Homework 11

Discrete Structures
Due Tuesday, March 23, 2021

Submit each question separately as .pdf

---

1. Suppose you have two dice, one with 6 sides (having the numbers $1, \ldots, 6$) and one with 8 sides (having the numbers $1, \ldots, 8$). You roll both at the same time. Let T be the sum total of numbers on both die that are rolled.
   (a) Express all the possible values of T and the number of ways each value could be rolled.
   (b) Let T_{avg} be the arithmetic mean of all the possible values of T from part (a). What is the probability of rolling exactly T_{avg}?
   (c) What is the probability of rolling exactly $T_{avg} - 1$ or exactly $T_{avg} + 1$?

2. A bank assigns 5 digit PIN's (for example, 02270) to bank cards (customers don't get to pick their own PIN's.) Assume all combinations of 5 digits are equally likely. You find a bank card belonging to this bank left in a bank machine.
   (a) What is the probability of guessing the PIN if you try three times using a different possible PIN each time?
   (b) What is the probability the PIN has 5 different digits?
   (c) What is the probability the PIN contains at least one repeated digit?
   (d) If a bank card is stolen do you think a PIN with no repeated digits is more safe or less safe than one with repeated digits? Why?

3. Assume that we roll two regular dice (they can roll numbers 1–6 with equal probabilities). Define the following events:

$$\begin{cases} A := \text{the sum of the points is 7,} \\ B := \text{the first die rolled a 2,} \\ C := \text{the second die rolled a 5.} \\ D := \text{the sum of the points is at least 7,} \end{cases}$$

   (a) What are the conditional probabilities $p(B|A)$ and $p(B|D)$?

   The entire event space contains $6 \cdot 6 = 36$ elementary events. Every event is a pair (k_1, k_2), where $k_1, k_2 \in \{1, 2, 3, 4, 5, 6\}$. Every event has the same probability $\frac{1}{36}$. We can count how many squares belong to each of the events in Figure 1. The conditional probabilities can be computed by its definition:

$$p(B|A) = \frac{p(B \cap A)}{pA} = \frac{\frac{1}{36}}{\frac{6}{36}} = \frac{1}{6}.$$

$$p(B|D) = \frac{p(B \cap D)}{pD} = \frac{\frac{2}{36}}{\frac{21}{36}} = \frac{2}{21}.$$

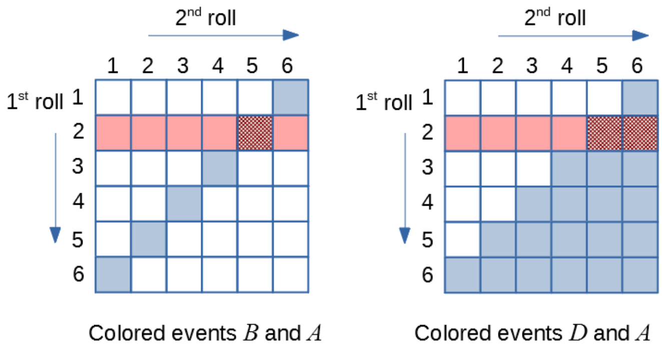

Figure 1: Event Space for two Dice Rolls.

Another way to see the conditional probability is to view only the cells shaded in blue (events B and D respectively). Then the conditional probability shows – the proportion of

□

(b) Are A, B, C pairwise independent?

We can use the method from the previous point **(a)** and count the number of events in the events A, B, C and also in their pairwise intersections $A \cap B$, $A \cap C$ and $B \cap C$. We get the following equalities:

$$\begin{cases} p(A \cap B) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p(A) \cdot p(B) \\ p(A \cap C) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p(A) \cdot p(C) \\ p(B \cap C) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p(B) \cdot p(C) \end{cases}$$

So, these pairs of events are independent.

Note. In everyday language we would call events B and C to be independent, since one dice roll does not affect the probability of the second dice roll. Event A (a statement about the sum of both dice rolls) would not be normally called *independent* from B and C, because the sum directly depends on these events. Yet the formal definition only needs to check the intersections such as $p(A \cap B) = p(A) \cdot p(B)$ and so on.

□

(c) Are A, B, C mutually independent?

In order to check the mutual independence of three events A, B, C, we need to check that they are pairwise independent (which was shown in **(b)**. But we also need the three events A, B, C to be independent.

So we need to check the equality:

$$p(A \cap B \cap C) \stackrel{?}{=} p(A) \cdot p(B) \cdot p(C).$$

The left side is satisfied by one event out of 36: it is the pair $(2; 5)$. On the other hand the probabilities $p(A) = p(B) = p(C) = \frac{1}{6}$. This equality fails, since $\frac{1}{36} \neq \frac{1}{216}$.

□

4. A chip factory *Intel* adds one toy animal to every bag of chips. There are three sorts of animals – Aligators, Bears or Cats (each one appears with probability p = 1/3). Let the random variable X denote the chip bags that someone needs to purchase in order to collect 3 different animals. Find the expected value $E(X)$.
   Hint. The total number of the bags to open to collect all three toy animals is denoted by random variable X. (It can take values $3, 4, 5, \ldots$ with certain probabilities). It may be tricky to compute these probabilities directly. Instead, you can express $X = X_1 + X_2 + X_3$ a sum of three simpler random variables:
   X_1 shows the bags that were needed to get the 1st unique animal (whatever it is),
   X_2 shows the bags needed to get the 2nd unique animal (already having the 1st one),
   X_3 shows the bags needed to get the 3rd unique animal (after getting the first two).
   All the random variables (except X_1) follow the geometric distribution (Textbook p.510, chapter 7.4.5).

   Clearly X_1 always equals 1: Whatever animal is found in the 1st bag of chips, it will be unique.

   $X_2 = 1$ with probability $p = \frac{2}{3}$ (a new unique animal is found in 2 cases of 3 – whenever it is not equal to the 1st animal.
   $X_2 = 2$ with probability $(1-p) \cdot p = \frac{1}{3} \cdot \frac{2}{3}$ (the first bag yielded an animal we already have, but the second bag had a unique one).
   In general, $X_2 = k$ with probability $(1-p)^{k-1} \cdot p = \left(\frac{1}{3}\right)^{k-1} \cdot \frac{2}{3}$ (the first $k-1$ bags were unsuccessful, but the last one was successful).

   $X_3 = 1$ with probability $p' = \frac{1}{3}$ (a new animal is found in 1 case out of 3 – we already have two animals and this one has to be different).
   In general, $X_3 = k$ with probability $(1-p')^{k-1} \cdot p' = \left(\frac{2}{3}\right)^{k-1} \cdot \frac{1}{3}$ (the first $k-1$ bags were unsuccessful with probability 2/3, but the last one was successful with probability 1/3).

   Random variable X_1 has expected value $E(X_1) = 1$ (it has only one value 1 with probability 1).
   Random variables X_2 and X_3 are *geometric distributions* with probabilities $p = \frac{2}{3}$ (and $p = \frac{1}{3}$ respectively). The expected values for X_1 and X_2 can be computed using the formula in Chapter 7.4.5 (Rosen2019, p.510).

   $$E(X_2) = \frac{1}{p} = \frac{1}{2/3} = 1\frac{1}{2}, \quad E(X_3) = \frac{1}{p'} = \frac{1}{1/3} = 3.$$

   Expected values are linear, so we can find $E(X) = E(X_1 + X_2 + X_3)$:

   $$E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = 1 + 1\frac{1}{2} + 3 = 5\frac{1}{2}.$$

   □

5. There are n people in some city after its reopening. Assume that health officials know that the fraction p of them still have a viral disease (here $p \in (0; 1)$ is a positive number very close to 0). There is a test for this disease which is 100% accurate; no false positives or false negatives. In order to save the number of tests needed, the people were grouped into groups of size k. (Assume that n is large and divisible by k.)

   For each group, the health officials blended k samples taken from the people in the group, and tested the blended sample. If everyone in the group is healthy, they need just one test. If the group tests positive for the disease, then they need to test all k people in the group, so they spend $k+1$ tests for that group. Let X be a random variable denoting the total number of tests spent to test all n people in the city in this manner.

(a) Express the expected value $E(X)$ in terms of n, the fraction of infected people p and the group size k?

Answer: See the expression (2).
Assume that we select k (the size of each group) so as the number of groups is much larger than the number of infected people:

$$\frac{n}{k} >> np. \tag{1}$$

We will justify this assumption in point **(b)**. Consider the "worst case" situation w.r.t. the number of tests necessary: Imagine that all the infected people fall into different groups (so that there are as many infected groups as possible and all of them need to be tested individually). In this case the number of necessary tests is the following:

$$X = \frac{n}{k} \cdot 1 + np \cdot k.$$

Namely, we spend 1 test per each of the $\frac{n}{k}$ groups; and we also spend k additional tests for each of the np infected people (as the entire group of k people has to be retested individually).
This X is a regular variable that depends on parameters n, k, p; and its value occurs with probability 1. It is not random (if we have the "worst case" assumption); so its expected value would be the same:

$$E(X) = \frac{n}{k} \cdot 1 + np \cdot k. \tag{2}$$

X becomes a random variable if we allow better than worst case distributions of the infected people (so that multiple infected people occur in the same group and we can save the total number of tests). Since p was assumed to be close to 0 (and we also assumed (1)), intuitively it won't be likely that many infected people happen in the same group. We will formalize our intuition in the next part of the answer.
□

(b) What group size k makes the value $E(X)$ as small as possible? (You can assume that p is so small that $(1-p)^k \approx 1$ and $\ln(1-p) \approx -p$.)

Introduce the function $f(k) = \frac{n}{k} \cdot 1 + np \cdot k$. Parameters n and p are constant, we can control the parameter k. To find the minimum of the function $f(k)$, find its derivative (with respect to the parameter k) and find where the derivative is 0:

$$f'(k) = -\frac{1}{k^2} + np = 0.$$

Solve this equation:

$$-\frac{n}{k^2} + np = 0; \quad -\frac{1}{k^2} + p = 0; \quad k^2 = \frac{1}{p}; \quad k = \frac{1}{\sqrt{p}}.$$

For example, if $p = 0.0001$ (one person out of 10000 is infected), then the optimal group size is $k = \frac{1}{\sqrt{p}} = 100$ regardless of the total number of people n.
We can check that for the values of k that are close to the optimal one, $\frac{n}{k} = n \cdot \sqrt{p}$ which is much larger than $n \cdot p$; therefore the assumption (1) in point **(a)** is reasonable.

Note. Considering only the "worst case" in **(a)** and **(b)**, where every infected person occurs in his/her own group alone, might seem arbitrary. Could the estimates for $E(X)$ be improved, if more than one infected person occurs in the same group we can save the number of tests considerably? We show that this is not the case. (This part is an optional piece of theory, it is not required as part of your solution for this question.)

Definition: Let $b(x; k, p)$ denote the *binomial distribution*: $b(x; k, p)$ is the probability of exactly x successes in k independent Bernoulli trials where the probability of success is p (Rosen2019, p.485).
(In the context of this problem a "success" means that some person tests positive for the disease.)

Statement 1: Let $q = 1 - p$ denote the failure (in our case the probability of the opposite event: testing negative for the disease). Then the binomial distribution can be computed like this:

$$b(x; k, p) = \binom{k}{x} p^x q^{k-x}.$$

This formula was derived in (Rosen2019, p.484). It shows that there are altogether $\binom{k}{x}$ ways to choose the locations of x successes, and this number has to be multiplied by $p^x q^{k-x}$ (the probability of getting one particular sequence of x successes and $k - x$ failures.

Statement 2: Consider the binomial distribution as the number of trials k increases indefinitely, but the product $\lambda = k \cdot p$ stays constant. In this case

$$\lim_{k \to \infty} b(x; k, p) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Proof. We want to get rid of the parameter p and replace it by λ using the expressions $p = \frac{\lambda}{k}$ and $q = 1 - p = 1 - \frac{\lambda}{k}$.

$$\begin{aligned} b(x; k, p) = \binom{k}{x} p^x q^{k-x} = \frac{k!}{(k-x)!x!} p^x q^{k-x} = \\ \frac{k!}{(k-x)!x!} \left(\frac{\lambda}{k}\right)^x \left(1 - \frac{\lambda}{k}\right)^{k-x} = \\ \frac{\lambda^x}{x!} \cdot \frac{k!}{(k-x)!k^x} \cdot \left(1 - \frac{\lambda}{k}\right)^k \left(1 - \frac{\lambda}{k}\right)^{-x}. \end{aligned} \tag{3}$$

Evaluate the limits for some factors separately:

$$\lim_{k\to\infty} \frac{k!}{(k-x)!k^x} = \lim_{k\to\infty} \frac{k(k-1)(k-2)\cdots(k-x+1)}{k^x} = 1.$$

$$\lim_{k\to\infty} \left(1 - \frac{\lambda}{k}\right)^k = e^{-\lambda}.$$

$$\lim_{k\to\infty} \left(1 - \frac{\lambda}{k}\right)^{-x} = 1.$$

Now, if we replace them in the expression (3):

$$\lim_{k\to\infty} b(x; k, p) = \frac{\lambda^x}{x!} \cdot 1 \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^x e^{-\lambda}}{x!}.$$

This limit is named *Poisson distribution*.

Definition: Random variable X is distributed accordingly to Poisson distribution, if for each nonnegative integer x

$$p(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Here the parameter λ denotes the expected value $E(X)$.

Conclusion. In our problem λ denotes the average number of infected people in a single group. If λ is small ($\lambda << 1$), then nearly all values of the Poisson-distributed random variable are either 0 or 1. (This justifies our "worst case" assumption. Even if we would compute the expected value of the necessary tests without this assumption, we would still get an expression that is very close to (2). $\square$