Problem Set 1 (Main classes 1-4)

Izzy Grosof

September 24, 2025 – Due October 1st

1. Suppose that days in Evanston are each mainly sunny, cloudy, or rainy. Each day has a 70% probability of being mainly sunny, 20% of being mainly cloudy, and 10% of being mainly rainy. On a mainly sunny day, I won't be rained on. On a mainly cloudy day, there's a 10% chance I'll be rained on. On a mainly rainy day, there's a 90% change I'll be rained on.

   Today, I was rained on. Given this information, what's the probability that today was mainly sunny? Mainly cloudy? Mainly rainy?

2. Brent is a logistics student, studying how full storage units are at a local self-storage business. They represent the fullness of a storage unit as a number between 0 (empty) and 1 (completely stuffed). They find that lower fullness numbers tend to be more common. They decide the model the fullness of a storage unit as a random variable X with density $f_X(x)$ proportional to $1 - x$. Specifically, Brent models the fullness as a continuous random variable with probability density function

$$f_X(x) = \begin{cases} c(1-x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

   where c is a constant. They need your help to find the value of c for which the formula in (1) gives a value probability distribution.

   (a) There is only one value of c for which the formula for $f_X(x)$ in (1) is a valid probability density function. What is that value of c?

   (b) What is the mean fullness of a storage unit, $\mathbb{E}[X]$?

   (c) What is the variance of the fullness of a storage unit, $\text{Var}[X]$?

3. We flip two fair coins, each heads or tails independently with 50% probability of either outcome. We define three events:

   A: First coin flip is heads.

   B: Second coin flip is heads.

   C: Between the two flips, exactly one coin is heads.

   We want to know which events are independent of each other.

   (a) Are A and B independent? Why?

   (b) Are A and C independent? Why?

   (c) Are B and C independent? Why?

   (d) Are A, B, and C mutually independent? Why?

4. Pearson's correlation coefficient (PCC, also known as r) is frequently used in statistics to measure the correlation between two random variables. The PCC of two random variables X and Y is defined by the following formula:

$$PCC(X,Y) := \frac{Cov(X,Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

+1.0	Perfect positive (+) association
+0.8 to 1.0	Very strong + association
+0.6 to 0.8	Strong + association
+0.4 to 0.6	Moderate + association
+0.2 to 0.4	Weak + association
0.0 to +0.2	Very weak + or no association
0.0 to -0.2	Very weak negative (-) or no association
-0.2 to -0.4	Weak - association
-0.4 to -0.6	Moderate - association
-0.6 to -0.8	Strong - association
-0.8 to -1.0	Very strong - association
-1.0	Perfect - association

Table 1: Pearson Correlation Coefficient, strength of association. Credit: Boston University School of Public Health.

The PCC of two random variables is always a number between -1 and 1, and can be interpreted as shown in Table 1.

Let X be a random variable representing the symptom severity for a person receiving treatment for flu infections, at the time they are first seen by a nurse. Let Y be a random variable represent the symptom severity for the same person after two weeks of treatment.

Let's model X as being distributed uniformly among severity levels $\{1, 2, 3, 4\}$. Let's model Y as changing by at most two severity levels from X. Specifically, let's model Y as being distributed uniformly among severity levels $\{\max(X - 2, 0), X - 1, X, X + 1, X + 2\}$. The $\max(\cdot, \cdot)$ function is included to ensure that the severity level is never negative.

(a) What is the Pearson Correlation Coefficient $PCC(X, Y)$ for these two random variables?

(b) Using Table 1, what is the qualitative strength of association between X and Y?

5. Emma has proposed a new formula for the variance of a random variable, in addition to the formulas we've seen in class. Suppose that X, X_1, and X_2 are all independent and identically distributed. Then Emma claims that

$$\text{Var}[X] = \frac{\mathbb{E}[(X_1 - X_2)^2]}{2}.$$

Is Emma always correct? Prove that her formula always holds, or provide a counterexample.

6. This is a programming problem – write a program in Python, as either a standalone file or as a Jupyter notebook. Include your answers in your solution directly, and also submit the code you write, either in the same document or as a separate upload.

Let U_1 and U_2 be two independent and identically distributed random variables with distribution Uniform$(0, 1)$. Define their sum S to be $U_1 + U_2$, and define their difference D to be $U_1 - U_2$.

(a) Using 10^6 samples, estimate $\mathbb{E}[S]$, $\mathbb{E}[D]$, and $\mathbb{E}[SD]$.

(b) Using your result in (a), estimate $Cov(S, D)$.