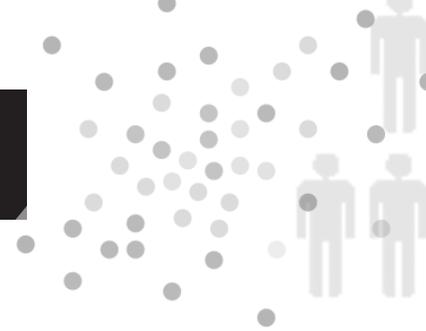


Unit 20: Random Variables



SUMMARY OF VIDEO

The video starts with the random phenomenon of a coin flip – more specifically, by examining what could happen if a fair coin is flipped four times. The sample space for this experiment is given below.

$$S = \{HHHH, HHHT, HHTH, HTHH, THHH, HHTT, HTHT, THHT, HTTH, TTHH, THTH, HTTT, THTT, TTHT, TTTH, TTTT\}$$

Each of these outcomes is equally likely. However, we are not interested in the actual outcomes, but rather on the number of heads in four flips. So, we'll define x as follows:

$$x = \text{number of heads in four flips of a coin}$$

We are now focusing on what statisticians call a random variable: the numerical outcome associated with the random phenomenon. The probability distribution of a random variable x tells us the values that the random variable can take on and the probabilities associated with each.

In our four coin tosses, the random variable x could equal 0, 1, 2, 3, or 4. It is a discrete random variable since it has a finite number of possible values. Although each of these values is possible, they are not equally likely as can be seen from the probability distribution in Table 20.1.

Value of x	0	1	2	3	4
Probability	0.0625	0.25	0.375	0.25	0.0625

Table 20.1. Probability distribution of x .

First, notice that the sum of all the probabilities is 1. Next, we need some notation – we use $p(x)$ to denote the probability associated with a particular value of x . So, for instance, $p(0)$ is the probability that the value of the random variable is 0. We represent this probability distribution graphically with the probability histogram in Figure 20.1.

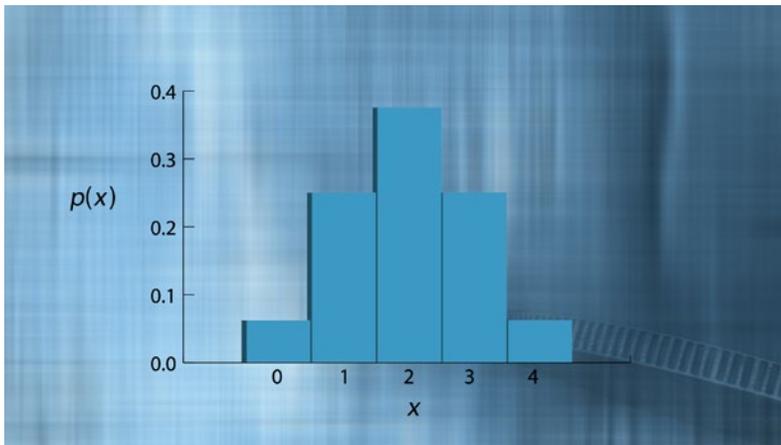


Figure 20.1. Probability histogram for x .

The horizontal axis shows the possible values of x , the bars have equal width, and the height of each bar represents the probability for that value. From the probability histogram, we can see that two is the most likely number of heads to come up in a string of four coin tosses. The histogram also tells us what we can expect from the data if we were to really run the experiment over and over again many times. However, instead of working with data, we can use the probability distribution.

The stakes are not very high when talking about coin tossing. But such calculations can be a matter of life and death when the events are critical equipment failures. On January 28, 1986, the space shuttle Challenger exploded shortly after liftoff. After the accident, President Ronald Reagan appointed a commission of experts to investigate its cause. Their eventual conclusion was that the accident was most likely caused by O-ring failure. O-rings sealed the field joints holding together the rocket boosters that would lift Challenger into orbit. The O-rings were supposed to contain hot, pressurized gases within the boosters. That morning, at least one failed to do so.

Could the disaster have been predicted? The first step in a probability analysis of field joint failure is to calculate the probability of failure in a single one of them. Under the Challenger flight conditions the probability of failure of a particular field joint was 0.023. That means that the probability of success of an individual field joint would be 0.977. However, there were six field joints. For the whole system to succeed, all six field joints had to succeed – in other words, zero failures. We form a probability distribution table for $x =$ the number of failures in the six field joints.

x	0	1	2	3	4	5	6
$p(x)$	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$

Table 20.2. Probability distribution for x , the number of field joint failures.

We want to determine $p(0)$, the probability of zero failures. We will need to use the Multiplication Rule:

If A and B are independent events, then $P(A \text{ and } B) = P(A)P(B)$

Remember, the probability that each field joint would succeed was 0.977 and there are six field joints that all need to succeed – no failures allowed. We assume that the field joints are independent. Failure of one field joint should not affect the likelihood that another fails.

$$\begin{aligned} p(0) &= P(\text{all six field joints succeed}) \\ &= (0.977) (0.977) (0.977) (0.977) (0.977) (0.977) \\ &\approx 0.87 \text{ or around } 87\% \end{aligned}$$

It is possible to complete all the other individual probabilities, but for now we will use the Complement Rule to calculate the likelihood of there being at least one field joint failure.

$$P(\text{at least one failure}) = 1 - p(0) = 1 - 0.87 = 0.13$$

So while the probability of an individual field joint failing is pretty low, the probability of at least one of the six failing is rather high, especially considering that astronauts' lives are at stake.

Over two hundred improvements were made to the next space shuttle after the Challenger disaster. NASA successfully launched shuttles almost 100 more times before retiring the space shuttle program in 2011. Of course a complex, state of the art technology like the shuttle system could never reduce the risk of failure to zero – and in fact, another disaster occurred in 2003 when the space shuttle Columbia disintegrated on re-entry to Earth's atmosphere. The Challenger and Columbia are tragic reminders of the risks of space exploration and the need for continued rigorous analysis.

STUDENT LEARNING OBJECTIVES

- A. Understand the concept of a random variable.
- B. Be able to differentiate between continuous and discrete random variables.
- C. Know how a random variable is characterized by its probability distribution.
- D. Be able to create probability distributions for discrete random variables in some simple situations, such as when all outcomes are equally likely.
- E. Be able to draw a probability histogram to represent a given probability distribution.
- F. Know how to calculate the mean and standard deviation of a discrete random variable given its probability distribution.

CONTENT OVERVIEW

Examples of random phenomena abound. Walk up to students at your school or campus and ask them to rate their happiness: Unhappy, So-So, Happy. Show up at Pete and Gerry’s farm, randomly select a hen and weigh it. As these two examples illustrate, sometimes the outcomes of a random phenomenon are categories and other times numbers. Next, we create random variables by mapping the outcomes of random phenomena to numbers.

Start with the example of the Happiness Survey that was actually given to residents of Somerville, Massachusetts. (Refer back to Unit 13, Two-Way Tables.) We can create a random variable x as follows:

$$x = \begin{cases} 0, & \text{if Unhappy} \\ 1, & \text{if So-So} \\ 2, & \text{if Happy} \end{cases}$$

Using the data from Somerville’s Happiness Survey, we can assign a probability, $p(x)$, to each of the numeric outcomes of x . The numeric outcomes of the random variable x together with their probability assignments form the probability distribution of x shown below.

x	0	1	2
$p(x)$	0.048	0.369	0.583

Table 20.3. Probability distribution for x .

The probabilities in Table 20.3 satisfy two properties required of all probability distributions:

$$0 \leq p(x) \leq 1$$

$$\sum p(x) = 1$$

In other words, the values for $p(x)$, 0.048, 0.369, and 0.583, are all between 0 and 1 and they sum to 1: $0.048 + 0.369 + 0.583 = 1$.

We can represent the probability distribution $p(x)$ graphically with the probability histogram in Figure 20.2. The possible values for x are on the horizontal axis and probability is on the vertical axis.

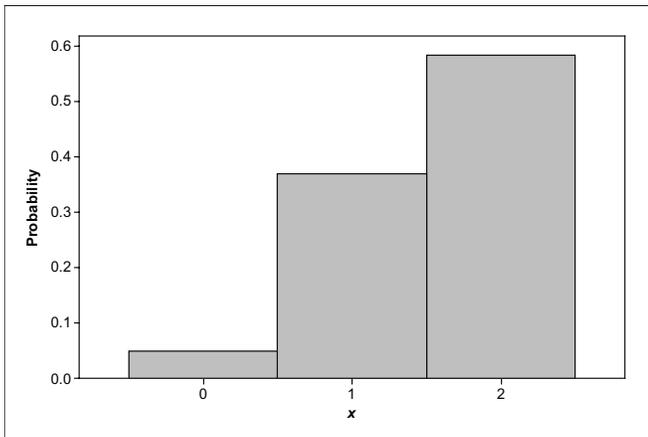


Figure 20.2. Probability histogram representing $p(x)$.

Next, consider the random phenomenon of the weight of 7½-week old hens. In this case, the outcome is already a number and so, we can define a new random variable w = hen weight. Notice that w takes on values in the interval from the weight of the smallest hen to the weight of the largest hen. An interval contains too many numbers to list them all – so, we can't assign probabilities to each possible weight as we did with x in Table 20.3. From the histogram of data on hen weights (Figure 20.3), we find that a normal density curve is a good approximation for the distribution of w .

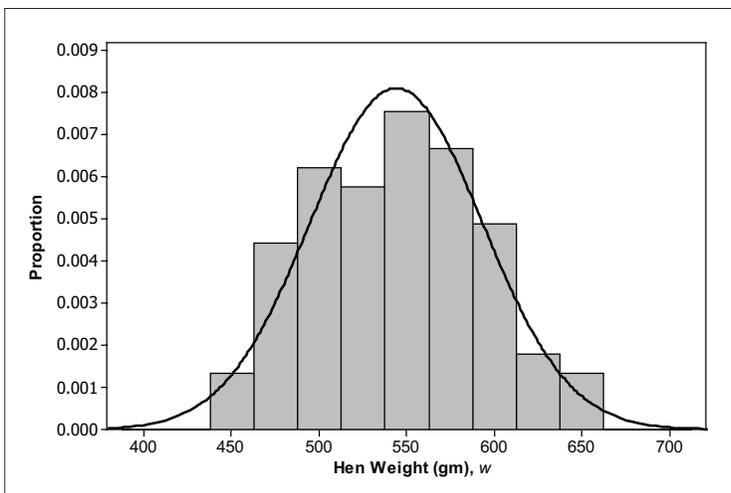


Figure 20.3. Approximating the distribution of w with a normal density curve.

The two random variables that we have looked at – x , happiness rating, and w , hen weight – are examples of two different types of random variables. Since it is possible to list all possible outcomes for x , x is called a discrete random variable. However, w takes on values in an interval – there are too many possible outcomes to list them all; w is an example of a continuous random variable.

Let's take a look at another discrete random variable. Table 20.4 gives a probability distribution for family size, y , in the United States. (Although there are some families that are bigger than 8, the likelihood is so small that we ignored them in this probability distribution model.)

y	1	2	3	4	5	6	7	8
$p(y)$	0.15	0.23	0.19	0.23	0.12	0.05	0.02	0.01

Table 20.4. Distribution of U.S. family size.

From the probability distribution for y , we can find the probability that a randomly selected family will consist of at least two people using the Complement Rule:

$$P(y \geq 2) = 1 - P(y < 2) = 1 - p(1) = 0.85$$

We can use the Addition Rule to find the probability that a randomly selected family will have two to four members:

$$P(2 \leq y \leq 4) = P(y = 2 \text{ or } y = 3 \text{ or } y = 4) = p(2) + p(3) + p(4) = 0.65$$

We can do more with probability distributions of discrete random variables than just compute probabilities. We can calculate the random variable's mean and standard deviation. All that's needed are the following formulas:

$$\mu = \sum x \cdot p(x)$$

$$\sigma^2 = \sum (x - \mu)^2 \cdot p(x) \quad \text{and} \quad \sigma = \sqrt{\sigma^2}$$

First, we calculate the mean family size. For the calculation, multiply each outcome by its probability and then sum the results:

$$\mu = 1 \times 0.15 + 2 \times 0.23 + 3 \times 0.19 + 4 \times 0.23 + 5 \times 0.12 + 6 \times 0.05 + 7 \times 0.02 + 8 \times 0.01$$

$$\mu = 3.22$$

Next, we calculate the variance. For this calculation, multiply each outcome's squared deviation from the mean by its probability and then sum the results:

$$\sigma^2 = (1 - 3.22)^2(0.15) + (2 - 3.22)^2(0.23) + \dots + (8 - 3.22)^2(0.01)$$

$$\sigma^2 \approx 2.512$$

To get the standard deviation, take the square root of the variance:

$$\sigma \approx \sqrt{2.512} \approx 1.58$$

Now, we return to the problem of determining probabilities for a continuous random variable, such as hen weight. For that we need the probability density curve. We will assume hen weight is normally distributed with mean 544 grams and standard deviation 49 grams. We can find probabilities by calculating areas under the density curve – for these calculations we need technology or must convert to z-scores and use the standard normal table. For example, suppose that we want to find the probability that a randomly selected hen weighs between 500 and 600 grams, $P(500 \leq w \leq 600)$. Figure 20.4 shows the area under the normal density curve over the interval from 500 to 600. That area gives the probability we are seeking.

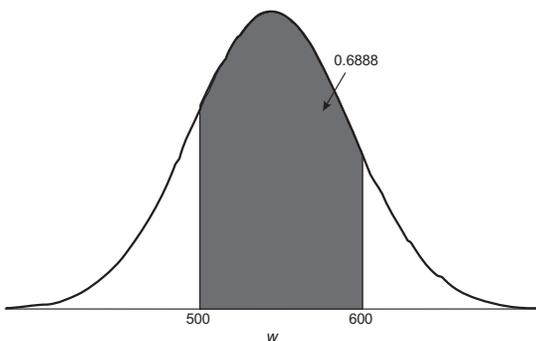


Figure 20.4. Shaded area under normal density curve.

We can use software to give us this area (as shown on Figure 20.4) or we can convert the endpoints of the interval into z-scores and use the standard normal table to get the probability. (This method was introduced in Unit 8, Normal Calculations.)

$$z = \frac{500 - 544}{49} \approx -0.90 \quad \text{and} \quad z = \frac{600 - 544}{49} \approx 1.14$$

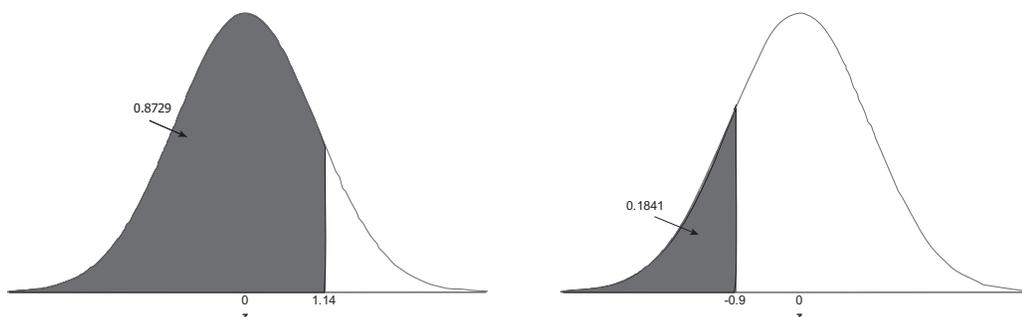


Figure 20.5. Areas to the left of $z = 1.14$ (a) and $z = -0.90$ (b).

We determine areas under the standard normal curve to the left of our two z-values. These areas are 0.8729 (a) and 0.0841 (b), corresponding to $z = 1.14$ and $z = -0.90$, respectively. By

subtracting these two areas, we get the area under the standard normal curve over the interval from -0.90 and 1.14:

$$0.8729 - 0.1841 = 0.6888.$$

This gives us the same value that we obtained using software in Figure 20.4.

Statisticians can also compute the mean and variance of a continuous random variable. All that's needed is a formula for the probability density curve and some calculus. The need for calculus puts computing the mean and variance of continuous random variables outside of the scope of this course.

KEY TERMS

A **random variable** is a variable whose possible values are numbers associated with outcomes of a random phenomenon.

A **discrete random variable** can take on only a countable number of distinct values – in other words, it is possible to list its possible values. Any random variable that can take on only a finite number of values is a discrete random variable. A **continuous random variable** can take on values in an interval.

The **probability distribution** of a discrete random variable x is a list of its possible values together with the probabilities associated with those values. The probability distribution is a model for the population distribution. The random variable's **mean** and **standard deviation** are computed as follows:

$$\mu = \sum x \cdot p(x)$$

$$\sigma^2 = \sum (x - \mu)^2 \cdot p(x); \sigma = \sqrt{\sigma^2}$$

THE VIDEO

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

1. For the tossing-four-coins experiment, x was defined to be the number of heads. What are the possible values for x ?
2. Given the probability distribution for x , what is the sum of the probabilities?
3. Based on the probability histogram, what is the most likely number of heads to come up in a string of four coin tosses?
4. What was the most likely cause of the Challenger disaster?
5. What rule of probability was used to find $p(0)$, the probability that none of the six independent field joints failed?
6. What is the relationship between the probability that none of the field joints failed and the probability that at least one of the field joints failed?

UNIT ACTIVITY:

CHILDREN IN HOUSEHOLDS AND SCHOOL LUNCH

The U.S. government collects data on many variables having to do with households. In this activity, you will examine issues related to the number of children in households and participation in free school lunch programs.

1. Whether or not a household has children is a random phenomenon. From this random phenomenon we define random variable μ as follows:

$$u = \begin{cases} 0, & \text{if no children in household} \\ 1, & \text{if at least one child in household} \end{cases}$$

The probability distribution of μ is given below. We have chosen to express the table with columns for the values of μ and corresponding values of $p(\mu)$. This will help with computation of mean and standard deviation later in this activity.

u	$p(u)$
0	0.586
1	0.432

Table 20.5. Probability distribution of μ .

- Draw a probability histogram for the probability distribution of u .
- The balance point of the probability histogram is the mean, μ_u , of the random variable u . Based on the probability histogram from (a), do you think the balance point of this distribution is less than 0.5, at 0.5, or above 0.5? Explain.
- Next, we walk you through one method of calculating $\mu_u = \sum u \cdot p(u)$. Make a copy of Table 20.6. Enter the products of the entries from the first two columns into the third column. Then compute the sum of the entries in the third column to get μ_u .

u	$p(u)$	$(u)(p(u))$
0	0.586	
1	0.432	
Sum =		

Table 20.6. Method for calculating μ_u .

d. You should have gotten $\mu_u = 0.432$ for (c). (If not, go back and redo part (c).) Next, we walk you through one method of calculating the variance: $\sigma_u^2 = (u - \mu_u)^2 \cdot p(u)$. Then you can calculate the standard deviation by taking the square root of the variance. Make a copy of Table 20.7.

- In column three, enter the squared difference between the values of μ and μ_u
- In column four, multiply the squared deviations by their corresponding probabilities.
- Sum column four to get σ_u^2 .
- Take the square root of σ_u^2 to get the standard deviation.

u	$p(u)$	$(u - 0.432)^2$	$((u - 0.432)^2)(p(u))$
0	0.586		
1	0.432		
Sum =			

Table 20.7. Method for calculating σ_u^2 .

2. In households with children, some of the school age children ate school lunches and others did not. Hence, we have another random phenomenon. Define a new random variable v as follows:

$$v = \begin{cases} 0, & \text{if none of the children in household ate school lunch} \\ 1, & \text{if at least one child in household ate school lunch} \end{cases}$$

a. Table 20.8 contains data collected from a U.S. government survey on random variable v . Calculate the proportions for the outcomes of v and enter them as estimates in the Probability column. Round probabilities to three decimals.

v	Frequency	Probability
0	32,491	
1	74,690	
Total		

Table 20.8. Data on random variable v .

b. Draw a diagram that represents all households broken down by whether or not there are children in the household, and then broken down further by whether or not at least one child in the household ate school lunch. Determine the proportion of all households that had at least one child who ate school lunch.

3. To ensure that children don't go hungry, some of the children who eat school lunches get free lunches. Define random variable x as follows:

$$x = \begin{cases} 0, & \text{if household in which children ate school lunch, none got it free} \\ 1, & \text{if household in which children ate school lunch, at least one got it free} \end{cases}$$

The probability distribution for x is given in Table 20.9.

x	$p(x)$
0	0.432
1	0.568

Table 20.9. Probability distribution of x .

From question 2, we know that 74,690 households in the survey had a least one child who ate school lunch. Estimate the number of those households in which at least one child participated in the free lunch program.

4. Next, for those households with at least one child participating in the free lunch program, some households have multiple children participating in the program. Define random variable y as follows:

$$y = \text{the number of children in the household participating in the free lunch program}$$

The probability distribution for y is given in Table 20.10.

y	$p(y)$
1	0.394
2	0.337
3	0.176
4	0.063
5	0.019
6	0.007
7	0.002
8	0.001

Table 20.10. Probability distribution of y .

- Draw a probability histogram representing the probability distribution of y .
- Visually estimate the mean μ_y as the balance point of your histogram in (a). Does the balance point appear closer to 1, 2, or 3?
- Calculate the mean of y , μ_y , using the method outlined in 1(c). (Round your answer to three decimals.)
- Calculate the standard deviation of y , σ_y , using the method outlined in question 1(d).

EXERCISES

1. The random variable x , defined below, gives the average grade of 12th grade students in U.S. high schools. The probability distribution for x is given in Table 20.11.

$$x = \begin{cases} 4, & \text{if A average} \\ 3, & \text{if B average} \\ 2, & \text{if C average} \\ 1, & \text{if D average} \end{cases}$$

x	1	2	3	4
$p(x)$	0.01	0.15	0.49	0.35

Table 20.11. Probability of grade averages.

- Find $P(x \geq 3)$, the probability that a randomly selected student has a B or better average.
- Find $P(x < 3)$, the probability that a randomly selected student has a below B average. How is this probability related to your answer to (a)?
- Make a probability histogram for the distribution of x . What does your graphic display tell you about the distribution of average grades?

2. The U.S. government collects data on many variables having to do with households. Let x = the number of children under 15 in a household. The probability distribution for x is shown in Table 20.12.

x	0	1	2	3	4	5	6	7	8
$p(x)$	0.468	0.200	0.199	0.087	0.031	0.009	0.003	0.002	0.001

Table 20.12. Probability distribution for x .

- What is the probability that a randomly selected household has at least one child under 15?
- What is the probability that a randomly selected household has between two and four children under 15? In other words, find $P(2 \leq x \leq 4)$.
- Draw a probability histogram that represents the probability distribution shown in Table 20.12. Describe the shape of the histogram.

d. What is the mean number of children under 15 per U.S. household? Show your calculations.

3. A DVD manufacturer receives certain components in lots of four from two different distributors. Let x and y represent the number of defective components in each lot from Distributor 1 and Distributor 2, respectively. The probability distributions for x and y are given in Tables 20.13.

x or y	$p(x)$	$p(y)$
0	0.40	0.15
1	0.33	0.70
2	0.18	0.15
3	0.05	0
4	0.04	0

Table 20.13. Probability distributions of x and y .

- Draw probability histograms for the probability distributions of x and y (Table 20.13).
- Find the mean number of defects in lots of four for both distributors. In other words, find the mean values of both x and y .
- Find the standard deviations of x and y . Show your calculations.
- Given the results in (b) and (c), which distributor should the DVD manufacturer rely on more heavily? Explain.

4. Assume that the distribution of weight for 7½-week old hens is normally distributed with mean 544 grams and standard deviation 49 grams. Let w = weight of a randomly selected hen.

- Sketch normal density curve representing the distribution of w .

Use technology or the standard normal table to find the probabilities in (b) – (d). On a copy of the normal density curve that you sketched for (a), shade the area under the curve that represents each probability.

- $P(w < 500)$
- $P(w \geq 580)$
- $P(500 \leq w \leq 580)$

REVIEW QUESTIONS

1. Let x represent the number of broken eggs in a randomly selected carton of 12 eggs. The probability distribution for x is given in Table 20.14. (The probability of 5 or more broken eggs in a carton is so small that this possibility is not included in the probability model.)

x	0	1	2	3	4
$p(x)$	0.73	0.15	0.07	0.03	?

Table 20.14.

- Determine the value of $p(4)$. Interpret this value in the context of broken eggs.
- Calculate the probability of randomly selecting a carton of eggs and finding that two or more of the eggs are broken.
- Draw a probability histogram for the distribution of x .
- Calculate μ = the mean number of broken eggs per carton. Interpret the meaning of μ in the context of broken eggs.

2. In each of the situations that follow decide if the random variable is discrete or continuous. Justify your answer.

- The number of unbroken Cheerios in a 9-ounce box of Cheerios.
- The time it takes to complete an exam.
- The hourly rate for a worker at a fast-food restaurant.
- The length of a fish.

3. Suppose you toss a fair coin three times.

- List the outcomes in the sample space.

Assume that the outcomes in the sample space that you determined for (a) are equally likely. Give the probability distribution for the random variables in (b – d). Then calculate their means and standard deviations.

b. Let x = the number of heads.

c. Let y = the absolute difference in the number of heads and tails: in other words, $|\text{number of heads} - \text{number of tails}|$.

d. Let w = the sum of the number of heads and tails.

4. A study of White-throated Sparrows indicates that their wingspan is normally distributed with mean 70 mm and standard deviation 3 mm. Let w = wingspan of a randomly selected White-throated Sparrow. Use technology or the standard normal table to find the probabilities in (a) – (c).

a. $P(w < 68)$

b. $P(w \geq 75)$

c. $P(68 < x < 75)$