2.1: Random variable: A random variable is a quantity that may take any of a given range of values that cannot be predicted exactly but can be described in terms of their probability. Random variables are named by capital letters, like $\boldsymbol{X}$. The same letter but lowercase, like $\boldsymbol{x}$, denotes a **data value** (a number).

Continuous and discrete random variables: A random variable is if it potentially can take on **any value** on some line segment or interval (that is, there are no "breaks" between possible values). A random variable is if the values it can potentially assume constitute a sequence of **isolated** or **separated** points on the real number axis. Continuous random variables usually measure the amount of something, whereas discrete random variables usually count something.

Examples of continuous and discrete random variables: a person's height, the length of time to run a marathon, the mass in kilograms of a celestial object (planet, star, meteor, piece of space dust, etc.). In this last example, we would consider the mass to be theoretically any value greater than 0, showing that sometimes the interval of possible values of a random variable is considered to be infinite in length, **whereas** the discrete random variables are the number of children in a family, the number of times a person catches a cold in a given year, and the number of tosses of a coin before a tail appears. This last example shows that sometimes the sequence of potential values of a discrete random variable can be infinite because the sequence in this example would be $\mathbf{1,2,3,\ldots}$.

We are interested in probabilities associated with various values of a random variable. A formula or table that enables us to find such probabilities is called a **probability distribution** for the random variable.

Discrete Probability Distributions:
In a study of families with one child, a researcher coded families as follows:

$$\boldsymbol{X}=\begin{cases}\mathbf{0};\ \textbf{if child is a boy}\\ \mathbf{1};\ \textbf{if child is a girl}\end{cases}$$

Imagine the experiment of randomly selecting a family with one child and recording whether the child is a boy ($\boldsymbol{B}$) or a girl ($\boldsymbol{G}$). The sample space is $\boldsymbol{S}=\{\boldsymbol{B},\boldsymbol{G}\}$ and $\boldsymbol{X}$ is a random variable on this sample space. We can view $\boldsymbol{X}$ as the number of girls in a randomly selected family with one child ($\mathbf{0}$ or $\mathbf{1}$). We assume that a boy and a girl are equally likely. Hence, the probability that $\boldsymbol{X}(\boldsymbol{B})=\mathbf{0}$ is $\frac{\mathbf{1}}{\mathbf{2}}$ and the probability that $\boldsymbol{X}(\boldsymbol{G})=\mathbf{1}$ is $\frac{\mathbf{1}}{\mathbf{2}}$. We sometimes write

$$\boldsymbol{P}(\mathbf{0})=\frac{\mathbf{1}}{\mathbf{2}}\ \&\ \boldsymbol{P}(\mathbf{1})=\frac{\mathbf{1}}{\mathbf{2}}$$

The specification of the probabilities associated with the distinct values of this random variable is called its **probability distribution**.

For a random variable $\boldsymbol{X}$ of the discrete type, the probability $\boldsymbol{P}(\boldsymbol{X}=\boldsymbol{x})$ is frequently denoted by $\boldsymbol{f}(\boldsymbol{x})$, and this function $\boldsymbol{f}(\boldsymbol{x})$ is called the **probability mass function**. Note that some authors refer to $\boldsymbol{f}(\boldsymbol{x})$ as the probability function, the frequency function, or the probability density function. In the discrete case, we shall use "probability mass function," and it is hereafter abbreviated **pmf**.

Properties of pmf: The pmf $f(x)$ of a discrete random variable X is a function that satisfies the following properties:

- **(a)** $P(X = x) = f(x) > 0, \; x \in S$
- **(b)** $\sum_{x \in S} f(x) = 1$
- **(c)** $P(X \in A) = \sum_{x \in A} f(x)$, where $A \subset S$
- **(d)** $f(x) = 0, \; x \notin S$

Cumulative distribution function (cdf): The cumulative distribution function (abbreviate it as **cdf**) is defined as: $F(x) = P(X \leq x), \quad -\infty < x < \infty$

Example: A college statistics class has 20 students. The ages of these students are as follows: ***One*** student is 16 years old, ***four*** are 18, ***nine*** are 19, ***three*** are 20, ***two*** are 21, and ***one*** is 30. Let x =the age of any student (randomly selected). Find the probability and Cumulative distribution for x.

Solution: Since each student has an equal likelihood of being selected, the probability of selecting a particular student is $\frac{1}{20}$. The probability of selecting a student that is 19 years old is $P(19) = \frac{9}{20}$, since there are ***nine*** students of that age. The probability and Cumulative distribution is summarized in the following table:

X	$P(X = x)$	$F(x) = P(X \leq x)$
16	$\frac{1}{20}$	$\frac{1}{20}$
18	$\frac{4}{20}$	$\frac{5}{20}$
19	$\frac{9}{20}$	$\frac{14}{20}$
20	$\frac{3}{20}$	$\frac{17}{20}$
21	$\frac{2}{20}$	$\frac{19}{20}$
30	$\frac{1}{20}$	$\frac{20}{20} = 1$

Example: Show that $P(X = x) = f(x) = \frac{x}{6}$; for $x = 1, 2, 3$ defines a probability distribution.

Sol: Here, $\sum f(x) = f(1) + f(2) + f(3) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} = 1$. So, $f(x)$ is a probability distribution.

Discrete Uniform Distribution: When a ***pmf*** is constant on the space or support, we say that the distribution is **uniform** over that space. Consider, X have a discrete uniform distribution over the first m positive integers, so that its ***pmf*** is $f(x) = \frac{1}{m}$; for $x = 1, 2, 3, \ldots, m$.

The ***cdf*** of X is defined as follows where $k = 1, 2, \ldots, m - 1$. We have

$$F(x) = P(X \leq x) = \begin{cases} \mathbf{0}; & x < 1 \\ \dfrac{k}{m}; & k \leq x < k+1 \\ \mathbf{1}; & x \geq m \end{cases}$$

Note that this is a step function with a jump of size $\frac{1}{m}$ for $x = 1, 2, \ldots, m$.

Example 2.1-3: Roll a fair four-sided die twice, and let X be the maximum of the two outcomes. Find the ***pmf*** of X.

Sol: The sample outcomes of rolling ***two*** four-sided die:

	1	2	3	4
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)

Here, $P(X = 1) = P[(1, 1)] = \frac{1}{16}$,

$P(X = 2) = P[\{(1, 2), (2, 1), (2, 2)\}] = \frac{3}{16}$,

Similarly, $P(X = 3) = \frac{5}{16}$ and $P(X = 4) = \frac{7}{16}$.

That is, the ***pmf*** of X can be written simply as

$$P(X = x) = f(x) = \frac{2x-1}{16};\ \text{for } x = 1, 2, 3, 4.$$

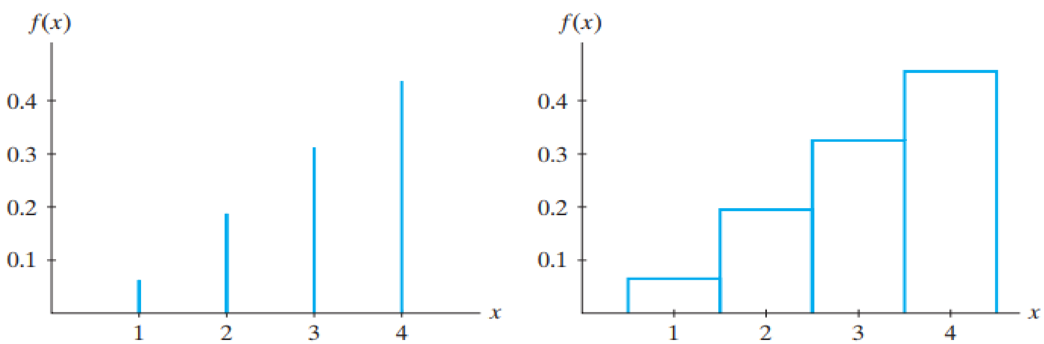

Figure: Line graph and probability histogram

Formula: The probability of selecting n objects from $N = N_1 + N_2$, which contains x objects from N_1 is:

$$P(X = x) = f(x) = \frac{\binom{N_1}{x} \times \binom{N_2}{n-x}}{\binom{N}{n}}$$

Examples: 2.1-1 to 2.1-7 (See yourself)

Exercises: 2.1-1 to 2.1-10 & 2.1-13 to 2.1-15 (Try yourself)