EXERCISE SET 2

1. Exercises

Exercise 1 (Computing conditional averages)**:** Suppose we have some data representing samples of a pair of random variables (Y_1, Y_2):

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

Compute the following (either by hand, with Python, or a calculator)

(a) $\mathbb{E}[Y_1]$

(b) $\mathbb{E}[Y_1|Y_2 = 2]$

(c) $\mathbb{E}[Y_2|Y_1 = 1]$

(d) $\mathbb{E}[Y_2|Y_1 > 1]$

Exercise 2: Do Exercises 3.1.3, 3.1.4, 3.1.10, 3.1.14 in [1] and for each one check your answer using simulations.

Exercise 3 (Independence and conditional expectation)**:** Let X and Y be two random variables with (discrete) sample spaces S_X and S_Y. (you can find these in the textbook, but give them a try yourself first).

(a) Show that if X and Y are independent $\mathbb{E}[X|Y = y] = \mathbb{E}[X]$ and $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$ for all $x \in S_X$ and $y \in S_Y$. You may assume S_X and S_Y have a finite number of elements, e.g. $S_X = \{1, 2, 3, 4\}$.

(b) Prove the tower property of expectation, which says that

$$\mathbb{E}[X] = \sum_{x \in S_Y} \mathbb{E}[X|Y = y]P(Y = y)$$

This is sometimes stated as $\mathbb{E}[\mathbb{E}[X|Y]]$ where the inner expectation is interpreted as a random variable depending on Y.

(c) Show that if X and Y are independent, then

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

Exercise 4 (A first taste of hypothesis testing)**:** Last spring, 1425 out of 2748 of student voted "I have no confidence in President Beilock's leadership", which comes out to 51.86%. Let's try to understand if this small margin could be a reflection of randomness in who voted rather than a true reflection of the student body's preference. To answer this, we make the *null hypothesis* that reality exactly 1/2 of the 6300 students at Dartmouth support the vote of no confidence. We can then view the 2748 students who voted as random samples from this pool of students. Under the null hypothesis, how likely is the vote margin to be at large as it was in reality? You may ignore the chance that we could in principle sample the same student twice and you can either use simulations or formulas.

Exercise 5 (Conditioning with continuous variables)**:** Let

$$Z_1 \sim \text{Normal}(0, 1)$$
$$Z_1 \sim \text{Normal}(1, 2)$$

Compute each of the following using Python

(a) $P(Z_1 + Z_2 > 3)$

(b) $P(Z_1 + Z_2 > 3 | Z_1 < -1)$

(c) $P(Z_2 Z_2 > 0 | Z_1 + Z_2 < 4)$

(d) Suppose we have a model of hemoglobin levels for men as

$$Z \sim \text{Normal}(15.8, 1.4)$$

(these numbers are in the ballpark but please don't try to diagnose your anemia based on this problem). Some has Polycythemia if $Z > 17.1$. Given that someone has does not have Polycythemia, what is the chance that they are anemic

Exercise 6: Do Exercise 2.4.2 in [1] using simulations. You can also check your answer using calculus if you wish.

Exercise 7: The **random walk** is a foundational model in nearly every area of science. It describes the "motion" of a variable which moves randomly over time without any memory of its past. Einstein developed a theory of the motion of microscopic particles based on random walks and they have been used as rudimentary models of stock prices.

We can define a random walk as follows. Let $X_0 = 0$ and define X_k for $k = 1, 2, 3, \ldots$ by the recursive formula

$$X_{k+1} = X_k + \Delta(2U_k - 1) \tag{1}$$

where Δ is a constant and

$$U_k \sim \text{Bernoulli}(1/2)$$

are iid random variables.

We can think of X_k as the position of a person who is randomly walking with 50-50 chance of the moving to the left or right by Δ at each time-step. The entire sequence $X_0, X_1, X_2, \ldots$ is referred to as the path of the random walker.

(a) Write a python function simulaterw(Delta,K) which simulates a random walk for N steps. Yours code should return the entire path in a numpy array. Make some plots of X_k vs. k.

(b) What are $E[X_k | X_{k-1} = 2]$ and $E[X_k]$?

(c) Using the central limit theorem, derive an approximation of the **mean squared displacement**

$$\text{MSD}(X_k) = E[X_k^2]$$

(you might notice this is just another name for the variance that is used in the context of random walks) Verify your approximation by plotting $\text{MSD}(X_k)$ as a function of N.

References

[1] Michael J Evans and Jeffrey S Rosenthal. *Probability and statistics: The science of uncertainty*. Macmillan, 2004.