Problem Set 1 Solutions (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?

   Solution. We want to find the probability of sunny, cloudy, or rainy, given that it rained. Let's write the mostly sunny event S, mostly cloudy event C, mostly rainy event A, and getting rained on R.

   We want to find $\mathbb{P}\{S \mid R\}, \mathbb{P}\{C \mid R\}, \mathbb{P}\{A \mid R\}$. We use Bayes' rule:

   $$\mathbb{P}\{S \mid R\} = \frac{\mathbb{P}\{S\&R\}}{\mathbb{P}\{R\}}, \mathbb{P}\{C \mid R\} = \frac{\mathbb{P}\{C\&R\}}{\mathbb{P}\{R\}}, \mathbb{P}\{A \mid R\} = \frac{\mathbb{P}\{A\&R\}}{\mathbb{P}\{R\}}$$

   We start by finding $\mathbb{P}\{R\}$, by using the law of total probability:

   $$\mathbb{P}\{R\} = \mathbb{P}\{S\&R\} + \mathbb{P}\{C\&R\} + \mathbb{P}\{A\&R\}$$

   We calculate $\mathbb{P}\{S\&R\}, \mathbb{P}\{C\&R\}, \mathbb{P}\{A\&R\}$ using conditioning:

   $$\begin{aligned} \mathbb{P}\{S\&R\} &= \mathbb{P}\{R \mid S\}\mathbb{P}\{S\} = 0 \cdot 0.7 = 0 \\ \mathbb{P}\{C\&R\} &= \mathbb{P}\{R \mid C\}\mathbb{P}\{C\} = 0.1 \cdot 0.2 = 0.02 \\ \mathbb{P}\{A\&R\} &= \mathbb{P}\{R \mid A\}\mathbb{P}\{A\} = 0.9 \cdot 0.1 = 0.09 \\ \mathbb{P}\{R\} &= 0 + 0.02 + 0.09 = 0.11 \end{aligned}$$

   Now we use Bayes' rule:

   $$\mathbb{P}\{S \mid R\} = \frac{0}{0.11} = 0, \mathbb{P}\{C \mid R\} = \frac{0.02}{0.11} = \frac{2}{11}, \mathbb{P}\{A \mid R\} = \frac{0.09}{0.11} = \frac{9}{11}$$

   Mostly sunny can't happen, mostly cloudy has $\approx 18.2\%$ probability, mostly rainy has $\approx 81.8\%$ probability. □

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

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

   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?

Solution. To be a valid probability density function, we need to guarantee that

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

With the given PDF, we have

$$\int_{x=0}^{1} c(1-x)dx = c[x - x^2/2]_{x=0}^{1} = c/2$$

We thus conclude that $c = 2$. □

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

Solution.

$$\begin{aligned}\mathbb{E}[X] &= \int_{x=-\infty}^{\infty} x f_X(x)dx = \int_0^1 x \cdot 2(1-x)dx = 2\int_0^1 x - x^2 dx \\ &= 2[x^2/2 - x^3/3]_{x=0}^{1} = 2(1/2 - 1/3) = 1/3\end{aligned}$$

□

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

Solution. We use the formula $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. We need to calculate $\mathbb{E}[X^2]$:

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{x=-\infty}^{\infty} x^2 f_X(x)dx = \int_0^1 x^2 \cdot 2(1-x)dx = 2\int_0^1 x^2 - x^3 dx \\ &= 2[x^3/3 - x^4/4]_{x=0}^{1} = 2(1/3 - 1/4) = 1/6\end{aligned}$$

Now, we plug this result into the $\mathrm{Var}[X]$:

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{6} - (\frac{1}{3})^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$$

□

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?

Solution. Yes, A and B are independent, because we were told that the coin flips were independent. □

(b) Are A and C independent? Why?

+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.

Solution. We can check if two events are independent by comparing $\mathbb{P}\{A\}$, $\mathbb{P}\{C\}$ and $\mathbb{P}\{A\&C\}$.

- We're told that $\mathbb{P}\{A\} = 1/2$.
- C can occur in two ways: with the sequence of coin flips HT or TH. Each happens with probability $1/4$, so $\mathbb{P}\{C\} = 1/2$.
- $A\&C$ can only happen in one way: HT, so $\mathbb{P}\{A\&C\} = 1/4$

As a result, we can verify that

$$\mathbb{P}\{A\} \cdot \mathbb{P}\{C\} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = \mathbb{P}\{A\&C\}$$

We have thus verified that A and C are independent. □

(c) Are B and C independent? Why?

Solution. By the same argument as in (b), we again conclude that B and C are independent. Note that $B\&C$ only happens if the coin flips turn out TH. □

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

Solution. For A, B, and C to be mutually independent, it must be the case that

$$\mathbb{P}\{A\&B\&C\} = \mathbb{P}\{A\}\mathbb{P}\{B\}\mathbb{P}\{C\} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

What is the actual probability $\mathbb{P}\{A\&B\&C\}$? For A and B to both occur, both coins must be heads, but in this case C does not occur. So $\mathbb{P}\{A\&B\&C\} = 0$. This doesn't match $\frac{1}{8}$, so the three events are not mutually independent. □

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:

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

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?

Solution. We must calculate $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, as well as $\text{Var}[X]$ and $\text{Var}[Y]$. We also use the facts that $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ and $\text{Var}[Y] = \mathbb{E}[Y^2] = \mathbb{E}[Y]^2$.
Let's start with quantities that don't involve Y.

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{4}(1+2+3+4) = \frac{10}{4} = \frac{5}{2}. \\
\mathbb{E}[X^2] &= \frac{1}{4}(1^2+2^2+3^3+4^3) = \frac{1}{4}(1+4+9+16) = \frac{30}{4} = \frac{15}{2} \\
\text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{15}{2} - (\frac{5}{2})^2 = \frac{5}{4}
\end{aligned}$$

Next, we'll calculate quantities that do involve Y. Our intention was that there were 19 joint possibilities for the pair (X, Y), corresponding to 4 possibilities for X and 5 for Y, with all pairs except $(1, 0)$ having probability $1/20$ and with the pair $(1, 0)$ having probability $2/20$. If you interpreted pairs $(1, 0), (1, 1), (1, 2)$, and $(1, 3)$ as each having probability $1/16$, we'll also count that as correct, as the problem was not completely clear.
We can calculate each expectation involving Y by summing over these 19 possibilities, which are:

$$\begin{aligned}
(X, Y) =& (1,0), (1,1), (1,2), (1,3), (2,0), (2,1), (2,2), (2,3), (2,4), \\
& (3,1), (3,2), (3,3), (3,4), (3,5), (4,2), (4,3), (4,4), (4,5), (4,6)
\end{aligned}$$

Whenever I perform a sum, I'll list the indices in this order, for clarity.

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{x,y} y\mathbb{P}\{X = x, Y = y\} \\
&= \frac{1}{20}(2\cdot 0 + 1 + 2 + 3 + 0 + 1 + 2 + 3 + 4 + 1 + 2 + 3 + 4 + 5 + 2 + 3 + 4 + 5 + 6) \\
&= \frac{1}{20}(6 + 10 + 15 + 20) = \frac{51}{20} \\
\mathbb{E}[Y^2] &= \sum_{x,y} y^2\mathbb{P}\{X = x, Y = y\} \\
&= \frac{1}{20}(2\cdot 0 + 1 + 4 + 9 + 0 + 1 + 4 + 9 + 16 + 1 + 4 + 9 + 16 + 25 + 4 + 9 + 16 + 25 + 36) \\
&= \frac{1}{20}(14 + 30 + 55 + 90) = \frac{189}{20} \\
\text{Var}[Y] &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{189}{20} - (\frac{51}{20})^2 = \frac{1179}{400} \\
\mathbb{E}[XY] &= \sum_{x,y} xy\mathbb{P}\{X = x, Y = y\} \\
&= \frac{1}{20}(2\cdot 0 + 1 + 2 + 3 + 0 + 2 + 4 + 6 + 8 + 3 + 6 + 9 + 12 + 15 + 8 + 12 + 16 + 20 + 24) \\
&= \frac{1}{20}(6 + 20 + 45 + 80) = \frac{151}{20}
\end{aligned}$$

Now, we can calculate our desired result:

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{151}{20} - \frac{15}{2}\frac{51}{20} = \frac{47}{40}$$
$$PCC(X,Y) = \frac{Cov(X,Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{47/40}{\sqrt{(1179/400)(5/4)}} = \frac{47}{3\sqrt{655}} \approx 0.612$$

□

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

Solution. X and Y have a strong positive association, in the $[0.6, 0.8]$ interval. □

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.

Solution. Emma correct. To prove it, we start by expanding the square.

$$\frac{\mathbb{E}[(X_1 - X_2)^2]}{2} = \frac{\mathbb{E}[X_1^2 - 2X_1X_2 + X_2^2]}{2}$$

Now, we use linearity of expectation to split up the expectation into simpler parts.

$$\frac{\mathbb{E}[X_1^2 - 2X_1X_2 + X_2^2]}{2} = \frac{\mathbb{E}[X_1^2] - 2\mathbb{E}[X_1X_2] + \mathbb{E}[X_2^2]}{2}$$

Next, we use the fact that X, X_1, and X_2 are identically distributed:

$$\frac{\mathbb{E}[X_1^2] - 2\mathbb{E}[X_1X_2] + \mathbb{E}[X_2^2]}{2} = \frac{2\mathbb{E}[X^2] - 2\mathbb{E}[X_1X_2]}{2} = \mathbb{E}[X^2] - \mathbb{E}[X_1X_2]$$

Finally, we use the fact that X_1 and X_2 are independent.

$$\mathbb{E}[X^2] - \mathbb{E}[X_1X_2] = \mathbb{E}[X^2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

This is one of our standard formulas for variance, so we have proven that the original formula is always correct. □

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 $\mathbb{E}[SD]$.

Solution. See solution program at the end of the document. I used the `random` package, it's fine if you used `numpy`.
The program shows that $\mathbb{E}[S] \approx 1, \mathbb{E}[D] \approx 0, \mathbb{E}[SD] \approx 0$. □

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

Solution. The program shows that $Cov(S, D) \approx 0$. □

Solution program for problem 6:

```
import random
random.seed(0)
trials = 1_000_000
u1_list = [random.random() for _ in range(trials)]
u2_list = [random.random() for _ in range(trials)]
s_list = [u1 + u2 for (u1, u2) in zip(u1_list, u2_list)]
d_list = [u1 - u2 for (u1, u2) in zip(u1_list, u2_list)]
es = sum(s_list) / trials
ed = sum(d_list) / trials

sd_list = [s * d for (s, d) in zip(s_list, d_list)]
esd = sum(sd_list)/trials
print("E[S]:", es, "E[D]:", ed, "E[SD]:", esd)

cov = esd - es * ed
print("Cov(S, D):", cov)
```

Output:

```
E[S]: 0.9996506078312181 E[D]: -0.00029408239190957373 E[SD]: -0.0002157764250784659
Cov(S, D): 7.820321674639795e-05
```