Unit 21: Binomial Distributions

SUMMARY OF VIDEO

In Unit 20, we learned that in the world of random phenomena, probability models provide us with a list of all possible outcomes and probabilities for how often they would each occur over the very long term. Probability models could be used to find out how many times we can expect to get heads on a coin toss, or how many daffodils we can expect to bloom in the spring given the number of bulbs planted in the fall, or even the number of children in a family that will have inherited a genetic disease.

There is a common thread to all of these examples. They are all concerned with random phenomena that have only two possible outcomes: heads versus tails, blooms versus none, sick versus healthy. Traditionally, we label one possible outcome as a success and the other as a failure. In this unit, what we are interested in is the number of successes. This count forms a particular kind of discrete probability model called the binomial distribution.

In addition to a situation having only two possible outcomes, there are three more conditions that must be true in order for it to be in a binomial setting. To demonstrate all four conditions, we turn to the context of basketball free throws. Free throws are mostly clear of any defensive pressure or other external factors during a basketball game. We only need to be concerned with whether or not the ball goes into the net.

The first trait of a binomial setting is that there is a repeated, fixed number, *n*, of trials, or observations. In this case, *n* equals the number of shots taken. Second, all of these trials are independent; meaning that the outcome of one trial does not change the probabilities of the other trials. Now, conventional basketball wisdom might tell you this is not true if a player has a “hot hand,” meaning that, because of the success of their previous shots, they are more likely to make successive baskets. But a 1985 study showed that this was not the case. The study found that players are not more likely to make second shots after making their first. The second shot’s outcome is independent of the first shot. Therefore, free throws do fit this trait of a binomial setting. The third trait that we must look at is whether or not each of the trials ends in one of two outcomes: success, *S*, or failure, *F*. Our basketball example fits – either the ball goes into the net (success) or it misses (failure). Lastly, the probability of success, or *p*, must be the same for all trials. Free throws are always shot from the same distance. There is no defensive pressure. Each time the player lines up at the free throw line he has the same probability of making his shot, which is based on his particular shooting skills. So, basketball free throws fit the binomial setting!

Next, we look at an example where the binomial setting comes up in genetics. Binomial distributions can help determine the probability of how many children in a family of six will inherit a genetic disease. Sickle cell disease is a genetic disorder of the red blood cells, estimated to affect around 100,000 people in the United States. It can cause a great deal of pain, life-threatening infections, strokes, and chronic organ damage. Inside the red blood cells is a protein called hemoglobin. There are different types of hemoglobin. For example, in the womb we all have fetal hemoglobin or hemoglobin F. After birth, we change over to normal hemoglobin A or, for people with sickle cell disease, to hemoglobin S. Like all genetic traits, the genes that determine an individual's hemoglobin type are inherited, one version from each parent. Since sickle cell is a recessive disease, a child needs to receive two bad versions of the gene, one from each parent, to have the disease.

In the typical situation for two parents who are carriers, each has one bad copy of the gene, S, and one good copy, A. Each time the mother makes an egg, there is a 50% chance that she will produce an egg that has the sickle mutation and a 50% chance she will produce a normal egg. On the other side, the father has a 50% chance that his sperm will have the sickle mutation and a 50% chance of normal. When the sperm and egg come together there are four possible outcomes for the gene from the mother and gene from the father (in that order): AA, AS, SA, SS. So, there is a 25% chance of the child inheriting SS and thus having the disease and a 75% chance of the child inheriting something other than SS and not having the disease. Because of the parents' genetic makeup, this probability never changes. With each pregnancy the probability of having a child with sickle cell disease is 1/4 or 0.25. So sickle cell disease fits the binomial setting. Here's a recap of why:

There are two possible outcomes – a sick child or a healthy child.

For this family, there are six children; $n = 6$.

The outcome for each child is independent.

The parents' genetic makeup never changes; $p = 0.25$ for each child.

Next, we look at a probability histogram for the number of children with sickle cell disease in families with six children where each parent is a carrier. We can use technology or a binomial table to calculate the probabilities and then use the values to form the histogram (See Figure 21.1.). The horizontal axis is marked with the values for *x*, the number of sick children in a family of six children. The vertical axis shows the probability for each possible value of *x*.

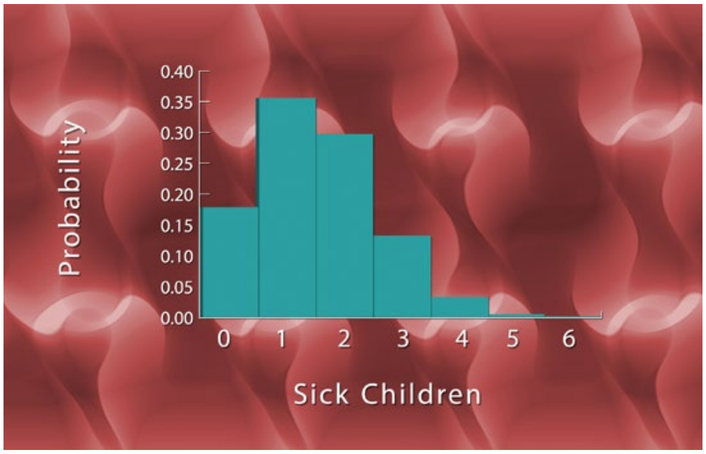

Figure 21.1. Probability histogram for a binomial distribution; n = 6 and p = 0.25.

We can see from this histogram that the probability of having three children with sickle cell, for example, is around 0.13 or 13%.

Public health officials might want to find the mean of this distribution. Fortunately that is easy to calculate. We just multiply the number of trials, *n*, times the probability of success, *p*:

$\mu = np$

So, the mean number of children with sickle cell disease, in families of six children, where both parents are carriers, is 1.5. Of course, no family has 1.5 children! This is the statistical average over many, many families.

STUDENT LEARNING OBJECTIVES

A. Be able to identify a binomial setting and define a binomial random variable.

B. Know how to find probabilities associated with a binomial random variable.

C. Know how to determine the mean and standard deviation of a binomial random variable.

CONTENT OVERVIEW

A basketball player shoots eight free throws. How many does he make? In a family of six children with two parents who are carriers of a genetic disease, how many of their children will inherit the disease? If you throw two dice four times, how many times will you throw doubles? In each of these situations, we can identify an outcome that we would like to count. We label that outcome a success (even if, as in the case of children inheriting a disease, that outcome is not good) and any other possible outcomes from these random phenomena as failure. So we begin our discussion with random phenomena in which all possible outcomes can be classified into one of two categories, success or failure.

In settings in which we can classify outcomes into successes and failures, we can define a random variable *x* to be the number of successes. The random variable *x* has a **binomial distribution** if the conditions of a **binomial setting** are satisfied. Those conditions are listed below.

The Binomial Setting

1. There is a fixed number of *n* trials or observations.
2. The trials are independent.
3. The trials end in one of two possible outcomes: Success (*S*) or Failure (*F*).
4. The probability of success, *p*, is the same for all trials.

If the conditions of the binomial setting are satisfied, then *x*, the number of successes, has a binomial distribution with parameters *n* and *p*; we express this distribution in shorthand as **b(*n, p*)**.

Now, we look at an example. As of August 2013, Jacqui Kalin was listed as having the top free throw percentage in Women's NCAA basketball. Her percentage of successful free throws was 95.5%. Suppose we test her skills on the basketball court and ask her to throw 20 free throws. Our random variable *x* is the number of free throws that she makes and we have a fixed number of *n* = 20 free throws. From the video, we know that consecutive free throws are independent. Each free throw either goes into the basket (success) or does not (failure). The probability of success depends on the shooting skills of the player; for Jacqui, *p* = 0.955. So, *x*, the number of successful free throws out of 20, fits the binomial setting and the distribution of *x* is b(20, 0.955).

Number 50 on the 2013 list of the highest free throws percentages in Women's NCAA basketball is Lauren Lenhardt. Her free throw percentage was listed as 83%. Suppose we test her skills on the basketball court and ask her to throw free throws until she gets a basket. We let *y* be the number of free throws until she gets one in the hoop. Unlike random variable *x* in the previous paragraph, *y* is not a binomial random variable because there is no fixed number of trials. If she makes her first free throw, then she stops and $n = 1$. If she misses and makes her second free throw, then $n = 2$.

One typical setting for the binomial distribution is with simple random samples.

Suppose that of the 2,500 adult patients at a healthcare center, 750 have high blood pressure and 1,750 do not have high blood pressure. Suppose that we select a random sample of 5 patients from the health center and let *x* equal the number of patients who have high blood pressure. This is not quite a binomial setting. Here's why. For trial 1, we randomly choose the first patient. The probability that Patient 1 has high blood pressure is $750/2500 = 0.3$. Now, for trial 2 we have 2,499 patients to choose from. The probability that Patient 2 has high blood pressure depends on the outcome for Patient 1. If Patient 1 had high blood pressure, then the probability that Patient 2 has high blood pressure is $749/2499 \approx 0.2997$. But if Patient 1 does not have high blood pressure, then the probability that Patient 2 has high blood pressure is $750/2499 \approx 0.3001$. So, the probability of success, having high blood pressure, is not constant for each trial. However, in this case, the distribution of *x* will be very close to b(5, 0.3). We summarize this result as follows.

> Choose a simple random sample of size *n* from a population with proportion *p* of successes. Let *x* be the number of successes in the sample. If the population is much larger than the sample, then *x*'s distribution is approximately b(*n*, *p*). (A good rule of thumb is that the population must be at least 20 times larger than the sample.)

Next, consider the context of children inheriting blue eyes from their brown-eyed parents, each of whom has a recessive gene for blue eyes. We'll label inheritance of blue eyes as success. In this case, each of their children has a 25% chance of inheriting blue eyes. In a family of six children whose parents have this genetic makeup, we would like to find the probability distribution for the number of their children, *x*, who will have blue eyes. In this situation, the distribution of *x* is b(6, 0.25).

Step 1: Find $p(0)$.

When $x = 0$, that means all six trials were failures: *FFFFFF*. From the Complement Rule, we know that $P(F) = 1 - P(S) = 0.75$. Since the trials are independent, we can use the Multiplication Rule to find the probability:

$p(0) = P(FFFFFF) = (P(F))^6 = (0.75)^6 \approx 0.178$

Step 2: Find $p(1)$.

When $x = 1$, one of the six trials is a success and the other five are failures. There are six possible ways that could happen:

SFFFFF, FSFFFF, FFSFFF, FFFSFF, FFFFSF, FFFFFS

We use the Multiplication Rule to find the probability that any outcome with one success and five failures occurs: $(0.25)(0.75)^5$, and there are six of these outcomes:

$p(1) = 6(0.25)(0.75)^5 \approx 0.336$

Step 3: Find $p(2)$.

When $x = 2$, two of the six trials are a success and the other four are failures. There are 15 possible ways that could happen:

SSFFFF, SFSFFF, SFFSFF, SFFFSF, SFFFFS,

FSSFFF, FSFSFF, FSFFSF, FSFFFS, FFSSFF,

FFSFSF, FFSFFS, FFFSSF, FFFSFS, FFFFSS

We use the Multiplication Rule to find the probability that any outcome with two success and four failures occurs: $(0.25)2(0.75)^4$ and there are 15 of these outcomes:

$p(2) = 15(0.25)^2(0.75)4 \approx 0.297$

At this point we could keep going and finish finding the probability distribution for x. However, our calculations to this point can serve as a pattern for a formula. First, we need to count the number of ways there are x successes in six trials. We can do that with combinations:

$\binom{6}{x} = \frac{6!}{x!(6-x)!}$ is the number of ways to choose x trials to label S from 6 trials.

In the case of x = 2:

$$\binom{6}{2} = \frac{6!}{2!4!} = \frac{6 \cdot 5 \cdot 4!}{2!4!} = 15$$

Step 4: Find p(3). Following the pattern established with p(0), p(1), and p(2), we get:

$$p(3) = \binom{6}{3}(0.25)^3(0.75)^3 = \frac{6!}{3!3!}(0.25)^3(0.75)^3 \text{ = 20(0.25)3(0.75)3} \approx 0.132$$

The general formula for calculating binomial probabilities is contained in the box below.

Binomial Probability

If the distribution of x is b(n, p), then the probabilities associated with x are given by

$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for x = 0, 1, 2, . . . , n.

Now that we can find the probability distribution for any binomial random variable, we could use the formulas given in Unit 20, Random Variables, to calculate its mean and standard deviation. However, there is a much easier way to calculate the mean and standard deviation when dealing with binomial random variables.

Mean and Standard Deviation

If the distribution of x is b(n, p), then the mean and standard deviation of x are

$\mu = np$

$\sigma = \sqrt{np(1-p)}$

So, for example, given x that is b(6, 0.25), the mean and variance of x are:

μ = (6)(0.25) = 1.5

$\sigma = \sqrt{(6)(0.25)(0.75)} = \sqrt{1.125} \approx 1.06$

We conclude this overview with one interesting feature of binomial distributions. Suppose we leave *p* at a fixed value and then allow *n* to increase – say from 5 to 30. Figure 21.2 shows histograms for the following distributions:

b(5, 0.3) (top left)

b(10, 0.3) (top right)

b(20, 0.3) (bottom left)

b(30, 0.3) (bottom right)

Notice that the probability histogram for b(5, 0.3) is skewed to the right. However, as *n* increases to 30, the probability histograms look more and more like bell-shaped curves. This example shows that if *x* has a binomial distribution and *n* is large, then we can approximate its distribution with a normal distribution.

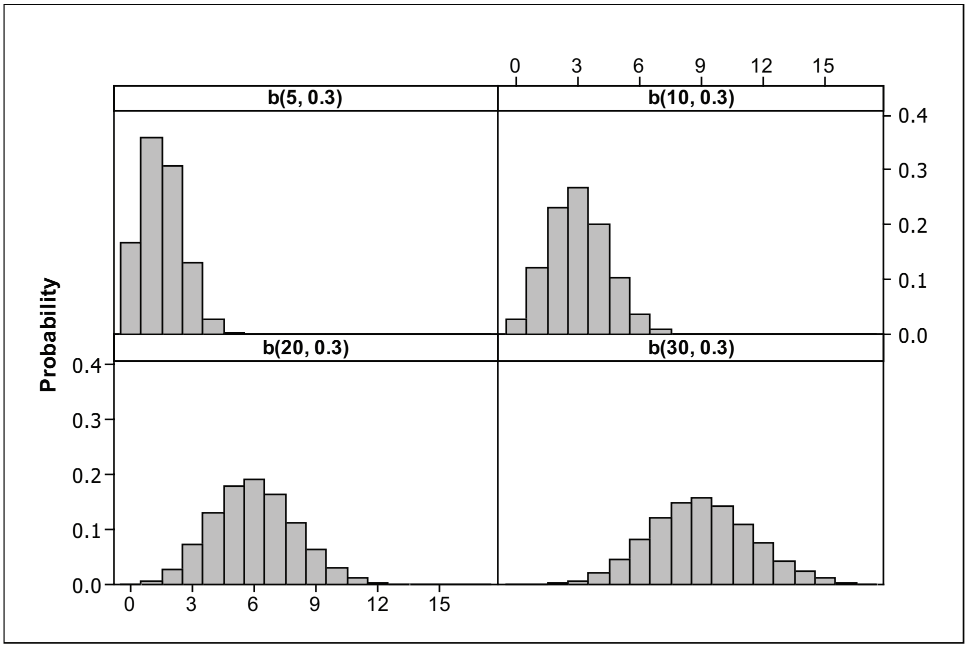

Figure 21.2. Probability histograms of four binomial distributions

Normal Approximation for Binomial Distributions

If *x* has the distribution b(*n*, *p*) and *n* is large, then the distribution of *x* is approximately normal with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. As a rule of thumb, this is a good approximation as long as *n* is sufficiently large so that $np \geq 10$ and $n(1-p) \geq 10$.

KEY TERMS

In a **binomial setting,** there is a fixed number of *n* trials; the trials are independent; each trial results in one of two outcomes, success or failure; the probability of success, *p*, is the same for each trial. In a binomial setting with *n* trials and probability of success *p*, *x* = the number of successes that has the **binomial distribution** with parameters *n* and *p*. Shorthand notation for this distribution is **b(*n*, *p*)**.

Probabilities for random variable *x* that has the distribution b(*n*, *p*) can be calculated from the following formula:

$$p(x)=\binom{n}{x}p^x(1-p)^{n-x} \text{ for } x = 0, 1, \ldots n,$$

$$\text{where } \binom{n}{x}=\frac{n!}{x!(n-x)!} \text{ .}$$

The mean and standard deviation of binomial random variable *x* can be calculated as follows:

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. In a random phenomenon with only two possible outcomes, traditionally what terms are used to label the two outcomes?

2. Give some reasons why the probability of success, *p*, for free throws is the same for each trial.

3. What is the probability of inheriting sickle cell disease for a child with two parents who are carriers? Why is this probability the same for each child in the family?

4. What is the formula for calculating the mean of a binomial random variable?

5. List the four conditions needed for a binomial distribution.

UNIT ACTIVITY:
INHERITING A GENETIC DISEASE

Each person has two copies of the eye color gene in their genome, one inherited from each parent. Let B stand for brown and b stand for blue. Brown is the dominant color, which means that a person with eye color genes Bb (B from mother and b from father) will have brown eyes. If brown eyed parents each have the recessive gene for blue eyes, b, then each of their children will have a 25% chance of having blue eyes, or genes bb.

In this activity, you will simulate the prevalence of blue eyes in families of four children, in which both parents have brown eyes but carry a recessive gene for blue eyes. We label the outcome as success if a child has blue eyes – not because blue eyes are better than brown, but because that is the outcome we are counting. In this case, the probability of success is $p = 0.25$.

For this activity, you will need to simulate selecting 30 samples of families with four children where both parents have brown eyes but have a recessive gene for blue eyes. If your instructor does not provide you with a method to simulate samples of four-children families, use one of the following methods:

> Method 1: You will need two coins of different denominations, say a nickel and a quarter. Let the nickel represent the gene the child inherits from the mother – if heads, the child receives gene b and if tails, gene B. Then let the quarter represent the gene that the child inherits from the father – if heads, the child receives gene b and if tails, gene B. Flip the two coins, if both land heads, then the child will inherit blue eyes and the outcome is labeled a success.
>
> Method 2: Generate a random number from the interval from 0 to 1. Excel's Rand() will work for this. If the number is 0.25 or less, call it a success. If the number is above 0.25 call it a failure.

You will generate outcomes for four children in a family. Each child is a new trial that ends in success, blue eyes, or failure, brown eyes. This process will be repeated 30 times. Save your data for use in Unit 28, Inference for Proportions.

1. In a copy of Table 21.1, record an *S* for each trial for which the child inherits blue eyes and an *F* for each trial for which the child inherits brown eyes. Let *x* be the number of successes in the family. Record the value of *x* for each family.

2. a. In a copy of Table 21.2, construct a relative frequency distribution for the number of successes in four trials. In the second column, enter the number of times 0, 1, 2, 3, and 4 successes were observed. In the third column, enter the proportion (relative frequency) of times 0, 1, 2, 3, and 4 successes were observed.

b. In the fourth column, enter the probabilities. To find these probabilities, you can use statistical or spreadsheet software, a graphing calculator, or an online binomial calculator. (You could also use the formulas from the Content Overview.)

3. a. Display the *x*-data in a histogram. Use proportion for the scaling on the vertical axis.

b. Represent the probability distribution with a probability histogram.

c. Compare your graphs from (a) and (b).

4. In a second copy of Table 21.2, combine the data from the class.

a. Display the *x*-data from the class in a histogram. Use proportion for the scaling on the vertical axis.

b. Compare your histogram in (a) to the probability histogram you drew for 3(b).

Sample	Trial #				
	1	2	3	4	*x*
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					

Table 21.1. Outcomes for each family of four children.

Number of Successes, *x*	Number of Families with *x* successes	Proportion of Families with *x* Successes	Theoretical Probability
0			
1			
2			
3			
4			

Table 21.2. Frequency, relative frequency, and probability table.

EXERCISES

1. Decide if each of the following situations fits the binomial setting. Give reasons for your answer in each case. If *x* has a binomial distribution, state the values of *n* and *p*.

a. Roll a fair die until it comes up six. Let *x* be the number of rolls needed to get the six.

b. Roughly 60% of Americans believe that extraterrestrial life exists on other planets. A random sample of 10 Americans is selected and during phone interviews are asked their opinion on extraterrestrial life. Let *x* be the number who say they believe in extraterrestrial life.

c. A deck of cards is shuffled. You draw a card and check to see if it is a red card and then put it aside. Then you draw a second card and check to see if it is a red card and put it on top of the first card drawn. You continue until you have drawn 5 cards. Let *x* be the number of red cards in the five that were drawn.

2. Blood banks are happy to receive blood donations from people who are type O+ (O positive) and O- (O negative). All people with Rh-positive blood can receive an O+ blood transfusion. The probability of having type O+ blood is 0.374. Suppose a random sample of four people show up during a blood drive to donate blood. Let *x* be the number of people with blood type O+.

a. Use the formulas given in the Content Overview to create a probability distribution table for *x*. Show your calculations. Round probabilities to two decimals.

b. In Unit 20, Random Variables, you learned how to calculate the mean of a discrete random variable: $\mu = \sum x \cdot p(x)$. Use this formula to compute the mean number of people with type O+ blood in a random sample of size four.

c. From this unit, you have learned another way to calculate the mean of binomial random variables: $\mu = np$. Use this formula to compute the mean. Compare your answer with your results in (b). Explain why there might be a slight discrepancy between your two answers.

d. Make a probability histogram to represent the probability distribution of *x*. Mark the mean, μ, on the horizontal axis.

3. Let w, x, and y be random variables with binomial distributions with $n = 5$ and probability of success p. Let $p = 0.2$ for w, $p = 0.5$ for x, and $p = 0.8$ for y.

a. Create probability distribution tables for w, x, and y. (Use software or tables to find the probabilities.)

b. Find the mean and standard deviation for each of the three distributions.

c. Draw probability histograms for the distributions of w, x, and y. Use the same scaling on all three histograms.

d. Compare the shapes of the three histograms. In two cases, the standard deviations of the random variables are the same (see solution (b)). Give an explanation based on the histograms for why that should be the case.

4. A drug manufacturer claims that its flu vaccine is 85% effective; in other words, each person who is vaccinated stands an 85% chance of developing immunity. Suppose that 200 people are vaccinated. Let x be the number that develops immunity.

a. What is the distribution of x?

b. What is the mean and standard deviation for x?

c. What is the probability that between 165 and 180 of the 200 people who were vaccinated develop immunity? (Hint: Use a normal distribution to approximate the distribution of x.)

REVIEW QUESTIONS

1. Decide if each of the following situations fits the binomial setting. Give reasons for your answer in each case. If *x* has a binomial distribution, state the values of *n* and *p*.

a. A random sample of 100 college students were asked: Do you routinely eat breakfast in the morning? Define *x* to be the number who respond Yes.

b. Classify the accidents each week into two categories, those involving alcohol and those not involving alcohol. Let *x* be the number of accidents involving alcohol in a week.

c. A particular type of heart surgery is successful 75% of the time. Five patients (who are not related) get this type of heart surgery. Let *x* be the number of successful surgeries.

2. People with O- blood are called universal donors because most people can receive an O- blood transfusion. The probability of having blood type O- is 0.066. Suppose a random sample of five people show up during a blood drive to donate blood. Let *x* be the number of people with blood type O-.

a. What is the probability that none of the five people has blood type O-?
Show your calculations.

b. What is the probability that exactly one of the five has blood type O-?
Show your calculations.

c. What is the probability that no more than one of the five people has blood type O-?

d. What is the probability that at least one of the five has blood type O-?

3. Let *x* be the number of children who have sickle cell disease in a family with three children, in which each parent is a carrier. Recall from the video that each child has a probability of 0.25 of having the disease.

a. Calculate the probability of each possible value of *x*. Round probabilities to two decimals.

b. Find the mean number of children who have the disease in families with 3 children where both parents are carriers. Interpret the value that you calculate in the context of this problem.

c. Draw a probability histogram for this distribution. Mark the location of the mean on your histogram.

4. According to the Centers for Disease Control and Prevention (CDC), 31% of American adults have high blood pressure. Suppose a random sample of 100 Americans is selected. Let *x* be the number with high blood pressure.

a. What is the distribution of *x*?

b. What is the mean and standard deviation of *x*?

c. What is the probability that fewer than 25 people in the sample have high blood pressure? Explain how you arrived at your answer.