STATS 116: Homework 2

Due: Thursday, July 13, 2023 at 10:00 pm PDT on Gradescope

There are 7 problems on this assignment, each worth 8 points, although subparts within a problem may not be equally weighted. Credit will be assigned primarily based on reasoning and work, not the final answer. You do not need to simplify arithmetic expressions unless otherwise noted. While you may discuss the problems on this assignment other students, you must write up your own solutions. As per the syllabus, you may occasionally use the Internet or other public resources to clarify concepts with citation when this information is used as part of your own solution to a homework problem. However, you may not search for direct solutions to any problems assigned for homework or exams. For example, you can ask ChatGPT to clarify a particular concept from lecture that may be related to a problem, but you cannot feed it any part of a course assignment or a substantively similar version.

1. Jane Villanueva is taking a pregnancy test because she was artificially inseminated. Suppose 20 percent of artificially inseminated women get pregnant, although it is known that 25 percent of artificially inseminated women test positive on this pregnancy test. The false positive rate of the test among artificially inseminated women is 10 percent (i.e. 10 percent of non-pregnant artificially inseminated women test positive on the test).

   (a) Compute the false negative rate of the test: that is, the probability that an artificially inseminated pregnant woman tests negative.

   All probabilities in this problem will implicitly condition on the event that Jane was artificially inseminated. Let T be the event that Jane tests positive and R be the event that Jane is pregnant. We are given $P(R) = 0.2$, $P(T) = 0.25$, and $P(T \mid R^c) = 0.1$. By LOTP we have

   $$P(T) = P(T \mid R)P(R) + P(T \mid R^c)P(R^c) \iff 0.25 = P(T \mid R) \cdot 0.2 + 0.1 \cdot 0.8$$

   Solving gives $P(T \mid R) = 0.85$ and hence the false negative rate is $P(T^c \mid R) = \boxed{0.15}$.

   (b) What is the probability that Jane is pregnant, given that she takes the test twice and gets a positive result both times? Assume the two test results are conditionally independent given Jane's pregnancy status.

Let $T_i, i = 1, 2$ be the event that Jane tests positive on test i. By Bayes' rule we have

$$P(R \mid T_1, T_2) = \frac{P(T_1, T_2 \mid R)P(R)}{P(T_1, T_2 \mid R)P(R) + P(T_1, T_2 \mid R^c)P(R^c)}$$
$$= \frac{P(T_1 \mid R)P(T_2 \mid R)P(R)}{P(T_1 \mid R)P(T_2 \mid R)P(R) + P(T_1 \mid R^c)P(T_2 \mid R^c)P(R^c)}$$

where the second equality is by conditional independence. With $P(T_1 \mid R) = P(T_2 \mid R) = 0.85$ as computed above and $P(T_1 \mid R^c) = P(T_2 \mid R^c) = 0.1$, we plug in numbers to get

$$P(R \mid T_1, T_2) = \frac{0.85 \cdot 0.85 \cdot 0.2}{0.85 \cdot 0.85 \cdot 0.2 + 0.1 \cdot 0.1 \cdot 0.8} \approx \boxed{0.95}$$

2. Harrison flips a fair coin repeatedly, recording the sequences of flips he observes.

(a) What is the probability he observes the sequence HH for the first time before observing the sequence HT for the first time?

Let A be the event that HH is observed before HT and B_i be the event that the i-th flip lands heads. Then

$$P(A \mid B_1) = P(B_2 \mid B_1) = P(B_2) = 1/2$$

by independence of B_1 and B_2 and the fact that when the first flip is heads, A occurs is and only if the second flip is heads. Furthermore $P(A \mid B_1^c) = P(A)$, since given the first flip is tails, the event A is the event that HH occurs before HT in the sequence of flips *starting from the second*, which are independent of the first flip. Then by LOTP

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_1^c)P(B_1^c) = 1/2 \cdot 1/2 + P(A) \cdot 1/2$$

which shows $P(A) = \boxed{1/2}$.

(b) What is the probability he observes the sequence HHT for the first time before observing the sequence HTH?

Let A be the event that HHT is observed before HTH and B_i be the event that the i-th flip lands heads. If the first flip is tails, the problem "resets" as in part (a), i.e. $P(A \mid B_1^c) = P(A)$. If the first flip is heads, then if the second flip is also heads, A happens for sure, as HHT will be observed on the next flip on which a tail occurs, and HTH cannot have occurred by then (as we have had no tails before the HHT occurrence). Thus $P(A \mid B_1, B_2) = 1$. If the first flip is heads and the second flip is tails, then clearly A never occurs whenever the third flip is heads, while the problem resets if the third flip is tails as then none of the first three flips can possibly be part of the "game-ending" sequence (the first HHT or HTH). Hence $P(A \mid B_1, B_2^c, B_3^c) = P(A)$ and $P(A \mid B_1, B_2^c, B_3) = 0$, so by LOTP with extra conditioning

$$\begin{aligned} P(A \mid B_1, B_2^c) &= P(A \mid B_1, B_2^c, B_3)P(B_3 \mid B_1, B_2^c) + P(A \mid B_1, B_2^c, B_3^c)P(B_3^c \mid B_1, B_2^c) \\ &= 0 \cdot 1/2 + P(A) \cdot 1/2 \end{aligned}$$

Another application of LOTP shows

$$\begin{aligned} P(A) &= P(A \mid B_1^c)P(B_1^c) + P(A \mid B_1, B_2)P(B_1, B_2) + P(A \mid B_1, B_2^c)P(B_1, B_2^c) \\ &= P(A) \cdot 1/2 + 1 \cdot 1/4 + 1/2 \cdot P(A) \cdot 1/4 \end{aligned}$$

where the second equality uses independence of the coin flips. Solving gives $P(A) = \boxed{2/3}$.

3. (BH 2.10) Fred is working on a major project. In planning the project, two milestones are set up, with dates by which they should be accomplished. This serves as a way to track Fred's progress. Let A_1 be the event that Fred completes the first milestone on time, A_2 be the event that he completes the second milestone on time, and A_3 be the event that he completes the project on time. Suppose that $P(A_{j+1} \mid A_j) = 0.8$ but $P(A_{j+1} \mid A_j^c) = 0.3$ for $j = 1, 2$, since if Fred falls behind on his schedule it will be hard for him to get caught up. Also, assume that the second milestone supersedes the first, in the sense that once we know whether he is on time in completing the second milestone, it no longer matters what happened with the first milestone. We can express this by saying that A_1 and A_3 are conditionally independent given A_2 and they're also conditionally independent given A_2^c.

(a) Find the probability that Fred will finish the project on time, given that he completes the first milestone on time. Also find the probability that Fred will finish the project on time, given that he is late for the first milestone.

We need to find $P(A_3 \mid A_1)$. To do so, let's use LOTP to condition on whether or not A_2 occurs:

$$P(A_3 \mid A_1) = P(A_3 \mid A_1, A_2)P(A_2 \mid A_1) + P(A_3 \mid A_1, A_2^c)P(A_2^c \mid A_1)$$

Using the conditional independence assumptions, the right-hand side becomes

$$P(A_3 \mid A_2)P(A_2 \mid A_1) + P(A_3 \mid A_2^c)P(A_2^c \mid A_1) = (0.8)(0.8) + (0.3)(0.2) = \boxed{0.7}$$

Similarly,

$$P(A_3 \mid A_1^c) = P(A_3 \mid A_2)P(A_2 \mid A_1^c) + P(A_3 \mid A_2^c)P(A_2^c \mid A_1^c) = (0.8)(0.3) + (0.3)(0.7) = \boxed{0.45}$$

(b) Suppose that $P(A_1) = 0.75$. Find the probability that Fred will finish the project on time.

By LOTP and part (a),

$$P(A_3) = P(A_3 \mid A_1)P(A_1) + P(A_3 \mid A_1^c)P(A_1^c) = (0.7)(0.75) + (0.45)(0.25) = \boxed{0.6375}$$

4. (BH 2.36) Suppose that in the population of college applicants, being good at baseball is independent of having a good math score on a certain standardized test (with respect to some measure of "good"). A certain college has a simple admissions procedure: admit an applicant if and only if the applicant is good at baseball or has a good math score on the test.

(a) Give an intuitive explanation of why it makes sense that among students that the college admits, having a good math score is negatively associated with being good at baseball, i.e., conditioning on having a good math score decreases the chance of being good at baseball.

Even though baseball skill and the math score are independent in the general population of applicants, it makes sense that they will become dependent (with a negative association) when restricting only to the students who are admitted. This is because within this sub-population, having a bad math score implies being good at baseball (else the student would not have been admitted). So having a good math score decreases the chance of being good in baseball (as shown in BH 2.16, if an event B is evidence in favor of an event A, then B^c is evidence against A).

As another explanation, note that 3 types of students are admitted: (i) good math score, good at baseball; (ii) good math score, bad at baseball; (iii) bad math score, good at baseball. Conditioning on having good math score removes students of type (iii) from consideration, which decreases the proportion of students who are good at baseball.

(b) Show that if A and B are independent and $C = A \cup B$, then A and B are conditionally dependent given C (as long as $P(A \cap B) > 0$ and $P(A \cup B) < 1$), with

$$P(A \mid B, C) < P(A \mid C)$$

This phenomenon is known as Berkson's paradox, especially in the context of admissions to a school, hospital, etc.

Note $B \cap C = B$ so $P(A \mid B, C) = P(A \mid B) = P(A)$ by independence of A and B. However

$$P(A \mid C) = \frac{P(A)}{P(C)} > P(A)$$

by definition of conditional probability since $P(A) > 0$ and $P(C) < 1$.

5. (BH 2.40) Consider the Monty Hall problem, except that Monty enjoys opening door 2 more than he enjoys opening door 3, and if he has a choice between opening these two doors, he opens door 2 with probability p, where $0.5 \leq p \leq 1$. To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is door 1. Monty Hall then opens a door to reveal a goat, and offers you the option of switching. Assume that Monty Hall knows which door has the car, will always open a goat door and offer the option of switching, and as above assume that if Monty Hall has a choice between opening door 2 and door 3, he chooses door 2 with probability p (with $0.5 \leq p \leq 1$).

   (a) Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of doors 2 or 3 Monty opens).

   Let C_j be the event that the car is hidden behind door j and let W be the event that we win using the switching strategy. Using the law of total probability, we can find the unconditional probability of winning:

$$\begin{aligned} P(W) &= P(W \mid C_1)P(C_1) + P(W \mid C_2)P(C_2) + P(W \mid C_3)P(C_3) \\ &= 0 \cdot 1/3 + 1 \cdot 1/3 + 1 \cdot 1/3 = 2/3 \end{aligned}$$

   (b) Find the probability that the strategy of always switching succeeds, given that Monty opens door 2.

   Let D_i be the event that Monty opens door i. Note that we are looking for $P(W \mid D_2)$, which is the same as $P(C_3 \mid D_2)$ as we first choose door 1 and then switch to door 3. By Bayes' rule and the law of total probability,

$$\begin{aligned} P(C_3 \mid D_2) &= \frac{P(D_2 \mid C_3)P(C_3)}{P(D_2)} \\ &= \frac{P(D_2 \mid C_3)P(C_3)}{P(D_2 \mid C_1)P(C_1) + P(D_2 \mid C_2)P(C_2) + P(D_2 \mid C_3)P(C_3)} \\ &= \frac{1 \cdot 1/3}{p \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} \\ &= \boxed{\frac{1}{1+p}} \end{aligned}$$

(c) Find the probability that the strategy of always switching succeeds, given that Monty opens door 3.

The structure of the problem is the same as Part (b) (except for the condition that $p \geq 1/2$, which was not needed above). Imagine repainting doors 2 and 3, reversing which is called which. By Part (b) with $1-p$ in place of p,

$$P(C_2 \mid D_3) = \frac{1}{1+(1-p)} = \boxed{\frac{1}{2-p}}$$

6. (BH 3.1) People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, i.e. before person X arrives no two people have the same birthday, but when person X arrives there is a match. Find the PMF of X.

We will make the usual assumptions as in the birthday problem (e.g., exclude February 29). The support of X is $\{2, 3, \dots, 366\}$ since if there are 365 people there and no match, then every day of the year is accounted for and the 366th person will create a match. Let's start with a couple simple cases and then generalize:

$$P(X = 2) = \frac{1}{365}$$

since the second person has a 1/365 chance of having the same birthday as the first,

$$P(X = 3) = \frac{364}{365} \cdot \frac{2}{365}$$

since $X = 3$ means that the second person didn't match the first but the third person matched one of the first two. In general, for $2 \leq k \leq 366$ we have

$$\begin{aligned} P(X = k) &= P(X > k-1, X = k) \\ &= P(X > k-1)P(X = k \mid X > k-1) \\ &= \frac{365 \cdot 364 \dots (365-k+2)}{365^{k-1}} \cdot \frac{k-1}{365} \\ &= \boxed{\frac{(k-1) \cdot 364 \cdot 363 \dots (365-k+2)}{365^{k-1}}} \end{aligned}$$

7. Suppose X and Y are discrete random variables with finite support that have the same distribution.

   (a) Is it possible that $P(X = Y) = 0$? Give an example, or show that it's not possible.

Yes, it is possible. Suppose I flip a fair coin so the sample space has two equally likely outcomes, H and T. Let $X(H) = Y(T) = 1$ and $X(T) = Y(H) = 0$. Then both X and Y are supported on $\{0, 1\}$ and both have PMF p with $p(0) = p(1) = 1/2$, however $X \neq Y$ always.

(Note a more compact way of writing this solution, in terms of the material of Lecture 8, is to take any $X \sim \text{Bern}(1/2)$ and $Y = 1 - X$).

(b) Is it possible that $P(X > Y) = 1$? Give an example, or show that it's not possible.

No, it is not possible. Let the common support of X and Y be $\{a_1, \ldots, a_n\}$ where $a_1 < \cdots < a_n$. Then we have $X \leq Y$ whenever $X = a_1$, so $P(X \leq Y) \geq P(X = a_1) > 0$ since a_1 is in the support of X, showing $P(X > Y) = 1 - P(X \leq Y) < 1$.