Lecture 5: Discrete Random Variables

Yunwen Lei

School of Computer Science, University of Birmingham

1 Random Variable

Random variables are important concepts in probability. We first introduce some examples to motivate the introduction of random variables.

Motivation: Mathematicians Are Lazy. Suppose we toss a coin and are interested in the probability of getting a head. The formulation of this probability is easy

$$\text{“probability of getting a head”} = \mathbb{P}[\text{“Head”}]$$

Suppose we toss 3 coins and are interested in the probability of getting 3 heads. The formulation needs more space

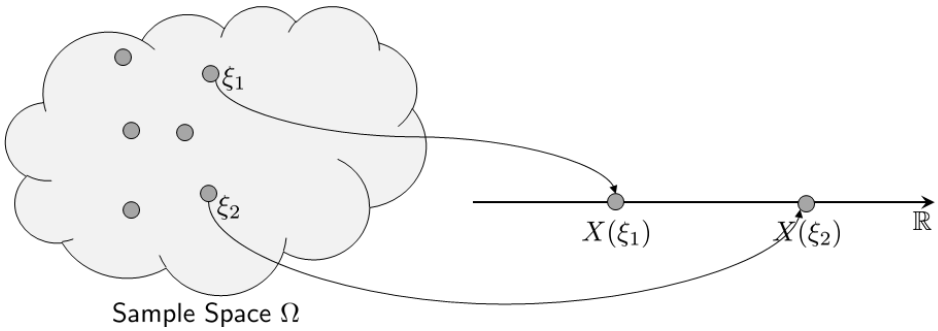
$$\text{“probability of getting 3 heads”} = \mathbb{P}[\text{“Head”} \cap \text{“Head”} \cap \text{“Head”}]$$

Mathematicians are lazy: Call “Head”=1 and “Tail”=0. Let X be the number of heads. Then the formulation becomes easy

$$\text{“probability of getting a head”} = \mathbb{P}[X = 1]$$
$$\text{“probability of getting 3 heads”} = \mathbb{P}[X = 3]$$

Then, with the introduction of a new variable X, it is equally easy to formulate the probability of getting n heads for any n. This X is an example of *random variable*. In a random experiment, the outcome can be of any type, i.e., words. Mathematicians are interested in numbers since numbers are more direct to handle. Intuitively speaking, random variables are functions that translate words to numbers. In the above example, “Head” is a word description. “$X = 1$” is numerical description

Definition 1 (Random Variable)**.** A Random Variable X is a function $X : \Omega \mapsto \mathbb{R}$ that maps an outcome $\xi \in \Omega$ to a number $X(\xi) \in \mathbb{R}$.



We use $X(\Omega)$ or R_X to denote the range of X, i.e., $X(\Omega) = \{X(\xi) : \xi \in \Omega\}$. The range consists of all the possible realizations of a random variable in the experiment. We have two types of random variables depending on the range

- X is a discrete random variable if $X(\Omega)$ is countable
- X is a continuous random variable if $X(\Omega)$ is uncountable

Example 1 (Toss 3 Coins)**.** For the experiment of tossing 3 coins, we define three variables as follows

	Sample Space Ω								
	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT	
$P(\xi)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	
$\mathbf{X}(\xi)$	3	2	2	1	2	1	1	0	← number of heads
$\mathbf{Y}(\xi)$	1	0	0	0	0	0	0	1	← matching tosses
$\mathbf{Z}(\xi)$	8	2	2	$\frac{1}{2}$	2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$	← H: double your money T: halve your money

- X denotes the number of heads
- Y is an indicator whether the three numbers in the three toss are the same
- In the beginning, we have 1 pound. We double our money if a coin shows the head, and halve our money if a coin shows the tail. Z denotes the money after 3 coin toss.

A natural question is how to compute the probability of a random variable taking each value. Consider the random variable X in the above example. How to calculate $\mathbb{P}(X = 1)$? The first problem is that $\{X = 1\}$ is not an event in the event space. Note the sample space is

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$$

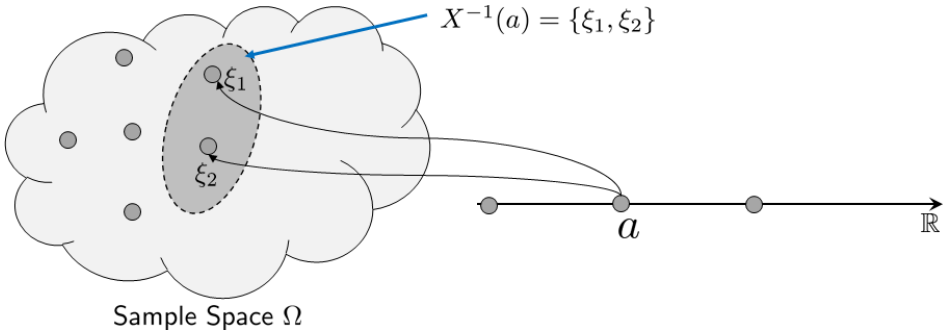
Recall that the event space is the combination of these outcomes. Then "$\{X = 1\}$" is not the same as HTT,THT or TTH, although each of the three outcomes corresponds to $X = 1$. Then, how to measure the probability. The basic idea is to map $\{X = 1\}$ back to the sample space, and compute the probability as follows

$$\mathbb{P}(\{X = 1\}) = \mathbb{P}\big(\{\text{HTT, THT, TTH}\}\big) = \frac{3}{8}$$

Essentially, you are finding ξ such that $X(\xi) = 1$

$$\mathbb{P}(X = 1) = \mathbb{P}(X(\xi) = 1) = \mathbb{P}(\xi \in X^{-1}(1)) = \mathbb{P}\big(\{\text{HTT, THT, TTH}\}\big) = \frac{3}{8}$$

The reason to mapping back to the sample space is that we only have the probability definition in the event space. We should map the realization of the random variable back to the event space. Then we can use the original definition of probability. This is called the *inverse mapping*: when calculating the probability, go backward to the sample space!



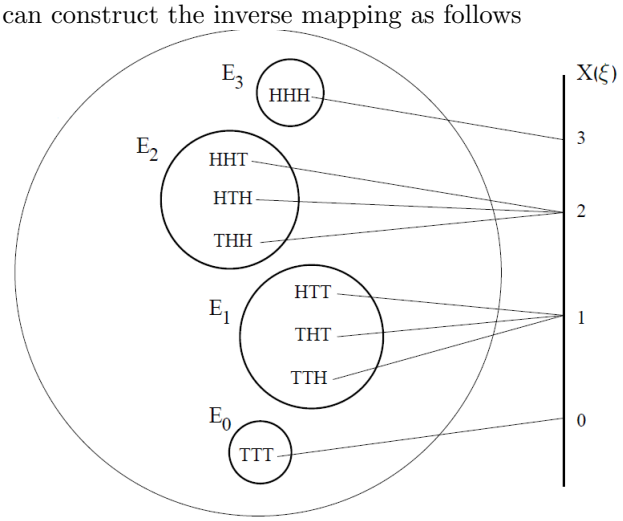
- Important! $X^{-1}(a)$ is a set; it is an event in the event space

$$X^{-1}(a) = \{\xi \in \Omega : X(\xi) = a\}$$

- You can measure an event using $\mathbb{P}$

$$\mathbb{P}(X = a) = \mathbb{P}(X^{-1}(a))$$

Example 2. Let consider the experiment of tossing three coins and define X as the number of heads. We can construct the inverse mapping as follows



Graphical representation of X

$$X^{-1}(3) = \{\text{HHH}\} = E_3$$
$$X^{-1}(2) = \{\text{HHT,HTH,THH}\} = E_2$$
$$X^{-1}(1) = \{\text{HTT,THT,TTH}\} = E_1$$
$$X^{-1}(0) = \{\text{TTT}\} = E_0$$

- The events E_0, E_1, E_2, E_3 are disjoint since $X(\xi)$ is a function! This guarantees that X cannot map any s in the sample space to *two* values.
- $X : \Omega \mapsto \mathbb{R}$ must be defined for all $\xi \in \Omega$.

Then we can use the probability defined in the sample space Ω to define the probability for the *random variable*:

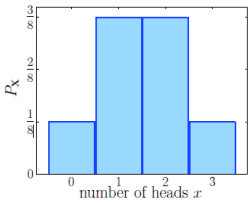
$$\underbrace{\{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}}_{\Omega} \xrightarrow{X} \underbrace{\{0, 1, 2, 3\}}_{X(\Omega)}$$

Each possible value x of the random variable X corresponds to an event

x	0	1	2	3
Event	{TTT}	{HTT, THT, TTH}	{HHT, HTH, THH}	{HHH}

For each $x \in X(\Omega)$, compute $\mathbb{P}(X = x)$ by adding the outcome-probabilities. The left hand side illustrates the probability via table, while the right hand side shows the probability via a figure

	possible values $x \in \mathbf{X}(\Omega)$			
x	0	1	2	3
$P_{\mathrm{X}}(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



This is indeed called the probability mass function.

Definition 2 (Probability mass function (PMF))**.** The Probability Mass Function $P_X(a)$ is the probability for the random variable X to take value a

$$P_X(a) = \mathbb{P}(X = a).$$

Note there are two functions here

- Function X: the random variable which translates words to numbers
- Function P_X: the mapping from event $\{X = a\}$ to a probability. This defines the probability for the random variable.

Note the above involves two variables X and a. Here is the difference

- X is a random variable (a map). Technically it should be $X(\xi)$
- a is a state, which is not random. The notation $X = a$ means X is taking the state a.

Below we show a basic property of the PMF.

Theorem 1. *A PMF P_X should satisfy the condition $\sum_{x\in X(\Omega)} P_X(x) = 1$.*

Proof. According to the definition of probability for random variable, we know

$$\sum_{x\in X(\Omega)} \mathbb{P}(X = x) = \sum_{x\in X(\Omega)} \mathbb{P}\big(\{\xi \in \Omega : X(\xi) = x\}\big).$$

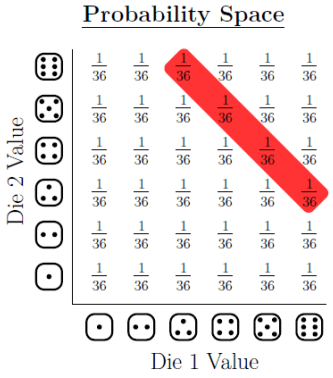
Since the event $\{\xi \in \Omega : X(\xi) = x\}$ are disjoint for different x, we can use the additivity property on disjoint events to show

$$\sum_{x\in X(\Omega)} \mathbb{P}(X = x) = \mathbb{P}\Big(\cup_{x\in X(\Omega)} \{\xi \in \Omega : X(\xi) = x\}\Big) = \mathbb{P}(\Omega) = 1.$$

This finishes the proof. □

Example 3 (PMF for the Sum of Two Dice)**.** We now calculate the PMF for a random variable X as the sum of two dice. The left hand side shows the sample space as well as the probability for each outcome. The red area shows the event corresponding to $X = 9$, i.e., $X^{-1}(9)$. Then

$$P_X(9) = \mathbb{P}(X = 9) = \mathbb{P}(\{\xi : X(\xi) = 9\}) = \mathbb{P}(\{(3,6),(4,5),(5,4),(6,3)\}) = 1/9.$$

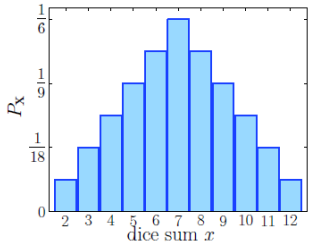


$X = 9$ has four outcomes,

$$\mathbb{P}[X = 9] = 4 \times \tfrac{1}{36} = \tfrac{1}{9}.$$

Possible sums are $X \in \{2, 3, \ldots, 12\}$ and

x	2	3	4	5	6	7	8	9	10	11	12
$P_X(x)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$



The PMF can only describe the distribution of a discrete random variable. Can we describe the distribution for any kind of random variable (discrete, continuous, and mixed)? This question motivates the introduction of Cumulative Distribution Function (CDF).

Definition 3 (Cumulative Distribution Function)**.** The Cumulative Distribution Function $F_X(x)$ is the probability for the random variable X to be at most x

$$F_X(x) = \mathbb{P}(X \le x).$$

Below we list some properties of CDF.

- $F_X(x)$ is a non-decreasing function of x.
- $F_X(-\infty) = 0$ and $F_X(\infty) = 1$
- $\mathbb{P}(a < X \le b) = F_X(b) - F_x(a)$.

The first property is clear from the definition. As x increases, the event $\{\xi : X(\xi) \le x\}$ is becoming larger and larger. $F_X(-\infty) = 0$ shows that it is impossible for X to be smaller than or equal to $-\infty$, while $F_X(\infty) = 1$ shows that it is certain X would be smaller than or equal to ∞. Below we give a proof for the last property.

Proof. According to the definition of probability for random variable we know

$$\mathbb{P}(a<X\le b)+\mathbb{P}(X\le a)=\mathbb{P}\big(\{\xi : X(\xi)\le b\}\cap\{\xi : X(\xi)>a\}\big)+\mathbb{P}\big(\{\xi : X(\xi)\le a\}\big).$$

The event $\{\xi : X(\xi) \le b\} \cap \{\xi : X(\xi) > a\}$ and the event $\{\xi : X(\xi) \le a\}\big)$ are disjoint. Therefore, by the additivity of probability on disjoint events, we know

$$\begin{aligned}&\mathbb{P}\big(\{\xi : X(\xi)\le b\}\cap\{\xi : X(\xi)>a\}\big)+\mathbb{P}\big(\{\xi : X(\xi)\le a\}\big)\\ &\qquad=\mathbb{P}\big(\big(\{\xi : X(\xi)\le b\}\cap\{\xi : X(\xi)>a\}\big)\cup\{\xi : X(\xi)\le a\}\big)=\mathbb{P}\big(\{\xi : X(\xi)\le b\}\big),\end{aligned}$$

where the last identity uses the fact

$$\big(\{\xi : X(\xi)\le b\}\cap\{\xi : X(\xi)>a\}\big)\cup\{\xi : X(\xi)\le a\}=\{\xi : X(\xi)\le b\}.$$

This shows that

$$\mathbb{P}(a<X\le b)+\mathbb{P}(X\le a)=\mathbb{P}(X\le b).$$

□

Example 4 (CDF for Tossing Three Coins)**.** We now consider the experiment of tossing three coins. The sample space is

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}.$$

Let $X(\xi) =$ the number of heads. According to a previous example, we know

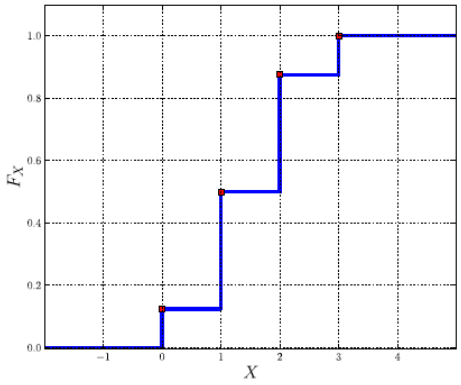
$$P_X(0)=\frac{1}{8},\quad P_X(1)=\frac{3}{8}\quad P_X(2)=\frac{3}{8},\quad P_X(3)=\frac{1}{8}$$

According to the definition, we then have the cumulative distribution function. The right hand side shows the graph for the PDF

$$\begin{aligned}F_X(-1)&=\mathbb{P}(X\le -1)=0\\ F_X(0)&=\mathbb{P}(X\le 0)=1/8\\ F_X(1)&=\mathbb{P}(X\le 1)=4/8\\ F_X(2)&=\mathbb{P}(X\le 2)=7/8\\ F_X(3)&=\mathbb{P}(X\le 3)=1\end{aligned}$$

We see, for example, that

$$\mathbb{P}(0<X\le 2)=F_X(2)-F_X(0)=6/8$$



The graph of the PDF for X

One can transform the information of PMF to CDF and vice versa. Below we show how to compute CDF from PMF and how to compute PMF from CDF.

Remark 1 (Connection Between PMF and CDF)**.** Suppose X has range $X(\Omega) = \{x_1, x_2, x_3, \ldots\}$ with $x_i < x_{i+1}$.

CDF from PMF: According to the definition of CDF, we know

$$F_X(x)=\mathbb{P}(X\le x)=\sum_{x_k\le x}\mathbb{P}(X=x_k)=\sum_{x_k\le x}P_X(x_k).$$

That is, we just take the summation of PMF over all $x_k \le x$

PMF from CDF: Since $\{X = x_k\} \cup \{X \le x_{k-1}\} = \{X \le x_k\}$ we know (X cannot take any number in (x_{k-1}, x_k))

$$P_X(x_k)=\mathbb{P}(X=x_k)=\mathbb{P}(X\le x_k)-\mathbb{P}(X\le x_{k-1})=F_X(x_k)-F_X(x_{k-1}).$$

2 Special Distributions

In this section, we introduce several common distributions.

2.1 Bernoulli Distribution

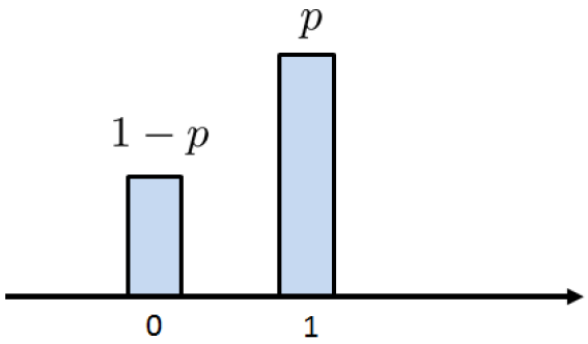
Bernoulli Distribution is related to a random variable which has only two outcomes: a success or a fail. It is motivated from the experiment of tossing a coin. In this case, one can interpret the head as the success and the tail as the failure.

Definition 4 (Bernoulli Distribution)**.** Suppose you have a coin where the probability of a heads is p and we define the random variable

$$X = \text{“the number of heads showing on one tossed coin”}$$

Then we say that X is distributed according to the Bernoulli Distribution with parameter p, and write this as

$$X \sim \text{Bernoulli}(p).$$



2.2 Binomial Random Variable

The Binomial Random Variable is a sum of Bernoulli random variable. We define

$$X = \text{the number of heads in } n \text{ independent coin tosses with probability } p \text{ of heads}$$

$$X = X_1 + X_2 + \cdots + X_n \quad \leftarrow \text{sum of } n \text{ independent Bernoullis, } X_i \sim \text{Bernoulli}(p)$$

n=5,X=3: HHHTT HHTTH HTTHH TTHHH HHTHT HTHTH THTHH HTHHT THHTH THHHT ← each has probability $p^3(1-p)^2$ (independence)

$$\mathbb{P}[X = 3 \mid n = 5] = 10p^3(1-p)^2 \quad \leftarrow \text{add outcome probabilities}$$

In general, there are $\binom{n}{k}$ sequences with k heads. Each sequence has probability $p^k(1-p)^{n-k}$: each head happens with probability p, each tail happens with probability $1-p$. Then, the probability of having k heads out of n coin toss is

$$\mathbb{P}[X = k|n] = \binom{n}{k} p^k (1-p)^{n-k}.$$

This motivates the definition of binomial distribution as follows.

Definition 5 (Binomial Distribution)**.** X is the number of successes in n independent trials with success probability p on each trial: $X = X_1 + \cdots + X_n$, where $X_i \sim \text{Bernoulli}(p)$

$$P_X(k) = B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.1}$$

We use the notation $B(n, p)$ to denote a binomial distribution with parameter n and p.

Remark 2 (Origin of Binomial Random Variables)**.** Toss a coin 3 times with probability p of heads.

- Find the probability of getting 3 heads

$$P_X(3) = \mathbb{P}(\{\text{HHH}\}) = p^3$$

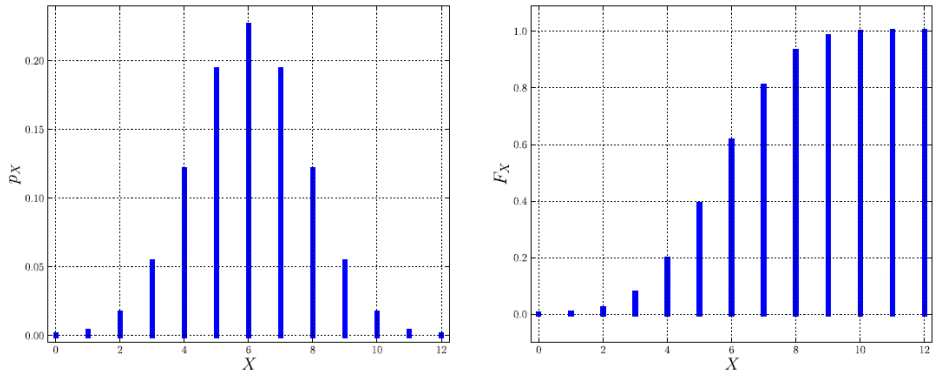
- Find the probability of getting 2 heads

$$\begin{aligned} P_X(2) &= \mathbb{P}\big(\{\text{HHT}, \text{HTH}, \text{THH}\}\big) \\ &= p^2(1-p) + p^2(1-p) + p^2(1-p) = 3p^2(1-p). \end{aligned}$$

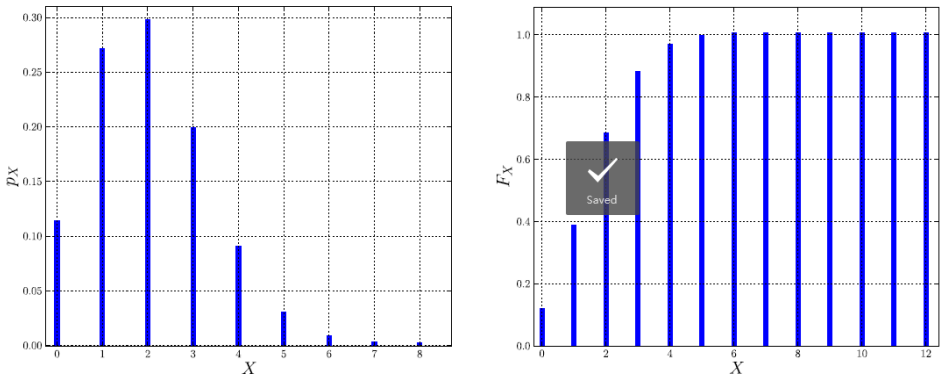
- In general,

$$P_X(k) = \underbrace{\binom{n}{k}}_{\text{number of combinations}} \underbrace{p^k}_{\text{prob getting } k \text{ H's}} \underbrace{(1-p)^{n-k}}_{\text{prob getting } n-k \text{ T's}}$$

Below we plot the PMF for Binomial Random Variables with different n and p:



Binomial Mass and Distribution Functions for $n = 12$ and $p = 1/2$



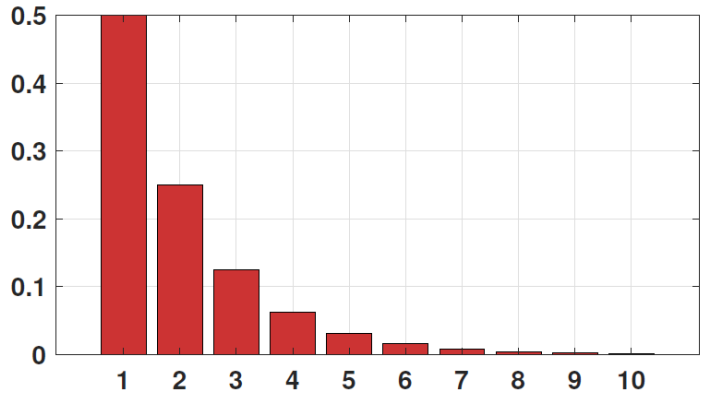
Binomial Mass and Distribution Functions for $n = 12$ and $p = 1/6$

2.3 Geometric Distribution

Suppose you toss a coin repeatedly until you see a head. This requires you to have $n-1$ tails, and then followed by a head in the n-th coin toss. The probability of this sequence of events are $1/2, 1/4, 1/8, \ldots$, which forms an infinite series

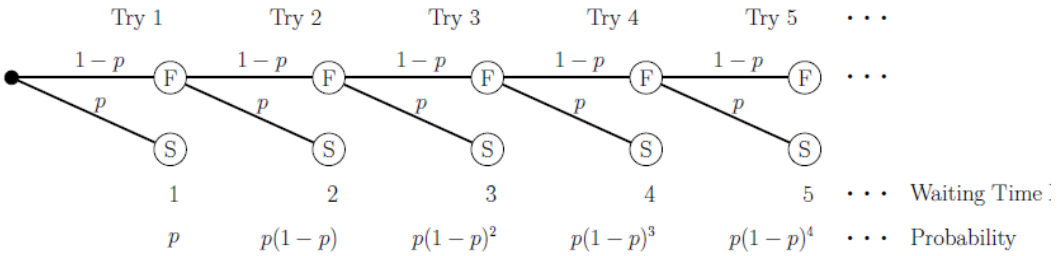
$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{8}$ $\frac{1}{16}$

We can plot the PMF of the above random variable in the following figure



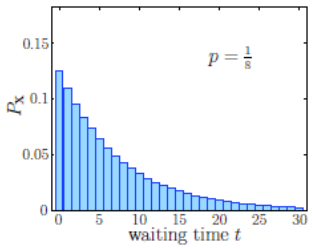
The x-axis denotes the number of coins we need to toss, and the y-axis denotes the probability.

The Geometric Distribution is related to the *Waiting Time to Success*. Let p be the probability to succeed on a random trial. Let X be the number of trials that appear until the first success.



$X = k$ means that the first $k-1$ trials all lead to fails while the last trial is a success. This leads to the following computation

$$\mathbb{P}(X = k) = \mathbb{P}(k \text{ trials}) = \underbrace{(1-p)^{k-1}}_{\text{the first } k-1 \text{ fails}} \underbrace{p}_{\text{last success}}$$

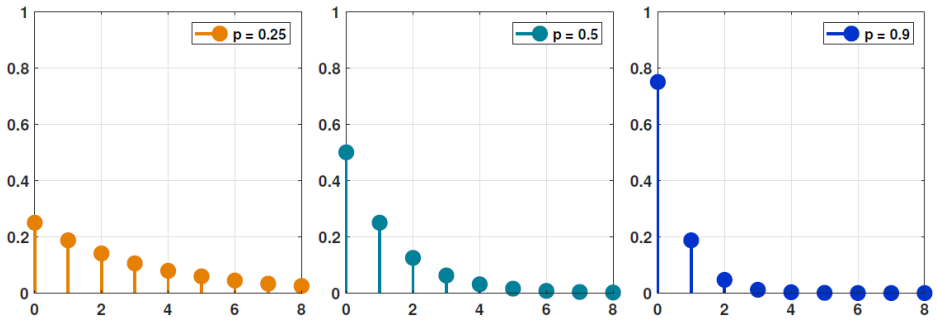


Based on the above deduction, we can introduce the following distribution.

Definition 6 (Geometric Distribution)**.** We say $X \sim \text{Geometric}(p)$ with the range $X(\Omega) = \{1, 2, 3, \ldots\}$ iff

$$P_X(k) = (1-p)^{k-1}p \qquad \text{for } k = 1, 2, 3, \ldots.$$

Below we plot the shape of a Geometric Random Variable with different parameters



PMFs of a geometric random variable $X \sim \text{Geometric}(p)$