

# Data Sets and Distributions

Accelerated 6  
Unit 8

## 11 Synthesis

What are the advantages and disadvantages of visualizing a data set as a dot plot?

Use the examples if they help with your thinking.

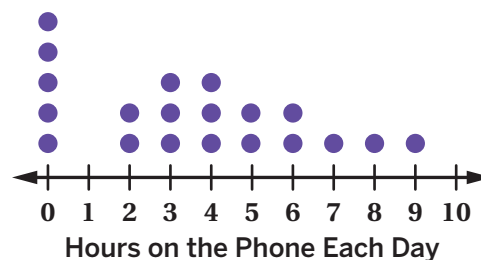
Advantages:

Disadvantages:

### Data Set

0, 3, 4, 0, 0, 0, 4, 5, 6, 2, 5, 6, 2, 9, 3, 3, 7, 4, 0, 8

### Dot Plot



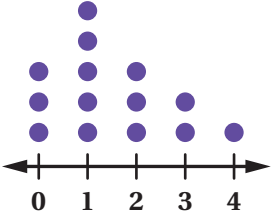
## Summary

A **statistical question** is a question that needs more than one piece of data to answer it.

Here is an example.

Statistical Question	Non-Statistical Question
<p>"Which classroom in your school has the most books?"</p> <p>You need to count all of the books in each classroom of your school to answer this question.</p>	<p>"How many books are in your classroom?"</p> <p>This is not a statistical question, because you only need to count the books in one classroom.</p>

Organizing data into lists or **dot plots** is helpful for describing a data set, seeing patterns in the data, and answering questions. For example, consider the data about the number of siblings a group of sixth graders have:

List	0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4	Lists allow you to see all of the data. Lists can be used for both numerical or categorical data.												
Dot Plot	 <table><thead><tr><th>Number of Siblings</th><th>Frequency (Number of Dots)</th></tr></thead><tbody><tr><td>0</td><td>3</td></tr><tr><td>1</td><td>5</td></tr><tr><td>2</td><td>3</td></tr><tr><td>3</td><td>2</td></tr><tr><td>4</td><td>1</td></tr></tbody></table>	Number of Siblings	Frequency (Number of Dots)	0	3	1	5	2	3	3	2	4	1	Dot plots are a visual representation of numerical data and allow you to compare multiple data sets.
Number of Siblings	Frequency (Number of Dots)													
0	3													
1	5													
2	3													
3	2													
4	1													

## 12 Synthesis

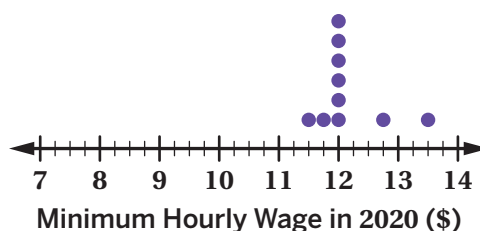
What do the center, spread, and shape tell you about a dot plot?

Use the example if it helps with your thinking.

Center:

Spread:

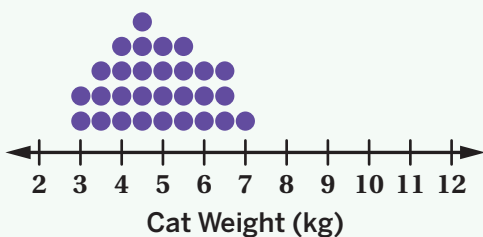
Shape:



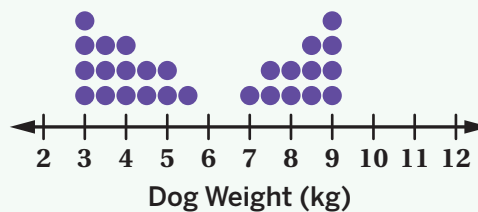
## Summary

**Center**, **spread**, and **shape** are helpful ways to describe the distribution of data on a dot plot. The center is a single value in the middle of a data set that represents a typical value. The spread describes how alike or different the values in a distribution are, often in relationship to the center. We can describe the shape of a distribution (or what the distribution looks like) using formal or informal language.

Here are some examples.



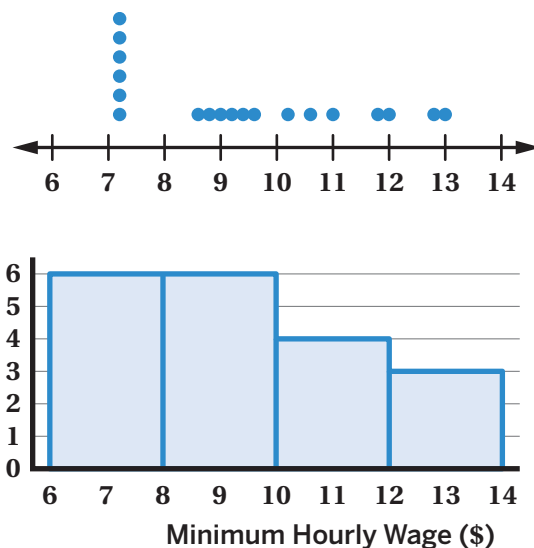
- The data is symmetric.
- There are no gaps in the data.
- The data is centered around 4.5 kilograms.
- Most of the data is clumped together.
- The data is shaped like a mountain with the most dots on 4.5 kilograms.



- There is a gap in the data between 5.5 and 7 kilograms.
- The data is spread out in two clumps, one with a peak at 3 kilograms and another peak at 9 kilograms.
- The center of the data is around 6 kilograms.
- The data is shaped like two triangles with the most dots on 3 kilograms and 9 kilograms.

## 10 Synthesis

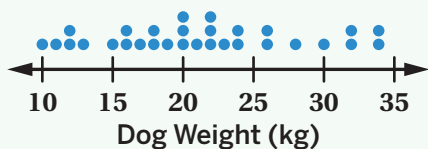
What is important to remember about a histogram?



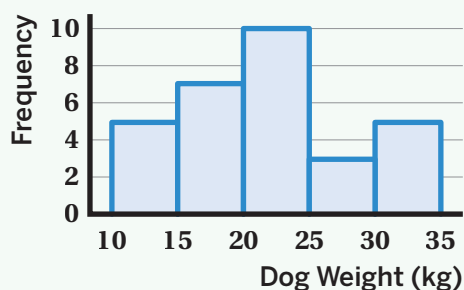
## Summary

Both dot plots and **histograms** can be used to visualize numerical data. Here is an example of a data set of the weights of 30 dogs presented in a dot plot and in a histogram.

Dot Plot



Histogram

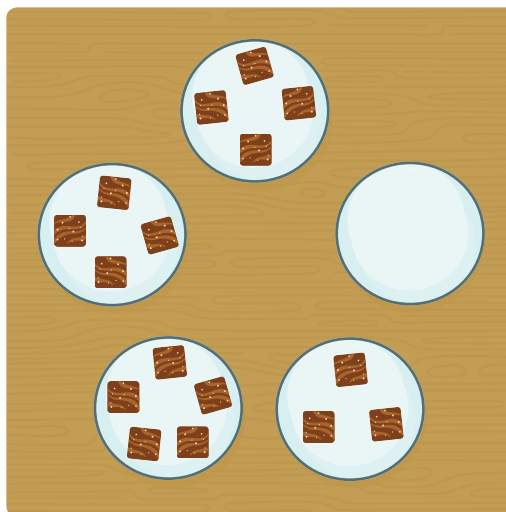


In a histogram, data values are grouped into bins that cover a range of values, and each **bin** has the same width. The height of each bar represents the total number of values in that range, including the left boundary (least value) but excluding the right boundary (greatest value). For example, the height of the tallest bar, from 20 to 25, represents weights of 20 kilograms up to (but not including) 25 kilograms.

## 12 Synthesis

How can you determine the mean of a data set?

Use the example if it helps with your thinking.



## Summary

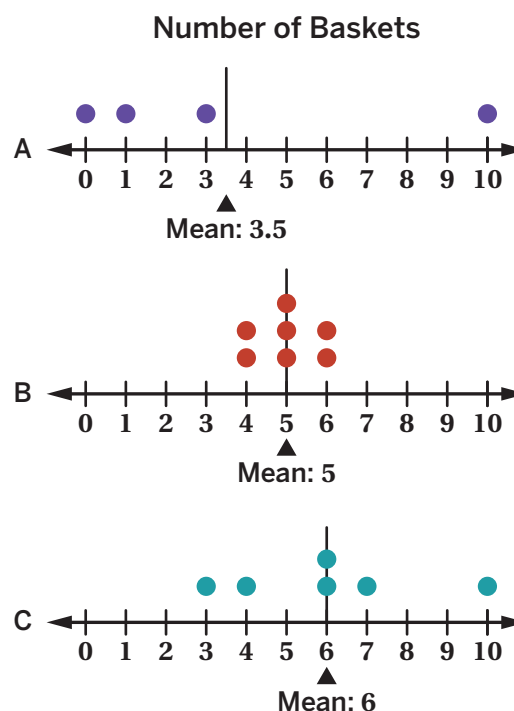
A **statistic** is a single number that measures something about a data set. A measure of *center* is a single number that summarizes all of the values in a data set. One way to measure the center of a data set is by determining the **mean**, or *average*, of all the data values. You can think of it as “determining an equal share.”

For example, suppose this data set represents how many liters of water are in 5 bottles: 1, 4, 2, 3, 0. To calculate the mean, you first add up all of the values to determine the total (10 liters), then divide that sum by the number of values (5 bottles). This example can be represented by the expression  $(1 + 4 + 2 + 3 + 0) \div 5$ , or  $10 \div 5$ . So, the mean amount of water in the 5 bottles is 2 liters (per bottle). The mean is a whole number in this example, but it is possible for the mean to be a decimal number.

## 12 Synthesis

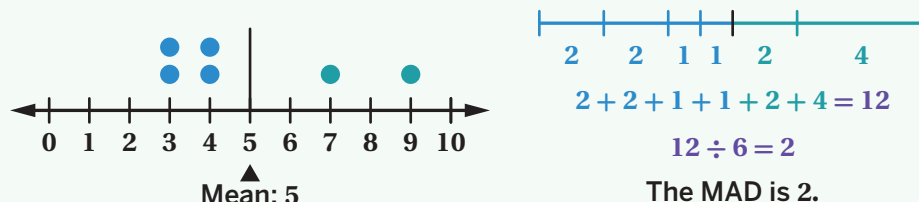
How does the mean absolute deviation (MAD) help you compare data sets?

Use the examples if they help with your thinking.



## Summary

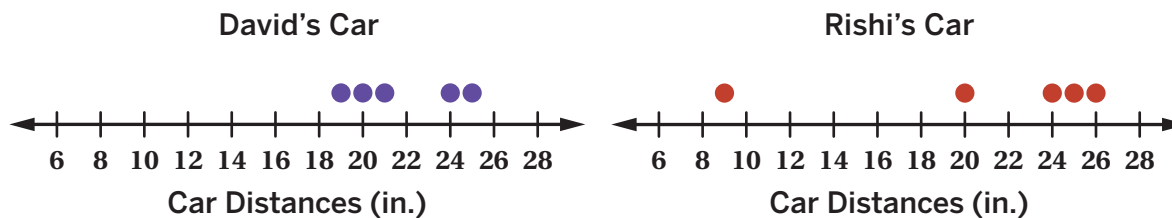
The **mean absolute deviation (MAD)** of a data set is a single number that describes how spread out the values in the data set are around its center. The MAD is calculated by determining the average of the distances between each data value and the mean. In other words, it is calculated by determining the *mean* of the *absolute deviations*.



In a previous lesson, you explored the mean as a measure of center. The mean absolute deviation is an example of a measure of spread. A measure of spread is a way to measure the consistency of the values in a data set. The smaller the value of the MAD, the less spread out the data points are around the mean, which indicates the data values are more consistent. The larger the MAD, the more spread out the data points are around the mean, which indicates the data values are less consistent.

## 11 Synthesis

Describe how to determine the median of a data set. Use the examples if they help with your thinking.



## Summary

Another measure of *center* that can be used to describe a data set is called the **median**. The median is literally the “middle” value in a data set when the values are listed in order from least to greatest (or greatest to least). Half of the data values have values less than or equal to the median, and half of the data values have values greater than or equal to the median.

To determine the median from an ordered representation of the data, such as a list or a dot plot, you repeat a process of eliminating the pairs of least and greatest values.

Here are some examples.

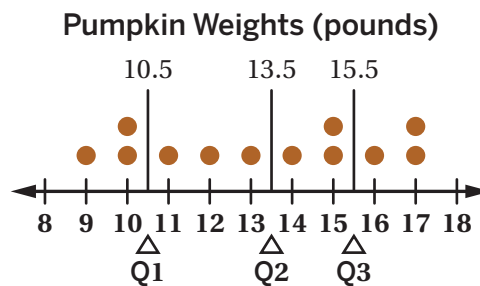
<b>Odd Number of Values</b>	<del>0</del> <del>1</del> <del>1</del> <u>2</u> <del>2</del> <del>4</del> <del>5</del> <p>Once all pairs have been eliminated, only one value remains in the middle, making it the median.</p> <p>Median: 2</p>
<b>Even Number of Values</b>	<del>0</del> <del>1</del> <del>1</del> <u>1</u> <u>2</u> <del>2</del> <del>4</del> <del>5</del> <p>Once all pairs have been eliminated, two values remain. Their average is the median.</p> <p><math>(1 + 2) \div 2 = 1.5</math></p> <p>Median: 1.5</p>

## 12 Synthesis

Discuss both questions with a partner.  
Then select *one* and write your response.

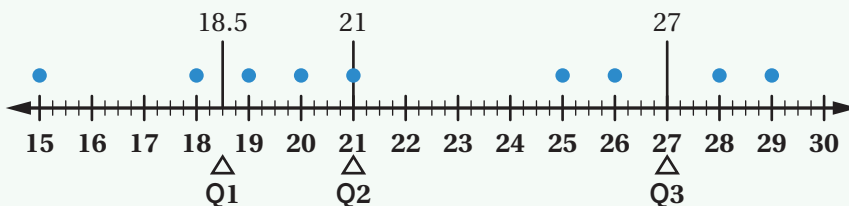
Use the example if it helps to show  
your thinking.

- How do quartiles relate to the middle half of a data set?
- How can you determine the value of the quartiles for a data set?



## Summary

**Quartiles** divide a data set into four equal sections to help us identify and describe the middle half of a data set.



The *first quartile* ( $Q1$ ) is the median of the lower half of the data set.

The *second quartile* ( $Q2$ ) is the median of the entire data set.

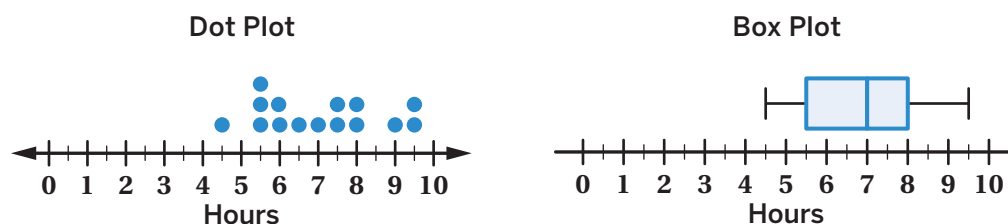
The *third quartile* ( $Q3$ ) is the median of the upper half of the data set.

You can determine the value of the quartiles by splitting the entire data set in half and then splitting the halves again. The middle half is all the data points that are between  $Q1$  and  $Q3$ . Representations such as dot plots are helpful for identifying quartiles to describe data sets.



## Synthesis

10. Here is a dot plot, a box plot, and several statistics for the car data.



Statistics			
Median: 7 hours	Number of Data Points: 15	Range: 5 hours	IQR: 2.5 hours

Which statistics are more visible in the dot plot? In the box plot? Why do you think that is?

## Summary

You can create a **box plot** to visualize a data set. While a box plot shows the same data as a dot plot, it gives us new information about the data. Rather than showing every data point, a box plot separates the data into quarters by plotting the *minimum*, *Q1*, *median* (*Q2*), *Q3*, and the *maximum*.

We can use box plots to describe the spread of the data in two ways.

- The **range** represents the difference between the maximum *and* minimum values of a data set. It gives you an idea of the overall spread of the data.

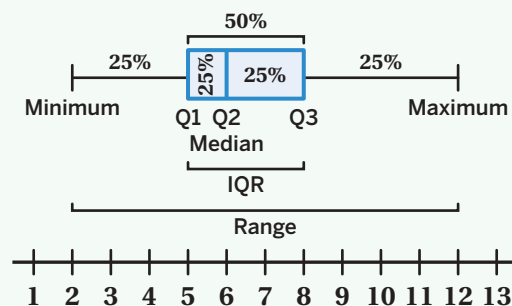
$$\text{Range: } 12 - 2 = 10$$

- The **interquartile range (IQR)** represents the range of the middle 50% of the data. It gives you an idea of how spread out the middle of the data is.

$$\text{IQR: } 8 - 5 = 3$$

Box plots do not show you how many data points are in each set, or the values of any individual data points, except the minimum and maximum. They do, however, make it possible to compare minimums, maximums, and three *quartiles* (*Q1*, *Q2*, and *Q3*).

Min.	Q1	Median	Q3	Max.
2	5	6	8	12



## 12 Synthesis

What are some advantages and disadvantages of using samples to answer a question about a population?

Sample	Population
The 50 students	The students in the school

## Summary

To answer a question about a population of data, it is sometimes unreasonable to collect data from the entire **population**. Instead, data is often collected from a **sample** of the population.

- A *population* is a set of people or things that we want to study.
- A *sample* is part of a population.

The sample you choose should be large enough to be able to draw conclusions about the population.

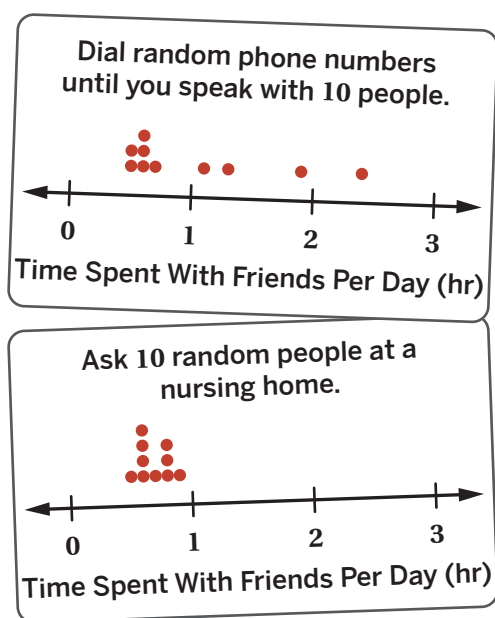
Here are some examples of populations and samples.

Population	Sample
All of the people who watch basketball.	The people at a basketball game.
All 7th grade students in your school.	The 7th graders in your school who are in a band.
All oranges grown in the U.S.	The oranges in your local grocery store.

## 10 Synthesis

Explain how collecting a representative sample or an unrepresentative sample can affect someone's understanding of a population.

Use the examples if they help with your thinking.



## Summary

*Samples* are useful when a population is too large to survey or measure. Depending on the strategy you use to sample, it is more or less likely that the sample will be **representative** of the population. Some samples are not good representations of the population.

A *representative sample* has a distribution that closely resembles the distribution of the population. Representative samples are useful for making predictions about the whole population.

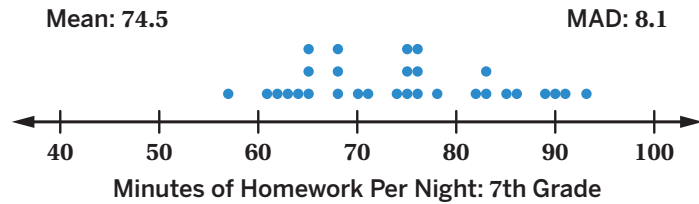
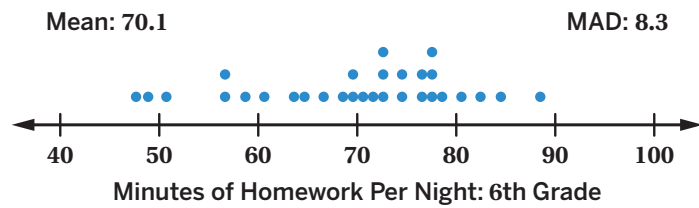
For example, if you were curious about all middle school students' favorite sport to play, the *population* would be all middle schoolers. A representative sample of this population might be randomly selecting 5 students from each class or 15 students from each grade to ask. A sample that is *not* representative of this population would be asking students in the tennis club because their responses might lead someone to believe that tennis is the favorite sport among all middle schoolers.



## Synthesis

8. How can you use the MAD to determine how different two populations are?

Use the data from Median Middle School if it helps with your thinking.



## Summary

Comparing two individuals or objects is fairly straightforward. For example, you can answer the question, “Which 7th grader is taller?” by measuring the heights of two 7th grade students and comparing them directly. But comparing two populations or two data sets requires some additional analysis. For example, to answer the question, “Do elementary school students sleep longer than middle school students per night?” you need to use the measures of center and the variability of both samples.

When you’re trying to decide whether two sets of data are very different from each other, you can look for the amount of overlap on their graphs, or you can calculate the difference in their *means* and express it as a multiple of the *mean absolute deviation (MAD)*.

Generally, we can say that two data sets are very different from each other if the difference in their means is more than 1 times the larger MAD.

## Synthesis

10. Select one question to answer:

- a How can statistics and sampling help us make sense of topics like air quality?
- b What other questions might you ask to investigate this topic more?

## Summary

Sometimes you can use samples to ask questions about our world. Statistics and sampling, along with visual representations of the data, allow you to make sense of real-world topics such as the air quality in different geographical locations. The more we understand the world around us, the more we can take action to improve it.

Let's say you want to compare the average daily high temperatures of Chicago and Mexico City throughout a whole year. The mean and median tell one story about the temperature in each location. Including measures of spread, like IQR and MAD, show a fuller picture of how the temperature in each location compares. Using representations, such as box plots, also allows for a visual comparison. In this instance, while Chicago and Mexico City may have similar means and medians, looking at the spread of the data would tell a much different story.

## 12 Synthesis

Explain how you can use a sample space to help you determine the *probability* of an event.



## Summary

The **probability** of an event is a number that represents how likely the event is to occur. One way to calculate the probability is to look at all of the possible *outcomes* for an *experiment*, which is known as the **sample space**.

When all of the outcomes are equally likely, the probability of an event is a ratio.

$$\frac{\text{number of favorable outcomes}}{\text{total possible number of outcomes}}$$

Probabilities are numbers between 0 and 1 written as fractions, decimals, or *percentages*. A probability of 1 means the event will always happen. A probability of 0 means the event will never happen.

Here are several examples of events and their probabilities.

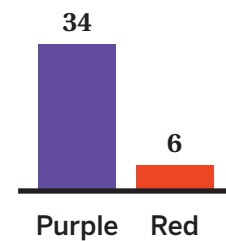
Example	Probability
Picking a green marble out of a bag that contains only red and yellow marbles.	0
Rolling a 1 on a number cube.	$\frac{1}{6}$ (or equivalent)
Tossing a coin and it landing heads up.	50% (or equivalent)
Picking a yellow marble from a bag of 10 marbles, where 8 of the marbles are yellow.	0.8 (or equivalent)
Picking a green marble in a bag that only contains green marbles.	1

## 10 Synthesis

Describe how you can use results from a repeated experiment to make predictions.

Use the results shown if that helps with your thinking.

**Total Picks: 40**



## Summary

In situations where you don't know the sample space, you can use data from experiments and proportional reasoning to predict what the *sample space* will be. The number of repeated experiments can affect how accurate your prediction may be.

There is a mystery bag that has 8 marbles in it, 2 red marbles and 6 blue marbles.

Here are the results from picking a marble out of the bag 10 times.

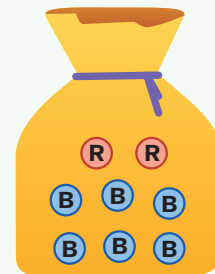
Only 1 out of the 10 marbles was red, so the *constant of proportionality* is 0.1.

Multiplying 0.1 times the number of marbles in the bag (8), may lead someone to predict that there is only  $0.1 \cdot 8 = 0.8$  or 1 red marble in the bag.

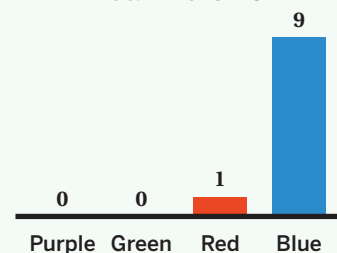
Here are the results from 50 picks.

The constant of proportionality 0.24 times the number of marbles in the bag is  $0.24 \cdot 8 = 1.92$  which is close to 2.

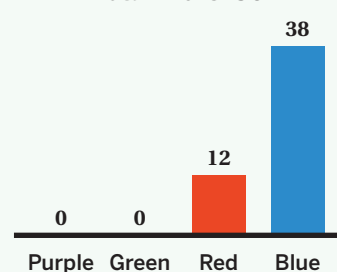
This is a more accurate prediction of the number of red blocks in the bag.



**Total Picks: 10**



**Total Picks: 50**

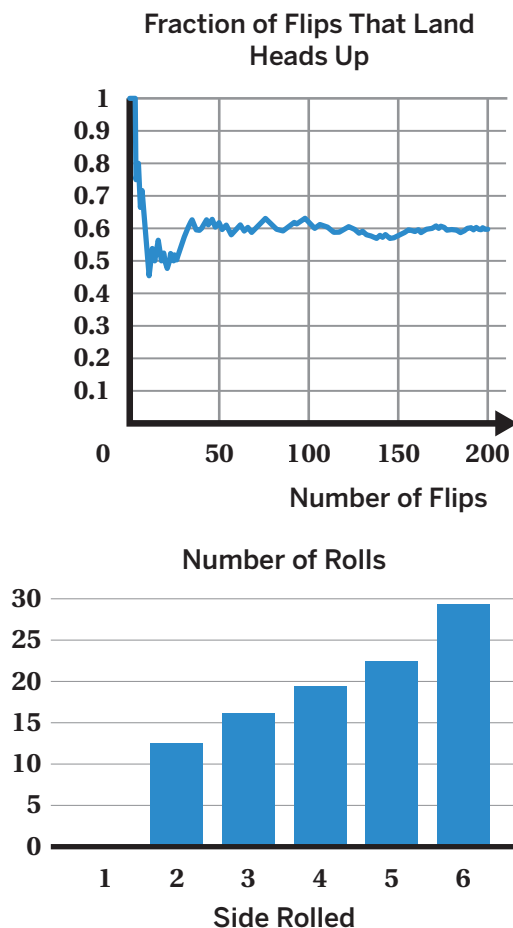




## 12 Synthesis

Describe how you can use a repeated experiment to decide whether an object is fair.

Use the results of these experiments if they help with your thinking.



## Summary

Repeated experiments can help you decide if an object is fair.

If an experiment is repeated only a few times, the results may not be what you expect, even if the object is fair. The more times you repeat the experiment (i.e., hundreds or thousands of times), the closer the probability in the experiments should get to the expected probability. This can allow you to make a better decision about whether the object is fair.

For example, here is a fair coin. The probability of this coin landing heads up is  $\frac{1}{2}$ .

If the coin is tossed only 3 times, it may land heads up all 3 times, leading to the belief that the coin is heads on both sides, or is unfair.

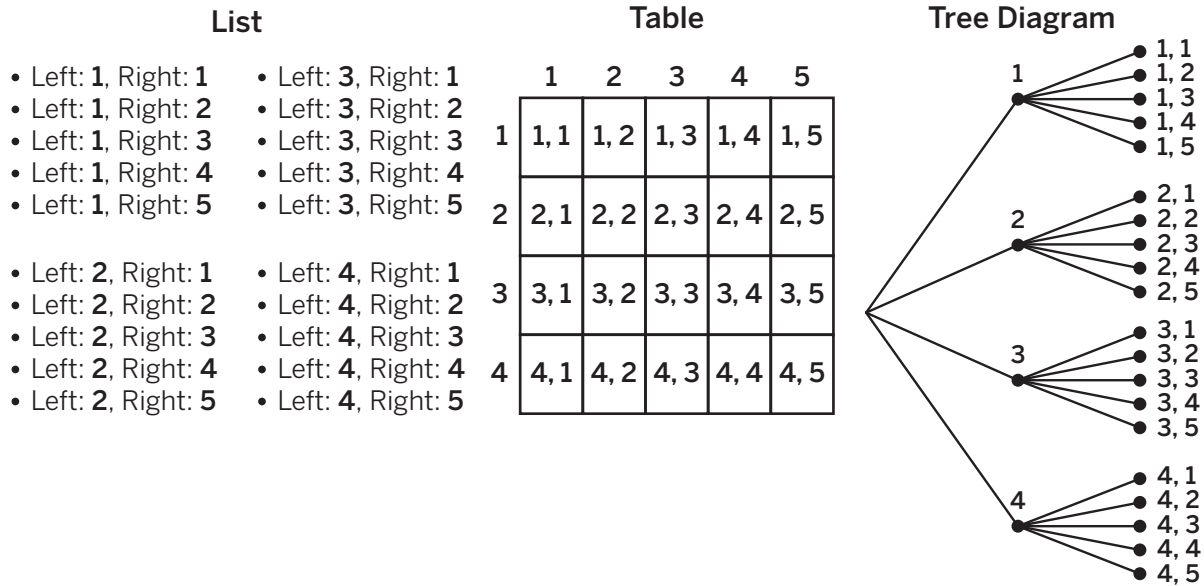
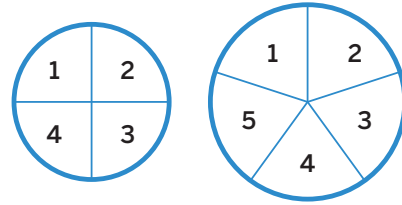
However, if this coin is tossed 1,000 times, it is expected to land heads up about half of the time because the sample space of this event is “heads” and “tails”.



## 12 Synthesis

Here are three representations of the sample space for this pair of spinners.

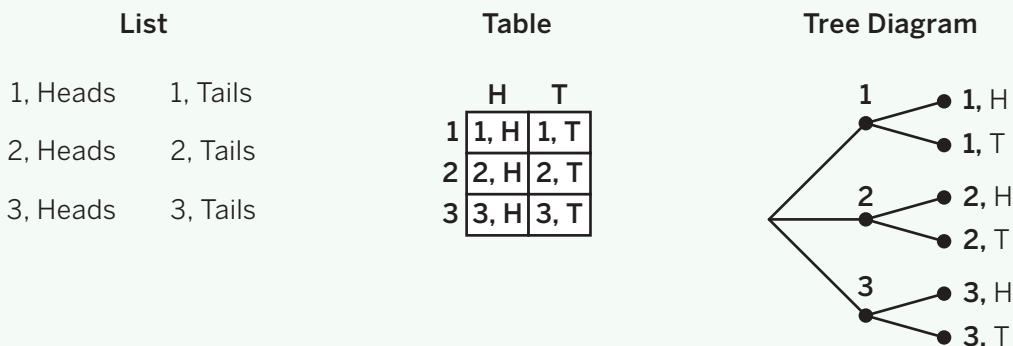
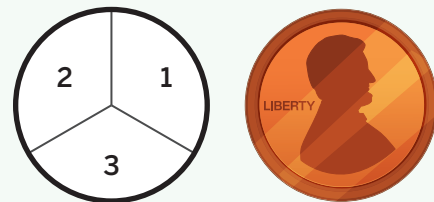
Choose a representation and describe an advantage and a disadvantage.



## Summary

There are several different ways to make sense of **compound events**, or events that involve multiple steps.

Here is one example: Let's spin a spinner and toss a fair coin. There are 6 outcomes in the *sample space* of this multistep event, which you can see in a list, a table, and a **tree diagram**.



## 12 Synthesis

Describe how simulations can be designed and used to estimate the probability of a real-world event.

Use these spinners if they help with your thinking.



1 day of rain

## Summary

**Simulations** are experiments that are used to estimate the probability of a real-world event. In order to design a good simulation, first determine the probability of the individual events occurring.

For example, you could use a coin, number cube, or spinner to simulate a 50% probability of rain.

### Flipping a Coin

Landing heads up  
 $\left(\frac{1}{2} = 50\%\right)$



### Rolling a Number Cube

Rolling an even number  
 $\left(\frac{3}{6} = 50\%\right)$



### Using a Spinner

Spinning a raindrop  
 $\left(\frac{5}{10} = 50\%\right)$



To simulate the probability of rain over three days where each day has a 50% chance of rain, you can use three coins, number cubes, or spinners and repeat the experiment many times.