BIST P8110: Applied Regression II

4. Estimating Quantiles and Mean of Survival Time

Qixuan Chen

Department of Biostatistics
Columbia University

This lecture's big ideas

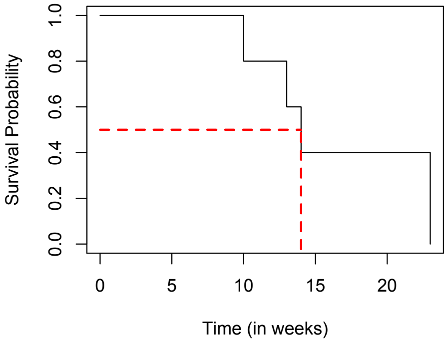
- Have you heard of “median survival time"?
- Quantiles estimation for survival data
- Mean estimation for survival data

Point Estimation of Quantiles

- Instead of estimating the survival probability at a given time, it is often useful to supplement the presentation with point and interval estimates of key quantiles.
  - For example, we want to know the median survival time, t, such that $S(t) = 0.50$.
- The estimated survival function can be used to estimate quantiles of the survival time.
- Quantiles can easily be estimated graphically.

Graphical Presentation of Quantiles

- To obtain graphic estimates, begin on the survival probability axis at the quantile of interest, and draw a horizontal line until it first touches the estimated survival function. A vertical line is drawn to the time axis to obtain the estimated quantile.



Point Estimation of Quantile

- The general formula for estimating the 100pth percentile is

$$\hat{t}_p = min\{t_j : \hat{S}(t_j) < 1 - p\}$$

- The three quartiles of survival time are defined as follow:

$$\hat{t}_{.25} = min\{t_j : \hat{S}(t_j) < 0.75\}$$
$$\hat{t}_{.50} = min\{t_j : \hat{S}(t_j) < 0.5\}$$
$$\hat{t}_{.75} = min\{t_j : \hat{S}(t_j) < 0.25\}$$

- If $\hat{S}(t_j)$ is exactly equal to $1 - p$, the 100pth percentile is taken to be $(t_j + t_{j+1})/2$.

NOTE: The quantile estimators here are defined the same as SAS, but differently from the definition in Hosmer et al.

Cancer Recurrence Example

Revisit the cancer recurrence example, among the 8 subjects, 5 have cancer returned at 10, 13, 14, 17, and 23 weeks, respectively after treatment. The other 3 subjects drop out at 13, 19, and 25 weeks.

t_j	Interval	n_j	d_j	$p_j = \frac{n_j - d_j}{n_j}$	$\hat{S}(t_j)$	95% CI
0	$[0, 10)$	8	0	1	1	
10	$[10, 13)$	8	1	$7/8$	0.875	(0.387, 0.981)
13	$[13, 14)$	7	1	$6/7$	0.75	(0.315, 0.931)
14	$[14, 17)$	5	1	$4/5$	0.6	(0.196, 0.852)
17	$[17, 23)$	4	1	$3/4$	0.45	(0.108, 0.751)
23	$[23, \infty)$	2	1	$1/2$	0.225	(0.012, 0.602)

Point Estimation of Quantiles

- Estimate median cancer recurrence survival time:

$$\hat{t}_{.50} = min\{t_j : \hat{S}(t_j) < 0.50\}$$

We have

$$\hat{S}(17) = 0.45,\ \hat{S}(23) = 0.225$$

$$\rightarrow \hat{t}_{.50} = 17$$

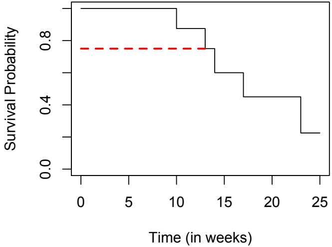
- Interpretation: We estimate that half of patients will survive more than 17 weeks without cancer recurrence after treatment.

Point Estimation of Quantiles

- Estimate the first quartile (or the 25th percentile) of survival time:

$$\hat{t}_{.25} = min\{t_j : \hat{S}(t_j) < 0.75\}$$

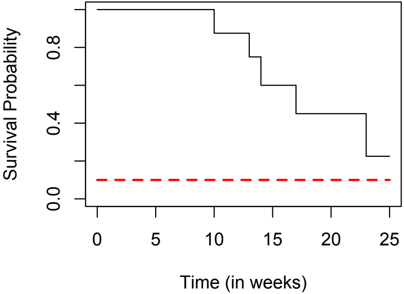
$$\hat{S}(13) = 0.75 \quad \hat{S}(14) = 0.6 \rightarrow \hat{t}_{.25} = (13 + 14)/2 = 13.5$$



- Note: $\hat{S}(t_j) = 0.75$, so $\hat{t}_{.25} = \frac{t_j + t_{j+1}}{2}$.
- Interpretation: We estimate that 75 percent of patients will survive more than 13.5 weeks without cancer recurrence after treatment or 25 percent of patients will have cancer recurrence by 13.5 weeks after treatment.

More on Quantile Estimation

- Only quantiles within the observed range of the estimated survival function may be estimated.
  - Example: Since the last censoring time (25) is greater than the largest event time (23), the survival function cannot go to zero ($\hat{S}(23) = 0.225$). In this case, we cannot estimate 90th percentile.



Interval Estimation for Quantiles

- The 95% CI for the $(100p)$th percentile survival time is the set of all points t satisfying

$$\left| \frac{\log[-\log \hat{S}(t)] - \log[-\log(1-p)]}{\sqrt{\widehat{Var}(\log[-\log \hat{S}(t)])}} \right| \leq 1.96 \qquad (1)$$

if we use a log-log transformation on $S(t)$.

Interval Estimation for Quantiles

- We can re-write

$$\left| \frac{\log[-\log \hat{S}(t)] - \log[-\log(1-p)]}{\sqrt{\widehat{Var}(\log[-\log \hat{S}(t)])}} \right| \leq 1.96$$

Interval Estimation for Quantiles

- Equation (1) can be re-written as

$$\exp[-\exp(\hat{C}_u)] \leq 1 - p \leq \exp[-\exp(\hat{C}_l)] \qquad (2)$$

  where,

  - $\hat{C}_u = \log[-\log \hat{S}(t)] + 1.96\sqrt{\widehat{Var}(\log[-\log \hat{S}(t)])}$
  - $\hat{C}_l = \log[-\log \hat{S}(t)] - 1.96\sqrt{\widehat{Var}(\log[-\log \hat{S}(t)])}$
- The 95% CI for the $(100p)$th percentile survival time is the set of all points t satisfying equation (2).
- This is the Brookmeyer and Crowley (1982) method.

Interval Estimation for Quantiles

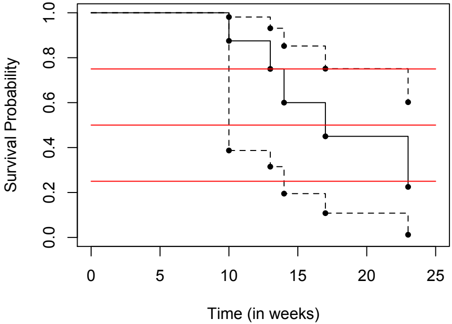
- Revisit the cancer recurrence example:

t_j	Interval	n_j	d_j	$p_j = \frac{n_j - d_j}{n_j}$	$\hat{S}(t_j)$	95% CI
0	$[0, 10)$	8	0	1	1	
10	$[10, 13)$	8	1	$7/8$	0.875	(0.387, 0.981)
13	$[13, 14)$	7	1	$6/7$	0.75	(0.315, 0.931)
14	$[14, 17)$	5	1	$4/5$	0.6	(0.196, 0.852)
17	$[17, 23)$	4	1	$3/4$	0.45	(0.108, 0.751)
23	$[23, \infty)$	2	1	$1/2$	0.225	(0.012, 0.602)

- The 95% CI for
  - $t_{0.25}$ (first quartile) is [10, 23)
  - $t_{0.5}$ (median) is [10, .)
  - $t_{0.75}$ (third quartile) is [14, .)

Interval Estimation for Quantiles

- Brookmeyer-Crowley CI can be demonstrated graphically. The 95% CI for $\hat{t}_{50}$ is $[10, .)$, $\hat{t}_{75}$ is $[14, .)$, and $\hat{t}_{25}$ is $[10, 23)$.



More on CI Estimate for Quantiles

- An alternative method for calculating the 95% CI of quantiles is discussed in Hosmer et al.

Point Estimate of Mean

- In survival analysis, the sample mean is not so important as quantiles, because censoring survival time data are most often skewed to right.

Point Estimate of Mean

- The estimate of mean is based on the observed range of event times

$$\hat{\mu}(t_J) = \sum_{j=1}^{J} \hat{S}(t_{j-1})(t_j - t_{j-1})$$

where, $J = \#$ of unique time points with events
$\hat{S}(t_0) = 1$ and $t_0 = 0$

Point Estimate of Mean

- The value of $\hat{\mu}(t_J)$ is the area under the step function defined by K-M estimate of $S(t)$.
- When the largest observation is censored, the mean is under-estimated. Why?

Variance Estimate of Sample Mean

- The variance estimate for the mean estimate is:

$$\widehat{Var}\left(\hat{\mu}(t_J)\right) = \frac{n_d}{n_d - 1} \sum_{j=1}^{J-1} \frac{A_j^2 d_j}{n_j(n_j - d_j)}$$

where, $n_d = \sum_{j=1}^{J} d_j$
$A_j = \sum_{k=j+1}^{J} \hat{S}(t_{k-1})(t_k - t_{k-1})$

Mean Estimation

- Revisit the cancer recurrence example:

j	t_j	n_j	d_j	$\hat{S}(t_j)$
0	0	8	0	1
1	10	8	1	0.875
2	13	7	1	0.75
3	14	5	1	0.6
4	17	4	1	0.45
5	23	2	1	0.225

- $\hat{\mu}(23) = 1 \times (10 - 0) + 0.875 \times (13 - 10) + 0.75 \times (14 - 13) + 0.6 \times (17 - 14) + 0.45 \times (23 - 17) = 17.875$

Mean Estimation

- Revisit the cancer recurrence example:

j	t_j	n_j	d_j	$\hat{S}(t_j)$	A_j
0	0	8	0	1	
1	10	8	1	0.875	7.875
2	13	7	1	0.75	5.25
3	14	5	1	0.6	4.5
4	17	4	1	0.45	2.7
5	23	2	1	0.225	

- $\widehat{Var}[\hat{\mu}(23)] = \frac{5}{5-1}\left[\frac{7.875^2\times 1}{8(8-1)} + \frac{5.25^2\times 1}{7(7-1)} + \frac{4.5^2\times 1}{5(5-1)} + \frac{2.7^2\times 1}{4(4-1)}\right] = 4.22959 \rightarrow \sqrt{\widehat{Var}[\hat{\mu}(23)]} = 2.0566$
- 95% CI: $17.875 \pm 1.96 \times 2.0566$

Suggested Readings:

- Chapter 2.3 (Hosmer et al.)