ST102 Outline solutions to Exercise 11

1. (a) The sum of n independent and identically distributed $\text{Pois}(\lambda)$ random variables follows the $\text{Pois}(n\lambda)$ distribution (see page 136 of the Michaelmas term lecture notes).

   (b) Since $\sum_{i=1}^{n} X_i$ has possible values $0, 1, 2, \ldots$, the possible values of $\bar{X} = \sum_{i=1}^{n} X_i/n$ are:

   $$0, \quad \frac{1}{n}, \quad \frac{2}{n}, \quad \ldots$$

   which note is just a rescaling of the values of $\sum_{i=1}^{n} X_i$ by dividing by n.
   The probabilities of these values are determined by the probabilities of the values of $\sum_{i=1}^{n} X_i$, which are obtained from the $\text{Pois}(n\lambda)$ distribution. Therefore, the probability function of $\bar{X}$ is:

   $$p_{\bar{X}}(\bar{x}) = \begin{cases} \dfrac{e^{-n\lambda}(n\lambda)^{n\bar{x}}}{(n\bar{x})!} & \text{for } \bar{x} = 0, 1/n, 2/n, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

   (c) For $X_i \sim \text{Pois}(\lambda)$ we have $\text{E}(X_i) = \text{Var}(X_i) = \lambda$, so the general results for $\bar{X}$ give:

   $$\text{E}(\bar{X}) = \lambda \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\lambda}{n}.$$

   When $\lambda = 5$ and $n = 100$, $\text{E}(\bar{X}) = 5$ and $\text{Var}(\bar{X}) = 5/100 = 0.05$.

   (d) Due to the randomness in the simulation (sample selection), every person doing this exercise will get slightly different results. However, the mean and variance of the 10 simulated sample means would not be expected to be very far from 5 and 0.05, respectively, even though we are simulating only 10 random samples here. For example, the $\bar{x}$s in my 10 random samples had a mean of 4.935 and a variance of 0.057. If we repeated the exercise with many more than 10 random samples, these figures would approach the theoretical values of 5 and 0.05, respectively.

2. Let X_i denote the weight of the ith object, for $i = 1, 2, \ldots, 200$. The X_is are assumed to be independent and identically distributed with $\text{E}(X_i) = 1$ and $\text{Var}(X_i) = (0.03)^2$. We require:

   $$P\left(\sum_{i=1}^{200} X_i > 200.5\right) = P\left(\sum_{i=1}^{200} \frac{X_i}{200} > 1.0025\right) = P(\bar{X} > 1.0025).$$

   If the weights are not normally distributed, then by the central limit theorem (CLT):

   $$P(\bar{X} > 1.0025) \approx P\left(Z > \frac{1.0025 - 1}{0.03/\sqrt{200}}\right) = P(Z > 1.18) = 0.1190.$$

   If the weights are normally distributed, then this is the *exact* (rather than an *approximate*) probability.

3. (a) Here the individual responses, the X_is, follow a Bernoulli distribution with probability parameter $\pi = 0.49$. The mean and variance of this distribution are:

$$\mathrm{E}(X) = \pi = 0.49 \quad \text{and} \quad \mathrm{Var}(X) = \pi(1-\pi) = 0.49 \times 0.51 = 0.2499.$$

Therefore, the CLT approximation of the sampling distribution of $\bar{X}$ is:

$$\bar{X} \sim N\left(0.49, \frac{0.49 \times 0.51}{n}\right) = N\left(0.49, \left(\frac{0.4999}{\sqrt{n}}\right)^2\right).$$

(b) When $n = 50$, the CLT approximation from (a) is $\bar{X} \sim N(0.49, (0.0707)^2)$. With this, we get:

$$P(\bar{X} > 0.50) = P\left(\frac{\bar{X} - 0.49}{0.0707} > \frac{0.50 - 0.49}{0.0707}\right) = P(Z > 0.14) = 0.4443$$

using Table 3 of Murdoch and Barnes' *Statistical Tables.*

(c) Here we need the smallest integer n such that:

$$P(\bar{X} > 0.50) = P\left(\frac{\bar{X} - 0.49}{0.4999/\sqrt{n}} > \frac{0.50 - 0.49}{0.4999/\sqrt{n}}\right) = P(Z > 0.0200\sqrt{n}) < 0.01.$$

According to Table 3, the smallest z such that $P(Z > z) < 0.01$ is $z = 2.33$. Hence we need:

$$0.0200\sqrt{n} \geq 2.33 \quad \Leftrightarrow \quad n \geq \left(\frac{2.33}{0.0200}\right)^2 = 13{,}572.25$$

which means that we need at least $n = 13{,}573$ likely voters in the sample – which is a very large sample size! Of course, the reason for this is that the population of likely voters is almost equally split between those supporting leaving the European Union, and those opposing. Hence such a large sample size is necessary to be sufficiently confident of obtaining a representative sample.

4.* (a) The probability that a single randomly-selected observation of X is at most y is $P(X_i \leq y) = F_X(y)$. Since the X_is are independent, the probability that they are all at most y is:

$$F_Y(y) = P(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) = \prod_{i=1}^{n} P(X_i \leq y) = (F_X(y))^n.$$

(b) The pdf is the first derivative of the cdf, so:

$$f_Y(y) = F_Y'(y) = n(F_X(y))^{n-1} f_X(y)$$

since $f_X(x) = F_X'(x)$.

5. Here $X_i \sim N(174.9, (7.39)^2)$. Therefore:

$$F_X(192) = P(X \leq 192) = P\left(Z \leq \frac{192 - 174.9}{7.39}\right) \approx P(Z \leq 2.31)$$

where $Z \sim N(0,1)$. According to Table 3:

$$P(Z \leq 2.31) = 1 - 0.01044 = 0.98956.$$

Therefore, the probability we need is:

$$P(Y > 192) = 1 - P(Y \leq 192) = 1 - (F_X(192))^{60} = 1 - (0.98956)^{60} = 0.4672.$$