Survival Analysis: Homework 1

Xuange Liang, xl3493

1. Exponential Density and Survival-related Functions

a. λ estimation and interpretation

i. $\hat{\lambda}_{repapse} = \frac{number\ of\ events}{total\ number\ of\ time\ units\ ovserved\ on\ all\ individuals} = \frac{6}{186} \approx 0.032$

Interpretation: The parameter λ represents the instantaneous hazard rate (risk) of relapse per month. A value of 0.032 per month indicates that at any given time, a patient has approximately a 3.2% risk of relapse per month, assuming the exponential distribution holds.

ii. $\hat{\lambda}_{death} = \frac{number\ of\ events}{total\ number\ of\ time\ units\ ovserved\ on\ all\ individuals} = \frac{3}{223} \approx 0.013$

Interpretation: The parameter λ represents the instantaneous hazard rate (risk) of death per month. A value of 0.013 per month indicates that at any given time, a patient has approximately a 1.3% risk of death per month, assuming the exponential distribution holds.

b. Using this parameter to estimate quantities

i. Mean time to relapse: $\frac{1}{\hat{\lambda}_{repapse}} = \frac{1}{0.032} = 31.25$ months

Mean time to death: $\frac{1}{\hat{\lambda}_{death}} = \frac{1}{0.013} = 76.92$ months

ii. Median time to relapse: $\frac{1}{\hat{\lambda}_{repapse}} = \frac{\log(2)}{0.032} = 21.66$ months

Mean time to death: $\frac{1}{\hat{\lambda}_{death}} = \frac{\log(2)}{0.013} = 53.32$ months

iii. $S_R(12) = e^{-\hat{\lambda}_{repapse} \times 12} = 0.6811,\ S_R(24) = e^{-\hat{\lambda}_{repapse} \times 24} = 0.4639$

$S_D(12) = e^{-\hat{\lambda}_{death} \times 12} = 0.8556,\ S_R(12) = e^{-\hat{\lambda}_{death} \times 24} = 0.732$

iv. $F_R(12) = 1 - S_R(12) = 0.3189,\ F_R(24) = 1 - S_R(24) = 0.5361$
$F_D(12) = 1 - S_D(12) = 0.1444,\ F_R(12) = 1 - S_D(24) = 0.268$

v. $P(T > 24 | T > 12) = S_R(12) = 0.6811$, which is the same with $S_R(12)$ in iii, demonstrating the memoryless property of the exponential distribution.

c. Median time of relapse: 27
Median time of death: Not estimable

2. Kaplan-Meier Survival Estimate
    a. Calculate the Kaplan-Meier estimate of the survival function by hand

    Method:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

    Result:

Time	At Risk (n_i)	Events (d_i)	Censored	$1 - \frac{d_i}{n_i}$	$\hat{S}(t)$
0	17	0	0	1.0000	1.0000
2	17	1	0	16/17 = 0.9412	0.9412
3	16	1	0	15/16 = 0.9375	0.9412 × 0.9375 = 0.8824
4	15	1	0	14/15 = 0.9333	0.8824 × 0.9333 = 0.8235
12	14	1	0	13/14 = 0.9286	0.8235 × 0.9286 = 0.7647
22	13	1	0	12/13 = 0.9231	0.7647 × 0.9231 = 0.7059
48	12	1	0	11/12 = 0.9167	0.7059 × 0.9167 = 0.6471
51†	11	0	1	1.0000	0.6471
56†	10	0	1	1.0000	0.6471
80	9	2	0	7/9 = 0.7778	0.6471 × 0.7778 = 0.5033
90	7	1	0	6/7 = 0.8571	0.5033 × 0.8571 = 0.4314
94†	6	0	1	1.0000	0.4314
160	5	1	0	4/5 = 0.8000	0.4314 × 0.8000 = 0.3451
161	4	1	0	3/4 = 0.7500	0.3451 × 0.7500 = 0.2588
180	3	1	0	2/3 = 0.6667	0.2588 × 0.6667 = 0.1725
180†	2	0	1	1.0000	0.1725

    b. Calculate the Kaplan-Meier estimate of the survival function by R

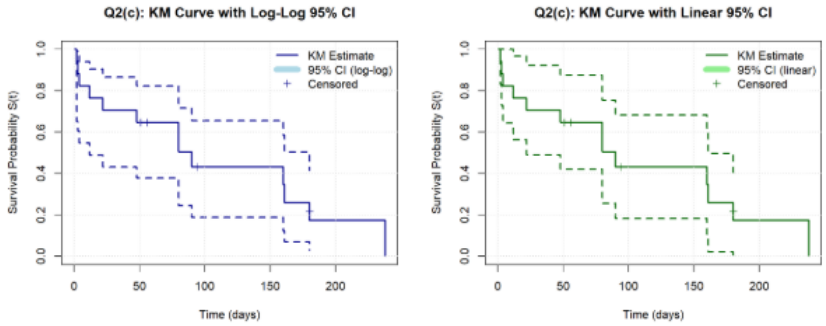
```r
# Create survival object
surv_q2 <- Surv(time = q2_data$Value, event = q2_data$Binary)

# Fit Kaplan-Meier estimator
km_q2 <- survfit(surv_q2 ~ 1, data = q2_data)

# Display detailed KM table
km_summary <- summary(km_q2)
```

    Log-log & Linear:

Time	At Risk	Events	S(t)	Log-Log Lower	Log-Log Upper	Linear Lower	Linear Upper
2	17	1	0.9412	0.6502	0.9915	0.8293	1.0000
3	16	1	0.8824	0.6060	0.9692	0.7292	1.0000
4	15	1	0.8235	0.5471	0.9394	0.6423	1.0000
12	14	1	0.7647	0.4883	0.9045	0.5631	0.9663
22	13	1	0.7059	0.4315	0.8656	0.4893	0.9225
48	12	1	0.6471	0.3771	0.8234	0.4199	0.8742
80	9	2	0.5033	0.2436	0.7162	0.2541	0.7525
90	7	1	0.4314	0.1870	0.6560	0.1811	0.6817
160	5	1	0.3451	0.1216	0.5844	0.0942	0.5960
161	4	1	0.2588	0.0691	0.5048	0.0204	0.4973
180	3	1	0.1725	0.0296	0.4159	0.0000	0.3831
238	1	1	0.0000	NA	NA	NaN	NaN

The log-log transformation maintains bounds within [0,1] because it works on the log-log scale and back-transforms, ensuring valid probabilities. The linear (plain) approach may produce invalid bounds, especially when S(t) is close to 0 or 1.

c. KM Plot using R



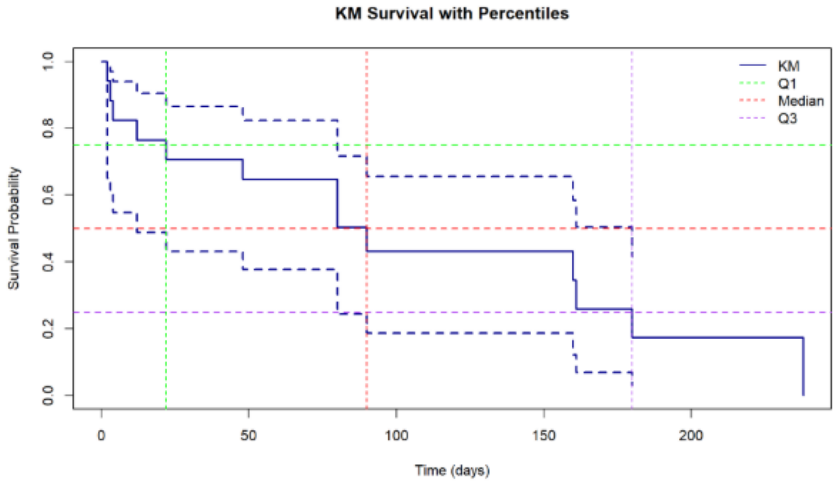
d. Provide the estimated median survival, the estimated 25th and 75th per-centiles

25th (Q1): 22 day

50th (Median) 90 day

75th (Q3) 180 day

The KM survival estimated corresponding is the same with $(1 - percentage)$



e. Cumulative Hazard from KM

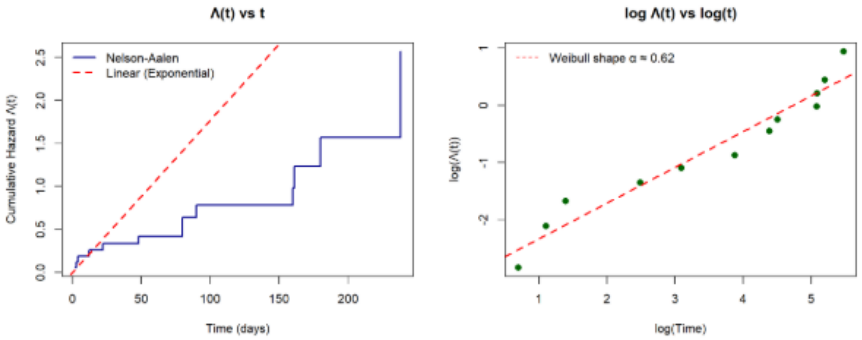
Time	S(t)	Λ(t) = -log(S(t))
2	0.9412	0.0606
3	0.8824	0.1252
4	0.8235	0.1942
12	0.7647	0.2683
22	0.7059	0.3483
48	0.6471	0.4353
80	0.5033	0.6866
90	0.4314	0.8408
160	0.3451	1.0639
161	0.2588	1.3516
180	0.1725	1.7571
238	0.0000	Inf

f. Nelson-Aalen Estimator

Where $\hat{\Lambda}_{NA}(t) = \sum \frac{d_j}{r_j}$:

Time	n_risk	events	Λ_NA(t)	Λ_KM(t)
2	17	1	0.0588	0.0606
3	16	1	0.1213	0.1252
4	15	1	0.1880	0.1942
12	14	1	0.2594	0.2683
22	13	1	0.3363	0.3483
48	12	1	0.4197	0.4353
80	9	2	0.6419	0.6866
90	7	1	0.7848	0.8408
160	5	1	0.9848	1.0639
161	4	1	1.2348	1.3516
180	3	1	1.5681	1.7571
238	1	1	2.5681	Inf

g. Plot different cumulative hazards and it's log



h. Fleming-Harrington estimator

$$S_{FH}(t) = e^{-\Lambda_{NA}(t)}$$

Time	Ŝ_FH(t)	Ŝ_KM(t)	Diff
2	0.9429	0.9412	0.0017
3	0.8857	0.8824	0.0034
4	0.8286	0.8235	0.0051
12	0.7715	0.7647	0.0068
22	0.7144	0.7059	0.0085
48	0.6573	0.6471	0.0102
80	0.5263	0.5033	0.0230
90	0.4562	0.4314	0.0249
160	0.3735	0.3451	0.0284
161	0.2909	0.2588	0.0321
180	0.2084	0.1725	0.0359
238	0.0767	0.0000	0.0767

Comment: The two estimators show excellent agreement (max difference ≈ 0.0767), which is expected for well-behaved survival data.

3. Life (Actuarial) Survival Estimate

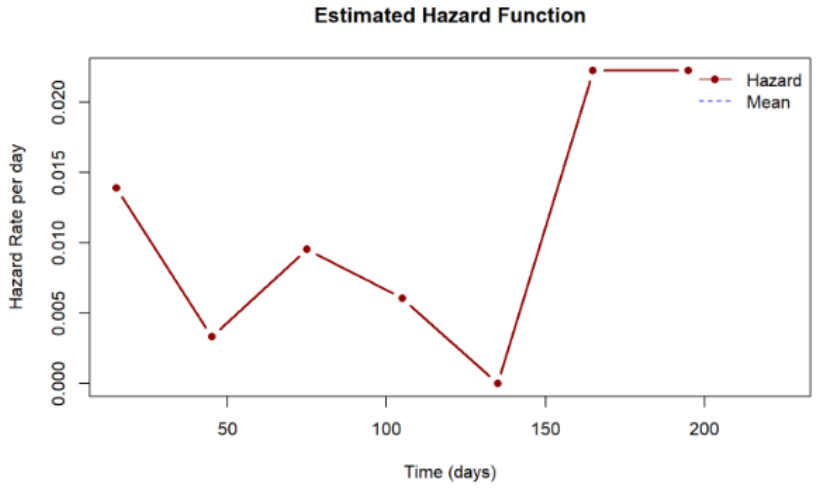
a. Calculate the actuarial estimate of the survival function

Where $\hat{S}(t_j) = \prod_{i \le j} \left(1 - \frac{d_i}{r_i'}\right)$

Interval	n at start	Events (d)	Censored (c)	n' = n - c/2	q_j	p_j = 1-q_j	S(t)
0-30	17	5	0	17.0	0.2941	0.7059	0.7059
30-60	12	1	2	11.0	0.0909	0.9091	0.6417
60-90	9	2	0	9.0	0.2222	0.7778	0.4991
90-120	7	1	1	6.5	0.1538	0.8462	0.4223
120-150	5	0	0	5.0	0.0000	1.0000	0.4223
150-180	5	2	0	5.0	0.4000	0.6000	0.2534
180-210	3	1	1	2.5	0.4000	0.6000	0.1520
210-240	1	1	0	1.0	1.0000	0.0000	0.0000

b. Calculate the estimated hazard function at the midpoint of each time interval and plot.

Interval	Midpoint (days)	h(t) per day
0-30	15	0.013888
30-60	45	0.003333
60-90	75	0.009523
90-120	105	0.006058
120-150	135	0.000000
150-180	165	0.022222
180-210	195	0.022222
210-240	225	Inf



c. According to the plot, the hazard varies substantially, therefore, the exponential model is not appropriate because the exponential distribution assumption is not held.

4. Non-Informative and Informative Censoring

T_1 is defined as the time from being placed on the waiting list to receiving a liver transplant. Censoring happens when patients die during the waiting period, withdraw from the waiting list, or have not received a transplant by the end of the study. Non-informative censoring

means that the occurrence of censoring is independent of survival time and does not affect statistical estimation of survival outcomes. In this case, if patients withdraw from the list for non-health-related reasons (such as relocation or choosing not to transplant), censoring may be non-informative. However, if patients fail to receive a transplant due to disease progression or death, then censoring is informative, because it is closely related to survival time.

Censoring mechanism for T_1: the censoring mechanism is likely informative, because patients may be unable to receive transplantation due to changes in their medical condition.

Censoring mechanism for "time until at least one HLA-mismatched organ becomes available": the censoring mechanism is also likely informative, because organ availability and patient health status may be correlated.