DATA 1010
Problem Set 5
Due 12 October 2018 at 11 PM

Problem 1

Suppose that X and Y are random variables whose joint distribution is given by the following table.

		X			
		-1	0	1	2
	-1	0	$\frac{1}{36}$	$\frac{1}{6}$	$\frac{1}{12}$
Y	0	$\frac{1}{18}$	0	$\frac{1}{18}$	0
	1	0	$\frac{1}{36}$	$\frac{1}{6}$	$\frac{1}{12}$
	2	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{1}{6}$

(a) Find $P(X \geq 1 \text{ and } Y \leq 0)$.

(b) What is the conditional probability of the event $\{Y \leq 0\}$ given that $X = 2$?

(c) Are X and Y independent?

(d) What is the distribution of $Z = XY$?

Solution

(a) We have $X(\omega) \geq 1$ and $Y(\omega) \leq 0$ if and only if ω maps under (X, Y) to one of the points in the top-right 2×2 corner. The total mass of these ω's is

$$\frac{1}{6} + \frac{1}{12} + \frac{1}{18} = \frac{11}{36}.$$

(b) Conditioning on the event $\{X = 2\}$ zeros out all the masses except the ones in the rightmost column. Each of the remaining masses gets boosted by a factor of $(\frac{1}{12} + \frac{1}{12} + \frac{1}{6})^{-1} = 3$. So the probability that $Y \leq 0$ is $3 \cdot (\frac{1}{12} + 0) = \frac{1}{4}$.

(c) Definitely not. There is a positive probability that $X = 0$ and a positive probability that $Y = 0$, but the probability that $(X, Y) = (0, 0)$ is equal to zero.

(d) The product $Z = XY$ takes values in $\{-2, -1, 0, 1, 2, 4\}$. We can find the probability mass associated with each of these values by looping through the table and assigning the probability mass of each entry to the corresponding product value. Here's an implementation in Julia (though it's also doable by hand).

```
A = [0 1//36 1//6 1//12; 1//18 0 1//18 0; 0 1//36 1//6 1//12; 1//12 0 1//12 1//6]

masses = Dict{Int64,Rational{Int64}}(-2=>0,-1=>0,0=>0,1=>0,2=>0,4=>0)
for i=1:size(A,1)
    for j=1:size(A,2)
        yval, xval = i-2, j-2
        masses[xval*yval] += A[i,j]
    end
end
masses
```

It turns out that Z is uniformly distributed on $\{-2, -1, 0, 1, 2, 4\}$.

Problem 2

A die is rolled twice. Let X denote the sum of the two numbers that turn up, and Y the difference of the numbers (first roll minus second). Show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ but that X and Y are not independent.

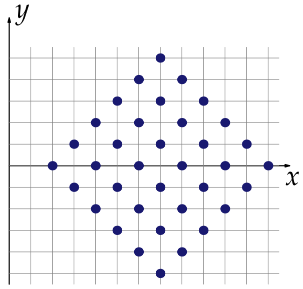
Solution

Let R_1 and R_2 be the two rolls. Then

$$\mathbb{E}[XY] = \mathbb{E}[(R_1+R_2)(R_1-R_2)] = \mathbb{E}[R_1^2 - R_2^2] = \mathbb{E}[R_1^2] - \mathbb{E}[R_2^2] = 0,$$

since R_1 and R_2 have the same distribution. Similarly, $\mathbb{E}[X]\mathbb{E}[Y] = 0$, since $\mathbb{E}[Y] = \mathbb{E}[R_1 - R_2] = \mathbb{E}[R_1] - \mathbb{E}[R_2] = 0$.

The random variables X and Y are not independent, since $X = 1$ and $Y = 2$ each have positive probability, but their intersection has zero probability.



Problem 3

Consider m zeros and n ones arranged in order uniformly at random. For example, if $m = 3$ and $n = 7$, then

$$1, 0, 1, 1, 0, 0, 1, 1, 1, 1$$

is one such ordering. This ordering has 3 runs of ones (in the first position, in the third through fourth positions, and in the final 4 positions).

Let the random variable R be the number of runs of ones. Find $\mathbb{E}[R]$.

(Hint: Write R as $R_1 + \cdots + R_{m+n}$, where R_i is defined to be 1 if a run begins at slot i and 0 otherwise.)

Solution

By linearity of expectation, we have

$$\mathbb{E}[R] = \mathbb{E}[R_1] + \cdots + \mathbb{E}[R_{m+n}],$$

where R_i is the indicator of the event that a run begins at slot i. A run begins in the first position if and only if the number in that position is a 1. Since there are n ones out of $m+n$ total bits, and each is equally likely to be in the first position, we have

$$\mathbb{E}[R_1] = \frac{n}{m+n}.$$

For each of the remaining positions, a run begins in that position if and only if the position contains a one and the previous position contains a zero. This happens with probability

$$\mathbb{E}[R_i] = \overbrace{\frac{m}{m+n}}^{\text{pos. } i-1 \text{ contains } 0} \times \overbrace{\frac{n}{m+n-1}}^{\text{pos. } i \text{ contains 1 given pos. } i-1 \text{ contains } 0} = \frac{mn}{(m+n)(m+n-1)}.$$

Therefore, the expected value of R is

$$\frac{n}{m+n} + \frac{mn}{(m+n)(m+n-1)}(m+n-1) = \boxed{\frac{m(n+1)}{(m+n)}}.$$

We can confirm this result by Monte Carlo:

Julia

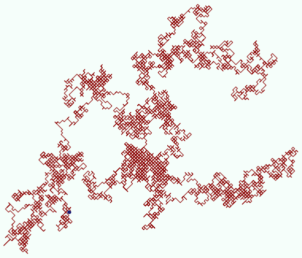
```julia
using Random, Statistics
function countruns(A)
    s = 0
    if A[1] == 1
        s += 1
    end
    for i=2:length(A)
        if A[i-1] == 0 && A[i] == 1
            s += 1
        end
    end
    s
end
m = 18; n = 24
mean(countruns(shuffle(vcat(fill(0,m),fill(1,n)))) for i=1:10^6), (m*n + n)/(m+n)
```

Problem 4

A particle begins at the origin at time 0. It repeatedly and independently chooses its steps uniformly at random from the set $\{(1,1),(1,-1),(-1,1),(-1,-1)\}$. (For example, after 2 steps the particle might be at the point $(2,0)$, perhaps after taking a $(1,-1)$ step and then a $(1,1)$ step.)

(i) Find the particle's expected squared distance from the origin after taking n steps.

(ii) Perform a Monte Carlo simulation to verify your result from part (i), for $n \in \{100, 1000, 10{,}000\}$.



Solution

(i) If the ith step is (X_i, Y_i), then the location of the particle after n steps is

$$(X_1 + X_2 + \cdots + X_n, Y_1 + Y_2 + \cdots + Y_n).$$

The expected squared distance from this point to the origin is

$$\mathbb{E}[(X_1 + X_2 + \cdots + X_n)^2 + (Y_1 + Y_2 + \cdots + Y_n)^2].$$

Squaring out the sum, we find that the expectation of the first term is

$$\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \cdots + \mathbb{E}[X_n^2] + 2\mathbb{E}[X_1 X_3] + \cdots + 2\mathbb{E}[X_{n-1} X_n].$$

Since X_i takes values in $\{-1,+1\}$, we have $X_i^2 = 1$ everywhere on the sample space. So each of the first n terms in this expression is equal to 1. The remaining terms are all zero, since X_i is independent of X_j whenever $i \neq j$. So we have

$$\mathbb{E}[(X_1 + X_2 + \cdots + X_n)^2] = n$$

and similarly for the Y terms. So the expected squared distance from the origin is $\boxed{2n}$.

(ii) We can check this using Monte Carlo simulation:

Julia

```julia
steps = [[1,1],[1,-1],[-1,1],[-1,-1]]
iterations = 10^6
n = 10
mean(norm(sum(rand(steps) for i=1:n))^2 for j=1:iterations), 2n
```

Problem 5

Suppose that $\Omega = \{A, B, C\} \times \mathbb{Z}$. We will represent an element of Ω as $\omega = (\omega_1, \omega_2)$, where $\omega_1 \in \{A, B, C\}$ and $\omega_2 \in \mathbb{Z}$. Suppose that $\mathbb{P}(\omega_1 = A) = \frac{1}{4}$, $\mathbb{P}(\omega_1 = B) = \frac{1}{5}$, and $\mathbb{P}(\omega_1 = C) = \frac{11}{20}$.

Suppose further that (i) the conditional distribution of ω_2 given $\{\omega_1 = A\}$ has probability mass function $n \mapsto \frac{1}{3}2^{-|n|}$, (ii) the conditional distribution of ω_2 given $\{\omega_1 = B\}$ is the uniform distribution on $\{-2, -1, 0, 1, 2\}$, and (iii) the conditional distribution of ω_2 given $\{\omega_1 = C\}$ has probability mass function $n \mapsto \mathbf{1}_{\{n \geq 1\}} \frac{6}{\pi^2 n^2}$.

Find the conditional distribution of ω_1 given:

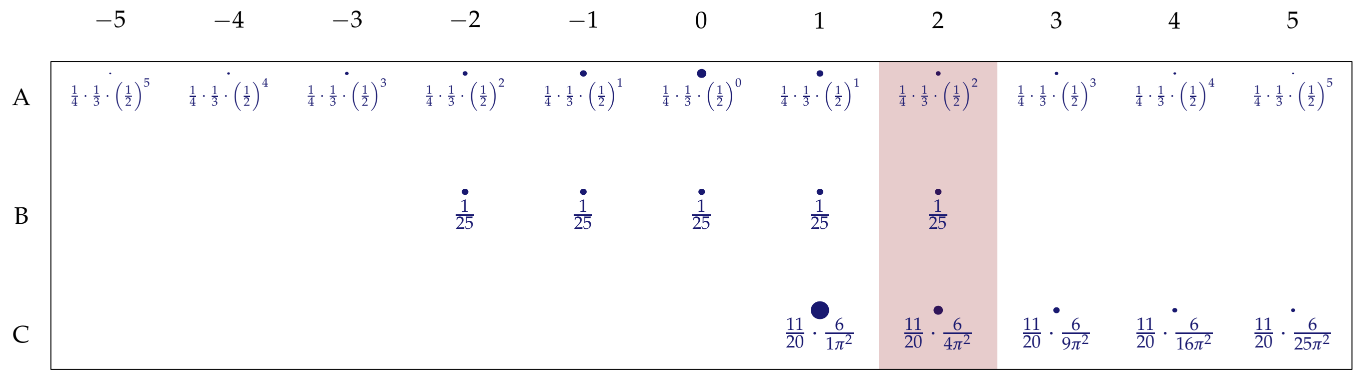
(i) $\omega_2 = 2$
(ii) $\omega_2 = 7$
(iii) $\omega_2 \geq 10$

(In each case, express the conditional probabilities as percentages rounded to the nearest hundredth of a percent.)

Solution

(i) The information given in the problem can be used to work out the probability mass of each $\omega \in \Omega$. If we arrange the elements of Ω into three infinitely long rows, then the information $\mathbb{P}(\omega_1 = A) = \frac{1}{4}$, $\mathbb{P}(\omega_1 = B) = \frac{1}{5}$, and $\mathbb{P}(\omega_1 = C) = \frac{11}{20}$ says that the first row has a total mass of $\frac{1}{4}$, the second row as a total mass of $\frac{1}{5}$, and the third row has a total mass of $\frac{11}{20}$.

The conditional information tells us how the mass along each row is distributed. For example, the $\frac{1}{5}$ mass on the second row is shared equally among $\{-2, -1, 0, 1, 2\}$. Therefore, each of those ω's has a mass of $\frac{1}{25}$. Applying similar reasoning to the first and third rows, we get the following picture:



	-5	-4	-3	-2	-1	0	1	2	3	4	5
A	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^5$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^4$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^3$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^2$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^1$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^0$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^1$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^2$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^3$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^4$	$\frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^5$
B				$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$			
C							$\frac{11}{20} \cdot \frac{6}{1\pi^2}$	$\frac{11}{20} \cdot \frac{6}{4\pi^2}$	$\frac{11}{20} \cdot \frac{6}{9\pi^2}$	$\frac{11}{20} \cdot \frac{6}{16\pi^2}$	$\frac{11}{20} \cdot \frac{6}{25\pi^2}$

Conditioning on $\{\omega_2 = 2\}$ means restricting the sample space to the column shown in red. All other masses are set to zero, and these three masses are each multiplied by the reciprocal of their sum. So the conditional probability of $\omega_1 = A$ given $\omega_2 = 2$ is

$$\frac{\frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{4}}{\frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{25} + \frac{11}{20} \cdot \frac{6}{4\pi^2}} = 14.43\%,$$

and similarly the conditional probability of $\omega_1 = B$ given $\omega_2 = 2$ is 27.70%, and the conditional probability of $\omega_1 = C$ given $\omega_2 = 2$ is 57.88%.

(ii) Replacing 2 with 7, we get masses of

$$(m(A), m(B), m(C)) = (8.71\%, 0\%, 91.29\%).$$

(iii) Conditioning on $\omega_2 \geq 10$ means restricting the sample space to all of the columns to the right of column 9. Under this conditional measure, the total mass of the first row is

$$\frac{\sum_{k=10}^{\infty} \frac{1}{4} \cdot \frac{1}{3} \cdot \left(\frac{1}{2}\right)^k}{\sum_{k=10}^{\infty} \frac{11}{20} \cdot \frac{6}{\pi^2 k^2}} = \frac{2 \cdot \frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{2^{10}}}{\frac{11}{20} - \sum_{k=1}^{9} \frac{11}{20} \cdot \frac{6}{\pi^2 k^2}} = \frac{0.000162760417}{0.035163406} = 0.46\%,$$

and the total mass in the last row is $1 - 0.46\% = 99.54\%$.

Problem 6

Consider a random independent sequence of letters uniformly distributed in $\{\texttt{a}, \texttt{b}, \ldots, \texttt{z}\}$. Use Monte Carlo simulation to estimate the expected number of letters that appear in the sequence up to the first appearance of `aa`.

Repeat with `ab` in place of `aa`. Based on your findings, is the expected time to the first `aa` different from the expected time to the first `ab`?

Solution

Using the Monte Carlo results below, we estimate the mean of the number of letters till `aa` around 700 and the mean of the number of letters till `ab` around 680.

Julia

```julia
function rand_letters_till(word::AbstractString)
    letters = String([rand(collect('a':'z')) for i=1:length(word)])
    ctr = length(word)
    while letters[end-length(word)+1:end] ≠ word
        letters = letters[2:end] * rand(collect('a':'z'))
        ctr += 1
    end
    ctr
end
mean(rand_letters_till("aa") for i=1:10^5) # returns ~700
mean(rand_letters_till("ab") for i=1:10^5) # returns ~680
```

If we run this block repeatedly, the `aa` estimate is consistently 20-30 letters larger than the `ab` estimate. This gives us confidence that the mean is genuinely larger.

Bonus: an explanation

One way to see that the mean should be larger for `aa` is to think about it progressively: you wait for an `a` to appear, and then you hope the next letter is `a`. If it isn't, you're back where you started. For `ab`, you're hoping the next letter after the first `a` you see is `b`, but if it isn't, there's a chance that it's another `a`. In that case, you're in much better position hoping the next letter is `b` than you would be if you were starting over from scratch.

We can actually calculate the expectations exactly. Imagine infinitely many bettors, one for each integer $n \geq 1$, placing a wager on whether the nth letter is `a`. The nth bettor pays one dollar to receive 26 dollars if they win and nothing if they lose. If the bettor wins (and if the game isn't over yet), then they wager all of their winnings on the next letter being an `a`. So they will end up with either 26^2 dollars or zero dollars in that case.

Once we have seen the first `aa` in positions $N-1$ and N, the bettor on position $N-1$ will have won 26^2 dollars, and the bettor on position N will have won 26 dollars. So the winnings will be exactly $26^2 + 26 = 702$ dollars. Since each bet was fair, the amount of money paid in by the bettors must be equal in expectation to the winnings. Therefore, the expected number of dollars paid in is 702.

If we switch from `aa` to `ab`, the same analysis holds except that the final bettor loses. So the total winnings are $26^2 = 676$, and therefore so is the expected number of letters that have appeared when we see the first `ab`.

This argument uses a theorem called the *optional stopping theorem*, which says that if you're playing a progressive game which is fair at each step you decide to stop at some point, the stopped game is still fair. Those who are interested may read about this theorem in any book which covers martingales (my favorite is David Williams' *Probability with Martingales*).