Statistics 215B Assignment 2 Due: 5pm Tue 30 Jan 2024

Models for time to failure

Define the *hazard function* for any disease as

$$H(t) = \frac{f(t)}{1 - F(t)}$$

where $F(t)$ and $f(t)$ are the distribution function and density, respectively, for the time to first onset. Thus

$$H(t) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} P(\text{Get disease during time } t \text{ to } t + \epsilon \mid \text{Never had disease up to time } t)$$

What is the hazard function if time to first onset is modelled with the exponential distribution? Is this a reasonable model? Explain.

One alternative is to use the *Weibull distribution*, which has cumulative distribution function

$$G_{\alpha,\beta}(u) = 1 - \exp\{-(u/\alpha)^{\beta}\}, \quad u > 0,\ \alpha > 0,\ \beta > 0.$$

What is the hazard function if time to failure has distribution $G_{\alpha,\beta}$?

Survival curves

Consider the following stylized model of a study to test the effectiveness of a new surgical procedure. One thousand individuals are enrolled into the study, half receiving the surgery and half serving as the control group—no surgery. The time to death by any cause is measured in years, up to a maximum of five years, at which time the study ends.

The ith participant in the study carries two random values: time to death if assigned to receive the surgery, X_i, and time to death if assigned to the control group, Y_i. Assume that $\{(X_i, Y_i) : i = 1, 2 \ldots, N\}$ are drawn independently with X_i and Y_i having distributions $G_{3,2}$ and $G_{2,2}$, respectively, as defined in part one.

Two *survival functions* can be defined as

$$S_X(t) = P(\text{Individual } i \text{ lives past time } t \mid \text{Individual } i \text{ receives surgery}) = P(X_i > t)$$

and

$$S_Y(t) = P(\text{Individual } i \text{ lives past time } t \mid \text{Individual } i \text{ is in control group}) = P(Y_i > t)\,.$$

1. Write an R function to simulate draws of a $G_{\alpha,\beta}$ random variable. Your function should take as input a sample size n as well as α and β. It should return a length-n vector of independent realizations of $G_{\alpha,\beta}$. Your function must not use any of R's random sampling routines apart from `runif`.

2. Write an R function that creates a Kaplan-Meier survival-function estimate. The input is two length-n vectors. The first contains event times. The second vector, parallel to the first, contains logical values: `TRUE` for observed deaths, `FALSE` for censoring events. The return

value should be the estimated survival function: an R function which, when evaluated at t, returns the Kaplan-Meier estimate of surviving past t. Your function must not use anything from the `survival` package or any similar package: the requirement is for you to build a Kaplan-Meier estimate "from scratch". If you are in any doubt about whether a supporting function is permissible, check with the course staff.

3. Simulate the performance of the clinical trial by drawing a sample of 500 failure times from $G_{3,2}$ and 500 from $G_{2,2}$. Estimate S_X and S_Y using Kaplan-Meier. Graphically compare these estimates to the true curves. (Recall that the study has a length of five years.) **Note**: do not be surprised by low survival rates at the five-year horizon. The people in this study are very sick.

4. The simulation in (3) did not include the possibility of censoring. Assume that for individual i there is another random variable, denoted Z_i, that gives the time at which i will be censored—if he lives that long. Consider the case when the Z_i are i.i.d. exponential random variables with mean 10, chosen independently of X_i and Y_i. Simulate censoring times under this scenario and create new Kaplan-Meier survival curves. Compare these with the true survival curves.

5. Repeat the previous exercise, but now suppose the Z_i are independent exponential random variables whose mean depends on the individual's time of death. If the time of death is less than two years, the mean of the distribution of Z_i is 10; otherwise, the mean is 5. This could arise in a study where the sicker patients are more likely to remain under the care of their doctors. Discuss the results. What is the key difference between this censoring scenario and the previous one?