STATG006: Exercise Sheet #1

The exercises in this sheet have two main motivations: first, as a sample of questions on probability that are an incentive for further revision, if necessary (but please see the suggested exercises from Rice for further, if more advanced, questions). The second type of questions concern statistical reasoning, its different points of view and communication. We start with these questions.

1. Statistical inference is much more than quantitative modelling and computation. In what follows, I raise questions on the interpretation of data, models and the consequences. For this part of the exercise, I strongly suggest that you submit your own answers to a (hopefully clearly indicated) topic in our discussion board "Questions and Answers", with links to external sources encouraged. Examples posted by myself can be found in the corresponding topics.

   (a) Find examples in the news, blogs, or textbooks, of statistical estimates of a same quantity that disagree among themselves, even though the different analyses are seemingly defensible. Provide your own views concerning the reasons for that to happen, and what consequences it may have.

   (b) Find examples of graphs that display data in a misleading or ambiguous way. Discuss how you would present them otherwise. (Students in Statistical Science will get in more in depth about this in *STATG099*).

   (c) This may be a bit harder than the previous questions as it starts to get more technical: find an example of a generally sound statistical method that was misleading due to a lack of understanding from the data scientists of its inadequacy for a particular problem.

   (d) As above, this might take some effort: find an example of a scientific discovery whose significance may have been presented in an exaggerated way due to overinterpretation of statistical methods.

   (e) Provide your views on one or more papers suggested as Overview Articles in the Moodle page (Cox and/or Donoho and/or Breiman).

2. Solve these assorted questions concerning probability. Consult the notes of *STAT1005* if you want a refresher for some of the probability models mentioned below.

   (a) Two fair dice are thrown. Let X be the smallest of the two numbers obtained (or the common value if the same number is obtained on both dice). Find the probability mass function of X. Find $P(X > 3)$.

   (b) Let X be a random variable with expectation μ and variance σ^2. Find the expectation and variance of the random variable $Y = (X - \mu)/\sigma$.

   (c) On a coral reef, S species of fish are present in proportions $p_1, \ldots, p_S$. A biologist wishes to take a sample of the fish, and wants to know how many species of fish she should expect to find in a sample of a given size.

      (i) Suppose a sample of size n is taken. Let X_i $(i = 1, \ldots, S)$ be a random variable taking the value 1 if species i occurs, 0 otherwise. Find an expression for $E(X_i)$ (assume that, first, the sample is small relative to the population of any fish species, so that taking the sample has a negligible effect on the proportions of fish remaining; second, species are distributed randomly so that successive fish in the sample can be regarded as independent draws from the population).

      (ii) Now let Y be the number of fish species present in the sample. Express Y in terms of the X_i's, and deduce that the expected number of species is

$$S - \sum_{i=1}^{S} (1 - p_i)^n.$$

      Are you making any further assumptions in obtaining this result? Check that the formula gives the correct result for a couple of different sample sizes where the answer is "obvious".

   (d) For each case below, state whether the binomial distribution is suitable. If not, give your reasons; if it would, state the values of parameters n and p.

      (i) The number of sixes obtained in three successive throws of a fair die.

      (ii) The number of girls in the families of British prime ministers.

      (iii) The number of aces in a hand of four cards dealt from a standard pack of cards.

      (iv) The number of students in a class of 40 whose birthday falls on a Sunday this year.

      (v) The number of throws of a fair coin until the first head is obtained.

   (e) Consider a jury trial in which it takes 8 out of the 12 jurors to convict. That is, in order for the defendant to be convicted, at least 8 of the 12 jurors must vote him guilty. Assume that jurors act independently and each makes the right decision with probability p. Let α denote the probability that the defendant is guilty. What is the probability that the jury renders a correct decision?

(f) An exam paper consists of ten multiple choice questions, each offering four choices of which only one is correct. If a candidate chooses his answers completely at random, what is the probability that

(i) he gets at least 8 questions right,

(ii) the last of the ten questions is the eighth one he gets right,

(iii) in six such exams, he gets at least 8 questions right in at most one exam?

(g) If X is a geometric random variable with parameter p, show that $P(X = n + k \mid X > n) = P(X = k), k = 1, 2, 3, \ldots$. In the light of the interpretation of a geometric random variable in terms of independent Bernoulli trials, explain why this result is "obvious".

(h) Suppose that the number of distinct uranium deposits in a given area is a Poisson random variable with parameter $\mu = 10$. If, in a fixed period of time, each deposit is independently discovered with probability 1/50, find the probability that (i) exactly one, (ii) at least one and, (iii) at most one deposit is discovered during that time.

(i) The proportion of time during a 40-hour week that an industrial robot is in operation is modelled by a random variable X with probability density function

$$f(x) = \begin{cases} cx, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where c is a constant. Find c. Find $P(X < 12)$ and $P(X > 1/3 \mid X < 1/2)$.

(j) The probability density function of X, the lifetime in hours of a certain type of electronic device, is given by

$$f(x) = \begin{cases} 10/x^2, & \text{if } x > 10, \\ 0, & \text{otherwise.} \end{cases}$$

Find $P(X > 15)$. What is the probability that, out of 5 such devices, at least 4 will function for at least 15 hours? Assume that the life times of the devices are independent.