W203 Statistics for Data Science

Aditya Mengani

aditya.mengani@berkeley.edu

Aditya Mengani

Monday, 4pm

Summer 2020

Lab 1: Probability Theory

June 1, 2020

1. Meanwhile, at the Unfair Coin Factory...

You are given a bucket that contains 1000 coins. 999 of these are fair coins, but one of them is a trick coin that always comes up heads. You select one coin from this bucket at random. Let T be the event that you select the trick coin. This means that $P(T) = 0.001$.

a. Suppose you flip the coin k times. Let H_k be the event that the coin comes up heads all k times. If you see this occur, what is the conditional probability that you have the trick coin? In other words, what is $P(T|H_k)$. (3 points)

b. How many heads in a row would you need to observe in order for the conditional probability that you have the trick coin to be higher than 99.9%? (3 points)

Let $T = \{$ event that you select a trick coin $\}$

$H_K = \{$ event that coin comes up head all k times $\}$

Given $P(T) = 0.001$

thus $P(\bar{T}) = 0.999$ (from rules of probability)

$P(H_K|\bar{T}) = \frac{1}{2^K}$

(since all possible events for a toss are head, tail, for H_K)

$P\left(\frac{H_1}{\bar{T}}\right) = \frac{1}{2}$, $P\left(\frac{H_2}{\bar{T}}\right) = \frac{1}{2} \times \frac{1}{2}$ $P\left(\frac{H_3}{\bar{T}}\right) = \frac{1}{2^3}$ ....

We can also infer that,

$P\left(\frac{H}{T}\right) = 1$, (Since when you toss a trick coin, you will always get a head)

<u>Problem a:-</u> Find $P(T/H_K)$

This can be solved by applying the "Total Law of probability", which states,

$A_1 \cdots A_K$ be M.E. events, then for B which is another event

$$\boxed{P(B) = \sum_{i=1}^{k} P(B/A_i) \cdot P(A_i)}$$

Thus, $P(H_K) = P\left(\frac{H_K}{T}\right) \cdot P(T) + P\left(\frac{H_K}{\bar{T}}\right) \cdot P(\bar{T})$

$= \cancel{\frac{1}{k}(0.001)} + \left(P(H_K|T) = 1, \text{ for any } K\right)$

$$\boxed{P(H_K) = 1 \cdot (0.001) + \frac{1}{2^K} \cdot (0.999)} \quad \text{— (a)}$$

$$P\left(\frac{T}{H_K}\right) = \frac{P(T \cap H_K)}{P(H_K)} = \frac{P(H_K|T) \cdot P(T)}{P(H_K)} \quad \left(\text{According to Baye's formula}\right)$$

$$= \frac{1 \cdot P(T)}{P(H_K)} \quad \left(\text{since } P(H_K|T) = 1\right) \quad \text{— (b)}$$

Substituting (a) in (b)

$$\boxed{P(T/H_K) = \frac{0.001}{0.001 + \frac{1}{2^K}(0.999)}}$$

Problem 1(b) :-

The question asks to find total heads in a row tossed to find the conditional probability that,

$$P(T/H_K) \geq 0.999$$

$$= \frac{0.001}{0.001 + \frac{1}{2^K}(0.999)} \geq 0.999$$

$$= 0.001 \geq 0.999(0.001) + \frac{1}{2^K}(0.999)(0.999)$$

$$\Rightarrow \frac{1}{2^K}(0.999)(0.999) < (0.001) - (0.999)(0.001)$$

$$= \frac{1}{2^K}(0.999)^2 < (0.001)^2$$

$$= 2^K > \frac{(0.999)^2}{(0.001)^2}$$

$$= 2^K > (999)^2$$

$$= K \log 2 > 2 \log 999$$

$$= K > \frac{2 \cdot (\log 999)}{\log 2}$$

$$= K > \frac{2 \cdot (2.999)}{(0.301)}$$

$$= \boxed{K > 19.99} \rightarrow$$ Thus we need atleast 19.99 (approx 20) in a row tosses of Heads for $P\left(\frac{T}{H_K}\right)$ to be greater than 99.99%

2. Broken Rulers

You have a ruler of length 4 and you choose a place to break it using a uniform probability distribution. Let random variable X represent the length of the left piece of the ruler. X is distributed uniformly in $[0, 4]$. You take the left piece of the ruler and once again choose a place to break it using a uniform probability distribution. Let random variable Y be the length of the left piece from the second break.

a. Draw a picture of the region in the X-Y plane for which the joint density of X and Y is non-zero. (3 points)

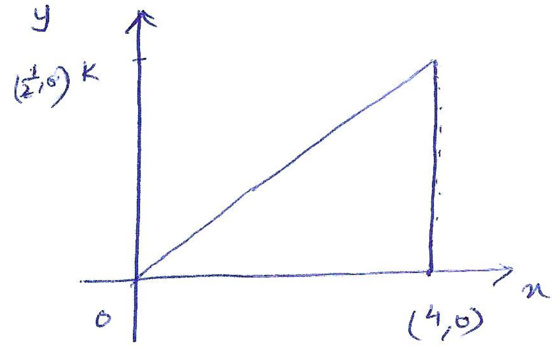
b. Find the conditional expectation of Y given X, $E(Y|X)$. (3 points)

c. Find the unconditional expectation of Y. (3 points)

d.* Give a complete expression for the conditional distribution of X, conditional on Y. (3 points)

e.* Compute $cov(X, Y)$. (3 points)

(a). The 2-d representation of the graph is as follows:-



Let K be a point in y axis, where area under the curve becomes 1

Thus area $= \frac{1}{2} \times K \times 4 = 1$

$\Rightarrow K = \frac{1}{2}$

Thus $y = f(x) = \left(\frac{1}{2}, 0\right)$ is the point at height $\frac{1}{2}$ above point (x, y) in a 3-dimensional coordinate system.

Hence the joint pdf can be written as

$$f(x,y) = \begin{cases} 1/2 & \text{where } 0<y<x<4 \\ 0 & \text{otherwise} \end{cases}$$

<u>Note:-</u> The problem says 'x' represents the random variable that represents the left piece of the ruler

So, the range of x can be between [0,4]

Similarly 'y' represents the left piece of ruler once again broken after the left piece is broken initially.

So, the value of Y is between (0, x)

So, f(x,y) will be a surface at height 1/2 above point (x,y) in 3-d coordinate system.

2(b). Find $E(Y/X)$

$E(Y/X)$ can be written as

$$\boxed{E(Y/X) = \int_{-\infty}^{\infty} Y \cdot f_{(Y/X)}(Y/X) \cdot dy}$$

the limits of y are between $(0, x)$ thus,

$$\boxed{E(Y/X) = \int_{0}^{x} Y \cdot \frac{f(x,y)}{f_X(x)} \cdot dy} \quad \text{——— (a)}$$

The marginal pdf of X, defined as $f_X(x)$ can be found by below rule:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) \cdot dy \quad \text{for } -\infty < x < \infty$$

As $f(x,y) = 1/2$ and positive limits/density of Y is $0 < y < x$,

$$f_X(x) = \int_0^x 1/2 \cdot dy$$

$$= 1/2 \int_0^x 1 \cdot dy = \frac{1}{2}\Big[y\Big]_0^x = \frac{x}{2}$$

Substituting $\boxed{f_X(x) = \frac{x}{2}}$ and $\boxed{f(x,y) = \frac{1}{2}}$ in (a)

$$E(Y/x) = \int_0^x y \cdot \frac{(1/2)}{(x/2)} dy = \int_0^x 1/x \cdot y \cdot dy$$

$$= \frac{1}{x}\left[\frac{y^2}{2}\right]_0^x = \frac{x^2}{2x} = \frac{x}{2}$$

$$\boxed{E(Y/x) = \frac{x}{2}}$$

(c). Find $E(Y)$

The $E(Y/x)$ calculated in (b) can be used to derive $E(Y)$

The unconditional probability for Y

$$E(Y) = E\left[\frac{X}{2}\right] \quad \left(\text{Since } E(Y/X) = \frac{X}{2}\right)$$

$$= \frac{E[X]}{2} = \frac{\int_{-\infty}^{\infty} x \cdot f(x)\, dx}{2}$$

$$= \frac{1}{2}\int_0^4 x \cdot \frac{x}{2} dx \quad \left(\text{since } f_X(x) = \frac{x}{2}\right)$$

$$= \frac{1}{4}\int_0^4 x^2 dx = \frac{1}{4}\left[\frac{x^3}{3}\right]_0^4$$

$$= \frac{4^3}{4 \times 3} = \frac{16}{3}$$

$$\boxed{E(Y) = \frac{16}{3}}$$

problem (d):-

Find $E(X/Y)$

$E(X/Y)$ can be written as

$$\boxed{E(X/Y) = \int_{-\infty}^{\infty} X \cdot f_{(X/Y)}(X/Y)\, dx}$$

the limits of X are between (0, 4),

thus $E(X/Y)$ can be written as,

$$\boxed{E(X/Y) = \int_{0}^{4} X \cdot \frac{f(x,y)}{f_y(y)}\, dx}$$

The marginal pdf of Y, defined as $f_Y(y)$ can be written by below rule:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) \cdot dx \quad \text{for } 0 < y < x$$

$$f_Y(y) = \int_0^4 1/2 \cdot dx \qquad \left(\text{Since } f(x,y) = 1/2\right)$$

$$= \frac{1}{2}\left[x\right]_0^4 = 2$$

$$\boxed{f_Y(y) = 2}$$

Thus $E(X/Y) = \int_0^4 x \cdot \frac{f(x,y)}{f_Y(y)} \cdot dx$

$$= \int_0^4 x \cdot \frac{1/2}{2} \cdot dx = \frac{x}{\cancel{4}}\Bigg|_0^4$$

$$\boxed{E(X/Y) = 1}$$

So, the $E(X/Y) = 1$ is the expression for, conditional distribution of X, conditional on Y.

problem(e):-

Compute $\operatorname{cov}(X,Y)$

$$\boxed{\operatorname{cov}(X,Y) = E(XY) - E(X) \cdot E(Y)}$$

The unconditional expectation of X, $E(X)$ can be written as,

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$$

Substituting $f_X(x) = \frac{x}{2}$, for $0 < x < 4$

$$E(X) = \int_0^4 x \cdot \frac{x}{2} \cdot dx$$

$$= \frac{1}{2} \int_0^4 x^2 dx = \frac{1}{2} \left[\frac{x^3}{3} \right]_0^4$$

$$= \frac{1}{2} \times \frac{4^3}{3} = \frac{32}{3}$$

$$\boxed{E(x) = \frac{32}{3}}$$

$E(XY)$ can be derived as below,

$$E(XY) = E(E(XY/X)) \quad (\text{since } E(Y) = E(E(Y/X)))$$

$$\boxed{E(XY/X) = X \cdot E(Y/X)}$$

$$E(XY) = E(E(XY|X)) = E(X E(Y/X)) =$$

$$\boxed{E(Y/X) = \frac{x}{2}}$$ from problem (b)

$$E(XY) = E\left(x \cdot \frac{x}{2}\right) = E\left(\frac{x^2}{2}\right) = \int_{-\infty}^{\infty} \frac{x^2}{2} \cdot \left(\frac{x}{2}\right) \cdot dx$$

$$\left(\text{since } E(h(x)) = \int_{-\infty}^{\infty} h(x) \cdot f_x(x) \cdot dx\right)$$

$$= E(XY) = \int_0^4 \frac{x^3}{4} dx = \frac{1}{4} \cdot \left| \frac{x^4}{4} \right|_0^4 = \frac{4^4}{4^2} = 16$$

$$\boxed{E(XY) = 16}$$

Thus $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$

Substituting $E(X) = \frac{32}{3}$, $E(Y) = \frac{16}{3}$ $E(XY) = 16$

$$\operatorname{cov}(XY) = 16 - \frac{32}{3}\cdot\frac{16}{3}$$

$$= \frac{144 - 32\cdot 16}{9} = \frac{16(9-32)}{9} = -40$$

3. Probabilities for Patients

You might be wondering how probability theory is applied to real world problems. One example has to do with predicting patient outcomes based on their current state. As you can see, there are big prizes for this sort of analysis (https://www.cmschallenge.ai/).

Suppose a set of patients is enrolled in yearly checkups: checkup 1, checkup 2, and so on. For checkup i, let W_i be the event that the patient is well, I_i be the event that the patient is ill, and D_i be the event that the patient is dead. Transitions can be modeled as a set of conditional probabilities. For example the probability that a patient that is well in checkup i is well at the next checkup is, $P(W_{i+1}|W_i)$.

In the basic model, individuals can only stay in their current state or get sicker, so $P(W_{i+1}|I_i) = P(W_{i+1}|D_i) = P(I_{i+1}|D_i) = 0$. Dead is an abosorbing state, meaning $P(D_{i+1}|D_I) = 1$.

You can see this model drawn out in this paper, which propsed this as a method of clinical decision making (Med Decis Making 3:419-458, 1983). The notation is a little different for example $P(W_{i+1}|W_i)$ is written as P_{ww}.

A particular patient is *Well* at checkup 1 (W_1 occurs), but given their age and risk factors, $P(W_{i+1}|W_i) = 0.2$ for all i.

Let L be the random variable representing the number of checkup in which the patient is well (a well checkup). For example, here is a sample trajectory for a patient:

$$W_1, W_2, W_3, I_4, D_5$$

This patient was well at checkup 1, 2, and 3, then ill at checkup 4, and dead at checkup 5. For this patient, $L = 3$.

a. Write a complete expression for the probability mass function of L. (3 points)

(a). Given 'L' represents the rv where the no. of checks for a patient is well

Thus $P(1) = P(L=1) = P(W_1) \cdot P(\bar{W_1}) = 0.8$

$P(2) = P(L=2) = P(W_1) \cdot P(\bar{W_2})$

According to total law of probability,

$$P(W_2) = \underbrace{P\left(\frac{W_2}{W_1}\right)}_{0.2} P(W_1) + \underbrace{P\left(\frac{W_2}{I_1}\right)}_{0} P(I_1) + \underbrace{P\left(\frac{W_2}{D_1}\right)}_{0} \cdot P(D_1)$$

so, $P(W_2) = 0.2$

similarly $P(W_3) = 0.2 \cdot 0.2$

$P(W_n) = 0.2 \cdot 0.2 \cdots n-1$

so $P(x) = P(L=x) = P(W_1) \cdot P(W_2) \cdots P(W_{x-1}) \cdot P(\bar{W})$

So, $P(l) = P(L=l) = (0.2)^{l-1} \cdot (0.8)$

Hence general formula for pmf is;

$$P(l) = \begin{cases} (0.2)^{l-1}(0.8) & l = 1,2,3,\ldots\infty \\ 0 & \text{otherwise} \end{cases}$$

(b). Compute $E(L)$

$$E(L) = \mu_L = \sum_{l \in D} l \cdot p(l)$$

$$= \sum_{l=1}^{\infty} l \cdot (0.2)^{l-1}(0.8)$$

From the rule of infinite sums,

$$\sum_{k=1}^{\infty} k \cdot z^{k-1} = \frac{1}{(1-z)^2}$$

Thus, $$E(L) = (0.8)\sum_{l=1}^{\infty} l \cdot (0.2)^{l-1} = \frac{0.8}{(1-0.2)^2} = \frac{0.8}{(0.8)^2} = \frac{1}{0.8}$$

$$= \frac{10}{8} = \frac{5}{4} = 1.25$$

Thus $\boxed{E(L) = 1.25}$

```
#problem 3.c
#This method generates a random variable between 1-100 assuming
#that rv represents the number of checkups patient is well.
# if x = 3 then,  w1,w2,w3, I4 or D4 where 4th state is the exit
#state. It then calculates the pdf for each x, x=1,x=2,x=3
# finally it does the sum(x.pdf(x)) for x = 1,2,3 and returns
#the expectation of random variable


cal_single_trajectory <- function(){
  #generate a sample rv and create a list
  x <- sample(1:100,1,replace=T)
  z <- sample(0,x,replace=T)
  x_transition <- sample(0,x,replace=T)

  for(i in 1:x)
  {
    #define intial pdf variable, well and ill state probabilities
    well_state <- 0.2
    ill_state <- 0.8

    # create a transition trajectory for x
    x_transition[i] <- i

    if(i==1){
      x_pdf <- 1
    }
    else{
      exponent_well <- well_state^(i-1)
      x_pdf <- exponent_well*ill_state
    }
    z[i] <- x_pdf
  }
  print("printing pdf distribution vector z")
  print(z)
  print("transitioning vector x")
  print(x_transition)
  print("The expectation of the rv X is ")
  print(sum(x_transition*z))
  return (sum(x_transition*z))
}
```

```
#problem 3.d
##method to simulate 1000 patient trajectories and calculate their mean
sim_1000_trajectory <- function(){
  patients_1000 <- sample(0,1000,replace=T)
  for(i in 1:1000)
  {
    patients_1000[i] <- cal_single_trajectory()
  }
  print("Sample mean for each patient")
  print(patients_1000)
  patients_1000_mean <- mean(patients_1000)
  print("Sample mean for 1000 patients is")
  return (patients_1000_mean)
}
```

(e). Given $P\left(\frac{W_{i+1}}{W_i}\right) = 0.5$ for all i

from (a), we can still apply the same derivation that was implemented in (a), and pmf can be written as

$$P(\lambda) = \begin{cases} (0.5)^{\lambda-1} \cdot (0.5) & \lambda = 1, 2, 3 \ldots \\ 0 & \text{otherwise} \end{cases}$$

Now, expected number of well checkups

$$= E(L) = M_L = \sum_{\lambda \in D} p(\lambda) = \sum_{\lambda=1}^{\infty} (0.5)^{\lambda-1} (0.5) \cdot \lambda$$

$$= 0.5 \left[\frac{1}{(1-0.5)^2}\right] = \frac{0.5}{(0.5)^2} = \frac{1}{0.5} = 2$$

$$\boxed{E(L) = 2}$$

The old expectation when $P\left(\frac{W_{i+1}}{W_i}\right) = 0.2$ was, 1.25 where as new expectation when $P\left(\frac{W_{i+1}}{W_i}\right) = 2$

Hence, 0.75 checkups were added due to the use of the walker (f).

Given $p(W_3) = 0.8$

According to total law of probabilities,

$$P(W_3) = P\left(\frac{W_3}{W_2}\right)P(W_2) + \underbrace{P\left(\frac{W_3}{I_2}\right)P(I_2) + P\left(\frac{W_3}{D_2}\right).P(D_2)}_{0 \text{ according to problem definition}}$$

Hence $P(W_3) = P\left(\frac{W_3}{W_2}\right).P(W_2)$

Similarly $P(W_2) = P\left(\frac{W_2}{W_1}\right).P(W_1)$

$$\Rightarrow P(W_3) = 0.8 = P\left(\frac{W_3}{W_2}\right).P\left(\frac{W_2}{W_1}\right).P(W_1)$$

$$\Rightarrow 0.8 = P\left(\frac{W_3}{W_2}\right).P\left(\frac{W_2}{W_1}\right).1 \quad \text{——— (a)}$$

It is mentioned that the transition probability $P\left(\frac{W_{i+1}}{W_i}\right)$ is a constant

Hence $P\left(\frac{W_3}{W_2}\right) = P\left(\frac{W_2}{W_1}\right)$

Thus $\boxed{P\left(\frac{W_{i+1}}{W_i}\right)^2 = 0.8}$ (by simplifying (a))

$$P\left(\frac{W_{i+1}}{W_i}\right) = \sqrt{0.8} = 0.8944$$

$$\boxed{P\left(\frac{W_{i+1}}{W_i}\right) = 0.8944}$$

(g). Given $P(W_7) = 0.4$

18

From problem (f), we can derive a general rule,

$$\boxed{P(W_{i+1}) = P\left(\frac{W_{i+1}}{W_i}\right)^i}$$

Thus $P(W_7) = P\left(\frac{W_{i+1}}{W_i}\right)^6$

$$P\left(\frac{W_{i+1}}{W_i}\right)^6 = 0.4$$

$$P\left(\frac{W_{i+1}}{W_i}\right) = (0.4)^{1/6} = 0.858$$

thus $\boxed{P\left(\frac{W_{i+1}}{W_i}\right) = 0.858}$

(h). Study 1 and 2 have transitioning of 0.8944 and 0.858 values. Hence study 1 has P(0.106) or 10.6 patients getting unwell every year similarly study 2 has p(0.142) or 14.2 patients getting unwell every year

These values are different from the average of 10 patients becoming unwell due to the introduction of bias every time we take a root for calculating transitional probability.

The bigger the root, the bigger the deviation from mean/average of 10.

(i).

Given $P\left(\frac{W_{i+1}}{W_i}\right) = 0.2$ $P\left(\frac{I_{i+1}}{W_i}\right) = 0.8$ $P\left(\frac{D_{i+1}}{W_i}\right) = 0$

$P\left(\frac{I_{i+1}}{I_i}\right) = 0.2$ $P\left(\frac{D_{i+1}}{I_i}\right) = 0.8$

we can put the conditional probabilities into the below matrix as below:

$$\begin{array}{cc} & \begin{array}{ccc} W & I & D \end{array} \\ (i) \begin{array}{c} W \\ I \\ D \end{array} & \begin{bmatrix} 0.2 & 0.8 & 0 \\ 0 & 0.2 & 0.8 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \quad (i+1) \qquad \text{---} (a)$$

Let 'A' be a rv that represents no. of checkups in which patient is alive

$$\boxed{A = W + I}$$

$$E(A) = E(W+I) = E(W) + E(I)$$

$$E(W) = \mu_W = \sum_{w \in D} w \cdot p(w)$$

(from problem (a), we know,

$$p(x) = \begin{cases} (0.2)^{x-1}(0.8) & x = 1, 2, 3 \ldots \\ 0 & \text{otherwise} \end{cases}$$)

Thus $P(W_1) = 1$,

$$P(W_2) = P\left(\frac{W_2}{W_1}\right) \cdot P(W_1) + P\left(\frac{W_2}{I_1}\right) \cdot P(I_1) + P\left(\frac{W_2}{D_1}\right) P(D_1)$$

$$= 0.2 + 0 + 0 = 0.2 \qquad \text{— (from (a))}$$

Similarly $P(W_i) = \begin{cases} (0.2)^{w-1}(0.8) & w = 1,2,3\ldots \\ 0 & \text{otherwise} \end{cases}$

thus $E(W) = \sum_{w=1}^{\infty} w\cdot(0.2)^{w-1}(0.8) = \frac{(0.8)}{(1-0.2)^2} = \frac{5}{4} = 1.25$

$E(I_2)$ can be derived as below:-

$P(I_2) = P\left(\frac{I_2}{I_1}\right)P(I_1) + P\left(\frac{I_2}{W_1}\right)\cdot P(W_1) + \underbrace{P\left(\frac{I_2}{D_1}\right)P(D_1)}_{0}$ (since in matrix (a) $P\left(\frac{I_2}{D_1}\right) = 0$)

$= (0.8)\cdot 1 + (0.2)\cdot 1$

$P(I_2) = 1$

we can derive

$P(I_n) = \begin{cases} 1 & \text{when } i = 1,2,3\ldots \\ 0 & \text{otherwise} \end{cases}$

Thus $E(I_2) = \frac{1+0}{2} = \frac{1}{2}$ (since I is uniform distribution $E(I) = \frac{a+b}{2}$)

$E(A) = E(W) + E(I) = 1.25 + 0.5$

$= 1.75$

Thus expected no. of checkups in which patient is alive (W & I) is $\underline{1.75}$

4. Wise Investments

You invest in two startup companies focused on data science, Company 1 and Company 2. Let U_i be the event that Company i reaches unicorn status (valued at \$1 billion). Thanks to your growing expertise in this area, $P(U_1) = P(U_2) = 3/4$. U_1 and U_2 may or may not be independent of each other. Let random variable X be the total number of companies that reach unicorn status. X can take on the values 0, 1, and 2.

a. Compute $E(X)$. (3 points)

b. Compute the maximum value for $var(X)$. (3 points)

c.* Compute the minimum value for $var(X)$. (3 points)

Now suppose there are four companies, Company 1 through Company 4. Again, let U_i be the event that Company i reaches unicorn status. $P(U_i) = 3/4$ for all i. Once again, let random variable X be the total number of companies that reach unicorn status.

d.* Compute the minimum and maximum values for $var(X)$. (3 points)

(a). Given $X = \{$total no. of companies reaching unicorn status$\}$

Let $Y_1 = \{$~~event~~ RV that U_1 reaches unicorn status$\}$

$Y_2 = \{$~~event~~ RV that U_2 reaches unicorn status$\}$

It can be easily inferred that Y_1 and Y_2 are two Bernoulli random variables

$$Y_1 = \begin{cases} 1 & \text{if } U_1 \text{ reaches unicorn status} \\ 0 & \text{if } U_1 \text{ does not reach unicorn status} \end{cases}$$

$$Y_2 = \begin{cases} 1 & \text{if } U_2 \text{ reaches unicorn status} \\ 0 & \text{if } U_2 \text{ does not reach unicorn status} \end{cases}$$

If U_1 and U_2 are not independent :- Random variable X can be defined as a sum of two RVs Y_1 and Y_2, where $Y_1 = 1$ and $Y_2 = 1$

$$X = Y_1 + Y_2$$

Hence $E(X) = E(Y_1 + Y_2)$

$= E(Y_1) + E(Y_2)$ (Rules of expectation. $E(A+B) = E(A) + E(B)$)

$$E(Y_1) = \sum y_1 P(y_1) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{3}{4} = \frac{3}{4}$$

$$E(Y_2) = \sum y_2 \cdot P(y_2) = 0 \cdot 1/4 + 1 \cdot 3/4 = 3/4$$

Hence, $E(Y_1) + E(Y_2) = 2 \cdot 3/4 = 3/2$

Thus $\boxed{E(X) = \frac{3}{2}}$

(b). maximum $var(X)$

$E(X) = 3/2$

$E(X) = E(Y_1) + E(Y_2)$

similarly

variance $V(X) = V(Y_1) + V(Y_2) + 2\,cov(Y_1, Y_2)$

Y_1 and Y_2 are perfectly co-related as their variance is same ~~var~~ variance

Hence, the maximum and minimum values of variance depends on the covariance of Y_1 and Y_2

According to rules of corelation, for any two rv's X and Y, $\left[-1 \leq corr(X, Y) \leq 1\right]$ — (a)

$$corr(Y_1, Y_2) = \frac{cov(Y_1, Y_2)}{\sqrt{var(Y_1) \cdot var(Y_2)}}$$

(According to Bernoulli rule $var(X_i) = P(1-P) = Pq$)

$$-1 \leq \frac{cov(Y_1, Y_2)}{\sqrt{var(Y_1) \cdot (var(Y_2))}} \leq 1 \quad \text{— from (a)}$$

$$cov(Y_1, Y_2) = \sqrt{var(Y_1) \cdot var(Y_2)} \quad \text{(when correlation of } Y_1, Y_2 \text{ is 1)}$$

$$cov(Y_1, Y_2) = \sqrt{\left(\tfrac{3}{16}\right) \cdot \left(\tfrac{3}{16}\right)} = \frac{3}{16} \quad \left(var(X) = \left(\tfrac{3}{4}\right)\left(\tfrac{1}{4}\right) = \left(\tfrac{3}{16}\right)\right)$$

substituting $cov(Y_1, Y_2) = \frac{3}{16}$ in variance,

$$var(X) = var(Y_1) + var(Y_2) + 2\,cov(Y_1, Y_2)$$
$$= \frac{3}{16} + \frac{3}{16} + \left(\frac{3}{16}\right) \cdot 2$$

[Note:- According to Bernoulli rule $var(X) = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$]

$$var(X) = 4\left(\frac{3}{16}\right) = \frac{3}{4}$$

Thus maximum $Var(X) = \frac{3}{4}$

(c). minimum $Var(X)$

from (b), correlation for two variables Y_1 and Y_2 is minimum of -1

Thus $$\frac{cov(Y_1, Y_2)}{\sqrt{Var(Y_1) \cdot Var(Y_2)}} \geq -1$$

$$\Rightarrow cov(Y_1, Y_2) \geq -\sqrt{\left(\frac{3}{16}\right)\left(\frac{3}{16}\right)}$$

$$cov(Y_1, Y_2) = -\frac{3}{16}$$

substituting $cov(Y_1, Y_2) = -\frac{3}{16}$ in variance,

$$Var(X) = Var(Y_1) + Var(Y_2) + 2\,cov(Y_1, Y_2)$$

$$= \frac{3}{16} + \frac{3}{16} - 2 \cdot \left(\frac{3}{16}\right)$$

$$= 0$$

$\boxed{Var(X) = 0}$ is the minimum value of variance

$Var(X) = 0$ for two companies

(d). var(X) max and minimum

From problem (c) and (b),

for 2 rv's

$$var(X) = var(Y_1) + var(Y_2) + 2\,cov(Y_1, Y_2)$$

$$= \left(\frac{3}{16}\right) + \left(\frac{3}{16}\right) + 2\,cov(Y_1, Y_2)$$

For problem (d), Let Y_1, Y_2, Y_3, Y_4 be 4 rv's representing 4 companies

$$Var(X) = Var(Y_1) + Var(Y_2) + Var(Y_3) + Var(Y_4) + 2 \cdot \sum_{1<i<j<4}^{4} cov(Y_i, Y_j)$$

For minimum:-

$$cov(Y_1, Y_2) = \frac{-3}{16}$$ (when correlation for 2 rv's is -1)

$$Var(X) = 4\left(\frac{3}{16}\right) - 2\left(\frac{3}{16}\right) \cdot 6$$

(Since all possible values of cov are (1,2), (1,3), (1,4), (2,3), (2,4), (3,4)) × 2

$$= \frac{12}{16} - \frac{36}{16} = \frac{-24}{16} = -\frac{3}{2} = -1.5$$

$$\boxed{\text{minimum of } var(X) = -1.5}$$

maximum:-

For maximum variance $cov(Y_1, Y_2) = \frac{3}{16}$ when

Y_1, Y_2 are zrv's

Similarly for 4 rv's (Y_1, Y_2, Y_3, Y_4), $\left[Var(X) = 4\left(\frac{3}{16}\right) + 2 \sum_{1 \le i < j \le 4} cov(Y_i, Y_j)\right]$

$$var(X) = 4\left(\frac{3}{16}\right) + 2 \cdot 6 \cdot \left(\frac{3}{16}\right)$$

$$= \frac{36}{16} + \frac{12}{16} = \frac{48}{16} = 4$$

maximum of 4 random Y_1, Y_2, Y_3, Y_4 is $\boxed{Var(X) = 4}$