Chapter 2

Discrete Random Variables

2.1 Random Variables

A random variable is similar to a function; indeed, the most common definition of a random variable is a function which assigns a value in the space of real numbers $\Re$ to each outcome in Ω. Recall that functions can assign only one value to each outcome.

Definition 2.1
A random variable X in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ is a function $X : \Omega \to \Re$, such that, for any interval (a, b), the set $\{\omega \in \Omega : a < X(\omega) < b\}$ belongs to the event space $\mathcal{E}$.

By constraining random variables to functions where the inverse image of an interval (a, b) is an event in $\mathcal{E}$, we can compute $\mathbb{P}[\{\omega \in \Omega : a < X(\omega) < b\}]$. As we discussed earlier in 1.8, the smallest σ-field in $\Re$ that contains the open intervals (a, b) is known as the Borel σ-field $\mathcal{B}$. Using limits and the continuity of probability measures, we can then compute for any Borel set $A \in \mathcal{B}$, the probability $\mathbb{P}[\{\omega \in \Omega : X(\omega) \in A\}]$. That is, the inverse image using the function $X(\omega)$ of any Borel set A will be an event in $\mathcal{E}$. In a more formal mathematical definition, we such functions **measurable functions** from $(\Omega, \mathcal{E})$ into $(\Re, \mathcal{B})$. Figure 2.1 illustrates the concept of a random variable.

Random variables provide a useful abstraction in probability. First, by assigning numbers to outcomes, they allow us to map outcomes onto a quantitative scale, which will allow us to compute interesting statistics. More important, they allow us to recognize that many different experiments give rise to the same type of random variables, and thus can be analyzed by a common methodology without worrying about the individual details of the experiments. For instance, we discussed in the previous chapter the concept of Bernoulli trials as an experiment with two outcomes. That experiment can be a coin flip, a race between two people, a bet, a roll of a pair of dice to get a total of 7, a shot at a target, etc. By mapping one outcome to the number 1, and the other outcome to 0 we get a Bernoulli random variable. Thus, the analysis of Bernoulli random variables provides the tools for analysis in all the diverse experiments that give rise to such random variables. Similar abstractions will allow us to use a common set of random variables to analyze measurement errors that arise in acoustic, aerospace, electronic and biomedical measurements.

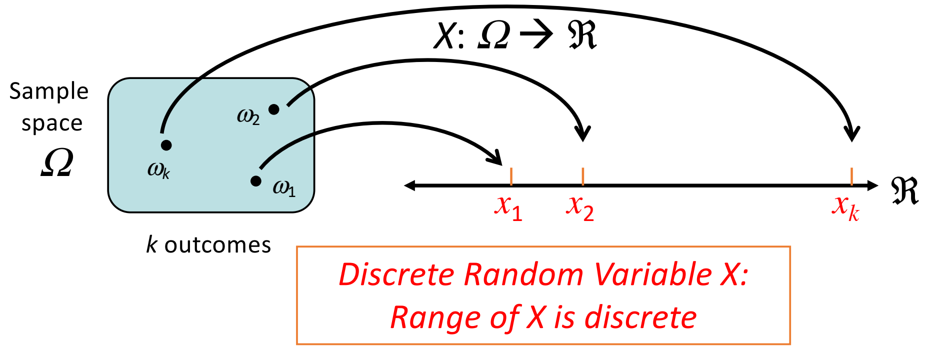

Figure 2.1: Discrete random variables map Ω into a discrete set of values in the real line.

We introduce some notation that we will use throughout the book: We use capital letters (e.g. X, Y, Z) to denote random variables, and we use lower case letters *(e.g. x, y, z) to denote the values that a random variable takes.

We denote by R_X the image of the sample space Ω (the **range**) as mapped by the random variable $X(\cdot)$. That is, $R_X \subset \Re$ is the set of possible values of $X(\omega), \omega \in \Omega$.

Definition 2.2
A **discrete random variable** in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ is a random variable X such that the range of X, denoted by $R_X = X(\Omega)$, has at most a countable number of elements.

We sometimes make a distinction to refer to a random variable as **finite** if R_X has a finite number of elements. Random variables that are not discrete can be either continuous or hybrid as described in the next chapter.

Example 2.1
Turn on a light source, and use a CCD detector to count the number of photons that hit the detector in an interval of one second. In this experiment, $\Omega = \{0, 1, 2, \ldots\}$. We define the random variable $X(\omega) = \omega$, as the outcomes are already numeric. The range R_X of this random variable is $R_X = \{0, 1, 2, \ldots\}$. X is a discrete random variable, as its range is discrete.

Example 2.2
Turn on a light source, and have a CCD detector that measures the time between the arrival of the first photon and the arrival of the second photon. In this experiment, $\Omega = [0, \infty)$. We define the random variable $X(\omega) = \omega$, as the outcomes are already numeric. The range R_X of this random variable is $R_X = [0, \infty)$, which is not countable. This random variable X is not a discrete random variable.

Suppose we define a different random variable $Y(\omega)$ as follows:

$$Y(\omega) = \begin{cases} 0 & \omega < 2ns, \\ 1 & \text{elsewhere.} \end{cases}$$

In this case, the range $R_Y = \{0, 1\}$, which is finite, so Y is a discrete random variable.

Most card experiments, dice experiments and coin flip experiments give rise to discrete random variables. We list some examples of discrete random variables below.

- The number of X-Ray photons detected in a pixel by an X-ray radiograph.
- The number of defective parts in a manufacturing process in 10 minutes.
- . The presence of a disease in a patient.
- The correctness of a software implementation of an algorithm.
- The number of parts that fail in an automobile in the course of a year.

Typical examples of random variables that are not discrete are the time until a part fails in an assembly plant, the error in location given by a GPS system, the error in measuring the distance to an obstacle using a LIDAR sensor and the time of arrival of customer at a service station.

A random variable X induces a probability measure $\mathbb{P}_X$ on $(\Re, \mathcal{B})$ using the function mapping. For any intervals $(a, b) \in \Re$, this probability is given by

$$\mathbb{P}_X((b, a)) = \mathbb{P}[\{\omega \in \Omega : b < X(\omega) < a\}]$$

and, more generally, for any set $B \in \mathcal{B}$, we have

$$\mathbb{P}_X(B) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}].$$

Indeed, with this induced probability, we can show that $(\Re, \mathcal{B}, \mathbb{P}_X)$ is also a probability space. We call this space the sample space. The abstractions that random variables provide will allow us to use the same induced probability space for many different random experiments.

Example 2.3
Consider the experiment of tossing two unbiased coins. In the original space Ω, there are four outcomes: HH, HT, TH and TT, where H denotes a heads outcome and T denotes a tails outcome. We define a random variable X as follows:

$$X(\omega) = \begin{cases} -1 & \text{if } s \neq \text{HH, TT.} \\ 1 & \text{otherwise.} \end{cases}$$

In this experiment, $R_x = \{-1, 1\}$. The induced probability $\mathbb{P}_X$ can be defined on its atoms, so that $\mathbb{P}_X[\{1\}] = \mathbb{P}[\{\omega \in \Omega : X(\omega) = 1\}] = \mathbb{P}[\{HH, TT\}] = 0.5$. Similarly, $\mathbb{P}_X[\{-1\}] = 0.5$.

Now, consider a second experiment, consisting of tossing a single unbiased coin, with sample space $\Omega_1 = \{H, T\}$, and define variable Y as

$$Y(\omega) = \begin{cases} -1 & \text{if } \omega = H \\ 1 & \{textotherrwise. \end{cases}$$

The sample space and induced probability of this random experiment and random variable Y are the same as those of the first experiment and random variable X. Rather than treating these random variables a different, by using the sample space, we can treat them as identical random variables.

2.2 Discrete Random Variables

Consider a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, with a discrete random variable X defined on it, with values in $\{x_1, x_2, x_3, \ldots\}$. Since every set $\{x_i\}$ containing a singleton value is a Borel set, we can compute the probability $\mathbb{P}[\{\omega \in \Omega : X(\omega) = x_i\}]$. We can use this to define the induced probability measure $\mathbb{P}_X$ on R_X. We define this formally next.

2.2.1 Probability Mass Function

Definition 2.3
The **probability mass function** of a discrete random variable X defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, taking values in $\{x_1, x_2, x_3, \ldots\}$ is the function $P_X(x_i) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x_i\}]$.

To keep the notation simple, we refer to the set $\{X = x_1\} \equiv \{\omega \in \Omega : X(\omega) = x_i\}$. Thus, we will write equivalently the following forms for the probability mass function of a discrete random variable:

$$P_X(x) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] \equiv \mathbb{P}[\{X = x\}] \equiv \mathbb{P}[X = x].$$

In each case, it should be clear that this is computing the probability of an event $A \in \mathcal{E}$ defined all possible solutions of the equation $X(\omega) = x$. Figure 2.2 illustrates a probability mass function for a discrete random variable.

The probability mass function (PMF) of a random variable X satisfies the following basic properties:

1. **Non-negativity:** $P_X(x) \geq 0$ for all x.
2. **Normalization:** $\sum_{x \in R_X} P_X(x) = 1$.

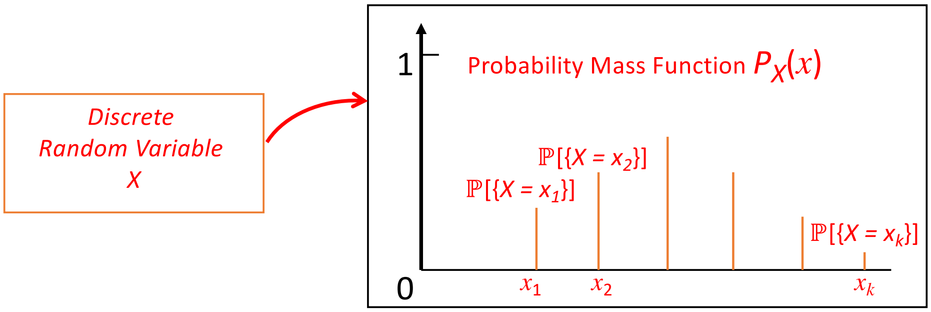
Discrete
Random Variable
X
1
0
Probability Mass Function $P_X(x)$
$\mathbb{P}[\{X = x_1\}]$
$\mathbb{P}[\{X = x_2\}]$
$\mathbb{P}[\{X = x_k\}]$
x_1
x_2
x_k

Figure 2.2: Illustration of a Probability Mass Function.

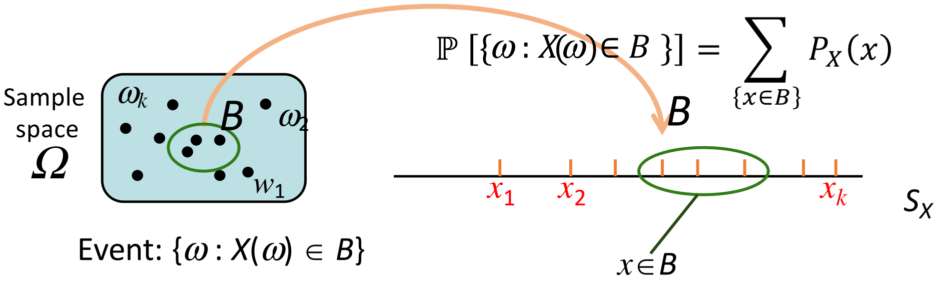
$\mathbb{P}\,[\{\omega : X(\omega) \in B\}] = \sum_{\{x\in B\}} P_X(x)$
Sample
space
Ω
ω_k
B
ω_2
w_1
Event: $\{\omega : X(\omega) \in B\}$
B
x_1
x_2
x_k
S_X
$x \in B$

Figure 2.3: Computing the probability of events using the PMF.

3. **Additivity:** For any subset $B \subset R_X$, the probability that X falls in B is

$$\mathbb{P}_X[B] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}] = \sum_{x\in B} P_X(x).$$

Note that $\mathbb{P}_X[B]$ implicitly refers to the event $\mathbb{P}_X[B] = \mathbb{P}\big[\{X \in B\}\big]$.

The Additivity property follows because the event $\{\omega \in \Omega : X(\omega) \in B\}$ can be decomposed into disjoint events $\{\omega \in \Omega : X(\omega) = x_i\}$ for each $x_i \in B$. These events are disjoint because $X(\omega)$ is a function and thus can only assign a single value to each $\omega \in \Omega$. Then, the countable additivity property of the probability measure shows

$$\mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}] = \sum_{x\in B} \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] = \sum_{x\in B} \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] = \sum_{x\in B} P_X(x).$$

Figure 2.3 illustrates the approach at computing probabilities of events using the additivity property of the probability mass function. Any event in R_X will contain discrete elements x_i on which the probability mass function is defined. The induced probability of the event is the sum of the probability mass function on the elements that are in B.

Example 2.4

In this experiment, we roll two four-sided dice, with all outcomes on each dice being equally likely. Note that these dice are tetrahedral, so the number that a die rolls is the number at the bottom. We define the random variable X to be the sum of the numbers at the bottom of the dice.

The sample space is $\Omega = \{(i,j) : i,j = 1,2,3,4\}$. The image $R_X = \{2,3,4,5,6,7,8\}$. Since this is a discrete set, we compute the PMF as:

$$\begin{aligned}
P_X(2) &= \mathbb{P}[\{(1,1)\}] = \frac{1}{16}\\
P_X(3) &= \mathbb{P}[\{(1,2),(2,1)\}] = \frac{1}{8}\\
P_X(4) &= \mathbb{P}[\{(1,3),(2,2),(3,1)\}] = \frac{3}{16}\\
P_X(5) &= \mathbb{P}[\{(1,4),(2,3),(3,2),(4,1)\}] = \frac{1}{4}\\
P_X(6) &= \mathbb{P}[\{(2,4),(3,3),(4,2)\}] = \frac{3}{16}\\
P_X(7) &= \mathbb{P}[\{(3,4),(4,3)\}] = \frac{1}{8}\\
P_X(8) &= \mathbb{P}[\{(4,4)\}] = \frac{1}{16}
\end{aligned}$$

Using this, define the event $B = \{X \text{ is even }\}$. Then,

$$\mathbb{P}_X[B] = P_X(2) + P_X(4) + P_X(6) + P_X(8) = \frac{8}{16} = \frac{1}{2}.$$

Define the event $C = \{X \text{ is a multiple of 3 }\}$. Then,

$$\mathbb{P}_X[C] = P_X(3) + P_X(6) = \frac{5}{16}.$$

Example 2.5
In an experiment, we have a biased coin with two outcomes, H and T, with probability of H $= p > 0$. We are going to toss that coin an infinite number of times, so that an outcome of the experiment is an infinite sequence of Hs and Ts; e.g. HHHTHTTHTHHHTTTTTH. . . . The outcomes of each coin toss are independent, so this defines the outcomes in the original probability space as well as the underlying probabilities, e.g. we have $(\Omega, \mathcal{E}, \mathbb{P})$. On this probability space, we define a random variable $X(\omega)$ for an outcome $\omega \in \Omega$ as the position of the first H in ω. That is, $X(\text{TTHTHT}\ldots) = 3, X(\text{THTTH}\ldots) = 2$, etc. Note the possible values of X are discrete and countable. Find the probability mass function $P_X(x)$, and compute the induced probability of the event $B = \{X(\omega) \in [2,3]\}$.

We note that all outcomes for $X(k) = 3$ have to start with TTH, and the rest of the outcomes after the first toss result in the same $X(k)$. Using this reasoning, we can derive

$$P_X(x) = p(1-p)^{x-1}, x = 1, 2, \ldots \ .$$

The induced probability $\mathbb{P}_X[B] = P(2) + P(3) = p(1-p) + p(1-p)^2 = p(1-p)(2-p)$.

Random variables of this type are called geometric random variables, because of the geometric decay of the PMF as x increases.

Example 2.6
This example shows we don't need to know anything about the underlying experiment if we know the probability mass function to compute probabilities for events defined in terms of the random variable. Assume $R_X = \{1,2,3,4\}$, and let the probability mass function be $P(x) = \frac{c}{x}$ for some $c > 0$. Find the value of c, and find the probabilities of the events $A = \{X \geq 2\}$ and $B = \{X < 3\}$.

We use the normalization property to compute c, since

$$P(1) + P(2) + P(3) + P(4) = c + \frac{c}{2} + \frac{c}{3} + \frac{c}{4} = \frac{25}{12}c = 1.$$

Hence, $c = \frac{12}{25}$. Next, we compute $\mathbb{P}_X[A] = P(2) + P(3) + P(4) = \frac{13}{25}$, and $\mathbb{P}_X[B] = P(1) + P(2) = \frac{18}{25}$.

2.2.2 Cumulative Distribution Function

The **cumulative distribution function (CDF)** of a random variable X in a probability space returns the probability that a random variable X is less than or equal to a value x:

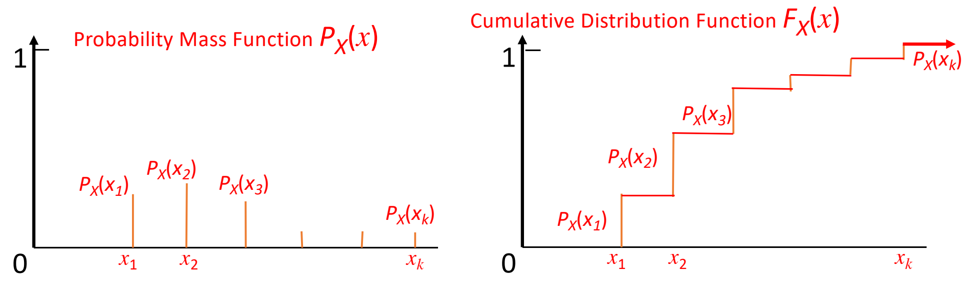

Figure 2.4: Relationship between the PMF and CDF of a random variable.

Definition 2.4 (Cumulative Distribution Function)
The **Cumulative Distribution Function** of the random variable X is defined as the function $F_X : \Re \to [0, 1]$ which satisfies:

$$F_X(a) \equiv \mathbb{P}_X(\{X \in (-\infty, a\}]) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\}].$$

We will sometimes use the notation $F(a)$ instead of $F_X(a)$ when it is clear which random variable we are referring to. In particular, for a generic argument, this is often written as $F_X(x)$ or just $F(x)$.

Figure 2.4 shows the relationship of the PMF and the CDF. In essence, the CDF is the sum of the PMF starting from the left at the smallest value of $x \in R_X$.

The CDF is a non-negative real-valued function $F_X(x) \in [0, 1]$, defined for all real values of its argument $x \in \Re$. The CDF of any discrete random variable is a staircase function. If X takes on values $x_1, x_2, \ldots x_k$ with probabilities $P(x_1), P(x_2), \ldots, P(x_k)$, then the CDF has jumps at $x_1, x_2, \ldots x_k$ with heights $P(x_1), P(x_2), \ldots, P(x_k)$ and is flat in between the jumps.

Cumulative distribution functions have the following properties:

1. $F_X(\infty) = \lim_{x\to\infty} F_X(x) = 1, F_X(-\infty) = \lim_{x\to-\infty} = 0$.
2. $a \leq b$ implies that $F_X(a) \leq F_X(b)$, so $F(x)$ is non-decreasing in x.
3. $F_X(x)$ is piecewise constant and jumps at values of $x \in R_X \subset \Re$ such that $P(x) > 0$.
4. For all $b > a, \mathbb{P}_X[\{a < X \leq b\}] = F_X(b) - F_X(a)$.
5. $\lim_{\epsilon\to 0^+} F_X(a + \epsilon) = F_X(a) \quad$ (continuity from the right)

Proof: The first properties follow from the continuity of probabilities. Define the events as $A_n = \{\omega \in \Omega : X(\omega) \leq n\}$. These form a non-decreasing sequence, so by Lemma 1.1

$$\lim_{n\to\infty} \mathbb{P}[A_n] = \lim_{n\to\infty} F(n) = \mathbb{P}[\cup_{k=1}^{\infty} A_n] = \mathbb{P}[\Omega] = 1$$

Similarly, the sequence $B_n = \{\omega \in \Omega : X(x) \leq -n\}$ forms a non-increasing sequence with an empty intersection, so

$$\lim_{n\to\infty} \mathbb{P}[B_n] = \lim_{n\to\infty} F(-n) = 0$$

The second property follows from the fact that $\{\omega \in \Omega : X(\omega) \leq a\} \subset \{\omega \in \Omega : X(\omega) \leq b\}$. The final property can be shown as follows: Define the sets $A_n = \{\omega \in \Omega | X(\omega) \leq a + 1/n\}$. Again, these sets are non-increasing, so

$$\lim_{n\to\infty} \mathbb{P}[A_n] = \lim_{n\to\infty} F(a + 1/n) = \mathbb{P}[\cap_{n=1}^{\infty} A_n] = F(a)$$

Example 2.7
Consider the example 2.4 with two quadrilateral dice. The CDF of X is given by:

$$F_X(x) = \begin{cases} 0 & x < 2; \\ \frac{1}{16} & 2 \leq x < 3; \\ \frac{3}{16} & 3 \leq x < 4; \\ \frac{3}{8} & 4 \leq x < 5; \\ \frac{5}{8} & 5 \leq x < 6; \\ \frac{13}{16} & 6 \leq x < 7; \\ \frac{15}{16} & 7 \leq x < 8; \\ 1 & 8 \leq x. \end{cases}$$

Thus, the CDF is piecewise constant, and jumps at each integer value in R_X by the amount $P(x)$.

In general, cumulative distributions of a random variable are not very useful for computing statistics. Almost every computation uses the probability mass function instead. So why do we bother with defining CDFs and their properties? Recall that, in chapter 1, we constructed many cases where the probability of every atom is zero. For those cases, such as those involving continuous random variables, it is impossible to define a PMF. However, the concept of CDF applies to all random variables, continuous or discrete, and has nearly the same properties in all cases.

2.3 Statistics of Discrete Random Variables

We are used to seeing sample statistics in many different fields. In data science, samples are collected by repeating the same experiment independently many times, and generating the random variables associated with each of these experiments. Social statisticians work hard to select samples that correspond to the true population at large. Given a set of sample values for a random variable generated this way, a sample statistic maps these values into a single real number.

For instance, suppose the experiment is a student in EK 381 taking a midterm exam. The random variable maps the student answers into a number grade. If 80 students take the same exam, this can be viewed as repeating the experiment of selecting a student randomly 80 times and getting a value for the random variable. We assume the grading is done in whole numbers from 0 to 100, so the possible values for the random variables are discrete.

The first class after every exam, the professor is asked the same question: "What was the class average?" The class average is an example of a sample statistic. If $x_i, i = 1, \ldots, N$ are the values of the random variable X in N repetitions of the same experiment, the sample average or sample mean is defined as:

$$m_X = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Similarly, the sample variance is defined by $\mathsf{Var}[X] = \frac{1}{N} \sum_{i=1}^{N} (x_i - m_X)^2$, and the sample standard deviation is computed as $\sigma_X = \sqrt{\mathsf{Var}[X]}$. However, note that those statistics will change as N changes. In essence, they are random also, in a manner that will be made more precise later in the course. What we hope is that, as N grows, the statistics approach a limit and become constant, and thus represent an intrinsic property of a random variable.

There is another way to write the sample statistics, in terms of a sample probability mass function $\tilde{P}(x)$. In essence, compute a sample probability mass function as:

$$\tilde{P}_X(x) = \frac{1}{N} \sum_{i=1}^{N} I[x_i = x]$$

where $I[x_i = x]$ is the indicator function that is 1 if it is true, and 0 otherwise. This computes the relative frequency that the value x appears in the sample of size N. Then, using the sample probability mass function, the sample statistics can be written as:

$$m_X = \sum_{x \in R_X} \tilde{P}_X(x)x$$

$$\mathsf{Var}[X] = \sum_{x \in R_X} \tilde{P}_X(x)(x - m_X)^2$$

and $\sigma_X = \sqrt{\mathsf{Var}[X]}$.

This new form suggests that each random variable has a true statistic that can be defined in terms of its probability mass function. The sample statistics are random approximations of these true statistics.

Definition 2.5
A statistic of a discrete random variable is a map from its probability mass function to a real-valued quantity.

Below we define some of the most common statistics associated with discrete random variables.

2.3.1 Expected Value

The **expected value** of a discrete random variable X is defined as

$$\mathbb{E}[X] = \sum_{x \in R_X} x\, P_X(x).$$

This is also known as the **mean** or **average**. In these notes, we also sometimes use $\mu_X = \mathbb{E}[X]$.

The expected value has many interpretations: It is the weighted average of all possible values, using the PMF weights. It can be viewed as the center of “mass” of the PMF. Ideally, it would also be the sample average after one performs a large number repetitions of the experiment (to be substantiated later in this course): the sample mean should approach the true mean as number of samples increases!

Example 2.8
Consider the two quadrilateral dice example 2.4. Then,

$$\mathbb{E}[X] = sum_{x \in R_X} x\, P_X(x) = 2 \cdot \frac{1}{16} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{3}{16} + 5 \cdot \frac{1}{4} \\ + 6 \cdot \frac{3}{16} + 7 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} = \frac{80}{16} = 5.$$

Note that, for some random variables where the range R_X is infinite, the expected value cannot be defined because the sum may not be finite! This is illustrated in the examples below:

Example 2.9
Assume we have a discrete random variable X with range $R_X = \{1, 2, \ldots\}$ and PMF given by $P_X(k) = \frac{6}{k^2 \pi^2}, k = 1, 2, \ldots.$

It is easy to verify that this is a valid PMF, as it is non-negative, and normalized properly because $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. This formula was derived by Leonard Euler in the early part of the 18th century. For this random variable, note that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \frac{6}{\pi^2} \frac{k}{k^2} = \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k} = \infty\ .$$

Thus, for statistics defined using expected values, it is possible that the statistics won’t be defined if the required sums do not converge.

Example 2.10
We consider a signaling example where we want to transmit a single bit using a DC voltage. If the bit is 1, we transmit a voltage +1 volts. For the bit being 0, we transmit the voltage -1 volts. Assume that the bit is equally likely to be a 0 or a 1.

We construct the sample space for the experiment as $\Omega = \{0,1\}$, the value of the bit. We define the random variable as the voltage $X(0) = -1, X(1) = 1$, so $R_X = \{-1,1\}$. The probability measure in the original space is $\mathbb{P}[\{0\}] = \mathbb{P}[\{1\}] = 0.5$. The resulting PMF is given as

$$P_X(-1) = 0.5; \quad P_X(1) = 0.5.$$

Using this, we compute $\mathbb{E}[X] = 0.5 \cdot (-1) + 0.5 \cdot (1) = 0$.

Assume that we wanted to transmit two bits at a time. In this case, the sample space $\Omega = \{00, 01, 10, 11\}$, with each outcome having probability 0.25. Now, we define a new random variable Y corresponding to the voltage used for signaling, so that $Y(00) = -3, Y(01) = -1, Y(10) = 1, Y(11) = 3$.

The range space $R_Y = \{-3,-1,1,3\}$. The induced PMF is $P_Y(-3) = P_Y(-1) = P_Y(1) = P_Y(3) = 0.25$. Then,

$$\mathbb{E}[Y] = 0.25 \cdot (-3) + 0.25 \cdot (-1) + 0.25 \cdot (1) + 0.25 \cdot (3) = 0.$$

Thus, the two signaling schemes X, Y have the same expected value 0. However, they will differ in other statistics, such as average energy, where you can expect that the energy is proportional to X^2 or Y^2. To do this, we need to be able to compute averages of functions of random variables such as X^2.

2.4 Functions of a Random Variable

Consider a random variable X defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. X is a function mapping outcomes in Ω into real numbers in $\Re$. Suppose we now define another function $g(\cdot)$ mapping a real number into another real number (e.g. $g : \Re \to \Re$.) Then, the composition of the two functions, $g(X(\omega))$ also maps outcomes in Ω into real numbers in $\Re$, so that each outcome is only mapped into a single real number. That is, the composition of the two functions is also a function. As long as the function $g(\cdot)$ is well behaved (measurable in the context discussed earlier), this composite function also defines a random variable in $(\Omega, \mathcal{E}, \mathbb{P})$! We denote this random variable as $Y = g(X)$ to indicate that the variable Y is derived by a function transformation of the random variable X, and the underlying random variable map $Y(\omega) \equiv g(X(\omega))$.

Note that this raises an interesting observation: we can define multiple random variables on the same probability space. We will explore this fully in later chapters. For the moment, let's focus on the case where $Y(\omega) = g(X(\omega))$. This case is often referred to as a **derived random variable**.

What is the range of Y as a random variable? It is derived from the range of X: $R_Y = \{g(x) : x \in R_X\}$. If X is a discrete random variable, then R_X is a countable, discrete range, and therefore R_Y will also be at most countable and discrete. Note that R_X is countably infinite does not imply R_Y will be, as the function $g(\cdot)$ may map many numbers in R_X into a single number in R_Y. For example, consider the function $g(\cdot)$ defined below that maps $\{1,2,3,\ldots,\}$ into $\{0,1\}$:

$$g(x) = \begin{cases} 1 & \text{if } x \text{ is an odd positive integer} \\ 0 & \text{elsewhere} \end{cases}$$

If we know the function $g(\cdot)$ and the probability mass function of X, $P_X(x)$, we can compute the probability mass function for Y directly as $P_Y(y) = \displaystyle\sum_{x:g(x)=y} P_X(x)$, where the notation $\sum_{x:g(x)=y}$ means sum over each value of $x \in R_X$ such that $g(x) = y$ is satisfied. This is exactly the same approach we took to computing the probability mass function $P_X(x)$: The event $\{Y = y\}$ has an inverse image through the function g which is composed of a subset of R_X, which is $\{x \in R_X : g(x) = y\}$. Since R_X is discrete, this is a discrete set, and

$$P_Y(y) = \mathbb{P}_X(\{x \in R_X : g(x) = y\}) = \sum_{x:g(x)=y} P_X(x).$$

As long as we have the properties of X, as summarized by its probability mass function P_X, we can compute all the properties of Y without having to refer to the original probability space $(\Omega, \mathcal{E}, \mathbb{P})$. We illustrate this with examples.

Example 2.11
Consider a discrete random variable X with values in $R_X = \{1, 2, 3, 4\}$ and probability mass function

$$P_X(x) = \begin{cases} 1/3 & x \leq 2 \\ 1/6 & x > 2 \end{cases}$$

Let $g(x) = x^3$ be a function, and define $Y = g(X)$ as a derived random variable. In this case, the range of Y is $R_Y = g(R_X) = \{1, 8, 27, 64\}$. The probability mass function of Y is now

$$P_Y(y) = \sum_{x:g(x)=y} P_X(x) = \begin{cases} 1/3 & y \leq 8 \\ 1/6 & y > 8 \end{cases}$$

Now, let's repeat the exercise for a different function: let $h(x) = 0$ for $x \leq 3$, and $h(x) = 1, x > 3$. Define $Z = h(X)$ be the resulting derived random variable. Then, $R_Z = h(R_X) = \{0, 1\}$, and the resulting probability mass function is

$$P_Z(z) = \sum_{x:h(x)=z} P_X(x) = \begin{cases} P_X(1) + P_X(2) + P_X(3) = 5/6 & z = 0 \\ P_X(4) = 1/6 & z = 1 \end{cases}$$

Example 2.12
Consider now the signaling example 2.10. Let $U = X^2$. Then, $U(-1) = 1, U(1) = 1$, so $R_U = \{1\}$. Hence, $P_U(1) = P_X(-1) + P_X(1) = 1$. Hence, $\mathbb{E}[U] = 1$.

Define $V = Y^2$. Then, $V(-3) = V(3) = 9, V(-1) = V(1) = 1$. Thus, $R_V = \{1, 9\}$, and $P_V(1) = P_Y(-1) + P_Y(1) = 0.5; P_V(9) = P_Y(-3) + P_Y(3) = 0.5$. The average is:

$$\mathbb{E}[V] = 1P_V(1) + 9P_V(9) = 5.$$

So, on average, signaling with two bits at a time in this scheme takes much more energy than signaling each bit separately.

For a derived random variable Y, we can compute all of its statistics using its probability mass function $P_Y(y)$. However, there is a simpler approach that avoids the need for computation of $P_Y(y)$. Consider computation of the expected value of Y (its mean). Using the approach in subsection 2.3, we compute $\mathbb{E}[Y]$ as

$$\mathbb{E}[Y] = \sum_{y \in R_Y} y P_Y(y)$$

However, note that, using the definition of $P_Y(y)$

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{y \in R_Y} y P_Y(y) = \sum_{y \in R_Y} y \sum_{x:g(x)=y} P_X(x) \\ &= \sum_{y \in R_Y} \sum_{x:g(x)=y} g(x) P_X(x) \quad \text{(since } y = g(x) \text{)} \\ &= \sum_{x \in R_X} g(x) P_X(x) \quad \text{(since } g \text{ is a function, and every } x \in R_X \text{ is mapped into some } y \in R_Y) \end{aligned}$$

Thus, we can compute $\mathbb{E}[Y]$ directly using the definition of the function $g(\cdot)$ and the probability mass function P_X without having to compute P_Y.

Example 2.13
Back to the signaling example 2.10, we compute directly:

$$\mathbb{E}[X^2] = (-1)^2 P_X(-1) + (1)^2 P_X(1) = 1.$$

$$\mathbb{E}[Y^2] = (-3)^2 P_X(-3) + (-1)^2 P_X(-1) + (1)^2 P_X(1) + (3)^2 P_X(3) = 5.$$

Let's focus now on a special class of functions: affine functions $g(x) = ax + b$. Let $Y = g(X) = aX + b$. Then,

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{x \in R_X} g(x) P_X(x) = \sum_{x \in R_X} (ax + b) P_X(x) \\ &= a \sum_{x \in R_X} x P_X(x) + b \sum_{x \in R_X} P_X(x) \\ &= a\mathbb{E}[X] + b \quad \text{(using the definition of } \mathbb{E}[X] \text{ and the normalization property of } P_X.)\end{aligned}$$

Thus, when a random variable Y is defined by an affine transformation of a random variable X, its expected value is computed by the same affine transformation of the expected value of X, avoiding having to do any summations over P_X.

An important statistic that we use to characterize the randomness in random variables is the variance. The **variance** measures how spread out a random variable is around its mean, and is defined by

$$\mathsf{Var}[X] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \sum_{x \in R_X} (x - \mu_X)^2 P_X(x).$$

Note that $\mathbb{E}[X]$ is a number, not a random variable. Hence, $Z = (X - \mathbb{E}[X])^2$ is transformation of the variable X. The variance of X is often referred to as $\sigma_X^2 = \mathsf{Var}[X]$, where σ_X is the positive square root of the variance of X, and is known as the **standard deviation.**

Example 2.14
Let X be a random variable, with $R_X = \{1, 3, 5\}$ and PMF $P_X(1) = P_X(3) = P_X(5) = \frac{1}{3}$. Then,

$$\mathbb{E}[X] = (1)P_X(1) + (3)P_X(3) + (5)P_X(5) = 3.$$

$$\mathsf{Var}[X] = (1-3)^2 P_X(1) + (3-3)^2 P_X(3) + (5-3)^2 P_X(5) = \frac{8}{3}.$$

The standard deviation is $\sigma_X = \sqrt{\frac{8}{3}}$.

There is an alternative formula for computing the variance of a random variable which is $\mathsf{Var}[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2$. We can show this as follows:

$$\begin{aligned}\mathsf{Var}[X] &= \sum_{x \in R_X} (x - \mu_X)^2 P_X(x) \\ &= \sum_{x \in R_X} (x^2 - 2x\mu_X + \mu_X^2) P_X(x) \\ &= \sum_{x \in R_X} x^2 P_X(x) - 2 \sum_{x \in R_X} x\mu_X P_X(x) + \sum_{x \in R_X} \mu_X^2) P_X(x) \\ &= \sum_{x \in R_X} x^2 P_X(x) - 2\mu_X \sum_{x \in R_X} x P_X(x) + \mu_X^2 sum_{x \in R_X} P_X(x) \\ &= \sum_{x \in R_X} x^2 P_X(x) - 2\mu_X^2 + \mu_X^2 = \sum_{x \in R_X} x^2 P_X(x) - \mu_X^2\end{aligned}$$

where the last line follows from the definition $\mu_X = \sum_{x \in R_X} x P_X(x)$ and the normalization property of the probability mass function $sum_{x \in R_X} P_X(x) = 1$.

Assume again we have an affine transformation $Y = aX + b$. We know that $\mathbb{E}[Y] = a\mathbb{E}[X] + b$. Can we compute a relationship for the variance of Y in terms of the variance of X? Reasoning as above, we obtain

$$\begin{aligned} \mathsf{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] &= \sum_{y \in R_Y} (y - \mathbb{E}[Y])^2 P_Y(y) \\ &= \sum_{x \in R_X} (ax + b - \mathbb{E}[Y])^2 P_X(x) \\ &= \sum_{x \in R_X} (ax + b - a\mathbb{E}[X] - b)^2 P_X(x) \\ &= \sum_{x \in R_X} a^2 (x - \mathbb{E}[X])^2 P_X(x) \\ &= a^2 \mathsf{Var}[X] \end{aligned}$$

Note that the constant b in the transformation $Y = aX + b$ affects the mean $\mathbb{E}[Y]$, but does not affect the variance, because the variance is a measure of the variation of Y about its mean. Notice also that the scaling factor a is squared in the variance, as the variance is a quadratic statistic. In terms of standard deviations, we have $\sigma_Y = |a|\sigma_X$.

To illustrate that the constant b does not affect the variance, consider the special transformation $Y = X - \mathbb{E}[X]$, where $a = 1$ and $b = -\mathbb{E}[X]$. In this special case,

$$\mathbb{E}[Y] = \mathbb{E}[X] - \mathbb{E}[X] = 0; \mathsf{Var}[Y] = \mathbb{E}[Y^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathsf{Var}[X]$$

which highlights that the variance of a random variable does not change when it is shifted by a constant.

These results provide a shortcut for computing statistics of derived random variables when the transformation $Y = aX + b$ is an affine transformation:

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b; \quad \mathsf{Var}[Y] = a^2 \mathsf{Var}[X]$$

The above results also highlight an important property of expectations. Suppose the function $g(x) = g_1(x) + g_2(x)$, and we define $Y = g(X) = g_1(X) + g_2(X)$. In the above linear case, $g_1(x) = ax, g_2(x) = b$. Then,

$$\mathbb{E}[Y] = \sum_{x \in R_X} (g_1(x) + g_2(x))\, P_X(x) = \mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)]$$

because the sum is a linear operation, and can be separated into two sums. Also, if $Y = ag_1(X) + bg_2(X)$, then

$$\mathbb{E}[Y] = \mathbb{E}[ag_1(X) + bg_2(x)] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(x)].$$

Thus, the expectation operator is a *linear operator*. We will exploit this property throughout the rest of this course.

There are other useful statistics that can be computed for a random variable X. We list a few below:

- n^{th} **Moment:** $\mathbb{E}[X^n] = \sum_{x \in R_X} x^n P_X(x)$.
- n^{th} **Central Moment:** $\mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^n\Big] = \sum_{x \in R_X} (x - \mu_X)^n P_X(x)$.
- **Median:** The median is a number $x_{med} \in \Re$ such that $\mathbb{P}_X[\{X < x_{med}\}] = \mathbb{P}_X[\{X > x_{med}\}]$. Note that such a number may not exist and, if it existed it may not be unique. For instance, consider a

random variable with two possible values, 0 or 1, and $P_X(0) = 0.1, P_X(1) = 0.9$. There is no median for this random variable. Similarly, consider another random variable with four possible values 0, 1, 2, 3, with $P_X(0) = P_X(1) = P_X(2) = P_X(3) = 0.25$. In this case, any number strictly between 1 and 2 serves as a median.

- **Mode**: The mode of a random variable X is any number x_{mod} such that $P_X(x_{mod}) \geq P_X(x)$ for all $x \in R_X$. Unlike the median, the mode of a discrete random variable must exist, but it may not be unique. The last example in the previous bullet has four possible values for the mode.

2.5 Important Families of Discrete Random Variables

Many experiments in engineering problems have the same underlying probability structure and give rise to the same type of random variable. In this section, we discuss several classes of discrete random variables that arise in many engineering applications. These classes of random variables have probability mass functions that can be described by a few parameters. Hence, they provide useful models for physical processes, as those parameters can be readily estimated from available sample data. Learning the properties of these random variables helps us avoid repetitive calculations.

The classes of random variables we discuss are:

- Bernoulli
- Uniform
- Binomial
- Geometric
- Poisson

For each family, we compute its statistics, so that we can avoid tedious summations when we can recognize the type of random variable involved.

2.5.1 Bernoulli(p) Random Variables

Let A be an event related to the outcome of some random experiment, such as a toss of a biased coin. Define the random variable X as the *indicator* function of A as:

$$X(\omega) = I_A(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is not in } A \\ 1 & \text{if } \omega \text{ is in } A. \end{cases}$$

Thus, X is one if the event A occurs, and zero otherwise. X is a random variable, with discrete values in range $\{0, 1\}$, and with probability mass function given by:

$$P_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1\ . \end{cases}$$

where $p = \mathbb{P}[A]$ in the original probability space. Such a random variable is called a *Bernoulli* random variable, since it identifies the outcome of a Bernoulli trial, which is 1 if the event A occurs.

The range of a Bernoulli random variable is $R_X = \{0, 1\}$. Its CDF is computed as:

$$F_X(x) = \begin{cases} 0 & x < 0, \\ (1-p) & x \in [0, 1) \\ 1 & x > 1. \end{cases}\ .$$

Note that $F_X(x)$ is defined for all real values of x.

The expected value and other statistics of Bernoulli random variables are easily computed, since R_X only has two entries:

$$\mathbb{E}[X] = \sum_{x=0}^{1} x P_X(x) = 0(1-p) + 1p = p.$$

$$\mathbb{E}[X^2] = \sum_{x=0}^{1} x^2 P_X(x) = 0^2(1-p) + 1^2 p = p.$$

Its variance is computed as

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p).$$

Bernoulli random variables are characterized by a single parameter p, which is easy to estimate from sample outcomes of the experiment. A summary of their properties is given below.

- X is a **Bernoulli**(p) random variable if it has PMF

$$P_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1 \; . \end{cases}$$

- Range: $R_X = \{0, 1\}$.
- Expected Value: $\mathcal{E}[X] = p$.
- Variance: $\mathsf{Var}[X] = p(1-p)$.
- Interpretation: single trial with success probability p.

2.5.2 Discrete Uniform(a, b) Random Variables

Suppose we have a discrete random variable X, with range in $R_X = \{a, a+1, a+2, \ldots, b\}$, where $a \leq b$ are integers, so it can take $b - a + 1$ values. We assume that the probability mass function $P_X(x)$ is the same for each value $x \in R_X$, so that each of the values is equally likely. In this case, $P_X(x) = \frac{1}{b-a+1}, x \in R_X$, as there are $b - a + 1$ possible values, and the normalization property requires $\sum_{x \in R_X} P_X(x) = 1$.

Discrete Uniform(a, b) random variables are used commonly in models of games of chance, such as coin tosses, roulette wheels, dice rolls, where there is no assumption of bias towards any of the outcomes. The outcomes in R_X are ordered in increasing order, and are separated by one unit.

We compute the statistics of a Discrete Uniform(a, b) random variable X as follows: Its CDF is given by

$$F_X(x) = \frac{\lfloor x \rfloor - a + 1}{b - a + 1}$$

where the notation $\lfloor x \rfloor$ refers to the largest integer less than or equal to x. The expected value of X is computed as:

$$\mathbb{E}[X] = \sum_{x \in R_X} x P_X(x) = \sum_{j=a}^{b} j \frac{1}{b - a + 1}$$

To do this sum, it helps to remember some summation equalities:

$$\sum_{j=1}^{n} j = \frac{n(n+1)}{2}$$

Let's define the derived random variable $Y = X - a$. Note that $R_Y = \{0, 1, 2, \ldots, b-a\}$, and $P_Y(y) = \frac{1}{b-a+1}, y \in R_y$. Now,

$$\mathbb{E}[Y] = \sum_{y \in R_Y} y P_Y(y) = \sum_{k=0}^{b-a} k \frac{1}{b-a+1} = \frac{1}{b-a+1} \frac{(b-a)(b-a+1)}{2} = \frac{b-a}{2}$$

For the original variable X, we know $\mathbb{E}[Y] = \mathbb{E}[X] - a$, so $\mathbb{E}[x] = a + \frac{b-a}{2} = \frac{a+b}{2}$.

We can compute the variance of Y as follows: First, we compute $\mathbb{E}[Y^2]$ as

$$\mathbb{E}[Y^2] = \sum_{k=0}^{b-a} k^2 \frac{1}{b-a+1}$$

To sum this, we use another summation formula:

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

Since the $k = 0$ term does not contribute to the sum $(k^2 = 0)$, we get:

$$\mathbb{E}[Y^2] = \sum_{k=0}^{b-a} k^2 \frac{1}{b-a+1} = \frac{(b-a)(b-a+1)(2(b-a)+1)}{6(b-a+1)} = \frac{(b-a)(2(b-a)+1)}{6}.$$

We compute the variance $\mathsf{Var}[Y]$ as

$$\mathsf{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{(b-a)(2(b-a)+1)}{6} - \frac{(b-a)^2}{4} = \frac{4(b-a)^2 + 2(b-a) - 3(b-a)^2}{12} = \frac{(b-a)b-a+2}{12}$$

Since $Y = X - k$, we know $\mathsf{Var}[Y] = \mathsf{Var}[X]$.

Uniform random variables are characterized by two parameters, k and n. Their properties are summarized below:

- X is a **Discrete Uniform**(a, b) random variable if it has PMF

$$P_X(x) = \begin{cases} \dfrac{1}{b-a+1} & x = a, a+1, \ldots, b, \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{a, a+1, \ldots, b\}$.
- Expected Value: $\mathcal{E}[X] = \dfrac{a+b}{2}$.
- Variance: $\mathsf{Var}[X] = \dfrac{(b-a)(b-a+2)}{12}$.
- Interpretation: equally likely to take any integer value between a and b.

2.5.3 Binomial(n, p) Random Variables

Suppose that a random experiment with a binary outcome of success or failure is repeated n times. Let x denote the number of times that such an experiment was a success. In terms of the notation used above in the context of Bernoulli random variables, let A denote an event, and let x denote the number of times that such an event occurs out of n independent trials of the same experiment. Then, X is a random variable with discrete range $\{0, 1, \ldots, n\}$. Define the parameter p to be the probability of success in a single trial of the experiment, as in Bernoulli random variables.

A simple representation of X is given by

$$X = I_1 + I_2 + \ldots + I_n, \tag{2.1}$$

where I_k is the indicator that event A occurs at the independent trial k.

We have seen this problem worked out in Section 1.4.4. The probability of any outcome with k successes out of n is $p^k(1-p)^{n-k}$. There are $\binom{n}{k}$ outcomes with k successes. Thus, the probability mass function of X is given by

$$P_X(k) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = k\}] = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

Thus, the CDF of X is given by

$$\sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{(n-k)},$$

where $\lfloor x \rfloor$ is the largest integer that is less than or equal to x.

Binomial(n, p) random variables arise in various applications where there are two types of outcomes, and we are interested in the number of outcomes of one type. Such applications include repeated coin tosses, correct/erroneous bits, good/defective items, active/silent stations, etc. The important statistics of binomial random variables are derived below:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} k p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \text{ (factor } np \text{ from sum)} \\
&= np \sum_{k'=0}^{n-1} \frac{(n-1)!}{(k')!(n-1-k')!} p^{k-1} (1-p)^{n-1-k'} \text{ (substitute } k' = k-1) \\
&= np
\end{aligned}$$

because the terms in the sum are the PMF for a Binomial$(n-1, p)$ RV, which add to 1 by normalization.

Similarly, to compute the variance of X, compute first the following expectation:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=0}^{n} \binom{n}{k} k(k-1)p^k(1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{n!}{k!(n-k)!} k(k-1)p^k(1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{n!}{(k-2)!(n-k)!} p^k(1-p)^{n-k} \\
&= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2}(1-p)^{n-k} \text{ (factor } n(n-1)p^2 \text{ from sum)} \\
&= n(n-1)p^2 \sum_{k'=0}^{n-2} \frac{(n-2)!}{(k')!(n-2-k'!} p^{k-2}(1-p)^{n-2-k'} \text{ substitute } k' = k-2 \\
&= n(n-1)p^2
\end{aligned}$$

because the last sum is again the sum of the PMF of a Binomial$(n-2, p)$ random variable, which is 1 by normalization.

Note now that $\mathbb{E}[X(X-1)] = \mathbb{E}[X^2] - \mathbb{E}[X]$, so $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = n(n-1)p^2 + np$. Now we use the identity

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = n^2p^2 - np^2 + np - n^2p^2 = n(p - p^2) = np(1-p).$$

In the above derivations, we have used extensive knowledge of binomial distributions to recognize identities, and to figure out how to factor terms so we can compute the sums. There is an alternative way of deriving these formulas, as discussed below.

Note that we can write $X = I_1 + I_2 + \ldots + I_n$, where I_k is the Bernoulli random variable indicating success in the k-th attempt. Then, using the linearity property of expectations, we have

$$\mathbb{E}[X] = \mathbb{E}[I_1 + I_2 + \ldots + I_n] = \mathbb{E}[I_1] + \mathbb{E}[I_2] + \ldots + \mathbb{E}[I_n] = np\ .$$

Note that we have avoided computing a difficult sum by using the fact that expectation is a linear operation, and the fact that, for Bernoulli random variables, $\mathbb{E}[I_k] = p$. To compute the variance, we use a property that we will derive in Chapter 5, that shows that the variance of a sum of **independent** random variables is the sum of the variances:

$$\mathsf{Var}[X] = \mathsf{Var}[I_1 + I_2 + \ldots + I_n] = \mathsf{Var}[I_1] + \ldots + \mathsf{Var}[I_n] = np(1-p).$$

Binomial random variables are characterized by the two parameters n and p. Their statistics are summarized below:

- X is a **Binomial**(n, p) random variable if it has PMF

$$P_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & x = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{0, 1, \ldots, n\}$.
- Expected Value: $\mathcal{E}[X] = np$.
- Variance: $\mathsf{Var}[X] = np(1-p)$.
- Interpretation: # of successes in n independent Bernoulli(p) trials.

2.5.4 Geometric(p) Random Variables

The binomial random variable is obtained by fixing the number of Bernoulli trials and counting the number of successes. A different random variable is obtained by counting the number of trials until the first success occurs. Denote this random variable as X; this is a *geometric* random variable, and it takes values in the discrete infinite set $\{1, 2, \ldots\}$.

Note that $X = 1$ if and only if the first Bernoulli trial is successful. Hence, $P_X(1) = p$, where p is the single trial probability of success For $X = 2$, the first Bernoulli trial must fail, but the second one must succeed. Since the trials are independent, $P_X(2) = (1 - p)p$. Reasoning along the same lines, $X = k$ if and only if the first $k - 1$ Bernoulli trials failed, but the k-th Bernoulli trial succeeded. Using the independence properties, we get

$$P_X(k) = (1 - p)^{k-1}p, k = 1, 2, \ldots$$

The corresponding CDF is

$$F_X(x) = 1 - (1 - p)^{\lfloor x \rfloor} \ .$$

Geometric(p) random variables arise in applications where one is interested in the time between occurrence of events in a sequence of independent experiments. Such random variables have broad applications in different aspects of queuing theory. The important statistics of geometric random variables are summarized below:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kP_X(k) = \sum_{k=1}^{\infty} kp(1 - p)^k - 1$$

To sum the above expression, we use the following summation for geometric series for $0 < q < 1$:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1 - q} \ .$$

Differentiating both sides with respect to q (which is justified by the summability of the series for $p < 1$) yields:

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1 - q)^2}$$

Using this formula, we get:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kp(1 - p)^k - 1 = p\frac{1}{p^2} = \frac{1}{p} \ .$$

To compute the variance, we take another derivative of the summation equality, to get

$$\begin{aligned} \frac{d}{dq}\sum_{k=1}^{\infty} kq^{k-1} &= \sum_{k=1}^{\infty} \frac{d}{dq}kq^{k-1} \\ &= \sum_{k=1}^{\infty} k(k-1)q^{k-2} = \sum_{k=1}^{\infty} k(k-1)q^{k-2} \\ &= \frac{d}{dq}\frac{1}{(1 - q)^2} = \frac{2}{(1 - q)^3} \end{aligned}$$

Substituting $1 - p = q$ yields $\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} = \frac{2}{p^3}$.

Using these formulas allows us to compute $\mathbb{E}[X^2]$ as

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2p(1 - p)^{k-1} = \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} + \sum_{k=1}^{\infty} kp(1 - p)^{k-1} \\ &= \frac{2(1 - p)}{p^2} + \frac{1}{p} = \frac{(2 - p)}{p^2} \end{aligned}$$

Hence,

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

The geometric distribution is specified by a single parameter, p. Its properties are summarized below:

- X is a **Geometric**(p) random variable if it has PMF
$$P_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$
- Range: $R_X = \{1, 2, \dots\}$.
- Expected Value: $\mathcal{E}[X] = \frac{1}{p}$.
- Variance: $\mathsf{Var}[X] = \frac{1-p}{p^2}$.
- Interpretation: # of independent Bernoulli(p) trials until first success.

2.5.5 Poisson(λ) Random Variables

In many applications, we are interested in counting the number of occurrences of an event in a certain time period or in a certain region of space. The Poisson random variable arises in situations where the events occur "completely at random" in time or space; that is, where the likelihood of an event occurring at a particular time is equal to and independent of the event occurring at a different time. For example, Poisson random variables arise in counts of emissions from radioactive substances, in the number of photons emitted as a function of light intensity, in counts of demands for telephone connections, and in counts of defects in a chip.

One of the applications of the Poisson random variable is as an approximation to the binomial probabilities when the number of trials is large. If the number of trials n is large, and if p is small, then, letting $\lambda = np$, Simeón Poisson established this limit:

$$\lim_{n\to\infty, np=\lambda} \frac{n|}{k!(n-k)!} p^k (1-p)^{n_t - k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

We briefly overview his proof below. Let K_n be the binomial random variable for n trials, each of which has probability λ/n of succeeding. The probability mass function of K_n is

$$P_{K_n}(k) = nchoosek(\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} (1 - \frac{\lambda}{n})^{n-k}$$

Note the following limits:

$$\lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} = 1 \text{(same highest order power in numerator, denominator)} .$$

$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^n = e^{-\lambda} \text{ (Definition of exponential).}$$

$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^k = 1 .$$

Thus,

$$\lim_{n\to\infty} P_{K_n}(k) = \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \lim_{n\to\infty} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Poisson(λ) random variables have an infinite, countable sample space $R_X = \{0, 1, 2, \ldots\}$ with the probability mass function

$$P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda} \ .$$

where λ is the average number of event occurrences in the specified time interval or region of space. The corresponding CDF of X is

$$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\lambda^k e^{-\lambda}}{k!} \ .$$

To compute the mean and variance of a Poisson(λ) random variable, we use a well-known summation formula

$$\sum_{k=0}^{\infty} \frac{(\lambda)^k}{k!} = e^{\lambda}$$

Then,

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^{\infty} k P_X(k) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \quad \text{since the last sum is equal to } e^{\lambda}. \end{aligned}$$

To compute the variance, we compute the second moment first:

$$\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 P_X(k) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda}$$

In order to get an expression for this sum, we differentiate the exponential summation twice with respect to λ, to obtain

$$\frac{d^2}{d\lambda^2} e^{\lambda} = e^{\lambda} = \sum_{k=0}^{\infty} \frac{d^2}{d\lambda^2} \frac{\lambda^k}{k!} = \sum_{k=2}^{\infty} (k^2 - k) \frac{\lambda^{k-2}}{k!}$$

Therefore,

$$\mathbb{E}[X^2] = \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{(k)!} e^{-\lambda} = \sum_{k=1}^{\infty} (k^2 - k) \frac{\lambda^k}{(k)!} e^{-\lambda} + \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k)!} e^{-\lambda} = \lambda^2 + \lambda$$

We now compute the variance of X as

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda \ .$$

Poisson(λ) random variables are specified by a single parameter λ. Its properties are summarized below:

- X is a **Poisson**(λ) random variable if it has PMF

$$P_X(x) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda} & x = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{0, 1, \ldots\}$.
- Expected Value: $\mathcal{E}[X] = \lambda$.
- Variance: $\mathsf{Var}[X] = \lambda$.

- Interpretation: # of arrivals in a fixed time window.

Example 2.15
Suppose we are at a service facility, with a total number of five servers. Assume there are seven potential customers in the facility, and the probability that any of them will require service is p, where each customer will require service independent of any other customers' requirements. Let X be the random variable denoting the number of service requests. What type of random variable is X? What is the expected number of requests? What is the probability that there will be more requests than available servers?

First, we recognize that the random variable X is a binomial random variable, as the sum of independent Bernoulli random variables (0-1 requests), with parameters $n = 7$ and p. The expected number of requests is thus $7p$. The probability that there will be more requests than available servers is

$$\mathbb{P}_X[\{X > 5\}] = P_X(6) + P_X(7) = \binom{7}{6} p^6(1-p) + \binom{7}{7} p^7 = 7p^6(1-[p) + p^7 \ .$$

Example 2.16
You are waiting for a taxi at the corner of St. Mary's street and Commonwealth Avenue. When a taxi goes by the corner, there is a 0.9 probability that the taxi is occupied, and will not stop to pick you up. Assume that whether a taxi is occupied or not is independent of whether other taxis are occupied. Let X denote the number of taxis that come by the corner until one of them picks you up. What type of random variable is X? What is the expected number of taxis that you will see until you are picked up?

We recognize that whether each taxi is occupied or not is a Bernoulli trial, and the probability of success is $p = 0.1$. The random variable X is thus a geometric random variable. The expected number of taxis that you should expect to see until being picked up is thus $\mathbb{E}[X] = 10$.

Example 2.17
. Assume you have an X-ray source generating an X-ray beam with intensity equal to 10^5 photons/second towards a detector. Let X denote the number of photons collected by the detector photons over a period of a millisecond. If X is a Poisson random variable, what are its mean and standard deviation?

We compute the parameter $\lambda = 10^5 \cdot 10^{-3} = 100$ for the Poisson distribution of X. In this case, $\mathbb{E}[X] = 100, \text{Var}[X] = 100$. Thus, the standard deviation is $\sigma_X = \sqrt{\text{Var}[X]} = 10$.

Example 2.18
Suppose each episode of Game of Thrones includes a death of a major character with probability $3/4$, independent of whether deaths happen in any other episode. Assume there are an infinite number of episods to watch (it felt that way sometimes...) Define X to be the number of episodes you watch until you see the death of a major character. What type of random variable is X?

X is a Geometric($\frac{3}{4}$) random variable, where we explicitly provide the value for the parameter. Then, we know its statistics:

$$\mathbb{E}[X] = \frac{1}{p} = \frac{4}{3}; \text{Var}[X] = \frac{1-p}{p^2} = \frac{\frac{1}{4}}{\frac{9}{16}} = \frac{4}{9}.$$

What is the probability that $X \geq 3$? Sometimes it is easier to compute the probability of the complement: the probability that $X \leq 2$. We know

$$\mathbb{P}_X[\{X \leq 2\}] = P_X(1) + P_X(2) = p + p(1-p) = \frac{3}{4} + \frac{3}{4} \cdot \frac{1}{4} = \frac{15}{16}.$$

Hence, $\mathbb{P}_X[\{X \geq 3\}] = 1 - \mathbb{P}_X[\{X \leq 2\}] = \frac{1}{16}$.

Let Y denote the number of episodes out of the first six you watch that contain a major character death. What type of random variable is Y? Y is a Binomial($6, \frac{3}{4}$) random variable. Hence, its key statistics are:

$$\mathbb{E}[Y] = np = 6 \cdot \frac{3}{4} = \frac{9}{2}; \quad \text{Var}[Y] = np(1-p) = 6 \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{9}{8}.$$

What is $E[Y^2]$? We know that $E[Y^2] = E[Y]^2 + \text{Var}[Y] = \frac{81}{4} + \frac{9}{8} = \frac{171}{8}$.

What is the probability that less than half of the six episodes include a death? This is $\mathbb{P}_Y[\{Y < 3\}]$, which we compute as

$$\begin{aligned}\mathbb{P}_Y[\{Y<3\}] &= P_Y(0) + P_Y(1) + P_Y(2) = (\frac{1}{4})^6 + 6(\frac{1}{4})^5(\frac{3}{4}) + \binom{6}{2}(\frac{1}{4})^4(\frac{3}{4})^2 \\ &= \frac{1+18+135}{4^6} = \frac{77}{2048}\end{aligned}$$

Given that less than half the episodes contain a death, what is the probability that exactly two of the six episodes contain a death? This is a conditional probability question: Let's define this in terms of events in the original probability space. Define event $A = \{Y = 2\}, B = \{Y \leq 2\}$. Then,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[\{Y=2\}]}{\mathbb{P}[B]} = \frac{\frac{135}{6^4}}{\frac{154}{6^4}} = \frac{135}{154}.$$

We specialize the techniques of computing conditional probabilities to handle events defined in terms of random variables in the next section.

2.6 Conditional Probability Models

In Chapter 1, given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and an event $B \in \mathcal{E}$, we defined the conditional probability of any other event $A \in E$, conditioned on observing B, as

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

provided $\mathbb{P}[B] > 0$. Otherwise, we left the conditional probability as undefined. A discrete random variable X in $(\Omega, \mathcal{E}, \mathbb{P})$ with range in $R_X = \{x_i, i \in 1, 2, \ldots,\}$ defines events $A_i = \{\omega \in \Omega : X(\omega) = x_i\} \in \mathcal{E}$. Those events were used to define the probability mass function $P_X(x_i) = \mathbb{P}(A_i)$.

Assume we observe an event $B \in E$. We can define the conditional probability mass function of X given B as

$$P_{X|B}(x_i) = \mathbb{P}[A_i|B] = \begin{cases} \frac{\mathbb{P}[\{\omega \in \Omega : X(\omega) = x_i\} \cap B]}{\mathbb{P}[B]} & \text{if } \mathbb{P}[B] > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

This conditional probability mass function will have all the properties of a probability mass function on R_X, satisfying the basic properties of non-negativity, normalization and additivity:

$$\begin{aligned} P_{X|B}(x) &> 0\text{for all } x \in R_x \\ \sum_{x \in R_x} P_{X|B}(x) &= 1 \\ \sum_{x \in C} P_{X|B}(x) &= m_{X|B}[C] \text{ for all } C \subset R_X \end{aligned}$$

There is a special case of interest, where we observe the event that X takes it values in a set $B \subset R_X$, and the conditioning event is $B_1 = \{\omega \in \Omega : X(\omega) \in B_1\}$. We are guaranteed that $B_1 \in \mathcal{E}$ is an event because X is a random variable, and $\mathbb{P}[B_1] = \mathbb{P}_X[B]$. In this special case, the conditional probability mass function simplifies: Specifically, note that

$$\{\omega \in \Omega : X(\omega) = x\} \cap B_1 = \begin{cases} \{\omega \in \Omega : X(\omega) = x\} & \text{if } x \in B\} \\ \emptyset & \text{if } x \notin B \end{cases}$$

We write, with a small abuse of notation, the conditional PMF $P_{X|B}(x)$ as

$$P_{X|B}(x) = \begin{cases} \frac{\mathbb{P}[\{\omega\in\Omega : X(\omega)=x\}]}{\mathbb{P}_X[B]} = \frac{P_X(x)}{\mathbb{P}_X(B)} & \text{if } x \in B \text{ and } \mathbb{P}_X[B] > 0 \\ 0 & \text{if } x \notin B \text{ and } \mathbb{P}_X[B] > 0 \\ \text{undefined} & \text{if } \mathbb{P}_X[B] = 0. \end{cases}$$

Thus, the conditional PMF $P_{X|B}(x)$ is proportional to the unconditional PMF $P_X(x)$, restricted to $x = B$, and rescaled to satisfy the normalization property. It is zero for any values $x \notin B$.

The conditional probability mass function has a range $R_{X|B} \subset B$, and satisfies all the properties of probability mass functions.

- (Non-negativity) $P_{X|B}(x) \geq 0$.
- (Normalization) $\sum_{x\in B} P_{X|B}(x) = 1$.
- (Additivity) For any set $C \in R_X$, the conditional probability that $X \in C$ given B is

$$\mathbb{P}[\{\omega \in \Omega : X(\omega) \in C\}|\{X(\omega) \in B\}] \equiv \mathbb{P}_{X|B}[C] = \sum_{x\in C} P_{X|B}(x).$$

Note that $\mathbb{P}_X[B] = \sum_{x_k\in B} P_X(x_k)$. Thus, we can write the conditional probability mass function of X given B entirely in terms of the random variable X and its probability mass function, as

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{sum_{x_k\in B} P_X(x_k)} & \text{if } x \in B \text{ and } sum_{x_k\in B} P_X(x_k) > 0 \\ 0 & \text{if } x \notin B \text{ and } sum_{x_k\in B} P_X(x_k) > 0 \\ \text{undefined} & \text{if } sum_{x_k\in B} P_X(x_k) = 0. \end{cases}$$

Now that we have a conditional probability mass function, we can define conditional statistics for the random variable X. For instance, the conditional expected value of X given an event B is given as

$$\mathbb{E}[X|B] = \sum_{x\in R_X} x P_{X|B}(x)$$

and the conditional variance as

$$\mathsf{Var}[X|B] = \mathbb{E}[(X - \mathbb{E}[X|B_1])^2|B] = \mathbb{E}[X^2|B_1] - (E[X|B])^2$$

For any function $g(X)$ that defines a derived random variable $Y = g(X)$, we can define the conditional expectation as

$$\mathcal{E}\big[g(X)|B\big] = \sum_{x\in R_X} g(x) P_{X|B}(x) \ .$$

Example 2.19
Assume X is a Binomial$(5, \frac{1}{3})$ random variable. Define $B = \{X \leq 2\}$. Compute $P_{X|B}(x)$, $\mathbb{E}[X|B]$ and $\mathsf{Var}[X|B]$.

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{\mathbb{P}_X[B]} & x \in B \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{P}_X[B] = P_X(0) + P_X(1) + P_X(2) = (\frac{2}{3})^5 + 5(\frac{2}{3})^4(\frac{1}{3}) + 10(\frac{2}{3})^3(\frac{1}{3})^2 = \frac{32+80+80}{3^5} = \frac{64}{81}.$$

$$P_{X|B}(0) = \frac{P_X(0)}{\mathbb{P}_X[B]} = \frac{32}{192} = \frac{1}{6}; P_{X|B}(1) = P_{X|B}(2) = \frac{80}{192} = \frac{5}{12}$$

Thus,

$$P_{X|B}(x) = \begin{cases} \frac{1}{6} & x = 0, \\ \frac{5}{12} & x = 1, 2, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[X^2|B] = 0P_{X|B}(0) + 1P_{X|B}(1) + 2P_{X|B}(2) = \frac{15}{12} = \frac{5}{4}.$$

$$\mathbb{E}[X^2|B] = 0^2P_{X|B}(0) + 1^2P_{X|B}(1) + 2^2P_{X|B}(2) = \frac{25}{12}.$$

$$\text{Var}[X|B] = \mathbb{E}[X^2|B] - (\mathbb{E}[X^2|B])^2 = \frac{25}{12} - \frac{25}{16} = \frac{25}{48}.$$

Example 2.20
Consider a manufacturing station, where the random arrival time X of a part to be processed is uniformly distributed in $R_X = \{1, 2, 3, \ldots, 20\}$. Thus, the probability mass function of X is $P_X(x) = \frac{1}{20}, x \in R_X$. Assume that you wait for the first 6 slots and the part X has not arrived yet. Equivalently, you observe the event $B = \{X > 6\}$. Compute the conditional probability mass function of X given B and its conditional expected value and variance.

Since B is defined in terms of R_X, we can use the simpler formula, restricting and rescaling the original $P_X(x)$. Note that $\mathbb{P}_X[B] = \sum_{x\in B} P_X(x) = \frac{14}{20}$. Then,

$$P_{X|B}(x) = \begin{cases} 0 & x \leq 6, \\ \frac{\frac{1}{20}}{\frac{14}{20}} = \frac{1}{14} & x > 6. \end{cases}.$$

Note that this is now a uniform distribution from 7 to 20, so we can use formulas for uniform distribution to compute mean and variance. The conditional expected value is

$$\mathbb{E}[X|B] = \sum_{x\in B} xP_{X|B}(x) = \frac{7+20}{2} = 13.5$$

The conditional variance is

$$\text{Var}[X|B] = \frac{(20-7)(20-7+2)}{12} = \frac{(13)(15)}{12} = \frac{65}{4} = 16.25.$$

Example 2.21
One of the interesting properties of a geometric random variable X is that it is "memoryless". Let X be a geometric random variable with parameter p. Assume we observe the event $B = \{X > k\}$ for some value k. What is the conditional mass distribution of X given B? Recall that $R_X = \{1, 2, \ldots, \}$, and $B = \{k+1, k+2, \ldots, \}$.

We compute

$$\mathbb{P}_X[B] = \sum_{k=7}^{\infty} P_X(k) = \sum_{k=7}^{\infty} p(1-p)^{k-1} = (1-p)^6 \sum_{k=1}^{\infty} p(1-p)^{k-1} = (1-p)^6$$

because we know, from normalization, that $\sum_{k=1}^{\infty} p(1-p)^{k-1} = 1$. Hence,

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{\mathbb{P}_X[B]} & x \geq 7 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p(1-p)^{x-7} & x \geq 7 \\ 0 & \text{otherwise} \end{cases}$$

Define the additional wait time random variable $T = X - 6$. Then, note that

$$P_{T|B}(t) = \begin{cases} p(1-p)^t & t \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus, conditioned on B, T is a geometric random variable with the same parameter p as the original random variable X.

In words, the above expression states that, if a success has not occurred in the first j trials, the probability of having to perform at least k more trials until a success is the same as the probability of initially having to perform at least k trials. Thus, the system "forgets" the past failures and begins anew as if it were performing the first trial.

Hence, if you are waiting for a bus that should arrive in 10 minutes, and you have already waited two hours, the expected arrival time of the bus is still 10 minutes from now...as long as the arrival time was a geometric random variable.

Note that conditional probability mass functions obey the usual laws that probability mass functions obey. For instance, for a random variable X defined in the probability space $(\Omega, \mathcal{E}, \mathbb{P})$, we have:

- **Multiplication Rule:** For a random variable X and event $B \in \mathcal{E}$,

  $$\mathbb{P}\big[\{X = x\} \cap B\big] = P_{X|B}(x)\,\mathbb{P}[B]\ .$$

  If $B \subset R_X$, then

  $$\mathbb{P}\big[\{X = x\} \cap \{X \in B\}\big] = P_{X|B}(x)\mathbb{P}_X[B] = \begin{cases} P_X(x) & x \in B \\ 0 & \text{otherwise.} \end{cases}$$

- **Law of Total Probability:** For a partition of R_X as $B_1, \ldots, B_n$, we can write the probability mass function as a weighted sum of conditional probability mass functions, as:

  $$P_X(x) = \sum_{i=1}^{n} P_{X|B_i}(x)\mathbb{P}_X[B_i]\ .$$

- **Bayes' Rule:** We can "flip" the conditioning, as in Bayes' Rule, with some care. Let $B \subset R_X$. Then,

  $$\mathbb{P}_X\big[B \big| \{X = x\}\big] = \frac{P_{X|B}(x)\mathbb{P}_X[B]}{P_X(x)}\ .$$