MATH-449 - Biostatistics
EPFL, Spring 2024
Problem Set 3

1. Consider the DAG below, which we discussed in lectures during Week 3. Unless directed otherwise, ignore the **red** arrow in the graph (pretend it is not there, until told otherwise). Suppose you are consulting with a clinician who is interested in estimating the expected potential outcomes in their study population under treatment level $a = 1$ and comparing it to that under treatment level $a = 0$. Suppose the investigator has access to data on $\{L, A, Y\}$ but that C represents a variable indicating that patients were lost to follow-up and data on Y for patients with $C = 1$ is not recorded.

   (a) Consider an intervention that sets treatment $A = a$. What are the back-door paths for this intervention with respect to outcome Y?

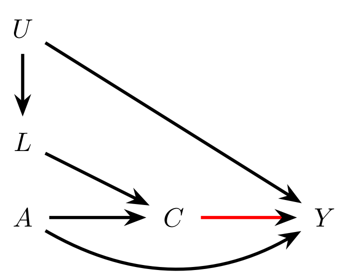
   (b) Is the backdoor criterion satisfied? If so provide a functional of parameters of the distribution of $\{L, A, Y\}$ that identifies $\mathbb{E}[Y^a]$. If so, describe an estimation strategy for this functional that you would share with your collaborators, or name some issues that would arise.

   (c) Consider an intervention that sets $A = a$ and $C = 0$. What are the back-door paths for this intervention with respect to outcome Y?

   (d) Is the backdoor criterion satisfied? If so provide a functional of parameters of the distribution of $\{L, A, Y\}$ that identifies $\mathbb{E}[Y^{a,c=0}]$. If so, describe an estimation strategy for this functional that you would share with your collaborators.

   (e) This parameter has a different interpretation than the one that your collaborators first described. Can you convince them that this parameter is still of interest, and if so, how?

   (f) Suppose now that the red arrow is present. Repeat the previous steps (c)-(e) for $\mathbb{E}[Y^{a,c=0}]$ with this modified graph.



2. A survival time T is exponentially distributed with rate parameter $\beta > 0$ if its survival function, $S(t) = P(T > t)$, takes the form $S(t) = e^{-\beta t}$ for $t \geq 0$. Note additional definitions at the end of the problem set.

   a) Find the density function $f(t) = -\frac{d}{dt}S(t)$.

   b) Find the hazard function and the cumulative hazard function.

   c) A waiting time T is *memoryless* if $P(T > t + s | T > t) = P(T > s)$ for all $t, s \geq 0$, i.e. if the waiting time distribution does not depend on how much time has already elapsed. Show that an exponentially distributed waiting time is memoryless.

3. Consider the following definition:

   Definition 1 (Discrete martingale)**.** *Let $M = \{M_0, M_1, M_2, \dots\}$ be a discrete stochastic process adapted to $\{\mathcal{F}_n\}$. The discrete process M is a martingale if*

$$\mathbb{E}(M_n \mid \mathcal{F}_{n-1}) = M_{n-1}$$

.

(Exercise 2.1 in ABG 2008) Let M_n be a discrete time martingale with respect to the filtration $\mathcal{F}_n$, for $n \in \{0, 1, 2, \cdots\}$. By definition of M being a martingale we have that $E[M_n|\mathcal{F}_{n-1}] = M_{n-1}$ for all $n \geq 1$. Show that this is equivalent to $E[M_n|\mathcal{F}_m] = M_m$ whenever $n \geq m \geq 0$.

4. a) Find $E[T]$ when T is a Weibull distributed variable, i.e. when the hazard function of T is $\alpha(t) = \lambda k t^{k-1}$ for $\lambda, k > 0$[‡]

   b) (Exercise 1.3 in ABG 2008) Suppose T is a survival time with finite expectation. Show that[†]

$$E[T] = \int_0^\infty P(T > s)ds.$$

Additional definitions

Definition 2 (Survival function)**.** *The survival function is $S(t) = P(T > t)$, that is, the probability that the survival time T exceeds t.*

Definition 3 (Hazard rate)**.** *The hazard rate $\alpha(t) = \lim_{dt\to 0} \frac{1}{dt} P(t + dt > T > t \mid T \geq t)$ is the rate of events per unit of time.*

Definition 4 (Cumulative Hazard rate)**.** *Define the cumulative hazard,*

$$H(t) = \int_0^t \alpha(s)ds.$$

[‡]Hint: Express the solution using the gamma function, which is given by $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$.
[†]Hint: Write $T = \int_0^\infty I(T > u)du$.