Chapter 6: Discrete Probability Distributions

Source: Wikipedia

In Chapter 6, we expand on the probability concepts we learned in Chapter 5, and introduce the idea of a **random variable**. Random variables are useful because they help us determine if playing a game like roulette (shown to the right) is profitable in the long-term. (It isn't.)

Random variables also help us determine how insurance companies set the premiums for their policies. They can also help an investor decide whether or not to invest in a company.

Source: stock.xchng

In Section 6.2, we'll introduce a specific type of random variable called a **binomial random variable**. Binomial random variables will help us answer questions like "What's the probability of getting 3 questions right on a multiple choice test if we're just guessing?" Or "If a basketball player makes 80% of her free throws, how often will she make less than 8 of 10?"

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

Section 6.1: Discrete Random Variables

Objectives

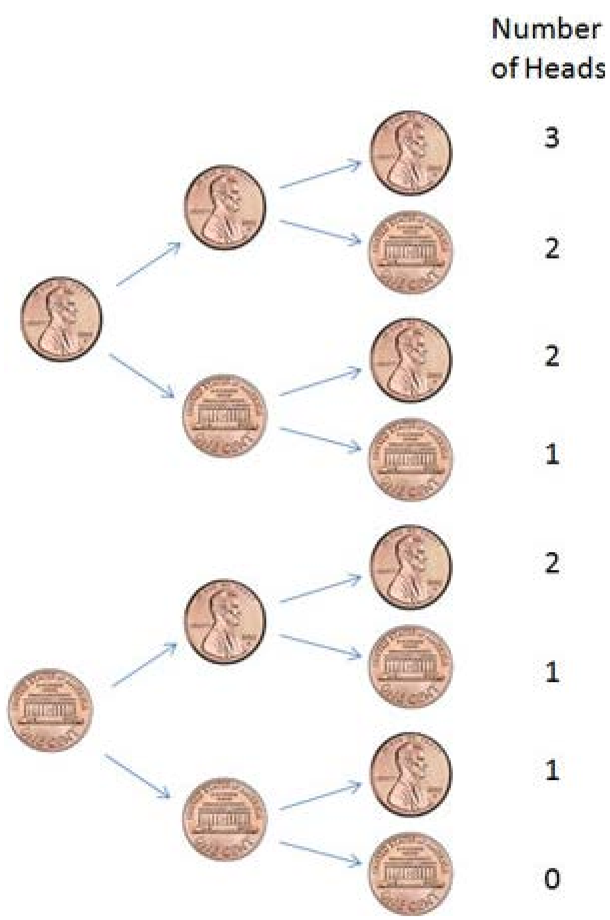
By the end of this lesson, you will be able to...

1. distinguish between discrete and continuous random variables
2. identify discrete probability distributions
3. construct probability histograms
4. compute and interpret the mean of a discrete random variable
5. interpret the mean of a discrete random variable as an expected value
6. compute the variance and standard deviation of a discrete random variable*

* You will not be tested on this objective.

Random Variables

Many probability experiments can be characterized by a numerical result. In Example 1, from Section 5.1, we flipped three coins. Instead of looking at particular outcomes (HHT, HTT, etc.), we might instead be interested in the total number of heads. Something like this:



In this case, the *number of heads* is called a **random variable**.

A **random variable** is a numerical measure of the outcome of a probability experiment whose value is determined by chance.

Example 1

Another example might be when we roll two dice, as in Example 2, from Section 5.1. Rather than looking at the dice individually, we can instead look at the *sum* of the dice, which would be a random variable.

Source: stock.xchng

In this case, if we let X = the sum of the two dice, x = 2, 3, 4, ..., 12. (We usually use a capital X to represent the random variable, and a lower case x to represent the particular values it can take on.)

One common goal with random variables is to know what the probability of each value is. This is called a **probability distribution**.

The **probability distribution** of a discrete random variable X provides the possible value of the random variable along with their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

Let's look at the earlier coin example to illustrate.

Example 2

If we flip three fair coins and let X = the number of heads, we know that x = 0, 1, 2, 3. We also know that since the coin is fair, each of the strands in the tree diagram shown earlier is equally likely. Since there are 8 total outcomes (HHH, HHT, HTH, etc), the probability distribution would look something like this:

Source: stock.xchng

x	P(x)
0	1/8
1	3/8
2	3/8
3	1/8

You'll notice that if find the sum of the P(x) values, we get 1.

Why don't you try an example now:

Example 3

Consider again the probability experiment where a fair six-sided die is rolled twice, and X = the sum of the two dice. Find the probability histogram for X.

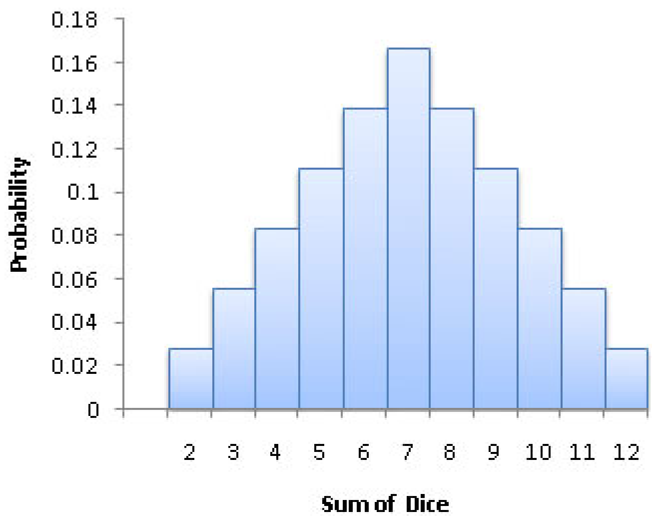
[reveal answer]

Probability Histograms

A probability histogram is similar to a histogram for single-valued discrete data from Section 2.2, except the height of each rectangle is the probability rather than the frequency or relative frequency.

Example 4

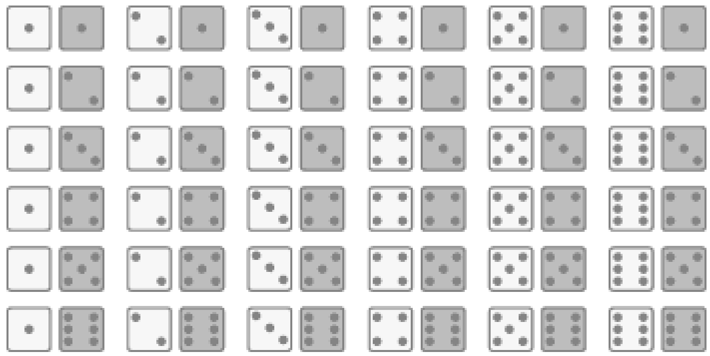
Looking again at the previous example about rolling two dice, the probability histogram would look something like this:



Since random variables represent numbers, it seems reasonable that we should be able to find the mean of those numbers, much like we did in Section 3.1. In fact, we can, but it has a much different meaning in this new context.

The Mean of a Random Variable

Let's consider our example again with the two dice.



We know x = 2, 3, 4, ... , 12, but there aren't equal numbers of each. If we were to calculate the mean of x like we usually would, we'd get something like this:

$$
\begin{aligned}
\mu &= \frac{2+3+3+4+4+4+\cdots+11+11+12}{36} \\
&= \frac{2\cdot 1+3\cdot 2+4\cdot 3+\cdots+11\cdot 2+12\cdot 1}{36} \\
&= \frac{2\cdot 1}{36}+\frac{3\cdot 2}{36}+\frac{4\cdot 3}{36}+\cdots+\frac{11\cdot 2}{36}+\frac{12\cdot 1}{36} \\
&= 2\cdot\frac{1}{36}+3\cdot\frac{2}{36}+4\cdot\frac{3}{36}+\cdots+11\cdot\frac{2}{36}+12\cdot\frac{1}{36} \\
&= 2\cdot P(X=2)+3\cdot P(X=3)+4\cdot P(X=4)+\cdots+11\cdot P(X=11)+12\cdot P(X=12) \\
&= 7
\end{aligned}
$$

What an interesting result! In fact, this isn't coincidental. The mean of a random variable will always be the sum of the values of the random variable multiplied by their corresponding probabilities. More formally:

The Mean of a Discrete Random Variable

The mean of a discrete random variable is given by the formula

$$u_X = \sum [x \cdot P(x)]$$

where x is the value of the random variable and P(x) is the probability of observing the random variable x.

All right, let's try one.

Example 5

Let's go back to the three coins again. If we flip three fair coins and let X = the number of heads, what is the mean of X?

[reveal answer]

Expected Value

So what does the mean of a random variable actually ... mean? When we found a mean of 7 in the example above, it certainly didn't mean that we'll get 7 every time. And it doesn't mean that our average after say 20 rolls will be 7.

No - the mean of a random variable is like a *long-term expectation*. Think of the mean being what to expect in the long run. In fact, the mean is usually referred to as the **expected value** of the random variable. If we continue throwing the two dice in the earlier example, our long-term average will get closer and closer to 7. In Example 5, the more times we toss three coins, the closer our long-term average will approach 1.5.

Expected values are used in lots of calculations in the business and finance world. Poker players use expected values to help make decisions on whether to continue playing in a hand. Hopefully by the end of this section, you'll use the idea of expected value to *not* play games of chance like roulette!

Let's illustrate with some examples.

Example 6

In the game of roulette, a wheel consists of 38 slots numbered 0, 00, 1, 2, ... , 36. To play the game, a metal ball is spun around the wheel and

Source: Wikipedia

is allowed to fall into one of the numbered slots. A *dozen bet* is betting that one of a particular dozen numbers hits on the next spin of the wheel. The wheel is divided into 3 different groups; 1-12, 13-24, and 25-36. If the payout for a *dozen bet* is 2 to 1 (the original bet is returned, along with twice its value), what is the expected value of $1 *dozen bet*?

Solution: To solve this problem, let's first define a random variable. In a situation like this, we typically let X = amount won. X can then take on values of either -$1 (we lose) or +$2 (we get our $1 plus $2 more).

P(X = 2) = 12/38, since there are 12 ways to land on our *dozen* (regardless of which dozen we choose), and

P(X = -1) = 26/38, since there are 26 other numbers

The expected value of X is then:

E(X) = (-1)•P(X = -1) + (2)•P(X = 2)
= (-1)(26/38) + (2)(12/38)
= -2/38 ≈ -$0.05

So on average, we expect to lose 5¢ for every $1 bet we make. Of course, this doesn't mean we'll lose 5¢ *every* time - just that in the long run, we'll average a loss of 5¢ per $1 bet.

You can see examples of other Roulette expected values here.

Example 7

Source: everystockphoto

Consider a car owner who has an 80% chance of no accidents in a year, a 20% chance of being in a single accident in a year, and no chance of being in more than one accident in a year.

For simplicity, assume that there is a 50% probability that after the accident the car will need repairs costing $500, a 40% probability that the repairs will cost $5,000, and a 10% probability that the car will need to be replaced, which will cost $15,000.

What is the expected loss for the car owner per year?

Solution: This one is a little trickier. If we let X = loss for the year, X can be $0, $500, $5,000, or $15,000. P(X=0) = 0.8, but P(X = $500) is actually (0.2)(0.5), since there's a 20% chance of being in an accident, and a 50% chance of that accident causing repair costs of $500. The complete list of X and its corresponding values is:

x	P(X= x)
$0	0.80
$500	(0.2)(0.5) = 0.1
$5,000	(0.2)(0.4) = 0.08
$15,000	(0.2)(0.1) = 0.02

E(X) = (0)(0.8) + (500)(0.1) + (5,000)(0.08) + (15,000)(0.02) = 750

So the owner can expect to lose $750 on average per year. Any insurance company must charge more than this amount to make a profit.

And while $750 might not sound that bad, is the risk of losing $15,000 really worth it? That's why most car owners buy more than the basic insurance - not because it's a good deal (it isn't - the insurance cost will be more than the expected loss) but because it gives the owner piece of mind.

Example 8

Source: CardPlayer

Gambling is full of expected value calculations. We already did one earlier about roulette (and saw that the game has a negative expectation for the player). Let's try another one, but this time make it more a game of skill.

There are many variations of poker, but most involve rounds of betting, where players have to choose what to bet, and whether to "call" (pay the amount bet by another player). One relatively simple situation to look at is the final round of betting.

Let's suppose you're in a poker game with the great Doyle Brunson. It's the last play of the hand - Doyle has bet $50 and you have to decide whether or not to call. There's $250 in the middle of the table if you call and win. If you call and lose, you'll lose another $50.

Doyle has been around a long time, so you can't be sure if he's bluffing. If you call and he is, you win the $250 in the middle and his additional $50. If you call and he's not bluffing, you lose your additional $50.

Based on the way the hand has played out, you think there's about a 20% chance he has nothing, and an 80% chance that he has you beat. What is the expected value of calling Doyle's bet?

Solution: There's a lot going on here, but let's start like we've done in the past. We'll make a random variable, call it X, with possible values of $300 (you call and win), and -$50 (you call and lose).

E(X) = ($300)(0.2) + (-$50)(0.8) = $20

So on average, if we were to replay this hand over and over, we'd expect an average profit of $20 over time. So yes, we should call. Keep in mind, we're still going to lose 80% of the time, but the 20% of the time we win is enough to make it worth the while.

Here's one more for you.

Example 9

Source: stock.xchng

Suppose you have an investment opportunity. A new small business in town is looking for investors, and they're asking for a $20,000 investment from you. After some investigating, you determine that with the current economic climate, there's a 40% chance the company will fail in the first year, and you'll lose the full $20,000. There's about a 50% chance the company will struggle but survive, and you'll have a loss of about $5,000. There is an opportunity in the field, though, and you guess there's about a 10% chance that the company can make it big, and you'll quadruple your investment in the first year.

Based on these estimates, should you make the investment?

[reveal answer]

The Standard Deviation of a Random Variable

Although we're usually much more interested in the expected value (mean) of a random variable, there are times (especially later on in the course) that we'll also be interested in the standard deviation. The formula is similar to the mean in that it weights each value by its corresponding probability.

The Variance and Standard Deviation of a Discrete Random Variable

The variance of a discrete random variable is given by the formula

$$\sigma_X^2 = \sum[(x - \mu_x)^2 \cdot P(x)]$$

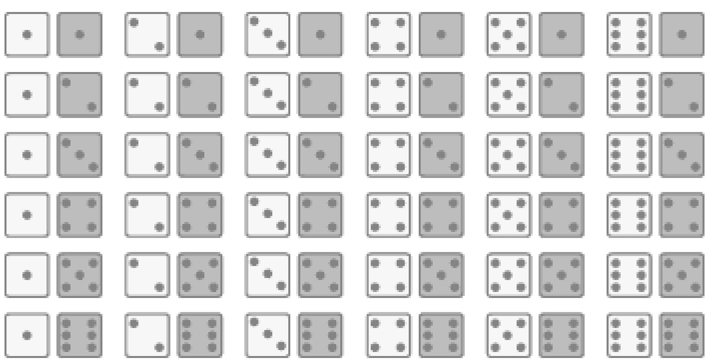
where x is the value of the random variable and P(x) is the probability of observing the random variable x.

To find the standard deviation of the discrete random variable, take the square root of the variance.

We'll just do one quick example of standard deviation.

Example 10

Let's consider our example again with the two dice.



We know from earlier this section that the expected value is 7. Using that, we can find the variance and standard deviation.

$$\sigma_X^2 = (2-7)^2 \cdot \frac{1}{36} + (3-7)^2 \cdot \frac{2}{36} + \cdots + (11-7)^2 \cdot \frac{2}{36} + (12-7)^2 \cdot \frac{1}{36} \approx 5.83$$

And so the standard deviation is: $\sigma_X = \sqrt{\sigma_X^2} \approx 2.42$

Section 6.2: The Binomial Probability Distribution

Objectives

By the end of this lesson, you will be able to...

1. determine whether a probability experiment is a binomial experiment
2. compute probabilities of binomial experiments
3. compute and interpret the mean and standard deviation of a binomial random variable

Binomial Experiments

In the last section, we talked about some specific examples of random variables. In this next section, we deal with a particular type of random variable called a **binomial random variable**. Random variables of this type have several characteristics, but the key one is that the experiment that is being performed has only two possible outcomes - *success* or *failure*.

An example might be a free kick in soccer - either the player scores or goal or she doesn't. Another example would be a flipped coin - it's either heads or tails. A multiple choice test where you're totally guessing would be another example - each question is either right or wrong.

Let's be specific about the other key characteristics as well:

Criteria for a Binomial Probability Experiment

A **binomial experiment** is an experiment which satisfies these four conditions:

- A fixed number of trials
- Each trial is independent of the others
- There are only two outcomes
- The probability of each outcome remains constant from trial to trial.

In short: *An experiment with a fixed number of independent trials, each of which can only have two possible outcomes*.

(Since the trials are independent, the probability remains constant.)

If an experiment is a binomial experiment, then the random variable X = the number of successes is called a **binomial random variable**.

Let's look at a couple examples to check your understanding.

Example 1

Consider the experiment where three marbles are drawn without replacement from a bag containing 20 red and 40 blue marbles, and the number of red marbles drawn is recorded. Is this a binomial experiment?

[reveal answer]

Source: stock.xchng

Example 2

A fair six-sided die is rolled ten times, and the number of 6's is recorded. Is this a binomial experiment?

[reveal answer]

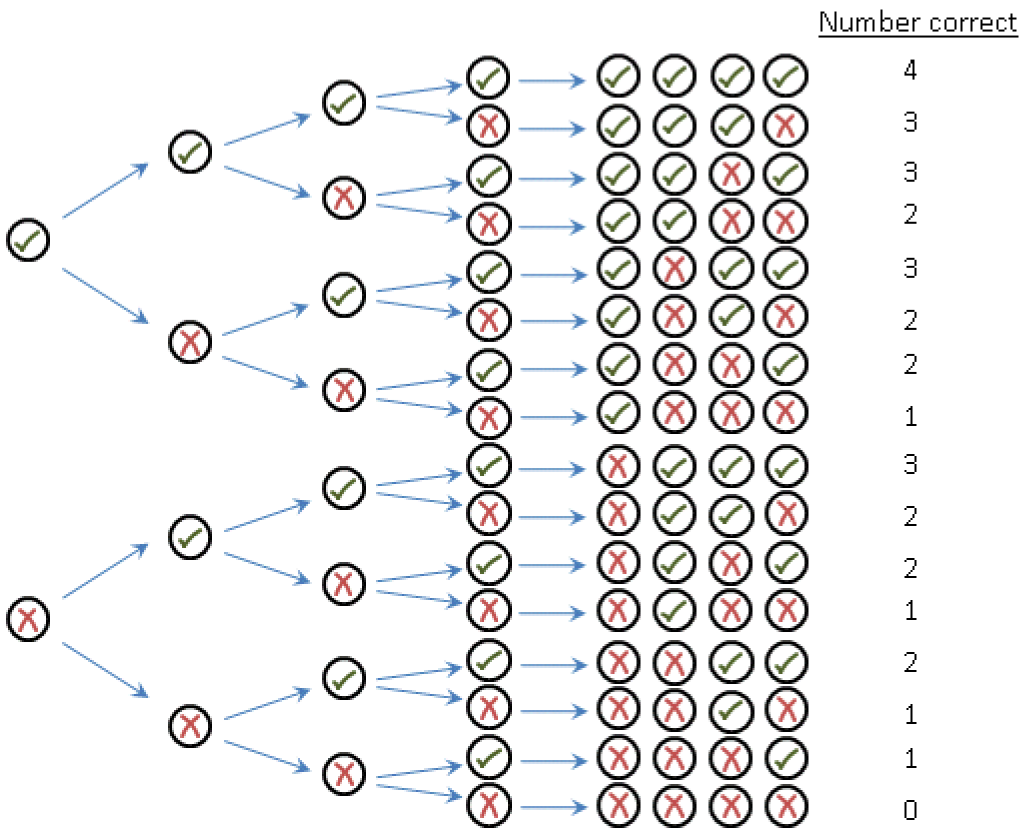
The Binomial Distribution

Once we determine that a random variable is a binomial random variable, the next question we might have would be how to calculate probabilities.

Let's consider the experiment where we take a multiple-choice quiz of four questions with four choices each, and the topic is something we have absolutely no knowledge. Say... theoretical astrophysics. If we let X = the number of correct answer, then X is a binomial random variable because

- there are a fixed number of questions (4)
- the questions are independent, since we're just guessing
- each question has two outcomes - we're right or wrong
- the probability of being correct is constant, since we're guessing: 1/4

So how can we find probabilities? Let's look at a tree diagram of the situation:



Finding the probability distribution of X involves a couple key concepts. First, notice that there are multiple ways to get 1, 2, or 3 questions correct. In fact, we can use combinations to figure out how many ways there are! Since P(X=3) is the same regardless of which 3 we get correct, we can just multiply the probability of one line by 4, since there are 4 ways to get 3 correct.

Not only that, since the questions are *independent*, we can just multiply the probability of getting each one correct or incorrect, so P(✓✓✓✗) = $(3/4)^3(1/4)$. Using that concept to find all the probabilities, we get the following distribution:

x	P(x)
0	$P(0) = \left(\frac{1}{4}\right)^4$
1	$P(1) = 4\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^3$
2	$P(2) = 6\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^2$
3	$P(3) = 4\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^1$
4	$P(4) = \left(\frac{1}{4}\right)^4$

We should notice a couple very important concepts. First, the number of possibilities for each value of X gets multiplied by the probability, and in general there are ${}_4C_X$ ways to get X correct. Second, the exponents on the probabilities represent the number correct or incorrect, so don't stress out about the formula we're about to show. It's essentially:

P(X) = (ways to get X successes)•(prob of success)$^{\text{successes}}$•(prob of failure)$^{\text{failures}}$

The Binomial Probability Distribution Function

The probability of obtaining x successes in n independent trials of a binomial experiment, where the probability of success is p, is given by

$$P(x) = {}_nC_x p^x (1-p)^{n-x}$$

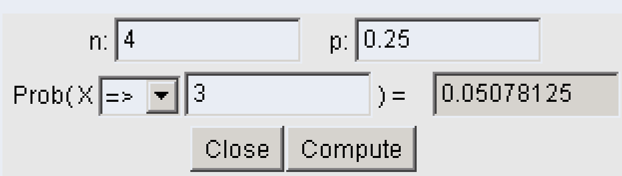
Where x = 0, 1, 2, ... , n

Technology

Here's a quick overview of the formulas for finding binomial probabilities in StatCrunch.

Click on **Stat** > **Calculators** > **Binomial**

Enter n, p, the appropriate equality/inequality, and x. The figure below shows P(X≥3) if n=4 and p=0.25.



Let's try some examples.

Example 3

Consider the example again with four multiple-choice questions of which you have no knowledge. What is the probability of getting exactly 3 questions correct?

[reveal answer]

Example 4

A basketball player traditionally makes 85% of her free throws. Suppose she shoots 10 baskets and count the number she makes. What is the probability that she makes less than 8 baskets?

[reveal answer]

Source: stock.xchng

Example 5

Traditionally, about 70% of students in a particular Statistics course at ECC are successful. Suppose 20 students are selected at random from all previous students in this course. What is the probability that more than 15 of them will have been successful in the course?

[reveal answer]

The Mean and Standard Deviation of a Binomial Random Variable

Source: stock.xchng

Let's consider the basketball player again. If she takes 100 free throws, how many would we expect her to make? (Remember that she historically makes 85% of her free throws.)

The answer, of course, is 85. That's 85% of 100.

We could do the same with any binomial random variable. In Example 5, we said that 70% of students are successful in the Statistics course. If we randomly sample 50 students, how many would we expect to have been successful?

Again, it's fairly straightforward - 70% of 50 is 35, so we'd expect 35.

Remember back in Section 6.1, we talked about the mean of a random variable as an expected value. We can do the same here and easily derive a formula for the mean of a binomial random variable, rather than using the definition. Just as we did in the previous two examples, we multiply the probability of success by the number of trials to get the expected number of successes.

Unfortunately, the standard deviation isn't as easy to understand, so we'll just give it here as a formula.

The Mean and Standard Deviation of a Binomial Random Variable

A binomial experiment with n independent trials and probability of success p has a mean and standard deviation given by the formulas

$$\mu_X = np \text{ and } \sigma_X = \sqrt{np(1-p)}$$

Let's try a quick example.

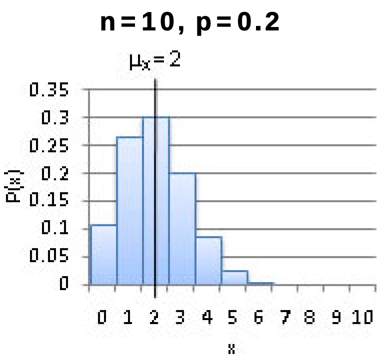
Example 6

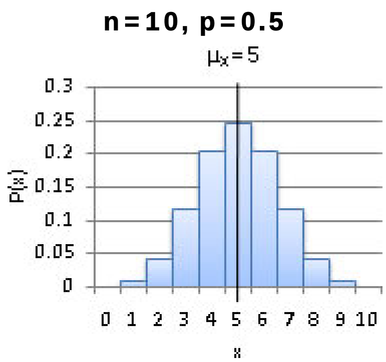
Suppose you're taking another multiple choice test, this time covering particle physics. The test consists of 40 questions, each having 5 options. If you guess at all 40 questions, what are the mean and standard deviation of the number of correct answers?

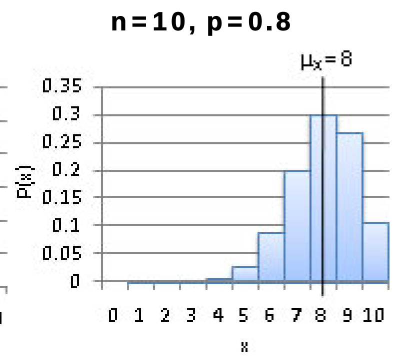
[reveal answer]

The Shape of a Binomial Probability Distribution

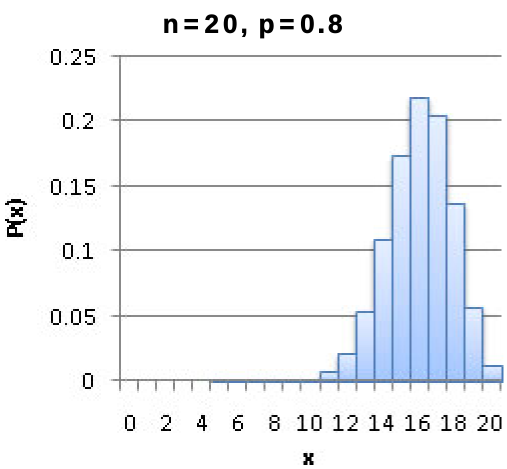
The best way to understand the effect of n and p on the shape of a binomial probability distribution is to look at some histograms, so let's look at some possibilities.

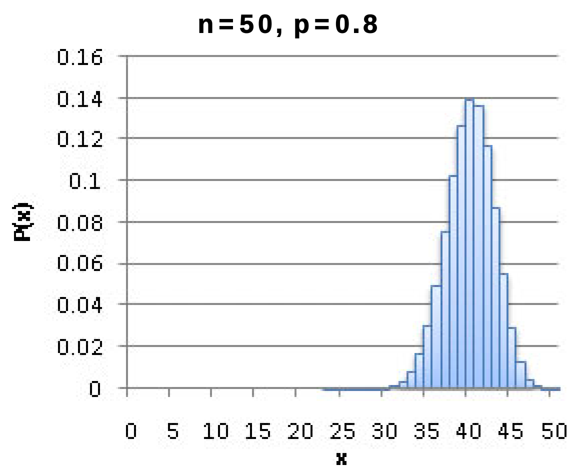






Based on these, it would appear that the distribution is symmetric only if p=0.5, but this isn't actually true. Watch what happens as the number of trials, n, increases:





Interestingly, the distribution shape becomes roughly symmetric when n is large, even if p isn't close to 0.5. This brings us to a key point:

> As the number of trials in a binomial experiment increases, the probability distribution becomes bell-shaped. As a rule of thumb, if $np(1-p) \geq 10$, the distribution will be approximately bell-shaped.