Syllabus

1. Introduction
   - Survival data
   - Censoring mechanism
   - Application in medical field
2. Concepts and definitions
   - Survival function
   - Hazard function
3. **Non-parametric approach**
   - **Life table**
   - **Kaplan-Meier survival estimate**
   - **Hazard function**
   - **Median and percentile survival time**
4. Hypothesis testing
   - Overview – hypothesis, test statistics, p-values
   - Log-rank
   - Wilcoxon
   - Gehan test
5. Study design and sample size estimation
   - Overview
   - Survival sample size estimation
   - Accrual time and Study duration
6. Semiparametric model – proportional hazard model
   - Partial likelihood
   - Inference
   - Time varying covariates
   - Stratification
7. Model checking in the PH model
   - Model checking
   - Residuals
8. Parametric model
   - Parametric proportional hazard model
   - Accelerate failure model
9. Other topics
   - Competing risk
   - Recurrent events
   - Non-proportional hazard ratio
   - Interval censoring

Homework 1 Issues

- Missing Citations
- Citation styles

References

Edmunson, J.H., Fleming, T.R., Decker, D.G., Malkasian, G.D., Jefferies, J.A., Webb, M.J., and Kvols, L.K., Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma vs. Minimal Residual Disease. Cancer Treatment Reports, 63:241-47, 1979.

Topics

3. Nonparametric Estimation

- Survival function
- Hazard function
- Median and percentile survival time

- Life-table approach
- Kaplan-Meier survival estimate
- Nelson-Aalen/Fleming-Harrington

Reading

- Klein, J.P. and Moeschberger, M.L. "Survial Analysis – Techniques for Censored and Truncated Data", Springer 2003, ISBN #0-387-95399-x.
  - Chapters 4 & 5
- Collett, D. *Modeling Survival Data in Medical Research* , London: Chapman & Hall 1994.
  - Chapter 2
- Cox DR and Oakes D. *Analysis of Survival Data*. London: Chapman & Hall, 1984.
  - Chapter 4
- Kalbfleisch JD and Prentice RL. *The Statistical Analysis of Failure Time Data.* New York: Wiley, 2003
  - Chapter 1
- Lawless, JF. *Statistical Models and Methods for Lifetime Data.* New York: Wiley, 1980.
  - Chapter 3
- Miller
  - Chapter 3

Estimating Survival Functions

- $S(t) = P(T > t)$

- Survival data of n subjects: $T_i, i = 1, 2, \dots, n$
- If no censoring, empirical survival function

$$\hat{S}(t) = \frac{\#\,\{T_i > t\}}{n}$$

  - $\hat{S}(t)$ is the empirical cumulative distribution function

An Example

- Consider the following survival data in years
  - 1,2,3,5,6,9,10,11,12,13,14,17,17, 18,19,21,23,24,24,24
- Without censoring
  - $S(3) = P(T > 3) = \frac{17}{20}$
  - $S(17) = \frac{7}{20}$

- What if there is censoring?

An Example

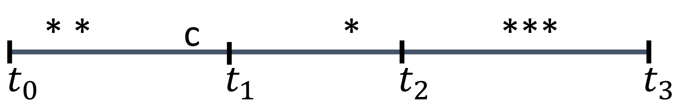
- Consider the following survival data in years
  - 1,2,3,5,6,9,10,11,12,13,14,17,17, 18,19,21,23,24,24,24
- Without censoring
  - $S(3) = P(T > 3) = \frac{17}{20}$
  - $S(17) = \frac{7}{20}$
- What if there is censoring?

The SAS System

Obs	Subjid	Years	Event
1	14	1	1
2	8	2	0
3	2	3	1
4	18	5	1
5	17	6	0
6	19	9	0
7	15	10	0
8	3	11	0
9	13	12	0
10	6	13	0
11	7	14	1
12	10	17	0
13	20	17	1
14	9	18	0
15	4	19	0
16	12	21	0
17	16	23	1
18	1	24	0
19	5	24	0
20	11	24	0

Life-table Estimate

- Also know as the actuarial estimate
  - Used in continuous survival data
  - Grouped data – similar to discrete survival time
- Divide survival data *T* into intervals, for the i^{th} interval
  - $t_{i-1} \leq t < t_i \;\; or \;\; [t_{i-1}, t_i) \;\; i = 1, \ldots, s$
  - The intervals may or may not be of equal length



Life-table Estimate

- Within the i^{th} interval
  - d_i, number of events
  - c_i, number of censors
  - n_i, number of subjects at risk at t_i
  - $n_i' = n_i - c_i/2$, average number of subjects at the interval

- Why $n_i' = n_i - c_i/2$,

Life-table Estimate – Conditional Probability

- For the i^{th} interval
  - Conditional probability of surviving through the i^{th} interval
$$\hat{p}_i = \frac{n_i' - d_i}{n_i'}$$
  - Conditional probability of experiencing an event in the i^{th} interval
$$\hat{q}_i = 1 - \hat{p}_i = \frac{d_i}{n_i'}$$

- Why $n_i' = n_i - c_i/2$,
  - Not $n_i' = n_i$, underestimate the risk $\hat{q}_i$
  - Not $n_i' = n_i - c_i$, overestimate the risk $\hat{q}_i$
  - $n_i' = n_i - c_i/2$, assuming constant censoring rate

Life-table Estimate – Survival Function

- In the i^{th} interval
  - Survival function at the end of the i^{th} interval

$$\hat{S}_L(t_0) = 1$$
$$\hat{S}_L(t_i) = \hat{S}_L(t_{i-1})\left(1 - \frac{d_i}{n_i'}\right)$$

$$\mathrm{var}\{\hat{S}_L(t_{i-1})\} = \hat{S}_L{}^2(t_{i-1}) \sum_{j=1}^{i-1} \frac{d_j}{n_j'(n_j' - d_j)}$$

Life-table Estimate - PDF

- For the i^{th} interval
  - Probability density function at $t_{mi} = \frac{t_i + t_{i-1}}{2}$

$$\hat{f}(t_{mi}) = \frac{\hat{S}_L(t_{i-1}) - \hat{S}_L(t_i)}{t_i - t_{i-1}}$$

$$\hat{S}_L(t_{mi}) = \frac{\hat{S}_L(t_{i-1}) + \hat{S}_L(t_i)}{2}$$

Life-table Estimate – Hazard Function

- Number of events per person-time-units

$$\hat{h}(t_{mi}) = d_i/[(t_i - t_{i-1})\,(n_i' - d_i/2)]$$

- Based on the definition

$$\hat{h}(t_{mi}) = {}^{\hat{f}(t_{mi})}\!/_{\hat{S}(t_{mi})} = {}^{2\hat{f}(t_{mi})}\!/_{[\hat{S}(t_i)+\hat{S}(t_{i-1})]}$$

- Variance

$$\mathrm{var}\{h(t_{mi})\} = \frac{(h(t_{mi}))^2}{n_i' q_i}\left\{1 - \left[\frac{h(t_{mi})(t_i - t_{i-1})}{2}\right]^2\right\}$$

Lifetime Table

Interval	Time Period	Events d_i	Censor c_i	At risk at the beginning of the interval n_i	Average number at risk in the interval n'_i	Survival probability	PDF	Hazard	$se(S(t))$

Example Data

The SAS System

Obs	Subjid	Years	Event
1	14	1	1
2	8	2	0
3	2	3	1
4	18	5	1
5	17	6	0
6	19	9	0
7	15	10	0
8	3	11	0
9	13	12	0
10	6	13	0
11	7	14	1
12	10	17	0
13	20	17	1
14	9	18	0
15	4	19	0
16	12	21	0
17	16	23	1
18	1	24	0
19	5	24	0
20	11	24	0

Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
20	6	14	70.00

Example – SAS Code

```
* Call in data from an excel file to a SAS
dataset;
proc import out=example
     datafile="&lib.\Datasets.xlsx"
     dbms=xlsx replace;
  getnames=yes;
  sheet="Sheet1";
  run;

* Check the range of the time;
proc sort data=example out=ex;
  by years;
  run;

proc print data=ex; run;
```

```
* Check the number of events and censoring;
proc freq data=ex;
  table event;
  run;

* Life tables;
ods graphics on;
proc lifetest data=example method=lt
     intervals=(0 to 25 by 5)
     plots=(s,h,p);
  time years*event(0);
  run;
ods graphics off;
```

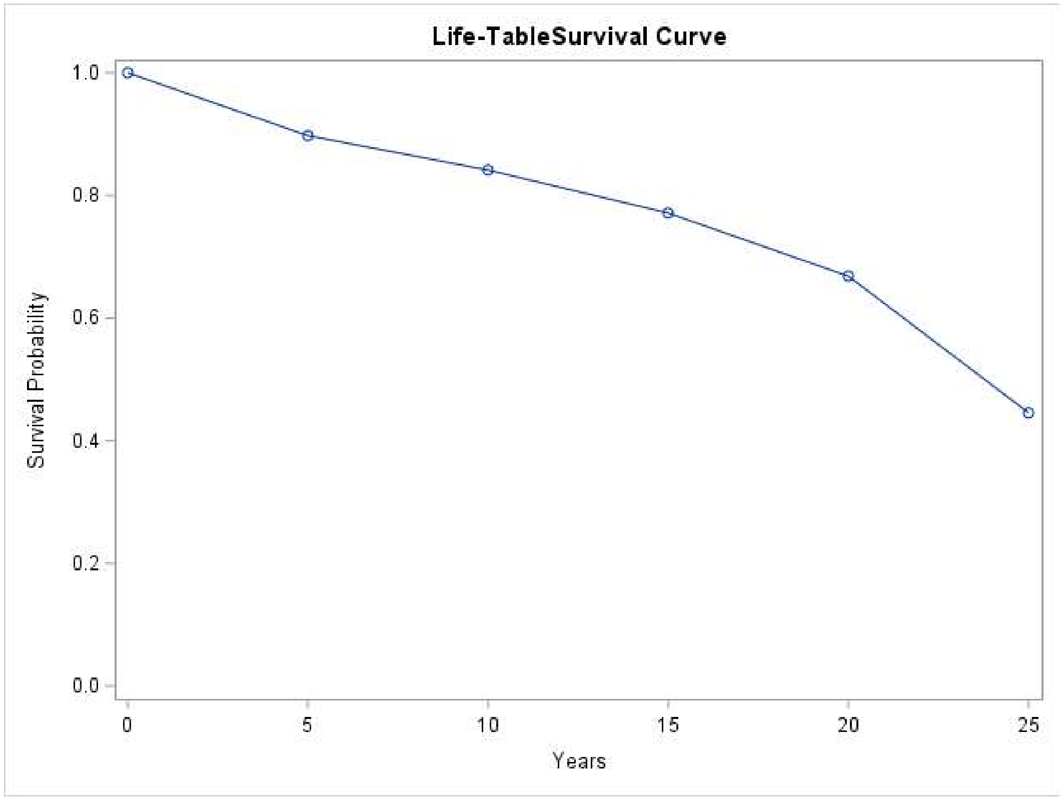
Example Data – Life-Table

The SAS System

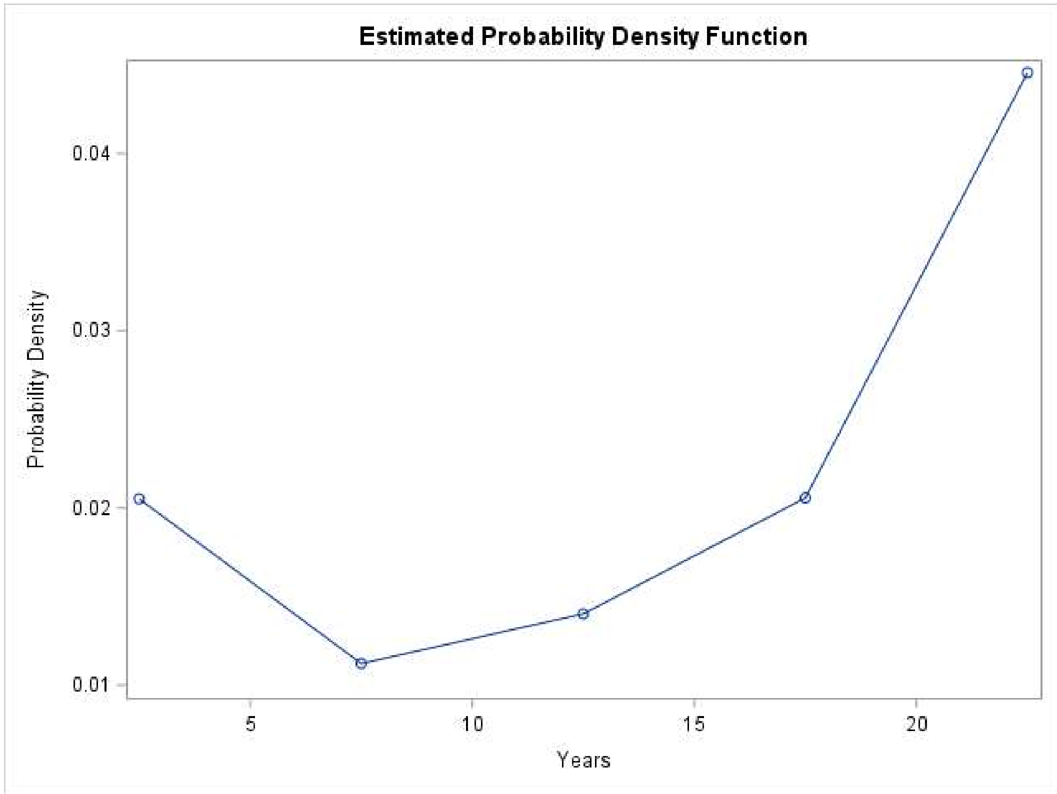
The LIFETEST Procedure

Life Table Survival Estimates															
Interval												Evaluated at the Midpoint of the Interval			
[Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	5	2	1	19.5	0.1026	0.0687	1.0000	0	0	23.7792	2.5410	0.0205	0.0137	0.021622	0.015266
5	10	1	2	16.0	0.0625	0.0605	0.8974	0.1026	0.0687	19.9301	2.5175	0.0112	0.0109	0.012903	0.012897
10	15	1	4	12.0	0.0833	0.0798	0.8413	0.1587	0.0843	.	.	0.0140	0.0135	0.017391	0.017375
15	20	1	3	7.5	0.1333	0.1241	0.7712	0.2288	0.1023	.	.	0.0206	0.0193	0.028571	0.028498
20	25	1	4	3.0	0.3333	0.2722	0.6684	0.3316	0.1305	.	.	0.0446	0.0374	0.08	0.078384
25	.	0	0	0.0	0	0	0.4456	0.5544	0.2016	.	.	.	.	.	.

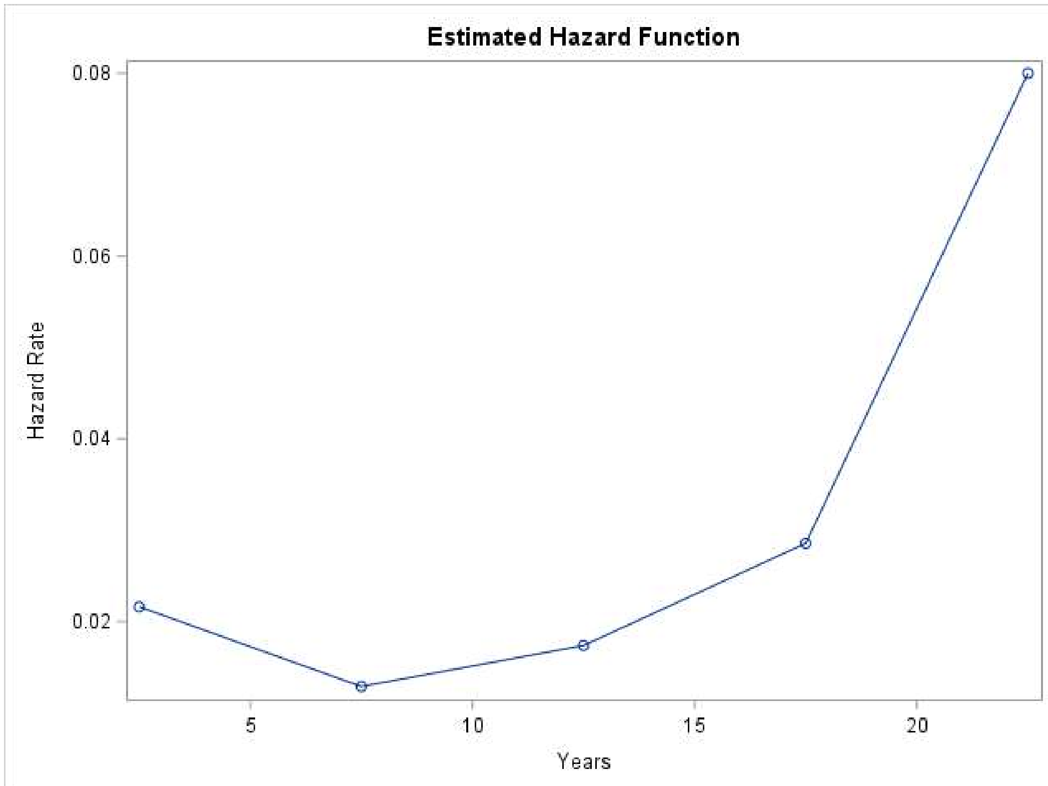
Example Data – Life-Table Survival Curve



Example Data – Estimated Probability Density



Example Data – Estimated Hazard Function



Grouped Survival Data

Obs	Years	Censored	Freq
1	0.5	0	456
2	0.5	1	0
3	1.5	0	226
4	1.5	1	39
5	2.5	0	152
6	2.5	1	22
7	3.5	0	171
8	3.5	1	23
9	4.5	0	135
10	4.5	1	24
11	5.5	0	125
12	5.5	1	107
13	6.5	0	83
14	6.5	1	133
15	7.5	0	74
16	7.5	1	102
17	8.5	0	51
18	8.5	1	68
19	9.5	0	42
20	9.5	1	64
21	10.5	0	43
22	10.5	1	45
23	11.5	0	34
24	11.5	1	53
25	12.5	0	18
26	12.5	1	33
27	13.5	0	9
28	13.5	1	27
29	14.5	0	6
30	14.5	1	23
31	15.5	0	0
32	15.5	1	30

```
proc lifetest data=Males  method=lt intervals=(0 to 15 by 1)
plots=(s,h,p);
  time Years*Censored(1);
  freq Freq;
run;
```

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
2418	1625	793	32.80

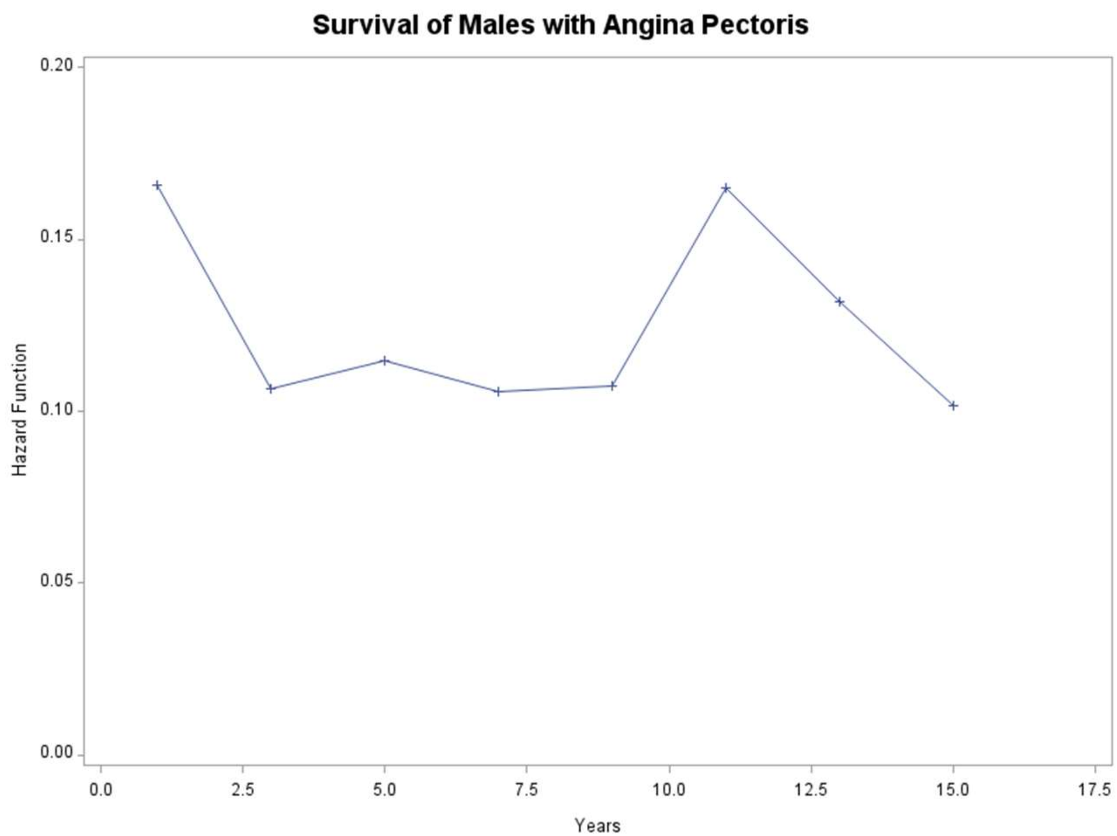
Grouped Survival Data – Life-Table

The LIFETEST Procedure

Life Table Survival Estimates

Interval												Evaluated at the Midpoint of the Interval			
[Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	2	682	39	2398.5	0.2843	0.00921	1.0000	0	0	5.3059	0.1718	0.1422	0.00461	0.165735	0.006259
2	4	323	45	1674.5	0.1929	0.00964	0.7157	0.2843	0.00921	6.3657	0.2433	0.0690	0.00356	0.106742	0.005905
4	6	260	131	1263.5	0.2058	0.0114	0.5776	0.4224	0.0101	6.2431	0.1919	0.0594	0.00345	0.114689	0.007066
6	8	157	235	820.5	0.1913	0.0137	0.4588	0.5412	0.0104	5.6468	0.1892	0.0439	0.00330	0.105795	0.008396
8	10	93	132	480.0	0.1938	0.0180	0.3710	0.6290	0.0105	5.1597	0.3393	0.0359	0.00350	0.107266	0.011059
10	12	77	98	272.0	0.2831	0.0273	0.2991	0.7009	0.0108	4.9856	0.5971	0.0423	0.00436	0.164882	0.018533
12	14	27	60	116.0	0.2328	0.0392	0.2144	0.7856	0.0113	.	.	0.0250	0.00441	0.131707	0.025126
14	16	6	53	32.5	0.1846	0.0681	0.1645	0.8355	0.0121	.	.	0.0152	0.00571	0.101695	0.041302
16	.	0	0	0.0	0	0	0.1341	0.8659	0.0149	.	.	.	.	.	.

Grouped Survival Data – Hazard Function



Colon Cancer - Life-table Example

install.package(“biostat3”)

Library(biostat3)

print(lifetab2(Surv(floor(surv_yy), status == "Dead: cancer")~1, colon_sample, breaks=0:10), digits=2)

```
##         tstart tstop nsubs nlost nrisk nevent surv   pdf hazard se.surv se.pdf se.hazard
## 0-1          0     1    35     1  34.5      7 1.00 0.203   0.23   0.000  0.068     0.085
## 1-2          1     2    27     3  25.5      1 0.80 0.031   0.04   0.068  0.031     0.040
## 2-3          2     3    23     4  21.0      5 0.77 0.182   0.27   0.073  0.073     0.120
## 3-4          3     4    14     1  13.5      2 0.58 0.086   0.16   0.090  0.058     0.113
## 4-5          4     5    11     1  10.5      0 0.50 0.000   0.00   0.095    NaN       NaN
## 5-6          5     6    10     0  10.0      0 0.50 0.000   0.00   0.095    NaN       NaN
## 6-7          6     7    10     3   8.5      0 0.50 0.000   0.00   0.095    NaN       NaN
## 7-8          7     8     7     1   6.5      0 0.50 0.000   0.00   0.095    NaN       NaN
## 8-9          8     9     6     4   4.0      1 0.50 0.124   0.29   0.095  0.110     0.283
## 9-10         9    10     1     1   0.5      0 0.37 0.000   0.00   0.129    NaN       NaN
## 10-Inf      10   Inf     0     0   0.0      0 0.37    NA     NA   0.129     NA        NA
```

Kaplan-Meier Estimator

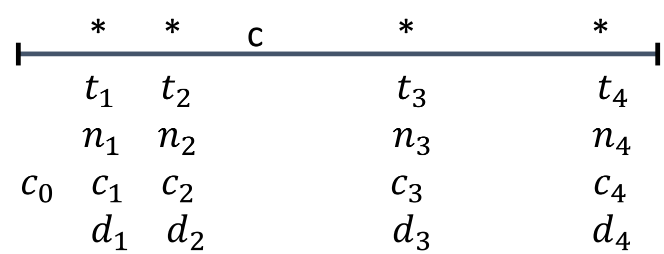
- Also known as product-limit estimator
- Observed
  - (T_i, δ_i) in n subjects, $i = 1,2,\ldots,n$
  - r – number of events
  - $n - r$ – number of censored
  - d distinct event times among r events, $r \geq d$
- Order the d event times: $t_1 < t_2 < \cdots < t_d$
  - Create intervals at distinct event times

Kaplan-Meier Estimator

- Let

$$d_i = \#\ of\ failure\ at\ time\ t_i$$
$$n_i = \#\ at\ risk\ at\ t_i^-$$
$$c_i = \#censored\ during\ the\ interval\ [t_i, t_{i+1})$$

$$i = 1, 2, \ldots, D$$



- Important relationship

$$n_i = n_{i-1} - c_{i-1} - d_{i-1}$$
$$n_i = \sum_{j>i} (c_j + d_j)$$

Kaplan-Meier Estimator

0	t_1	t_2	t_3	t_4
n	n_1	n_2	n_3	n_4
c_0	c_1	c_2	c_3	c_4
0	d_1	d_2	d_3	d_4

$$n_1 = n - c_0$$

$$n_1 = c_1 + d_1 + c_2 + d_2 + c_3 + d_3 + c_4 + d_4$$

$$n_2 = n_1 - c_1 - d_1$$

$$n_2 = c_2 + d_2 + c_3 + d_3 + c_4 + d_4$$

$$n_3 = n_2 - c_2 - d_2$$

$$n_3 = c_3 + d_3 + c_4 + d_4$$

$$n_4 = n_3 - c_3 - d_3$$

$$n_4 = c_4 + d_4$$

Kaplan-Meier Estimator

- At the distinct event time t_i
- The probability of surviving beyond t_i is

$$\hat{p}_i = \frac{n_i - d_i}{n_i}$$

Recall the Discrete Survival Function

$$S(t_j) = P(T > t_j)$$

$$= P(T > t_j | T \geq t_j) \times P(T > t_{j-1} | T \geq t_{j-1}) \times \cdots \times P(T > t_1 | T \geq t_1)$$

$$h_i = P(T = t_i | T \geq t_i) = P(T \geq t_i | T \geq t_i) - P(T > t_i | T \geq t_i)$$

$$P(T > t_i | T \geq t_i) = 1 - h_i$$

$$S(t_j) = \prod_{i:t_i \leq t_j} (1 - h_i)$$

Bayes Theorem

$$
\begin{aligned}
P(A_1 \cap A_2 \cap \cdots \cap A_k) = {} & P(A_k | A_{k-1} \cap \cdots \cap A_1) \times \\
& P(A_{k-1} | A_{k-2} \cap \cdots \cap \ A_1) \times \\
& \ldots \times P(A_2 | A_1) \times P(A_1)
\end{aligned}
$$

Product-Limit Estimator

- Similarly, let $t_{j-1} < t < t_j$

$S_K(t_j) = P(T > t_j)$

$= P(T > t_j \cap T > t_{j-1} \cap \cdots \cap T > t_1)$

$= P(T > t_j | T \geq t_j^-) \times P(T > t_{j-1} | T \geq t_{j-1}^-) \times \cdots \times P(T > t_1 | T \geq t_1^-)$

- The survival function can be estimated as

$$\hat{S}_K(t) = \begin{cases} 1 & if\ t < t_1 \\ \prod_{t_i \leq t} [1 - \frac{d_i}{n_i}] & if\ t \geq t_1 \end{cases}$$

Product-Limit Estimator

- If there is not censoring, $n_i - d_i = n_{i+1}$

$$\hat{S}_L(t) = \frac{n_i}{n_1},$$

where $n_i =$ the number of subjects at risk at $t_i^- = t_{i-1}$.

$$\mathrm{var}\{\hat{S}_L(t)\} = \hat{S}_L(t)(1 - \hat{S}_L(t))/n_1$$

Product-Limit Estimator

- A step function with jumps at event times
- The size of the jump t_i at depends on
  - Number of events at t_i
  - Number of subjects at risk after t_{i-1}
  - Number of subjects censored in the interval $[t_{i-1}, t_i]$

- When there is no censor, KM estimator is empirical estimator

K-M Example

```
ods graphics on;
proc lifetest data=example method=KM
plots=(s,h,p);
  time years*event(0);
  run;
ods graphics off;
```

The SAS System

Obs	Subjid	Years	Event
1	14	1	1
2	8	2	0
3	2	3	1
4	18	5	1
5	17	6	0
6	19	9	0
7	15	10	0
8	3	11	0
9	13	12	0
10	6	13	0
11	7	14	1
12	10	17	0
13	20	17	1
14	9	18	0
15	4	19	0
16	12	21	0
17	16	23	1
18	1	24	0
19	5	24	0
20	11	24	0

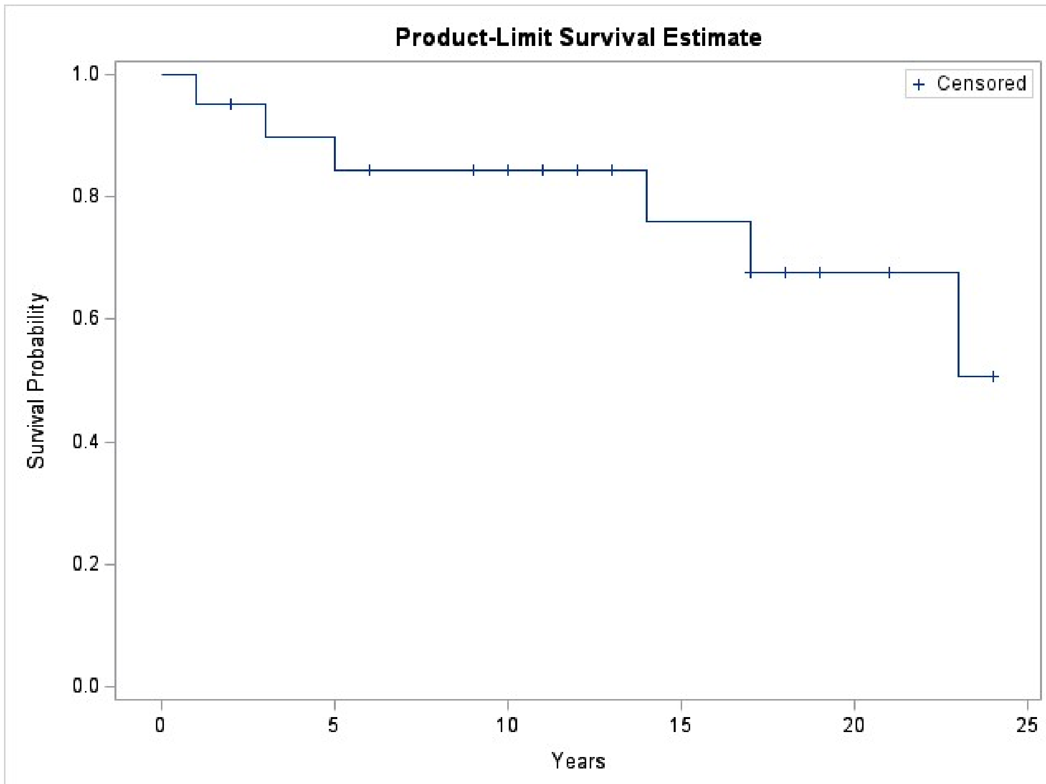
K-M Example

Summary Statistics for Time Variable Years

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	23.0000	.
50	.	LOGLOG	14.0000	.
25	17.0000	LOGLOG	1.0000	.

Product-Limit Survival Estimates						
Years		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	20
1.0000		0.9500	0.0500	0.0487	1	19
2.0000	*	.	.	.	1	18
3.0000		0.8972	0.1028	0.0689	2	17
5.0000		0.8444	0.1556	0.0826	3	16
6.0000	*	.	.	.	3	15
9.0000	*	.	.	.	3	14
10.0000	*	.	.	.	3	13
11.0000	*	.	.	.	3	12
12.0000	*	.	.	.	3	11
13.0000	*	.	.	.	3	10
14.0000		0.7600	0.2400	0.1093	4	9
17.0000		0.6756	0.3244	0.1256	5	8
17.0000	*	.	.	.	5	7
18.0000	*	.	.	.	5	6
19.0000	*	.	.	.	5	5
21.0000	*	.	.	.	5	4
23.0000		0.5067	0.4933	0.1740	6	3
24.0000	*	.	.	.	6	2
24.0000	*	.	.	.	6	1
24.0000	*	.	.	.	6	0

K-M Example

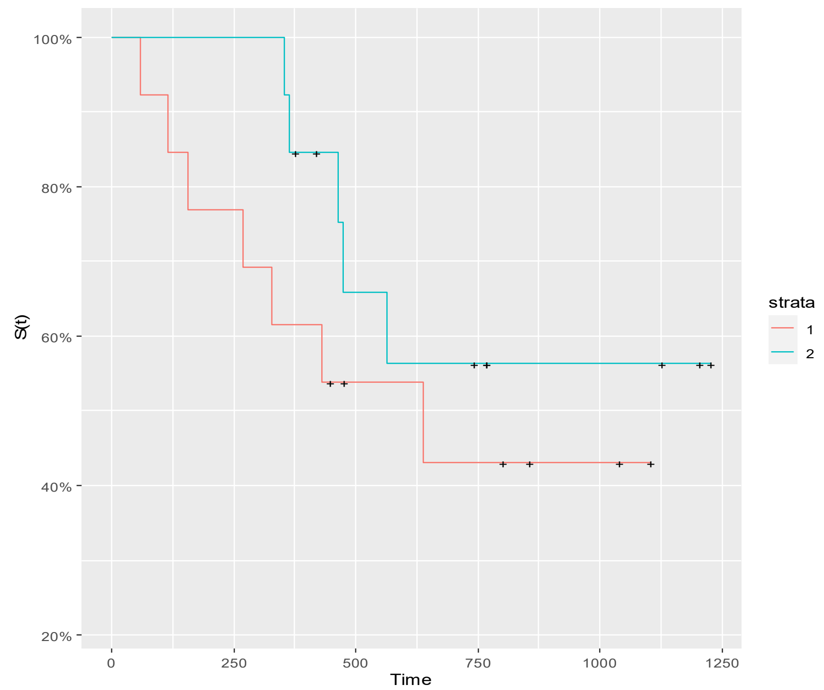

Product-Limit Survival Estimate
+ Censored
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
Years

Ovarian Cancer Example

```
library("survival")
library(ggplot)
ibrary(tidyverse)
library("ggfortify")

ovarian.survfit <-
  ovarian %>%
  survfit(Surv(futime, fustat) ~ rx, data = .)

ovarian.survfit %>%
  autoplot() +
  ylab("S(t)") +
  xlab("Time")
```



Variance – Greenwood Formular

For $t_{(k)} \leq t \leq t_{(k+1)}$

Let $\hat{p}_i = 1 - \frac{d_i}{n_i}$

$$\hat{S}_K(t) = \prod_{i=1}^{k} \hat{p}_i\, , k = 1,2, \dots , d$$

Taking log of $\hat{S}_K(t)$ - what is the reason?

$$log\hat{S}_K(t) = \sum_{i=1}^{k} log\, \hat{p}_i$$

Variance – Greenwood Formular

- $\hat{p}_i$ is asymptotically independence
- What is asymptotically independence?
  - Two sequences X_n, Y_n
  - For finite n, $P(X_n \cap Y_n) \neq P(X_n)P(Y_n)$
  - Loosely speaking,

$$sup\{P(X_n \cap Y_n) - P(X_n)P(Y_n)\} \to 0 \ \ as\ n \to \infty$$

- For rigorous definitions, read
- https://arxiv.org/pdf/1910.04243.pdf

Remarks on asymptotic independence

Youri Davydov[1] and Svyatoslav Novikov [2]
[1] St. Petersburg State University
and Université de Lille, Laboratoire Paul Painlevé
[2] Chebyshev Laboratory, St. Petersburg State University

Greenwood Formular

$$log\hat{S}_K(t) = \sum_{i=1}^{k} log\,\hat{p}_i$$

$$\mathrm{var}\{log\hat{S}_K(t)\} \approx \sum_{i=1}^{k} \mathrm{var}\{\log \hat{p}_i\}$$

$$\mathrm{var}\{\log \hat{p}_i\} = ?$$

Greenwood Formular

- Review of Delta method

- $X \sim N(\mu, \sigma^2)$
- $Y = g(X)$ is a differentiable and $g'(\mu) \neq 0$, then approximately,
- $Y \sim N\left(g(\mu), \left(g'(\mu)\right)^2 \sigma^2\right)$

- $g(\hat{p}_i) = \log \hat{p}_i$
- $g'(\hat{p}_i) = \frac{1}{\hat{p}_i}$

Greenwood Formular

Apply Delta method

$$var[log\hat{p}_i] \approx var(\hat{p}_i)\frac{1}{\hat{p}_i{}^2}$$

As

$$var(\hat{p}_i) = \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$$

We have

$$var(log\hat{p}_i) = \frac{(1-\hat{p}_i)}{n_i\hat{p}_i} = \frac{d_i}{n_i(n_i-d_i)}$$

Greenwood Formular

Therefore,

$$\text{var}\{log\hat{S}_K(t)\} \approx \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}$$

Apply Delta method one more time: $Y = g(X) = e^X,\ g'(X) = e^X = Y$

$$X = log\hat{S}_K(t), \qquad Y = \hat{S}_K(t)$$

$$\text{var}\{\hat{S}_K(t)\} \approx \{\hat{S}_K(t)\}^2 \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}$$

$$\text{se}\{\hat{S}_K(t)\} \approx \hat{S}_K(t) \left\{\sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}\right\}^{1/2}$$

Confidence Interval

- At time point t, $2-\text{sided } 100(1-\alpha)\% \text{ CI}$
- Assuming normal distribution

$$\{\hat{S}(t) - z_{\alpha/2}\mathrm{se}(\hat{S}(t)), \hat{S}(t) + z_{\alpha/2}\mathrm{se}(\hat{S}(t))\}$$

- For example, 2-sided 95% confidence interval for K-M estimator

$$\hat{S}_K(t) \pm z_{1-0.025}se[\hat{S}_K(t)]$$

Confidence Intervals

- Issues: the confidence limits can be <0 or >1
- To avoid such issues, we can do the following
  - $\{\max\{0, \hat{S}(t) - z_{\alpha/2}\text{se}(\hat{S}(t)\}, \min\{1, \hat{S}(t) + z_{\alpha/2}\text{se}\left(\hat{S}(t)\right)\}\}$
  - Transform $\hat{S}(t)$ to $(-\infty, \infty)$
    1. $\log[\hat{S}(t)/(1- \hat{S}(t))]$
    2. $\log\{-\log[\hat{S}(t)]\}$
       why double log?

Log-log Transformation

- Let

$$B(t) = \log\left(-\log(\hat{S}_K(t))\right)$$

- Calculate 95% CI

$$B(t) \pm 1.96\ se[B(t))]$$

- The bounds for $B(t)$

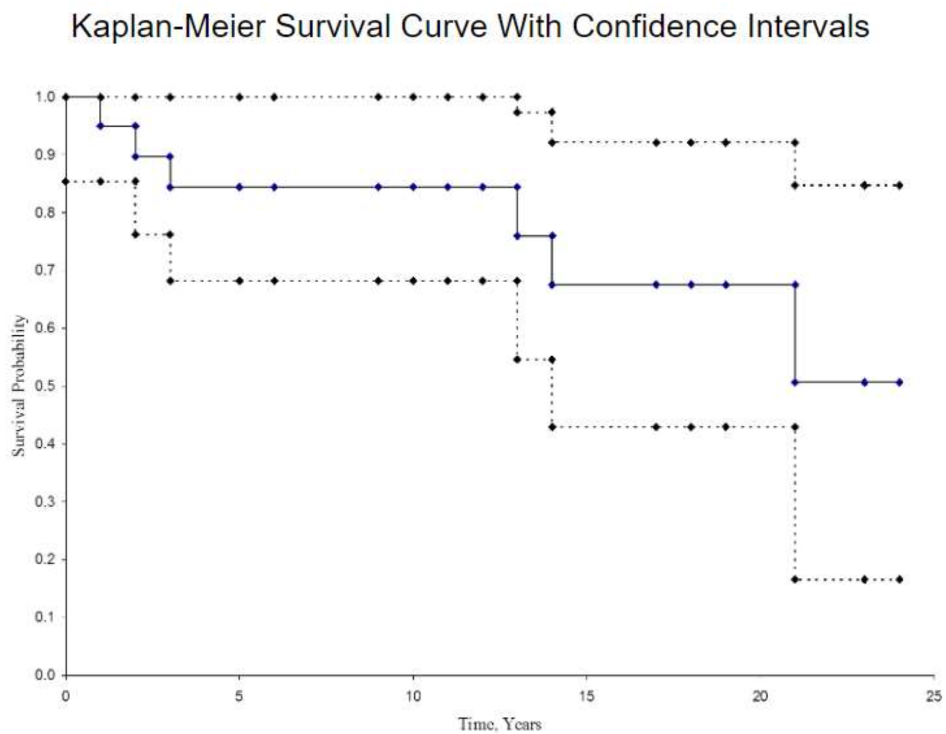
  The lower bound $L = B(t) - 1.96\ se[B(t))]$

  The upper bound $U = B(t) + 1.96\ se[B(t))]$

- Since $\hat{S}_K(t) = \exp(-\exp(B(t))$
- The bounds for $\hat{S}_K(t)$ are

$$[\exp(-\exp(L))\,, \exp(-\exp(U))]$$

Example With Pointwise CI

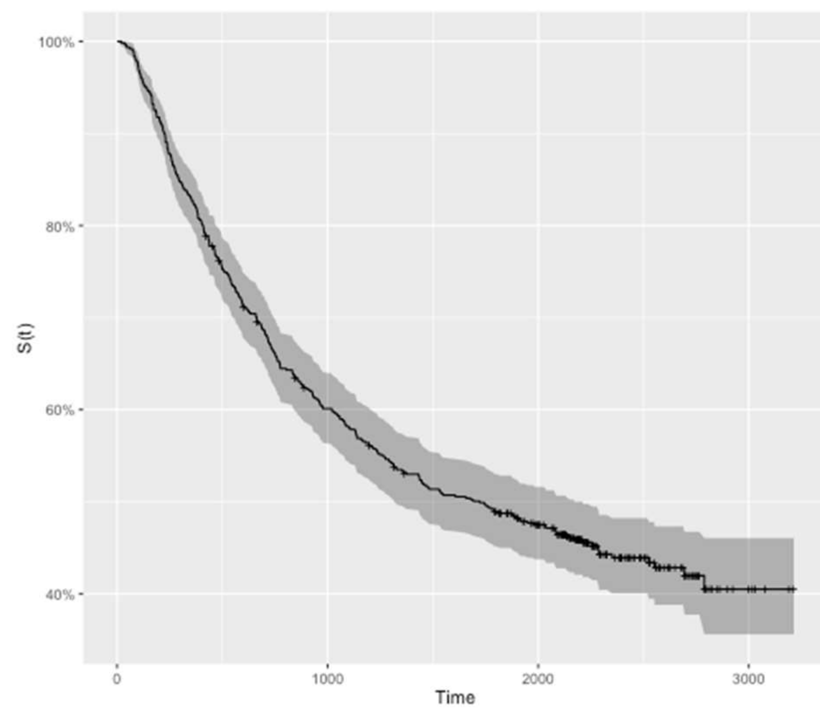


Example – Colon Cancer

```
library("survival")
library("ggfortify")

colon.survfit <-
  colon %>%
  filter(rx == "Obs") %>%
  survfit(Surv(time, status) ~ 1, data = .)

colon.survfit %>%
  autoplot() +
  ylab("S(t)") +
  xlab("Time")
```



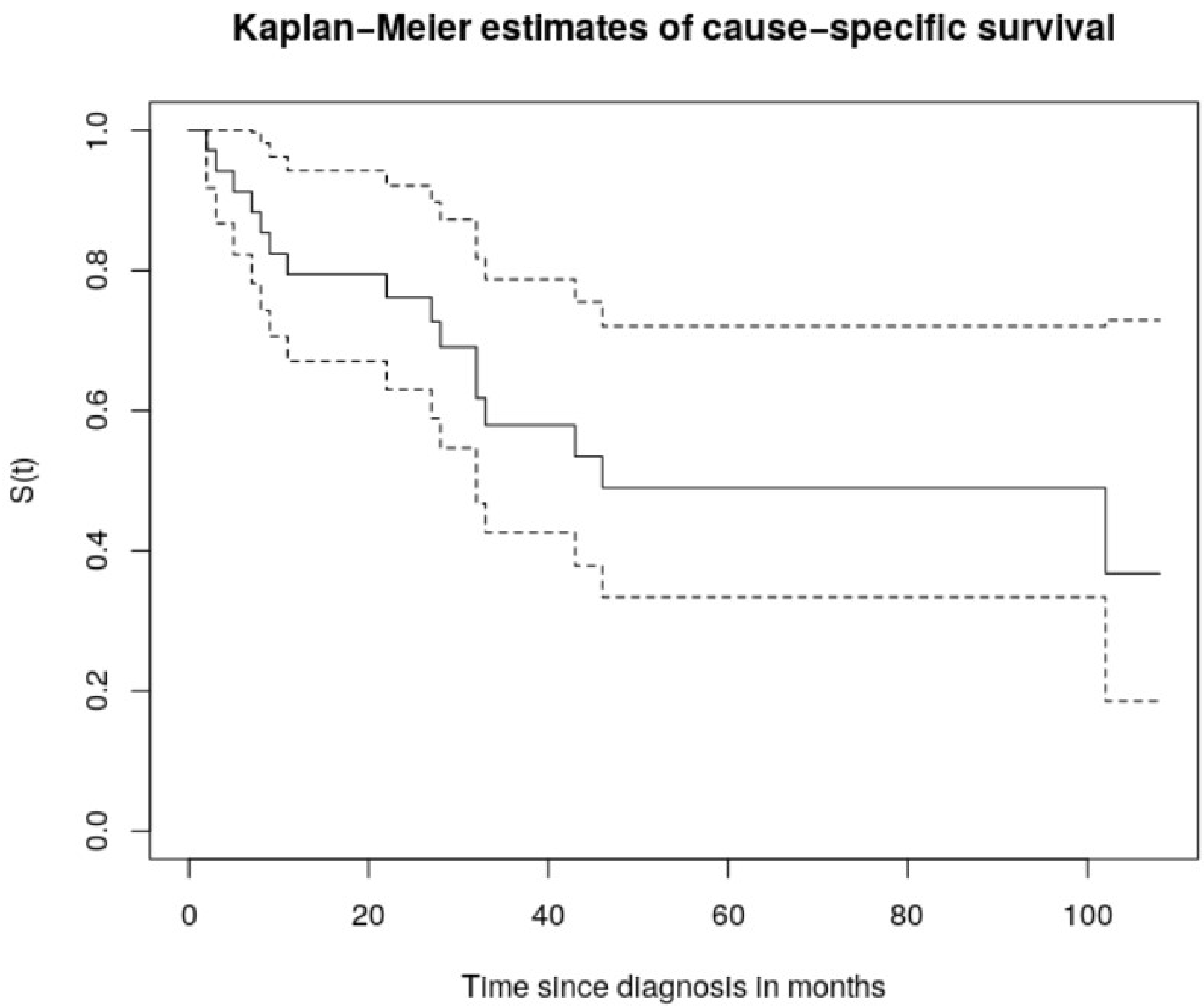
Example – Death Events in Colon Cancer

```
# make Kaplan-Meier estimates
mfit <- survfit(Surv(surv_mm, status == "Dead: cancer") ~ 1, data = colon_sample)

# print Kaplan-Meier table
summary(mfit)

# plot Kaplan-Meier curve
 plot(mfit,
    ylab="S(t)",
    xlab="Time since diagnosis in months",
    main = "Kaplan−Meier estimates of cause-specific survival")
```

```
## Call: survfit(formula = Surv(surv_mm, status == "Dead: cancer") ~ 1,
##     data = colon_sample)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2     35       1    0.971  0.0282        0.918        1.000
##     3     33       1    0.942  0.0398        0.867        1.000
##     5     32       1    0.913  0.0482        0.823        1.000
##     7     31       1    0.883  0.0549        0.782        0.998
##     8     30       1    0.854  0.0605        0.743        0.981
##     9     29       1    0.824  0.0652        0.706        0.962
##    11     28       1    0.795  0.0692        0.670        0.943
##    22     24       1    0.762  0.0738        0.630        0.921
##    27     22       1    0.727  0.0781        0.589        0.898
##    28     20       1    0.691  0.0823        0.547        0.872
##    32     19       2    0.618  0.0882        0.467        0.818
##    33     16       1    0.579  0.0908        0.426        0.788
##    43     13       1    0.535  0.0941        0.379        0.755
##    46     12       1    0.490  0.0962        0.334        0.720
##   102      4       1    0.368  0.1284        0.185        0.729
```


Kaplan−Meier estimates of cause−specific survival
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Time since diagnosis in months

K-M Estimator Median and Quantiles

- Recall, the p^{th} quantile of the survival function is

$$S(t_p) = P\left(T > t_p\right) = p$$

$$t_p = \inf\{\mathrm{t}: S(t_p) \leq p\}$$

- So, for the K-M estimator, $\hat{S}_K\left(t_p\right) = p$

$$t_p = \inf\{\mathrm{t}: \hat{S}_K\left(t_p\right) \leq p\}$$

Recall the example of the death events from the colon cancer

K-M Estimator Median and Quantiles

Recall the example of the death events from the colon cancer

```
##    33     16       1    0.579  0.0908          0.426          0.788
##    43     13       1    0.535  0.0941          0.379          0.755
##    46     12       1    0.490  0.0962          0.334          0.720
##   102      4       1    0.368  0.1284          0.185          0.729
```

$$\hat{S}_K(43) = 0.535$$
$$\hat{S}_K(46) = 0.490$$

The median survival time is 46 months

Nelson-Aalen Estimator

- Focus on estimating the cumulative hazard function $H(t)$
- Recall, $H(t) = \int_0^t h(x)dx$
- $H(t)$ can be approximated as

  $H(t) \approx \sum_{i:t_i \leq t} h(i)\Delta_i$ and $\quad h(i) = \frac{d_i}{n_i \Delta_i}$

  for $i = 1, \ldots, d$

  where Δ_i are intervals small enough to contain 1 event except ties

Nelson-Aalen Estimator

- Note, the focus is on the hazard estimation
  - Hazard function can be unstable and depending upon the length of the intervals
- Therefore, the focus is the $H(t)$, which can be estimated as

$$\widehat{H}(i) = \sum_{i:t_i \leq t} \frac{d_i}{n_i}$$

Fleming-Harrington Estimator

- Once the $\widehat{H}(i)$ is obtained
- Fleming-Harrington estimator for survival function can be obtained

$$\hat{S}_F(t) = e^{-\widehat{H}(t)}$$
$$= e^{-\sum_{i:t_i \le t} \frac{d_i}{n_i}}$$

Fleming-Harrington Estimator

- Therefore

$$\hat{S}_F(t) = \begin{cases} 1 & if\ t < t_1 \\ \prod_{t_i \le t} exp[-\frac{d_i}{n_i}] & if\ t \ge t_1 \end{cases}$$

- Variance

$$\text{var}\,\{\hat{S}_F(t)\} = \{\hat{S}_F(t)\}^2 \sum_{i=1}^{k} \frac{d_i}{{n_i}^2}$$

Nelson-Aalen(Fleming-Harrington) and K-M Estimators

- Survival function

$$\hat{S}_F(t) = \begin{cases} 1 & if\ t < t_1 \\ \prod_{t_i \le t} exp[-\frac{d_i}{n_i}] & if\ t \ge t_1 \end{cases}$$

$$\hat{S}_M(t) = \begin{cases} 1 & if\ t < t_1 \\ \prod_{t_i \le t} [1 - \frac{d_i}{n_i}] & if\ t \ge t_1 \end{cases}$$

Nelson-Aalen(Fleming-Harrington) and K-M Estimators

- Taylor expansion

$$exp\left[-\frac{d_i}{n_i}\right] = 1 - \frac{d_i}{n_i} + \frac{1}{2!}\left(\frac{d_i}{n_i}\right)^2 - \frac{1}{3!}\left(\frac{d_i}{n_i}\right)^3 + \ldots$$

$$exp\left[-\frac{d_i}{n_i}\right] \approx 1 - \frac{d_i}{n_i} \quad \text{when } \frac{d_i}{n_i} \text{ is small}$$

$$exp\left[-\frac{d_i}{n_i}\right] \geq 1 - \frac{d_i}{n_i}$$

- Fleming-Harrington estimator can be larger than K-M esimator

Example

```
ods graphics on;
proc lifetest data=example nelson
method=FH  alpha=0.05 plots=(s,h,p);
  time years*event(0);
  run;
ods graphics off;
```

The SAS System

Obs	Subjid	Years	Event
1	14	1	1
2	8	2	0
3	2	3	1
4	18	5	1
5	17	6	0
6	19	9	0
7	15	10	0
8	3	11	0
9	13	12	0
10	6	13	0
11	7	14	1
12	10	17	0
13	20	17	1
14	9	18	0
15	4	19	0
16	12	21	0
17	16	23	1
18	1	24	0
19	5	24	0
20	11	24	0

Example

Survival Function and Cumulative Hazard Rate								
Years		Fleming-Harrington			Nelson-Aalen		Number Failed	Number Left
		Survival	Failure	Survival Standard Error	Cumulative Hazard	Cum Haz Standard Error		
0.0000		1.0000	0	0	0	.	0	20
1.0000		0.9512	0.0488	0.0488	0.0500	0.0500	1	19
2.0000	*	.	.	.	.	.	1	18
3.0000		0.8998	0.1002	0.0691	0.1056	0.0747	2	17
5.0000		0.8484	0.1516	0.0830	0.1644	0.0951	3	16
6.0000	*	.	.	.	.	.	3	15
9.0000	*	.	.	.	.	.	3	14
10.0000	*	.	.	.	.	.	3	13
11.0000	*	.	.	.	.	.	3	12
12.0000	*	.	.	.	.	.	3	11
13.0000	*	.	.	.	.	.	3	10
14.0000		0.7677	0.2323	0.1104	0.2644	0.1380	4	9
17.0000		0.6870	0.3130	0.1277	0.3755	0.1772	5	8
17.0000	*	.	.	.	.	.	5	7
18.0000	*	.	.	.	.	.	5	6
19.0000	*	.	.	.	.	.	5	5
21.0000	*	.	.	.	.	.	5	4
23.0000		0.5350	0.4650	0.1837	0.6255	0.3064	6	3
24.0000	*	.	.	.	.	.	6	2
24.0000	*	.	.	.	.	.	6	1
24.0000	*	.	.	.	.	.	6	0

Product-Limit Survival Estimates						
Years		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	20
1.0000		0.9500	0.0500	0.0487	1	19
2.0000	*	.	.	.	1	18
3.0000		0.8972	0.1028	0.0689	2	17
5.0000		0.8444	0.1556	0.0826	3	16
6.0000	*	.	.	.	3	15
9.0000	*	.	.	.	3	14
10.0000	*	.	.	.	3	13
11.0000	*	.	.	.	3	12
12.0000	*	.	.	.	3	11
13.0000	*	.	.	.	3	10
14.0000		0.7600	0.2400	0.1093	4	9
17.0000		0.6756	0.3244	0.1256	5	8
17.0000	*	.	.	.	5	7
18.0000	*	.	.	.	5	6
19.0000	*	.	.	.	5	5
21.0000	*	.	.	.	5	4
23.0000		0.5067	0.4933	0.1740	6	3
24.0000	*	.	.	.	6	2
24.0000	*	.	.	.	6	1
24.0000	*	.	.	.	6	0

Homework 2

- Construct the first 4 rows of the life table table by hand
  - using the example data set by 4-year intervals

Interval	Time Period	Events d_i	Censor c_i	At risk at the beginning of the interval n_i	Average number at risk in the interval n_i'	Survival probability $S(t)$	PDF f(t)	Hazard h(t)	$se(S(t))$
1	[0,4)								
2	[4,8)								
3	[8,12)								
4	[12,16)								

The SAS System

Obs	Subjid	Years	Event
1	14	1	1
2	8	2	0
3	2	3	1
4	18	5	1
5	17	6	0
6	19	9	0
7	15	10	0
8	3	11	0
9	13	12	0
10	6	13	0
11	7	14	1
12	10	17	0
13	20	17	1
14	9	18	0
15	4	19	0
16	12	21	0
17	16	23	1
18	1	24	0
19	5	24	0
20	11	24	0

Show that the following two ways of deriving hazard are equivalent

- Number of events per person-time-units

$$\hat{h}(t_{mi}) = d_i/[(t_i - t_{i-1})\,(n_i' - d_i/2)]$$

- Based on the definition

$$\hat{h}(t_{mi}) = {}^{\hat{f}(t_{mi})}/_{\hat{S}(t_{mi})} = {}^{2\hat{f}(t_{mi})}/_{[\hat{S}(t_i)+\hat{S}(t_{i-1})]}$$

Ovarian Data

	futime	fustat	age	resid.ds	rx	ecog.ps
1	59	1	72.3315	2	1	1
2	115	1	74.4932	2	1	1
3	156	1	66.4658	2	1	2
4	421	0	53.3644	2	2	1
5	431	1	50.3397	2	1	1
6	448	0	56.4301	1	1	2
7	464	1	56.9370	2	2	2
8	475	1	59.8548	2	2	2
9	477	0	64.1753	2	1	1
10	563	1	55.1781	1	2	2
11	638	1	56.7562	1	1	2
12	744	0	50.1096	1	2	1
13	769	0	59.6301	2	2	2
14	770	0	57.0521	2	2	1
15	803	0	39.2712	1	1	1
16	855	0	43.1233	1	1	2
17	1040	0	38.8932	2	1	2
18	1106	0	44.6000	1	1	1
19	1129	0	53.9068	1	2	1
20	1206	0	44.2055	2	2	1
21	1227	0	59.5890	1	2	2
22	268	1	74.5041	2	1	2
23	329	1	43.1370	2	1	1
24	353	1	63.2192	1	2	2
25	365	1	64.4247	2	2	1
26	377	0	58.3096	1	2	1

- Dataset available in R survival package
  - futime: survival or censoring time (day)
  - fustat: censoring status (censor=0)
  - age: in years
  - resid.ds: residual disease present (1=no,2=yes)
  - rx: treatment group
  - ecog.ps: ECOG performance status (1 is better, see reference)
- Perform survival analyses
  - Create life-table stratified by rx
  - Plot hazard function by rx based on life-table estimate
  - Plot K-M survival function by rx
  - What is the median survival time for each treatment group?
  - Compare survival function estimations between K-M and F-H methods
- Describe your analyses and write conclusions based on your analyses

References

Edmunson, J.H., Fleming, T.R., Decker, D.G., Malkasian, G.D., Jefferies, J.A., Webb, M.J., and Kvols, L.K., Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma vs. Minimal Residual Disease. Cancer Treatment Reports, 63:241-47, 1979.