EXERCISE SET 2

1. Exercises

Exercise 1 (Computing conditional averages)**:** Consider a random variable $Y = (Y_1, Y_2)$ which takes values in the sample space

$$S = \mathbb{N} \times \mathbb{N} = \{(i,j) : i,j \in \mathbb{N}\}$$

That is, the sample space consists of all possible pairs of numbers (i,j). Now suppose we have some data:

$$\{(1,2),(1,2),(3,1),(1,4),(3,3),(2,2),(1,5)\}$$

Give you best estimates of the following (either by hand, with Python, or a calculator)

(a) $E[Y_1]$
(b) $E[Y_1|Y_2 = 2]$
(c) $E[Y_2|Y_1 = 1]$
(d) $E[Y_2|Y_1 > 1]$

Exercise 2 (Independence and conditional expectation)**:** Let X and Y be two random variables with (discrete) sample spaces S_X and S_Y.

(a) Prove that if X and Y are independent $E[X|Y = y] = E[X]$ and $E[Y|X = x] = E[Y]$ for all $x \in S_X$ and $y \in S_Y$. You may assume S_X and S_Y are countable (i.e. discrete) spaces.
(b) Prove the tower property of expectation that is stated in the class notes.
(c) Show that if X and Y are independent, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

Hint: use the formula

$$\text{var}(X) = E[X^2] - E[X]^2$$

Exercise 3 (Aspects of the binomial distribution)**:** Suppose Y_1 and Y_2 are two independent binomial distributions:

$$Y_1 \sim \text{Binomial}(N_1, p_1)$$
$$Y_2 \sim \text{Binomial}(N_2, p_2)$$

with $p_1, p_2 \in (0,1)$ and $N_1, N_2 \in \mathbb{N}$.

(a) If $p = p_1 = p_2$ then what is the distribution of $Y_1 + Y_2$? Explain your reasoning.
(b) Confirm you answer to part (a) with simulations with $N_1 = 100$, $N_2 = 10$ and $p = 0.3$.
(c) Now suppose $p_1 \neq p_2$. Let

$$Y_3 \sim \text{Binomial}\left(N_1 + N_2, \frac{N_1}{N_1 + N_2} p_1 + \frac{N_2}{N_1 + N_2} p_2\right)$$

Here is an **erroneous** argument for why Y_3 might have the same distribution as $Y_1 + Y_2$ (it doesn't!):

> $Y_1 + Y_2$ is the sum of N Bernoulli random variables. Denote these as $X_1, X_2, \dots, X_N$ where $N = N_1 + N_2$. Assume they are in order, so that the first N_1 terms are the Bernoulli random variables corresponding to the first binomial distribution (the one with success probability p_1). Note that with this notation we are not specifying whether X_i comes from the first or second Bernoulli sequence. If we randomly select one of these, X_i, then the chance it is equal to 1 is
>
> $$P(X_i = 1) = P(X_i = 1|i \leq N_1)P(i \leq N_1) + P(X_i = 1|i > N_1)(i > N_1).$$
>
> Now observe that
>
> $$P(i \leq N_1) = N_1/(N_1 + N_2)$$
> $$P(i > N_1) = N_2/(N_1 + N_2)$$
> $$P(U_i = 1|i \leq N_1) = p_1$$
> $$P(U_i = 1|i > N_1) = p_2.$$
>
> Plugging these into the formula for $P(X_i = 1)$ gives the probability of success in the definition of Y_3.

(d) Explain why this argument above is flawed. Hint: Is the variable X_i and X_j independent for all $i \neq j$? If not, why is independence important?
(e) Confirm that the argument above is incorrect using simulations; that is, confirm via simulations of an example that Y_3 does not have the same probability distribution as $Y_1 + Y_2$. You can do this many ways, for example, by plotting $P(Y_3 > k)$ as a function of k and comparing to $P(Y_1 + Y_2 > k)$.

Exercise 4 (Election modeling)**:** Suppose again that we are interested in predicting the outcome of an election with two candidates and N voters. Based on or polling data, people's preferences are equally split between the two candidates ($q = 1/2$). However, there is one particular person – person 1 – who is particularly influential. If person 1 votes for candidate one, then everyone else votes for candidate one, while if person 1 votes for candidate 2, everyone sticks with their original preference.

The vote total for candidate 1 can be written as

$$Y = \sum_{i=1}^{N} y_i y_1$$

where

$$Y_i \sim \text{Bernoulli}(1/2), \quad i = 1, \dots, N$$

Each Y_i is 0 if they vote for candidate 1 and 1 if they vote for candidate 2. Y_1 represents the vote of the very influential person The code below simulates this model.

```
> def sampleY(N,n_samples):
>   y = np.zeros(n_samples)
>   for i in range(n_samples):
>     ys = np.random.choice([0,1],N)
>     y[i] = np.sum(ys)*ys[0]
>   return y
```

(a) Let ϕ denote the fraction of votes for candidates 1. How do you think the CV (coefficient of variation) of ϕ depends on N as N becomes large? Test you hypothesis with simulations.

(b) What are $E[\phi]$ and $E[\phi|Y_1 = 0]$. Confirm your answers with simulations of the model.