DATA SCIENCE CSE558

MONSOON SEMESTER 2023

ASSIGNMENT 1, HARJEET SINGH YADAV, 2020561

Q3

a) To determine whether the given measure $P(A) = \frac{|A|}{|\Omega|}$ is a probability measure, we need to check if it satisfies the three axioms of probability measures: non-negativity, additivity and normalisation.

1) **Non-negativity:** The given probability measure is defined as the ratio of the cardinality of A and Ω, both are non-negative integers. Therefore, $P(A) \geq 0$ for all event A in the σ - algebra F.

2) **Additivity:** The measure P(A) is additive if for a set of disjoints events $A_i \epsilon F$, it satisfies: $P(\bigcup_{I=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

$$\Rightarrow P(\bigcup_{I=1}^{\infty} A_i) = \frac{|\bigcup_{I=1}^{\infty} A_i|}{|\Omega|}$$

By, the properties of the union of disjoint sets, we know that:

$$|\bigcup_{I=1}^{\infty} A_i| = \sum_{i=1}^{\infty} |A_i|$$

Substituting into the above equation:

$$\Rightarrow P\left(\bigcup_{I=1}^{\infty} A_i\right) = \frac{\sum_{i=1}^{\infty} |A_i|}{|\Omega|} = \sum_{i=1}^{\infty} \frac{|A_i|}{|\Omega|} = \sum_{i=1}^{\infty} P(A_i)$$

$$\Rightarrow P\left(\bigcup_{I=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Hence we can see that the measure satisfies the additivity property.

3) **Normalization condition:** A probability measure must satisfy the normalisation condition, which means that the probability of the entire sample space Ω is 1: $P(\Omega) = 1$.

In this case, we have $P(\Omega) = \frac{|\Omega|}{|\Omega|} = 1$

So, the normalisation condition is satisfied.

Therefore, based on the analysis of the three axioms, the defined measure $P(A) = \frac{|A|}{|\Omega|}$ is indeed a probability measure.

b) Upper bound:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k)$$

Proof by Induction:

n = 1 & 2 & 3 given with the condition and principal of inclusion & exclusion.

n = 4

$$P\left(\bigcup_{i=1}^{4} A_i\right) \le \sum_{1\le i\le 4} P(A_i) - \sum_{1\le i<j\le 4} P(A_i \cap A_j) + \sum_{1\le i<j<k\le 4} P(A_i \cap A_j \cap A_k)$$

Let it valid for n = k

$$P\left(\bigcup_{i=1}^{k} A_i\right) \le \sum_{1\le i\le k} P(A_i) - \sum_{1\le i<j\le k} P(A_i \cap A_j) + \sum_{1\le i<j<m\le k} P(A_i \cap A_j \cap A_k)$$

To show it is valid for n = k+1

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) = P\left(\left\{\bigcup_{i=1}^{k} A_i\right\} \cup A_{k+1}\right) = P\left(\left\{\bigcup_{i=1}^{k} A_i\right\}\right) + P(A_{k+1}) - P\left(\left\{\bigcup_{i=1}^{k} A_i\right\} \cap A_{k+1}\right)$$

$$\le \sum_{1\le i\le k} P(A_i) + P(A_{k+1}) - \sum_{1\le i<j\le k} P(A_i \cap A_j) - P\left(\bigcup_{i=1}^{k} A_i \cap A_{k+1}\right) + \sum_{1\le i<j<m\le k} P(A_i \cap A_j \cap A_k)$$

$$\le \sum_{1\le i\le k+1} P(A_i) - \sum_{1\le i<j\le k+1} P(A_i \cap A_j) + \sum_{1\le i<j<m\le k} P(A_i \cap A_j \cap A_k)$$

$$+ \sum_{1\le i<j\le k} P(A_i \cap A_j \cap A_{k+1})$$ (Adding this term won't affect the equality sign)

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) \le \sum_{1\le i\le k+1} P(A_i) - \sum_{1\le i<j\le k+1} P(A_i \cap A_j) + \sum_{1\le i<j<m\le k+1} P(A_i \cap A_j \cap A_k)$$

This equation is valid for n = k+1. Therefore, we can generalise using the proof of induction for any n.

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i<j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i<j<k \leq n} P(A_i \cap A_j \cap A_k)$$

Lower Bound:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i<j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i<j<k \leq n} P(A_i \cap A_j \cap A_k)$$

$$- \sum_{1 \leq i<j<x<y \leq n} P(A_i \cap A_j \cap A_x \cap A_y)$$

Proof by Induction:

n = 1 & 2 & 3 given with the condition and principal of inclusion & exclusion.

n = 5

$$P\left(\bigcup_{i=1}^{5} A_i\right) \geq \sum_{1 \leq i \leq 5} P(A_i) - \sum_{1 \leq i<j \leq 5} P(A_i \cap A_j) + \sum_{1 \leq i<j<k \leq 5} P(A_i \cap A_j \cap A_k)$$

$$- \sum_{1 \leq i<j<x<y \leq 5} P(A_i \cap A_j \cap A_x \cap A_y)$$

Let it valid for n = k

$$P\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{1\leq i\leq k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<x\leq k} P(A_i \cap A_j \cap A_x)$$

$$- \sum_{1\leq i<j<x<y\leq k} P(A_i \cap A_j \cap A_x \cap A_y)$$

To show it is valid for n = k+1

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) = P\left(\left\{\bigcup_{i=1}^{k} A_i\right\} \cup A_{k+1}\right) = P\left(\left\{\bigcup_{i=1}^{k} A_i\right\}\right) + P(A_{k+1}) - P\left(\left\{\bigcup_{i=1}^{k} A_i\right\} \cap A_{k+1}\right)$$

$$\geq \sum_{1\leq i\leq k} P(A_i) + P(A_{k+1}) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) - P\left(\bigcup_{i=1}^{k} A_i \cap A_{k+1}\right)$$

$$+ \sum_{1\leq i<j<m\leq k} P(A_i \cap A_j \cap A_k) - \sum_{1\leq i<j<x<y\leq n} P(A_i \cap A_j \cap A_x \cap A_y)$$

$$\geq \sum_{1\leq i\leq k+1} P(A_i) - \sum_{1\leq i<j\leq k+1} P(A_i \cap A_j) + \sum_{1\leq i<j<m\leq k} P(A_i \cap A_j \cap A_k)$$

$$- \sum_{1\leq i<j<x<y\leq k} P(A_i \cap A_j \cap A_x \cap A_y)$$ (Using eq for n = k and 2 given eqs.)

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) \geq \sum_{1\leq i\leq k+1} P(A_i) - \sum_{1\leq i<j\leq k+1} P(A_i \cap A_j) + \sum_{1\leq i<j<x\leq k+1} P(A_i \cap A_j \cap A_x)$$

$$- \sum_{1\leq i<j<x<y\leq k+1} P(A_i \cap A_j \cap A_x \cap A_y)$$

This equation is valid for n = k+1. Therefore, we can generalise using the proof of induction for any n.

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{1 \le i \le n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < x \le n} P(A_i \cap A_j \cap A_x)$$

$$- \sum_{1 \le i < j < x < y \le n} P(A_i \cap A_j \cap A_x \cap A_y)$$

Q5

a) Let Xi be the random variable: Number of times we need to roll the dice to get $\lfloor\sqrt{k}\rfloor$.

Clearly, Xi is a geometric random variable {n, $(1 - p)^{(n-1)} p$ } where p = 1/k (uniformly distributed)

E[Xi] = $1p + 2(1 - p)p + 3(1 - p)^2 p + 4(1 - p)^3 p +$ ......

E[Xi] = $\sum_{n=1}^{\infty} n(1 - p)^{(n-1)} p = p * \sum_{n=1}^{\infty} n(1 - p)^{(n-1)}$

E[Xi] = $p * \frac{1}{(1-1+p)^2} = p * \frac{1}{p^2} = 1/p$

Over Expectation = $1/p = 1/(1/k) = k$ i.e we need k rolls on average until we see $\lfloor\sqrt{k}\rfloor$.

b) Let Y be a random variable that governs number of rolls we need to get each face at least once.

Let Xi be the random variable: Number of times we need to roll the dice to get i^{th} new face with $p_i = \frac{k}{k-i+1}$ success. Clearly, Xi is a geometric random variable. With the proof used in part a), we can see that the expectation of E[Xi] = 1/p.

So, Expectation of Y (no of rolls we need to see at least one coupon of each type)

$$E[\sum_{i=1}^{k} X_i] = \sum_{i=1}^{k} E[X_i] = \sum_{i=1}^{k} (1/p_i) = \sum_{i=1}^{k} \frac{k}{k-i+1} \simeq k\, log(k)$$

c) Solution 1:

Assume there are four faces on the die i.e. add one more face to the die with face value 2. Now the die has [1, 2, 2, 3] faces and now the probability of $P(2) = 1/4$ has been divided between two 2s, and each face has equal and uniform probability.

of $P(1) = P(2) = P(2) = P(3) = 1/4$

Now, this problem is similar to part b) where k = 4 and the probability of each face is ¼.

Using the formula we proved in the previous part:

$$\sum_{i=1}^{k} E[X_i] = \sum_{i=1}^{k} (1/p_i) = (4/4 + 4/3 + 4/2 + 4/1) = 25/3$$

Now, we need to subtract the expectation of extra (2) face. I.e.

$$\sum_{i=1}^{k} E[X_i] - E[X = (extra\ 2)] = 25/3 - (1/(1/2)) = 25/3 - 2 = 19/3$$

So, over expectation we need 19/3 = 6.33 rolls to see number 1 to 3 at least once on its upward face.

Solution 2:

Treating it as a Poisson random variable and calculating the expectation we get.

$E[T] = \int_0^{\infty} \left(1 - \prod_{j=1}^{m} \left(1 - e^{-p_i t}\right)\right) dt$ (Reference: 10th edition of *Introduction to Probability Models* by Sheldon Ross (page 322).)

where there are m events with probability $p_j\ j = 1, 2, 3, .. m$

In our case 1 <= j <= 3. There are three events. With probability

$P(1) = P(3) = 1/4\ and\ P(2) = 1/2$

And putting them into integral and solving, we get.

$$E[T] = \int_0^{\infty} \left(1 - \left(1 - e^{-t/4}\right)\left(1 - e^{-t/4}\right)\left(1 - e^{-t/2}\right)\right) dt = 19/3$$

THE END