Example sheet 0

Remembering IA Maths for NST

Foundations of Data Science—DJW—2019/2020

Foundations of Data Science builds on the probability theory you learnt in IA *Maths for the Natural Sciences Tripos*. All of the questions below (apart from the last two) are taken from that course. Please look through and make sure you can still answer them! Solutions will be provided.

For supervisors: it isn't intended that you supervise this example sheet.

Question 1. A card is drawn at random from a pack. Event A is 'the card is an ace', event B is 'the card is a spade', event C is 'the card is either an ace, or a king, or a queen, or a jack, or a 10'. Compute the probability that the card has (i) one of these properties, (ii) all of these properties.

Question 2. A biased die has probabilities p, $2p$, $3p$, $4p$, $5p$, $6p$ of throwing 1, 2, 3, 4, 5, 6 respectively. Find p. What is the probability of throwing an even number?

Question 3. Consider drawing 2 balls out of a bag of 5 balls: 1 red, 2 green, 2 blue. What is the probability of the second ball drawn from the bag being blue given that the first ball was blue if (i) the first ball is replaced, (ii) the first ball is not replaced?

Question 4. Two cards are drawn from a deck of cards. What is the probability of drawing two queens, given that the first card is not replaced?

Question 5. A screening test is 99% effective in detecting a certain disease when a person has the disease. The test yields a 'false positive' for 0.5% of healthy persons tested. Suppose 0.2% of the population has the disease. (i) What is the probability that a person whose test is positive has the disease? (ii) What is the probability that a person whose test is negative actually has the disease after all?

Question 6. What is the probability that in a room of r people at least two have the same birthday?

Question 7. Out of 10 physics professors and 12 chemistry professors, a committee of 5 people must be chosen in which each subject has at least 2 representatives. In how many ways can this be done?

Question 8. What is the probability of throwing exactly 3 heads out of 6 tosses of a fair coin? How about at least one head?

Question 9. A bag contains 6 blue balls and 4 red balls. Three balls are drawn from the bag without replacement. Let X be the number of these three balls that are red. Find the density function $f(x) = \mathbb{P}(X = x)$.

Question 10. Let X be a random variable. Show that $\operatorname{Var} X = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2$.

Note: by convention, $\mathbb{E}$ is taken to have lower precedence than multiplication and power, and higher precedence than addition and subtraction. So the expression of interest is $(\mathbb{E}(X^2)) - (\mathbb{E}(X))^2$.

Question 11. Find the mean and variance of the Exponential distribution, which has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that X takes a value in excess of two standard deviations from the mean?

Question 12. Players A and B roll a six-sided die in turn. If a player rolls 1 or 2 that player wins and the game ends; if a player rolls 3 the other player wins and the game ends; otherwise the turn passes to the other player. A has the first roll. What is the probability (i) that B gets a first throw and wins on it? (ii) that A wins before A's second throw? (iii) that A wins, if the game is played until there is a winner?

Question 13. A coal bunker is to be constructed on the side of a house. Assuming that it is a cuboid of given volume V, find the shape that minimizes the external surface area.

Question 14. *More examples of finding stationary points. These two specific cases crop up again and again in data science. You should be able to solve them blindfolded.*

(a) Find the value of $p \in [0, 1]$ that maximizes $p^a(1-p)^b$, where a and b are both positive.

(b) Find the value of $p \in [0, 1]$ that maximizes $\log\bigl(p^a(1-p)^b\bigr)$, where a and b are both positive.

(c) Find the values of $\mu \in \mathbb{R}$ and $\sigma > 0$ that jointly maximize

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Bigl(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\Bigr)$$

(d) Find the values of $\mu \in \mathbb{R}$ and $\rho > 0$ that jointly maximize

$$\frac{1}{\sqrt{2\pi\rho}} \exp\Bigl(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\rho}\Bigr)$$

What do you notice about the solutions to (a) *versus* (b), and about the solutions to (c) *versus* (d)?

Question 15. Bayes's rule says that, for any events A and B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\,\mathbb{P}(B \mid A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\,\mathbb{P}(B|A)}{\mathbb{P}(A)\,\mathbb{P}(B|A) + \mathbb{P}(\neg A)\,\mathbb{P}(B|\neg A)}.$$

There are four core definitions and laws in probability theory. Derive Bayes's rule from them, and explain carefully which of the four you are using at each step.

(a) $\mathbb{P}(\Omega) = 1$ where Ω is the entire sample space

(b) Conditional probability: $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\,\mathbb{P}(B)$, when $\mathbb{P}(B) > 0$

(c) Sum rule: If $\{B_1, B_2, \dots\}$ partition Ω then $\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i)$
Law of total probability: $\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)$

(d) A and B are said to be *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$

Note: "$\{B_1, B_2, \dots\}$ partition Ω" means the B_i are mutually exclusive and $\bigcup_i B_i = \Omega$.