



Kaplan-Meier (KM) Methods

Md. Moyazzem Hossain

Professor,

Department of Statistics and Data Science

Jahangirnagar University

1



Introduction to Kaplan-Meier

- In 1958, Product-Limit (P-L) method was introduced by Kaplan and Meier (K-M).
- As you move from left to right in estimation of the survival curve first assign equal weights to each observation. Do not jump at the censored observations
- Redistribute equally the pre-assigned weight to the censored observations to all observations to the right of each censored observation

Md. Moyazzem Hossain

2

2



Introduction to Kaplan-Meier

- Non-parametric estimate of the survival function.
- Commonly used to describe survivorship of study population/s.
- Commonly used to compare two study populations.
- Intuitive graphical presentation.

Md. Moyazzem Hossain

3

3



KM estimates of survival function

- Let us consider a sample where all of the patients are observed to death so that the survival times are exact and known (i.e. no censoring).
- Let, t_1, t_2, \dots, t_n be the exact survival times of the “n” individuals under study such that $t_1 \leq t_2 \leq \dots \leq t_n$ then the empirical survival function is defined as

$$\hat{S}(t) = \frac{\text{Number of observations} \geq t}{n} ; t \geq 0 \quad \dots \quad (1)$$

- But when censored observations are present, let d_i represent the number of deaths at t_i , then the KM estimate of the survival function at time t is

$$S(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

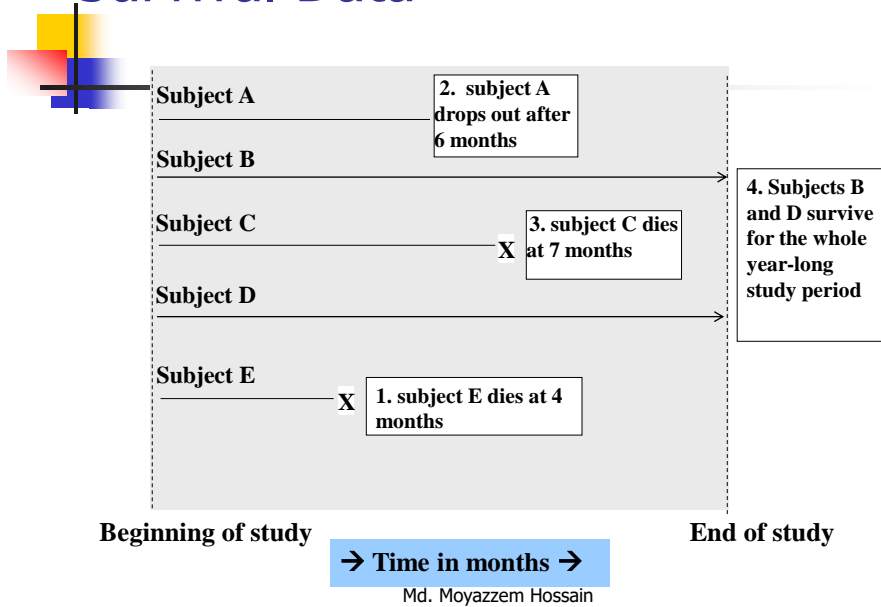
n_i is the number of individuals at risk just before time t_i

Md. Moyazzem Hossain

4

4

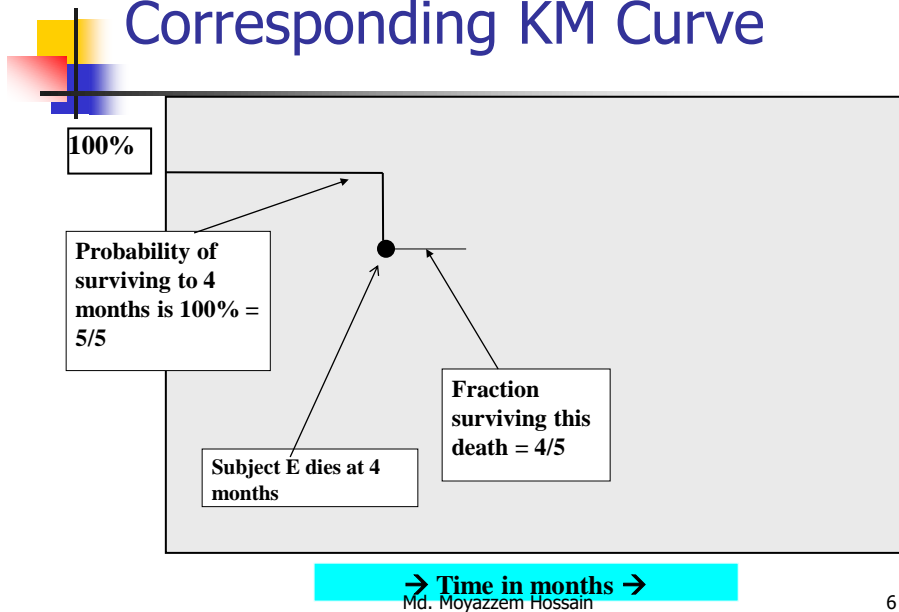
Survival Data



5

5

Corresponding KM Curve

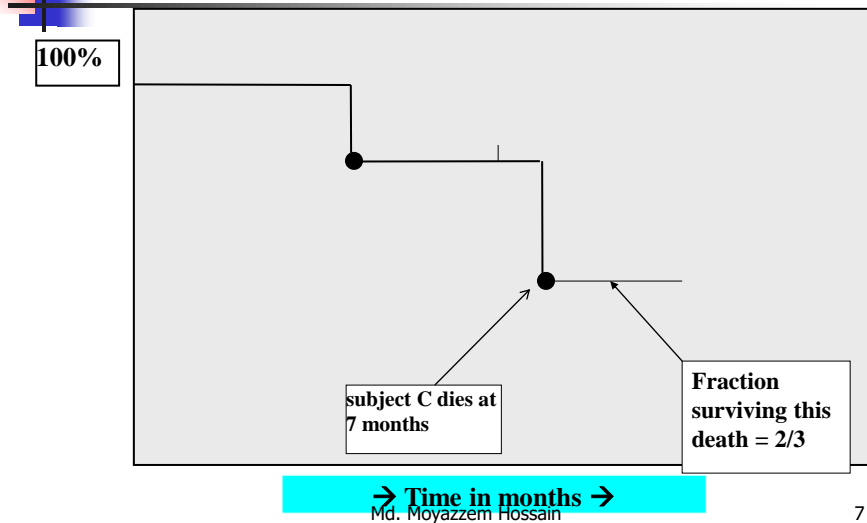


6

6



Corresponding KM Curve

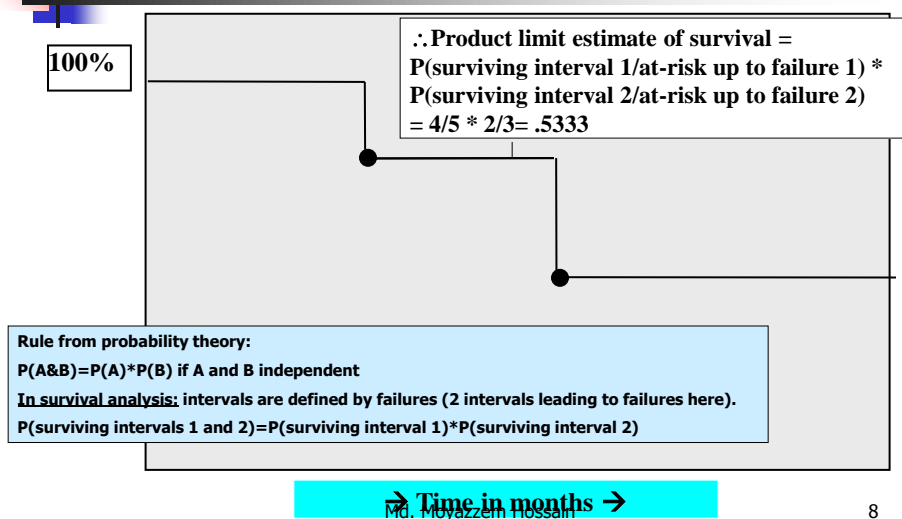


7

7



Corresponding KM Curve



8

8

Example

Weeks to death or censoring (*) in 20 adults with recurrent astrocytoma:

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149*	202	219

ID	weeks	death
1	6	1
2	13	1
3	21	1
4	30	1
5	31	0
6	37	1
7	38	1
8	47	0
9	49	1
10	50	1
11	63	1
12	79	7
13	80	0
14	82	0
15	82	0
16	86	1
17	98	1
18	149	0
19	202	1
20	219	1

Data reproduced from BMJ 2004; 328:1073.

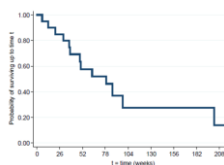
Md. Moyazzem Hossain

9

9

Example First death

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149*	202	219



- 20 individuals in study at $t=0$.
- First death at $t=6$ weeks.
- No individuals censored before $t=6$.
- Probability of death for each individual: $1/20=0.05$
- Therefore probability of surviving beyond $t=6$ is $(1-0.05)=0.95=19/20$.

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
0	20	0	0	1	1
6	20	1	0.05	0.95	$1 \times 0.95 = 0.95$

"Risk set" at time t

$1/20$

$19/20$

Md. Moyazzem Hossain

10

10



Example Second death

	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149*	202	219

- **19** individuals in study between $t=6$ and $t=13$.
- Second death at $t=13$.
- No individuals censored between $t=6$ and $t=13$.
- Probability of death for each individual: $1/19 = 0.053$ (with $19/20$ and $18/19$)
- Therefore probability of surviving beyond $t=13$ is $0.95 \times 0.947 = 0.90$.
 - with $0.95 = (1 - (1/20))$ and $0.947 = (1 - (1/19))$

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
6	20	1	0.05	0.95	0.95
13	19	1	0.053	0.947	$0.95 \times 0.947 = 0.90$

$1/19$

$1 - (1/19) = 18/19$

Md. Moyazzem Hossain

11

11



Example Third and fourth death

		21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149*	202	219

- **18** individuals in study between $t=13$ and $t=21$.
- Probability of death for each individual: $1/18 = 0.056$ (From $t=13$: 0.95×0.947)
- Probability of surviving beyond $t=21$ is $0.90 \times (1 - (1/18)) = 0.85$.
- **17** individuals in study between $t=21$ and $t=30$.
- Probability of death for each individual: $1/17 = 0.059$
- Probability of surviving beyond $t=30$ is $0.85 \times (1 - (1/17)) = 0.80$.

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
13	19	1	$1/19 = 0.053$	0.947	0.90
21	18	1	$1/18 = 0.056$	0.944	0.85
30	17	1	$1/17 = 0.059$	0.941	0.80

Md. Moyazzem Hossain

12

12

Example Fifth and sixth death

				31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149*	202	219

- **16** individuals in study between $t=30$ and $t=31$.
- 1 individual censored at $t=31$.
- **Probability of surviving beyond $t=31$ remains at 0.80.**
- **15** individuals in study between $t=31$ and $t=37$.
- Probability of surviving beyond $t=37$ is $0.80 \times (1 - (1/15)) = 0.747$.

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
30	17	1	0.059	0.941	0.80
31	16	0	0	1	$0.80 \times 1 = 0.80$
37	15	1	$1/15=0.067$	0.933	$0.80 \times 0.933 = 0.747$

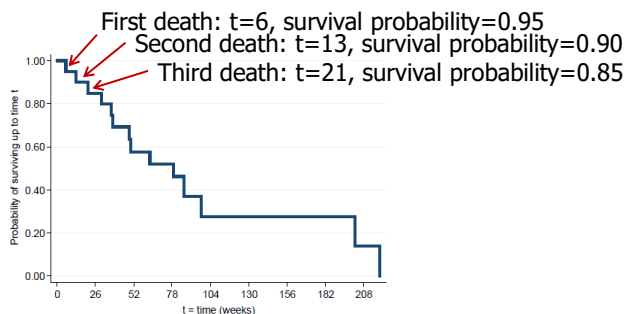
Md. Moyazzem Hossain

13

13

K-M plot of survivor function


- Continue these calculations until reaching the longest event time.
- K-M plot drawn as a step function:



Md. Moyazzem Hossain

14

14




Example

- We consider the “lung” dataset in the `survival` package of R that contains information about survival times and censoring status for patients with advanced lung cancer. In this dataset, the following variables are available:
 - `inst`: Institution code.
 - `time`: This variable represents the survival time or the time until death (measured in days).
 - `status`: This variable indicates the censoring status. A value of 1 represents an observed event (death), and a value of 0 represents censoring (individuals who were still alive at the end of the study).
 - `sex`: The gender of the patient, coded as 1 for male and 2 for female.
 - `age`: The age of the patient at the time of diagnosis.
 - `ph.ecog`: The performance status of the patient, measured on the ECOG scale (Eastern Cooperative Oncology Group). It is a categorical variable representing the overall health and activity level of the patient. Common values include 0 (fully active), 1 (restricted activity but ambulatory), 2 (ambulatory but unable to work), and so on.
 - `ph.karno`: The Karnofsky performance score, another measure of the patient's ability to perform normal daily activities.
 - `pat.karno`: The Karnofsky performance score for the patient's spouse or partner.
 - `meal.cal`: The number of calories consumed during a meal.
 - `wt.loss`: Weight loss in the last six months.

Md. Moyazzem Hossain

15

15



Example

- To perform a Kaplan-Meier survival analysis in R, we use the `survival` package. In order to install and load the `survival` package as well as to see the structure of the “lung” dataset from the `survival` package, the following R-code can be used:

```
# Install and load the survival package
install.packages("survival")
library(survival)
# Explore the structure of the lung dataset
head(lung)
```

Md. Moyazzem Hossain

16

16

Example

```
Console Terminal Background Jobs x
R 4.3.2 . ~/
> library(survival)
> # Explore the structure of the lung dataset
> head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74  1      1      90      100    1175      NA
2    3  455      2  68  1      0      90      90     1225      15
3    3 1010      1  56  1      0      90      90       NA      15
4    5  210      2  57  1      1      90      60     1150      11
5    1  883      2  60  1      0     100      90       NA       0
6   12 1022      1  74  1      1      50      80      513       0
> |
```

Md. Moyazzem Hossain

17

17

Example

```
# Create a Surv object with survival times and
event/censoring indicator
surv_object <- with(lung, Surv(time, status))
# Fit the Kaplan-Meier estimator
fit_km <- survfit(surv_object ~ 1) #The formula ~ 1
indicates a single group
summary(fit_km)
# Plot the Kaplan-Meier curve
plot(fit_km, main = "Kaplan-Meier Survival Curve",
     xlab = "Time", ylab = "Survival Probability",
     col="blue")
```

Md. Moyazzem Hossain

18

18

Example

```

R 4.3.2 ~ /
> library(survival)
> # Explore the structure of the lung dataset
> head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1     3   306     2   74  1       1       90      100    1175      NA
2     3   455     2   68  1       0       90      90     1225      15
3     3  1010     1   56  1       0       90      90      NA      15
4     5   210     2   57  1       1       90      60     1150      11
5     1   883     2   60  1       0      100      90      NA       0
6    12  1022     1   74  1       1       50      80      513       0
> # Create a surv object with survival times and event/censoring indicator
> surv_object <- with(lung, surv(time, status))
> # Fit the Kaplan-Meier estimator
> fit_km <- survfit(surv_object ~ 1) #The formula ~ 1 indicates a single group
> summary(fit_km)
call: survfit(formula = surv_object ~ 1)

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
5     228      1      0.9956 0.00438  0.9871  1.000
11    227      3      0.9825 0.00869  0.9656  1.000
12    224      1      0.9781 0.00970  0.9592  0.997
13    223      2      0.9693 0.01142  0.9472  0.992
15    221      1      0.9649 0.01219  0.9413  0.989

735    12      1      0.0979 0.02660  0.0575  0.167
765    10      1      0.0881 0.02568  0.0498  0.156
791     9      1      0.0783 0.02462  0.0423  0.145
814     7      1      0.0671 0.02351  0.0338  0.133
883     4      1      0.0503 0.02285  0.0207  0.123
> # Plot the Kaplan-Meier curve
> plot(fit_km, main = "Kaplan-Meier survival curve", xlab = "time", ylab = "Survival Probability", col="blue")
> |

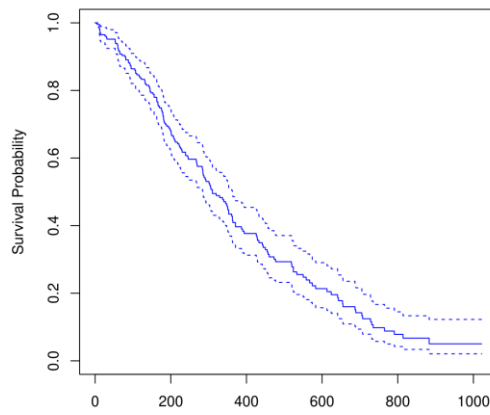
```

Md. Moyazzem Hossain

19

19

Example



Md. Moyazzem Hossain

20

20

Example_Multiple group

- We may also consider multiple groups, and we can compare survival curves. Suppose we consider the “ph.ecog” variable to compare the survival curves for different treatment groups.

```
# Fit Kaplan-Meier estimator for different treatment groups
# Convert 'ph.ecog' to a factor
lung$ph.ecog <- as.factor(lung$ph.ecog)
# Check the levels of 'ph.ecog'
print(levels(lung$ph.ecog))
fit_km_group <- survfit(surv_object ~ ph.ecog, data = lung)
summary(fit_km_group)
# Plot the Kaplan-Meier curves for each group
plot(fit_km_group, main = "Kaplan-Meier Survival Curves by Treatment
Group", col = 1:length(levels(lung$ph.ecog)) , xlab = "Time", ylab =
"Survival Probability")
legend("topright", legend = levels(lung$ph.ecog), col =
1:length(levels(lung$ph.ecog)), bty = "n", lty = 1)
```

Md. Moyazzem Hossain

21

21

Example_Multiple group

```
R 4.3.2 - E:/Book/Chapter 12 /
> # Fit Kaplan-Meier estimator for different treatment groups
> # Convert 'ph.ecog' to a factor
> lung$ph.ecog <- as.factor(lung$ph.ecog)
> # Check the levels of 'ph.ecog'
> print(levels(lung$ph.ecog))
[1] "0" "1" "2" "3"
> fit_km_group <- survfit(surv_object ~ ph.ecog, data = lung)
> summary(fit_km_group)
call: survfit(formula = surv_object ~ ph.ecog, data = lung)

1 observation deleted due to missingness
ph.ecog=0
  time n.risk n.event survival std.err lower 95% CI upper 95% CI
5      63      1  0.9841  0.0157   0.9537      1.000
11     62      1  0.9683  0.0221   0.9259      1.000
15     61      1  0.9524  0.0268   0.9012      1.000
31     60      1  0.9365  0.0307   0.8782      0.999
53     59      1  0.9206  0.0341   0.8562      0.990

524    6      1  0.1392  0.0533   0.0656     0.295
533    5      1  0.1113  0.0494   0.0466     0.266
654    3      1  0.0742  0.0448   0.0227     0.242
707    2      1  0.0371  0.0345   0.0060     0.229
814    1      1  0.0000      NA      NA         NA

      ph.ecog=3
  time n.risk n.event survival std.err lower 95% CI upper 95% CI
118    1      1      0      NA      NA         NA
upper 95% CI NA

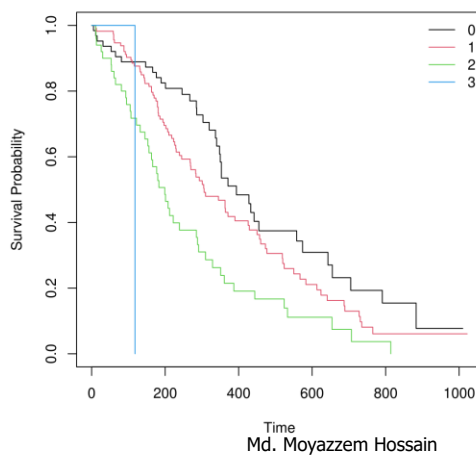
> # Plot the Kaplan-Meier curves for each group
> plot(fit_km_group, main = "Kaplan-Meier Survival Curves by Treatment Group", col = 1:length
h(levels(lung$ph.ecog)) , xlab = "Time", ylab = "Survival Probability")
> legend("topright", legend = levels(lung$ph.ecog), col = 1:length(levels(lung$ph.ecog)), bty
y = "n", lty = 1)
> |
```

Md. Moyazzem Hossain

22

22

Example_Multiple group



23

23

Comparison

Survival chances appear to be different among different groups, however, is the **difference between groups statistically significant?**

Md. Moyazzem Hossain

24

24



Log-rank test

- We use a test that compares survivor functions over the whole follow-up period.
- **Log rank test:** tests the null hypothesis of no difference between samples in the probability of an event (death in this example) at any time point during follow-up.
- **Log rank test statistic:**
 - based on calculating the expected number of events that would occur under the null hypothesis at each event time, and comparing to the observed number of events.
 - under the null hypothesis it follows a Chi-square distribution.

Md. Moyazzem Hossain

25

25



Log-rank test

Null Hypothesis (H0)

The null hypothesis states that there is no difference in the survival experience between the groups being compared. Mathematically, this can be expressed as:

$$H_0: S_A(t) = S_B(t)$$

for all times t , where $S_A(t)$ and $S_B(t)$ are the survival functions of groups A and B, respectively.

Alternative Hypothesis (H1)

The alternative hypothesis states that there is a difference in the survival experience between the groups. Mathematically, this can be expressed as:

$$H_0: S_A(t) \neq S_B(t)$$

Interpretation

- If the p-value obtained from the log-rank test is less than the chosen significance level (typically 0.05), we reject the null hypothesis and conclude that there is a statistically significant difference in the survival distributions between the groups.
- If the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is no statistically significant difference in the survival distributions between the groups.

Md. Moyazzem Hossain

26

26

Log-rank test

The formula for the log-rank test statistic is as follows:

$$\chi^2 = \sum_{i=1}^k \left(\frac{O_i - E_i}{E_i} \right)^2$$

where:

- k is the total number of distinct event times.
- O_i is the observed number of events in group i at time t_i .
- E_i is the expected number of events in group i at time t_i , assuming that the null hypothesis is true.

1.Observed number of events O_i : The actual number of events (e.g., deaths) observed in each group at each time point.

2.Expected number of events E_i : The expected number of events in each group at each time point under the null hypothesis that the survival functions of the groups are the same. This is calculated using the formula:

$$E_i = \frac{R_i \times D_i}{N_i}$$

where:

- R_i is the number of individuals at risk in group i just prior to time t_i .
- D_i is the total number of events at time t_i across all groups.
- N_i is the total number of individuals at risk just prior to time t_i across all groups.

Md. Moyazzem Hossain

27

27

Log-rank test

Let's assume the following data:

Time	Group A Risk	Group A Events	Group B Risk	Group B Events	Total Events
5	3	1	3	1	2
12	2	1	2	1	2
20	1	1	1	1	2

Calculations for each time point:

- At time 5:
 - $R_A = 3, R_B = 3$
 - $O_A = 1, O_B = 1$
 - $D = 2$ (total events)
 - $E_A = \frac{3}{6} \times 2 = 1$
 - $E_B = \frac{3}{6} \times 2 = 1$
 - Contribution to $\chi^2 = \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} = 0$

Md. Moyazzem Hossain

28

28



Log-rank test

4. Calculate the expected number of events for each group:

$$E_{A,12} = \frac{R_{A,12}}{R_{A,12} + R_{B,12}} \times D_{12} = \frac{2}{2+2} \times 2 = \frac{2}{4} \times 2 = 1$$

$$E_{B,12} = \frac{R_{B,12}}{R_{A,12} + R_{B,12}} \times D_{12} = \frac{2}{2+2} \times 2 = \frac{2}{4} \times 2 = 1$$

5. Calculate the contribution to χ^2 at time 12:

$$\chi_{12}^2 = \frac{(O_{A,12} - E_{A,12})^2}{E_{A,12}} + \frac{(O_{B,12} - E_{B,12})^2}{E_{B,12}}$$

Substituting the values:

$$\chi_{12}^2 = \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} = \frac{0^2}{1} + \frac{0^2}{1} = 0 + 0 = 0$$

So, the contribution to χ^2 at time 12 is 0.

Md. Moyazzem Hossain

29

29



Log-rank test

4. Calculate the expected number of events for each group:

$$E_{A,20} = \frac{R_{A,20}}{R_{A,20} + R_{B,20}} \times D_{20} = \frac{1}{1+1} \times 2 = \frac{1}{2} \times 2 = 1$$

$$E_{B,20} = \frac{R_{B,20}}{R_{A,20} + R_{B,20}} \times D_{20} = \frac{1}{1+1} \times 2 = \frac{1}{2} \times 2 = 1$$

5. Calculate the contribution to χ^2 at time 20:

$$\chi_{20}^2 = \frac{(O_{A,20} - E_{A,20})^2}{E_{A,20}} + \frac{(O_{B,20} - E_{B,20})^2}{E_{B,20}}$$

Substituting the values:

$$\chi_{20}^2 = \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} = \frac{0^2}{1} + \frac{0^2}{1} = 0 + 0 = 0$$

Md. Moyazzem Hossain

30

30



Log-rank test

```
# Perform the log-rank test in R
log_rank_test <- survdiff(surv_object ~ ph.ecog, data= lung)
print(log_rank_test)

> print(log_rank_test)
Call:
survdiff(formula = surv_object ~ ph.ecog, data = lung)

n=227, 1 observation deleted due to missingness.

      N Observed Expected (O-E)^2/E (O-E)^2/V
ph.ecog=0  63    37  54.153   5.4331   8.2119
ph.ecog=1 113    82  83.528   0.0279   0.0573
ph.ecog=2  50    44  26.147  12.1893  14.6491
ph.ecog=3   1     1   0.172   3.9733   4.0040

Chisq= 22 on 3 degrees of freedom, p= 7e-05
>
```

Md. Moyazzem Hossain

31

31



Limitations of Kaplan-Meier

- Does not control for covariates
- Requires categorical predictors
- Can not accommodate time-dependent variables

Md. Moyazzem Hossain

32

32



Thank you all.

Md. Moyazzem Hossain

33

33