Mid-Term Review

- Homework discussion
- Key concepts for mid-term

Homework 5

1. A study design has been discussed in a study team for treating B-cell lymphoma in the second line patient population. The study will randomize subjects in 1:1 ratio

- After thorough literature search, the study team would like to assume 20 months for the median survival time in the standard of care
- The expected median survival time in the new treatment arm is 28 months
- The enrollment period is 18 months
- The minimum follow-up time for each subject is 24 months

How many events will be needed to reach 90% power at the 1-sided significant level of 0.025?

How many subjects should be planned?

What is the number of subjects if more investigate sites are available and the enrollment period is shortened to 12 months?

What do you think of the power loss if the hazard ratio is 1 during the first 4 months of treatment? What strategies would you like to recommend to the study team?

Please add your own assumption on the rates of loss of follow-up and re-answer the questions above.

2. The study team learned from clinicaltrial.gov that a competitor's trial for the same indication requires only 300 subjects and will take only 3 years to complete. Discuss what you think of the competitor's trial design.

3. Let $T_i \sim \exp(h_i)$, $T_j \sim \exp(h_j)$, and $T_i \perp T_j$. Show $P(T_i \geq T_j) = \frac{h_j}{h_i + h_j}$.

Homework 6

skewed data → log transform.

Biomarker geometric mean → log transformation

ratio → log-transform

1. A trial is designed to evaluate a biomarker reduction post-treatment from baseline. The biomarker is a continuous variable. The reduction in the geometric mean is 33% with 95% CI (15%,47%) reported in literature. The team would like to know the sample size with 30% and 50% risk reduction. The statistician in the team suggests to calculate sample size for 80% and 90% power at the 2-sided significance levels of 0.05 and 0.1. Please show steps for the sample size calculation.

Let

Y_{pi} – post-treatment biomarker measure

Y_{bi} – pre-treatment biomarker measure

$i = 1, \ldots, n$

arithmetic = log (geometric)

$$\frac{1}{n}\sum_{i=1\ to\ n}\{\log Y_{pi} - \log Y_{bi}\} = \log \sqrt[n]{\prod_{i=1\ to\ n}\frac{Y_{pi}}{Y_{bi}}} = \log \sqrt[n]{\frac{Y_{p1}\cdot Y_{p2}\cdot Y_{p3}\cdots Y_{pn}}{Y_{b1}\cdots Y_{bn}}}$$

$$= \log\frac{\bar{Y}_p}{\bar{Y}_b}$$

The reduction $1 - \frac{\bar{Y}_p}{\bar{Y}_b} = 33\%$, $\frac{\bar{Y}_p}{\bar{Y}_b} = 67\%\ with\ 95\%\ CI = (53\%, 85\%)$

Take log transformation for the ratio and CI, $\log \bar{Y}_p - \log \bar{Y}_b = \log 0.67\ \ with\ 95\%\ CI = (\log 0.53\,, \log 0.85)$

Find SD for 1-sample t-test.

(-0.63 -0.16)

Homework 6

$\alpha/2$

2. A randomized, controlled, and three-arm phase II study for Alzheimer's disease is designed using 2 active doses and a placebo control. The randomization ratio is 1:1:1. The primary endpoint will be clinical dementia rating scale sum of boxes (CDR-SB). Two recent phase III studies for Aducanumab showed a mean change from baseline to be 1.74 with standard error of 0.11 in the placebo arm and 548 subjects in Study Emerge 302; and a mean change from baseline of 1.56 with standard error also 0.11 calculated from 545 subjects. Please choose the significance level and power and calculate the sample for each arm and the total sample size. Please explain your rationale of the sample size calculation, including the assumptions used and if multiplicity adjustment should be considered.

$$n = \frac{2(Z_{1-\alpha} - Z_{\beta})^2 \sigma^2}{\delta^2}$$

$\delta_1 = 1.74, \delta_2 = 1.56,$

Homework 6

4. Obtain the PBC dataset from R survival package. Fit the following PH model

$$h(t,Z) = h_0(t)\exp(\beta_1 sex + \beta_2 edema + \beta_3 bili + \beta_4 albumin + \beta_5 copper + \beta_6 stage) \quad (1)$$

   a) Identify the information matrix
   b) Construct likelihood ratio test for hypothesis $H_0: \beta_4 = \beta_5 = \beta_6 = 0$ (show steps)
   c) Compare the p-values for sex between Model (1) and the log-rank test
   d) Suppose that the team plans to publish a paper about the study result. You need to describe the statistical methods in the method section of the paper. Please write all needed analyses in the method section.
   e) You are also responsible for the result section of the potential publication. Please describe the results, including interpretation of the hazard ratios, p-values, 2-sided 95% CIs of each covariate.

5. Derive PH model score test is the same as the log-rank test for an indicator covariate when no ties.

6. The observed survival data $(T_i, \Delta_i, Z_i)\ \ i = 1,2,3,4,5,6$ are

   (16,1,1), (20, 0,1), (12, 1,0), (14, 0,0), (11, 1, 0), (9, 1, 1).

   Please construct the partial likelihood.

Survival and Hazard Functions

Hazard Function and the Survival Distribution

$$h(t_j) = P(T = t_j | T \geq t_j)$$

$$= \frac{P(T = t_j \cap T \geq t_j)}{P(T \geq t_j)}$$

$$= \frac{P(T = t_j)}{P(T \geq t_j)} = \frac{P(T = t_j)}{P(T > t_{j-1})}$$

$$= \frac{f(t_j)}{S(t_{j-1})}$$

Discrete Survival Time

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

Since

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

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

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

Continuous Survival Time

- The density function

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

Continuous Survival Time

- The survival function

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= \int_t^{\infty} f(x)dx \end{aligned}$$

$$f(x) = -\frac{dS(t)}{dt}$$

Hazard Function Definition

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, T \ge t)}{\Delta t P(T \ge t)}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, T \ge t)}{\Delta t} \cdot \frac{1}{P(T \ge t)}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} \cdot \frac{1}{P(T \ge t)}$$

$$= \frac{f(t)}{S(t^-)} \qquad \text{where} S(t^-) = \lim_{t \to t^-} S(t^-)$$

- The limit $\lim_{\Delta t \to 0}$ indicates instantaneous risk after t.
- $P(t \le T < t + \Delta t | T \ge t)$: survival probability in the interval of $(t, t + \Delta t)$ given that the person has survived up to t.
- $h(t)$ >0

Cumulative Hazard Function

$$H(t) = \int_0^t h(x)dx$$

$$= \int_0^t -\frac{dlogS(x)}{dx}dx$$

$$= -logS(t)$$

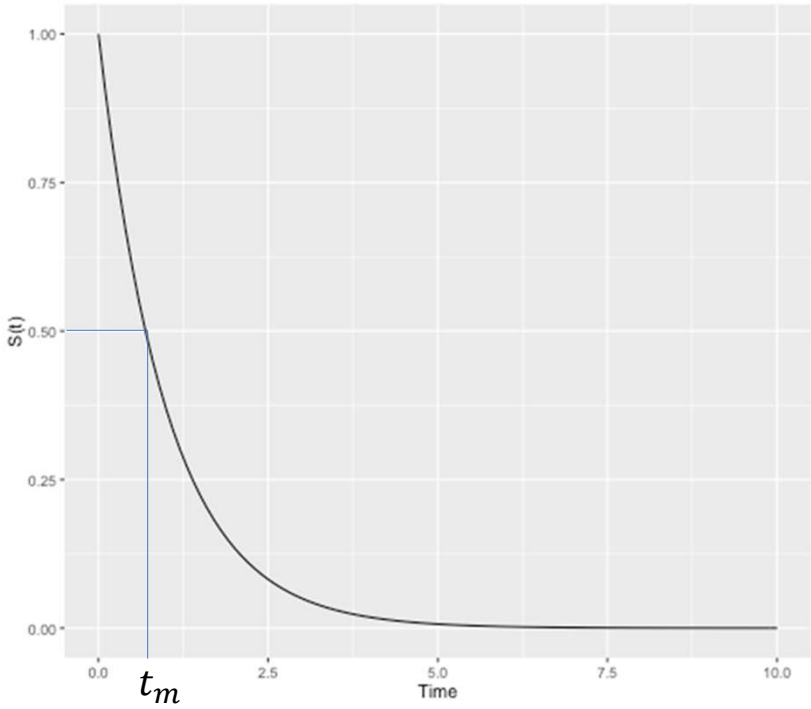
$$S(t) = e^{-H(t)}$$

Survival Quantiles

- Median Survival time t_m

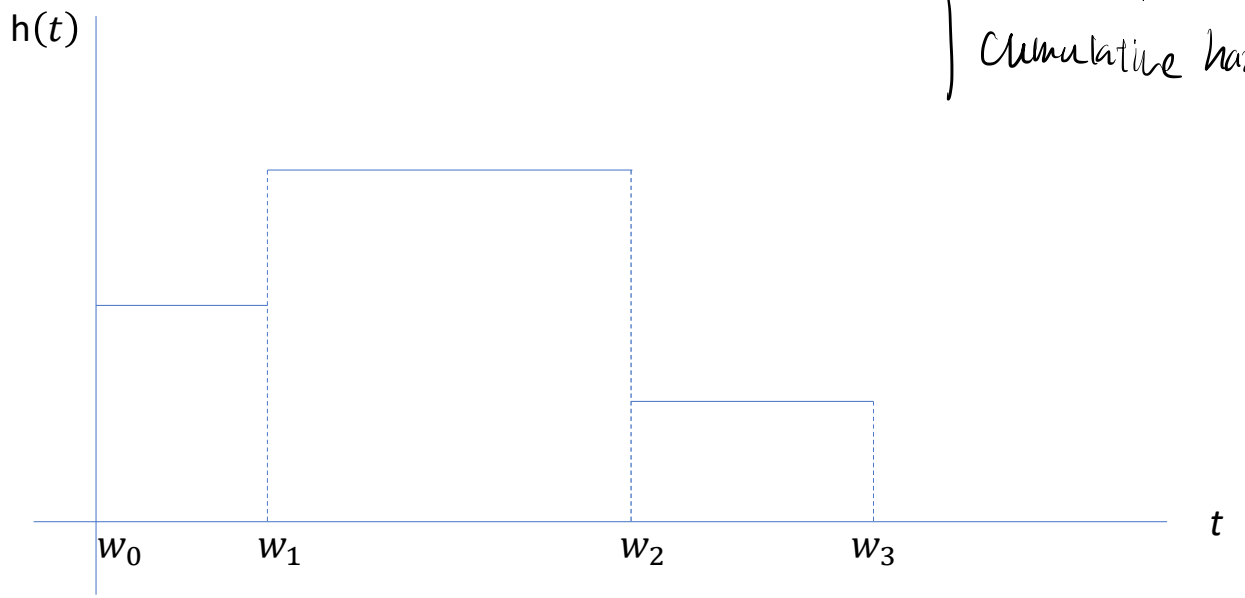
- $S(t_m) = P(T > t_m)$=0.5
- $t_m = \inf\{t: S(t) \leq 0.5\}$

- The p^{th} quantile
- $S(t_p) = P(T > t_p) = p$
- $t_p = \inf\{t: S(t) \leq p\}$



Piecewise Hazard Function

Piecewise exponential
{ hazard function?
  cumulative hazard function?



Piecewise Exponential –The Cumulative Hazard Function

- The hazard function

$$\mathrm{h}(t) = h_1 I(t \le w_1) + h_2 I(w_1 < t \le w_2) + \cdots$$

where $w_1, w_2, \ldots,$ are fixed time intervals, $w_1 = 0$

- At time $t \in (w_j, w_{j+1})$, the cumulative hazard function can be written as

$$H(t) = \sum_{i<j} h_i(w_i - w_{i-1}) + h_j\big(t - w_j\big)\mathrm{I}(t \in (w_j, w_{j+1}))$$

given piecewise hazard ⇒ derive cumulative / survival / PDF

Piecewise Exponential – The Survival Function

- Recall $S(t) = e^{-H(t)}$
- Therefore, for $t \in (w_j, w_{j+1})$

$$S(t) = e^{-\{\sum_{i<j} h_i(w_i - w_{i-1}) + h_j(t - w_j)\}}$$

$$= \prod_{i<j} e^{-h_i(w_i - w_{i-1})} \, e^{-h_j(t - w_j)}$$

Piecewise Exponential - PDF

- Recall $f(t) = -\frac{dS(t)}{dt}$
- Therefore, for $t \in (w_j, w_{j+1})$

$$f(t) = -\frac{d}{dt} e^{-\{\sum_{i<j} h_i(w_i - w_{i-1}) + h_j(t - w_j)\}}$$

$$= h_j \prod_{i<j} e^{-h_i(w_i - w_{i-1})} e^{-h_j(t - w_j)}$$

Life-Table

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

  Which side of survival { beginning / end.

  - Survival function at the end of the i^{th} interval (in SAS, always the begining)

$$\hat{S}_L(t_0) = 1$$

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

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

Life-table Estimate – Hazard Function

- Number of events per person-time-units

events / # of patient-time

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

- 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

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

K-M Estimator

Kaplan-Meier Estimator

- Also known as product-limit estimator
- Observed
  - (T_i, δ_i) in n subjects, $i = 1,2, \dots, n$
  - 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

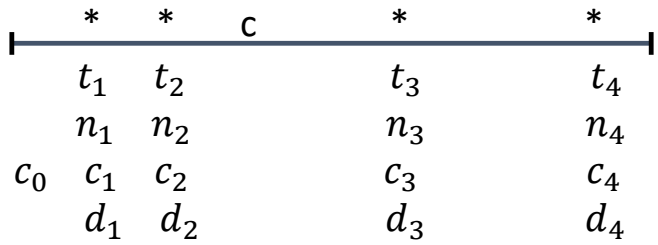
- Let

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

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

- 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

- At the distinct event time t_i

- The probability of surviving beyond t_i is

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

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

K-M Example

Summary Statistics for Time Variable Years

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

fill in the K-M table

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 Estimator Median and Quantiles

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

$$S(t_p) = P(T > t_p) = p$$

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

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

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

Recall the example of the death events from the colon cancer

```
## 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
```

K-M Estimator Median and Quantiles

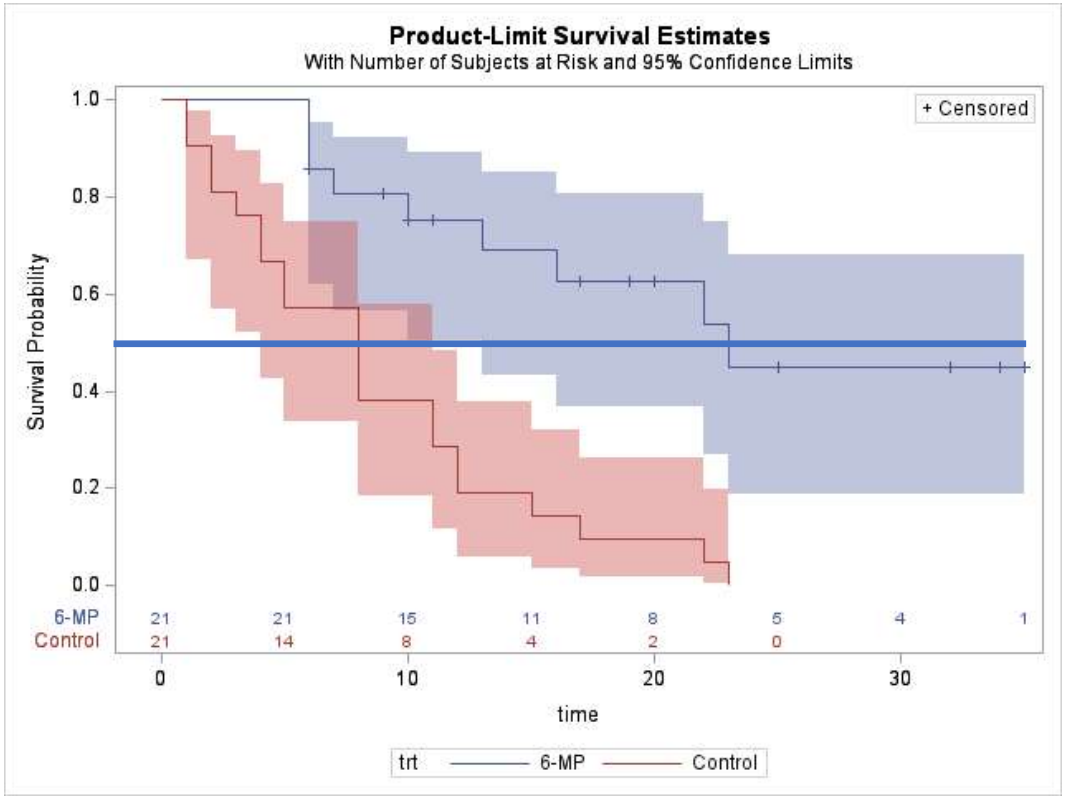
Recall the example of the death events from the colon cancer

```
##    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

Leukemia Data



Leukemia Data

```
ods graphics on;
proc lifetest data=example method=KM
plots=survival (cl atrisk=0 to 35 by 5)
conftype=loglog outsurv=survival;
  time time*event(0);
  strata trt;
  run;
ods graphics off;
```

Obs	trt	time	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL	STRATUM
1	6-MP	0	.	1.00000	1.00000	1.00000	1
2	6-MP	6	0	0.85714	0.61972	0.95155	1
3	6-MP	6	1	0.85714	.	.	1
4	6-MP	7	0	0.80672	0.56315	0.92281	1
5	6-MP	9	1	0.80672	.	.	1
6	6-MP	10	0	0.75294	0.50320	0.88936	1
7	6-MP	10	1	0.75294	.	.	1
8	6-MP	11	1	0.75294	.	.	1
9	6-MP	13	0	0.69020	0.43161	0.84907	1
10	6-MP	16	0	0.62745	0.36751	0.80491	1
11	6-MP	17	1	0.62745	.	.	1
12	6-MP	19	1	0.62745	.	.	1
13	6-MP	20	1	0.62745	.	.	1
14	6-MP	22	0	0.53782	0.26778	0.74679	1
15	6-MP	23	0	0.44818	0.18805	0.68014	1
16	6-MP	25	1	.	.	.	1
17	6-MP	32	1	.	.	.	1
18	6-MP	32	1	.	.	.	1
19	6-MP	34	1	.	.	.	1
20	6-MP	35	1	.	.	.	1
21	Control	0	.	1.00000	1.00000	1.00000	2
22	Control	1	0	0.90476	0.67005	0.97529	2
23	Control	2	0	0.80952	0.56891	0.92389	2
24	Control	3	0	0.76190	0.51939	0.89326	2
25	Control	4	0	0.66667	0.42535	0.82504	2
26	Control	5	0	0.57143	0.33798	0.74924	2
27	Control	8	0	0.38095	0.18307	0.57779	2
28	Control	11	0	0.28571	0.11656	0.48182	2
29	Control	12	0	0.19048	0.05948	0.37743	2
30	Control	15	0	0.14286	0.03566	0.32116	2
31	Control	17	0	0.09524	0.01626	0.26125	2
32	Control	22	0	0.04762	0.00332	0.19704	2
33	Control	23	0	0.00000	.	.	2

Logrank

The Log-rank Test

- Construct k 2x2 tables at each distinct event time

- At i^{th} event time,
  - n_{0i} number of subjects at risk at $t_{(i)}^{-}$ in Group 0
  - n_{1i} number of subjects at risk at $t_{(i)}^{-}$in Group 1
  - $n_i = n_{0i} + n_{1i}$ total number of subjects at risk at $t_{(i)}^{-}$

  - d_{0i} number of subjects at risk at $t_{(i)}$ in Group 0
  - d_{1i} number of subjects at risk at $t_{(i)}$ in Group 1
  - $d_i = d_{0i} + d_{1i}$ total number of events at $t_{(i)}$

2x2 Table at the i^{th} Event Time

Be able to construct the 2x2 table for log-rank tests.

Group	Events occurred at $t_{(i)}$	Number of subjects Survival at $t_{(i)}^{+}$	Number of subject at risk at $t_{(i)}^{-}$
0	d_{0i}	$n_{0i} - d_{0i}$	n_{0i}
1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
Total	d_i	$n_i - d_i$	n_i

- n_{0i}, n_{1i}, n_i are fixed at $t_{(i)}^{-}$
- d_i is fixed at $t_{(i)}$
- $d_{0i} \sim$ hypergeometric distribution

The Log-rank Test

- The total deviation from null

$$L = \sum_{i=1}^{k} (d_{0i} - e_{0i})$$

- Variance

$$\text{var}(L) = var\left(\sum_{i=1}^{k} (d_{0i} - e_{0i})\right) \approx \sum_{i=1}^{k} var(d_{0i})$$

From hypergeometric distribution

$$var(d_{0i}) = \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$\text{var}(L) = \sum_{i=1}^{k} \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$\frac{L}{\sqrt{var(L)}} \sim N(0,1)$$

The Log-rank Test

- If there is no tie at each event time, log-rank test statistics reduces to

$$\frac{\sum_{i=1}^{k}(d_{0i}-n_{0i}/n_i)}{\sum_{i=1}^{k} n_{0i}n_{1i}/{n_i}^2}$$

where d_{0i} can be 0 or 1

General Class of Weighted Log-rank Tests

- Can be written as

$$L_w = \sum_{i=1}^{k} \omega_i (d_{0i} - e_{0i})$$

$$\text{var}(L_w) = \sum_{i=1}^{k} \omega_i^2 \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

Don't need to remember

Test	Weight ω_i
Log-rank	$\omega_i = 1$
Gehan's Wilcoxon	$\omega_i = n_i$
Peto/Prentice	$\omega_i = S(t_i)$
Fleming-Harrington	$\omega_i = S(t_{i-1})^{\rho}(1 - S(t_{i-1}))^{q} \quad \rho, q \geq 0$
Tarone-Ware	$\omega_i = \sqrt{n_i}$

When to use log-rank Wilcoxon

Comparisons

- All tests will control type I error under null
- May obtain optimal power under certain alternatives
- The Log-rank test
  - Tends to be sensitive to distributional differences which are separated late in time
  - Has the optimal power when the hazard ratio is constant
- The Wilcoxon test
  - Tends to be more powerful in detecting early separation between survival functions

Reference: A pretest for choosing between logrank and wilcoxon tests in the two-sample problem. International Journal of Statistics, 2010, vol. LXVIII, n. 2, pp. 111-125

The Cox Regression Model

The Cox Regression Model

- Survival data (T, Δ, Z), the hazard function can be written as the following based on the Cox model

$$h(t|Z = z) = h_0(t)e^{\beta' z}$$

  - where $h_0(t)$ is a baseline hazard function,
    Z can a vector of p covariates,
    β is a vector of p coefficients

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Understanding the Coefficient

Let Z be a univariate covariate and take values of 0,1

$$h(t|Z=1) = h_0(t)e^{\beta} \qquad h(t|Z=0) = h_0(t)$$

$$\frac{h(t|Z=1)}{h(t|Z=0)} = e^{\beta}$$

$$\beta = \log\frac{h(t|Z=1)}{h(t|Z=0)}$$

Interpretation

- β is the log hazard ratio for unit increase
- e^{β} is the hazard ratio
- $1 - e^{\beta}$ represents risk reduction if $Z = 0$ represents the control arm.

Partial Likelihood

- The partial likelihood can be written as

don't need to remember when there are ties.

$$L_P(\beta) = \prod_{j=1}^{J} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_l}}$$

$$= \prod_{i=1}^{n} \left\{ \frac{e^{\beta' z_i}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right\}^{\Delta_i}$$

Example – Construct Partial Likelihood

Survival data

- Group 0: 6+,7,9+,10,11+,13,16+,17+, 20+ $n_0 = 9$
- Group 1: 4, 5+, 8+,11+,12,15,17+,22+,23+ $n_1 = 9$
- $Z = 0,1$
- 7 Distinct event time: 4, 7, 10, 12, 13, 15,

$$L_P(\beta) = \frac{e^{\beta}}{9e^{\beta}+9} \cdot \frac{1}{7e^{\beta}+7} \cdot \frac{1}{6e^{\beta}+6} \cdot \frac{e^{\beta}}{5e^{\beta}+4} \cdot \frac{1}{4e^{\beta}+4} \cdot \frac{e^{\beta}}{4e^{\beta}+3}$$

Risk sets

Group 0	6+,7,9+, 10,11+,13, 16+,17+, 20+	7,9+,10, 11+,13, 16+,17+, 20+	10, 11+, 13,16+, 17+, 20+	13,16+, 17+, 20+	13,16+, 17+, 20+	16+, 17+, 20+
Group 1	4, 5+, 8+,11+, 12,15,17+, 22+,23+	8+,11+, 12,15,17+, 22+,23+	11+,12, 15,17+, 22+,23+	12,15, 17+,22+, 23+	15,17+, 22+,23+	15,17+,22+, 23+
Event Time	4	7	10	12	13	15

MLE

- MLE $\hat{\beta}$ can be obtained from the score function

$$U(\hat{\beta}) = 0 \quad \text{Solve } \hat{\beta}$$

$$Var(\hat{\beta}) = I(\hat{\beta})^{-1} \rightarrow \text{Construct Wald test.}$$

Inference

- Estimation and confidence intervals
- Hypothesis testing,
  - Hazard ratio <1 represents risk reduction
  - $\beta < 0$ mean risk reduction

$$H_0: \beta \geq 0 \quad vs \quad H_A: \beta < 0$$

- Association
- Prediction

Confidence Intervals

- 95% CI confidence interval for the coefficient estimates

$$\hat{\beta} \pm z_{0.975} se(\hat{\beta}) = \hat{\beta} \pm 1.96 se(\hat{\beta})$$

$\hat{\beta}$ is a unit log hazard ratio, to covert to unit hazard ratio, lower and upper bound of the 95% CI for the unit hazard ratio is

$$\left[e^{\hat{\beta} - 1.96 se(\hat{\beta})}, e^{\hat{\beta} + 1.96 se(\hat{\beta})} \right]$$

Ward Test

- For components of $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$,

- A Wald test for one covariate is

$$Z_i = \frac{\left(\widehat{\beta}_i - \beta_i\right)}{\sqrt{I\left(\widehat{\beta}_i\right)^{-1}}} \sim N(0,1)$$

$$for\ i = 1, \dots, p$$

A composite test can be written as

$$\chi_p^2 = \hat{\beta}' I\left(\hat{\beta}\right)^{-1} \hat{\beta}$$

Score Test

Under null $\approx$ log-rank test

Under null H_0: $\beta = 0$, score test

$$U(0)/\sqrt{I(0)} \sim N(0,1)$$

$$U(0) = \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l e^{\beta' z_l}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right\}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l}{\sum_{l \in R(t_i)} 1} \right\}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l}{n_i} \right\}$$

Likelihood Ratio Test

- Let full covariates be

$$Z^{K\prime} = (Z_1, Z_2, \dots, Z_k, Z_{k+1}, \dots, Z_K)$$

- Subset of covariates for $k < K$

$$Z^{k\prime} = (Z_1, Z_2, \dots, Z_k)$$

- The full model is

$$L_P(\beta^K | Z^K) = \prod_{i=1}^{n} \left\{ \frac{exp(\beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_K Z_{iK})}{\sum_{l \in R(t_i)} exp(\beta_1 Z_{l1} + \beta_2 Z_{l2} + \cdots + \beta_K Z_{lK})} \right\}^{\Delta_i}$$

- The sub-model is

$$L_P\left(\beta^k | Z^k\right) = \prod_{i=1}^{n} \left\{ \frac{exp(\beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_k Z_{ik})}{\sum_{l \in R(t_i)} exp(\beta_1 Z_{l1} + \beta_2 Z_{l2} + \cdots + \beta_K Z_{lk})} \right\}^{\Delta_i}$$

Likelihood Ratio Test

- To test

$$H_0: \beta_{k+1} = \beta_{k+2} = \cdots = \beta_K = 0$$

- The likelihood Ratio test is

$$\Lambda = \frac{L_P\left(\hat{\beta}^K | Z^K\right)}{L_P\left(\hat{\beta}^k | Z^k\right)}$$

$$-2\log\Lambda = -2\left\{\log L_P\left(\hat{\beta}^K | Z^K\right) - \log L_P\left(\hat{\beta}^k | Z^k\right)\right\} \sim \chi^2_{(K-k)}$$

Example - Melanoma

Interpretation:

Overall tests – reject the null that there is no difference in age, sex, and tumor thickness

- Likelihood Ratio
- Score
- Wald

Wald tests shows

- The risk of death increased 2% with 1 year old
- The risk in male is 67% higher than that of female
- The risk increased 15% with 1 unit increase in tumor thickness

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	700.985	666.615
AIC	700.985	672.615
SBC	700.985	679.403

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.3703	3	<.0001
Score	41.8566	3	<.0001
Wald	38.2646	3	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	0.02221	0.00795	7.8071	0.0052	1.022	age
sex	1	0.51242	0.23877	4.6056	0.0319	1.669	sex
thickness	1	0.13499	0.03048	19.6188	<.0001	1.145	thickness

construct confidence interval for β
↓
then for HR.