# 2 Discrete Random Variables

## 2.1 Random Variables

A *random variable* is a real-valued function on a sample space[1]. A random variable generally is a quantity we wish to measure; the output of the random variable depends on which element of the sample space is chosen. In this section, we will look at discrete random variables.

A *discrete* random variable $Y$ is a random variable which can only take on a finite or countable set of distinct values.

Here are some examples of discrete random variables:

1. The number of voters in Rhode Island who prefer Hillary Clinton.
2. The number of defective lightbulbs out of a shipment of 1000 lightbulbs.
3. The number of times I play a slot machine in Las Vegas until I win.

We will use the following simple example to illustrate features of discrete random variables.

**Example.** Let $\mathcal{S}$ be the sample space representing the flip of two fair coins. Let $Y$ be the number of heads flipped. Then $Y$ is a discrete random variable, since it can only have the values 0, 1, or 2. We can illustrate it graphically below.

|   | *H* | *T* |
|---|---|---|
| *H* | 2 | 1 |
| *T* | 1 | 0 |

Uppercase letters, such as $Y$, are used to designate random variables. We use lowercase letters, such as $y$, to designate a value that a random variable can take. The expression $(Y = y)$ is shorthand for the set of all points in our sample space $\mathcal{S}$ for which the random variable $Y$ outputs the value $y$. Since $(Y = y)$ is a subset of $\mathcal{S}$, it is an event in our sample space. In the two-coin-toss problem, for example, the possible values of $Y$ are 0, 1, and 2, so we have:

- $(Y = 0) = \{(T, T)\}$
- $(Y = 1) = \{(H, T), (T, H)\}$
- $(Y = 2) = \{(H, H)\}$

[1] It's not really a variable at all, but we are stuck with the terminology.

Since $(Y = y)$ is an event in our sample space, we can talk about its probabiltiy, i.e. $\mathbb{P}(Y = y)$. In fact, the point of random variables is to do just this!

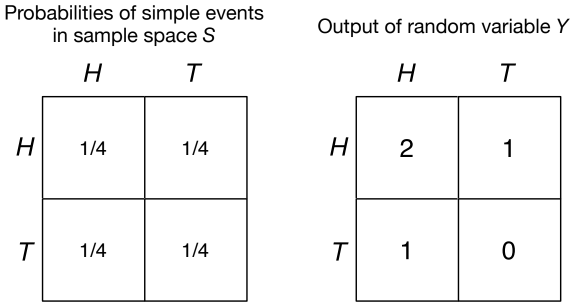
*Probability of a discrete random variable*

---

The probability that a discrete random variable $Y$ takes the value $y$, denoted $\mathbb{P}(Y = y)$, or $p(y)$ for short, is the probability of the event $(Y = y)$, the set of all points in the sample space $\mathcal{S}$ which output the value $y$.

$\mathbb{P}(Y = y)$ is the sum of the probabilities of all the simple events in $\mathcal{S}$ which are assigned the value $y$ by the random variable $Y$.

Back to our two-coin-toss problem, let's look at the probabilties of the random variable $Y$. For convenience, here is a picture of the sample space probabilities next to a graphical representation of $Y$. Since we are using the discrete uniform distribution for coin tosses, each simple event in our sample space has probability 1/2.



Probabilities of simple events in sample space *S*

| | *H* | *T* |
|---|---|---|
| *H* | 1/4 | 1/4 |
| *T* | 1/4 | 1/4 |

Output of random variable *Y*

| | *H* | *T* |
|---|---|---|
| *H* | 2 | 1 |
| *T* | 1 | 0 |

Using the rule above, we can compute the following probabilties for $Y$ by adding up the probabilities of the underlying simple events. In this case, since we are using the discrete uniform distribution, we could also just count the number of simple events which lead to each output of $Y$ and divide by 4, the size of the sample space.

- $\mathbb{P}(Y = 0) = 1/4$
- $\mathbb{P}(Y = 1) = 1/4 + 1/4 = 1/2$
- $\mathbb{P}(Y = 2) = 1/4$

*Probability mass function*

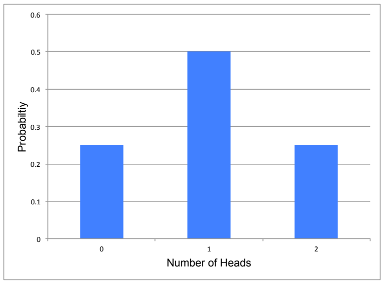
---

The *probablity mass function (pmf)* is a function which gives the probabiltiy that a discrete random variable $Y$ is exactly equal to a value $y$. The pmf can be represented as a function, table, or graph which gives the values $p(y) = \mathbb{P}(Y = y)$ for all possible values $y$ which $Y$ can take.

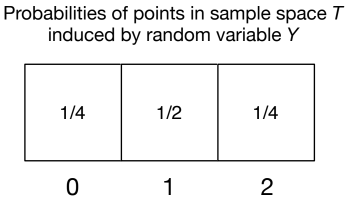
In the two-coin-flip example, we can represent the pmf of $Y$ in a table:

| $y$ | $p(y)$ |
|---|---|
| 0 | 1/4 |
| 1 | 1/2 |
| 2 | 1/4 |

We can also represent the pmf graphically as a histogram[2].



A discrete random variable induces a probability distribution on the sample space of all possible values the random variable can take. This is a different sample space from the original sample space. Back to our two-coin-flip example, the random variable $Y$ induces a probability distribution on a new sample space $\mathcal{T} = \{0, 1, 2\}$. The probabilities of the sample points in $\mathcal{T}$ are the probabilties $p(y)$ for $y = 0, 1, 2$. We can illustrate this new sample space in a picture.


Probabilities of points in sample space T
induced by random variable Y
1/4
1/2
1/4
0
1
2

Often (as we shall see), we care much more about the sample space induced by a random variable than the underlying sample space. Since a discrete random variable induces a probability distribution, the following must be true.

For any discrete random variable $Y$:

$$0 \leq p(y) \leq 1 \text{ for all } y$$
$$\sum_{\text{all } y} p(y) = 1$$

[2] I will undoubtedly lose some of my math street cred if I admit to using Microsoft Excel for these histograms, but in some cases it really is the easiest tool to use.

where $p(y) = \mathbb{P}(Y = y)$. Since we have a discrete sample space, the sum is finite or countable.

Let's look at two more examples, this time involving the rolls of two six-sided dice.

**Example.** Let $\mathcal{S}$ be the sample space representing the rolls of two six-sided dice. Consider the following two random variables:

1. $X$ = the sum of the two dice
2. $Y$ = the larger of the two die rolls (if they are the same, then it's just equal to both die rolls)

Let's look at these random variables graphically.

Two random variables on the sample space of two die rolls

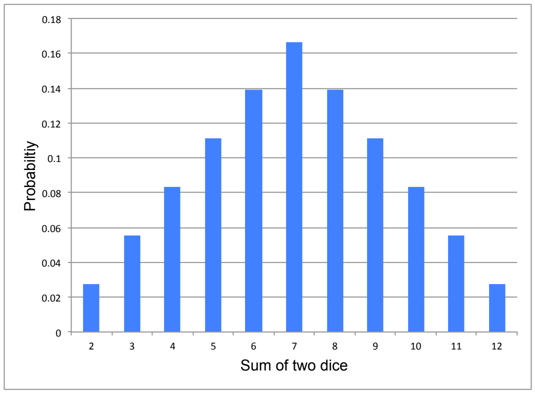
|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |

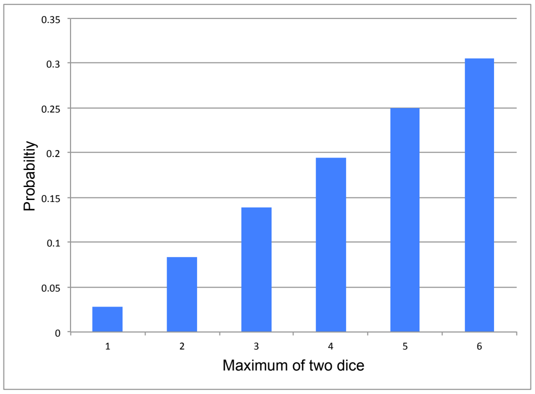
$X$

|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | 2 | 2 | 3 | 4 | 5 | 6 |
| 3 | 3 | 3 | 3 | 4 | 5 | 6 |
| 4 | 4 | 4 | 4 | 4 | 5 | 6 |
| 5 | 5 | 5 | 5 | 5 | 5 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 |

$Y$

The random variable $X$ induces a probability distribution on the set of integers $\{2, 3, 4, \ldots, 12\}$, and the random variable $Y$ induces a probability distribution on the set of integers $\{1, 2, 3, 4, 5, 6\}$. Let's look at the pmfs of both random variables using histograms.





Both of these distributions are nonuniform, even though the underlying distribution of the two dice is uniform. The first distribution, that of the random variable $X$, is a familiar one to afficionados of board games such at Settlers of Catan and Monopoly. We can also write the pmfs in table form. For the random variable $Y$, we have:

| $y$ | $p(y)$ |
|---|---|
| 1 | 1/36 |
| 2 | 3/36 |
| 3 | 5/36 |
| 4 | 7/36 |
| 5 | 9/36 |
| 6 | 11/36 |

The pmf for $X$ can be expressed similarly.

## 2.2 Expected Value

Given a discrete random variable, we can definite it's mean, or expected value.

*Expected value of a random variable*

For a discrete random variable $Y$ with probability function $p(y)$, the *expected value* or *mean* is defined to be

$$\mathbb{E}(Y) = \sum_{\text{all } y} y\, p(y)$$

where the sum is taken over all possible values $y$ can take. We can think of the expected value as a weighted average of the values of $Y$ with each possible output $y$ weighted by its probability $p(y)$. The expected value is sometimes written as $\mu$ (for mean)

Here is one interpretation of the expected value of a random variable. Think of a random variable $Y$ as an observation from an experiment. Suppose we perform the experiment $n$ times, and observe $n$ values of $Y$, which we shall designate $y_1, y_2, \ldots, y_n$. Then for large $n$,

$$\frac{y_1 + y_2 + \cdots + y_n}{n} \approx \mathbb{E}(Y)$$

where the approximation "gets better" as $n$ gets larger, i.e. as we perform more experiments. The quantity on the left hand side is known as the *empirical mean* or *sample mean* and looks like what we likely learned in high school (add a bunch of stuff up and divide by the number of things). The expected value is, in a sense, the limit of the empirical mean as the sample size approaches infinity. We will make this more precise later in the course, but this is a good concept to keep in mind.

**Example.** Let $X$ represent the roll of a standard, fair six-sided die. (In this case, the underlying sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ with the discrete uniform distribution, and the random variable $X$ is the same as the sample space element selected.) Then the expected value of $X$ is:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=1}^{6} x\mathbb{P}(X = x)\\ &= \sum_{x=1}^{6} x\frac{1}{6}\\ &= \frac{1}{6}\sum_{x=1}^{6} x\\ &= \frac{21}{6}\\ &= 3.5\end{aligned}$$

where used the fact from the discrete uniform distribution that $\mathbb{P}(X = x) = 1/6$ for all $x$. Note that the expected value of 3.5 is not a possible value of $X$, i.e. we cannot roll a 3.5 on a single die. Given our "long term average" interpretation, this is saying that we expect the empirical average to approach 3.5 as the number of rolls increases, not that a 3.5 is the most likely die roll.

**Example.** Let $Y$ be the random variable above representing the maximum of two dice. What is the expected value of $Y$.

To find the expected value, we do a weighted average using the probabities in the table above.

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y=1}^{6} y\mathbb{P}(Y = y)\\ &= 1\cdot\frac{1}{36} + 2\cdot\frac{3}{36} + 3\cdot\frac{5}{36} + 4\cdot\frac{7}{36} + 5\cdot\frac{9}{36} + 6\cdot\frac{11}{36}\\ &= \frac{1+6+15+28+45+66}{36}\\ &= \frac{161}{36} \approx 4.47\end{aligned}$$

## 2.3 Properties of Expectation

We will discuss several properties of the expected value. The first and and one of the most important is the *linearity of expectation.*

*Linearity of expectation*

Let $X$ and $Y$ be two random variables[3], and let $a$ and $b$ be constants. Then

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

This is called *linearity* in reference to linear algebra, i.e. we can separate addition and pull out constants. This holds whether or not $X$ and $Y$ are independent.

As a corollary of this, if we have random variables $X_1, X_2, \ldots, X_n$, then

$$\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)$$

Linearity of expectation is a really nice property since it does not require the random variables to be independent. Let's do a problem to illustrate the usefulness of linearity of expectation.

**Example.** One evening, $n$ customers dine at a restaurant. Each gives their hat to a hat-check person at a restaurant. (These are fashionable diners!) After dinner, the hat-check person gives the hats back to the customers in a random order, i.e. each customer receives one of the hats uniformly at random. What is the expected number of customers that get their own hat back?

Let $X$ be the number of customers who get their own hat back. Then $X$ is a discrete random variable taking values $0, 1, \ldots, n$. By the definition of expected value:

$$\mathbb{E}(X) = \sum_{i=1}^{n} i \, \mathbb{P}(X = i)$$

At this point, we have a mess! We have to compute $\mathbb{P}(X = i)$ for all $i$, which would involve a sophisticated combinatorial argument, as well as considerable time and mental energy. Luckily for us, there is another way.

We will use the method of indicator random variables, together with linearity of expectation, to solve this problem. An *indicator random variable* is a random variable $I$ which only takes the values 0 and 1. It is used to indicate whether (or not) an event takes place: $I = 1$ if the event happens, and $I = 0$ if the event does not happen.

We will define some indicator random variables for this problem. First, let's number the customers $1, 2, \ldots, n$. For $i = 1, \ldots, n$, let $X_i$ be the indicator random variable for the event that customer $i$ gets their own hat back. In other words,

$$X_i = \begin{cases} 1 & \text{customer } i \text{ gets their own hat back} \\ 0 & \text{otherwise} \end{cases}$$

From the way we have constructed these indicator random variables, we see that

$$X = X_1 + X_2 + \cdots + X_n$$

Does this make sense? If we add up the indicator random variables, we are adding a 1 whenever a customer gets their own hat back, which gives us the total number of customers who get their hat back. Now we use linearity of expectation. The indicator variables $X_i$ are not independent, since, for example, if customers $1, 2, \ldots, n-1$ all get their own hat back, then customer $n$ must also get their own hat back. But that doesn't matter, since linearity of expectation does not require independence.

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(X_1 + X_2 + \cdots + X_n) \\ &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) \end{aligned}$$

If we can compute the expected value of the indicator random variables, we are all set. By the definition of expected value:

$$\begin{aligned} \mathbb{E}(X_i) &= 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) \\ &= \mathbb{P}(X_i = 1) \end{aligned}$$

$\mathbb{P}(X_i = 1)$ is the probability that customer $i$ gets their own hat back. Since the hats are distributed uniformly at random, and there are $n$ hats to distribute, we must have $\mathbb{P}(X_i = 1) = 1/n$. Thus,

$$\mathbb{E}(X_i) = \frac{1}{n} \quad \text{for } i = 1, 2, \cdots, n$$

Note that this does *not* depend on $i$, i.e. this is exactly the same for all $n$ customers. Substituting this above:

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} && n \text{ terms in this sum} \\ &= 1 \end{aligned}$$

The key to this method is that $\mathbb{E}(X_i)$ does not depend on $i$. Why does this make sense? I like to think of this in terms of symmetry. We number the customers for convenience, but mathematically there is no distinction between the $n$ customers[4]. Imagine the customers lining up to leave the restaurant. The person at the front of the line is handed a hat uniformly at random, then the customer leaves. This is repeated until all customers have left. If we swap any two customers in line, nothing should change. The expected number of customers who receive their own hat should remain the same.

How do you know when to use the method of indicator random variables and linearity of expectation? Like everything in probability, it is difficult to come up with hard-and-fast rules for when to use a given tool. That being said, here are a few guidelines for when this method is useful:

1. You are looking for an *expected value* involving a group of people or objects (not a probabiltiy).

2. You have a symmetric group of people or objects, i.e. you can swap them around without affecting the result.

[4] If you are the restaurant proprietor, don't tell them this!