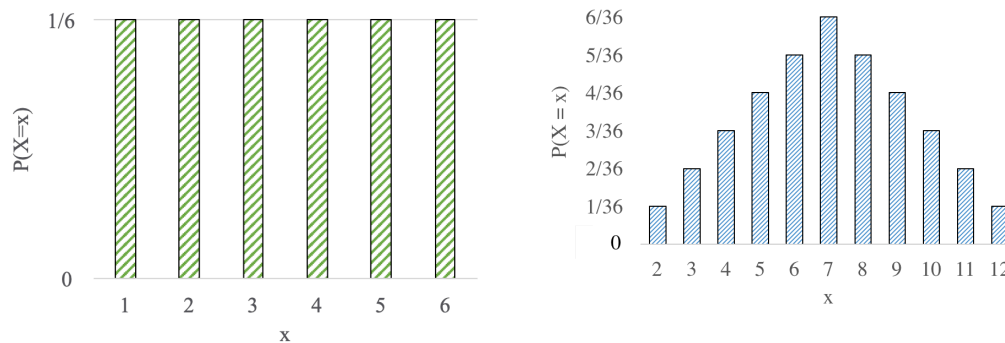


# Probability Mass Functions

For a random variable, the most important thing to know is: how likely is each outcome? For a discrete random variable, this information is called the "**probability mass function**". The probability mass function (PMF) provides the "mass" (i.e. amount) of "probability" for each possible assignment of the random variable.

Formally, the probability mass function is a mapping between the values that the random variable could take on and the probability of the random variable taking on said value. In mathematics, we call these associations functions. There are many different ways of representing functions: you can write an equation, you can make a graph, you can even store many samples in a list. Let's start by looking at PMFs as graphs where the  $x$ -axis is the values that the random variable could take on and the  $y$ -axis is the probability of the random variable taking on said value.

In the following example, on the left we show a PMF, as a graph, for the random variable:  $X$  = the value of a six-sided die roll. On the right we show a contrasting example of a PMF for the random variable  $X$  = value of the sum of two dice rolls:



Left: the PMF of a single six-sided die roll. Right: the PMF of the sum of two dice rolls.

The [sum of two dice example](#) in the equally likely probability section. Again, the information that is provided in these graphs is the likelihood of a random variable taking on different values. In the graph on the right, the value "6" on the  $x$ -axis is associated with the probability  $\frac{5}{36}$  on the  $y$ -axis. This  $x$ -axis refers to the event "the sum of two dice is 6" or  $Y = 6$ . The  $y$ -axis tells us that the probability of that event is  $\frac{5}{36}$ . In full:  $P(Y = 6) = \frac{5}{36}$ . The value "2" is associated with " $\frac{1}{36}$ " which tells us that,  $P(Y = 2) = \frac{1}{36}$ , the probability that two dice sum to 2 is  $\frac{1}{36}$ . There is no value associated with "1" because the sum of two dice can not be 1. If you find this notation confusing, revisit the [random variables](#) section.

Here is the exact same information in equation form:

$$P(X = x) = \frac{1}{6} \quad \text{if } 1 \leq x \leq 6 \qquad P(Y = y) = \begin{cases} \frac{(y-1)}{36} & \text{if } 1 \leq y \leq 7 \\ \frac{(13-y)}{36} & \text{if } 8 \leq y \leq 12 \end{cases}$$

As a final example, here is the PMF for  $Y$ , the sum of two dice, in Python code:

```
def pmf_sum_two_dice(y):  
    # Returns the probability that the sum of two dice is y  
    if y < 2 or y > 12:  
        return 0  
    if y <= 7:  
        return (y-1) / 36  
    else:  
        return (13-y) / 36
```

## Notation

You may feel that  $P(Y = y)$  is redundant notation. In probability research papers and higher-level work, mathematicians often use the shorthand  $P(y)$  to mean  $P(Y = y)$ . This shorthand assumes that the lowercase value (e.g.  $y$ ) has a capital letter counterpart (e.g.  $Y$ ) that represents a random variable even though it's not written explicitly. In this book, we will often use the full form of the event  $P(Y = y)$ , but we will occasionally use the shorthand  $P(y)$ .

## Probabilities Must Sum to 1

For a variable (call it  $X$ ) to be a proper random variable it must be the case that if you summed up the values of  $P(X = k)$  for all possible values  $k$  that  $X$  can take on, the result must be 1:

$$\sum_k P(X = k) = 1$$

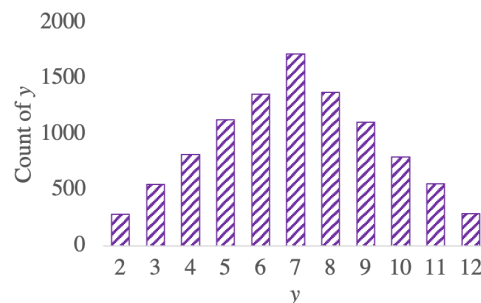
For further understanding, let's derive why this is the case. A random variable taking on a value is an event (for example  $X = 2$ ). Each of those events is mutually exclusive because a random variable will take on exactly one value. Those mutually exclusive cases define an entire sample space. Why? Because  $X$  must take on *some* value.

## Data to Histograms to Probability Mass Functions

One surprising way to store a likelihood function (recall that a PMF is the name of the likelihood function for *discrete* random variables) is simply a list of data. We simulated summing two die 10,000 times to make this example dataset:

[8, 4, 9, 7, 7, 7, 7, 5, 6, 8, 11, 5, 7, 7, 7, 6, 7, 8, 8, 9, 9, 4, 6, 7, 10, 12, 6, 7, 8, 9, 3, 7, 4, 9, 2, 8, 5, 8, 9, 6, 8, 7, 10, 7, 6, 7, 7, 5, 4, 6, 9, 5, 7, 4, 2, 11, 10, 11, 8, 4, 11, 9, 7, 10, 12, 4, 8, 5, 11, 5, 3, 9, 7, 5, 5, 5, 3, 8, 6, 11, 11, 2, 7, 7, 6, 5, 4, 6, 3, 8, 5, 8, 7, 6, 9, 4, 3, 7, 6, 6, 6, 5, 6, 10, 5, 9, 9, 8, 8, 7, 4, 8, 4, 9, 8, 5, 10, 10, 9, 7, 9, 7, 7, 10, 4, 7, 8, 4, 7, 8, 9, 11, 7, 9, 10, 10, 2, 7, 9, 4, 8, 8, 12, 9, 5, 11, 10, 7, 6, 4, 8, 9, 9, 6, 5, 6, 5, 6, 11, 7, 3, 10, 7, 3, 7, 7, 10, 3, 6, 8, 6, 8, 5, 10, 2, 7, 4, 8, 11, 8, 3, 4, 3, 8, 8, 6, 6, 12, 11, 10, 10, 10, 8, 4, 8, 4, 4, 6, 6, 7, 8]

Note that this data, on its own, represents an approximation for the probability mass function. If you wanted to approximate  $P(Y = 3)$  you could simply count the number of times that "3" occurs in your data. This is an approximation based on the [definition of probability](#). Here is the full [histogram](#) of the data, a count of times each value occurs:



A normalized histogram (where each value is divided by the length of your data list) is an approximation of the PMF. For a dataset of discrete numbers, a histogram shows the count of each value (in this case  $y$ ). By the definition of probability, if you divide this count by the number of experiments run, you arrive at an approximation of the probability of the event  $P(Y = y)$ . In our example, we have 10,000 elements in our dataset. The count of times that 3 occurs is 552. Note that:

$$\frac{\text{count}(Y = 3)}{n} = \frac{552}{10000} = 0.0552$$
$$P(Y = 3) = \frac{4}{36} = 0.0555$$

In this case, since we ran 10,000 trials, the histogram is a very good approximation of the PMF. We use the sum of dice as an example because it is easy to understand. Datasets in the real world often represent more exciting events.