SOC 7717 Event History Analysis and Sequence Analysis

Week 3: Nonparametric Methods

Wen Fan

Spring 2019

Outline

Setup

Life Table Method

Kaplan–Meier Method

Standard Errors, Confidence Intervals, and Cumulative Hazard Function

Comparison of the Life Table and Kaplan–Meier Methods

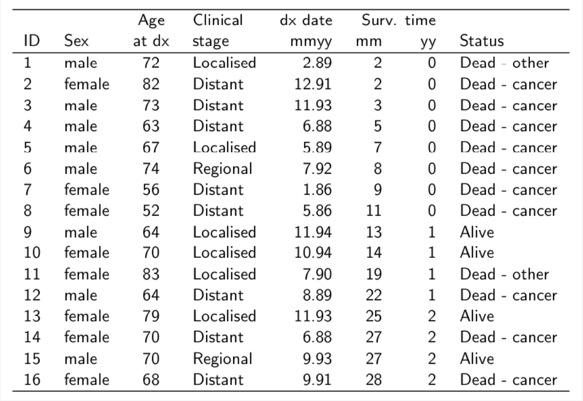
Comparison of Survival between Groups

Setup

Goal

- Consider the sample data for 35 colon cancer patients (see part of the data on the next slide, sorted by survival time)
- Our goal is to estimate $S(t)$, where the event of interest is death due to any cause
- Formally, let $S(t) = Pr(T > t)$ where T is the time of the event, i.e., the probability that an individual “survives” to time t
  - Recall that $0 \leq S(t) \leq 1$ is a non-increasing function of t

Data



ID	Sex	Age at dx	Clinical stage	dx date mmyy	Surv. time mm	 yy	Status
1	male	72	Localised	2.89	2	0	Dead - other
2	female	82	Distant	12.91	2	0	Dead - cancer
3	male	73	Distant	11.93	3	0	Dead - cancer
4	male	63	Distant	6.88	5	0	Dead - cancer
5	male	67	Localised	5.89	7	0	Dead - cancer
6	male	74	Regional	7.92	8	0	Dead - cancer
7	female	56	Distant	1.86	9	0	Dead - cancer
8	female	52	Distant	5.86	11	0	Dead - cancer
9	male	64	Localised	11.94	13	1	Alive
10	female	70	Localised	10.94	14	1	Alive
11	female	83	Localised	7.90	19	1	Dead - other
12	male	64	Distant	8.89	22	1	Dead - cancer
13	female	79	Localised	11.93	25	2	Alive
14	female	70	Distant	6.88	27	2	Dead - cancer
15	male	70	Regional	9.93	27	2	Alive
16	female	68	Distant	9.91	28	2	Dead - cancer

Scenarios

- When there is no censoring, it's easy to estimate the survivor function. For any t, simply estimate $S(t)$ by the proportion of cases still alive at selected values of t
  - E.g., eight of the 35 patients died during the first year of follow-up, so the estimate for $S(1)$ is $\hat{S}(1) = \frac{(35-8)}{35} \approx 0.771$
- If all event times are less than all censored times, the survivor function can again be estimated by the proportion surviving, for all times up to the lowest censored time
  - After that the estimate is undefined (though it must be between the last value and 0)

Scenarios

- If some censored times are less than some event times, problems occur
  - E.g., when we estimate $S(2)$, note that ten patients died within two years of follow-up, but 2 patients (patients 9 and 10) could not be followed-up for a full 2 years
- How do we deal with these two patients?
  - We could exclude these two patients from the analysis and let $\hat{S}(2) = \frac{(35-2-10)}{35-2}$ → problems?
  - We could instead use $\hat{S}(2) = \frac{(35-10)}{35}$ → problems?

A nonparametric solution

- Two common nonparametric methods for estimating $S(t)$ in the presence of censoring:
  - The life table (actuarial) method
  - The Kaplan-Meier (product-limit) method
- Advantage: we don't have to make constraining assumptions about the distribution of event times
  - Few researchers have a sound basis for preferring one distribution over another
  - Adopting an incorrect assumption can lead to erroneous conclusions

Life Table Method

Life Table Method

- Also known as the actuarial method
- The approach is to divide the period of observation into a series of time intervals and estimate the conditional (interval-specific) survival proportion for each interval
  - Typically each interval includes the initial time and excludes the concluding time
- The survivor function, $S(t)$, at the end of a specified interval is then given by the product of the interval-specific survival proportions for all intervals up to and including the specified interval

Life Table Method

- In the absence of censoring, the interval-specific survival proportion is $p = \frac{l-d}{l}$, where d is the number of events (deaths) observed during the interval and l is the number of people alive at the start of the interval
  - E.g., for the first interval, $l(1) = 35$ and $d(1) = 8$
  - $\rightarrow p(1) = \frac{(35-8)}{35} \approx 0.771$
  - $\rightarrow$ The estimated 1-year survival proportion is therefore $S(\hat{1}) = 0.771$

Life Table Method

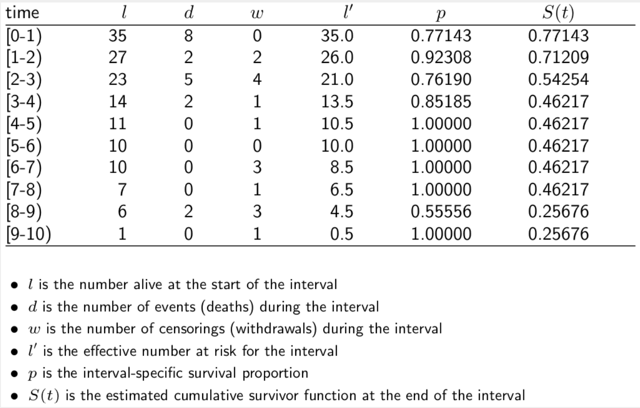
- In the presence of censoring, it is assumed that censoring occurs uniformly throughout the interval such that each censored individual is at risk for, on average, half of the interval
  - The effective number of individuals at risk during the interval is given by $l' = l - 0.5w$ where l is the number of individuals alive at the start of the interval and w is the number of censorings during the interval
  - The estimated interval-specific survival proportion is then given by $p = \frac{l'-d}{l'}$
  - E.g., for the second interval, $l'(2) = 27 - 0.5 \times 2 = 26$ and $p(2) = \frac{(26-2)}{26} \approx 0.923$

Life table method

- The probability of surviving through the 1st interval is $S(1) = p(1)$
- The probability of surviving through the 2nd interval is $S(2) = p(1) \times p(2)$
- The probability of surviving through the 3rd interval is $S(3) = p(1) \times p(2) \times p(3)$
- ...

E.g., if we want the probability of surviving past, say, the 2nd year, an appropriate estimate is $\hat{S}(2) = p(1) \times p(2) = 0.771 \times 0.923 \approx 0.712$

Life table with annual interval



time	l	d	w	l'	p	$S(t)$
[0-1)	35	8	0	35.0	0.77143	0.77143
[1-2)	27	2	2	26.0	0.92308	0.71209
[2-3)	23	5	4	21.0	0.76190	0.54254
[3-4)	14	2	1	13.5	0.85185	0.46217
[4-5)	11	0	1	10.5	1.00000	0.46217
[5-6)	10	0	0	10.0	1.00000	0.46217
[6-7)	10	0	3	8.5	1.00000	0.46217
[7-8)	7	0	1	6.5	1.00000	0.46217
[8-9)	6	2	3	4.5	0.55556	0.25676
[9-10)	1	0	1	0.5	1.00000	0.25676

- l is the number alive at the start of the interval
- d is the number of events (deaths) during the interval
- w is the number of censorings (withdrawals) during the interval
- l' is the effective number at risk for the interval
- p is the interval-specific survival proportion
- $S(t)$ is the estimated cumulative survivor function at the end of the interval

Life table method

The survival proportion, despite commonly being called the survival rate, is a proportion

To truly capture rate, we need the hazard function, which is estimated as:

$$\frac{d_j}{b_j(l_j - \frac{w_j}{2} - \frac{d_j}{2})}$$

where d_j is the number of deaths in the interval i, b_j is the length of the interval, l_j is the number still alive at the start of the interval, and w_j is the number censored in the interval

Note that we subtract $\frac{d}{2}$ assuming those who die are at risk for only half the entire interval

Exercise

Produce a life table based on the following survival data collected from 68 patients in the Stanford Heart Transplantation Program

Number of Days	Number of Deaths	Number Censored
0–49	16	3
50–99	11	0
100–199	4	2
200–399	5	4
400–699	2	6
700–999	4	3
1000–1299	1	2
1300–1599	1	3
1600+	0	1

Kaplan-Meier Method

Background

- Also known as the product-limit method but is more commonly known as the Kaplan-Meier (KM) method, after the two researchers who first published the method in 1958
- In essence, the KM method is a limiting form of the life table method with the interval size decreased towards zero. Each life table interval is of infinitesimal length, just enough for one event or time increment
- In practice, survival time is measured on a discrete scale (e.g. days, months, or years) so the interval length is limited by the accuracy to which survival time is measured

Estimation setup

- Only those intervals containing an event contribute to the KM estimate, so we can ignore all other intervals
  - Suppose that we calculate a life table in which the interval is 1 month. The life table probability of surviving past the 50th month is:

$$p_1 \times p_2 \times p_3 .... \times p_{50}$$

  - For 27 of these 50 months, however, no deaths occurred. So the estimate for p on those months is 0 → we only need to calculate terms for the months on which events occurred

Estimation setup

- To obtain KM estimates of survival, survival times are first ranked in increasing order
  - The times where events (deaths) occur are denoted by t_i, where $t_1 < t_2 < t_3 < ...$
- Unlike for the life table estimator, if observations are censored on the same month that events occurred, they are assumed to be at risk for the whole month rather than half the month

Kaplan-Meier Method

The general form of the KM estimator is

$$S(\hat{t}) = \prod_{t_j \leq t} (1 - \frac{d_j}{l_j})$$

where d_j is the number of events at time t_j and l_j is the number "at risk" at time t_j

- Censorings do not affect the estimate of $S(t)$, but contribute by decreasing l_i, the number of persons at risk, at the next death time
- If the largest observed survival time (t_z) is a censored survival time, then $S(\hat{t})$ is undefined for $t > t_z$, otherwise $S(\hat{t}) = 0$ for $t > t_z$

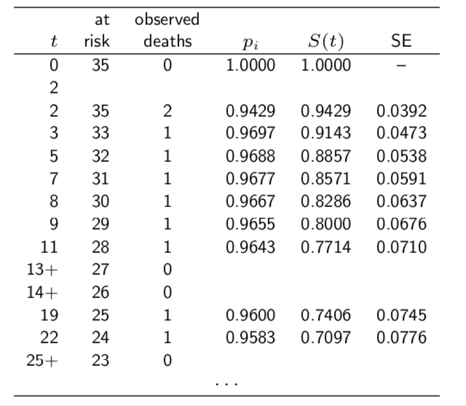
Kaplan-Meier Method

Consider the example we presented in the beginning:

- At $t = 2$ months we observed 2 deaths among the 35 patients at risk, so $p_1 = 1 - \frac{2}{35} \approx 0.943$
- At $t = 3$ months we observed 1 death among the 33 patients at risk, so $p_2 = 1 - \frac{1}{33} \approx 0.970$
- Subsequently, $S(\hat{t}) = 0.943 \times 0.970 = 0.914$ for $3 \leq t < 5$

When a censoring happens at the same time as a failure, Stata assumes that the failure occurred before the censoring (but you can change the assumption easily, see p.97 in CGM)

Kaplan-Meier estimates



	at	observed			
t	risk	deaths	p_i	$S(t)$	SE
0	35	0	1.0000	1.0000	–
2					
2	35	2	0.9429	0.9429	0.0392
3	33	1	0.9697	0.9143	0.0473
5	32	1	0.9688	0.8857	0.0538
7	31	1	0.9677	0.8571	0.0591
8	30	1	0.9667	0.8286	0.0637
9	29	1	0.9655	0.8000	0.0676
11	28	1	0.9643	0.7714	0.0710
13+	27	0			
14+	26	0			
19	25	1	0.9600	0.7406	0.0745
22	24	1	0.9583	0.7097	0.0776
25+	23	0			
		. . .			

Standard Errors, Confidence Intervals, and Cumulative Hazard Function

Estimating the Standard Error of S(t)

Obtained by the Greenwood's formula (1926)

- Appropriate for both the life table and KM methods
- The default method for most softwares
- Reduces to the binomial standard error in the absence of censoring

Confidence Intervals

To provide a measure of uncertainty associated with the point estimate

- A 95% confidence interval (CI) is an interval such that under repeated sampling, the true survival proportion will be contained in the interval 95% of the time
- A two-sided $100(1-\alpha)\%$ CI ranges from $p - z_{\alpha/2}\text{SE}(p)$ to $p + z_{\alpha/2}\text{SE}(p)$, where p is the estimated survival proportion, SE(p) the associated standard error, and $z_{\alpha/2}$ the upper $\alpha/2$ percentage point of the standard normal distribution

Confidence Intervals

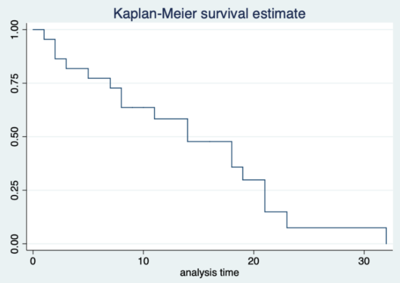
- For a 95% confidence interval, $z_{\alpha/2} = 1.96$
- For a 99% confidence interval, $z_{\alpha/2} = 2.58$
- Confidence intervals obtained in this way are symmetric with respect to the point estimate, and can sometimes contain implausible values for the survival proportion, i.e., values less than zero or greater than one
- One method to correct for this problem is to transform the estimate to a value in the range $[-\infty, \infty]$, obtain a confidence interval for the transformed value, and then back-transform the confidence interval to $[0, 1]$ (e.g., log-log transformation)

Confidence Intervals

- Pointwise confidence intervals: CI's at a specified time point
- Confidence bands: CI's for the entire survival function

Kaplan-Meier Estimates

A plot of the KM estimate of the survivor function takes the form of a step function, in which the survival probabilities decrease at each death time and are constant between adjacent deaths times



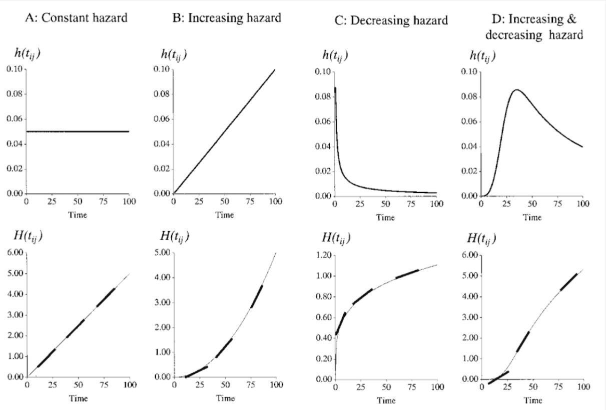
Cumulative hazard function

Cumulative hazard function (negative log survivor function)

$$\Lambda(t) = \int_0^t h(u)du = -\log S(t)$$

- If the hazard is constant over time, the cumulative hazard will be a straight line sloping upward
- If the cumulative hazard function curves upward, it's evidence for an increasing hazard
- If the cumulative hazard function curves downward, it's evidence for a decreasing hazard

Hazard and cumulative hazard functions



Nelson-Aalen estimates

One way to estimate $\Lambda(t)$ is to plug in the KM estimate of $S(t)$ in the the equation of the previous slide (the negative log survival function method)

But we can estimate $\Lambda(t)$ in another way that has better small-sample properties → Nelson-Aalen estimator

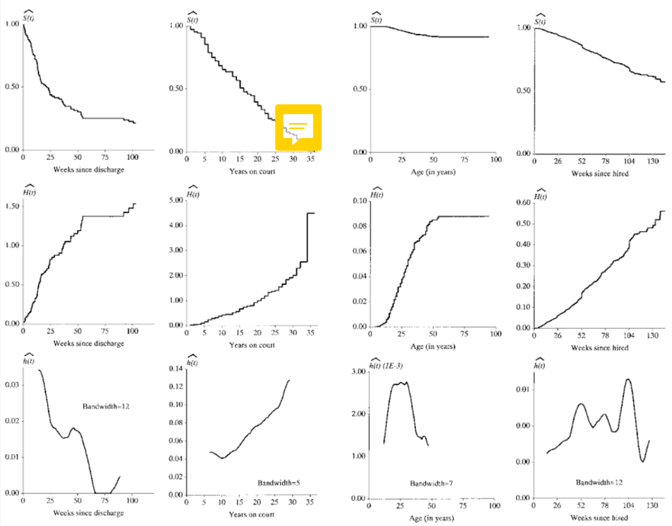
$$\hat{\Lambda(t)} = \sum_{t_j \leq t} \frac{d_j}{l_j}$$

where d_j is the number of events at time t_j and l_j is the number "at risk" at time t_j

Hazard function

It’s possible to get nonparametric estimates and graphs of hazard functions, but they tend to be very choppy. Two ways to avoid this

- Divide the time scale into intervals and use standard life table methods to get a hazard estimate for each interval
- Use smoothing techniques (e.g., weighted averages around each point)

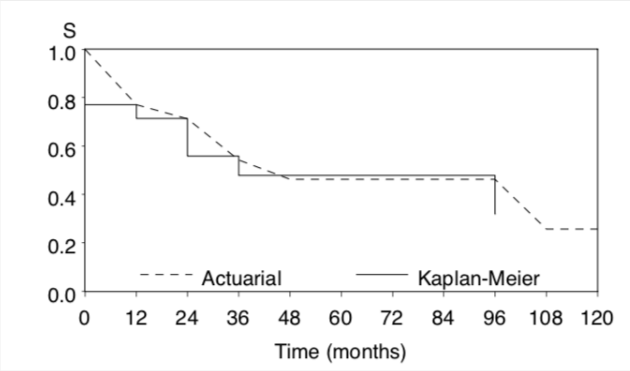


Comparison of the Life Table and Kaplan-Meier Methods

Tabular and graphical forms

- Estimates of the survivor function can be presented in either tabular or graphical form
- For tabular presentations, we rarely require estimates of the survivor function for interval lengths shorter than one year → the life table method suffices, but note that
  - The number of censored observations in an interval is not halved by the KM estimator, but is halved by the life table estimator
  - The interval boundaries for the KM estimator are determined by the event times themselves

Comparison of the two methods



Comparison of the two methods

1. How ties are handled
   - If two individuals have the same survival time (time to event or time to censoring), we say that the survival times are “tied”
   - The KM estimator was developed for applications where survival time is measured on a continuous scale, where ties are rare, although the KM estimator is discrete in nature
   - The life table method takes account of ties through the assumption that the censored individuals were at risk for half of the interval, whereas the KM method overestimates the survivor function in the presence of ties by assuming that censored individuals were at risk for all of the interval

Comparison of the two methods

2. How estimates are obtained and interpreted
   - The life table method provides estimates of the survivor function at the end of each interval, and no estimate of the survivor function is made between the interval endpoints
     - It estimates $S(t)$ for values of t between the interval endpoints through interpolation, which corresponds to assuming an approximately even distribution of deaths within each interval
     - This assumption may not always be valid (e.g., mortality after surgery)

Comparison of the two methods

2. How estimates are obtained and interpreted
   - The KM method provides an estimate of $S(t)$ for all values of t, although the estimate of $S(t)$ is constant between event times
     - In other words, the KM estimator changes only at an observed event time
     - The interpretation of the KM estimate of $S(t)$ presented in the previous figure is that 22% of the patients die immediately following diagnosis but no deaths occur for another 12 months

Comparison of the two methods

3. Computational time
   - The life table method with annual intervals has been popular in population-based survival analysis since it requires fewer arithmetic calculations than would be required for the KM method with survival time measured in months
   - With the advent of computers, however, this is no longer a significant advantage
   - Either method will suffice, although the life table method could be considered technically superior for large data sets or when the measurement of event times is crude (although the differences are insignificant in practice)

Comparison of Survival between Groups

Comparing Survival between Groups

- Comparing survival at a fixed time point (e.g., five years) wastes available information
- Various global tests are available (parametric and non-parametric) for testing equality of survival curves
- Different tests
  - Log rank test
  - Wilcoxon test
  - Tarone-Ware test
  - Peto-Peto-Prentice test
  - Fleming-Harrington test

Log rank test

- The most common is the log rank test, which is non-parametric
- Sensitive to departures from the null hypothesis in which the two hazards are proportional over time. Very insensitive to situations in which the hazard functions cross
- Puts equal weight on every failure (irrespective of the number at risk at the time of the failure) ($W(t_j) = 1$)

Wilcoxon test

- The Wilcoxon test is constructed by weighting the contribution of each failure time by the total number of individuals at risk ($W(t_j) = n_j$)
- Because the Wilcoxon test gives more weight to early times than to late times (the number at risk never increases with time), it is more sensitive to differences early in the follow-up period (when the number at risk is larger)
- More powerful than the log rank test if the proportional hazards assumption does not hold
- Neither the log rank nor the Wilcoxon test is good at detecting differences when survival curves cross

Other tests

- Tarone-Ware test: $W(t_j) = \sqrt{n_j} \rightarrow$ more weight is given to earlier times but not as much as the Wilcoxon test
- Peto-Peto-Prentice test: $W(t_j) = S(\tilde{t_j})$ where $S(\tilde{t_j})$ is similar to the KM estimator
- Fleming-Harrington test: $W(t_j) = [S(\hat{t_j})]^p[1 - S(\hat{t_j})]^q$ where $S(\hat{t_j})$ is the KM estimator, and p and q are chosen by the user
  - When $p > q$: more weight is given to earlier times
  - When $p < q$: more weight is given to later times
  - When $p = q = 0$: reduces to the log rank test