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. This question is very open ended. The following are just examples. More should be found in our "Question and Answering" forum.

   (a) In this 20/09/2016 New York Times article, "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results."[1], the same raw data was given to four different teams of data scientists. The data concerned voter preference in the 2016 US presidential election. The results were relatively close, but in a close race the difference has practical significance. In particular, one of the five estimates indicated a victory for Trump, while the other four indicate a victory for Clinton. The reason for the discrepancies are due to different ways of reweighing the raw data, as something as simple as the empirical average might not reflect the biases of the data collection process. The article does a good job of providing a light description of these different adjustments. We will not focus on these issues in *STATG006* itself, as they are better covered in *STATG002*.

   (b) Data visualisation is by itself a rich topic, and we cannot do justice to it in the restricted time of *STATG006*. To give a typical family of examples, the press may be particularly fond of pie charts, perhaps because they can be splashed with flashy colours. But pie charts are not a good way of communicating what is essentially "one dimensional data" (single numbers being compared across categories), as it uses two-dimensional visual information (areas of slices) to convey that. Some interesting discussion of misuses of pie charts is given by this article in Business Insider[2]. For an extreme misuse of pie charts, this is a classic example: `https://flowingdata.com/2009/11/26/fox-news-makes-the-best-pie-chart-ever/`.

   (c) There were many causes of the 2008 financial crises, but one that was highlighted in the press was the misuse of some statistical models for risk assessment

[1] `http://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-ab` `html?_r=0`

[2] `http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?IR=T`

based on the concept of "Gaussian copulas". We may have the chance of discussing the technical aspects of it later on, but the key message is that these models assert (in the context of finance) that the price of two assets will covary at a constant rate regardless of their values (that is, the variability of one asset will be the same regardless whether the other asset is at a typical price or at an unlikely low/high price). In contrast, reality would be that at more extreme values, asset prices tends to be more strongly associated. The implication is that this Gaussian copula would underestimate the probability of systematic failure: that is, many components of the system would be more likely to fail at the same time than the probability predicted by the model. This was not a failure of Gaussian copulas per se, which are used successfully in other domains such as spatial analysis, but the application of it to financial applications of risk management. A longer version of the story can be found in the following article in the Wired magazine[3]. As a reminder: popular accounts of scientific methods should also be taken with a grain of salt, as they oversimplify the implications of such methods – the situation was far too complexed to be explained just by a simple methodological failure (it also does not convey the fact that Gaussian copulas are simple, well-understood models with a very old history).

(d) The experiments Kramer et al. ("Experimental evidence of massive-scale emotional contagion through social networks", 2014, PNAS 111, 8788–8790) were in the news due to their unclear ethical implications, but let us not focus on that. The study experimented with Facebook users, manipulating the news on their news feed to contain more positive or negative stories, measuring how this affected the users' moods. Ethical implications aside, the questions asked can be considered relevant. What was not typically discussed in the media is that the effect of this manipulation was extremely small: a change of mood of the order of 0.07% (that's zero point seven percent, not 7 percent!).

2. For some questions, I will provide only a sketch of the solution. It is not my intention to provide any fully detailed solutions of any questions, but in many cases it will be fairly detailed.

   (a) If X is a random variable denoting the outcome of this process, it is clear the only values it can take are $1, 2, \dots, 6$. For each possible outcome of the two dice, $(1,1), (1,2), \dots, (6,6)$, we have 1/36 as the probability of it occurring (notice we are ordering the dice here, so $(1,2)$ and $(2,1)$ are two different outcomes). We only need to count each outcomes are mapped to $X = 1, 2, \dots, 6$ and multiply that by 1/36. The resulting pmf is

x	1	2	3	5	5	6
$P(X = x)$	11/36	9/36	7/36	5/36	3/36	1/36

[3] https://www.wired.com/2009/02/wp-quant/

(b) Let's attempt to answer a more general question here. What is the expectation of $aX + b$, where a and b are constants? Applying the definition

$$E[aX + b] = \int (ax + b)p(x)\ dx = a \int xp(x) + b \int p(x)\ dx = aE[X] + b$$

From that, it follows that $E[(X - u)/\sigma] = (\mu - \mu)/\sigma = 0$.
What about variance? We need to find

$$Var(aX + b) = \int (ax + b - E[aX + b])^2 p(x)\ dx = E[(aX + b)^2] - (E[aX + b])^2.$$

Do the calculation to show that $E[(aX + b)^2] = a^2E[X^2] + 2abE[X] + b^2$, and plug in the fact that $(E[aX + b])^2 = (aE[X] + b)^2$ to get to the conclusion that $E[(aX + b)^2] = a^2E[X^2]$. From there, conclude that $Var(aX + b) = a^2Var(X)$ so that $Var(Y) = 1$.

(c) The following concerns the coral reef question:

(i) The event corresponding to $X_i = 0$ is the event that species i is not in the sample. So $P(X_i = 0) = (1 - p_i)^n$. Its pmf can be written more explicitly as

x	0	1
$P(X = x)$	$(1 - p_i)^n$	$1 - (1 - p_i)^n$

$E[X_i]$ is just $0 \times p(0) + 1 \times p(1)$, that is, $E[X_i] = 1 - (1 - p_i)^n$.

(ii) It is clear that $Y = \sum_{i=1}^{S} X_i$. So $E[Y] = \sum_{i=1}^{S} E[X_i]$ without any further assumptions (see the reasoning of item (b) above), and the result follows immediately. "Obvious" cases: check what happens when $n = 1$ and its limit when n grows to ∞, that is $\lim_{n\to\infty}(S - \sum_{i=1}^{S}(1 - p_i)^n)$.

(d) The following concerns the question on the suitability of the binomial:

(i) Yes, $n = 3$ and $p = 1/6$.

(ii) No, n is not a constant.

(iii) No, p is not a constant.

(iv) Yes, $n = 40, p \approx 1/7$.

(v) No, n is not fixed.

(e) This is a case where it might be more straightforward to think in terms of events instead of random variables (but this doesn't matter that much: we could think of Bernoulli variables encoding the same information). Let X be the number of jurors who return a guilty verdict. Let E denote the event "correct decision is returned" and G the event "defendant is guilty". Then

$$P(E) = P(E \mid G)P(G) + P(E \mid \neg G)P(\neg G)$$

which is just an application of the Law of Total Probability (here "$\neg A$" means the complementary event "A did not happen"). We are given that $P(G) = \alpha, P(\neg G) = 1 - \alpha$.

Now comes the interesting part. $P(E \mid G)$ is the same as $P(X \geq 8 \mid G)$. If we denote by p the probability of each juror (independently0 returning a guilty verdict, then $X \sim Bin(12, p)$. Therefore,

$$P(E \mid G) = \sum_{k=8}^{12} \binom{12}{k} p^k (1-p)^{12-k}.$$

An analogous reasoning for $P(E \mid \neg G)$ gives

$$P(E \mid \neg G) = \sum_{k=0}^{7} \binom{12}{k} (1-p)^k p^{12-k},$$

so that

$$P(E) = \alpha \sum_{k=8}^{12} \binom{12}{k} p^k (1-p)^{12-k} + (1-\alpha) \sum_{k=0}^{7} \binom{12}{k} (1-p)^k p^{12-k}.$$

(f) The following refers to the ten multiple choice problem. We will denote the number of correct answers as X, so that $X \sim Bin(10, 1/4)$.

(i) $P(X \geq 8) \approx 0.0004$.

(ii) $P(\text{“7 right out of 9, then 1 right”}) = \binom{9}{7}(1/4)^7(3/4)^2(1/4) \approx 3.09 \times 10^{-4}$.

(iii) Let Y be the number of exams where candidate gets at least 8 questions right, so that $Y \sim Bin(6, p)$ where $p = 0.0004$. Therefore,

$$P(Y \leq 1) = \binom{6}{0}(1-p)^6 + \binom{6}{1}(1-p)^5 p \approx 0.9999974.$$

(g) $P(X = n + k \mid X > n) = P(X = n + k \text{and} X > n)/P(X > n) = P(X = n + k)/P(X > n)$ for $k \leq 1$. This gives $q^{n+k-1}p/q^n = q^{k-1}p$, which is equal to $P(X = k)$.

The result is “obvious” as the geometric distribution arises as the number of independent trials until a “success” is observed: the fact we've wait for n trials tells us nothing about the number of trials until success, as the trials are independent.

(h) Let N be the number of deposits in the area, so that $N \sim Poisson(10)$. Let Y be the number of deposits discovered. If $N = n$, $Y \sim Bin(n, 1/50)$. Hence for all integers $k \geq 0$, by the Law of Total Probability,

$$\begin{aligned}
P(Y = k) &= \textstyle\sum_{n=0}^{\infty} P(Y = k \mid N = n)P(N = n) = \sum_{n=k}^{\infty} P(Y = k \mid N = n)P(N = n) \\
&= \textstyle\sum_{n=k}^{\infty} \binom{n}{k} (1/50)^k (49/50)^{n-k} \times e^{-10} 10^n / n! \\
&= \frac{(1/5)^k e^{-10}}{k!} \textstyle\sum_{n=k}^{\infty} \frac{(49/5)^{n-k}}{(n-k)!} \\
&= \frac{(1/5)^k e^{-10}}{k!} \textstyle\sum_{n=0}^{\infty} \frac{(49/5)^n}{n!} = \frac{(1/5)^k e^{-1/5}}{k!}.
\end{aligned}$$

Notice this is the pmf of a Poisson with parameter 1/5, so $Y \sim Poisson(1/5)$. We now have (a) $P(Y = 1) = 0.164$, (b) $P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-1/5} = 0.181$ and (c) $P(Y \leq 1) = 0.983$.

(i) We must have

$$\int_{-\infty}^{\infty} p(x)\ dx = 1,$$

so $\int_0^1 cx\ dx - [cx^2/2]_0^1 = c/2 = 1$, so $c = 2$.

$P(X < 1/2) = \int_0^{1/2} 2x\ dx = [x^2]_0^{1/2} = 1/4$.

$P(X > 1/3 \mid X < 1/2) = P(1/3 < X < 1/2)/P(X < 1/2) = 4 \int_{1/3}^{1/2} 2x\ dx = 4(1/3 - 1/9) = 5/9$.

(j) $P(X > 15) = \int_{15}^{\infty} 10x^{-2}\ dx = [-10x^{-1}]_{15}^{\infty} = 2/3$. Let Y be the number of devices working after 15 hours; then $Y \sim Bin(5, 2/3)$. So $P(Y \geq 4) = \binom{5}{4}(2/3)^4 * 1/3) + \binom{5}{5}(2/3)^5 \approx 0.4609$.