Homework 1

1. ***Airline on-time performance***

We will consider a sample space consisting of all the United Airlines and America West flights landing at Chicago, Los Angeles, Phoenix, San Diego, or San Francisco.

Define events corresponding to the airlines

$$\begin{aligned} U &= \text{ flight is run by United} \\ W &= \text{ flight is run by America West} \end{aligned}$$

and also to the airports

$$\begin{aligned} C &= \text{ flight lands at Chicago} \\ L &= \text{ flight lands at Los Angeles} \\ X &= \text{ flight lands at Phoenix} \\ D &= \text{ flight lands at San Diego} \\ F &= \text{ flight lands at San Francisco} \end{aligned}$$

Also let

$$T = \text{ flight lands on time}$$

The conditional probabilities of on-time arrival are

$$\begin{aligned} &\mathrm{Prob}(T \mid U \cap C) = 0.85, &&\mathrm{Prob}(T \mid U \cap L) = 0.92, &&\mathrm{Prob}(T \mid U \cap X) = 0.95, \\ &\mathrm{Prob}(T \mid U \cap D) = 0.91, &&\mathrm{Prob}(T \mid U \cap F) = 0.83, && \\ &\mathrm{Prob}(T \mid W \cap C) = 0.78, &&\mathrm{Prob}(T \mid W \cap L) = 0.88, &&\mathrm{Prob}(T \mid W \cap X) = 0.92, \\ &\mathrm{Prob}(T \mid W \cap D) = 0.85, &&\mathrm{Prob}(T \mid W \cap F) = 0.73. && \end{aligned}$$

Looking at the data, it seems clear that United has the better on-time performance, since more of its flights land on time at each of the different airports.

(a) Use the fact that $\{C, L, X, D, F\}$ is a partition of the sample space to show that the average on-time arrival probability $\mathrm{Prob}(T \mid U)$ for United flights is given by

$$\begin{aligned} \mathrm{Prob}(T \mid U) = \mathrm{Prob}(T \mid U \cap C)\,\mathrm{Prob}(C \mid U) + \mathrm{Prob}(T \mid U \cap L)\,\mathrm{Prob}(L \mid U) \\ + \mathrm{Prob}(T \mid U \cap X)\,\mathrm{Prob}(X \mid U) + \mathrm{Prob}(T \mid U \cap D)\,\mathrm{Prob}(D \mid U) \\ + \mathrm{Prob}(T \mid U \cap F)\,\mathrm{Prob}(F \mid U) \end{aligned}$$

where $\mathrm{Prob}(C \mid U)$ is the conditional probability that the flight is landing at Chicago given that it is a United flight, etc.

(b) 60% of United Airlines flights land at its hub (snowy Chicago), 15% at each of LA and San Francisco, and 5% at each of Phoenix and San Diego. 75% of America West flights land at its hub (sunny Phoenix), 10% at LA, and 5% at each of the other three airports. Show that $\mathrm{Prob}(T \mid U) < \mathrm{Prob}(T \mid W)$, i.e., United has a worse average on-time performance even though it beats America West at all the five airports! Explain this discrepancy between the per-airport on-time performance and the overall on-time performance.

2. ***Comparing dice***

In this problem, we consider three dice with unusual numbering. The numbers on the dice are

die 1	5	6	7	8	9	18
die 2	2	3	4	15	16	17
die 3	1	10	11	12	13	14

All of the dice are unbiased; that is all 6 outcomes are equally likely. A game is played with these dice. Each player has a different one of the dice. They each roll their die, and whoever had the higher number wins.

(a) If dice 1 and 2 are rolled, what is the probability that die 1 beats die 2?

(b) If dice 2 and 3 are rolled, what is the probability that die 2 beats die 3?

(c) If dice 3 and 1 are rolled, what is the probability that die 3 beats die 1?

3. ***Random variables***

The term *random variable* can be very misleading. A random variable is just a *function* mapping one set Ω to another set V. The probabilities of elements in Ω are mapped to probabilities of elements in V.

Suppose $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $p : \Omega \to \mathbb{R}$ is a probability mass function on Ω given by $p(i) = p_i$ where

$$p = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.2 \\ 0.3 \\ 0.2 \\ 0.1 \end{bmatrix}$$

(a) Now let's define a random variable f on Ω, taking values in $V = \{1, 2, 3\}$. That is, $f : \Omega \to V$. The function f is given by

$$f = \begin{bmatrix} 1 \\ 3 \\ 1 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

where as for pmfs we understand that $f(i) = f_i$

Let $q \in \mathbb{R}^3$ be the *induced pmf* on V. That means

$$q_i = \text{Prob}(f = i)$$

which means

$$q_i = \text{Prob}\big(\{\, \omega \in \Omega \mid f(\omega) = i \,\}\big)$$

Another name for the *induced pmf* is the *derived distribution*. Find q.

(b) Whenever you have a random variable $f : \Omega \to V$, there are two ways to find probabilities. We can either use the pmf on Ω or the pmf on V. Suppose we now want to find the probability that $f \geq 2$. We can think of this as the event $A \subset \Omega$, where

$$A = \big\{\, \omega \in \Omega \mid f(\omega) \geq 2 \,\big\}$$

Find the set A, and hence the probability of A, which is defined by

$$\underset{\Omega,\, p}{\text{Prob}}(A) = \sum_{\omega \in A} p(\omega)$$

(c) Another way to think about this is as the event $B \subset V$ where

$$B = \big\{\, y \in V \mid y \geq 2 \,\big\}$$

The probability of B is

$$\underset{V,\, q}{\text{Prob}}(B) = \sum_{y \in B} q(y)$$

Find B, and hence the probability of B. Since $\omega \in A$ if and only if $f(\omega) \in B$, that is A and B represent the same set of outcomes, we know that A and B have the same probability.

(d) There are also two ways to find the expectation of the random variable f. First, find $\mathrm{E}\, f$ using the definition

$$\mathrm{E}\, f = \sum_{\omega \in \Omega} f(\omega) p(\omega)$$

Alternatively, we can work in V, where

$$\mathrm{E}\, f = \sum_{y \in V} y q(y)$$

Find $\mathrm{E}\, f$ using this approach also, and verify that the above two answers are equal.

4. ***Markov bounds and test scores***

Suppose n people take a test, and receive scores $x_1, x_2, \ldots, x_n$, all of which satisfy $x_i \geq 0$. Let's define as usual the average score

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Also define the function $h : \mathbb{R} \to \mathbb{R}$ by

$$h(a) = \frac{1}{n}(\text{no. of students scoring greater than or equal to } a)$$

Notice that there is no probability in this question; it's only about counting.

(a) Show that

$$h(a) \leq \frac{\mu}{a} \qquad \text{for all } a > 0$$

(b) Suppose $\mu = 80$. What is the largest fraction γ of the class that could have scored greater than or equal to 95?

(c) Give an example of a set of scores where the average is 80 and $n\gamma$ of the scores are greater than or equal to 95. Here γ is the fraction you gave in part (b).

5. ***Simulating Discrete Random Variables***

Suppose Ω is a sample space and $x : \Omega \to V$ is a random variable. In this question we will not specify Ω and x directly. Instead we will work with the *induced pmf* p^x on V. Here $V \subset \mathbb{R}$, given by

$$V = \{ v_1, v_2, \ldots, v_n \}$$

where

$$v = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 7 \\ 8 \\ 9 \end{bmatrix}$$

The random variable x has induced pmf $p^x : V \to [0, 1]$. We represent p^x by a vector $p^x \in \mathbb{R}^n$, so that

$$p^x(v_i) = p^x_i$$

Let

$$p^x = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.3 \\ 0.2 \\ 0.1 \end{bmatrix}$$

(a) Write code which takes v and p^x and returns a random element of V, generated according to the probabilities in p^x.

(b) Perform $m = 10,000$ trials, and collect data points $y(1), \dots, y(m)$, where each $y(i) \in V$. Let $s(j, n)$ denote the frequency of v_j in the first n data points, i.e., the $1/n$ multiplied by how many times v_j appears in the sequence $y(1), y(2), \dots, y(n)$. Plot $s(j, n)$ against n, for each j.