Homework 1 *Probability & Statistics for DS & AI, Michele Guindani, Fall 2022*

Topics: Basic concepts of Probability — Due on July 27, 2022

Note: Attach your code with your submission.

1. **Basic Probability** (Do both the following problems) [8+4=12 points]

   (a) Consider an experiment consisting of rolling a dice twice. The outcome of this experiment is an ordered pair whose first element is the first value rolled and whose second element is the second value rolled.

      i. Find the sample space.

      ii. Find the set A representing the event that the value on the first roll is greater than or equal to the value on the second roll.

      iii. Find the set B corresponding to the event that the first roll is a six.

      iv. Let C correspond to the event that the first valued rolled and the second value rolled differ by two. Find $A \cap C$.

   Note that A,B and C should be subsets of the sample space specified in Part(a).

   (b) Consider two urns, one containing 5 red and 10 white balls, and the other with 5 white and 10 red balls. Now choose one of the urns randomly, and take two random balls from the chosen urn. Let A be the event that the first ball is red. and B be the event that the second ball is white. Are A and B independent?

2. **Probability in Medical science:** [8 points]

   (a) A new medical test has been designed to detect the presence of the mysterious Brain lesserian disease. Among those who have the disease, the probability that the disease will be detected by the new test is 0.86 (sensitivity). However, the probability that the test will erroneously indicate the presence of the disease in those who do not actually have it is 0.08 (1-specificity). It is estimated that 16% of the population has the disease.

      i. If the test administered to an individual is positive, what is the probability that the person actually has the disease?
      Now suppose that the same individual was born in the exotic island of Nobrain, close to Antarctica, where 60% of the population is estimated to have the disease.

      ii. Would the probability that the person actually has the disease change in any way? If so, how?

   (b) In a Japanese cohort study, 5,322 male non-smokers and 7,019 male smokers were followed for four years. Of these men, 16 non-smokers and 77 smokers developed lung cancer.

      i. What is the probability that a randomly chosen non-smoker from this group developed lung cancer?

      ii. What is the probability that a randomly chosen smoker from this group developed lung cancer?

      iii. Are the events "smoking" and "lung cancer" in this example independent?

      iv. What is the conditional probability that the patient is a smoker if he has developed lung cancer?

3. **Simpson's Paradox** [6 points]

A researcher wants to determine the relative efficacies of two drugs. The results efficacies of two drugs. The results (differentiated between men and women) were as follows.

women	drug I	drug II
success	200	10
failure	1800	190

men	drug I	drug II
success	19	1000
failure	1	1000

We are now faced with the question which drug is better. Here are two possible answers:

- Drug I was given to 2020 people, of whom 219 were cured. Drug II was given to 2200 people, of whom 1010 were cured. Therefore, drug II is much better.
- Amongst women, the success rate of drug I is 1/10, and for drug II the success rate is 1/20. Amongst men, these rates are 19/20 and 1/2 respectively. In both cases, that is, for both men and women, drug I wins, and is therefore better.

Which of the two answers do you believe? Can you explain the paradox?

4. **Probability in sports:** (Do both the following problems.) [5+5=10 points]

   (a) Based on the results of the 2018 NFL season, the probability that a home team scores at least 30 points is 0.312, the probability that a visiting team scores at least 30 points is 0.219, and the probability that both teams score at least 30 points is 0.105.

      i. Suppose that, in a given game, the home team scores at least 30 points. What is the probability that the visiting team also scores at least 30.

      ii. Suppose the visiting team scores at least 30 points. What is the probability that the home team scores at least 30 points?

      iii. Are the events "the home team scores at least 30 points" and "the visiting team scores at least 30 points" independent?

   (b) Consider MLB games in the 1973 to 2017 seasons. The home team won 54.0% of these games and 15.4% of the games were tied after 5 innings. 14.9% of the home team wins came in games that were tied after 5 innings, while 15.9% of the visiting team wins came in games that were tied after 5 innings. These values were taken from the win probability calculator at Click This

   Based on these results, is it correct to conclude that if a game is tied after 5 innings, then the visiting team is more likely to win the game than is the home team? Why or why not?

5. **Simulation based problem:** [4 points]

Consider the following analysis of bus ridership, which (in more complex form) could be used by the bus company/agency to plan the number of buses, frequency of stops and so on. Again, in order to keep things easy, it will be quite oversimplified, but the principles will be clear. Here is the model:

- At each stop, each passenger alights from the bus, independently of the actions of others, with probability 0.2 each.

- Either 0,1 or 2 new passengers get on the bus, with probabilities 0.5, 0.4 and 0.1, respectively. Passengers at successive stops act independently.
- Assume the bus is so large that it never becomes full, so the new passengers can always board.
- Suppose the bus is empty when it arrives at its first stop.

Here and throughout, it will be greatly helpful to first name the quantities or events involved. Let L_i denote the number of passengers on the bus as it leaves its i^{th} stop, $i = 1, 2, 3, \ldots$ Let B_i denote the number of new passengers who board the bus at the i^{th} stop.

Find the probability that after visiting the tenth stop, the bus is empty.