Lesson 07: How do I describe a distribution?
Q: Think before you draw!


## Shape of a distribution

1. Does the histogram have a single, central hump or several separated humps?

These humps are called modes


Gaps help to see multiple modes or enable us to see that the data may be from multiple populations (groups)

Sugar in cereals. The histogram displays the sugar content (as a percent of weight) of 49 brands of breakfast cereals.


Is the histogram symmetric?


so that the two sides almost match.
pretty close!
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

Do any unusual features stick out?

$\qquad$

A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.


How would you describe the distribution?

It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape.
Where do you think the center might be? How spread out do you think the values will be?

1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving

Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.
2. Center of the distribution



The middle value that divides the histogram into two equal areas is called the median.

How can we find it for $n$ entries of data?

even odd

## 3. Spread

Range $=\max -\min$

Statistics pays close attention to what we don't know as well as what we do know. Understanding how spread out the data are is a first step in understanding what a summary cannot tell us about the data. It's the beginning of telling us what we don't know.
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


The range has the disadvantage that a single extreme value can make it very large, giving a value that doesn't really represent the data overall.

## Interquartile Range

One quarter of the data lies below the lower quartile, and one quarter of the data lies above the upper quartile, so half the data lies between them.

The quartiles border the middle half of the data.

$$
I Q R=\text { upper quartile }- \text { lowver quartile }
$$




Benefit of IQR over range??

