Lecture 4

Describing Numerical Distributions

Shape, Center, Spread, and Unusual Values

1. Describing a Numerical Distribution

When examining a numerical distribution, we want to describe its important features. A useful description should tell us both what the distribution looks like and where the observations are located.

A complete description of a numerical distribution considers:

$$\boxed{\text{Shape} \longrightarrow \text{Peaks} \longrightarrow \text{Unusual Values} \longrightarrow \text{Center} \longrightarrow \text{Spread}}$$

- **Shape:** Is the distribution symmetric or skewed?
- **Peaks:** How many major mounds does it have?
- **Unusual values:** Are there gaps, clusters, or outliers?
- **Center:** Where is the distribution located?
- **Spread:** How much do the observations vary?

This framework provides a consistent way to describe and compare histograms and other displays of numerical data.

2. Center and Spread

Two fundamental features of a numerical distribution are its **center** and **spread**.

- The **center** describes approximately where the data are located.
- The **spread** describes how much the observations vary around the center.

Two distributions can have approximately the same center but very different spreads.

2.1 Example: Same Center, Different Spread

Consider the two data sets

$$A: \quad 48,\ 49,\ 50,\ 50,\ 51,\ 52$$

and

$$B: \quad 35,\ 42,\ 48,\ 52,\ 58,\ 65.$$

Both data sets are centered near 50, but their spreads are very different.

Data set A is tightly concentrated around 50, whereas data set B extends much farther from 50.

A statistical description should not report only a typical value. We should describe both

$$\boxed{\text{where the data are located}} \quad \text{and} \quad \boxed{\text{how much the data vary.}}$$

3. Shape of a Distribution

The shape of a distribution describes its overall pattern.

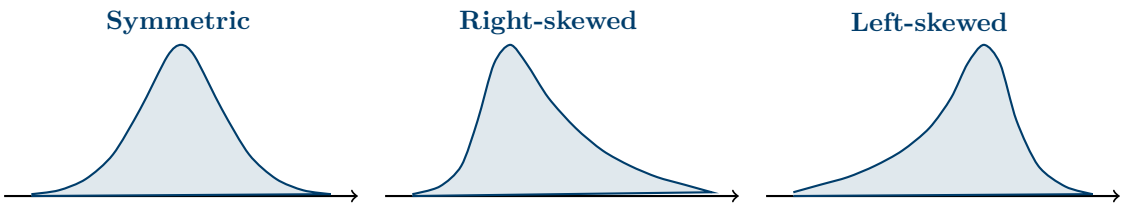
Three important questions are:

Is the distribution symmetric or skewed?
How many major peaks does it have?
Are there unusual features or values?

3.1 Symmetric and Skewed Distributions

Definitions.

- A **symmetric** distribution has approximately equal tails on both sides.
- A **right-skewed** distribution has a long tail extending toward larger values.
- A **left-skewed** distribution has a long tail extending toward smaller values.



Memory rule:

Tail right	$\Rightarrow$	right-skewed
Tail left	$\Rightarrow$	left-skewed
Tails approximately equal	$\Rightarrow$	symmetric

The skew is named for the direction of the tail, not the direction of the peak.

3.2 Number of Peaks: Modality

A distribution may have one or more major groups, or "mounds."

Definitions.

- **Unimodal**: one major mound.
- **Bimodal**: two major mounds.
- **Multimodal**: more than two major mounds.

When counting peaks, ignore small random ups and downs. Focus on the major high points of the overall distribution.

1 major peak	$\Rightarrow$	Unimodal
2 major peaks	$\Rightarrow$	Bimodal
$3+$ major peaks	$\Rightarrow$	Multimodal

How to count peaks

1. Look for the major high points of the distribution.
2. Ignore small random fluctuations between neighboring bars.
3. Count the major mounds.
4. Use the number of major mounds to describe the modality.

3.3 Why a Bimodal Distribution Can Be Important

A bimodal distribution may indicate that the data contain two different groups or populations.
For example:

- heights of children and adults;
- test scores from two different classes;
- waiting times during two different periods of the day.

If the heights of children and adults are combined into one data set, the resulting distribution might have one mound near children's heights and another near adults' heights.

Important:
A second mound is not automatically evidence of two populations. It is a clue that deserves further investigation.

4. Unusual Values and Outliers

Definition.
An **outlier** is an unusually large or small observation that does not fit the general pattern of the rest of the data.

Possible explanations for an outlier include:

- a data-entry error;
- an unusual individual;
- an unusual circumstance;
- a measurement problem.

Important:
Always investigate an outlier before removing it.

4.1 Example: Investigating an Outlier

Suppose the ages of seven people at a meeting are

$$21, \quad 22, \quad 23, \quad 24, \quad 25, \quad 26, \quad 72.$$

The value 72 is much larger than the other observations and may be considered unusual.
However, we should not automatically delete it. Instead, ask:

1. Was 72 entered correctly?
2. Was the measurement made correctly?
3. Does the person actually belong to the population being studied?

Suppose the meeting is for a college student organization and the person is actually 72 years old. Then 72 may be a genuine observation.

An unusual observation should therefore be investigated in context rather than automatically removed.

5. Choosing Measures of Center and Spread

The shape of a distribution helps determine which numerical summaries are most appropriate.

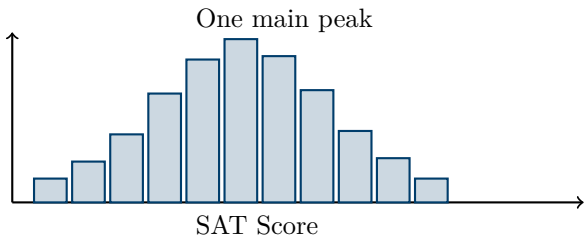
General rule for numerical data

Shape	Center	Spread
Approximately symmetric	Mean	Standard deviation
Skewed	Median	IQR

The reason is that the mean and standard deviation are sensitive to extreme observations, whereas the median and IQR are more resistant to them.

6. Example: SAT Scores

Suppose we have a histogram showing the SAT scores of a large group of students.



Shape:

The distribution is approximately symmetric and mound-shaped. The left and right sides are reasonably similar, so there is no strong evidence of skewness.

SAT scores: approximately symmetric

Peaks:

There is one main high point.

Unimodal

Unusual values:

There do not appear to be any major gaps. Extremely high or low scores could occur, but an observation would need to be sufficiently far from the rest of the data to be considered an outlier.

Center:

Because the distribution is approximately symmetric, the **mean** is an appropriate measure of center.

If the histogram supports such an estimate, the typical score might be described as approximately

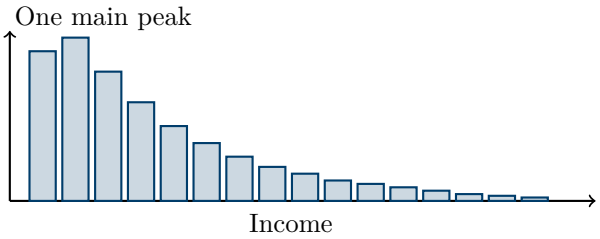
1050–1100.

The exact value should depend on the actual data represented by the histogram.

Spread:
The scores show a moderate amount of variability. For an approximately symmetric distribution, the preferred numerical measure of spread is the **standard deviation**.

7. Example: Income of U.S. Residents

Now consider a histogram showing income.



Shape:
Income is typically **right-skewed**. Most observations are concentrated toward lower and middle income levels, while a relatively small number of observations have extremely high incomes.
These high values produce a long tail extending to the right.

Income: right-skewed

Peaks:
There is one main high point, so the distribution is generally **unimodal**.
Unusual values:
Income distributions may contain high-income outliers. These observations can appear far to the right of the main group.
We should distinguish between a general long tail and individual observations that are unusually far from the rest of the data.
Center:
Because income is strongly right-skewed, the **median** is the preferred measure of center.
The mean can be pulled upward by a small number of extremely wealthy individuals.

Right-skewed distribution $\Rightarrow$ use the median

Spread:
Income can have a large amount of variability. Because the distribution is skewed, the preferred numerical measure of spread is the **interquartile range (IQR)**.

Summary of the Income Distribution

- **Shape:** Right-skewed with a long right tail.
- **Peaks:** One major peak; therefore, unimodal.
- **Unusual values:** Possible high-income outliers.
- **Center:** Median.
- **Spread:** IQR.

8. Comparing Two Distributions

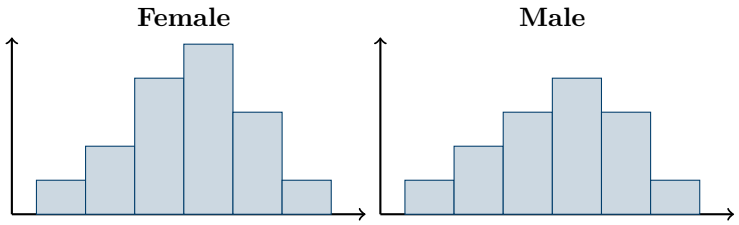
When comparing two numerical distributions, use the same framework for both distributions.

Ask:

1. What is the **shape**?
2. How many major **peaks** are there?
3. What is the approximate **center**?
4. How does the **spread** compare?
5. Are there any **unusual values**?

8.1 Example: Comparing Two Soccer Distributions

Suppose we compare the number of goals scored by first-year NCAA Division III female and male soccer players.



Shape:

Both distributions are approximately mound-shaped, with frequencies increasing toward the middle and then decreasing. Neither shows a strong skew.

Center:

The distributions are centered at different locations on the horizontal scale. The female distribution is centered around approximately 4 goals, while the male distribution is centered around approximately 11 goals on the scale shown.

Spread:

The distributions have approximately the same overall spread.

Unusual values:

There are no obvious unusual values or outliers in either distribution.

The important point is that distributions should be compared using their overall patterns rather than by focusing on a single numerical feature.

Contextual Examples of Shape

The direction of skewness can often be understood from the context.

8.2 Household Income

Most households may have moderate incomes, while a relatively small number have extremely high incomes.

Household income is often right-skewed.

8.3 Scores on an Easy Exam

If an exam is very easy, many students may score near 100, while only a few students have much lower scores.

The tail extends toward smaller values:

The distribution may be left-skewed.

8.4 Heights of Adults

For a reasonably homogeneous population, adult heights often form a roughly symmetric, mound-shaped distribution.

Adult height is often approximately symmetric.

9. Mode

The **mode** is the value that occurs most frequently in a data set.

Definition.
The **mode** is the value that occurs most frequently in a data set.

For example,

$$2, \quad 3, \quad 3, \quad 4, \quad 4, \quad 4, \quad 5, \quad 6$$

has mode

$$4.$$

The mode can be especially useful for categorical data, where the most frequent category may be of interest.

For numerical data, however, the mode is not always a useful measure of center. For example,

$$10, \quad 20, \quad 20, \quad 30, \quad 40, \quad 50, \quad 60$$

has mode 20, but 20 is not necessarily representative of the overall distribution.

The mode can also:

- be absent;
- occur more than once;
- change substantially when the data change slightly.

Thus, the mode should be interpreted in the context of the overall distribution.

When asked to describe a numerical distribution, use the following order:

Shape → Peaks → Unusual Values → Center → Spread

A complete description might look like this:

> The distribution is approximately symmetric and unimodal, with one main peak, no obvious outliers, a center around 50, and moderate spread.

For a skewed distribution:

> The distribution is right-skewed and unimodal, with one main peak and a long right tail. There may be high-value outliers. The median is an appropriate measure of center, and the IQR is an appropriate measure of spread.

Summary

Key ideas to remember

1. **Skewness is named for the direction of the tail.**
2. **Count major peaks** to determine modality: unimodal, bimodal, or multimodal.
3. **Investigate unusual values** before deciding whether they should be removed.
4. For an approximately **symmetric** distribution, use the **mean and standard deviation**.
5. For a **skewed** distribution, use the **median and IQR**.
6. Always describe a distribution using its **overall pattern**, not just one numerical summary.

Symmetric
$\Downarrow$
Mean + Standard Deviation

Skewed
$\Downarrow$
Median + IQR

Tail right $\Rightarrow$ right-skewed Tail left $\Rightarrow$ left-skewed

1 peak $\Rightarrow$ unimodal, 2 peaks $\Rightarrow$ bimodal, 3 + peaks $\Rightarrow$ multimodal