Survival Analysis: Kaplan-Meier Method

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we will learn …

- Goals of survival analysis
- Data layout
- Kaplan-Meier method

Goals of survival analysis

- To estimate and interpret survivor/hazard functions from survival data
- To compare survivor/hazard functions
- To assess the relationship of predictors to survival time
- Develop prognostic models

Basic data layout

Person	Time	Status	Factor 1	Factor 2	… Factor p
1	t_1	d_1	X_{11}	X_{12}	X_{1p}
2	t_2	d_2	X_{21}	X_{22}	X_{2p}
…	…	…	…	…	…
…	…	…	…	…	…
5	$t_5 = 3$	$d_5=1$	X_{51}	X_{52}	X_{5p}
…	…	…	…	…	…
n	t_n	d_n	X_{n1}	X_{n2}	X_{np}

An example

Group 1		Group 2	
t (weeks)	log WBC	t (weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

Actual data from a study

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

+ denotes:

in remission at the study end

lost of follow-up

withdraws

Data arrangement for analysis

Group 1, sorted by time: 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

Person	Time	Status (1=event, 0=censored)	Group
1	6	1	1
2	6	1	1
3	6	1	1
4	6	0	1
5	7	1	1
6	9	0	1
…			
20	34	0	1
21	35	0	1

Data arrangement for analysis

Group 2 (coded 0), sorted by time: 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Person	Time	Status (1=event, 0=censored)	Group
22	1	1	0
23	1	1	0
24	2	1	0
…			
41	22	1	0
42	23	1	0

Kaplan-Meier analysis (preparation)

Group 1, sorted by time: 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

Time (t_j)	m_j	q_j	$R(t_j)$
0	0	0	21 patients survive $\geq$ 0 weeks
6	3	1	21 patients survive $\geq$ 6 weeks
7	1	1	17 patients survive $\geq$ 7 weeks
10	1	2	15 patients survive $\geq$ 10 weeks
13	1	0	12 patients survive $\geq$ 13 weeks
16	1	3	11 patients survive $\geq$ 16 weeks
22	1	0	7 patients survive $\geq$ 22 weeks
23	1	5	6 patients survive $\geq$ 23 weeks

t_j = survival time; m_j = number of persons who failed (events); q_j = number of censored persons

Tied observation

- Note that there were 3 ties at 6 weeks
- When there are no ties, $m_j = 1$

Kaplan-Meier analysis (preparation)

Group 2, sorted by time: 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Time (t_j)	m_j	q_j	$R(t_j)$
0	0	0	21 patients survive $\geq$ 0 week
1	2	0	21 patients survive $\geq$ 1 week
2	2	0	19 patients survive $\geq$ 2 weeks
3	1	0	17 patients survive $\geq$ 3 weeks
4	2	0	16 patients survive $\geq$ 4 weeks
5	2	0	14 patients survive $\geq$ 5 weeks
8	4	0	12 patients survive $\geq$ 8 weeks
11	2	0	8 patients survive $\geq$ 11 weeks
12	2	0	6 patients survive $\geq$ 12 weeks
15	1	0	4 patients survive $\geq$ 15 weeks
17	1	0	3 patients survive $\geq$ 17 weeks
22	1	0	2 patients survive $\geq$ 22 weeks
23	1	0	1 patient survive $\geq$ 23 weeks

Descriptive measures of survival experience

Group 1	Group 2
6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 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
Sum of times (ignoring +'s): 359 weeks	Total follow-up: 182 weeks
T_1 (mean follow-up) = 359/21 = 17.1 weeks	T_2 = 8.6 weeks
h_1 = 9 / 359 = 0.025	h_1 = 21 / 182 = 0.115

Kaplan Meier Analysis

Life table (group 1)

Time (t_j)	n_j	m_j	q_j
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	-	-	-

Life table (group 2)

Time (t_j)	n_j	m_j	q_j
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

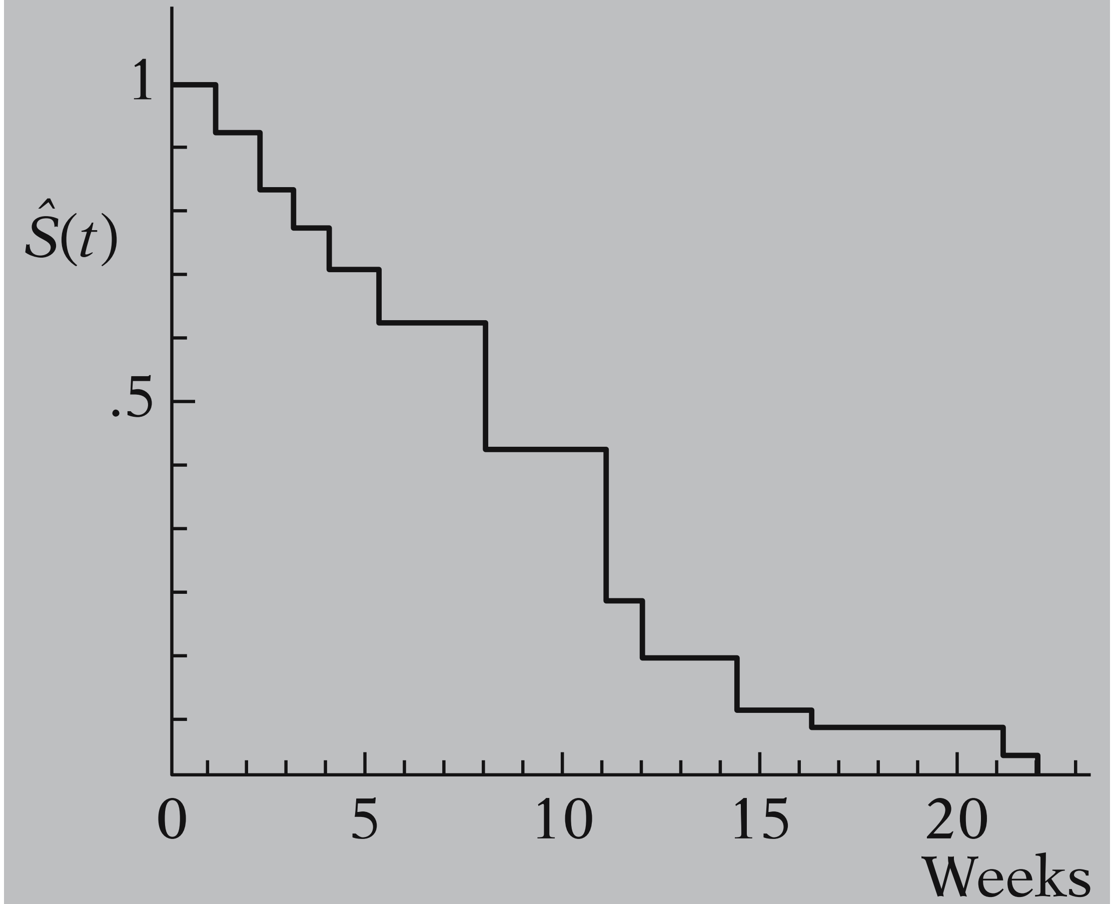
Kaplan Meier (KM) survival probabilities

Time (t_j)	n_j	m_j	q_j	S_{tj}
0	21	0	0	1
1	21	2	0	19 / 21 = 0.90
2	19	2	0	17 / 21 = 0.81
3	17	1	0	16 / 21 = 0.76
4	16	2	0	14 / 21 = 0.67
5	14	2	0	12 / 21 = 0.57
8	12	4	0	8 / 21 = 0.38
11	8	2	0	6 / 21 = 0.29
12	6	2	0	4 / 21 = 0.19
15	4	1	0	3 / 21 = 0.14
17	3	1	0	2 / 21 = 0.10
22	2	1	0	1 / 21 = 0.05
23	1	1	0	0 / 21 = 0.00

Kaplan Meier (KM) survival probabilities

Time (t_j)	S_{tj}
0	1
1	19 / 21 = 0.90
2	17 / 21 = 0.81
3	16 / 21 = 0.76
4	14 / 21 = 0.67
5	12 / 21 = 0.57
8	8 / 21 = 0.38
11	6 / 21 = 0.29
12	4 / 21 = 0.19
15	3 / 21 = 0.14
17	2 / 21 = 0.10
22	1 / 21 = 0.05
23	0 / 21 = 0.00

KM Curve for Group 2 (Placebo)

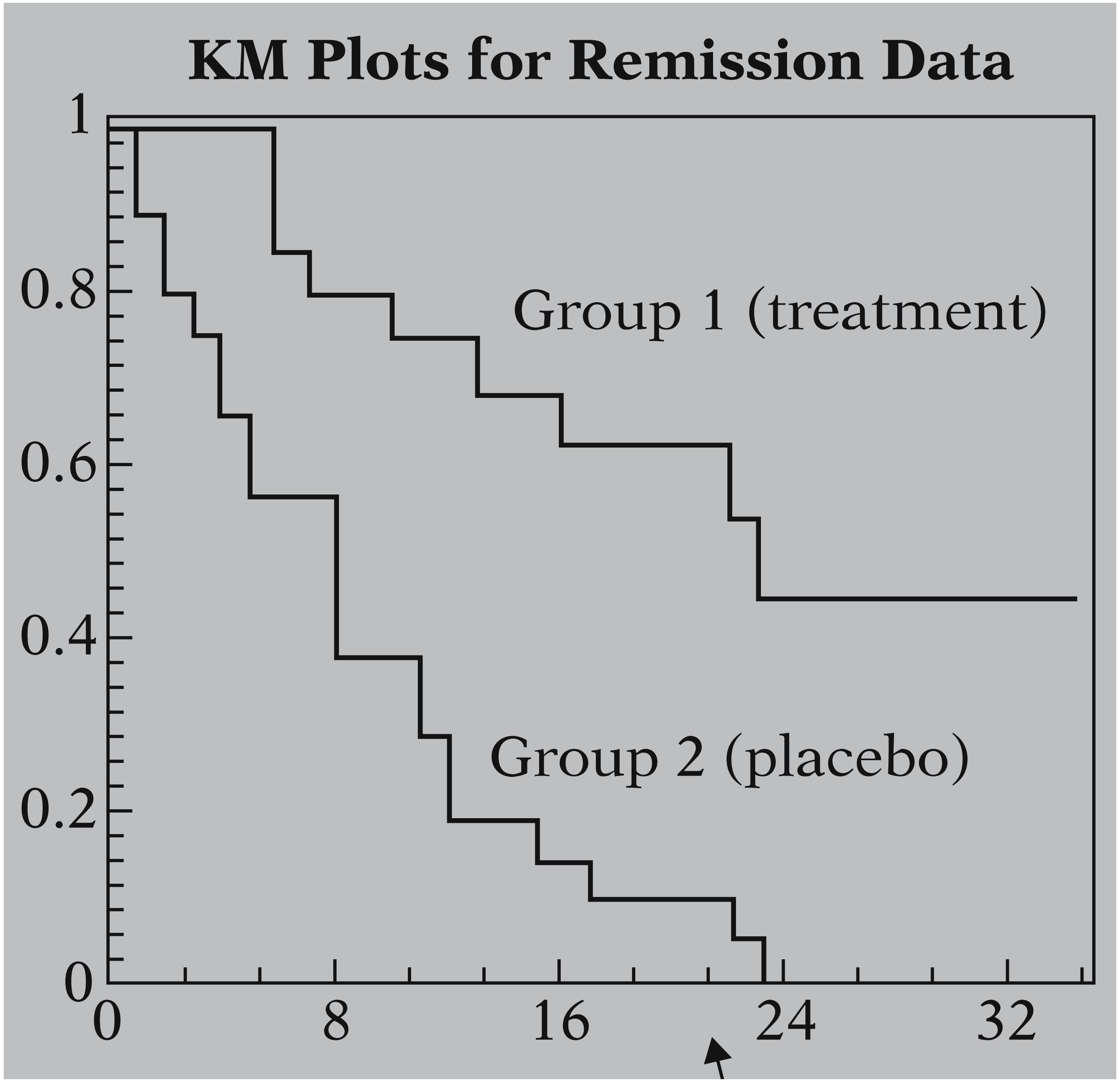


$$S(t) = \Pr(T > t)$$

KM survival probabilities (group 1)

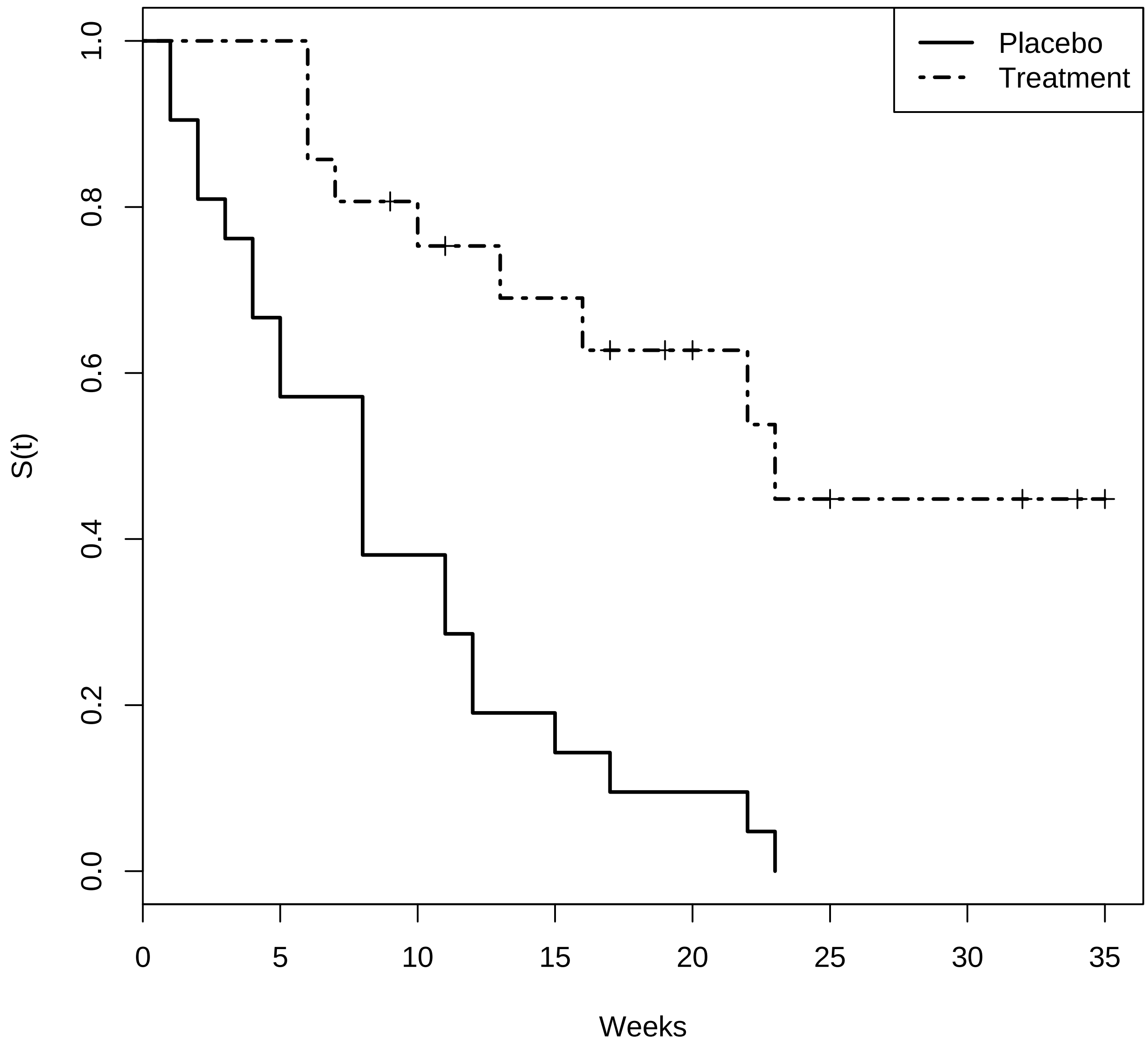
Time (t_j)	n_j	m_j	q_j	P(survive)	S_{tj}
0	21	0	0	1	1
6	21	3	1	18/21 = 0.8571	1 x 0.8571 = **0.8571**
7	17	1	1	16/17 = 0.9411	0.8571 x 0.9411 = **0.8067**
10	15	1	2	14/15 = 0.9333	0.8067 x 0.9333 = **0.7529**
13	12	1	0	11/12 = 0.9167	0.7529 x 0.9167 = **0.6902**
16	11	1	3	10/11 = 0.9091	0.6902 x 0.9091 = **0.5378**
22	7	1	0	6/7 = 0.8571	0.5378 x 0.8571 **= 0.5378**
23	6	1	5	5/6 = 0.8333	0.5378 x 0.8333 = **0.4482**

Comparison of KM survival curves



R codes

```
time = c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16, 17, 19, 20, 22,
23, 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)

status = c(0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1,
1, 1, 1, 1, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,0)

group = c(1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1,1,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,0)

dat = data.frame(time, status, group)

library)(survival)

km = survfit(Surv(time, status==0) ~ group); km

plot(km, lty=c(1,4), lwd=2, xlab="Weeks", ylab="S(t)")

legend("topright", c("Placebo", "Treatment"), lty=c(1,4),
lwd=2)
```



1.0
0.8
0.6
0.4
0.2
0.0
S(t)
0
5
10
15
20
25
30
35
Weeks
Placebo
Treatment

Comparison of survival curves

- Are the two survival curves significantly different?
- How to test for the difference?
- Answer: Log-rank statistic

Log-rank test statistic

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

Log-rank test statistic

EXAMPLE

Remission data: $n = 42$

	# failures		# in risk set	
$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}
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

Expected cell counts:

$$e_{1j} = \underbrace{\left(\frac{n_{1j}}{n_{1j}+n_{2j}}\right)}_{\text{Proportion in risk set}} \times \underbrace{(m_{1j}+m_{2j})}_{\text{\# of failures over both groups}}$$

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

Log-rank test statistic

EXAMPLE

Expanded Table (Remission Data)

		# failures		# in risk set		# expected		Observed−expected	
j	$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}	e_{1j}	e_{2j}	$m_{1j}-e_{1j}$	$m_{2j}-e_{2j}$
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

$$O_i - E_j = \sum_{j=1}^{17} (m_{ij} - e_{ij}), \quad i = 1, 2$$

EXAMPLE

$O_1 - E_1 = -10.26$

$O_2 - E_2 = 10.26$

Log-rank statistic

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

- **O = observed, E = expected**
- **Var = variance**

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

Log-rank statistic using R

```
time = c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16, 17, 19, 20, 22,
23, 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)
status = c(0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1,
1, 1, 1, 1, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,0)
group = c(1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1,1,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,0)
dat = data.frame(time, status, group)
library)(survival)
km = survfit(Surv(time, status==0) ~ group)
km
km.diff = survdiff(Surv(time, status==0) ~ group)
km.diff
```

Log-rank statistic using R

```
> km.diff
Call:
survdiff(formula = Surv(time, status == 0) ~ group)

         N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 21       21     10.7      9.77      16.8
group=1 21        9     19.3      5.46      16.8

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

Problem of with log-rank test

- Cannot take into account the effects of covariates
- We need another method: Cox's proportional hazards model