## Week 1 – Life Tables

The Kaplan-Meier method is used to calculate tables of the survival probability. When we plot the results, we end up with a stepped survival curve. Complete the table below using the Kaplan-Meier method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Time (t) in days** | **Number of patients known to be alive at time t** | **Number of patients who died at time t** | **Proportion of patients surviving past time t** | **Probability of survival *past* time t** |
| 0 (study start) | 8 | 0 | 1 | 1 |
| 1 | 8 | 2 | 0.75 | 0.75 |
| 4 | 6 | 1 | 0.83 | 0.623 |
| 5 | 5 | 1 | 0.8 | 0.498 |
| 6+ | 4 | 0 | (4-0)/4 = 1 | 0.498 \* 1 = 0.498 |
| 9 | 3 | 1 | (3-1)/3 =0.667 | 0.498 \* 0.667 = 0.332 |
| 9+ | 2 | 0 | (2-0)/2 = 1 | 0.332 \* 1 = 0.332 |
| 22 | 1 | 1 | (1-1)/1 = 0 | 0.332 \* 0 = 0 |

## Week 2 – Harzard Function and Ratio

1. In survival analysis, a hazard is:
   1. The probability of surviving at time t having survived up to time t.
2. The risk set comprises:
   1. The set of patients at time t that are at risk of experiencing the event.

## Final Assessment

1. In a birth cohort study, people are enrolled into the study at birth and followed up over time to see who gets the outcome of interest, e.g. some disease. Their age in days can be used as the time variable in a time-to-event analysis.
   1. True.
2. In survival analysis, censoring can occur if individuals drop out of the study and we don’t know whether they have the event of interest, e.g. death. They are handled as neither alive nor dead and are deducted from the number of patients alive.
   1. True.
3. With a Kaplan-Meier table and plot where death is the outcome of interest, if ten patients are alive on day 20 and then three die and one is censored, all on day 21, what is the proportion of patients under observation who are alive at the start of day 22?
   1. 6/6 because in the calculation, the patients who died are handled first, and then we remove the censored patients from those remaining alive.
4. What does the following R code produce?

km\_fit <- survfit(Surv(fu\_time, death) ~ 1)

plot(km\_fit)

* 1. Kaplan-Meier estimates of the probability of survival over time for all patients in the data set for all time points until the last patient’s “fu\_time”.

1. You run the following R code:

survdiff(Surv(fu\_time, death) ~ gender, rho = 0)

After the table of observed and expected counts, it gives you the following output:

Chisq = 2.9 on 1 degrees of freedom, p = 0.085

* 1. The null hypothesis tested here is that, at every given time point, each gender has the same hazard of death.
  2. By convention, we conclude there is insufficient evidence against the null hypothesis.

1. You run a Cox regression model on the time to relapse for cancer, with treatment as a binary predictor. The estimated hazard ratio for treatment A related to treatment B is 1.33. What can you conclude from this study from just these facts?
   1. The hazard for treatment A is estimated to be a third higher than the hazard for treatment B.
   2. The hazard for treatment A may in fact be lower than the hazard for treatment B, but we can’t tell just from the estimate.
   3. You should test that the hazards are proportional before even interpreting the hazard ratio.
2. You want to see whether the hazard for death differs by ethnic group using a Cox model. You import a comma-separated file and store the imported data in an R object “g”. Ethnic group takes the vales 1 to 20. The column “time” is the follow-up time. Which of the following statements is/are true?
   1. cox <-coxph(Surv(time, death) ~ ethnicgroup, data = g) assumes a linear relation between ethnic group and the outcome.
   2. cox <-coxph(Surv(time, death) ~ ethnicgroup, data = g) will do the job but only if you make “ethnicgroup” a categorical variable first.
3. You want to fix a multiple Cox model with hospital readmission as the outcome and five columns that could be used as predictors in the data set: age, gender, number of comorbidities, severity of disease on admission (categorical) and type of housing or accommodation (binary). Which of the following tasks are aspects of good statistical practice for the reason given?
   1. Cross-tabulate age and gender to get to know your sample, including whether these variables have any missing values.
   2. Plot the number of comorbidities against the overall proportion who were readmitted to get a very rough sense of the shape of the relation.
4. Which of the following statements are true regarding missing data?
   1. Data set documentation is a good place to start if you want to understand why that data set has missing data.
   2. With data that are missing not at random (MNAR), the probability that a value is missing depends partly on things not in your data set.
5. You run a Cox model using the same outcome and the same set of predictors as used in a previous study published in a good journal, but your hazard ratios, confidence intervals, p-values and indeed some of your conclusions differ from those for that previous study. Which of the following are plausible explanations for the differences?
   1. Your sample size was four times as large as theirs.
   2. Your data were of worse quality.
   3. You set males to be the reference category but they set females to be the reference.