MA52112: Statsistics for Data Science

Exercise sheet 1 (Probability)

1. Recall that the set difference of two sets A and B is defined as:

$$B \setminus A = \{x \in B : x \notin A\} = B \cap A^c$$

   Show that for any events A and B, the following holds:

$$P(B \setminus A) = P(B) - P(A \cap B)$$

2. Show that for any events A and B, the following holds:

$$P(B \setminus A) = P(A \cup B) - P(A)$$

3. Show that for any events A and B such that $A \subseteq B$ we have $P(A) \leq P(B)$.

4. Show that if two events A and B are independent, then their complements A^c and B^c are also independent.

5. Three events A, B, and C are said to be mutually independent if all the following conditions hold:

   - $P(A \cap B) = P(A)P(B)$
   - $P(A \cap C) = P(A)P(C)$
   - $P(B \cap C) = P(B)P(C)$
   - $P(A \cap B \cap C) = P(A)P(B)P(C)$

   Consider the experiment of tossing two dice. Assume that each outcome is equally likely. Define the following events:

   - A: The two throws give the same result
   - B: the sum of the numbers is between 7 and 10 (inclusive)
   - C: the sum is 2 or 7 or 8

   Show that the events A, B, and C are not mutually independent.

6. Consider the experiment of tossing three coins. Assume that each outcome is equally likely.Define the following events:

   - A: The first coin is a head
   - B: The second coin is a head
   - C: The third coin is a head

   Show that the events A, B, and C are mutually independent.

7. Consider the events A_1, A_2, A_3 and A_4, defined on the sample space Ω. Show that $p(A_1, A_2, A_3, A_4) = p(A_4 \mid A_3, A_2, A_1)\, p(A_3 \mid A_2, A_1)\, p(A_2 \mid A_1)\, p(A_1)$. This is the product rule for four events. If we further assume that A_1 and A_2 are independent. How does the expression simplify?

8. Again consider the experiment of tossing three coins. Assume that each outcome is equally likely. Define the random variable X as the number of heads obtained. Find the probability mass function of X as well as its cumulative distribution function.

9. Consider now the experiment that consists in tossing a coin until we get a head. Assume that the coin is biased and that the probability of getting a head is 0.6. Define the random variable X as the number of tosses before we get the first head. Find the probability mass function of X.

10. Prove that the following functions are CDFs of some random variables. Find the PDF or PMF in each case.

    a)
    $$F(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)\,, \quad x \in \mathbb{R}$$

    b)
    $$F(x) = \frac{1}{(1+e^{-x})}\,, \quad x \in \mathbb{R}$$

    c)
    $$F(x) = e^{-e^{-x}}\,, \quad x \in \mathbb{R}$$

    d)
    $$F(x) = \begin{cases} 0 & x < 0 \\ 1/3 & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

11. Some advanced large computer simulations are sensitive to initialisation. It is common that these simulations do not converge every time they are run, and no result is obtained for the specific run. For a specific type of simulation, the probability of convergence (and results obtained for the run) is 0.8. What is the probability results are obtained for (exactly) 10 runs, if 15 runs are performed. We can assume the runs are independent. Give your answer with 2 decimal points.

12. The continuous random variable X is uniformly distributed between -2 and 12. What is $p(1 < X < 6)$? Give your answer with 2 decimal points.

13. The continuous random variable X has the following probability density function (PDF):

$$f(x) = \begin{cases} kx^2 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

    a) Find the value of k.
    b) Find the cumulative distribution function (CDF) of X.
    c) What is $p(1 < X < 2)$? Give your answer with 2 decimal points.

14. A fashionable (and by some considered risk-prone) Data Science student has bought a new pair of suede shoes. Suede gets ruined by water, and hence by rain. If we assume the probability of rain on any day is 0.4 in England, where the student is pursuing the degree, what is the probability the shoes get ruined the third time they are worn (third time unlucky)? We can assume the student does not check the weather before putting on the shoes, and that the probability of rain is the same every day and independent on other days. Give your answer with 2 decimal points.

15. IVF is the treatment where eggs are fertilised with sperm in a laboratory, and the fertilised egg (embryo) is returned to the woman's womb. It is a treatment used by couples who are unable to get pregnant naturally or by women who wish to have children on their own. The probability an embryo survives in the womb, resulting in a successful pregnancy, varies depending on several factors. To increase the chance of a pregnancy, sometimes several embryos are inserted in the woman's womb simultaneously.

    a) If three embryos are inserted during a treatment, and if the probability an embryo survives is $p = 0.12$, what is the probability of triplets (all embryos survive)? What is the probability of twins (two embryos survive)? What is the probability exactly one embryo survives? What is the probability of no pregnancy (no embryos survive)? Assume the embryos are independent of each other. Give your answers with 2 decimal points.
    b) What is the probability of a pregnancy (one baby, twins or triplets) in the above treatment? Give your answer with 2 decimal points.
    c) If a treatment is not successful, the couple or woman can choose to undergo the treatment again. The price of a treatment can be £5,000 or more. Some clinics offer a discount if a fixed number of treatments are bought upfront. Using your obtained

probability from part b), what is the probability a pregnancy occurs in the first try? Within two tries? Within three tries?

16. Glowing seahorses are a rare, and sought after, sight by water wildlife explorers. Situated on an island in the Pacific Ocean, a tourist wildlife park offers supervised night group dives to visiting explorers. The dives take place in a specific region of the ocean, where the glowing seahorses can be found.

    Table 1 shows data collected for number of glowing seahorses spotted in the specific region of the ocean. The data collection took place during 8 (night) hours every day for one week.

    When signing up for a dive, the explorer gets to choose between a 30 minute dive, a one hour dive, a three hour dive or a four hour dive. The price for a 30 minute dive (per person) is £50, a one hour dive £100, a three hour dive £250, and a four hour dive £350.

    Table 1: Number of seahorses spotted

Day	Number
1	5
2	6
3	4
4	3
5	8
6	12
7	7

    We wish to model the probability of number of glowing seahorses spotted for a specific time interval. For the questions below, use the Poisson distribution, and estimate the intensity parameter using the data from the table.

    a) What assumptions do we make regarding the occurrence of seahorses when using this distribution?
    b) What is the probability estimate of seeing exactly two glowing seahorses during a one hour dive? Give your answer with 2 decimal points.
    c) What is the probability estimate of seeing two or more glowing seahorses during a one hour dive? Give your answer with 2 decimal points.
    d) What is the probability estimate of seeing exactly two glowing seahorses during a three hour dive? Give your answer with 2 decimal points.
    e) A nature passionate data science student is very excited to see the glowing seahorse, but is also on a tight holiday budget. The student decides to go for the cheapest option (length of diving chosen) while having the probability of at least 0.9 (at least a 90% chance) of seeing at least one glowing seahorse. What length of diving shall the student pay for?

f) What assumption do we make regarding the data when choosing to model the probability in this way? Give two examples of when this assumption would not be fulfilled.