ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

SUMMARIZING NUMERICAL DATA

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Summarizing Numerical Data

We will:

- Visualize data using scatterplots and histograms, and

- Succinctly summarize and visualize data using statistics:
  - Mean, median, mode
  - Variances
  - Quartiles

Scatterplots for Paired Data

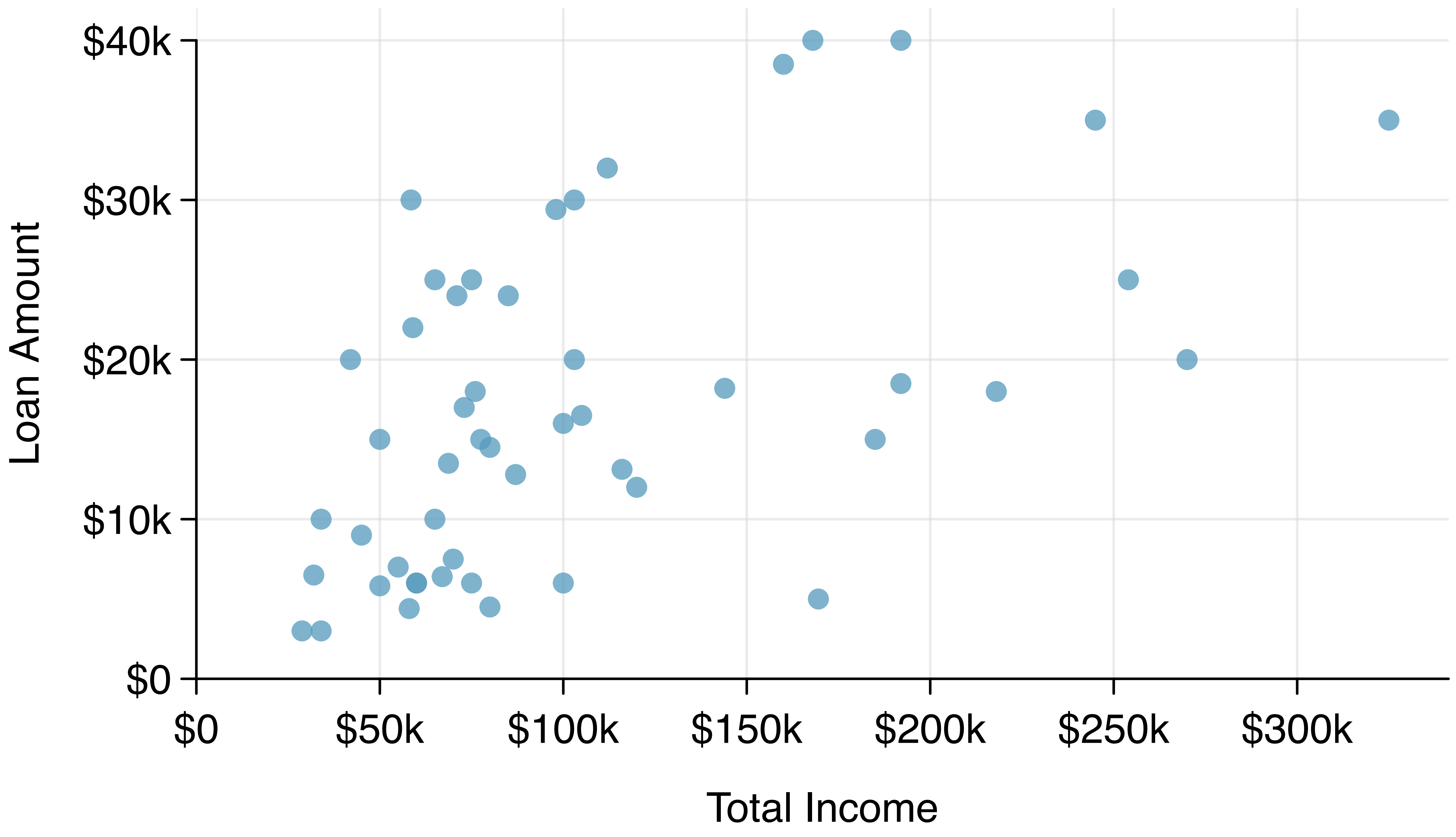

Figure from Open-intro Statistics textbook, Chapter 1.

- In a **scatterplot** each point represents a single case.

- **Scatterplots**
  - help visualize the relationship between two numerical variables.
  - help spot associations between variables, and identify whether relations are simple or complex.

Histograms for Single Variables

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	$\cdots$	25.0% - 27.5%
Count	11	15	8	4	$\cdots$	1

- When we are interested in a single numerical variable's distribution, we can use a **histogram** as a visual aid.
- We think of each numerical value as belonging to a **bin,** and we count the number of cases falling in that bin.
- Histograms provide a view of the **data density**: higher bars $\Longrightarrow$ data in the bin is more common

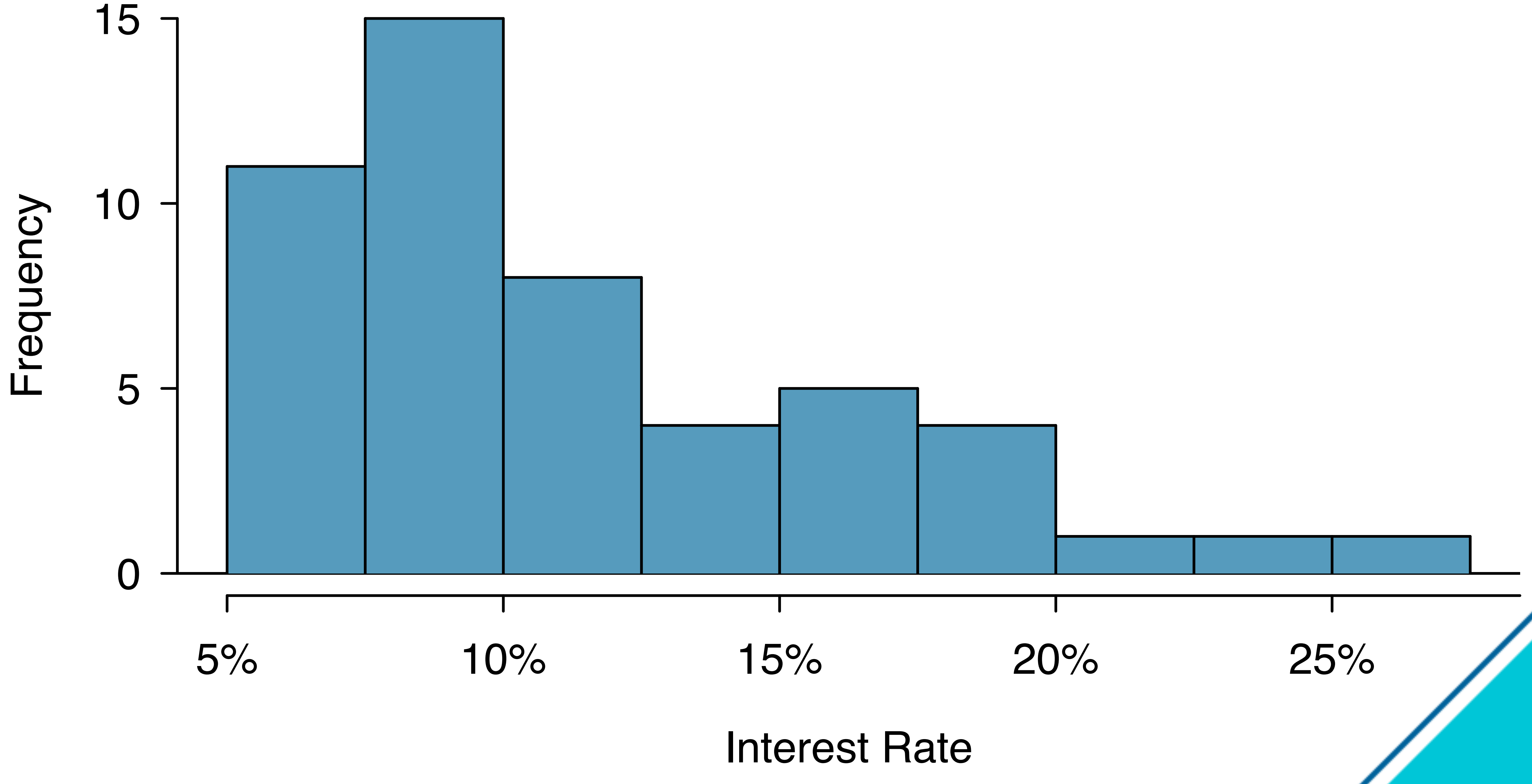

Figure from Open-intro Statistics textbook, Chapter 1.

Histograms

- Histogram suggests that most loans have rates under 15%. Few have rates above 20%.
- When data trail off to the right this way, we say it has a longer **right tail.** The shape is said to be **right skewed.**
- When data trail off to the left and has a longer left tail, the shape is said to be **left skewed.**
- Data sets that show roughly equal trailing off in both directions are called **symmetric**.

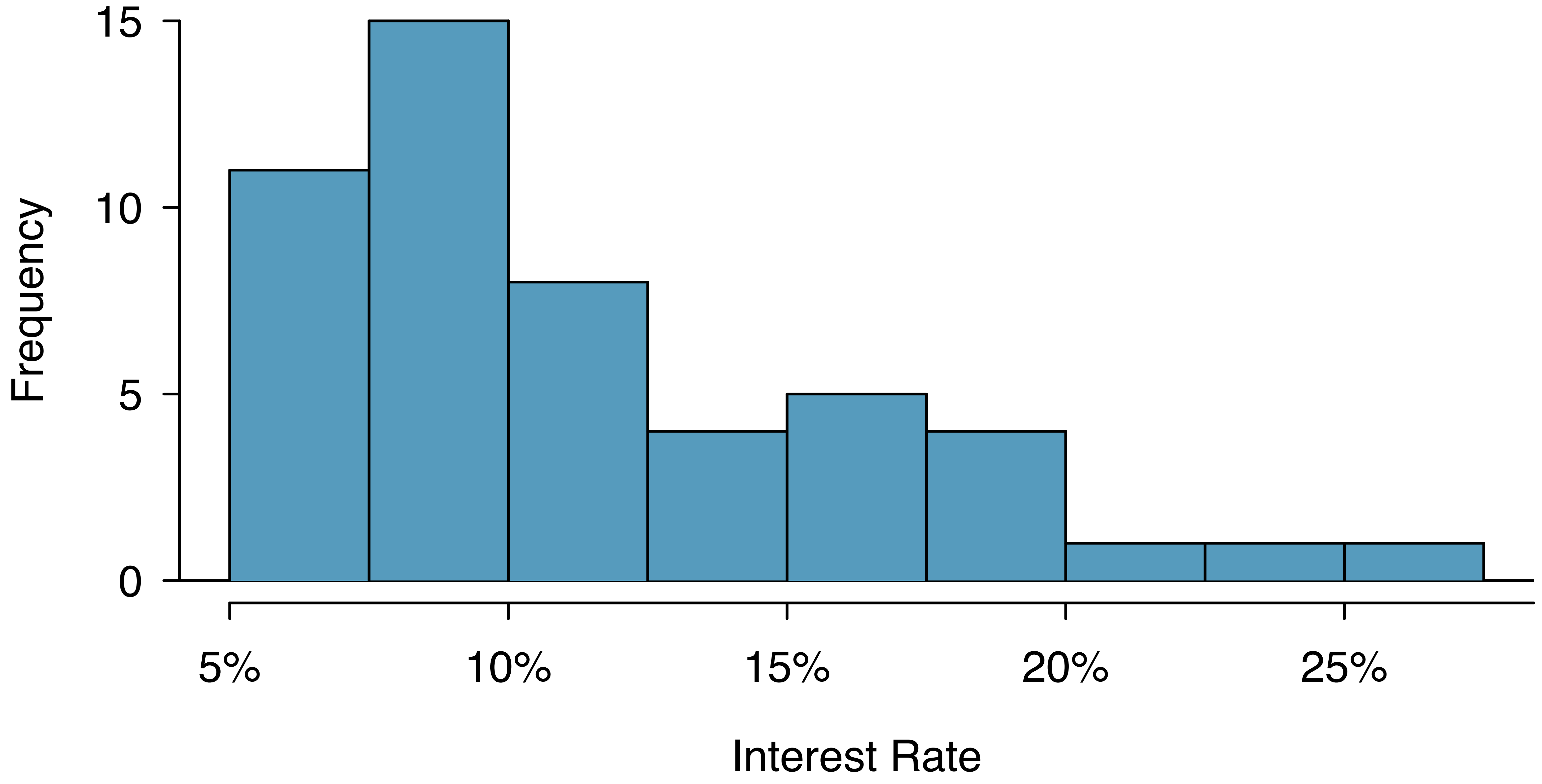

Figure from Open-intro Statistics textbook, Chapter 1.

Sample Mean (or Average)

- The **mean** (or **average**), is a common way to measure the center of a distribution of data.

- To compute the mean of a variable, we add up all the case values and divide by the number of observations, say, n:

$$\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Example: to compute the mean interest rate, we add all the interest rates and divide by the number of observations

$$\bar{x} := \frac{10.9 + 9.92 + \cdots + 6.08}{50} = 11.57\,\%$$

Sample Mean (or Average)

- Knowing the mean of a set of data can be useful because it allows us to make comparisons.

- For example:
  - We can compare average income of people with college degrees to the average income of people without college degrees.
  - We can compare average increase/decrease in blood pressure of people taking a certain drug to the average increase/decrease in blood pressure of people taking a placebo.

Mode

- The **mode** of a data set, strictly speaking, is the value that occurs the most.
- Instead, we think of it as a value (or bin) with a prominent peak in the distribution

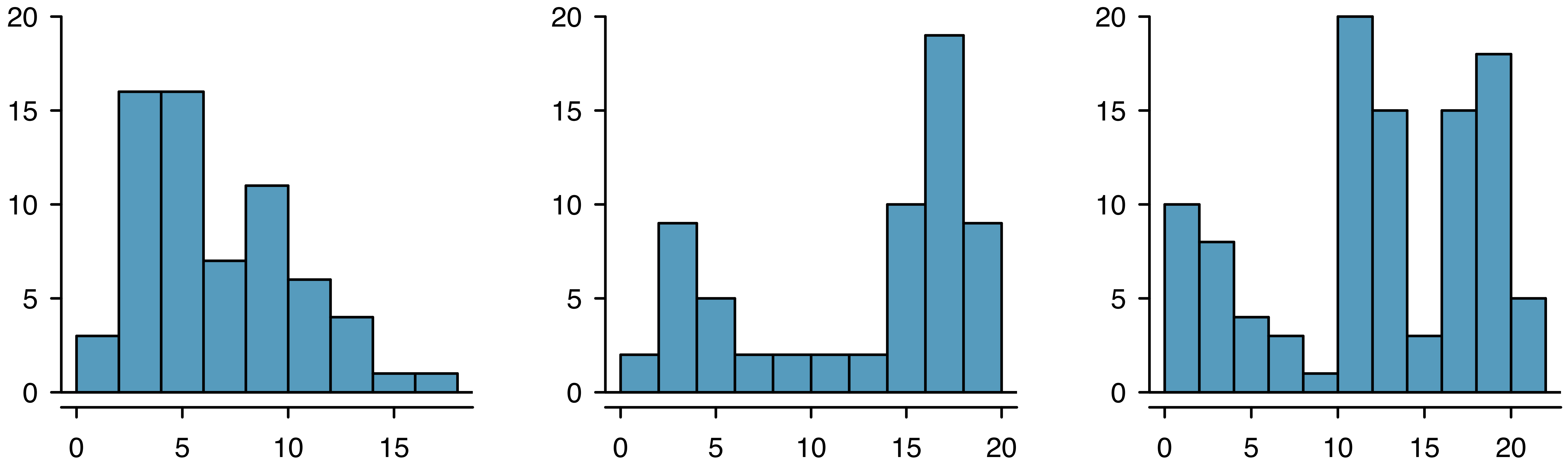

Figure from Open-intro Statistics textbook, Chapter 1.

- Figure shows histograms that have one, two, or three ***prominent peaks.***
- Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively.

Variance and Standard Deviation

- The mean, roughly speaking, describes the center of a data set.
- *Variability* is also of interest: How much does data deviate from the mean?
- Deviation of one data point = distance from the mean = $x_i - x$
- Sample **variance**: $$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- **Standard deviation** is the square root of **variance**, denoted s.
- Useful in considering how far the data are distributed away from the mean.

Mean and Standard Deviation

- Different data can have very similar (even identical) means and standard deviations.
- The distributions on the right all have mean 0, and standard deviation 1.

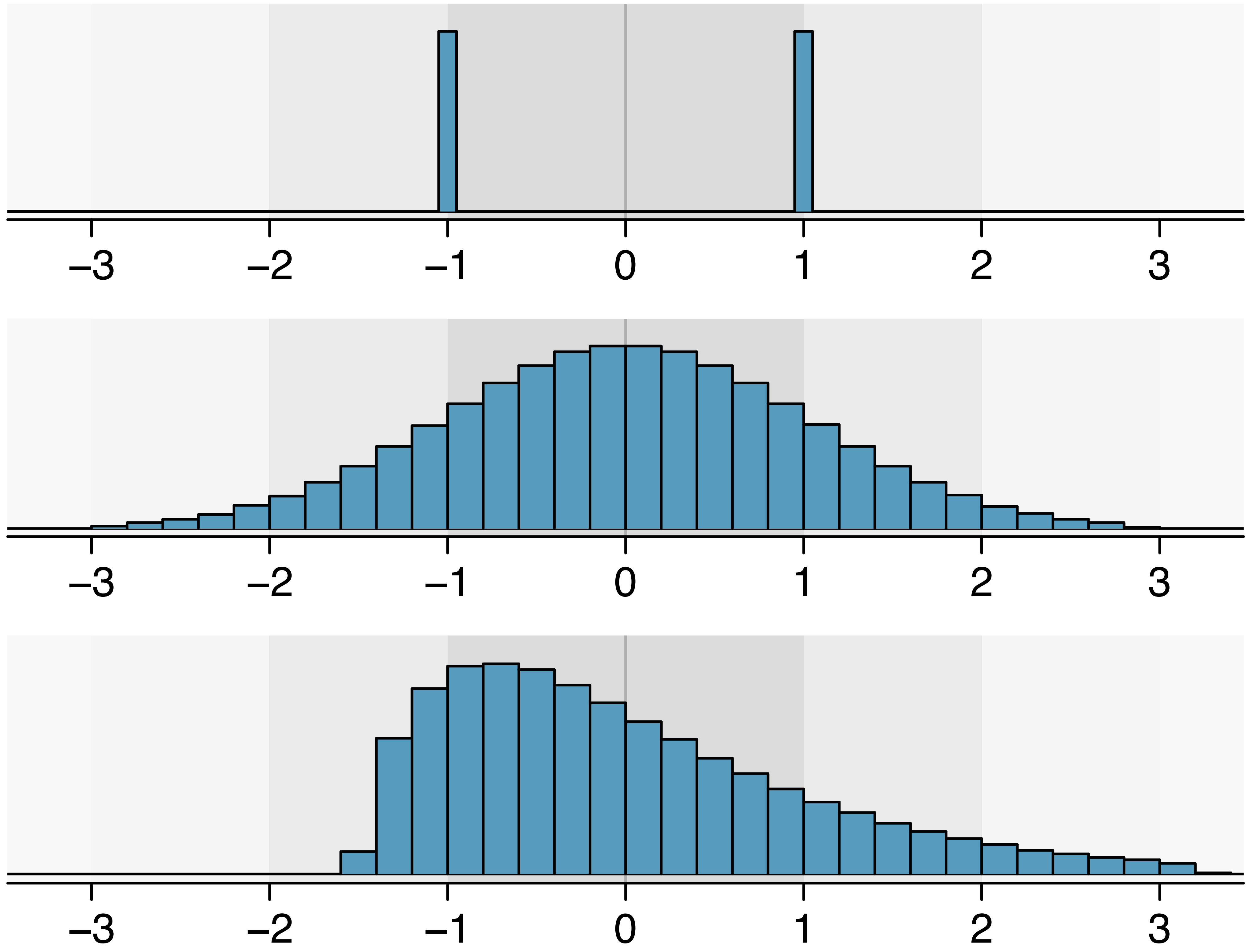

Figure from Open-intro Statistics textbook, Chapter 2.

The Median

- **The median** of a data set of size n is the number "in the middle" so ~ 50% of the data lie above it, and 50% lie below it.

- More precisely, if we sort the data (in increasing or decreasing order, the median is
  - the number in the middle, if n is odd. So 5 is the median of $\{1,7,5\}$
  - the average of the two numbers in the middle, if n is even. So 5.5 is the median of $\{1,7,5,6\}$.
- The median is a **robust statistic**: extreme observations have little effect on it.

- To see this: consider the mean net worth of 1000 individuals, and the median.
  - *How does each change if Bill Gates was in the sample?*

Quartiles and the Interquartile Range (IQR)

- **Quartiles,** like the median, are robust statistics.
- The **first quartile (Q_1)** is the middle number between minimum and median of a data set. $25\ \%$ of the data lies below it.
- The **second quartile** (Q_2) is simply the median. $50\ \%$ of the data lies below it.
- The **third quartile** (Q_3) is the middle number between median and maximum of a data set. $75\ \%$ of the data lies below it.
- **Interquartile range (IQR)**: $IQR = Q_3 - Q_1$

Box Plots

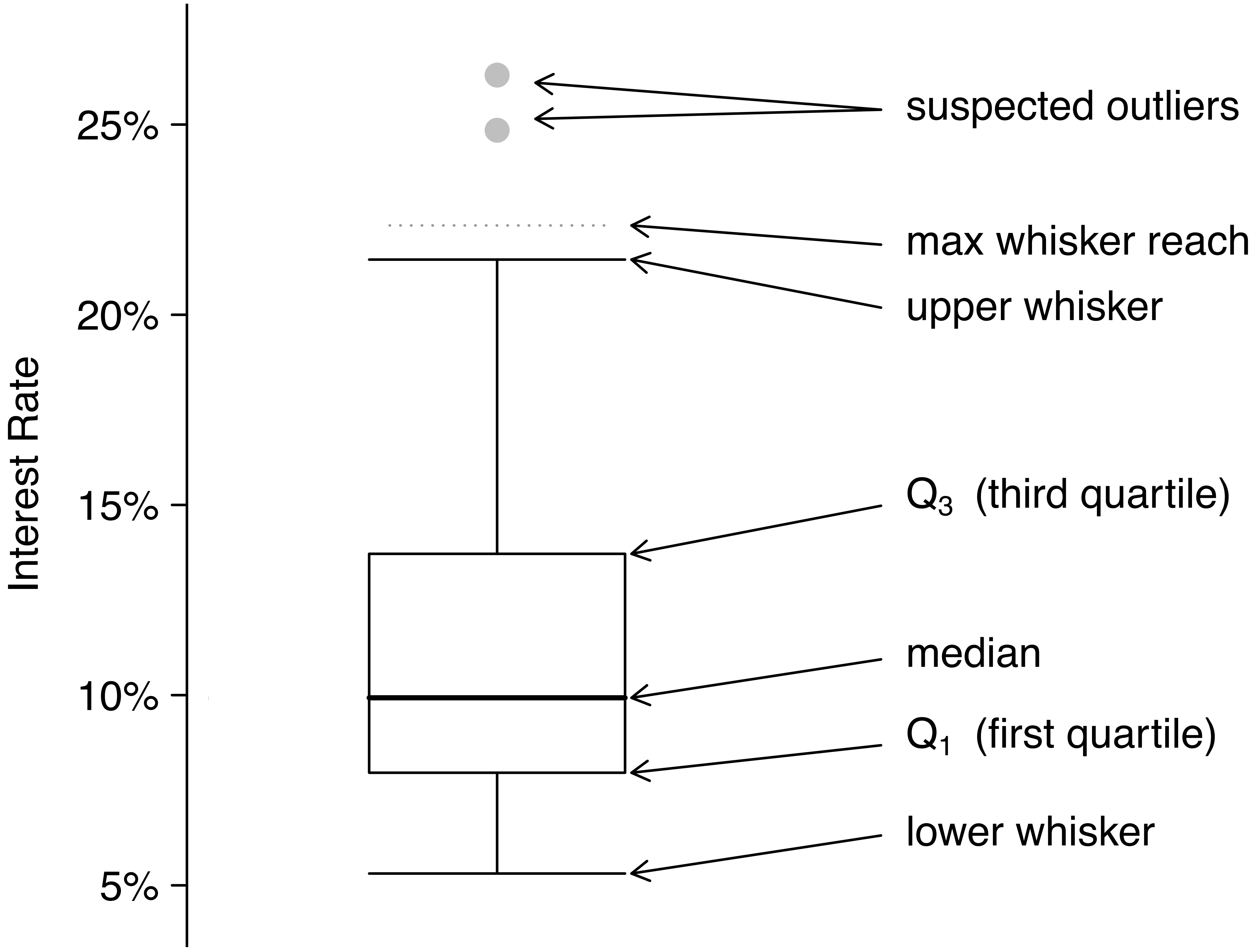

Figure from Open-intro Statistics textbook, Chapter 2.

- A **box plot** summarizes the data using 5 statistics:
- The median
  - Q_1 and Q_3
  - Upper whisker: marks largest data point below $Q_3 + 1.5 \times IQR$ $\leftarrow$ max whisker reach
  - Lower whisker: marks smallest data point above $Q_1 - 1.5 \times IQR$ $\leftarrow$ min whisker reach
- If no data lies below the lower whisker (or above the upper whisker), no need to show that reach.
- It often also plots suspected **outliers**: observations lying beyond the whiskers (unusually distant observations).

Outliers

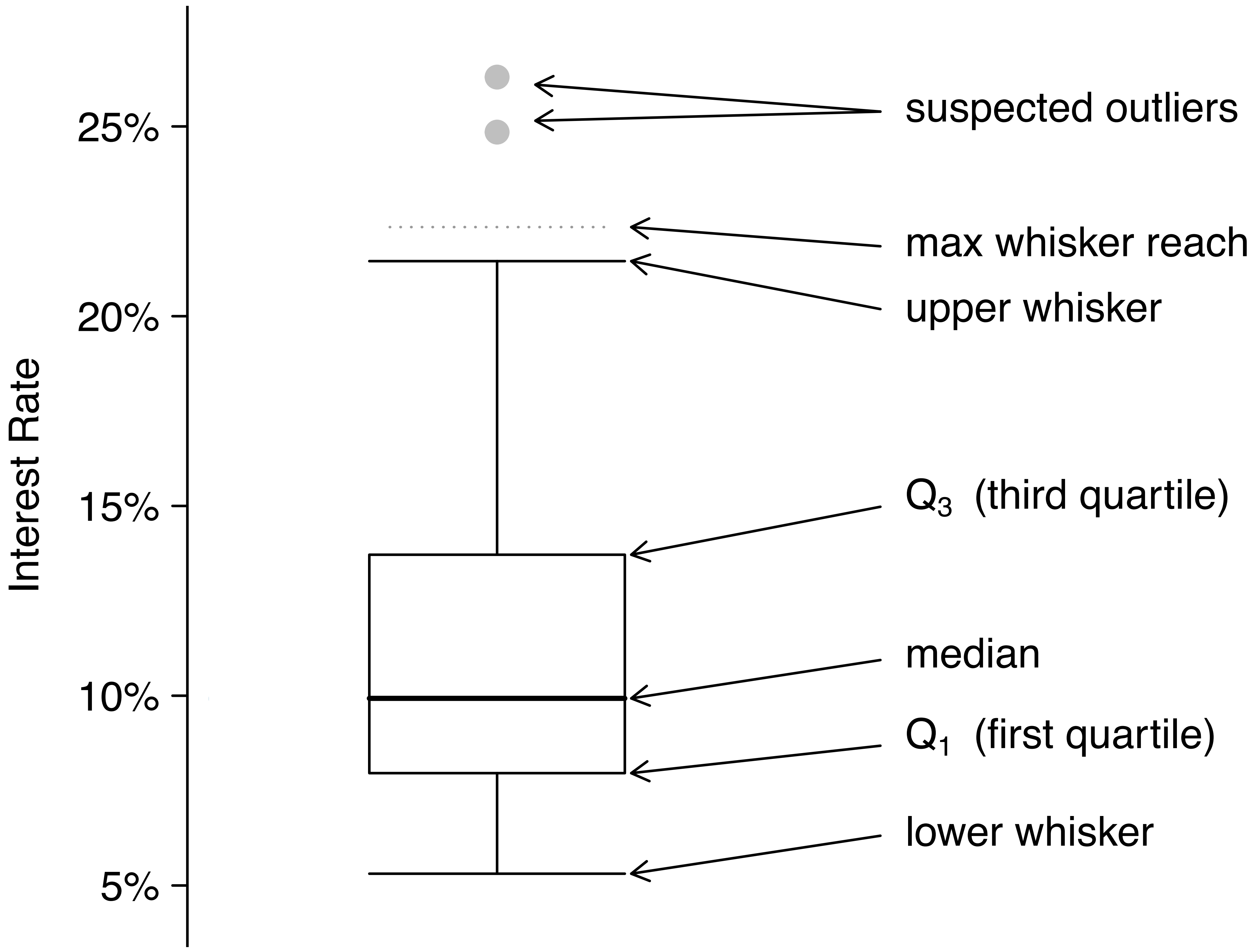

Figure from Open-intro Statistics textbook, Chapter 2.

- An **outlier** is an observation that appears extreme relative to the rest of the data.

- Examining data for outliers is important in:
  - Identifying strong skew in the distribution.
  - Identifying possible data collection or data entry errors.
  - Providing insight into interesting properties of the data.