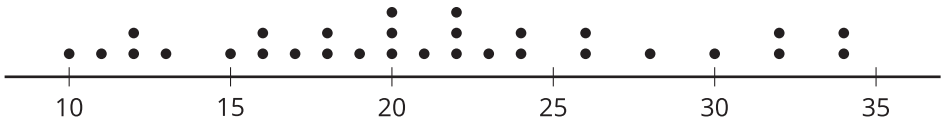
Unit 8 Family Support Materials

Data Sets, Distributions, and Sampling

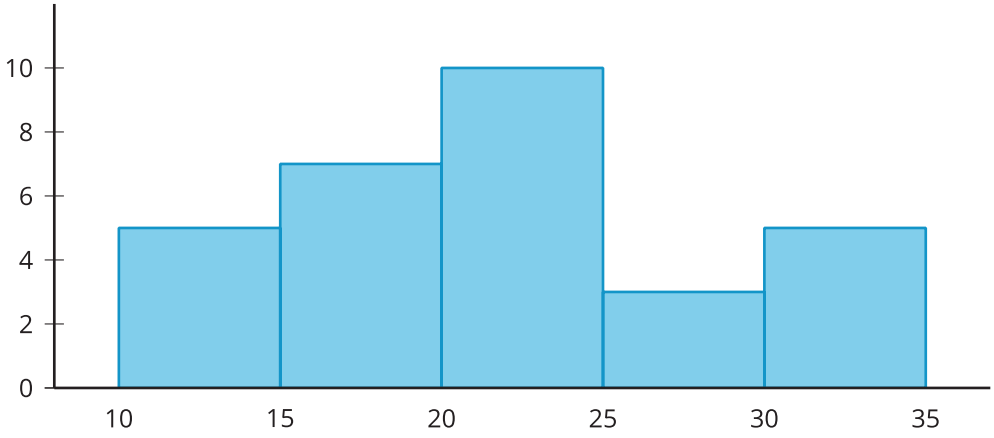
Section A: Dot Plots and Histograms

This week, your student will work with data and use data to answer **statistical questions**. Questions such as "Which band is the most popular among students in sixth grade?" or "What is the most common number of siblings among students in sixth grade?" are statistical questions. They can be answered using data, and the data are expected to vary (that is, the students do not all have the same musical preference or the same number of siblings).

Students have used bar graphs and line plots, or **dot plots**, to display and interpret data. Now they learn to use **histograms** to make sense of numerical data. The dot plot and histogram here display the distribution of the weights of 30 dogs.



dog weights in kilograms



dog weights in kilograms

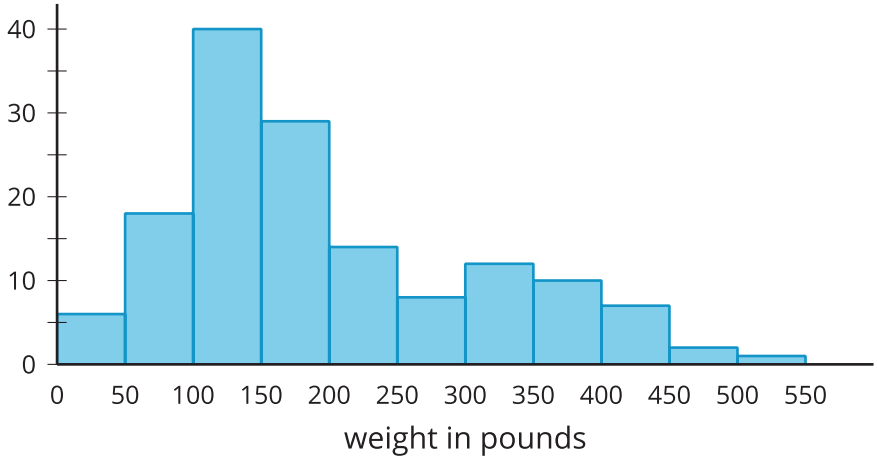
A dot plot shows individual data values as points. In a histogram, the data values are grouped. Each group is represented as a vertical bar. The height of the bar shows how many values are in that group. The tallest bar in this histogram shows that there are 10 dogs that weigh from 20 kilograms up to 25 kilograms (not including 25).

The shape of a histogram can tell us about how the data are distributed. For example, we can see that more than half of the dogs weigh less than 25 kilograms, and that a dog weighing from 25 to

30 kilograms is not typical.

Here is a task to try with your student:

This histogram shows the weights of 143 bears.



1. About how many bears weigh from 100 to 150 pounds?
2. About how many bears weigh less than 100 pounds?
3. Noah says that because almost all the bears weigh from 0 to 500 pounds, we can say that a weight of 250 pounds is typical for the bears in this group. Using the histogram, explain why this is incorrect.

Solution:

1. About 40 bears. This is the height of the tallest bar of the histogram.
2. About 24 bears. The two leftmost bars represent the bears that weigh less than 100 pounds. Add the heights of these two bars.
3. We can visually tell from the histogram that most bears weigh less than 250 pounds: The bars to the left of 250 are taller than those to the right. If we add the heights of bars, fewer than 40 bears weigh more than 250 pounds, while over 100 bears weigh less than 250 pounds, so it is not accurate to say that 250 pounds is a typical weight. A better typical weight might be around 150 pounds because most of the bears in this group seem to weigh around that much.

Section B: Measures of Center and Variability

This week, your student will learn to use measures of center and measures of spread to summarize the distribution of data.

- The *mean* and *median* are two different ways to describe the center of a data set and what is typical.
- The *mean absolute deviation (MAD)* and *interquartile range (IQR)* are two different ways to describe how spread out the data set is.

We can think of the **mean** of a data set as a fair share—what would happen if the numbers in the data set were distributed evenly. Suppose a runner ran 3, 4, 3, 1, and 5 miles over five days. The total number of miles she ran is 16 miles. If this were distributed evenly across the days, the distance run per day would be 3.2 miles ($16 \div 5 = 3.2$). To calculate the mean, we can add the data values and then divide the sum by how many values there are.

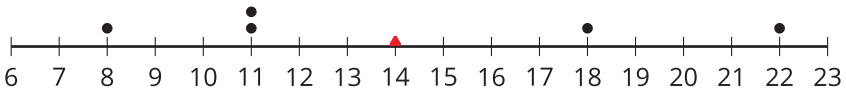
The **mean absolute deviation (MAD)** tells us the distance, on average, of a data point from the mean. When the data points are close to the mean, the MAD will be small. When points are more spread out, the MAD will be greater.

The **median** is the middle value of a data set whose values are listed in order. The runner's distances, listed in order, are: 1, 3, 3, 4, 5. The middle number is 3. This means that half of the runs were less than or equal to 3 miles, and the other half were greater than or equal to 3 miles.

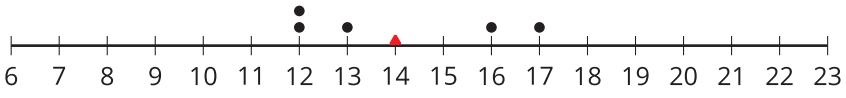
It can be helpful to break it down further. We can split each half to find the **quartiles**. The first quartile (Q1) is the median of the lower half of the data set. The third quartile (Q3) is the median of the upper half of the data set. The distance between the first and third quartiles is the **interquartile range (IQR)**. It tells us the spread of the middle half of the data.

Here is a task to try with your student:

The dots show the travel times, in minutes, of Lin and Andre. The triangles show each mean travel time.



Lin's travel time in minutes



Andre's travel time in minutes

1. Use the data on Lin's and Andre's dot plots to verify that the mean travel time for each student is 14 minutes.
2. Andre says that the mean for his data should be 13 minutes, because there are two numbers

to the left of 13 and two to the right. Explain why 13 minutes cannot be the mean.

3. Which data set, Lin's or Andre's, has a higher MAD (mean absolute deviation)? Explain how you know.

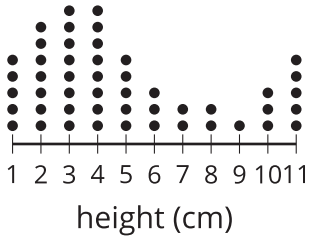
Solution:

1. For Lin's data, the mean is $\frac{8+11+11+18+22}{5} = \frac{70}{5}$, which equals 14. For Andre's data, the mean is $\frac{12+12+13+16+17}{5} = \frac{70}{5}$, which also equals 14.
2. Sample explanations:
   - The mean cannot be 13 minutes because it does not represent a fair share.
   - The mean cannot be 13 minutes because the data would be unbalanced. The two data values to the right of 13 (16 and 17) are much further away from the two that are to the left (12 and 12).
3. Lin's data has a higher MAD. Sample explanations:
   - In Lin's data, the points are 6, 3, 3, 4, and 8 units away from the mean of 14. In Andre's data, the points are 2, 2, 1, 2, and 3 units away from the mean of 14. The average distance of Lin's data will be higher because those distances are greater.
   - The MAD of Lin's data is 4.8 minutes, and the MAD of Andre's data is 2 minutes.
   - Compared to Andre's data points, Lin's data points are farther away from the mean.

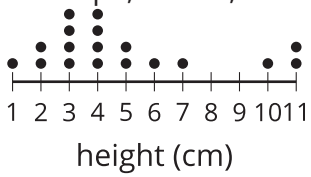
Section C: Sampling

This week your student will be working with data. Sometimes we want to know information about a group, but the group is too large for us to be able to ask everyone. It can be useful to collect data from a **sample** (some of the group) of the **population** (the whole group). It is important for the sample to resemble the population.

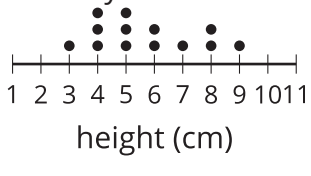
- For example, here is a dot plot showing a population: the height of 49 plants in a sprout garden.



- This sample is **representative** of the population, because it includes only a part of the data, but it still resembles the population in shape, center, and spread.



- This sample is not representative of the population. It has too many plant heights in the middle and not enough really short or really tall ones.



A sample that is selected at random is more likely to be representative of the population than a sample that was selected some other way.

Here is a task to try with your student:

A city council needs to know how many buildings in the city have lead paint, but they don't have enough time to test all 100,000 buildings in the city. They want to test a sample of buildings that will be representative of the population.

1. What would be a bad way to pick a sample of the buildings?
2. What would be a good way to pick a sample of the buildings?

Solution:

1. There are many possible answers.
    - Testing all the same type of building (like all the schools or all the gas stations) would not lead to a representative sample of all the buildings in the city.
    - Testing buildings all in the same location, such as the buildings closest to city hall, would also be a bad way to get a sample.
    - Testing all the newest buildings would bias the sample towards buildings that don't have any lead paint.
    - Testing a small number of buildings, like 5 or 10, would also make it harder to use the sample to make predictions about the entire population.
2. To select a sample at random, they could put the addresses of all 100,000 buildings into a computer and have the computer select 50 addresses randomly from the list. Another possibility could be picking papers out of a bag, but with so many buildings in the city, this method would be difficult.

Section D: Probability

This week your student will be working with probability. A **probability** is a number that represents how likely something is to happen. For example, think about flipping a coin.

- The probability that the coin lands somewhere is 1. That is certain.
- The probability that the coin lands heads up is $\frac{1}{2}$, or 0.5, because it is just as likely as not.
- The probability that the coin turns into a bottle of ketchup is 0. That is impossible.

Sometimes we can figure out an exact probability. For example, if we pick a **random** date, the chance that it is on a weekend is $\frac{2}{7}$, because 2 out of every 7 days fall on the weekend. Other times, we can estimate a probability based on what we have observed in the past.

Here is a task to try with your student:

People at a fishing contest are writing down the type of each fish they catch. Here are their results:

- Person 1: bass, catfish, catfish, bass, bass, bass
- Person 2: catfish, catfish, bass, bass, bass, bass, catfish, catfish, bass, catfish
- Person 3: bass, bass, bass, catfish, bass, bass, catfish, bass, catfish

1. Estimate the probability that the next fish that gets caught will be a bass.
2. Another person in the competition caught 5 fish. Predict how many of these fish were bass.
3. Before the competition, the lake was stocked with equal numbers of catfish and bass. Describe some possible reasons why the results do not show a probability of $\frac{1}{2}$ for catching a bass.

Solution:

1. About $\frac{15}{25}$, or 0.6. Of the 25 fish that have been caught, 15 of them were bass.
2. About 3 bass. $\frac{3}{5} = 0.6$, which is what we estimated for the probability of getting bass in this competition. It would also be reasonable if they caught 2 or 4 bass out of their 5 fish.
3. There are many possible answers. For example:
   - Maybe the lures or bait they are using are more likely to catch bass.
   - With data from only 25 total fish caught, we can expect the results to vary a little from the exact probability.