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 $1/4$.

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].
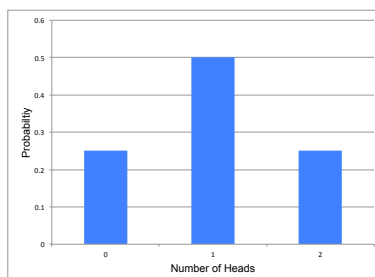
A discrete random variable induces a probability distribution on the sample space of all possible values the random variable can take. This is a different sample space from the original sample space. Back to our two-coin-flip example, the random variable Y induces a probability distribution on a new 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

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.
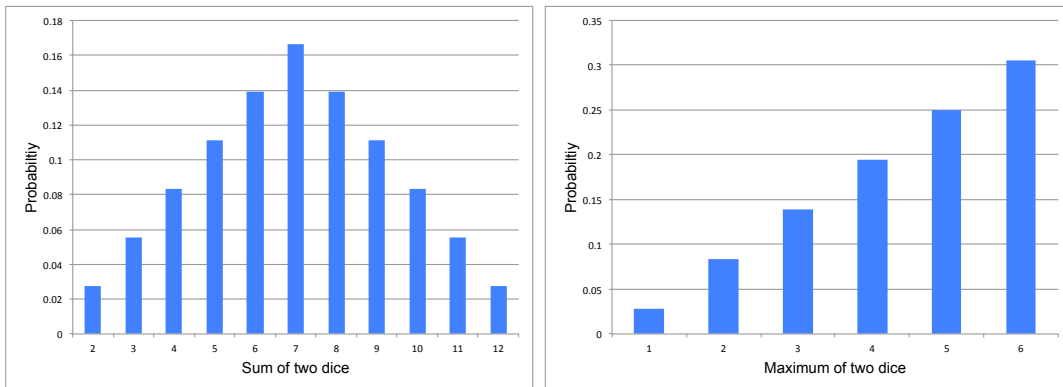
Both of these distributions are nonuniform, even though the underlying distribution of the two dice is uniform. The first distribution, that of the random variable X, is a familiar one

4

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 μ (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:

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

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.

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

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:

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

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!

We can also talk about functions of random variables. If Y is a random variable and $g(y)$ is a real-valued function, then $g(Y)$ is another random variable (since $g(Y)$ is also a function on our sample space). To evaluate $g(Y)$, just take the output of Y and run it through g.

Example. Let X be the output of a standard six-sided die, and let $g(x) = x^2$. Then $g(X)$ is also a discrete random variable. First, let's show X and $g(X)$ graphically:

	1	2	3	4	5	6
Sample space	1/6	1/6	1/6	1/6	1/6	1/6
X: output of die roll	1	2	3	4	5	6
$g(X) = X^2$	1	4	9	16	25	36

There are multiple ways to write the pmf for $g(X)$ in a table. I think the most useful way is to take the pmf table for X and add an extra column for $g(X)$. We will see why this makes sense shortly.

x	$g(x)$	$p(x)$
1	1	1/6
2	4	1/6
3	9	1/6
4	16	1/6
5	25	1/6
6	36	1/6

The expected value of a function of discrete random variable is computed in a similar fashion to the expected value of a random variable.

Expected value of a function of a discrete random variable

Let Y be a discrete random variable with probability function $p(y)$, and let g be a real-valued function. Then the expected value of the random variable $g(Y)$ is given by:

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

where the sum is taken over all possible values y can take. Note that the *only* difference from the expected value of Y is that we have $g(y)$ in place of y in the sum. The probabilties $p(y)$ are *unchanged*. This is again a weighted average, but this time we are taking a weighted average of the possible values of $g(Y)$.

Example. Let X be the output of a standard six-sided die, and let $g(x) = x^2$. Find $E[g(X)]$.

We can use the formula for the expected value of a function of a random variable, together with the pmf table above to get:

$$\mathbb{E}[g(X)] = \sum_{\text{all } x} g(x)\, p(x)$$

$$= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6} + 36 \cdot \frac{1}{6}$$

$$= \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36)$$

$$= \frac{91}{6} \approx 15.17$$

Before we go on, we mention one more property of expected value: the expected value of a constant.

Expected value of a constant

Let c be a constant. Then $\mathbb{E}(c) = c$. In other words, the expected value of a constant is just the constant itself.

Combining this with the linearity of expectation, we get the following expression for the expected value of a shifted and scaled random variable.

Expected value of $aY + b$

If Y is a random variable, and a and b are constants, then

$$\mathbb{E}(aY + b) = a\mathbb{E}(Y) + b$$

2.4   Variance

So far we have seen one quantititive descriptor of the distribution of a random variable: the expected value. In this section, we discuss another descriptor: the variance. To begin, let's compare the pmfs of two discrete random variables.

Example. Let X be the number of heads observed when a fair coin is flipped 6 times, and let Y be the uniform distribution on the integers $\{0, 1, 2, 3, 4, 5, 6\}$. Since Y is the uniform distribution on a finite set with 7 elements, $\mathbb{P}(Y = y) = 1/7$ for all Y. X takes the same values as Y since the number of heads observed in 6 coin flips is an integer between 0 and 6. What is $\mathbb{P}(X = x)$? Since there are two possibilities for each flip, there are 2^6 possible

flip-sequences. The number of flip-sequences which give us x heads is given by the binomial coefficient $\binom{6}{x}$ (why is this true?). Thus we have:
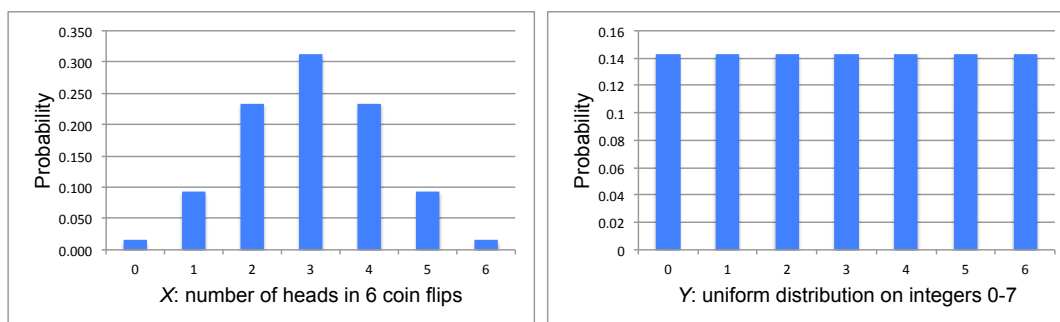
$$\mathbb{P}(X = x) = \frac{\binom{6}{x}}{2^6}$$

Putting both pmfs in tables we get:

x	$p(x)$
0	1/64
1	6/64
2	15/64
3	20/64
4	15/64
5	6/64
6	1/64

y	$p(y)$
0	1/7
1	1/7
2	1/7
3	1/7
4	1/7
5	1/7
6	1/7

Now let's look at the pmfs as histograms:

Both of these two random variables have an expected value (mean) of 3. (You can compute it yourself if you want, or reason that both distributions are symmetric, and the "middle" is 3.) Despite having the same mean, the two distributions are very different. X is relatively centered about the mean, while Y is all "spread out". We want a way to quantify this amount of "spread". There are many ways to do this. For example, we could use the

average distance from the mean. Statisticians have settled on a slightly different measure of "spread": the variance. In words, the variance of a random variable is the expected value of the squared-difference from the mean. Mathematically, we have the following definition:

Variance of a discrete random variable

Let Y be a discrete random variable with probability function $p(y)$, and let $\mu = \mathbb{E}(Y)$ be its expected value (mean). Then we define the *variance* of Y by

$$Var(Y) = \mathbb{E}[(Y - \mu)^2]$$

Using the formula for the expected value of a function of a random variable, this becomes

$$Var(Y) = \sum_{\text{all } y} (y - \mu)^2 \, p(y)$$

The variance is sometimes denoted by the symbol σ^2. The *standard deviation* of a random variable is the square root of the variance, and is denoted σ.

Let's compute the variance of the two random variables above. Since Y is more "spread out" than X, we expect it to have a higher variance.

Example. Let X be the number of heads observed when a fair coin is flipped 6 times, and let Y be the uniform distribution on the integers $\{0, 1, 2, 3, 4, 5, 6\}$. Compute the variances of both random variables.

The mean of both random variables is 3, so we have:

$$Var(X) = \sum_{\text{all } x} (x - 3)^2 \, p(x)$$

$$= (0-3)^2 \frac{1}{64} + (1-3)^2 \frac{6}{64} + (2-3)^2 \frac{15}{64} + (3-3)^2 \frac{20}{64} + (4-3)^2 \frac{15}{64} + (5-3)^2 \frac{6}{64} + (6-3)^2 \frac{1}{64}$$

$$= \frac{(9)(1)}{64} + \frac{(4)(6)}{64} + \frac{(1)(15)}{64} + \frac{(0)(20)}{64} + \frac{(1)(15)}{64} + \frac{(4)(6)}{64} + \frac{(9)(1)}{64}$$

$$= \frac{96}{64}$$

$$= \frac{3}{2}$$

$$Var(Y) = \sum_{\text{all } y} (y-3)^2 \, p(y)$$

$$= (0-3)^2\frac{1}{7} + (1-3)^2\frac{1}{7} + (2-3)^2\frac{1}{7} + (4-3)^2\frac{1}{7} + (4-3)^2\frac{1}{7} + (5-3)^2\frac{1}{7} + (6-3)^2\frac{1}{7}$$

$$= \frac{1}{7}(9+4+1+0+1+4+9)$$

$$= \frac{28}{7}$$

$$= 4$$

As predicted from the histograms, Y has a much higher variance.

The variance is often not computed by using the definition above, but by using a formula affectionately known as the *Magic Variance Formula*.

Magic Variance Formula

Let Y be a random variable and let $\mu = \mathbb{E}(Y)$ be its expected value (mean). Then the variance of Y is given by
$$Var(Y) = \mathbb{E}[Y^2] - \mu^2$$

To verify this formula, we expand out the binomial in the definition of variance, and use linearity of expectation.

$$
\begin{aligned}
Var(Y) &= \mathbb{E}[(Y-\mu)^2] \\
&= \mathbb{E}(Y^2 - 2\mu Y + \mu^2) \\
&= \mathbb{E}(Y^2) - 2\mu\mathbb{E}(Y) + \mathbb{E}(\mu^2) \\
&= \mathbb{E}(Y^2) - 2\mu^2 + \mu^2 \\
&= \mathbb{E}(Y^2) - \mu^2
\end{aligned}
$$

where we used the fact that $\mathbb{E}(Y) = \mu$, and the expected value of a constant is just the constant.

Variance behaves much differently from expectation. Unlike expectation, it is *not* linear. For a random variable Y, we have the following expression for the variance of $aY + b$.

Variance of $aY + b$

If Y is a random variable, and a and b are constants, then

$$Var(aY + b) = a^2 Var(Y)$$

To see this, we use the Magic Variance Formula. First we compute $\mathbb{E}[(aY + b)^2]$.

$$\mathbb{E}[(aY + b)^2] = \mathbb{E}(a^2Y^2 + 2abY + b^2)$$
$$= a^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(Y) + b^2$$

where we used the linearity of expectation and the fact that the expected value of a constant is the constant itself. Next, we compute $[\mathbb{E}(aY+b)]^2$. Once again using linearity of expectation, we get:

$$[\mathbb{E}(aY + b)]^2 = (a\mathbb{E}(Y) + b)^2$$
$$= a^2\mathbb{E}(Y)^2 + 2b\mathbb{E}(Y) + b^2$$

Using the Magic Variance Formula, we subtract these two to get:

$$Var(aY + b) = \mathbb{E}[(aY + b)^2] - [\mathbb{E}(aY + b)]^2$$
$$= \left(a^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(Y) + b^2\right) - \left(a^2\mathbb{E}(Y)^2 + 2b\mathbb{E}(Y) + b^2\right)$$
$$= a^2\left(E(Y^2) - [E(Y)]^2\right)$$
$$= a^2Var(Y)$$

where in the final line we have used the Magic Variance Formula one final time.

Note that the b disappears entirely. Why does this make sense? The variance represents the spread of a random variable around its mean. Shifting a random variable by a constant amount should not affect this spread.

We conclude this section with a result about the variance of a sum of *independent* random variables which we will need later. We have not formally defined independence for random variables, but the intuition is that the output of one does not affect the output of each other. This result is only true for independent random variables.

Variance of a sum of independent random variables

If $X_1, X_2, \ldots, X_n$ are *independent* random variables, then

$$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$$

i.e. the variance of a sum is the sum of the variances. This is only true for independent random variables.

2.5 Bernoulli Trials

One of the simplest models in probability is that of a sequence *Bernoulli trials*. Essentially Bernouilli trials are the generalization of repeated coin tossing.

> *Bernouilli trials*
> ---
> A sequence of Bernoulli trials is a sequence of experiements satisfying the following assumptions:
>
> 1. Each trial has exactly two possible outcomes, designated *success* and *failure.*
>
> 2. The trials are independent, i.e. the outcome of one trial does not influence the outcome of any other trial.
>
> 3. On each trial, the probability of success is p and the probability of failure is $1 - p$, where $0 \leq p \leq 1$. (Sometimes, for convenience, we define $q = 1 - p$).
>
> 4. These probabilities are constant for all trials, i.e. p does not change as the experiment is repeated.

Since Bernoulli trials have only two possible outcomes, we can think of them as "yes or no" events. Here are some examples of Bernoulli trials:

1. Flipping a coin of a coin, where we designate a flip of heads as a success. If it is a fair coin, then $p = 1/2$. If it is an unfair coin, then $p \neq 1/2$.

2. Rolling two standard, six-sided dice, where we designate a roll of doubles as success (and every other roll is failure). This is how you roll to get out of jail in Monopoly.

3. Playing a slot machine in Las Vegas.

4. Calling a registered voter uniformly at random and asking them if they are voting for Hillary Clinton.

5. Repeated free throw attempts by your instructor, assuming that he never improves.

Here are some examples of trials which are not Bernouilli trials. Why are these not Bernouilli trials?

1. Drawing cards one at a time from a standard deck of 52 playing cards, where we designate success as drawing an ace.

2. Calling a list of 100 registered voters one-at-a-time and asking them if they are voting for Hillary Clinton.

3. Repeated free throw attempts by your instructor, assuming his skill can improve (if only a little) with practice.

An individual Bernoulli trial modeled by a Bernoulli random variable.

> **Bernoulli random variable**
>
> ---
>
> A *Bernoulli random variable* Y with parameter p is a random variable which is 1 with probability p and 0 with probability $1 - p$. It models a single Bernoulli trial, where 1 indicates success and 0 indicates failure. We can write its pmf in the table below:
>
> | y | $p(y)$ |
> | --- | --- |
> | 0 | $1 - p$ |
> | 1 | p |
>
> To indicate that Y is a Bernoulli random variable with paramater p, we will sometimes write $Y \sim$ Bernoulli(p).

What is the mean and variance of a Bernoulli random variable? Let $Y \sim$ Bernoulli(p). Then for the expected value:

$$\mathbb{E}(Y) = \sum_{\text{all } y} y \, p(y)$$
$$= 0(1 - p) + 1(p)$$
$$= p$$

For the variance, we will use the Magic Variance Formula. This requires computing $\mathbb{E}(Y^2)$.

$$\mathbb{E}(Y^2) = \sum_{\text{all } y} y^2 \, p(y)$$
$$= 0^2(1 - p) + 1^2(p)$$
$$= p$$

By the Magic Variance Formula,

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

Summarizing, these results, we have:

> **Characteristics of a Bernoulli random variable**
>
> ---
>
> Let Y be a Bernouilli random variable with parameter p. Then:
>
> $$\mathbb{E}(Y) = p$$
> $$Var(Y) = p(1 - p) = pq$$
>
> where $q = 1 - p$.

Bernoulli trials are useful in many situations. Here are two questions we might want to ask about a sequence of Bernoulli trials.

1. How many successes do we have out of a *fixed* number of trials?

2. How many trials does it take to get our first success?

The first question is answered by the binomial distribution, and the second by the geometric distribution.

2.6   Binomial distribution

The binomial distribution models the number of successes in a fixed number of Bernoulli trials. Let's start with an example.

Example. Suppose we are playing a slot machine in Las Vegas. Suppose p is the probabiltiy of winning on one play[5]. Suppose you play 10 times. What is the probability that you win 2 times?

First, this is an example of Bernoulli trials, since each play of the slot machine has two possible outcomes (win and loss), is independent, and the probability of winning is does not change as you continue to play.

What is the sample space for this problem? Let us use the symbol W for a win and the symbol L for a loss. Then any sequence of 10 plays can be represented as a sequence of 10 Ws and Ls. One sequence might be LLWLLWLLLL. Is each sequence equally likely? Unless $p = 1/2$, different sequences can have different probabilities. For example, if $p < 1/2$, LLLLLLLLLL is much more probable than WWWWWWWWWW. What is the probability of a given sequence? Since we are looking at the probability of 2 wins (and thus 8 losses), let's look at a typical sequence where this is the case: LLWLLWLLLL. Since these are Bernoulli trials and thus each trial is independent, we just multiply together the probability of winning or losing at each trial.

$$\mathbb{P}(\texttt{LLWLLWLLLL}) = (1-p)(1-p)p(1-p)(1-p)p(1-p)(1-p)(1-p)(1-p) = p^2(1-p)^8$$

We see from this that the probability of a sequence depends only on the total number of wins and losses, not on the order in which those wins and losses occur. Thus the probability of *any* sequence of 10 trials composed of 2 wins and 8 losses will be $p^2(1-p)^8$.

The only question remaining is how many events in the sample space correspond to 2 wins and 8 losses, i.e. how many strings of length 10 contain exactly 2Ws and 8Ls? Recalling our combinatorics, this is given by the binomial coefficient $\binom{10}{2}$. Since each such string is a simple event in our sample space, we add up the probabilities of all $\binom{10}{2}$ of them to get:

$$\mathbb{P}(2 \text{ wins out of 10 trials}) = \binom{10}{2}p^2(1-p)^8$$

---

[5]This can be quite different depending on the type of slot machine you are playing, although all slot machines are designed so that you lose money on average. There are numerous websites devoted to slot machine odds.

We now generalize this. Suppose we have a sequence of n Bernoulli trials, with probability of success p for each trial. What is the probability of obtaining y successes, where y is an integer between 0 and n? If we designate success by S and failure by F, we can take the sample space to be strings of n letters, where each letter is either S or F. Just as in the example above, the probability of any string depends only on the number of Ss and Fs, not on their order. If we have y successes, this corresponds to a string of length n with y Ss and $(n-y)$ Fs. The probability of that event is $p^y(1-p)^{n-y}$. Using the binomial coefficient, the number of strings of length n composed of y Ss and $(n-y)$ Fs is $\binom{n}{y}$. Thus we have:

$$\mathbb{P}(y \text{ successes out of } n \text{ trials}) = \binom{n}{y} p^y(1-p)^{n-y}$$

This probability distribution, which repesents the number of successes in a fixed number n Bernouilli trials, is called the *binomial probability distribution*. A random variable Y with this distribution is called a *binomial random variable*.

Binomial distribution

A discrete random variable Y has a *binomial distribution* with n trials and probability of success p if

$$p(y) = \binom{n}{y} p^y(1-p)^{n-y} \qquad\qquad y = 0, 1, \ldots, n$$

Y is a *binomial random variable* with parameters n and p, written $Y \sim \text{Binomial}(n, p)$. This models the number of successes out of n Bernoulli trials, where the probability of a single success is p.

First let's check to make sure Y is a well-defined discrete random variable, i.e. the probabilities of all its possible outputs sum to 1. Using the binomial theorem, we see that

$$\sum_{y=0}^{n} p(y) = \sum_{y=0}^{n} \binom{n}{y} p^y(1-p)^{n-y} = [p + (1-p)]^n = 1^n = 1$$

What is the expected value and variance of a binomial random variable? There are many ways to compute this, but we will use a clever trick based on linearity. Let $Y \sim \text{Binomial}(n, p)$. For each trial $i = 1, 2, \ldots, n$ let:

$$Y_i = \begin{cases} 0 & \text{trial } i \text{ is a failure} \\ 1 & \text{trial } i \text{ is a success} \end{cases}$$

Since $\mathbb{P}(Y_i = 1) = p$, $Y_i \sim \text{Bernoulli}(p)$. Since each Bernoulli random variable is 1 only if that trial is a success and Y is the number of trials,

$$Y = Y_1 + Y_2 + \cdots + Y_n$$

Using the linearity of expectation,

$$\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n) \\
&= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \cdots + \mathbb{E}(Y_n) \\
&= p + p + \cdots + p = np
\end{aligned}$$

Furthermore, since Bernoulli trials are independent, the Y_i are all independent. Thus we can do the same thing for the variance:

$$\begin{aligned}
Var(Y) &= Var(Y_1 + Y_2 + \cdots + Y_n) \\
&= Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n) \\
&= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)
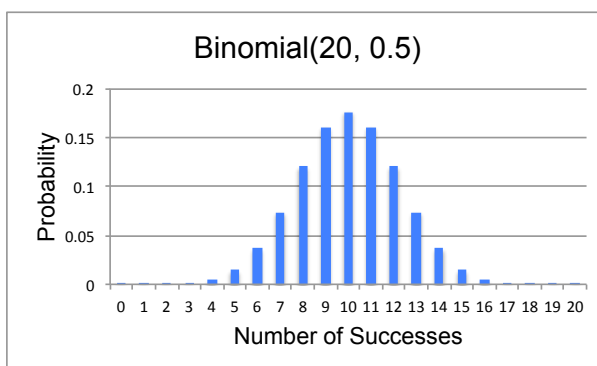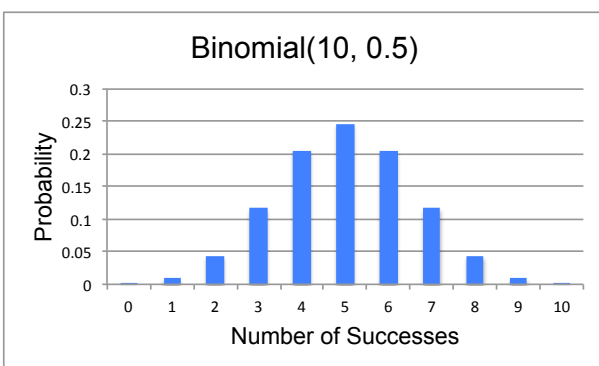\end{aligned}$$

Summarizing these properties, we have:

Properties of binomial distribution

Let Y be a binomial random variable with parameters n and p. Then

$$\mathbb{E}(Y) = np$$
$$Var(Y) = np(1-p)$$

Let's look at histograms of the binomial distribution for a few choices of parameters. First here are histograms for $p = 1/2$ (fair coin):

These distributions are perfectly symmetric about the mean, which is expected for the case where $p = 1/2$. Note that as n increases, these look more and more like "bell curves". While we have not yet talked about the normal distribution, keep in mind how these histograms look. For large enough n, we will be able to approximate the (discrete) binomial distribution by the (continuous) normal distribution.

The histograms look a bit different for p significantly different from $1/2$. These histograms are for $p = 0.2$.
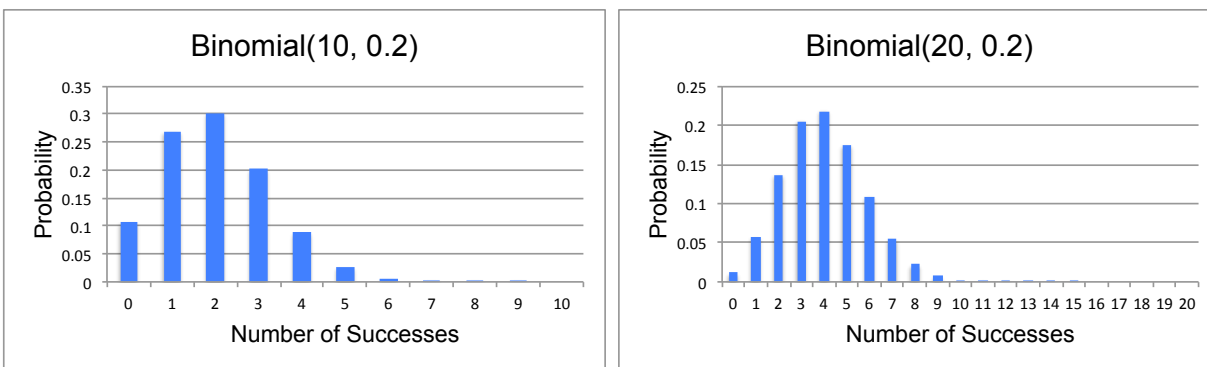
These are not symmetric. Since $p < 1/2$, the distributions are skewed to the left, which is what we expect since failure is more likely that success. Although not to the same extent as the case where $p = 1/2$, these also start to look like bell curves as n increases. We will also be able to approximate these by normal distributions for large n, but the farther p is from $1/2$, the larger n will need to be for this approximation to be reasonable.

Example. When a certain variety of pea plant with purple flowers is cross-fertilized, the offspring have purple flowers $3/4$ of the time and white flowers $1/4$ of the time[6].

1. You cross-fertilize 20 of these pea plants. What is the the probability that none of them are white?

   We can model this problem as a sequence of 20 Bernoulli trials, with "success" defined as a cross-fertilization producing a white flower. (Are all the assumptions of Bernouilli trials satisfied here?) Since we are looking for the number of successes in a *fixed* number of trials, this is a binomial distribution. Let $X \sim Binomial(20, 1/4)$. Then, using the binomial pmf:

   $$\mathbb{P}(X = 0) = \binom{20}{0} \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{20} = \left(\frac{3}{4}\right)^{20} \approx 0.0032$$

   In other words, it is highly unlikely that we will have no white flowers.

2. What is the probability that we will have at least 2 white flowers?

   Let A be the event that we get at least two white flowers. It is much easier here to compute the probability of the opposite event A^c, the event that we get 0 or 1 white flower.

   $$\mathbb{P}(A^c) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1)$$
   $$= \binom{20}{0} \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{20} + \binom{20}{1} \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^{19}$$
   $$= \left(\frac{3}{4}\right)^{20} + 20 \left(\frac{1}{4}\right) \left(\frac{3}{4}\right)^{19} \approx 0.024$$

---

[6]This is one of the varieties Gregor Mendel tested. In this problem, the pea plants are heterozygous for flower color, i.e. genotype Pp.

Thus we have $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 0.976$.

3. What is the expected number of white flowers?

The expected value of a binomial random variable is np. For this case, we have $n = 20$ and $p = 0.25$, thus
$$\mathbb{E}(X) = np = (20)(0.25) = 5$$