Chapter 5. Discrete Random Variables

5.1 Random Variables

Very often the outcomes in a sample space of an experiment are associated with numbers. Randomness or uncertainty in the outcome of the experiment then induces a randomness or uncertainty in the associated number. This is the idea of a random variable — a numerical quantity whose value is determined by a probabilistic process.

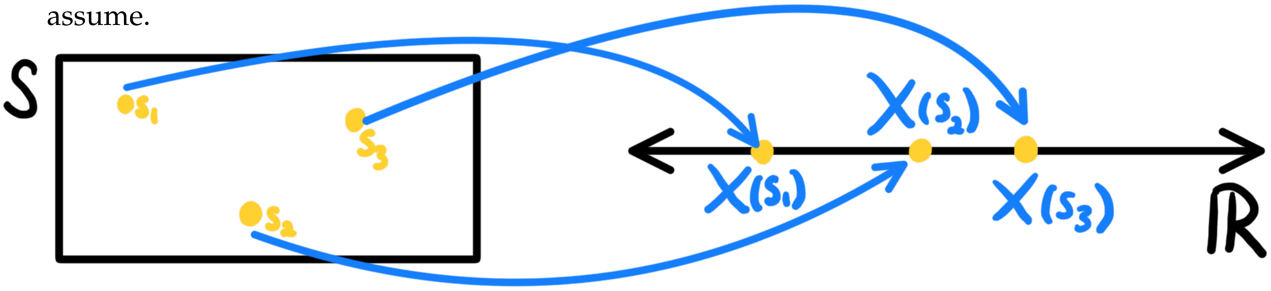
To be precise, a **random variable** can be thought of as a function which maps the sample space S to the real numbers $\mathbb{R}$.

Definition 5.1.1. *A* **random variable** *is a (measurable) mapping*

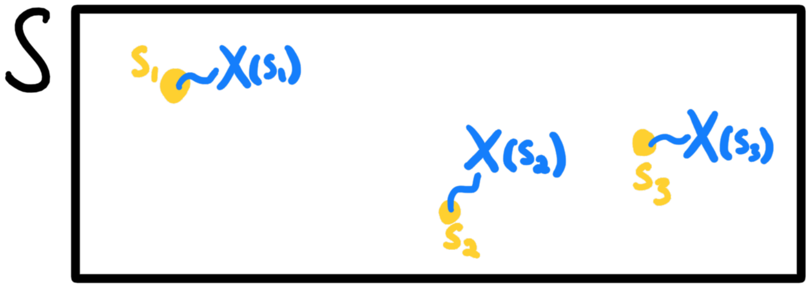
$$X : S \to \mathbb{R}$$

with the property that $\{s \in S : X(s) \leq x\} \in \mathcal{F}$ *for each* $x \in \mathbb{R}$.

If the experiment yields the specific outcome $s \in S$ then the outcome of the random variable is the particular real number $x = X(s)$. We try to always use a capital letter to denote a random variable and a lowercase letter to denote a particular value the random variable can assume.



or



The RV labels every outcome in S

Definition 5.1.2. *The image of S under X is called the* **range** *of the random variable:*

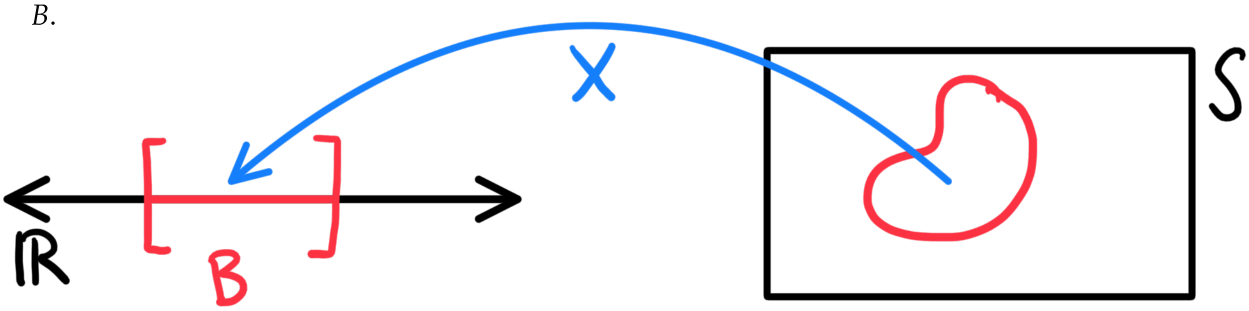
$$\mathbb{X} \equiv X(S) = \{X(s) \mid s \in S\} = \{x \in \mathbb{R} \mid \exists s \in S \text{ s.t. } X(s) = x\}$$

So as S contains all the possible outcomes of the experiment, $\mathbb{X}$ contains all the possible outcomes for the random variable X.

The value of the random variable is determined by the outcome of the experiment. This lets us transfer the probability measure P defined on $\mathcal{F}$ to the real numbers in a natural way:

$$\mathrm{P}_X\left(X \in \text{some set } B \text{ of real numbers}\right) := \mathrm{P}\left(\{s \in S : X(s) \in B\}\right),$$

where $\mathrm{P}_X(X \in B)$ is the probability that the random variable X will have a value in the set B.



The function P_X is called the **probability distribution** for the random variable X. It is really just a probability measure for a new probability space whose sample space is $\mathbb{R}$.

Examples

$$\mathrm{P}_X(X \leq b) = \mathrm{P}(\{s \in S : X(s) \leq b\})$$

$$\mathrm{P}_X(X = 7) = \mathrm{P}(\{s \in S : X(s) = 7\})$$

$$\mathrm{P}_X(a < X \leq b) = \mathrm{P}(\{s \in S : a < X(s) \leq b\})$$

Note The shorthand introduced above is standard in probability theory. In general, if $B \subset \mathbb{R}$,

$$\{X \in B\} \text{ denotes the event } \{s \in S : X(s) \in B\}$$

and

$$\mathrm{P}_X(X \in B) := \mathrm{P}_X(\{X \in B\}) = \mathrm{P}(\{s \in S : X(s) \in B\}).$$

If B is an interval such as $B = (a, b]$, we can write

$$\{a < X \leq b\} := \{X \in (a, b]\}, \text{ which denotes the event } \{s \in S : a < X(s) \leq b\}.$$

Analogous notation applies to intervals such as $[a, b]$, $[a, b)$, (a, b), $(-\infty, b)$, $(-\infty, b]$, (a, ∞), $[a, \infty)$, along with unions and intersections of such intervals.

The property $\{s \in S : X(s) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$ in Definition 5.1.1 ensures that any set $B \subseteq \mathbb{R}$ we might conceivably ask about corresponds to an event in the event space $\mathcal{F}$.

Example Let our random experiment be tossing a fair coin, with sample space $S = \{\mathrm{H}, \mathrm{T}\}$ and probability measure $\mathrm{P}(\{\mathrm{H}\}) = \mathrm{P}(\{\mathrm{T}\}) = \dfrac{1}{2}$. The event space is $\mathcal{F} = \{\emptyset, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H,T}\}\}$.

We can define a random variable $X : \{\mathrm{H}, \mathrm{T}\} \to \mathbb{R}$ taking values, say,

$$X(\mathrm{T}) = 0 \quad \text{and} \quad X(\mathrm{H}) = 1$$

In this case,

$$\{X \leq x\} \text{ denotes the events } \{s \in S : X(s) \leq x\} = \begin{cases} \emptyset & \text{if } x < 0; \\ \{\mathrm{T}\} & \text{if } 0 \leq x < 1; \\ \{\mathrm{H}, \mathrm{T}\} & \text{if } x \geq 1. \end{cases}$$

This lets us compute the values of the probability P_X on the continuum $\mathbb{R}$

$$\mathrm{P}_X(X \leq x) = \begin{cases} \mathrm{P}(\emptyset) = 0 & \text{if } x < 0; \\ \mathrm{P}(\{\mathrm{T}\}) = \frac{1}{2} & \text{if } 0 \leq x < 1; \\ \mathrm{P}(\{\mathrm{H}, \mathrm{T}\}) = 1 & \text{if } x \geq 1. \end{cases}$$

For example, we can say "The probability that X will be less than or equal to 0.2 is 1/2." ■

Example Consider counting the number of heads in a sequence of 3 coin tosses. The underlying sample space is

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

which contains the 8 possible sequences of tosses. However, since we are only interested in the number of heads in each sequence, we define the random variable X by

$$X(s) = \begin{cases} 0, & s = TTT, \\ 1, & s \in \{TTH, THT, HTT\}, \\ 2, & s \in \{HHT, HTH, THH\}, \\ 3, & s = HHH. \end{cases}$$

This mapping is illustrated in Figure 5.1 below.

Continuing this example, let us assume that the sequences are equally likely. Now lets find the probability that the number of heads X is less than 2. In other words, we want to find $\mathrm{P}_X(X < 2)$. But what does this precisely mean? $\mathrm{P}_X(X < 2)$ is the probability

$$\mathrm{P}(\{s \in S : X(s) < 2\}).$$

The first step in calculating the probability is therefore to identify the event $\{s \in S : X(s) < 2\}$. In Figure 5.1, the only lines pointing to the numbers less than 2 are the lines pointing to 0 and 1. Tracing these lines backwards from $\mathbb{R}$ into S, we see that

$$\{s \in S : X(s) < 2\} = \{TTT, TTH, THT, HTT\}.$$

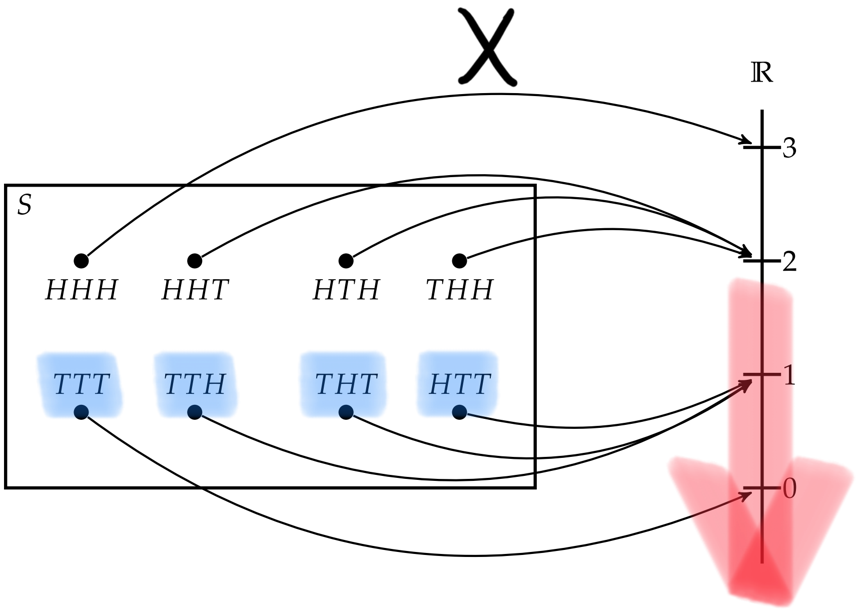


Figure 5.1: Illustration of a random variable X that counts the number of heads in a sequence of 3 coin tosses.

Since we have assumed that the sequences are equally likely

$$P_X(X < 2) = P(\{TTT, TTH, THT, HTT\}) = \frac{|\{TTT, TTH, THT, HTT\}|}{|S|} = \frac{4}{8} = \frac{1}{2}$$

On the same sample space, we can define another random variable able to describe the event that the number of heads in 3 tosses is even. Define this random variable, Y, as

$$Y(s) = \begin{cases} 0, & s \in \{TTT, THH, HTH, HHT\} \\ 1, & s \in \{TTH, THT, HTT, HHH\} \end{cases}.$$

The probability that the number of heads is less than two and odd is $P(X < 2, Y = 1)$, by which we mean the probability of the event

$$\{s \in S : X(s) < 2 \text{ and } Y(s) = 1\}.$$

This event is equal to

$$\{s \in S : X(s) < 2\} \cap \{s \in S : Y(s) = 1\}$$

which is

$$\{TTT, TTH, THT, HTT\} \cap \{TTH, THT, HTT, HHH\} = \{TTH, THT, HTT\}.$$

The probability of this event, assuming all sequences are equally likely, is 3/8. ■

5.1.1 Cumulative Distribution Function

For a random variable X we define the cumulative distribution function (CDF or sometimes just called the distribution function) as follows:

Definition 5.1.3. *The* **cumulative distribution function** (**CDF**) *of a random variable X is the function $F_X : \mathbb{R} \to [0,1]$, defined by*

$$F_X(x) = \mathrm{P}_X(X \le x) = \mathrm{P}(\{s \in S : X(s) \le x\})$$

For any random variable X, F_X is right-continuous, meaning if a decreasing sequence of real numbers $x_1, x_2, \ldots \to x$, then $F_X(x_1), F_X(x_2), \ldots \to F_X(x)$.

Given a right-continuous function $F_X(x)$, to check if it is a valid CDF, we need to make sure the following necessary and sufficient conditions hold:

i) $0 \le F_X(x) \le 1, \forall x \in \mathbb{R}$;

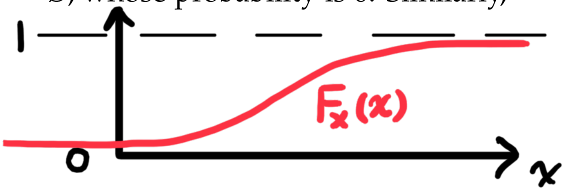
$F_X(x)$ is a probability, and probabilities are always in the range $[0,1]$.

ii) Monotonicity: $\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \Rightarrow F_X(x_1) \le F_X(x_2)$;

$F_X(x_1)$ is the probability of the event $E_1 = \{s \in S : X(s) \le x_1\}$ and similarly for $F_X(x_2)$. $E_1 \subseteq E_2$, and therefore $\mathrm{P}(E_1) \le \mathrm{P}(E_2)$.

iii) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

To be rigorous we'd need to get into limits of sequences of events. But you can think of $F_X(-\infty)$ as corresponding to the event $\{X \le -\infty\} = \emptyset$, whose probability is 0. Similarly, $\{X \le +\infty\} = S$, whose probability is 1.



For finite intervals $(a,b] \subseteq \mathbb{R}$, it is easy to check that

$$\mathrm{P}_X(a < X \le b) = F_X(b) - F_X(a).$$

The relevant events are $E_a = \{X \le a\}$, $E_b = \{X \le b\}$, and $E_{ab} = \{a < X \le b\}$. Check that $E_b = E_a \cup E_{ab}$ is a disjoint union and therefore $\mathrm{P}(E_b) = \mathrm{P}(E_a) + \mathrm{P}(E_{ab})$. But these three terms are just $F_X(b)$, $F_X(a)$, and $\mathrm{P}_X(a < X \le b)$, respectively.

Unless there is any ambiguity, we generally suppress the subscript of $\mathrm{P}_X(\cdot)$ in our notation and just write $\mathrm{P}(\cdot)$ for the probability measure for the random variable.

- That is, we forget about the underlying sample space S and event space $\mathcal{F}$. If there's any confusion we can always be careful and use the mapping between S and $\mathbb{R}$.
- Often, it will be most convenient to work this way and consider the random variable directly from the very start, with either the range of X or $\mathbb{R}$ being our sample space.

5.2 Discrete Random Variables

Definition 5.2.1. *A random variable X is* **discrete** *if the range of X, denoted by $\mathbb{X}$, is countable, that is*

$$\mathbb{X} = \{x_1, x_2, \ldots, x_n\} \quad \text{(FINITE)} \quad \text{or} \quad \mathbb{X} = \{x_1, x_2, \ldots\} \quad \text{(INFINITE)}.$$

Note The even numbers, the odd numbers and the rational numbers are countable; the set of real numbers between 0 and 1 is not countable.

Definition 5.2.2. *For a discrete random variable X, we define the* **probability mass function** *(often called the pmf) as*

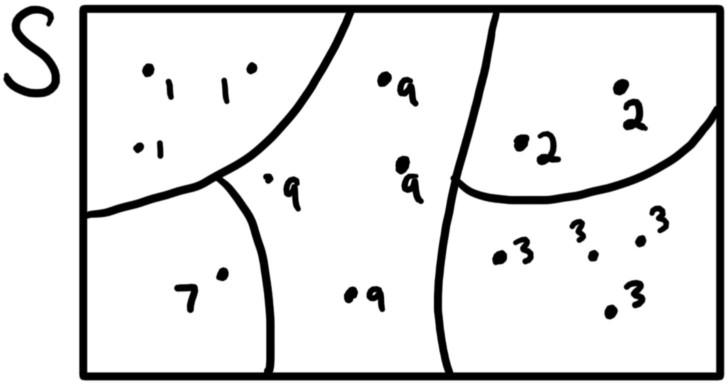
$$p_X(x) = \mathrm{P}_X(X = x), \quad x \in \mathbb{X}.$$

Note For completeness, we define

$$p_X(x) = 0, \quad x \notin \mathbb{X}.$$

so that p_X is defined for all $x \in \mathbb{R}$. Furthermore, we will refer to the **support** of random variable X as the set of $x \in \mathbb{R}$ such that $p_X(x) > 0$. The support is almost always the same as the range $\mathbb{X}$.

Label each outcome s in S according to the value of the RV, X(s).



→ partition of the Sample space

$$S = \bigcup_i \{s \in S : X(s) = x_i\},$$

where $\{x_1, x_2, \ldots\}$ = range of X

$$\overset{\text{pmf}}{p_X}(x) = \overset{\text{probability distribution}}{P_X}(X = x) = \overset{\text{probability measure}}{P}(\{s \in S : X(s) = x\})$$

5.2.1 Properties of Mass Function p_X

A function p_X is a probability mass function for a discrete random variable X with range $\mathbb{X}$ of the form $\{x_1, x_2, \dots\}$ if and only if

i) $p_X(x_i) \geq 0$;

ii) $\sum_{x \in \mathbb{X}} p_X(x) = 1$.

Proof of ii) Partition sample space. $1 = \mathrm{P}(S) = \mathrm{P}\left(\bigcup_{x \in \mathbb{X}} \{s \in S : X(s) = x\}\right) = \sum_{x \in \mathbb{X}} \mathrm{P}\left(\{s \in S : X(s) = x\}\right) = \sum_{x \in \mathbb{X}} \mathrm{P}_X(x)$.

5.2.2 Discrete Cumulative Distribution Function

The cumulative distribution function, or CDF, F_X of a discrete random variable X is defined by

$$F_X(x) = \mathrm{P}(X \leq x), \quad x \in \mathbb{R}.$$

5.2.3 Connection between F_X and p_X

Let X be a discrete random variable with range $\mathbb{X} = \{x_1, x_2, \dots\}$, where the x_i's are sorted into increasing order (i.e. $x_i < x_j$ if $i < j$). The probability mass function is p_X and the CDF is F_X. Then, for any real value x,

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i).$$

This relation is equivalent to $p_X(x_i) = F_X(x_i) - F_X(x_{i-1}), \forall i$. Thus we can compute the CDF from the PMF and vice versa.

5.2.4 Properties of Discrete CDF F_X

i) In the limiting cases,

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1.$$

ii) F_X is continuous from the right on $\mathbb{R}$, that is, for $x \in \mathbb{R}$,

$$\lim_{h \to 0^+} F_X(x+h) = F_X(x)$$

iii) F_X is non-decreasing, that is,

$$a < b \implies F_X(a) \leq F_X(b).$$

iv) For $a < b$

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a).$$

Note The key idea is that the functions p_X and/or F_X can be used to describe the probability distribution of the random variable X. A graph of the function p_X is non-zero only at the elements of $\mathbb{X}$. A graph of the function F_X is a step-function which takes the value zero at minus infinity, the value one at infinity, and is non-decreasing with points of discontinuity at the elements of $\mathbb{X}$.

Example Consider a coin tossing experiment where a fair coin is tossed repeatedly under identical experimental conditions, with the sequence of tosses independent, until a head is obtained. For this experiment, the sample space, S, consists of the set of sequences $S = \{H, TH, TTH, \ldots\}$ with associated probabilities for the elementary events 1/2, 1/4, 1/8, ...

Define the discrete random variable $X : S \to \mathbb{R}$, by $X(s) = x$, where the first head appears on toss x. Then

$$p_X(x) = \mathrm{P}(X = x) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, 3, \ldots$$

and zero otherwise. For $x \geq 1$, let $k(x)$ be the largest integer not greater than x (also denoted by the "floor" symbol $\lfloor x \rfloor$), then

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i) = \sum_{i=1}^{k(x)} p_X(i) = 1 - \left(\frac{1}{2}\right)^{k(x)} \quad \text{[sum of a geometric series]}$$

and $F_X(x) = 0$ for $x < 1$.

Figure 5.2 displays the probability mass function (top) and cumulative distribution function (bottom). Note that the PMF is only non-zero at points that are elements of $\mathbb{X}$ and the CDF is defined for all real values of x, but is only continuous from the right. F_X is therefore a step-function.

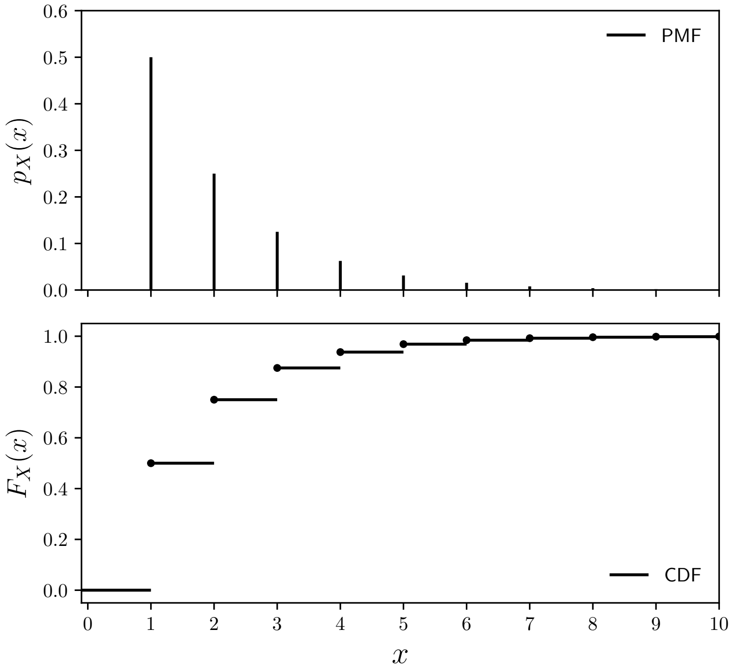


Figure 5.2: pmf $p_X(x) = \left(\frac{1}{2}\right)^x$, $x = 1, 2, \ldots$, and CDF $F_X(x) = 1 - \left(\frac{1}{2}\right)^{k(x)}$.

■

We can now see some connections between the numerical summaries and graphical displays we saw in earlier lectures and probability theory.

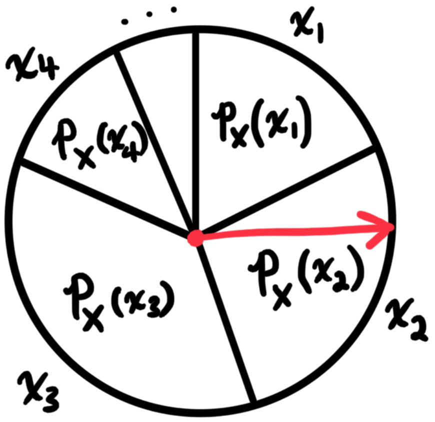
We can often think of a set of data $(x_1, x_2, \ldots, x_n)$ as n realizations of a random variable X (i.e. the outcomes of n copies of the experiment) or n random elements from a large population.

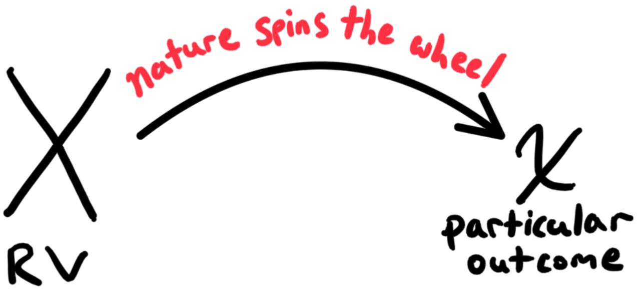
- Recall the frequency counts we considered for a set of data and their histogram. This can be seen as an *empirical estimate* for the PMF of the underlying population.
- Also recall the empirical cumulative distribution function. This too is an empirical estimate, but for the CDF of the underlying population.

5.3 Functions of a discrete random variable

Another visualization of random variables:

range of X:	x_1	x_2	x_3	...
probabilities:	$p_X(x_1)$	$p_X(x_2)$	$p_X(x_3)$	...

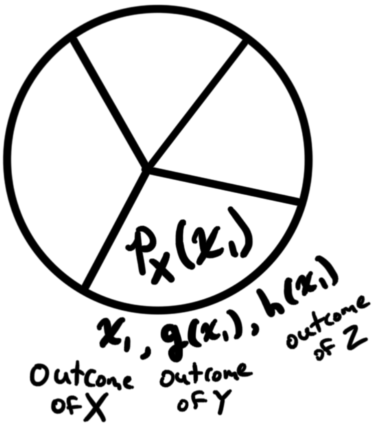




Expressions like X^2, $3X + 2$, $\exp(X)$, and $g(X)$ define new random variables. E.g. if X is a discrete random variable then $Y = X^2$ is another discrete random variable. When X takes the value x then Y takes the value x^2.

X is a RV. $Y = g(X)$, $Z = h(X)$ for some functions g and h.

prob	$p_X(x_1)$	$p_X(x_2)$	$p_X(x_3)$	...
X	x_1	x_2	x_3	...
Y	$g(x_1)$	$g(x_2)$	$g(x_3)$	...
Z	$h(x_1)$	$h(x_2)$	$h(x_3)$	...



The PMF of $Y = g(X)$ is found by grouping all the values in the range of x that correspond to the same value of Y.

$$p_Y(y) = \sum_{x \in \mathbb{X} : g(x) = y} p_X(x)$$

5.4 Mean and Variance

5.4.1 Expectation

The mean, or expectation, of a discrete random variable is the "average value" of X.

Definition 5.4.1. *The* **expected value**, *or* **mean** *of a discrete random variable X is defined to be*

$$\mathrm{E}_X(X) = \sum_{x \in \mathbb{X}} x p_X(x) = \sum_{x \in \mathbb{X}} x\, \mathrm{P}(X = x)$$

The expectation is a one-number summary of the distribution and is often just written $\mathrm{E}(X)$, $\mathrm{E}[X]$ or μ_X.

$\mathrm{E}(X)$ gives a weighted average of the possible values of the random variable X, with the weights given by the probability of each particular outcome.

We can use our illustration of the spinner to motivate the definition of expected value. Let's say when the spinner lands on x_i we win x_i points. We want to know the average amount we win "per spin". So we spin the wheel N times keeping track of our total winnings, and then divide the total by N.

Let $k(x_i)$ be the number of times the spinner landed on the outcome x_i (note that $\sum_i k_i = N$ since every spin has some outcome x_i). Then,

$$\begin{aligned}
\text{Average value per spin} &= \frac{k_1 x_1 + k_2 x_2 + \dots}{N} \\
&= x_1 \frac{k_1}{N} + x_2 \frac{k_2}{N} + \dots \\
&= x_1 p_X(x_1) + x_2 p_X(x_2) + \dots \\
&= \mathrm{E}(X),
\end{aligned}$$

where to go from the second to the third line we used the frequentist interpretation of probability: as $N \to \infty$ the fraction of times we obtain outcome x_i is just the probability that $X = x_i$, i.e. the pmf.

Examples:

1. If X is a random variable taking the integer value scored with a single roll of a fair die, then

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{x=1}^{6} x p_X(x) \\
&= 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{21}{6} = 3.5.
\end{aligned}
$$

2. If X is a score from a student answering a single multiple choice question with four options, with 3 marks awarded for a correct answer, -1 for a wrong answer and 0 for no answer, what is the expected value if they answer at random?

$$\mathrm{E}(X) = 3 \times \mathrm{P}_X(\text{Correct}) + (-1) \times \mathrm{P}_X(\text{Incorrect}) = 3 \times \frac{1}{4} - 1 \times \frac{3}{4} = 0.$$

Extension: Let $g : \mathbb{R} \to \mathbb{R}$ be a real-valued (measurable) function of interest of the random variable X; then we have the following result:
Theorem 5.5.

$$\mathrm{E}(g(X)) = \sum_{x\in\mathbb{X}} g(x) p_X(x)$$

Note: $E(X)$ is just a particular number. It is *not* a random variable.

Properties of Expectations

Let X be a random variable with pmf p_X. Let g and h be real-valued functions, $g, h : \mathbb{R} \to \mathbb{R}$, and let a and b be constants. Then

$$\mathrm{E}(ag(X) + bh(X)) = a\mathrm{E}(g(X)) + b\mathrm{E}(h(X))$$

Special Cases:

(i) For a linear function, $g(X) = aX + b$ for constants $a, b \in \mathbb{R}$, we have (from Theorem 5.5) that

$$
\begin{aligned}
\mathrm{E}(g(X)) &= \sum_x (ax+b) p_X(x) \\
&= a\sum_x x p_X(x) + b\sum_x p_X(x)
\end{aligned}
$$

and since $\sum_x x p_X(x) = \mathrm{E}(X)$ and $\sum_x p_X(x) = 1$ we have

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$$

(ii) Consider $g(x) = (x - \mathrm{E}(X))^2$. The expectation of this function wrt p_X gives a measure of spread or variability of the random variable X around its mean, called the **variance**.

Definition 5.5.1. *Let X be a random variable. The* **variance** *of X, denoted by σ^2 or σ_X^2 or $\mathrm{Var}_X(X)$ is defined by*

$$\mathrm{Var}_X(X) = \mathrm{E}_X\left[(X - \mathrm{E}_X(X))^2\right].$$

We can expand the expression $(X - \mathrm{E}(X))^2$ and exploit the linearity of expectation to get an alternative formula for the variance.

$$\begin{aligned} (X - \mathrm{E}(X))^2 &= X^2 - 2\mathrm{E}(X)X + \mathrm{E}(X)^2 \\ \implies \mathrm{Var}(X) &= \mathrm{E}\left[X^2 - 2\mathrm{E}(X)X + \mathrm{E}(X)^2\right] \\ &= \mathrm{E}(X^2) - 2\mathrm{E}(X)\mathrm{E}(X) + \mathrm{E}(X)^2 \end{aligned}$$

and hence

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2.$$

It is easy to show that the corresponding result is

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X), \qquad \forall a, b \in \mathbb{R}$$

Related to the variance is the standard deviation, which is defined as follows:

Definition 5.5.2. *The* **standard deviation** *of a random variable X, written $\mathrm{sd}_X(X)$ (or sometimes σ_X), is the square root of the variance.*

$$\mathrm{sd}_X(X) = \sqrt{\mathrm{Var}_X(X)}.$$

Lastly, we can define the skewness of a discrete random variable as follows:

Definition 5.5.3. *The* **skewness** *(γ_1) of a discrete random variable X is given by*

$$\gamma_1 = \frac{E_X[\{X - E_X(X)\}^3]}{\mathrm{sd}_X(X)^3}.$$

Example If X is a random variable taking the integer value scored with a single roll of a fair die, then

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}(X^2) - (E(X))^2 \\ &= \sum_{x=1}^{6} x^2 p_X(x) - 3.5^2 \\ &= 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \ldots + 6^2 \times \frac{1}{6} - 3.5^2 = 1.25. \end{aligned}$$

■

Example If X is a score from a student answering a single multiple choice question with four options, with 3 marks awarded for a correct answer, -1 for a wrong answer and 0 for no answer, what is the standard deviation if they answer at random?

$$\begin{aligned} &\mathrm{E}(X^2) = 3^2 \times \mathrm{P}_X(\text{Correct}) + (-1)^2 \times \mathrm{P}_X(\text{Incorrect}) = 9 \times \frac{1}{4} + 1 \times \frac{3}{4} = 3 \\ \Rightarrow\ &\mathrm{sd}(X) = \sqrt{3 - 0^2} = \sqrt{3}. \end{aligned}$$

■

Note We have met three important quantities for a random variable, defined through expectation – the mean μ, the variance σ^2 and the standard deviation σ.

Again we can see a duality with the corresponding numerical summaries for data which we met – the sample mean $\bar{x}$, the sample variance s^2 and the sample standard deviation s.

The duality is this: If we were to consider the data sample as the *population* and draw a random member from that sample as a *random variable*, this random variable would have CDF $F_n(x)$, the empirical CDF. The mean of the random variable $\mu = \bar{x}$, variance $\sigma^2 = s^2$ and standard deviation $\sigma = s$.

5.5.1 Sums of Random Variables

Let $X_1, X_2, \ldots, X_n$ be n random variables, perhaps with different distributions and not necessarily independent.

Let $S_n = \sum\limits_{i=1}^{n} X_i$ be the sum of those variables, and $\dfrac{S_n}{n}$ be their sample average. Both S_n and S_n/n are random variables themselves.

The mean of S_n and S_n/n are given by

$$\mathrm{E}(S_n) = \sum_{i=1}^{n} \mathrm{E}(X_i), \quad \mathrm{E}\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^{n} \mathrm{E}(X_i)}{n}.$$

If $X_1, X_2, \ldots, X_n$ are **independent** we can calculate the variance of S_n as well:

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i), \qquad \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^{n} \mathrm{Var}(X_i)}{n^2}.$$

So if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (**i.i.d.**) with $\mathrm{E}(X_i) = \mu_X$ and $\mathrm{Var}(X_i) = \sigma_X^2$ we get

$$\mathrm{E}\left(\frac{S_n}{n}\right) = \mu_X, \qquad \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma_X^2}{n}.$$

5.6 Some Important Discrete Random Variables

5.6.1 Bernoulli Distribution

Consider an experiment with only two possible outcomes, encoded as a random variable X taking value 1, with probability p, or 0, with probability $(1-p)$.

Example Tossing a coin, $X = 1$ for a head, $X = 0$ for tails, $p = \frac{1}{2}$. ■

Then we say $X \sim \text{Bernoulli}(p)$, where we must have $0 \leq p \leq 1$. The pmf is given by

$$\begin{aligned} p_X(x) &= \begin{cases} p & x = 1 \\ 1-p & x = 0 \\ 0 & \text{otherwise} \end{cases} \\ &= p^x(1-p)^{1-x} \quad \text{for } x \in \mathbb{X} = \{0, 1\} \end{aligned}$$

Note Using the formulae for mean and variance, it follows that

$$\mu \equiv \mathrm{E}(X) = p, \qquad \sigma^2 \equiv \mathrm{Var}(X) = p(1-p).$$

$\mathrm{E}(X) = \sum_{x \in \mathbb{X}} x p_X(x) = 0(1-p) + 1p = p$

$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \sum_{x \in \{0,1\}} (x-p)^2 p_X(x) = (0-p)^2(1-p) + (1-p)^2 p = p(1-p)$

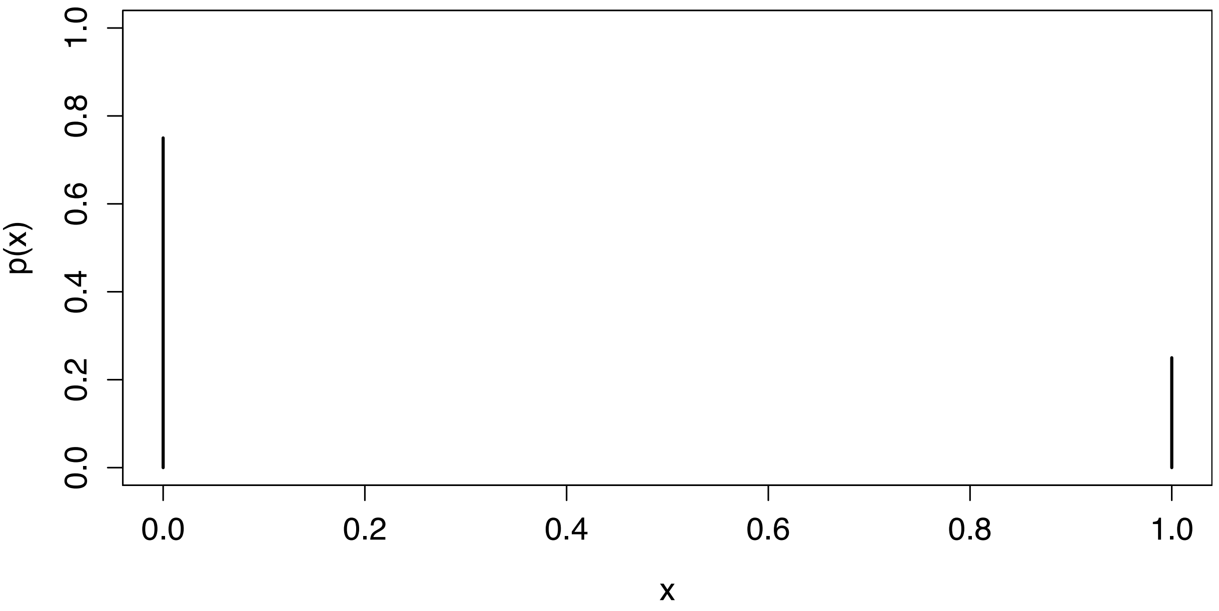


Figure 5.3: Example: pmf of Bernoulli(1/4).

5.6.2 Binomial Distribution

The Binomial Distribution is the pmf for the random variable X which equals the number of successes in n independent trials, each having a probability of success p.

In other words, consider n identical, independent Bernoulli(p) trials $X_1, \ldots, X_n$. Let $X = \sum_{i=1}^{n} X_i$ be the total number of 1s observed in the n trials.

Example Tossing a coin n times, X is the number of heads obtained, $p = \frac{1}{2}$. ■

Then X is a random variable taking values in $\mathbb{X} = \{0, 1, 2, \ldots, n\}$, and we say $X \sim$ Binomial(n, p).

From the Binomial Theorem we find the pmf to be

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \mathbb{X} = \{0, 1, 2, \ldots, n\}, \quad n \geq 1, \quad 0 \leq p \leq 1.$$

Notes

- To calculate the Binomial pmf we recall that the binomial coefficient is defined as $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ and the factorial function is $n! = 1 \cdot 2 \cdot 3 \cdots n$, (with $0! = 1$.)
- It can be shown, either directly from the pmf or from the results for sums of random variables, that the mean and variance are

$$\mu \equiv \mathrm{E}(X) = np, \qquad \sigma^2 \equiv \mathrm{Var}(X) = np(1-p).$$

- The skewness is given by

$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}.$$

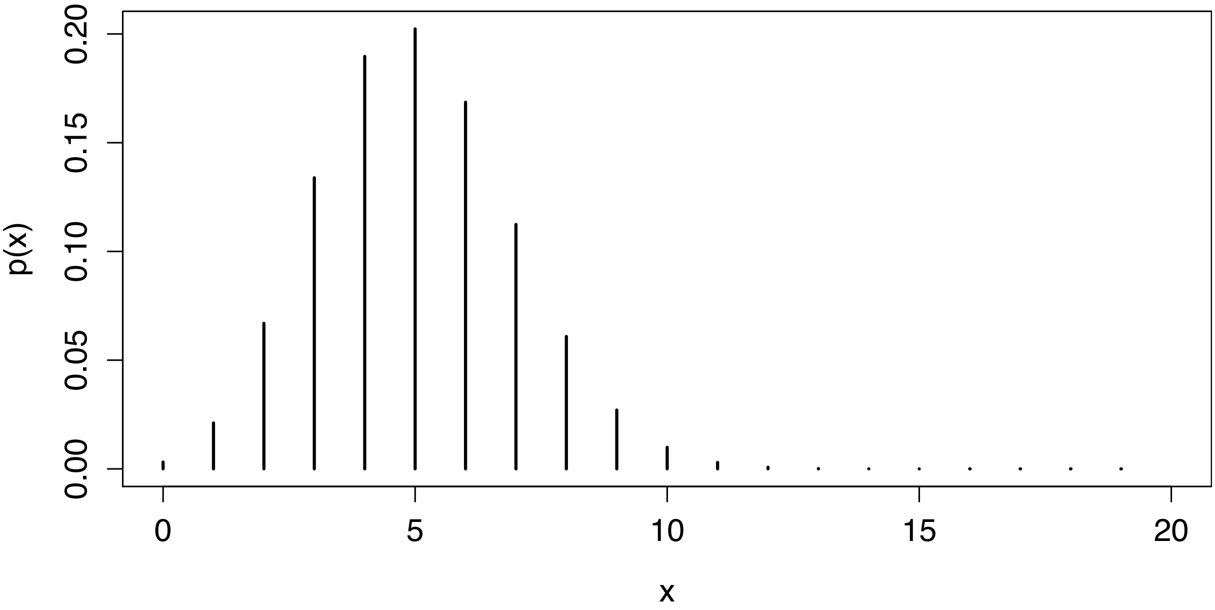


Figure 5.4: Example: pmf of Binomial$(20, 1/4)$.

Derivation of Binomial pmf:

We can write the outcome of n trials as a sequence of n 1's and 0's, where 1 represents success and 0 failure. E.g. for $n = 3$ the possible outcomes with two successes are 110, 101, and 011.

Each such sequence with k successes and $n - k$ failures occurs with probability $p^k(1-p)^{n-k}$ since each trial is independent. Therefore, the probability of getting a sequence with k successes in n trials is

$$p_X(k) = \begin{pmatrix} \text{\# of sequences of} \\ \text{length } n \text{ with } k \text{ 1s} \\ \text{and } n-k \text{ 0s} \end{pmatrix} p^k(1-p)^{n-k} = \binom{n}{k} p^k(1-p)^{n-k}.$$

Another way to picture it:
With p = probability of success let $q = 1 - p$ = probability of failure and consider

$$(p+q)^n = \underbrace{(p+q)(p+q)\cdots(p+q)}_{n \text{ factors}}.$$

Imagine multiplying out all n factors. You will get one term for each way of choosing either a p or a q from each factor, i.e. for each possible sequence of 1s and 0s of length n. Collect terms and look at the $p^k q^{n-k}$ term. Its coefficient is $\binom{n}{k}$. This also shows that the Binomial pmf is properly normalized since $(p+q)^n = 1^n = 1$.

Example Suppose that 10 users are authorised to use a particular computer system, and that the system collapses if 7 or more users attempt to log on simultaneously. Suppose that each user has the same probability $p = 0.2$ of wishing to log on in each hour.

Question: What is the probability that the system will crash during a given hour?

Solution

The probability that exactly x users will want to log on in any hour is given by Binomial(n, p) = Binomial$(10, 0.2)$.

Hence the probability of 7 or more users wishing to log on in any hour is

$$\begin{aligned} & p_X(7) + p_X(8) + p_X(9) + p_X(10) \\ &= \binom{10}{7} 0.2^7 0.8^3 + \ldots + \binom{10}{10} 0.2^{10} 0.8^0 \\ &= 0.00086. \end{aligned}$$

- A manufacturing plant produces chips with a defect rate of 10%. The quality control procedure consists of checking samples of size 50. Then the distribution of the number of defectives is expected to be Binomial$(50, 0.1)$.
- When transmitting binary digits through a communication channel, the number of digits received correctly out of n transmitted digits, can be modelled by a Binomial(n, p), where p is the probability that a digit is transmitted correctly.

Note The independence condition is necessary for these models to be reasonable. ■

5.6.3 Geometric Distribution

Consider a potentially infinite sequence of independent trials, each of which is either a failure or success. The probability of success on an individual trial is p. The trial number on which we have the first success is called a geometric random variable.

Concisely, we have a potentially infinite sequence of independent Bernoulli(p) random variables $X_1, X_2, \ldots$. We define a quantity X by

$$X = \min\{i \mid X_i = 1\}$$

to be the index of the first Bernoulli trial to result in a 1.

Example Tossing a coin, X is the number of tosses until the first head is obtained, $p = \frac{1}{2}$. ■

Then X is a random variable taking values in $\mathbb{X} = \mathbb{Z}^+ = \{1, 2, \ldots\}$, and we say $X \sim$ Geometric(p).

Clearly the pmf is given by

$$p_X(x) = p(1-p)^{x-1}, \qquad x \in \mathbb{X} = \{1, 2, \ldots\}, \quad 0 \le p \le 1.$$

Notes

- The mean and variance are

$$\mu \equiv \mathrm{E}(X) = \frac{1}{p}, \qquad \sigma^2 \equiv \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

- The skewness is given by

$$\gamma_1 = \frac{2-p}{\sqrt{1-p}},$$

  and so is always positive.

Useful trick for deriving the mean and variance:

By definition, the mean of a geometric RV is $\mathrm{E}(X) = \sum\limits_{k=1}^{\infty} kp(1-p)^{k-1}$.

Now define a function f by

$$f(q) = \sum_{k=1}^{\infty} pq^k,$$

and notice what happens if you differentiate $f(q)$ with respect to q: $df/dq = \sum_{k=1}^{\infty} kpq^{k-1}$. Therefore, $\mathrm{E}(X)$ is just the derivative of $f(q)$ evaluated at $q = 1-p$. But the sum in $f(q)$ is just the sum of a geometric series with first term pq and ratio between terms q. So $f(q) = pq/(1-q)$. Take the derivative of this with respect to q, set $q = 1-p$, and find that $\mathrm{E}(X) = 1/p$.

The same trick will work to get the variance: $\mathrm{E}(X^2) = \frac{d}{dq}\left(q\frac{df}{dq}\right)$, evaluated at $q = 1-p$.

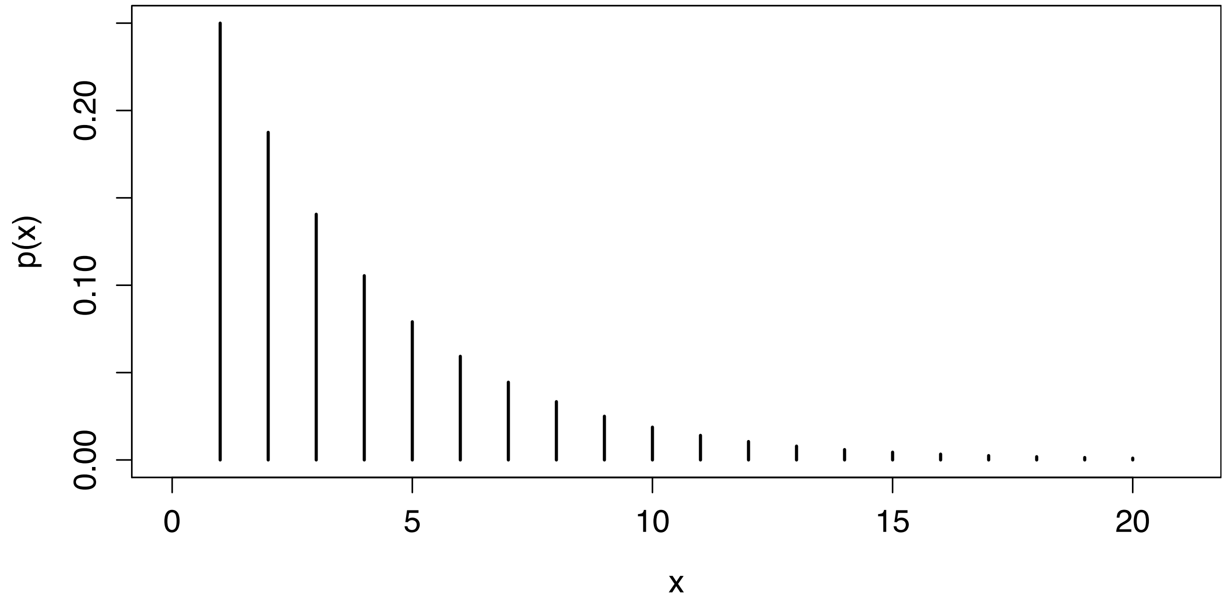


Figure 5.5: Example: pmf of Geometric$(1/4)$.

Alternative Formulation

If $X \sim$ Geometric(p), let us consider $Y = X - 1$.

Then Y is a random variable taking values in $\mathbb{N} = \{0, 1, 2, \dots\}$, and corresponds to the number of independent Bernoulli(p) trials *before* we obtain our first 1. (Some texts refer to *this* as the Geometric distribution.)

Note we have pmf

$$p_Y(y) = p(1-p)^y, \qquad y = 0, 1, 2, \dots,$$

and the mean becomes

$$\mu_Y \equiv \mathrm{E}_Y(Y) = \frac{1-p}{p}.$$

while the variance and skewness are unaffected by the shift.

Example Suppose people have problems logging onto a particular website once every 5 attempts, on average.

1. Assuming the attempts are independent, what is the probability that an individual will not succeed until the 4th?

$$p = \frac{4}{5} = 0.8. \quad p_X(4) = (1-p)^3 p = (0.2)^3\, 0.8 = 0.0064.$$

2. On average, how many trials must one make until succeeding?

$$\text{Mean} = \frac{1}{p} = \frac{5}{4} = 1.25.$$

3. What's the probability that the first successful attempt is the 7th or later?

$$\begin{aligned} p_X(7) + p_X(8) + p_X(9) + \ldots &= p(1-p)^6 + p(1-p)^7 + p(1-p)^8 + \ldots \\ &= \frac{p(1-p)^6}{1-(1-p)} \\ &= (1-p)^6 = 0.2^6. \end{aligned}$$

■

Example Again suppose that 10 users are authorised to use a particular computer system, and that the system collapses if 7 or more users attempt to log on simultaneously. Suppose that each user has the same probability $p = 0.2$ of wishing to log on in each hour.

Using the Binomial distribution we found the probability that the system will crash in any given hour to be 0.00086.

Using the Geometric distribution formulae, we are able to answer questions such as: On average, after how many hours will the system crash?

Mean $= \frac{1}{p} = \frac{1}{0.00086} = 1163$ hours. ■

Example A dictator, keen to maximise the ratio of males to females in his country (so he could build up his all male army) ordered that each couple should keep having children until a boy was born and then stop.

Calculate the number expected number of boys that a couple will have, and the expected number of girls, given that P(boy)=½.

Assume for simplicity that each couple can have arbitrarily many children (although this is not necessary to get the following results). Then since each couple stops when 1 boy is born, the expected number of boys per couple is 1.

On the other hand, if Y is the number of girls given birth to by a couple, Y clearly follows the alternative formulation for the Geometric(½) distribution.

So the expected number of girls for a couple is $\frac{1-\frac{1}{2}}{\frac{1}{2}} = 1$. ■

5.6.4 Poisson Distribution

Let X be a random variable on $\mathbb{N} = \{0, 1, 2, \ldots\}$ with pmf

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x \in \mathbb{X} = \{0, 1, 2, \ldots\}, \quad \lambda > 0.$$

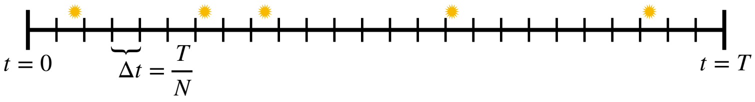
Then X is said to follow a Poisson distribution with *rate* parameter λ and we write $X \sim \text{Poisson}(\lambda)$.

Notes

- Poisson random variables almost always show up where there is an average underlying rate at which events occur per unit time or space. That is, we can imagine dividing a continuous range into very small segments and there is some small probability for an event to occur in each segment. The Poisson distribution is the pmf for the total number of events that occur. For example,
  - the number of minor car crashes per day in the U.K.;
  - the number of potholes in each mile of road;
  - the number of jobs which arrive at a database server per hour;
  - the number of particles emitted by a radioactive substance in a given time.

Derivation of Poisson pmf

Discrete events (e.g. car crashes) occur in a region we express as a 1-d line from 0 to T (but we could imagine any region capable of being divided up into small pieces). Divide region into N segments, each of size $\Delta T = T/N$.



The key is to realize that, as $\Delta T \to 0$, the probability of an event occurring in a segment is proportional to the size of the segment ΔT. Call the constant of proportionality r. It has units of probability per unit of T (e.g. probability of a pothole per foot of road).

The number of events X over the whole region T is then exactly a binomial random variable with N trials and probability $r\Delta T$ of success on each trial. Therefore, the probability of k events ("k successes") is given by the binomial pmf,

$$p_X(k) = \frac{N!}{k!(N-k)!}(r\Delta T)^k(1 - r\Delta T)^{N-k}.$$

Replace ΔT with T/N and then take the limit as $N \to \infty$ to find,

$$p_X(k) = \frac{N(N-1)(N-2)\cdots(N-k+1)}{N^k}\left(1 - \frac{rT}{N}\right)^N \frac{(rT)^k}{k!}\left(1 - \frac{rT}{N}\right)^{-k}.$$

The first fraction is the product of k factors $\left(\frac{N}{N}\right)\left(\frac{N-1}{N}\right)\cdots\left(\frac{N-k+1}{N}\right)$, each of which goes to 1 as $N \to \infty$. For the next term use $\left(1 + \frac{x}{N}\right)^N \to \exp(x)$ and note that the last term goes to 1 as $N \to \infty$. The result is the Poisson pmf with $\lambda = rT$. The product rT is intuitively interpreted as the average number of events over the whole region.

- An interesting property of the Poisson distribution is that it has equal mean and variance, namely
$$\mu \equiv \mathrm{E}(X) = \lambda, \qquad \sigma^2 \equiv \mathrm{Var}(X) = \lambda.$$
- The skewness is given by
$$\gamma_1 = \frac{1}{\sqrt{\lambda}},$$
so is always positive but decreasing as λ increases.

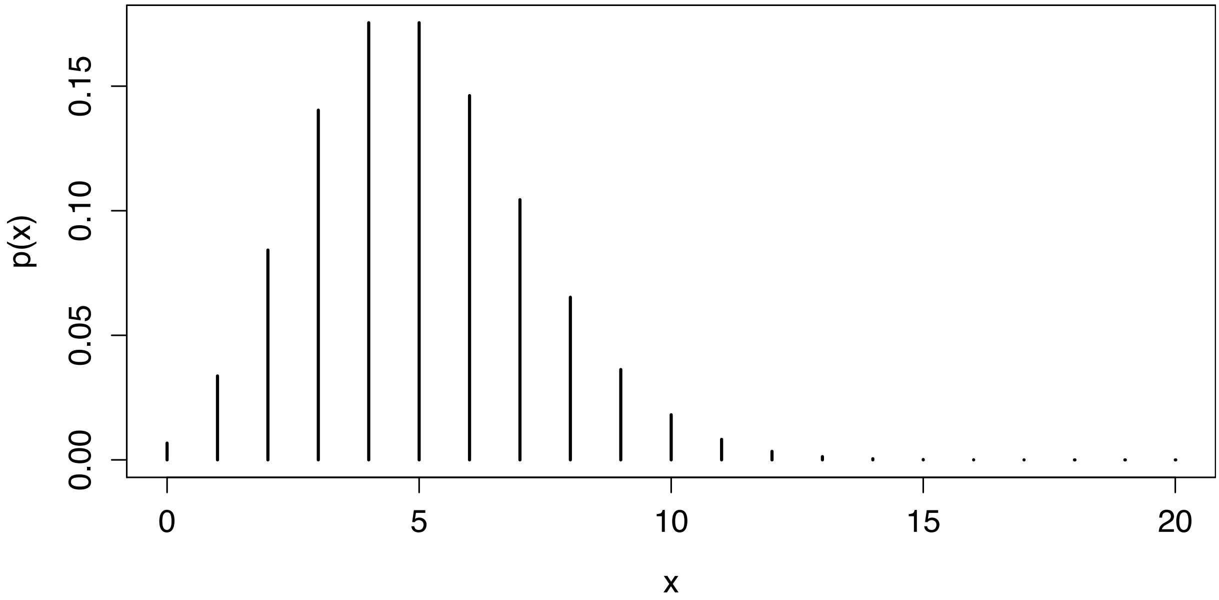


Figure 5.6: Example: pmf of Poisson(5).

Notice the similarity between the pmf plots for Binomial(20, 1/4) and Poisson(5) (Figures 5.4 and 5.6).

It can be shown that for Binomial(n, p), when p is small and n is large, this distribution can be well approximated by the Poisson distribution with rate parameter np, Poison(np).

The value of p in the above is not small, we would typically prefer $p < 0.1$ for the approximation to be useful.

The usefulness of this approximation is in using probability tables; tabulating a single Poisson(λ) distribution encompasses an infinite number of possible corresponding Binomial distributions, Binomial$(n, \frac{\lambda}{n})$.

Example A manufacturer produces VLSI chips, of which 1% are defective. Find the probability that in a box of 100 chips none are defective.

We want $p_X(0)$ from Binomial(100,0.01). Since n is large and p is small, we can approximate this distribution by Poisson$(100 \times 0.01) \equiv$ Poisson(1).

Then $p_X(0) \approx \dfrac{e^{-1}1^0}{0!} = 0.3679.$ ■

Example The number of particles emitted by a radioactive substance which reached a Geiger counter was measured for 2608 time intervals, each of length 7.5 seconds.

The (real) data are given in the table below:

x	0	1	2	3	4	5	6	7	8	9	$\geq$10
n_x	57	203	383	525	532	408	273	139	45	27	16

Do these data correspond to 2608 independent observations of an identical Poisson random variable?

The total number of particles, $\sum_x xn_x$, is 10,094, and the total number of intervals observed, $n = \sum_x n_x$, is 2608, so that the average number reaching the counter in an interval is $\frac{10094}{2608} = 3.870$.

Since the mean of Poisson(λ) is λ, we can try setting $\lambda = 3.87$ and see how well this fits the data.

For example, considering the case $x = 0$, for a single experiment interval the probability of observing 0 particles would be $p_X(0) = \frac{e^{-3.87}3.87^0}{0!} = 0.02086$. So over $n = 2608$ repetitions, our (Binomial) expectation of the number of 0 counts would be $n \times p_X(0) = 54.4$.

Similarly for $x = 1, 2, \ldots,$ we obtain the following table of expected values from the Poisson(3.87) model:

x	0	1	2	3	4	5	6	7	8	9	$\geq$10
$O(n_x)$	57	203	383	525	532	408	273	139	45	27	16
$E(n_x)$	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1

(O=Observed, E=Expected).

The two sets of numbers appear sufficiently close to suggest the Poisson approximation is a good one. Later, when we come to look at *hypothesis testing*, we will see how to make such judgements quantitatively. ■

5.6.5 Discrete Uniform Distribution

The discrete uniform random variable describes the situation when we choose from among n possibilities completely at random.

Let X be a random variable on $\mathbb{X} = \{1, 2, \ldots, n\}$ with pmf

$$p_X(x) = \frac{1}{n}, \qquad x \in \mathbb{X} = \{1, 2, \ldots, n\}.$$

Then X is said to follow a discrete uniform distribution and we write $X \sim \mathrm{U}(\{1, 2, \ldots, n\})$.

Note The mean and variance are

$$\mu \equiv \mathrm{E}(X) = \frac{n+1}{2}, \qquad \sigma^2 \equiv \mathrm{Var}(X) = \frac{n^2-1}{12}.$$

and the skewness is clearly zero.