- For the vector of unknown parameters $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ the likelihood function is:

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

  $\delta_i$is 1 if the failure of item i is observed and 0 if the failure of item i is right censored.

- We can express the likelihood function as:

  $$L(x, \theta) = \prod_{i=1}^{n} = \prod_{i \in U} f(x_i, \theta) \prod_{i \in C} S(x_i, \theta)$$

  where $S(x_i, \theta)$ is the survivor function of the population distribution with parameters θ evaluated at censoring time $x_i, i \in C$

Cont'd

- Taking the logarithm we have:

$$\ln L(x,\theta) = \sum_{i\in U} \ln f(x_i,\theta) + \sum_{i\in C} \ln S(x_i,\theta)$$

- Since the probability density function is the product of the hazard function and the survivor function, the log likelihood function can be simplified to

$$\ln L(x,\theta) = \sum_{i\in U} \ln h(x_i,\theta) + \sum_{i\in U} \ln S(x_i,\theta) + \sum_{i\in C} \ln S(x_i,\theta)$$

- Which can be expressed as:

$$\ln L(x,\theta) = \sum_{i\in U} \ln h(x_i,\theta) + \sum_{i=1}^{n} \ln S(x_i,\theta)$$

Log-Rank Test

- Are KM curves statistically equivalent?
- We now describe how to evaluate whether or not KM curves for two or more groups are statistically equivalent.
- 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.
- 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.

Log-Rank Test Cont'd

- 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.

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

Expected Cell Counts

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

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

- Proportion in risk set.
- Number of failures over both groups. $(m_{1f} + m_{2f})$

Assumptions of Log-Rank Test

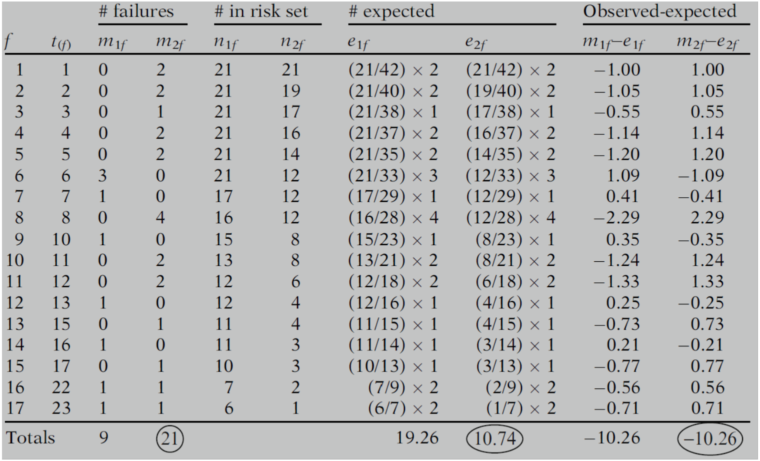
- Independence: The survival times or event times of individuals in each group should be independent to each other.
- Non-Informative Censoring: Censoring should not be related to the event being studied or to the group assignment. The log-rank test assumes that the probability of censoring should be the same for all individuals within each group.
- Proportional Hazards: The hazard rates (the risk of an event occurring) for the compared groups should be consistent over time. The ratio of the hazard rates should remain constant.

Example

Using the data below, compare the two groups using log-rank test at 5% level of significance.

Group 1										
6	6	6	7	10	13	16	22	23	6+	9+
10+	11+	17+	19+	20+	25+	32+	32+	34+	35+	
Group II										
1	1	2	2	3	4	4	5	5	8	8
8	8	11	11	12	12	15	17	22	23	

Solution



		# 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

Solution Cont'd

- $$\log \ \ Rank \ \ Stats = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}$$

- We can use either the values from group 1 or the values from group 2.

- The variance is given by:

$$Variance(O_2 - E_2) = \sum \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$

Cont'd

- The hypothesis is:

$$H_0 : No\ \ difference\ \ between\ \ survival\ \ curves$$

$$H_a : There\ \ is\ \ difference\ \ between\ \ survival\ \ curves$$

- The log-rank statistic follows χ^2_1 distribution under H_0
- Variance $= 6.2685$
- The log rank statistic

$$\frac{(O_2 - E_2)^2}{Var(O_2 - E_2)} = \frac{(10.26)^2}{6.2685} = 16.793$$

- The tabulated value is 3.84 less than 16.793 hence we reject H_0

Cont'd

- An approximate formula is given by:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

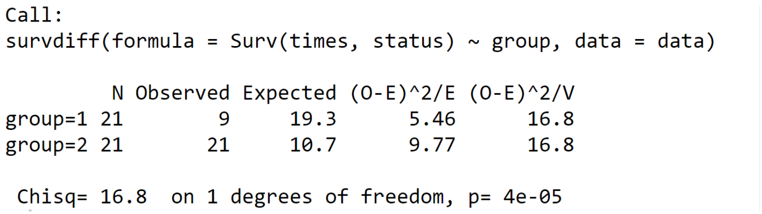
- We can compute for the above example as:

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

R Code

```r
# reading in the data
times <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,3
           1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23
status <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,
            1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
group <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
# data frame
data <- data.frame(times, status,group)
# fitting the model
fit <- survdiff(Surv(times, status) ~ group, data = data)
fit
```

Output



```
Call:
survdiff(formula = Surv(times, status) ~ group, data = data)

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

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

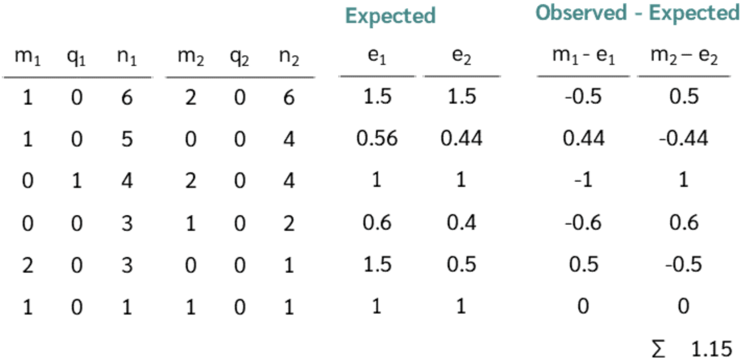
- The p-value is less than 0.05 hence we reject H_0 and conclude that the survival time are different.

Example 2

Suppose we have Group 1 and Group 2 and we want to test whether the two groups have the same survival function or not.
The data is as given:

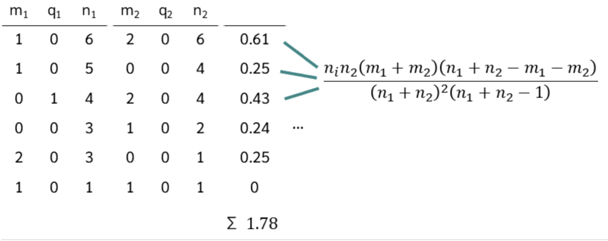
Group 1					
2	3	5+	7	7	8
Group II					
2	2	4	4	6	8

Solution



						Expected		Observed - Expected	
m_1	q_1	n_1	m_2	q_2	n_2	e_1	e_2	$m_1 - e_1$	$m_2 – e_2$
1	0	6	2	0	6	1.5	1.5	-0.5	0.5
1	0	5	0	0	4	0.56	0.44	0.44	-0.44
0	1	4	2	0	4	1	1	-1	1
0	0	3	1	0	2	0.6	0.4	-0.6	0.6
2	0	3	0	0	1	1.5	0.5	0.5	-0.5
1	0	1	1	0	1	1	1	0	0
									Σ 1.15

Solution Cont'd



m_1	q_1	n_1	m_2	q_2	n_2	
1	0	6	2	0	6	0.61
1	0	5	0	0	4	0.25
0	1	4	2	0	4	0.43
0	0	3	1	0	2	0.24
2	0	3	0	0	1	0.25
1	0	1	1	0	1	0
						Σ 1.78

$$\frac{n_i n_2(m_1+m_2)(n_1+n_2-m_1-m_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

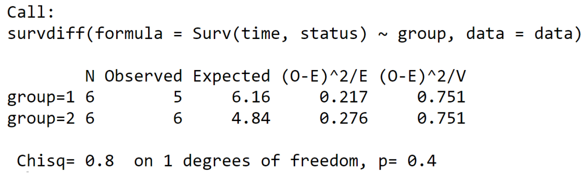
...

- Log rank $= \frac{1.32}{1.78} = 0.74$
- The calculated value is less than tabulated value hence we reject the null hypothesis.

R Code

```r
# reading the data
time <- c(2,3,5,7,7,8,2,2,4,4,6,8)

status <- c(1,1,0,1,1,1,1,1,1,1,1,1)

group <- c(1,1,1,1,1,1,2,2,2,2,2,2)

# data frame
library(survival)

data <- data.frame(time, status,group)

fit2 <- survdiff(Surv(time, status) ~ group, data = data)
fit2
```

Output



```
Call:
survdiff(formula = Surv(time, status) ~ group, data = data)

        N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 6        5     6.16     0.217     0.751
group=2 6        6     4.84     0.276     0.751

 Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

Exercise

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 24 participants. Group 1 has no history of chronic disease (CHR¼ 0), and group 2 has a positive history of chronic disease (CHR ¼ 1):

Data

Group 1											
12.3+	5.4	8.2	12.2+	11.7	10.0	5.7	9.8	2.6	11.0	9.2	12.1
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											
5.8	2.9	8.4	8.3	9.1	4.2	4.1	1.8	3.1	11.4	2.4	1.4
1.6	2.8	4.9	3.5	6.5	9.9	3.6	5.2	8.8	7.8	4.7	3.9

- Compute the K-M estimate for the two groups and plot
- Compute the log-rank test statistic
- Write an R code that does the above two.

Thank You!