Lecture 22: The log rank test.

- If don't use censored data & we want to compare survival times of treatment A vs. those of treatment B, what would you do?

Treatment A	Treatment B
10	20
	5
100	50
150	209
200	110
310	⋮
⋮	90+
110+	

However, data is now censored

- The main issue is that at any given point in time, our number of subjects can change. Either due to a death or due to a loss of tracking.

Suppose we have

Month	Treatment A At risk	Treatment A Died	Treatment B At risk	Treatment B Died
1	51	1	45	0
2	50	2	45	1
3	48	5	44	1
4	42	2	43	5
5	40	8	38	5
$\boxed{6}$	$\boxed{32}$	$\boxed{7}$	$\boxed{33}$	$\boxed{4}$

	Died	Survived	
TA	$y = 7$, $E(y) = 5.41$	25	$n_A = 32$
TB	4	29	$n_B = 33$
	$n_d = 11$	$n_s = 54$	$n = 65$

$$V = \frac{n_A n_B n_d n_s}{[n^2(n-1)]}$$

$$V = 2.31$$

Hyp. Test: Null hypothesis is that the hazard rate for month 6 are the same for Treat. A and Treat. B

$$H_0(6) : h_{A6} = h_{B6}$$

- we got n_A marbles labeled "A".
- n_B "B"
- we draw n_d marbles

⇒ what is the prob. distribution for the # of marbles labeled "A"?

• It is a hypergeometric distribution!

$$P(y \mid n_A, n_B, n_d) = \binom{n_A}{y} \cdot \binom{n_B}{n_d - y} \Big/ \binom{n}{n_d}$$

$$\Rightarrow E(y) = \frac{n_A \cdot n_d}{n} \quad \left(\text{Just the mean of a hypergeometric distr.}\right)$$

$$V(y) = \frac{n_A \cdot n_B \cdot n_d \cdot n_s}{[n^2(n-1)]}$$

Repeat this for all the $N = 47$ months, calculate y_i, E_i, V_i for month i and compute the log-rank statistic, which is defined as:

$$Z = \frac{\sum_{i=1}^{N} (y_i - E_i)}{\left(\sum_{i=1}^{N} V_i\right)^{1/2}}$$

& compare to normal critical values.

[illegible] Cancer Patients

· For our dataset ~ ___________

$$\sum_{i=1}^{N} y_i = 42 \, , \quad \sum_{i=1}^{N} E_i = 32.9$$

$$\sum_{i=1}^{N} r_i = 16$$

$$\Rightarrow \quad z = 2.27 \quad \ddot{o}!$$

---

① Survival rates using parametric estimation

② Proportional hazards model

③ Missing Data & the EM Algorithm.

, Land used $= x_1$

• # of labor hours $= x_2$

- Amount of fertilizer by Ha $= x_3$

• yield $= y$

$\Rightarrow$ didn't know how much rainfall ^.

$$y_i = \pi_R \cdot f_R(x_1, x_2, x_3) + \pi_N \cdot f_N(x_1, x_2, x_3)$$

$+ \pi_D \cdot f_D(x_1, x_2, x_3)$

repeat until convergence

Step 1 : Simulate Posterior β's for each of f_R, f_N, f_D

Step 2 : Simulate belonging probabilities for each farmer to each rain condition.

Step 3 : Simulate Posterior π_R, π_N, π_D