METHODS FOR CAUSAL INFERENCE: TUTORIAL 1

* = For formative assignment (marked but not evaluated)
** = For summative assignment (marked and evaluated, counting towards the 20% of student's final mark)

1. * Write down and discuss 3 case studies, in any context you like (*e.g.* , biomedical, education, economics, ...), where causal inference can play or already plays a major role in guiding policies. Try to explicitly state the causal question of interest and the variables/data that may be relevant for your case studies of interest.

2. * For each of the case studies in question 1, write down a counterfactual statement, imagining what would have happened if a different intervention was used?

3. "Cold weather increases the desire to spend more money in the shops and online." Discuss what is wrong with this sentence.

   Solution: Cold weather co-insides with Christmas and the new year sale, which might explain better why national spending increases during the winter months. One could also discuss price of winter vs summer clothing.

4. [*Causal Inference in statistics: A primer*, Chapter 1] In the following case study, determine if you should use the aggregate or the segregated data to determine the true effect. There are two doctors in a small town. Each has performed 100 surgeries in their career. Doctor A performs easier surgeries more often, while B performs difficult surgeries more often. You need surgery, but do not know if yours is going to be difficult or easy. In order to maximise the chance of a successful surgery, should you consult the success rate of each doctor over all cases, or consult their success rates for the easy and difficult cases separately?

   Solution: The difficulty of surgery is a confounder for the choice of doctor and a the rate of success of the surgery (plot the classical confounder graph). The difficulty of surgery affects the choice of doctor and also chances of recovery as more difficult cases can have less chance of success. Therefore, to make a causal conclusion, we need to consult the segregated data, by conditioning on the level of difficulty.

5. Suppose we perform the experiment of rolling two fair dice. Plot the outcome/event space. By shading the appropriate areas of the outcomes space, compute the following probabilities:

   (a) Getting a double 6
   (b) The sum of the rolls is even.
   (c) The sum of the rolls is odd.
   (d) The first roll is equal to the second
   (e) At least one roll is equal to 4
   (f) At least one roll is equal or larger than 4

   Solution: See Fig. 1.

6. Using the total law of probability:

$$p(X) = \sum_z p(X, Z = z) = \sum_z p(X|Z = z)p(Z = z) \ , \tag{1}$$

   show that:

$$p(X|Y) = \sum_z p(X|Y, Z = z)p(Z = z|Y)\cdot \tag{2}$$

   Solution:

$$\sum_z p(X|Y, Z = z)p(Z = z|Y) = \sum_z \frac{p(X, Y, Z = z)}{p(Y, Z = z)} \frac{p(Y, Z = z)}{p(Y)} \tag{3}$$

$$= \frac{1}{p(Y)} \sum_z p(X, Y, Z = z) = \frac{p(X, Y)}{p(Y)} = p(X|Y) \tag{4}$$

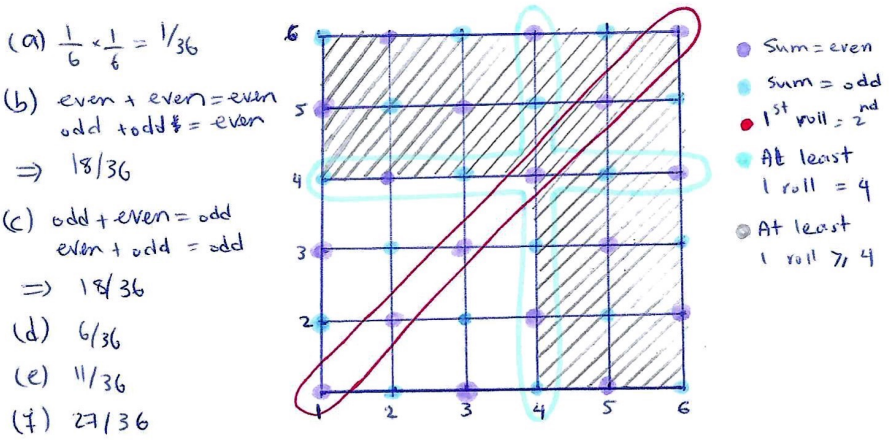

FIG. 1: Outcome space Ω, for two rolls of dice.

7. Consider the graph below, Fig. 2, indicating the relationship between gender and higher-education subject groups in the UK in 2017/2018. Estimate the following probabilities:

   (a) P("Mathematical and Computer Science")

   (b) P("Mathematical and Computer Science" OR Female)

   (c) P("Mathematical and Computer Science" | Female)

   (d) P(Female | "Mathematical and Computer Science")

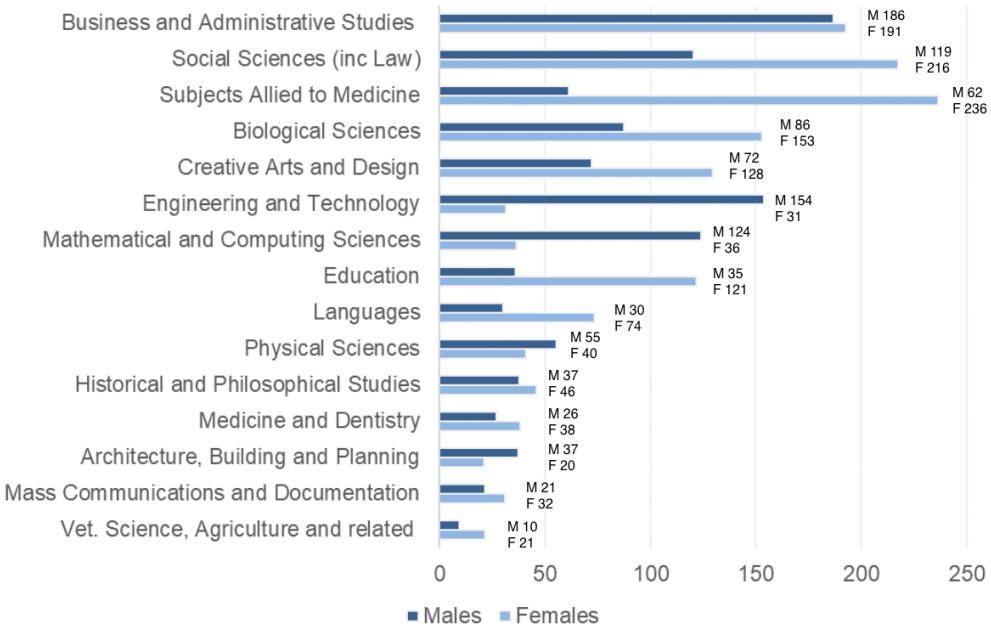

FIG. 2: "Students in higher education in the UK by gender and subject group, 2017/18 (thousands)". The figure from the Department of Education report, *Education and Training Statistics for the United Kingdom 2019*. The numbers are approximate.

8. * The NHS may invite individuals over 50 who have had abnormal bowel screens for further examination, e.g. via colonoscopy, in order to examine the size of polyps (abnormal growth of cells) in the bowel. This is done

in order to detect early signs of bowel cancer and increase survival rates. Suppose we are told[1] that if the examination forecast is 'abnormal' the probability of cancer actually being present is 80%. If the examination forecast is 'normal' then the probability of cancer actually being present is 5%.

Furthermore, individuals with a strong family history of bowel cancer have a forecast of 60% for 'abnormal' polyps, while those without evidence of family history have a forecast of 20% for 'abnormal' polyps. We may also assume that knowing the family history in addition to the forecast does not change our belief about cancer being present or not (*i.e.* , once we know the forecast leads to the above probabilities, irrespective of family history). Use the information provided to answer the following:

Suppose a patient forgot to to open the letter containing their colonoscopy results and unfortunately they have early cancer. What is the probability that the forecast is 'abnormal' if they had a strong family history of bowel cancer? What is the probability that the forecast is 'abnormal' if they did not have a strong family history of bowel cancer?

Hint 1: Before performing any calculations, write down the assumptions and conditional probabilities that are known or can easily be inferred.
Hint 2: The results in question 6 may be useful here.

Solution: The hint suggest we start by writing the following information:

$p(\text{cancer}|\text{'abnormal'}) = 0.8$
$p(\text{no cancer}|\text{'abnormal'}) = 0.2$
$p(\text{cancer}|\text{'normal'}) = 0.05$
$p(\text{no cancer}|\text{'normal'}) = 0.95$
$p(\text{'abnormal'}|\text{family history}) = 0.6$
$p(\text{'normal'}|\text{family history}) = 0.4$
$p(\text{'abnormal'}|\text{no family history}) = 0.2$
$p(\text{'normal'}|\text{no family history}) = 0.8$
Conditional independence: $p(\text{cancer/no cancer}|\text{forecast, (no) family history}) = p(\text{cancer/no cancer}|\text{forecast})$

We wish to calculate $p(\text{'abnormal'}|\text{cancer}, \text{family history})$, so we can use Bayes' Rule (fh is short for family history):

$$p(\text{'abnormal'}|\text{cancer}, \text{fh}) = \frac{p(\text{cancer}|\text{'abnormal', fh}) \times p(\text{'abnormal'}|\text{fh})}{p(\text{cancer}|\text{fh})} \tag{5}$$

$$= \frac{p(\text{cancer}|\text{'abnormal'}) \times p(\text{'abnormal'}|\text{fh})}{p(\text{cancer}|\text{fh})} \tag{6}$$

$$= \frac{0.8 \times 0.6}{p(\text{cancer}|\text{fh,'abnormal'})p(\text{'abnormal'}|\text{fh}) + p(\text{cancer}|\text{fh,'normal'})p(\text{'normal'}|\text{fh})} \tag{7}$$

$$= \frac{0.8 \times 0.6}{p(\text{cancer}|\text{'abnormal'})p(\text{'abnormal'}|\text{fh}) + p(\text{cancer}|\text{'normal'})p(\text{'normal'}|\text{fh})} \tag{8}$$

$$= \frac{0.8 \times 0.6}{0.8 \times 0.6 + 0.05 \times 0.4} = 0.96 \tag{9}$$

[1] All the numbers in this case study are made-up.

$$p(\text{'abnormal'}|\text{cancer, no fh}) = \frac{p(\text{cancer}|\text{'abnormal', no fh}) \times p(\text{'abnormal'}|\text{no fh})}{p(\text{cancer}|\text{no fh})} \tag{10}$$

$$= \frac{p(\text{cancer}|\text{'abnormal'}) \times p(\text{'abnormal'}|\text{no fh})}{p(\text{cancer}|\text{no fh})} \tag{11}$$

$$= \frac{0.8 \times 0.6}{p(\text{cancer}|\text{no fh,'abnormal'})p(\text{'abnormal'}|\text{no fh}) + p(\text{cancer}|\text{no fh,'normal'})p(\text{'normal'}|\text{no fh})} \tag{12}$$

$$= \frac{0.8 \times 0.2}{p(\text{cancer}|\text{'abnormal'})p(\text{'abnormal'}|\text{no fh}) + p(\text{cancer}|\text{'normal'})p(\text{'normal'}|\text{no fh})} \tag{13}$$

$$= \frac{0.8 \times 0.2}{0.8 \times 0.2 + 0.05 \times 0.8} = 0.8 \tag{14}$$

9. Create a simulation for the Monty Hall problem described in the lectures, and check if the probablity of winning is more if one were to switch door after the host reveals a goat. The process may be repeated a large number of times (e.g. 100 or 1000) to obtain a robust estimate of the 1/3 (no switching) vs 2/3 (switching) probabilities.

10. [*Causal Inference in statistics: A primer*, Chapter 1] For the graph in 3, determine the following:

    (a) All parents of Z.

    (b) All ancestors of Z.

    (c) All children of W.

    (d) All descendants of W.

    (e) Draw all simple paths between X and T (*i.e.* , no node should appear more than once).

    (f) Draw all the directed paths between X and T.

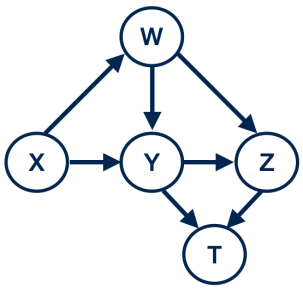

FIG. 3: A directed graph.

Solution:

(a) W, Y.

(b) X, W, Y.

(c) Y, Z.

(d) Y, Z, T.

(e) $\{X, Y, T\}$, $\{X, Y, Z, T\}$, $\{X, Y, W, Z, T\}$, $\{X, W, Y, T\}$, $\{X, W, Z, T\}$,$\{X, W, Y, Z, T\}$, $\{X, W, Z, Y, T\}$

(f) $\{X, Y, T\}$, $\{X, Y, Z, T\}$, $\{X, W, Y, T\}$, $\{X, W, Z, T\}$, $\{X, W, Y, Z, T\}$

11. * [*Causal Inference in statistics: A primer*, Chapter 1] Consider the chain of binary random variables $X_1 \to X_2 \to X_3 \to X_4$ with the following consecutive conditional probabilities:

$$P(X_i = 1|X_{i-1} = 1) = p, \tag{15}$$
$$P(X_i = 1|X_{i-1} = 0) = q, \tag{16}$$
$$P(X_1 = 1) = p_0 \cdot \tag{17}$$

Compute the following probabilities:

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0), \tag{18}$$
$$P(X_4 = 1|X_1 = 1), \tag{19}$$
$$P(X_1 = 1|X_4 = 1) \tag{20}$$
$$P(X_3 = 1|X_1 = 0, X_4 = 1) \cdot \tag{21}$$

Solution: Write the joint probability in terms of the conditional probabilities imposed by the chain graph:

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) =& P(X_4 = 0|X_3 = 1)P(X_3 = 1|X_2 = 0)P(X_2 = 0|X_1 = 1)P(X_1 = 1) \\ =& (1-p)q(1-p)p_0 = (1-p)^2 q p_0 \end{aligned} \tag{22}$$

$$P(X_4 = 1|X_1 = 1) = \frac{P(X_4 = 1, X_1 = 1)}{P(X_1 = 1)} = \frac{1}{p_0} \sum_{x_2,x_3} P(X_4 = 1, X_1 = 1, x_2, x_3) \tag{23}$$

$$\begin{aligned} &= \frac{1}{p_0} \sum_{x_2,x_3} P(X_4 = 1|x_3)P(x_3|x_2)P(x_2|X_1 = 1)P(X_1 = 1) \\ &= P(X_4 = 1|X_3 = 1)P(X_3 = 1|X_2 = 1)P(X_2 = 1|X_1 = 1) \\ &+ P(X_4 = 1|X_3 = 1)P(X_3 = 1|X_2 = 0)P(X_2 = 0|X_1 = 1) \\ &+ P(X_4 = 1|X_3 = 0)P(X_3 = 0|X_2 = 1)P(X_2 = 1|X_1 = 1) \\ &+ P(X_4 = 1|X_3 = 0)P(X_3 = 0|X_2 = 0)P(X_2 = 0|X_1 = 1) \\ &= p^3 + pq(1-p) + q(1-p)p + q(1-q)(1-p) = p^3 + 2pq(1-p) + q(1-q)(1-p) \end{aligned} \tag{24}$$

$$\begin{aligned} P(X_1 = 1|X_4 = 1) &= \frac{P(X_1 = 1, X_4 = 1)}{P(X_4 = 1)} = \frac{\sum_{x_2,x_3} P(X_4 = 1, X_1 = 1, x_2, x_3)}{\sum_{x_1,x_2,x_3} P(x_1, x_2, x_3, X_4 = 1)} \\ &= \frac{p_0 p^3 + 2p_0 pq(1-p) + p_0 q(1-q)(1-p)}{p_0 p^3 + 2p_0 pq(1-p) + p_0 q(1-q)(1-p) + \text{terms with } X_1 = 0} \end{aligned} \tag{25}$$

$$\text{terms with } X_1 = 0 : (1-p_0)\Big[p^2 q + q(1-p)q + pq(1-q) + q(1-q)^2\Big] \tag{26}$$

$$\begin{aligned} P(X_3 = 1|X_1 = 0, X_4 = 1) &= \frac{\sum_{x_2} P(X_3 = 1, X_4 = 1, X_1 = 0, x_2)}{P(X_4 = 1, X_1 = 0)} \\ &= \frac{\sum_{x_2} P(X_4 = 1|X_3 = 1)P(X_3 = 1|x_2)P(x_2|X_1 = 0)P(X_1 = 0)}{(1-p_0)\Big[p^2 q + q(1-p)q + pq(1-q) + q(1-q)^2\Big]} \\ &= \frac{p^2 q(1-p_0) + pq(1-q)(1-p_0)}{(1-p_0)\Big[p^2 q + q(1-p)q + pq(1-q) + q(1-q)^2\Big]} \end{aligned} \tag{27}$$