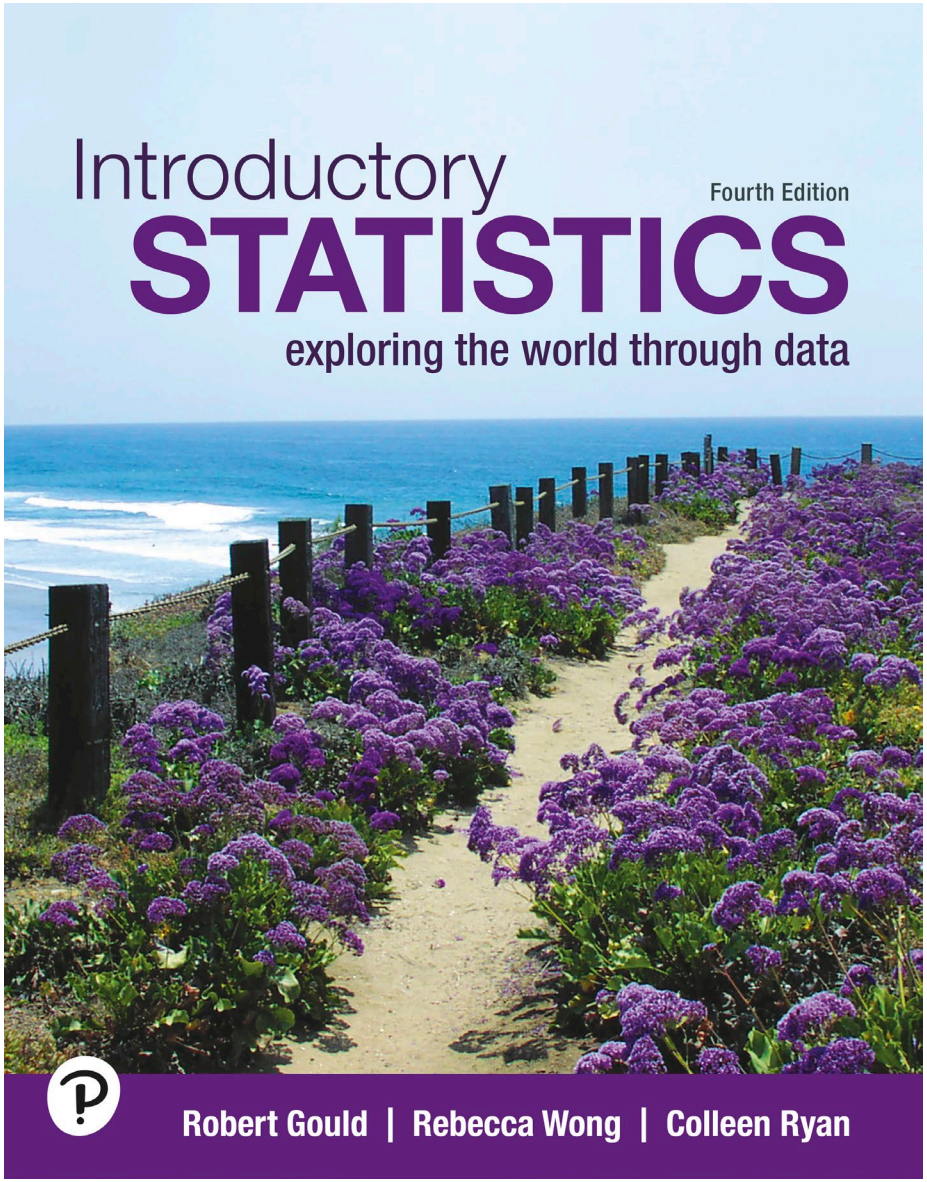
Chapter 2

Picturing Variation with Graphs



Section 2.1

Visualizing Variation in Numerical Data

- Distribution of Numerical Data
- Dotplots
- Histograms
- Density Plots

What Is a Distribution?

Recall:

Any collection of data will have variation within the data.

The most important tool for organizing the variation in data is called the **distribution**.

The **distribution of a data set** is a list that records the values that were observed and the frequencies (counts) of these values.

Example

The data set shows the number of goals scored by the 24 players of the San Diego Wave Football Club in 2022:

0, 16, 2, 1, 0, 1, 0, 3, 1, 1, 0, 0, 0, 2, 0, 3, 0, 0, 0, 1, 0, 0, 4, 0

This list includes only the values.

Tables

The distribution of data can be organized as a *table of frequencies* that lists all data values with their counts (frequencies).

Patterns may be difficult to identify.

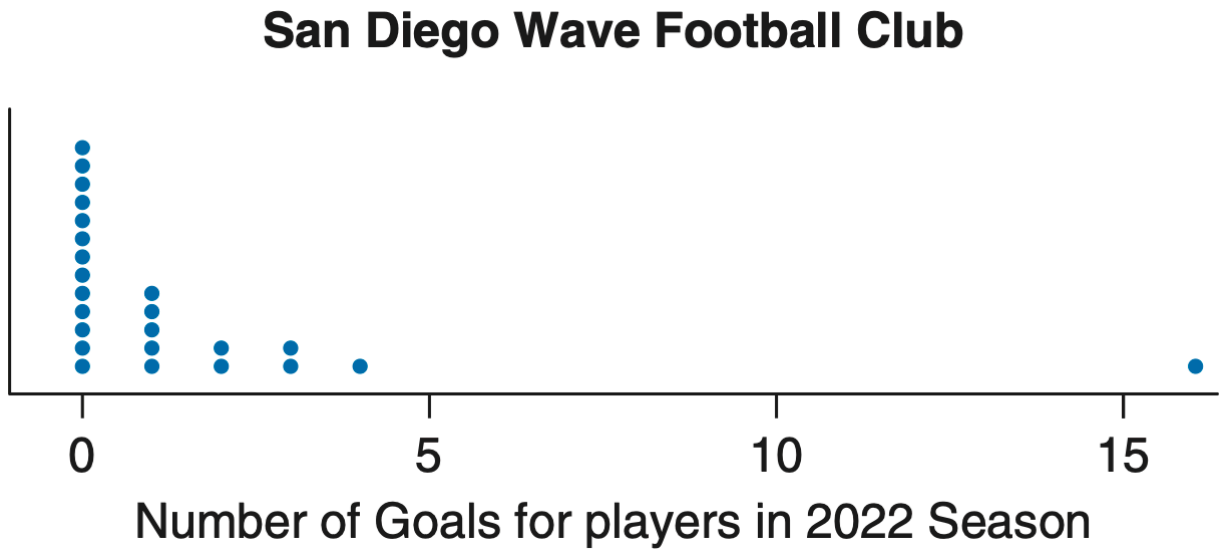
Value	Frequency
0	13
1	5
2	2
3	2
4	1
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	1

Visualizing Data

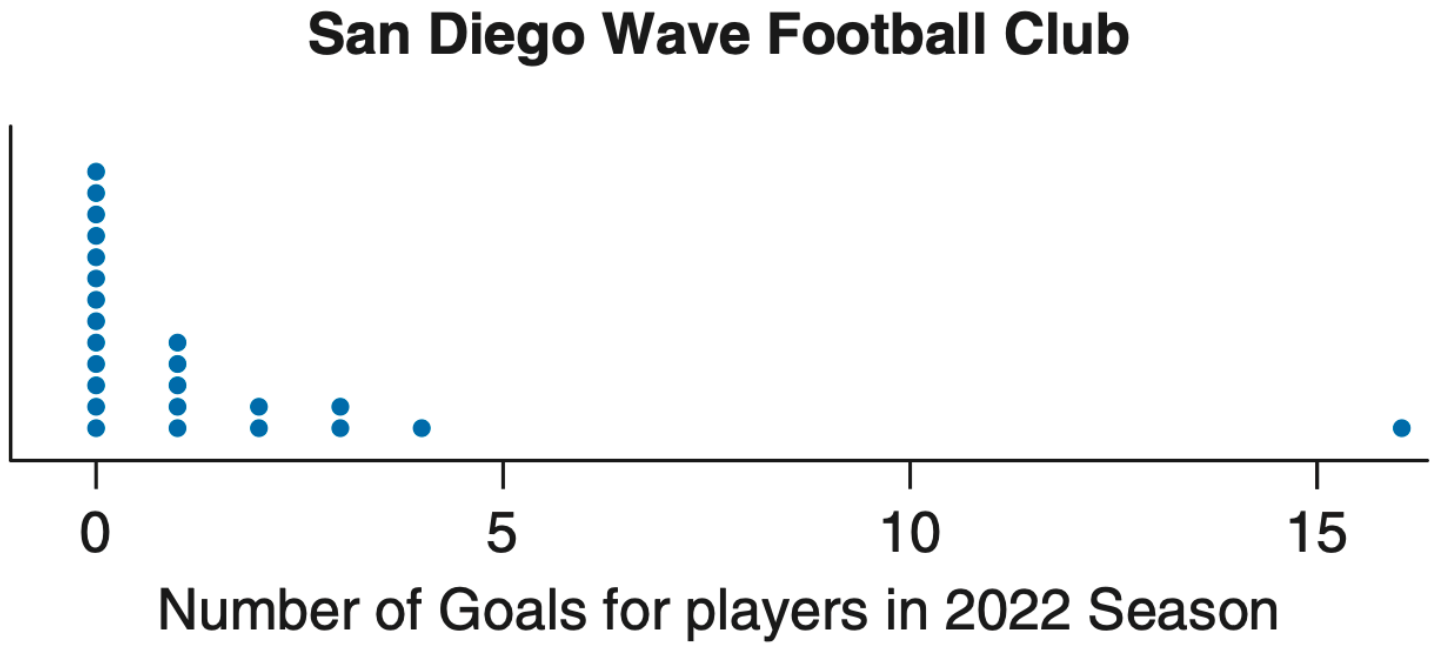
- A visual representation of the distribution must:
  - Record the data values.
  - Indicate the frequencies of the data values.
- Using a picture to display the data will help identify patterns.
- Different visual representations capture different aspects in the data.

Dotplot

In constructing a **dotplot**, we simply put a dot above a number line where each value occurs.



Dotplot Example



Each dot represents one player.

Observations:

- Most of the athletes scored no goal.
- Scoring 4 goals is rare. Scoring 16 goals is truly exceptional.

Dotplots: Pros and Cons

- **Advantages**
  - Shows individual data values
  - Helps investigate the shape of the distribution
- **Disadvantages**
  - Not great for data set with too many values

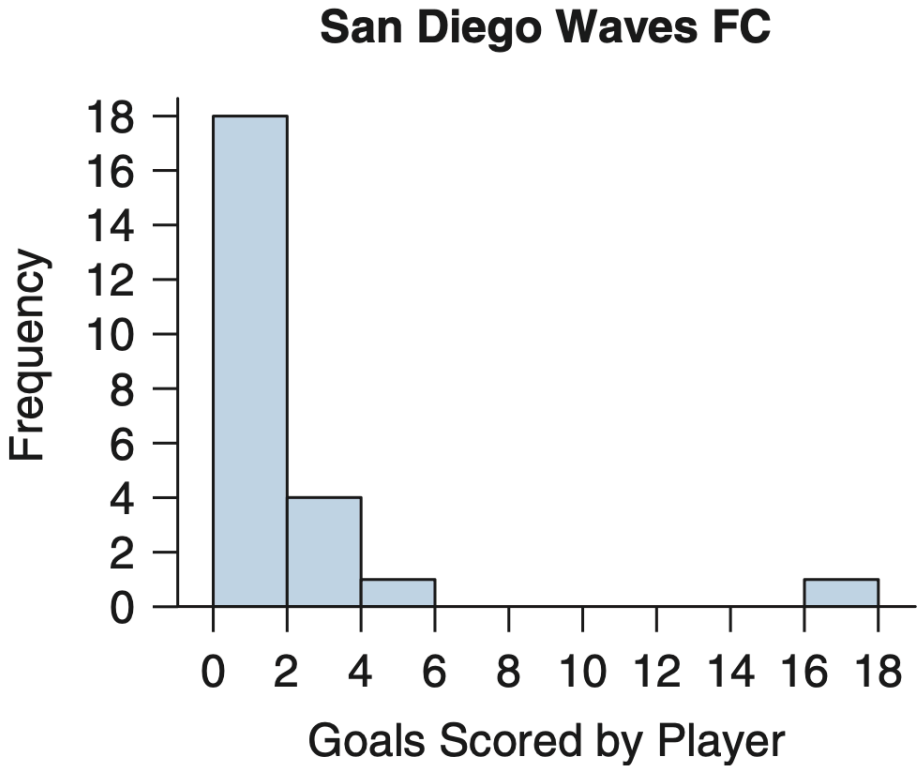
Histogram

- Group data into **intervals**, also called **bins**.
- Count how many data fall into each bin.
- Each rectangle has the following properties:
  - Consecutive bins touch
  - The height of each rectangle corresponds to the count.

Histogram Example

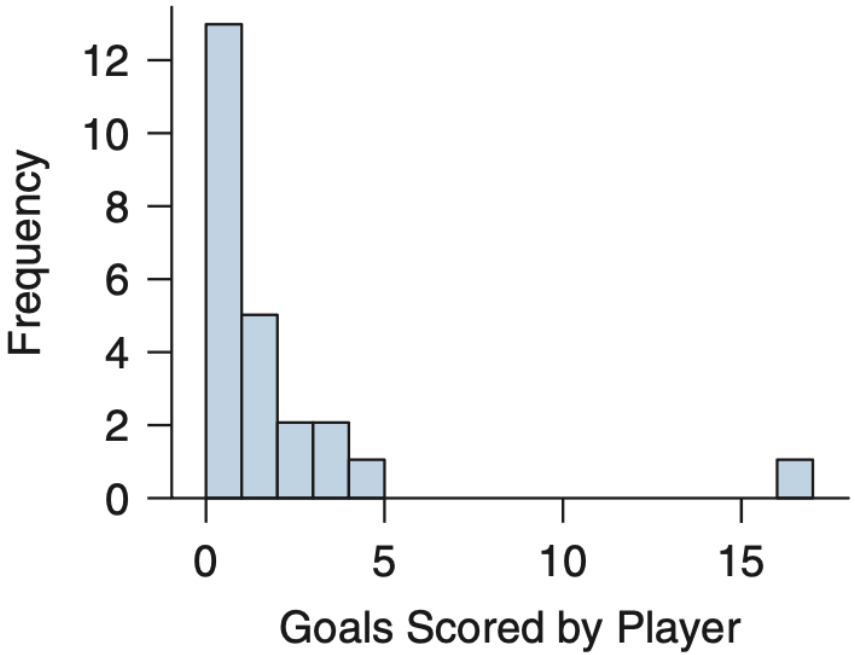
Bins: 0—2, 2—4, 4—6, …, 16—18.

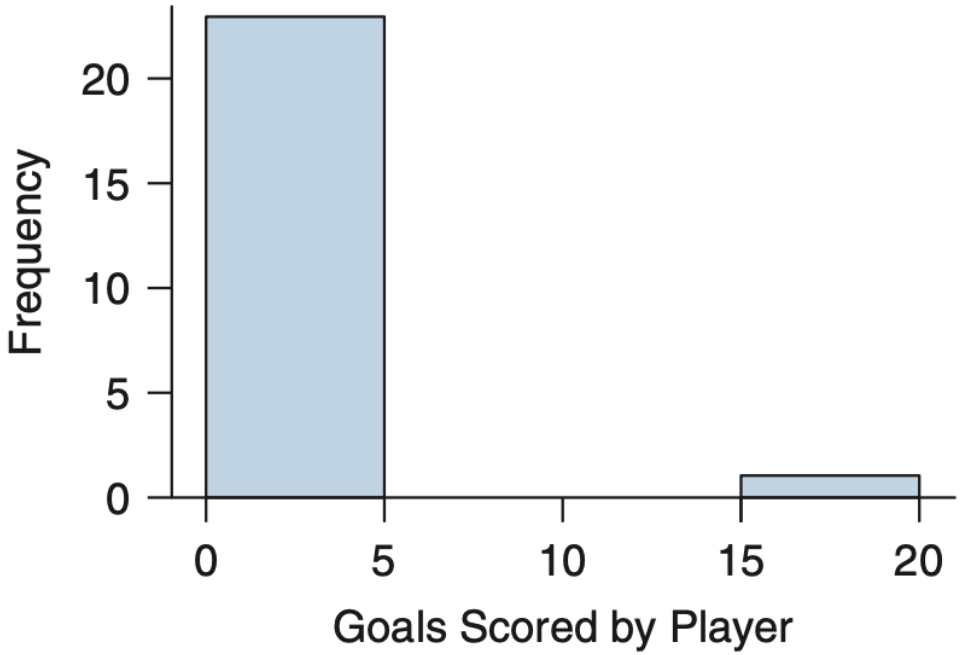
The rightmost value in each bin is counted toward the next bin.



Changing bin widths

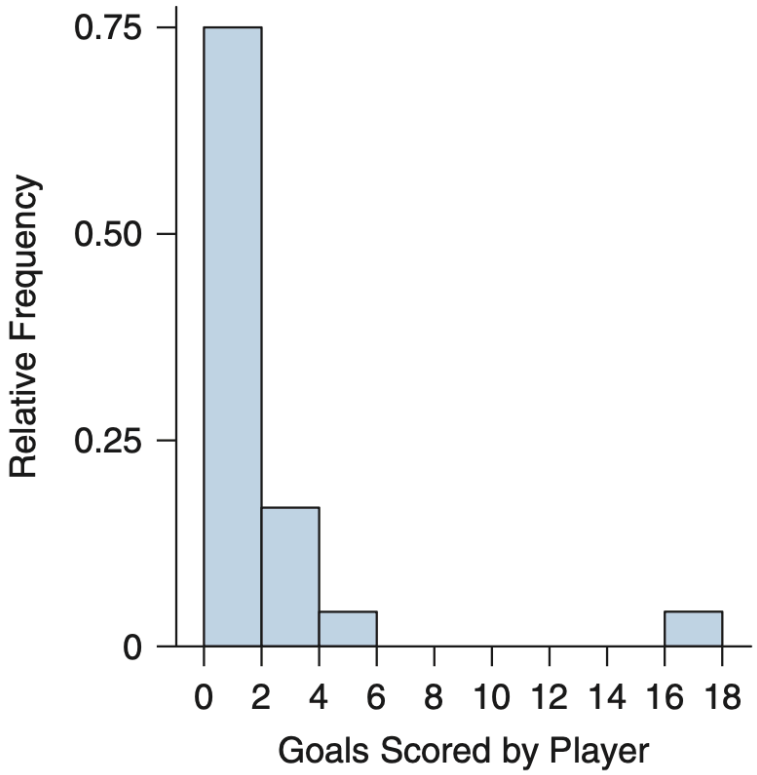
Changing the bin width changes the shape.





Relative Frequency Histogram

The rectangle heights are **relative frequencies** (proportion) rather than the count.



Histogram: Pros and Cons

- **Advantages**
  - Good for large data sets
  - Helps focus on the general shape of the data
  - Easy to spot outliers
- **Disadvantages**
  - Individual data values are not visible (lost)
  - Distribution shape affected by change in bin width

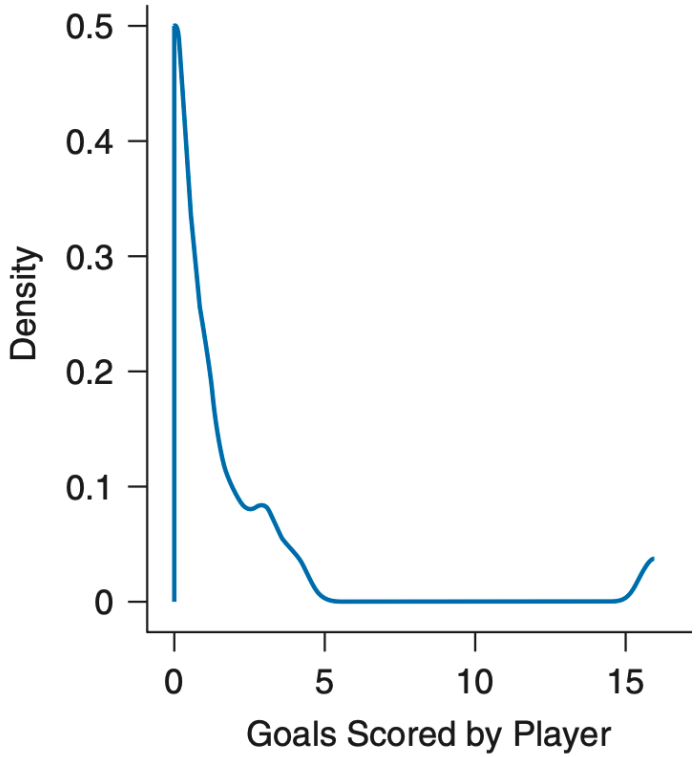
Density Plot

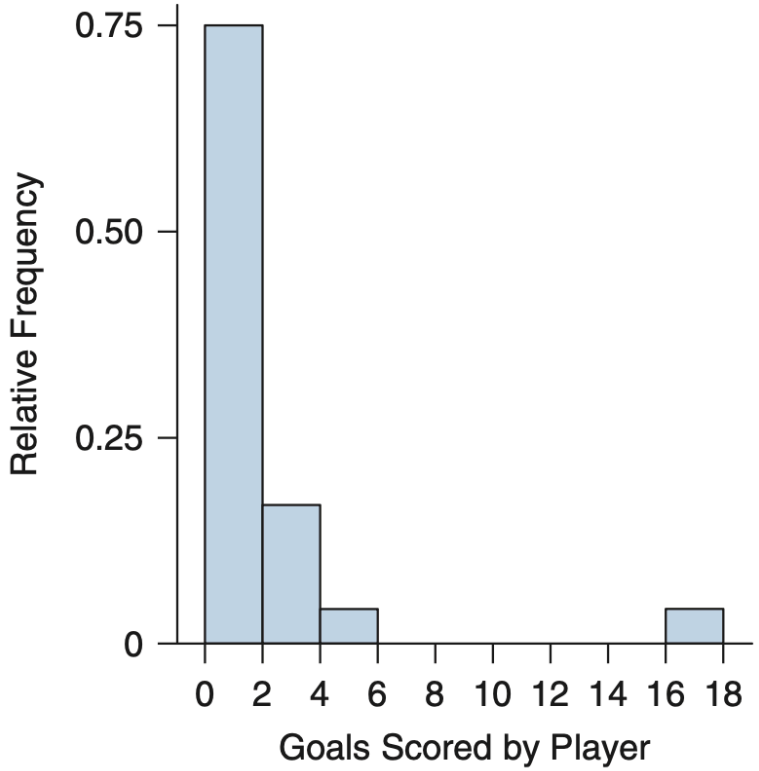
If a histogram is a smooth version of the dotplot, the **density plot** is a smooth version of the histogram. You can draw it by outlining the shape of a histogram.

The density scale on the horizontal axis is similar to relative frequency in that it is indicating (roughly speaking) proportions rather than actual counts.

Example

The density plot is along side the relative frequency histogram for the goals scored.





The density plot outlines the general shape of the histogram.

Section 2.2

Summarizing Numerical Distribution

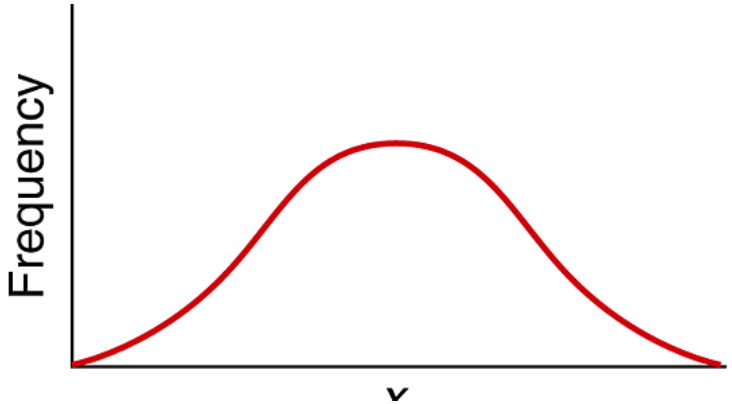
- **Shape**: symmetry, mounds, outliers
- **Center**: typical value
- **Spread**: variability of data

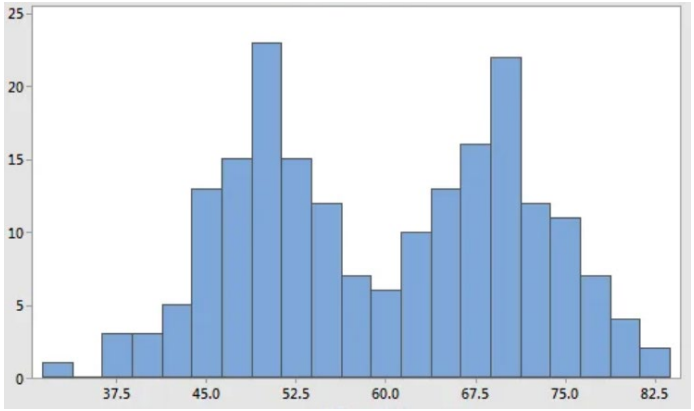
Shape

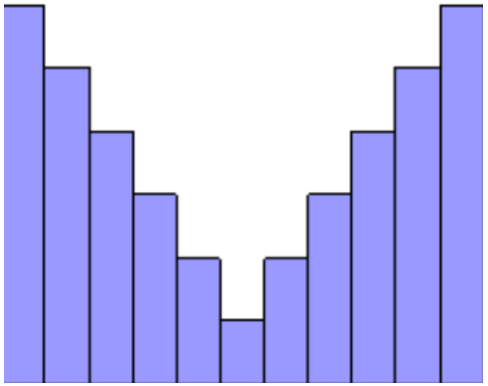
- Is the distribution symmetric or skewed?
- How many mounds (peaks) appear?
- Are there unusually large or small values (outliers)?

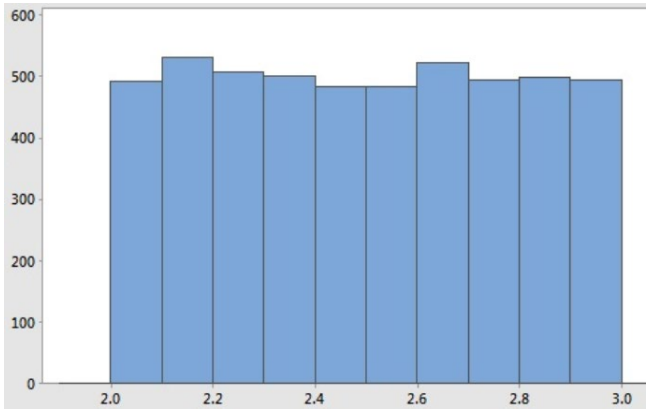
Shape: Symmetric

Symmetric: Left and right side roughly the same



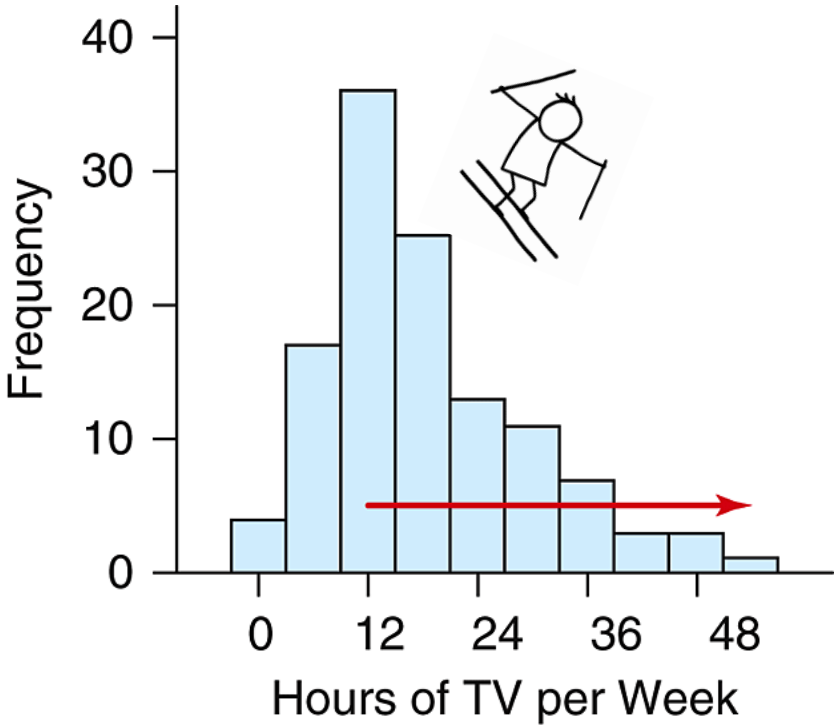




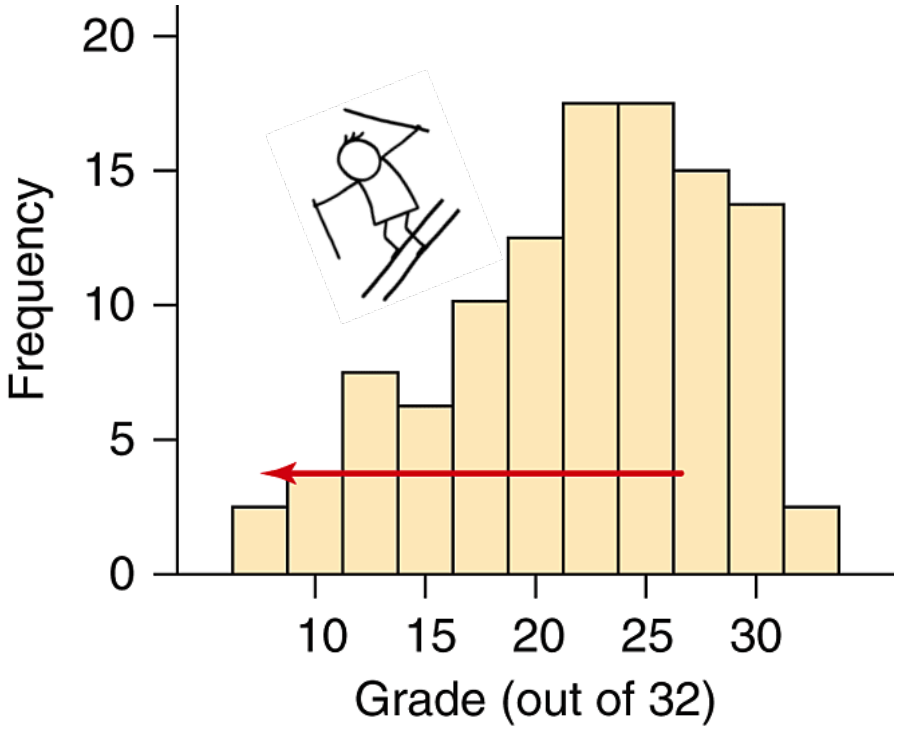


Shape: Skewed

Most of the data is on one side with a long tail.



Right-skewed distribution

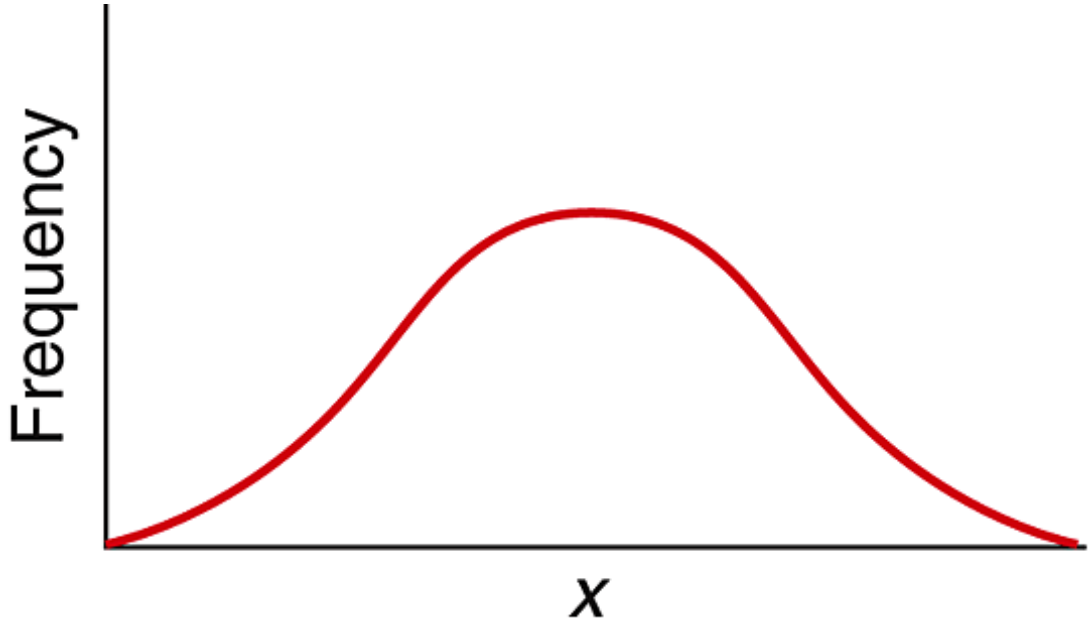


Left-skewed distribution

Shape: Mounds

How many mounds (peaks) are present?

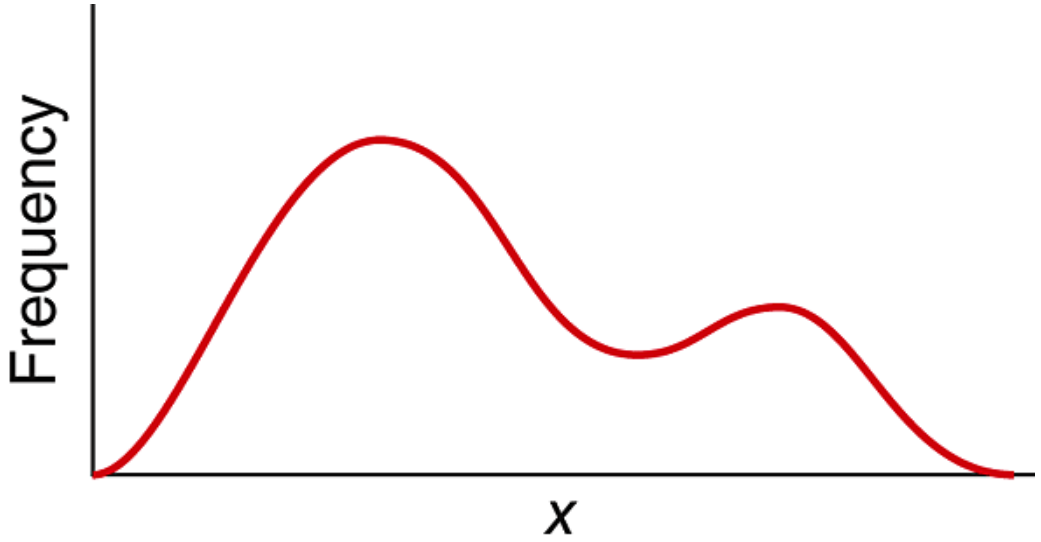
- Unimodal: "uni" means 1



Shape: Mounds

How many mounds are present?

- Bimodal: “bi” means 2

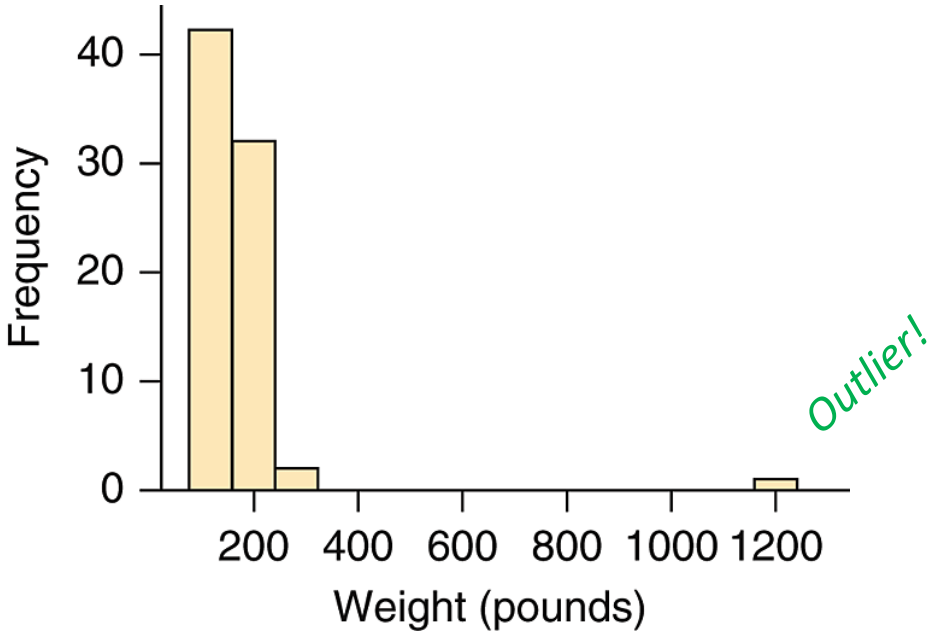


- Multimodal

More than two main mounds

Shape: Outliers

- Extremely large or small values
- Data values that don't fit the pattern of the rest of the data
- Outliers may be subjective

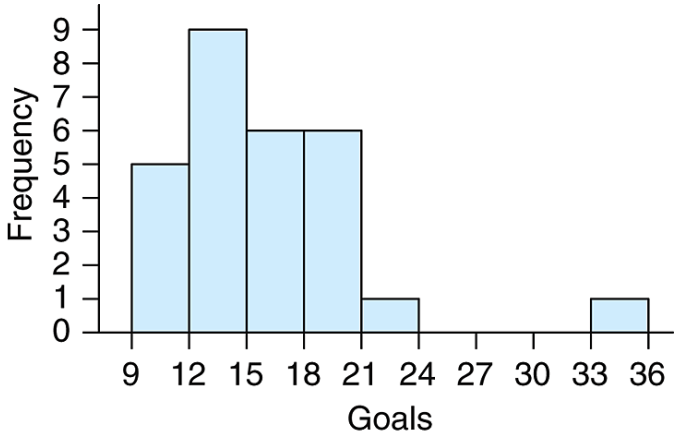


Center

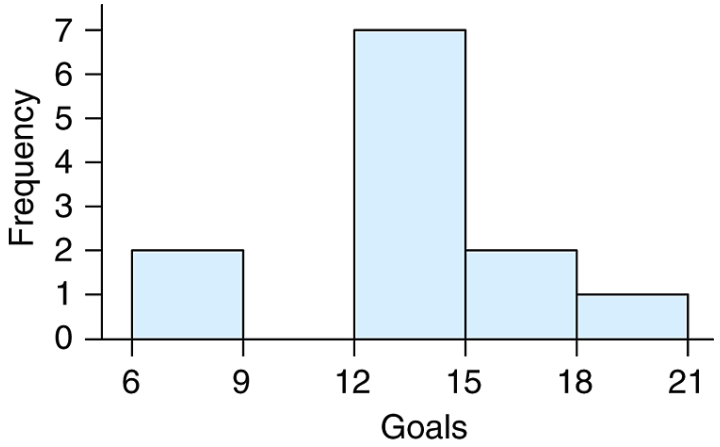
Center: the typical data value

The histograms for women and men soccer players in 2012:

Women



Men



- The typical scores:

  Women: 16 goals Men: 13 goals

- A typical male soccer player seems to score fewer goals than a typical female player.

Variability: spread

Look at the horizontal spread in the histogram:

Which data has the greater spread? How do you know?