MATH-449 - Biostatistics
EPFL, Spring 2022
Problem Set 1

1. Determine whether each of these questions are phrased as causal questions or not (yes or no).

   a) Does the Moderna vaccine reduce the risk of severe COVID-19 infection?

   b) Do women with breast cancer survive longer than men with prostate cancer?

   c) Is the life expectancy in Switzerland longer than the life expectancy in Italy?

   d) Does drinking 0.5 L beer compared to 0.5 L Coca Cola at 19h00 affect the quality of sleep?

   e) Would drinking a cup of coffee 2 hours before your exam improve your performance?

2. Based on the definition of a causal effects in the lecture slides, argue whether the following statements about a covariate $L \in \mathbb{R}$, a treatment $A = 0, 1$ and an outcome $Y \in \mathbb{R}$ are right or wrong (there is no guarantee that A is randomly assigned).

   a) $\mathbb{E}(Y^{a=1} \mid L = l) - \mathbb{E}(Y^{a=0} \mid L = l)$ is a causal effect .

   b) $\mathbb{E}(Y \mid A = 1, L = l) - \mathbb{E}(Y \mid A = a, L = l)$ is a causal effect .

   c) $\mathbb{E}(Y^{a=1} \mid A = 1, L = l) - \mathbb{E}(Y^{a=0} \mid A = 1, L = l)$ is a causal effect .

   d) $\frac{\mathbb{E}(Y^{a=1})}{\mathbb{E}(Y^{a=0})}$ is an average over individual level (additive) causal effects.

3. Translate these English sentences to mathematical (counterfactual) statements.

   a) The average causal effect of receiving a COVID-19 vaccine ($A = 1$) vs placebo ($A = 0$) on mortality after one year ($Y = 1$ is death, $Y = 0$ is alive) in the entire population of interest.

   b) The average causal effect of receiving a COVID-19 vaccine ($A = 1$) vs placebo ($A = 0$) on mortality after one year ($Y = 1$ is death, $Y = 0$ is alive) among those who received placebo in the observed (factual) data.

   c) The average causal effect of receiving a COVID-19 vaccine ($A = 1$) vs placebo ($A = 0$) on mortality after one year ($Y = 1$ is death, $Y = 0$ is alive) among those who received treatment in the observed (factual) data.

   d) The average causal effect of receiving a COVID-19 vaccine ($A = 1$) vs placebo ($A = 0$) on mortality after one year ($Y = 1$ is death, $Y = 0$ is alive) in men ($X = 1$).

   e) Are your answers in a)-d) estimands, estimators or estimates?

4. Suppose investigators had access to data from a study in which they observed for each patient a binary outcome Y, a binary treatment A and a 4-level baseline covariate L. The parameters of the joint density of (L, A, Y) were computed from the data and summarized in Table 1 (where we suppose that the sample size was so large, that sampling variability is not a concern).

   a) From the parameters in Table 1, compute $\mathbb{E}[Y]$.

   b) Suppose now that the data did not in fact arise from a regular observational study, but had instead come from a special trial. Upon recruitment into the study, each patient's covariate L is measured and then they are sorted into groups based on that covariate's value. In each group, the investigators conduct a separate experiment, which are identical except they use a special coin to randomize patients to either treatment ($a = 1$) or control ($a = 0$), with "heads" corresponding to treatment and "tails" corresponding to control. The probabilities for heads for each of these sub-trials is given by the column labeled $P(A = 1 \mid L = l)$. Assume consistency holds ($Y^A = Y$), and that patients perfectly complied with their assignments. With the information in the table, compute the effect of treatment $\mathbb{E}[Y^{a=1} - Y^{a=0} \mid L = l]$ for each subgroup $L = l$ that was targeted in each of the sub-trials. What additional assumptions did you use along the way, that was justified given the source of the data?

c) From the quantities computed in part a), use laws of probability to compute the average treatment effect, among the whole population, $\mathbb{E}[Y^{a=1} - Y^{a=0}]$.

d) Draw a directed acyclic graph (DAG) that could depict the mechanism that generated the observed data.

e) The data analyst for the study approaches you and said they made a terrible mistake: when preparing the column $P(A = 1 \mid L = l)$ in Table 1, they reverse coded the treatment variable, so in fact the true values of the treatment propensities are 1 minus those listed in the table. What will be the values of the previously computed parameters, and explain in words why these changes did (or did not occur).

	$P(Y = 1 \mid A = a, L = l)$		$P(A = 1 \mid L = l)$	$P(L = l)$
	$a = 1$	$a = 0$		
$l = 1$	.1	.8	.2	.2
$l = 2$	.2	.7	.4	.4
$l = 3$	.3	.6	.6	.1
$l = 4$	.4	.5	.8	.3

Table 1: Parameters of $P_{L,A,Y}$ observed in the conditionally randomized trial.

5. Consider a covariate $L \in \mathbb{R}$, a treatment $A = 0, 1$ and an outcome $Y \in \mathbb{R}$.

a) Investigator 1 claims that $A \perp\!\!\!\perp Y \implies A \perp\!\!\!\perp Y \mid L$. Show that the statement is wrong.

b) Investigator 2 claims that $A \perp\!\!\!\perp Y \mid L \implies A \perp\!\!\!\perp Y$. Show that the statement is wrong.