Section 1 Assignment

Instructions (read carefully):

- Each student must submit their own (independent) work.
- Assignments must be done using RMarkdown.
- Submissions must include the .pdf file and the reproducible .rmd file used to do the homework. R code for all applied questions must be provided and be executable in the .rmd file.
- This assignment is due electronically through CANVAS on Friday Jan 27 2023.
- Students should use the practices covered in Section 0 of the course to organize their folder structure and code (i.e., using RStudio Projects with the `here` package).

Question 1) Using the language of "censoring" and/or "truncation" (left, right, and/or interval), explain why a prospective cohort study is often seen as higher quality than a retrospective cohort study.

Question 2) Using Figure 1 from the Section 1 notes, draw the line diagram for ID = 0 that would result if this individual was left truncated.

Question 3) For a randomly selected subset of 25 observations out of the N = 100 observations in the "section1_cohort.csv" data, fit a line plot for the time-to-event outcome using ggplot. With this line plot, explore different themes. Pick your two favorite themes and compare them to the classic theme (i.e., `theme_classic()`). Title each plot with the name of the theme used. Plot each of the three themes in a grid with one row and three columns. Save the plot as a pdf or png file to directory. Include this plot in your homework output and provide an informative caption with the plot.

Question 4) Please do a basic exploratory analysis of the "section1_cohort.csv" dataset. No more than 1/2 page of results. Provide results for the exposure, the confounder, and the outcome.

Question 5) Describe, in words, the interpretation of the CDF:

$$F(t) = P(T \leq t)$$

AND the survival function:

$$S(t) = P(T > t)$$

if T represents age at death from all causes, and t represents 64 years of age.

Question 6) Using the first five observations from the synthetic data in Table 1 of the course notes, write out (but do not solve for) the terms for the Kaplan-Meier estimator $\hat{S}(t) = \prod_{k \in t_k \leq t}(1 - d_k/n_k)$. Assume that the total population at risk includes all 10 observations in Table 1.

Question 7) Please explain the difference between the `Surv()` function, and the `survfit()` function in the survival package.

Question 8) Refer to Figure 3 in the Section 1 course notes. Note that the dashed blue line in Figure 3 is from the Kaplan-Meier estimator, while the solid black line is from the simple calculation shown in the equations above the Figure (on page 10). Why don't these figures align exactly?

Question 9) Fit the `survfit()` function to the "section1_cohort.csv" data. Before you fit, be sure to re-code the outcome so that any non-zero event counts as an event (i.e., re-code outcome=2 to outcome=1). Examine the R object that you get from this fit. How many elements are in this object? What are the first six elements (describe them briefly, don't just provide their element names). Is there enough information in this object for you to determine the median survival time for the outcome? If so, what is the median survival time?

Question 10) Using the fit from Question 9, plot the cumulative distribution function (not the survival function) using the KM estimator. Interpret the curve assuming that the outcome is death from any cause and the time-scale is year on study.

Question 11) Referring to Figure 6 of the section 1 course notes, why is the cumulative risk represented by the dashed line higher than the cumulative risks represented by the solid black line, even though they are the same events?

Question 12) What is the main problem with using the cause-specific risk to understand the causal effects of exposures on outcomes of interest?

Question 13) Provide a single plot of the cause-specific and sub-distribution risk for "outcome = 1'' in the "section1_cohort.csv" using the Kaplan-Meier, Aalen-Johansen, and Gray's CIF estimators.