Censoring and Life Table Estimates

- With censored data, it is common to use nonparametric or semiparametric models, rather than parametric models like the Weibull or exponential.

Reasons for Censoring

- Some individuals are still alive at the end of the data collection. These subjects are right-censored.
- If we have staggered entry, we measure lifetimes from the point of entry into the study, so the censoring time may differ among individuals (random censoring).
- We may have loss to follow-up (e.g., patient moves away or stops coming to the clinic).
- We may have death from another cause, or competing risk (e.g., a cancer patient dies in a car accident).

Example 1: Consider the following grouped data in the form of a life table:

Here, $n(t)$ = # alive and under observation at beginning of interval

$d(t)$ = # dying during interval

$w(t)$ = # censored or withdrawn during interval

Time Interval	$n(t)$	$d(t)$	$w(t)$
[0,1)	80	12	2
[1,2)	66	8	4
[2,3)	54	10	5
[3,4)	39	5	4
[4,5)	30	1	2
[5,6)	27	2	7
[6,7)	18	3	5
[7,8)	10	3	7

- The r.v. T is the time until death.
- Consider estimating the 4-year survival probability, $S_T(4)$.

Naive estimate A:

Naive estimate B:

- Estimate A would be correct if all withdrawing individuals left the study

- This is not really true, so is an

  ______________________.

- Estimate B would be correct if all withdrawing individuals left the study

- This is not really true, so is an ______________________.

The Life-Table Estimate of $S_T(t)$

- Note $S_T(4) =$

- We need to estimate $m_T(i-1)$, for $i=1,\ldots,4$, using the data.

Possible Approaches to Estimate $m(t)$

① Assume the interval's censored subjects exited the study at the end of the interval. Then we would estimate $m(t)$ by:

② Assume the interval's censored subjects exited the study at the beginning of the interval. Then we would estimate $m(t)$ by:

- Most likely, neither ① nor ② reflect reality.

③ <u>Compromise</u>: Estimate $m(t)$ by:

- Let's fill in the first 4 rows of our life table using approach ③:

Time	$n(t)$	$d(t)$	$w(t)$	$\hat{m}(t)$	$1-\hat{m}(t)$	$\prod(1-\hat{m}(t))$
$[0,1)$	80	12	2			
$[1,2)$	66	8	4			
$[2,3)$	54	10	5			
$[3,4)$	39	5	4			

- So our life-table estimate $\hat{S}_T(4) =$

<u>Exercise</u>: Show that using approach ①, we obtain $\hat{S}_T(4) =$

<u>Exercise</u>: Show that using approach ②, we obtain $\hat{S}_T(4) =$

- We know is an ______________ and is an ______________ of $S_T(4)$, but they are not as bad as our "naive" estimates.

- We will define $\hat{S}_T(t)$ as the estimator using approach ③.

Sampling Distribution of $\hat{S}_T(t)$

- Clearly, $\hat{S}_T(t)$ is a random variable since it is a function of sample data.
- It can be shown that for a fixed t, $\hat{S}_T(t)$ is approximately ________ with mean and a variance which is consistently estimated by

- This is called Greenwood's formula.
- An approximate large-sample $100(1-\alpha)\%$ confidence interval for $S_T(t)$ is thus:

- Code on the course web page enables easy calculation of these quantities.

Example 1: An approximate 95% CI for $S_T(4)$ is:

- With 95% confidence,

The Kaplan-Meier Estimator

- Suppose that instead of having grouped data (as with the life table), we know the exact survival times (or censoring times).
- This is like making the interval widths so small that no more than one observation falls in any interval.
- The "limit" of the life-table estimator of $S_T(t)$, as interval-width → 0, is called the Kaplan-Meier (product-limit) estimator.

Example 1: Consider the following simple data set with $n = 6$ patients. The censoring indicator is "1" if the observation is a death time and "0" if it is a censoring time.

Time	2.5	5.5	6.5	9.5	11.5	13.5
Cens. Ind.	1	1	0	1	0	1

- We will estimate the mortality rate at time t as

Time	$\hat{m}(t)$	$1-\hat{m}(t)$	$\hat{S}(t)$

- The K-M estimator is a ______________ with jumps at the death times (the function is defined to be right-continuous):

Plot:

- when we have more data, this more closely resembles a continuous survival function.

General Formulas for K-M Estimator

Let $T_i =$

and $C_i =$

Note that if $T_i \leq C_i$, we observe

- If $T_i > C_i$, we only observe

- So for each subject, we actually observe

- Define the indicator

- Hence our data are pairs

- Define the number of individuals at risk at time t by $n(t)$.
- So $n(t)$ is the number of subjects who have neither died nor been censored by time t.
- Then the K-M estimator of $S_T(t)$ is:

- This formula works if there are no tied survival times in the sample.
- For continuous data, the probability of ties is , but ties can occur in reality since data are given in rounded form.

- If $d(t)$ is the number of observed deaths in the sample at time t, then $d(t)$ will always be ____ or ____ if there are no ties.
- But we could have $d(t)$ ____________ if ties are possible.
- Then the K-M estimator is:

where $A(u)$ is the set of all death times u that are less than or equal to t.

- A consistent estimator of $\text{var}[KM(t)]$ is the limit of Greenwood's formula:

- For a fixed t, it can be shown that $KM(t)$ is approximately normal for large samples.

- So a $100(1-\alpha)\%$ CI for $S_T(t)$ is:

- We can find and plot KM(t), and 95 % CIs, using the 'survfit' function in the 'survival' package in R.
- See the course web page for examples.

<u>Example 2</u>: The built-in 'cancer' data set in the 'survival' package in R gives survival and censoring times (in days) for 228 advanced lung cancer patients.

- We find the K-M estimate of the survival function (and pointwise 95 % CIs) using R.
- Estimate the one-year survival probability.

- Estimate the two-year survival probability.

Example 3: The built-in 'stanford2' data set gives survival and censoring times (in days) for 184 heart transplant patients.

- Estimate the one-year survival probability.

- Estimate the three-year survival probability.