Different types of Data

A variable is any characteristic observed on the subjects in a study.
Examples: Marital status, Height, Weight, IQ, Sqft, Price, NE.
A variable can be classified as either

- **Categorical** (in Categories), or
- **Quantitative** (Numerical)

- **Qualitative (Categorical) variable** each observation belongs to one of a set of categories order between them. **Categorical values** may be:
  - **binary** where there are two choices, e.g. Gender (Male - Female; Belief in Life After Death (Yes or No); NE - Located in northeast sector of city (1) or not (0))
  - **ordinal** where the categories imply levels with hierarchy or order of preference, e.g. Education (Primary, High school, College)
  - **nominal** where no hierarchy is implied, e.g. Religious Affiliation (Catholic, Jewish,...); Type of Residence (Apartment, Condo, ...).

- **Quantitative variable:** observations on it take numerical values that represent different magnitudes of the variable.
  Examples: Age, Number of Siblings, Annual Income, Selling price, Sqft.
  Quantitative values can be:
  - **discrete** if they are possible values form a set of separate numbers, such as 0,1,2,3,... The set of possible values is not dense. E.g., Number of pets in a household; Number of children in a family; Number of foreign languages spoken by an individual.
  - **continuous** if they are possible values from an interval. The set of possible values is dense. E.g., Height/Weight; Age; Blood pressure.

Exercise

Identify the variable type

1. Number of siblings in a family
2. County of residence
3. Distance (in miles) of commute to school
4. Marital status
5. Length of time to take a test
6. Number of people waiting in line
7. Number of speeding tickets received last year
8. Your dog's weight

Key features of a categorical versus a quantitative variable

- **For categorical variables**: a key feature is the percentage of observations in each of the categories
- **For quantitative variables**: key features are the center (location) and spread (variability)

Example: What percentage of students at a certain college are Democrats?

Example: What's a typical annual amount of precipitation? Is there much variation from year to year?

Proportion & Percentage (Relative Frequencies)

- The **proportion** of the observations that fall in a certain category is the **frequency** (**count**) of observations in that category divided by the total number of observations

$$\frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

- The **percentage** is the proportion multiplied by 100
- Proportions and percentages are also called **relative frequencies**

Example

Table classifies the 630 parliamentary seats of the Italian chamber of deputies by coalition (2018 elections).

Coalition	**Seats**		
	Freq.	Prop.	Perc.
Centre-right coalition	265	0.4206	42.06
Five Star Movement (M5S)	227	0.3603	36.03
Centre-left coalition	122	0.1937	19.37
Free and Equal (LeU)	14	0.0222	2.22
Associative Movement Italians Abroad (MAIE)	1	0.0016	1.60
South American Union Italian Emigrants (USEI)	1	0.0016	0.16
Total	**630**	**1**	**100**

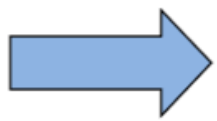
For example, for M5S:
227 is the **frequency**; **0.3603** $= 227/630$ is the **proportion and relative frequency**; while **36.03** is the **percentage** $0.3603 \times 100 = 36.03\%$.

Frequency table

A **frequency table** is a listing of possible values for a variable, together with the number of observations and/or relative frequencies for each value.

Raw data

Code	Gender
000001	F
000002	M
...	...
100000	F



Frequency table

Gender	n_i	f_i	p_i
F	1000	0.01	1
M	99000	0.99	99
	100000	1.00	100

Example

A stock broker has been following different stocks over the last month and has recorded whether a stock is up, the same, or down in value. The results were:

Performance of stock	Up	Same	Down
Count	21	7	12

- What are the subjects?
- What is the variable of interest?
- What type of variable is it?
- Add proportions to this frequency table.

Describe data using graphical summaries

Distribution

- A graph or frequency table describes a distribution.
- A distribution tells us the possible values/categories a variable takes as well as the occurrence of those values (frequency or relative frequency or percentage).

In the 2008 General Social Survey, 2020 respondents answered the question, "How many children have you ever had?"
The results were

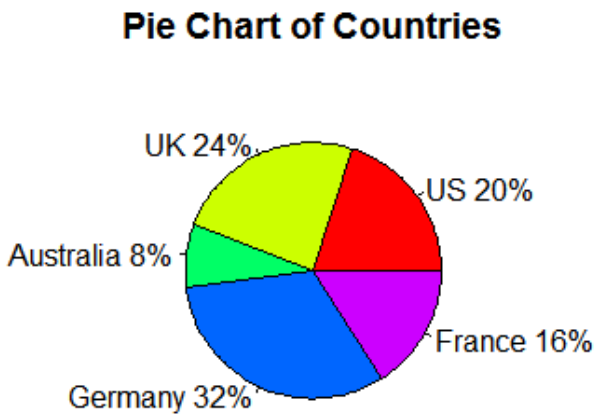
No. children	0	1	2	3	4	5	6	7	8+	Total
Count	521	323	524	344	160	77	30	19	22	2020

Graphs for categorical data: bar graphs and pie charts

Use pie charts and bar graphs to summarize categorical variables:

- **Pie chart**
  - a circle where each category is represented as a "slice of the pie"
  - area of the pie represents the percentage of that category
- **Bar chart**
  - bar graphs display a vertical bar for each category
  - the height of the bar for each category is equal to the frequency (number of observations) or percentages (relative frequencies) in the category
  - bar graphs are called Pareto Charts when the categories are ordered by their frequency, from the tallest bar to the shortest bar

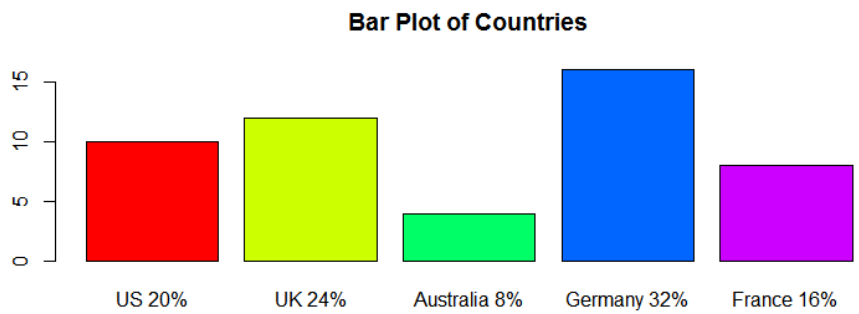
Graphs for categorical data: bar graphs and pie charts

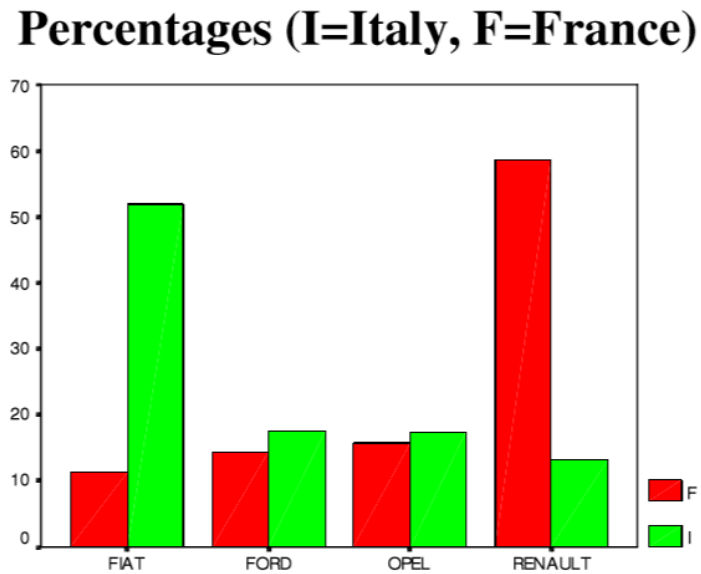


Pie chart

- It is easier to compare one category with the whole
- It may not be suitable for too many categories

Graphs for categorical data: bar graphs and pie charts



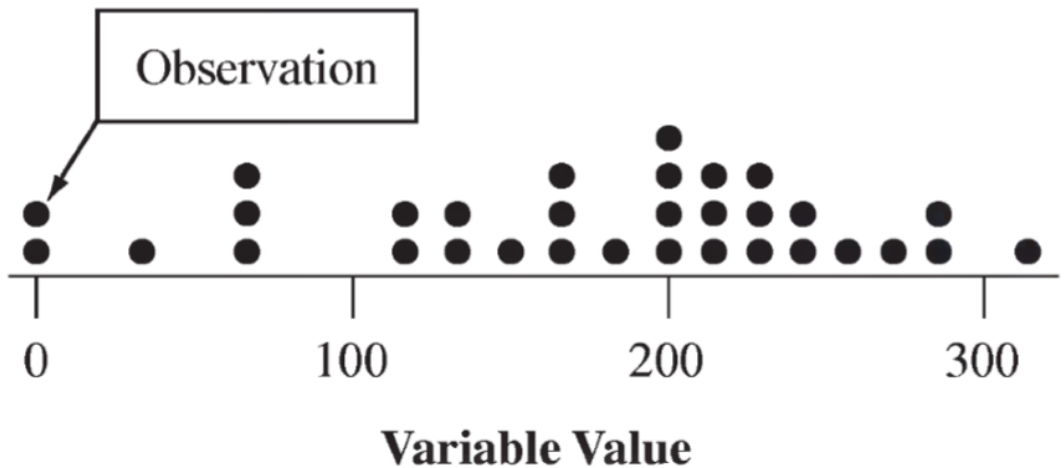


Bar chart

- It is easier to compare categories

Graphs for quantitative data: dot plot

- **Dotplot**
  Shows a dot for each subject (observation) placed above its value on a number line. To construct a dot plot
  - Draw a horizontal line and label it with the name of the variable.
  - Mark regular values of the variable on it.
  - For each observation, place a dot above its value on the number line.



It is useful to show the relative positions of the data.

Graphs for quantitative data: histograms

- **Histogram:** it displays the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

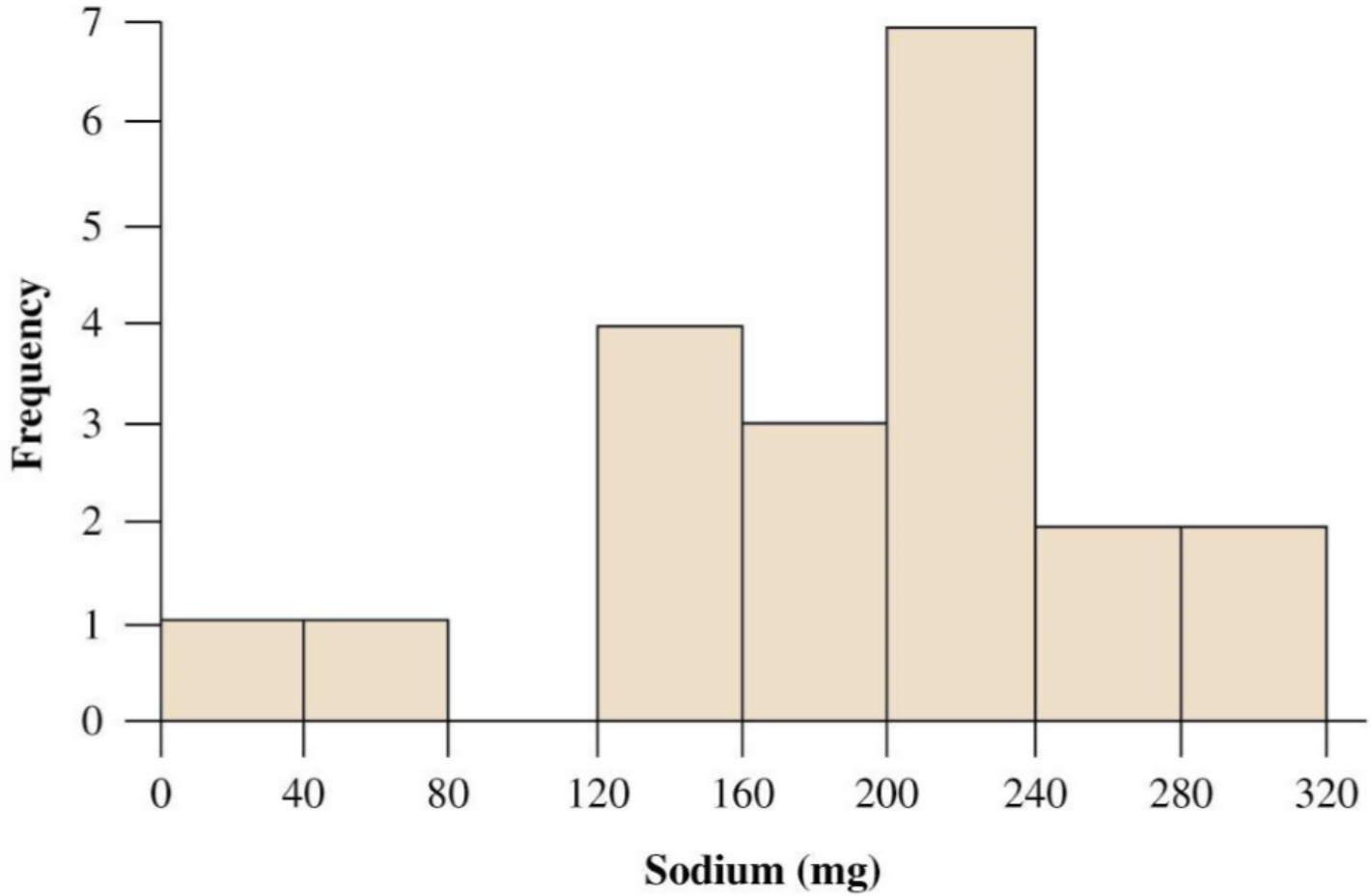
Steps for constructing a histogram

1. Divide the range of the data into intervals of equal width
2. Count the number of observations in each interval, creating a frequency table
3. On the horizontal axis, label the values or the endpoints of the intervals.
4. Draw a bar over each value or interval with height equal to its frequency (or proportion or percentage), values of which are marked on the vertical axis.
5. Label and title appropriately

TABLE 2.4: Frequency Table for Sodium in 20 Breakfast Cereals.

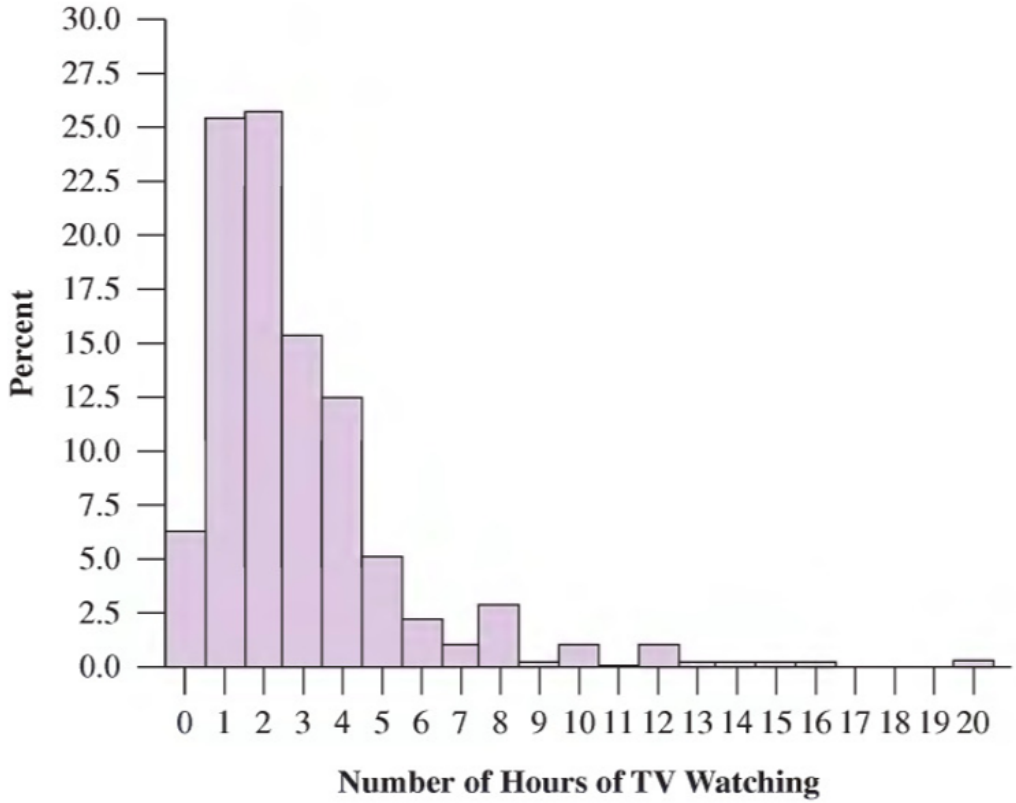
The table summarizes the sodium values using eight intervals and lists the number of observations in each, as well as the proportions and percentages.

Interval	Frequency	Proportion	Percentage
0 to 39	1	0.05	5%
40 to 79	1	0.05	5%
80 to 119	0	0.00	0%
120 to 159	4	0.20	20%
160 to 199	3	0.15	15%
200 to 239	7	0.35	35%
240 to 279	2	0.10	10%
280 to 319	2	0.10	10%


Frequency
0
1
2
3
4
5
6
7
0
40
80
120
160
200
240
280
320
Sodium (mg)

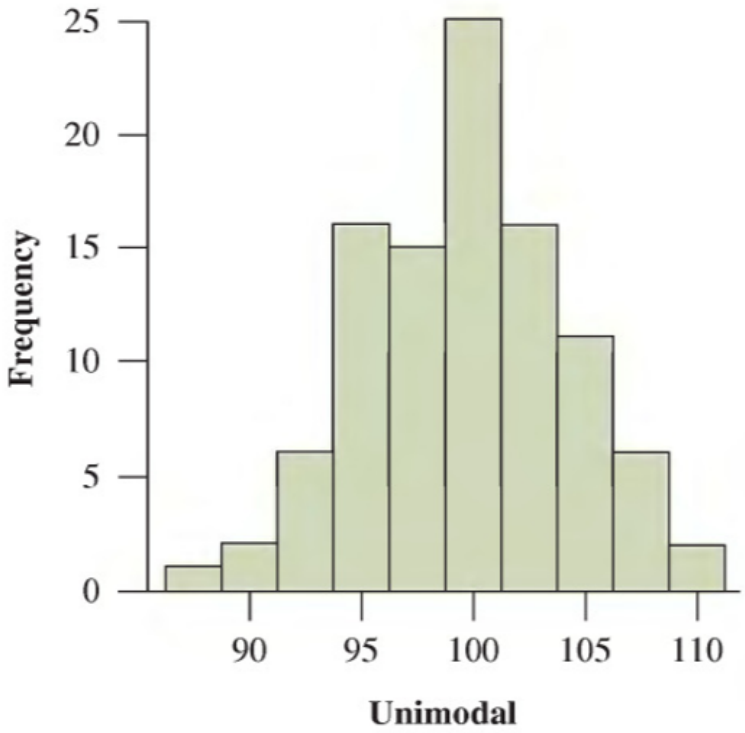
Histograms are sometimes used for discrete quantitative variables drawing a separate bar for each possible value.

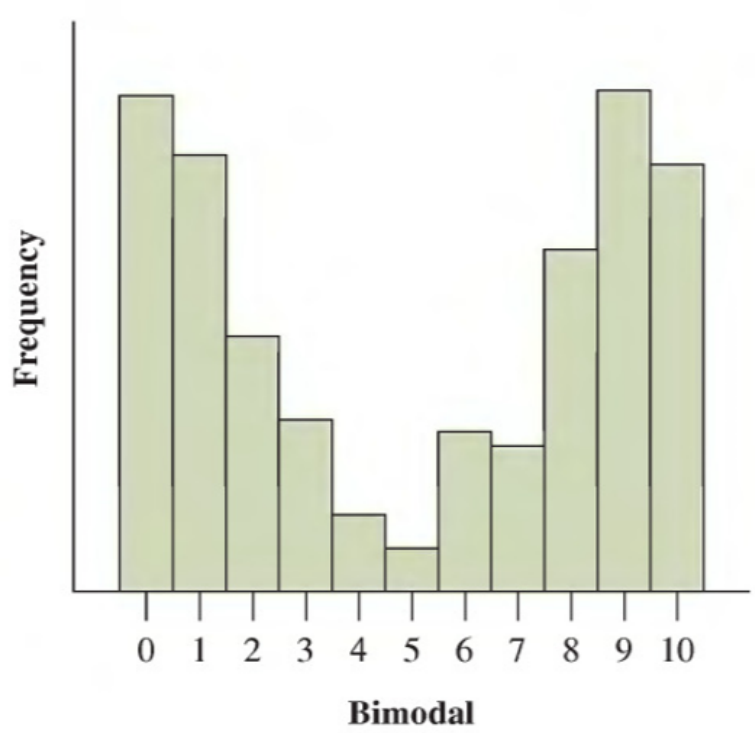
Example. The 2004 General Social Survey asked, "On an average day, about how many hours do you personally watch television?"



Shape of a distribution

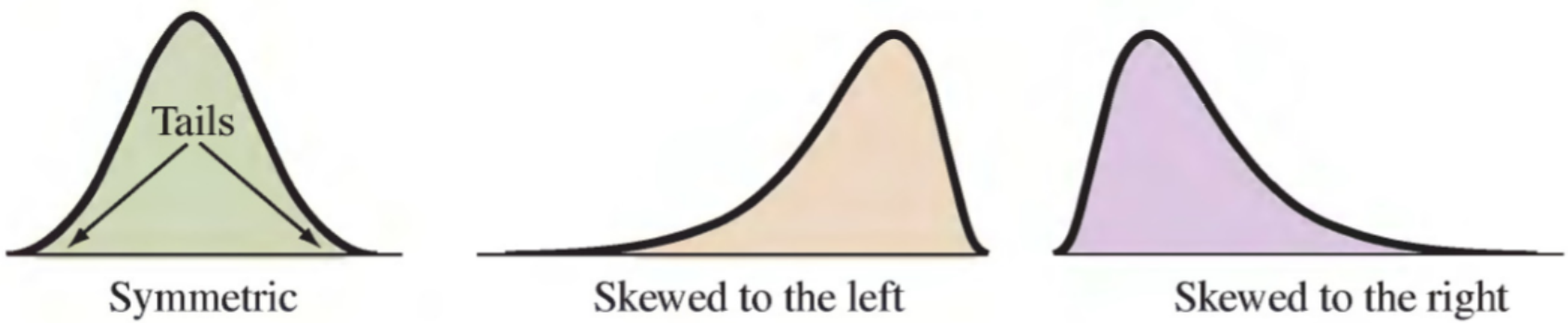
- **Unimodality:** A distribution of data with a single mound is called **unimodal**. The highest point is at the **mode**.
- **Bimodality:** A distribution with two distinct mounds is called **bimodal**.

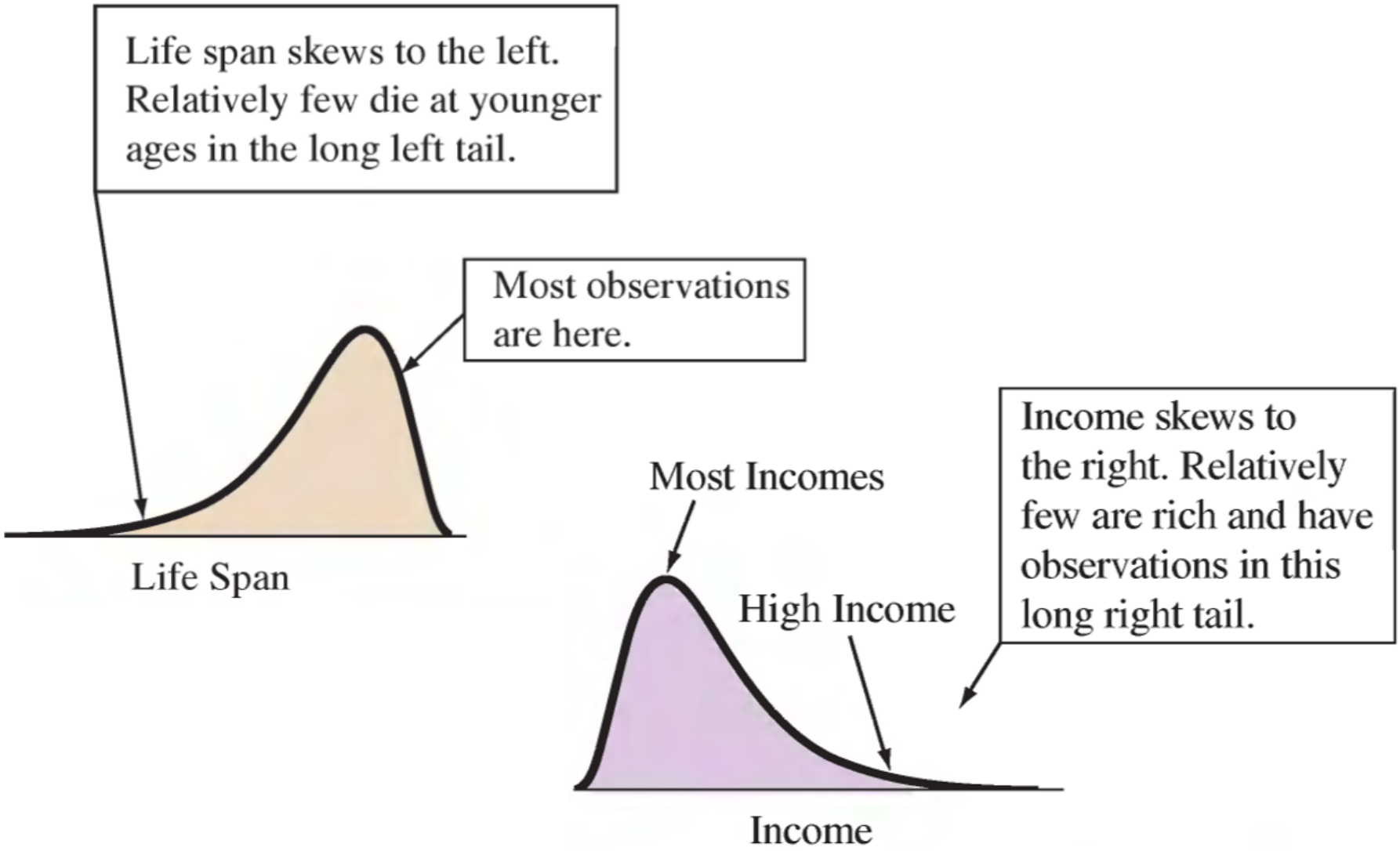




Symmetry and Skewness

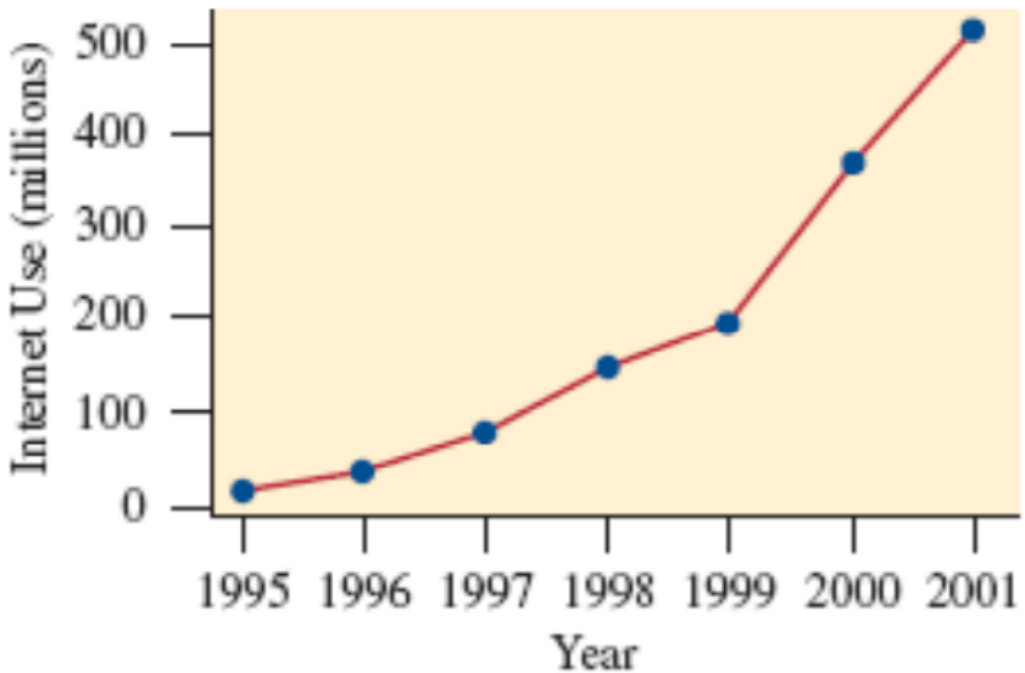
The shape of the distribution is often described as symmetric or skewed.
A distribution is symmetric if the side of the distribution below a central value is a mirror image of the side above that central value, otherwise is said skewed.




Life span skews to the left. Relatively few die at younger ages in the long left tail.
Most observations are here.
Life Span
Most Incomes
High Income
Income skews to the right. Relatively few are rich and have observations in this long right tail.
Income

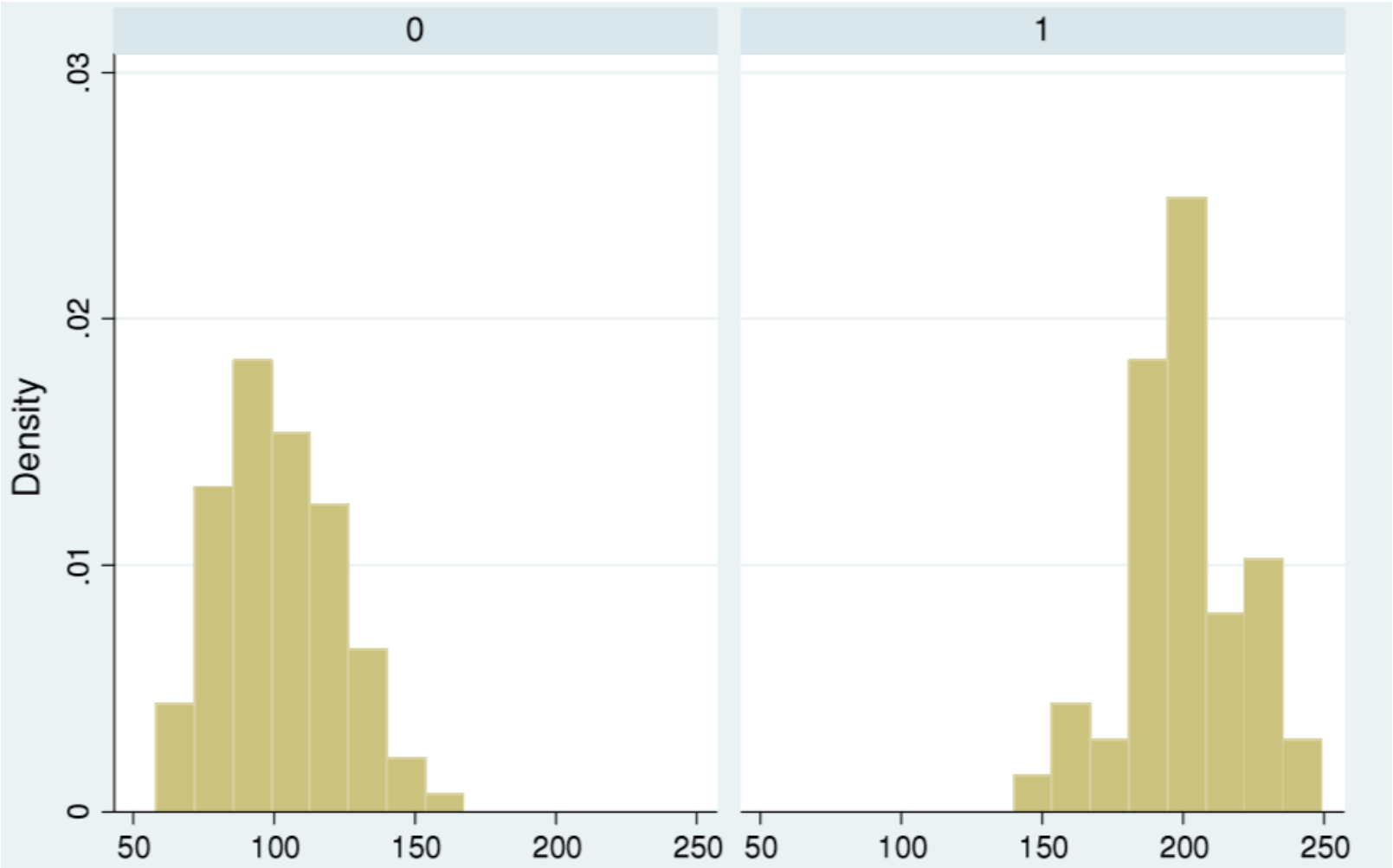
Single Time Series

- **Time Plot:** Used for displaying a time series, a data set collected over time. Plots each observation on the vertical scale against the time measured on the horizontal scale. Points are usually connected. Common patterns in the data over time, known as **trends**, should be noted.



Measuring the Center of Quantitative Data

Distribution of reselling home prices (1000$) in city 0 and city 1



Where are more expensive the houses?

In order to give an answer to the previous question we have to identify a single value (center, location) who represent the whole distribution.

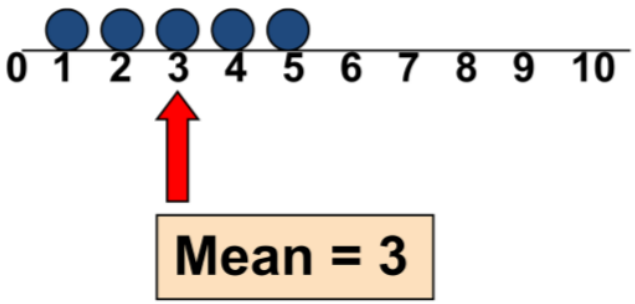
There are several ways to compute the center of a distribution:

- **Mean**
- **Median**
- **Mode**

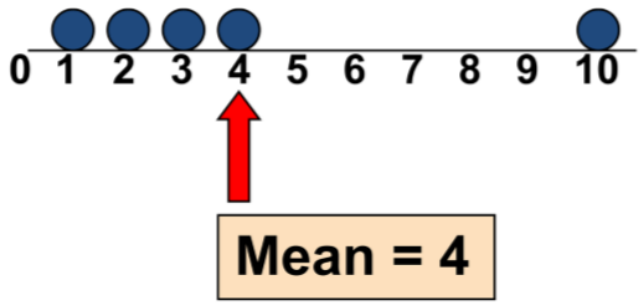
Calculating the mean (only quantitative variables)

The (arithmetic) mean is the sum of the observations divided by the number of observations

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



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Affected by extreme values

Properties

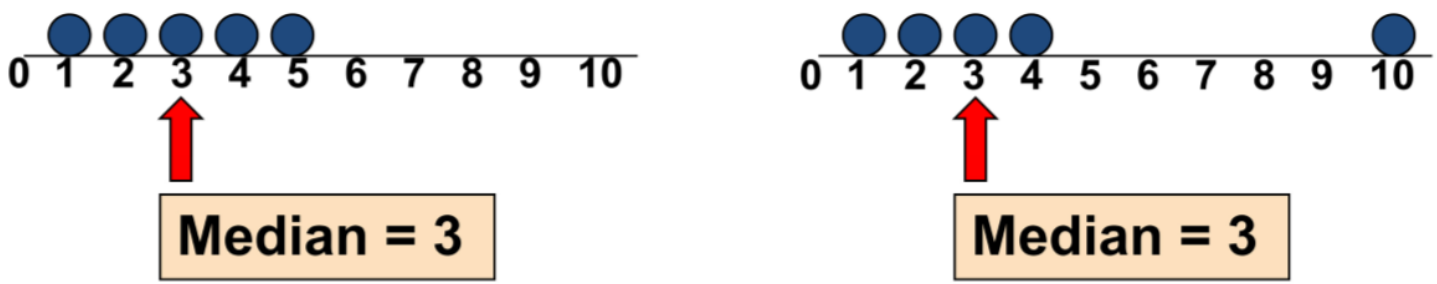
1. **it is internal**, the mean is always in between the minimum and the maximum;
2. **it is the fair value**, i.e. substituting the value of each observation with the mean we preserve the total amount.
3. **the sum of deviations** $(x_i - \bar{x})$ **is zero**;
4. **it is linear**. In formulas: if $\bar{x}$ is the mean of $x_1, x_2, \ldots, x_n$, and $\bar{y}$ is the mean of $y_1, y_2, \ldots, y_n$, where $y_i = a + bx_i$, then $\bar{y} = a + b\bar{x}$.

Calculating the median

The median is the middle value of the observations when they are ordered from the smallest to the largest (or from the largest to the smallest).

Steps to finding the median for a set of data:

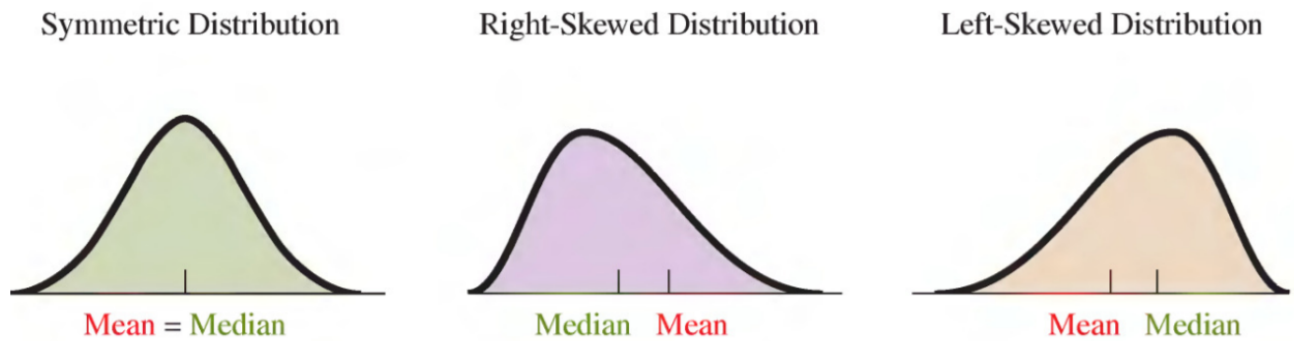
1. Order the observations
2. Find the location of median in the ordered data by $(n+1)/2$
3. The value that represents the location found in Step 2 is the median.
4. **NOTE**: if the sample size is an odd number then the median is the middle observation. If sample size is an even number, then the median is the average of the two middle observations . The result may or may not be an observed value.



NOT affected by extreme values

Comparing the Mean & Median

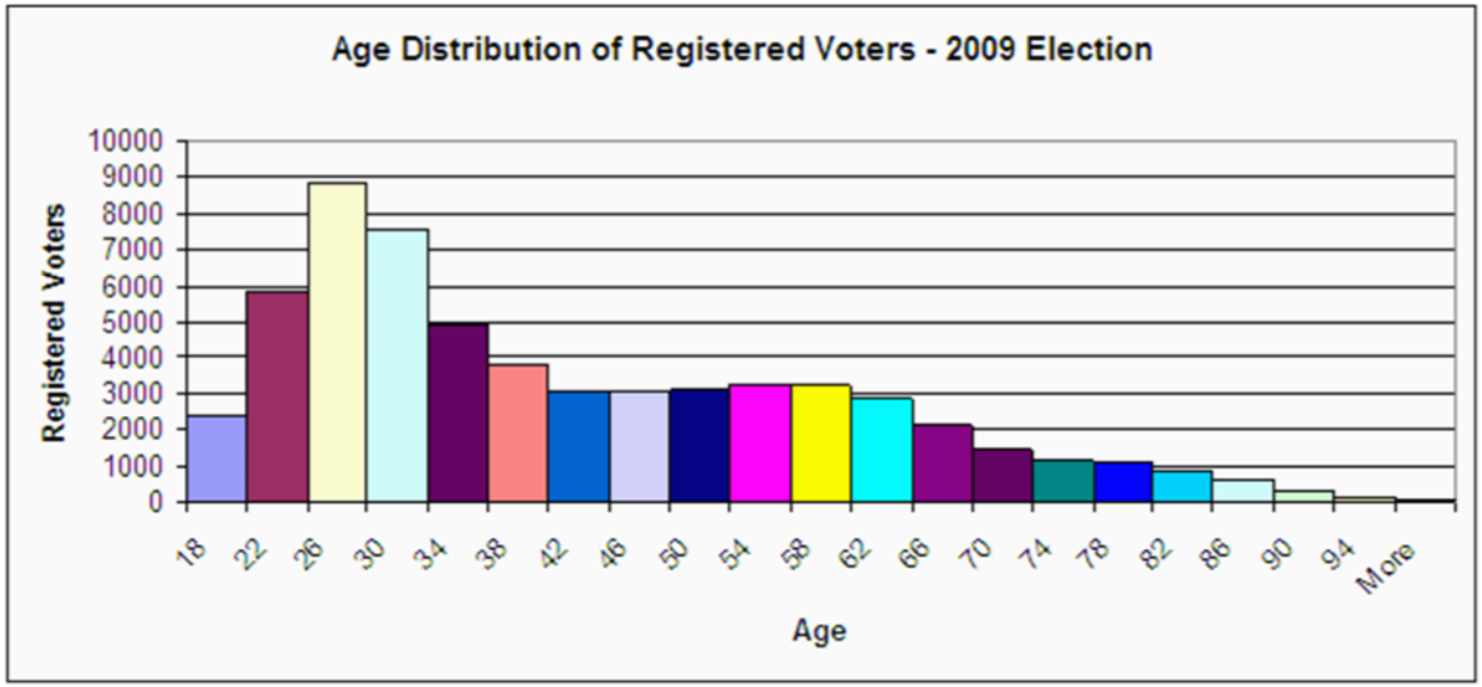
- **Mean, median and mode are usually not equal**
  In a skewed distribution, the mean is farther out in the long tail than is the median.



Example: in the income distribution usually the mean is higher than the median.

For skewed distributions, the **median** is preferred because it is better representative of a typical observation.

Cambridge municipal election



- Average age of all registered voters: 43.7
- Median age of all registered voters: 38.4

Comparing the Mean & Median

Resistant Measures

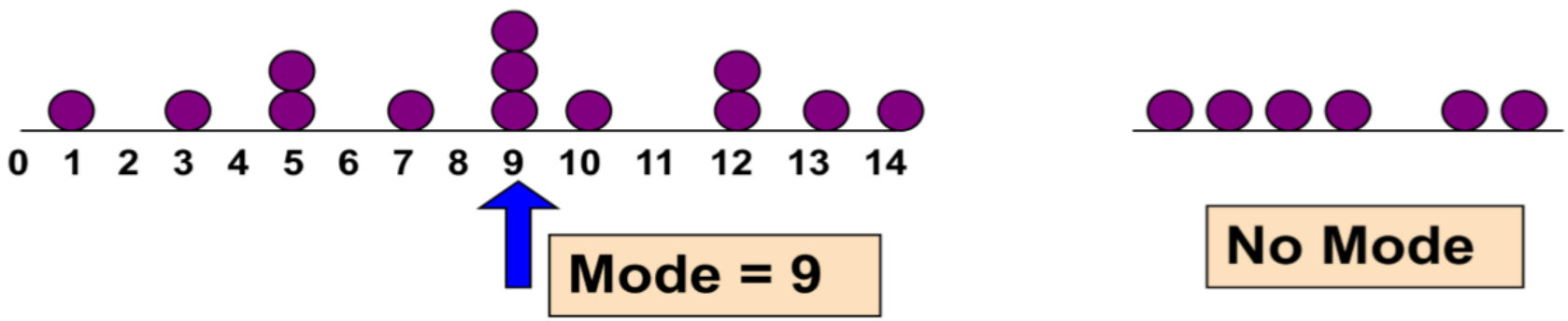
A numerical summary measure is resistant if extreme observations (outliers) have little, if any, influence on its value.

- **Mean is affected by extreme values. It is not resistant to outliers**.
  Given this data set, $95, 78, 69, 91, 82, 76, 76, 86, 88, 80$, the mean is $\bar{x} = (95 + 78 + 69 + 91 + 82 + 76 + 76 + 86 + 88 + 80)/10 = \mathbf{82.1}$. If the entry **69** is mistakenly recorded as **9**, the mean would be **76.1**, which is very different from 82.1.
- **Median is resistant to outliers, i.e. it is not affected by extreme values**
  Ordered original data set is: $69, 76, 76, 78, 80, 82, 86, 88, 91, 95$. With $n = 10$, the median is the average of the fifth (80) and sixth (82) ordered value and the median is **81**. Ordered new data set (with 69 coded as 9) is: $9, 76, 76, 78, 80, 82, 86, 88, 95$ where the median is still **81**.

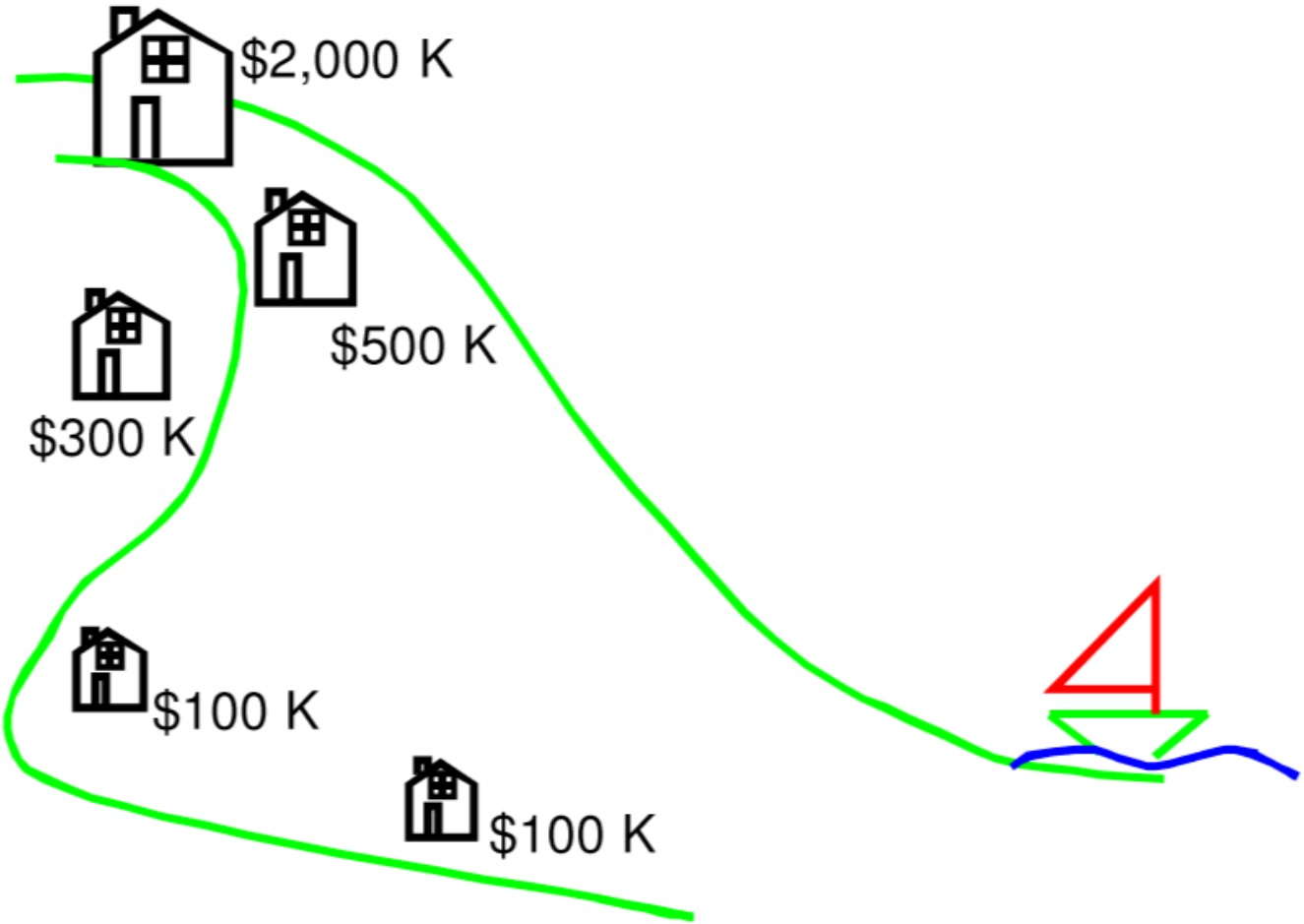
The mode of a distribution (all types of variables)

Mode = value/category that occurs most often

- The mode is most often used with categorical data
- Not affected by extreme values
- There may be no mode
- There may be several modes



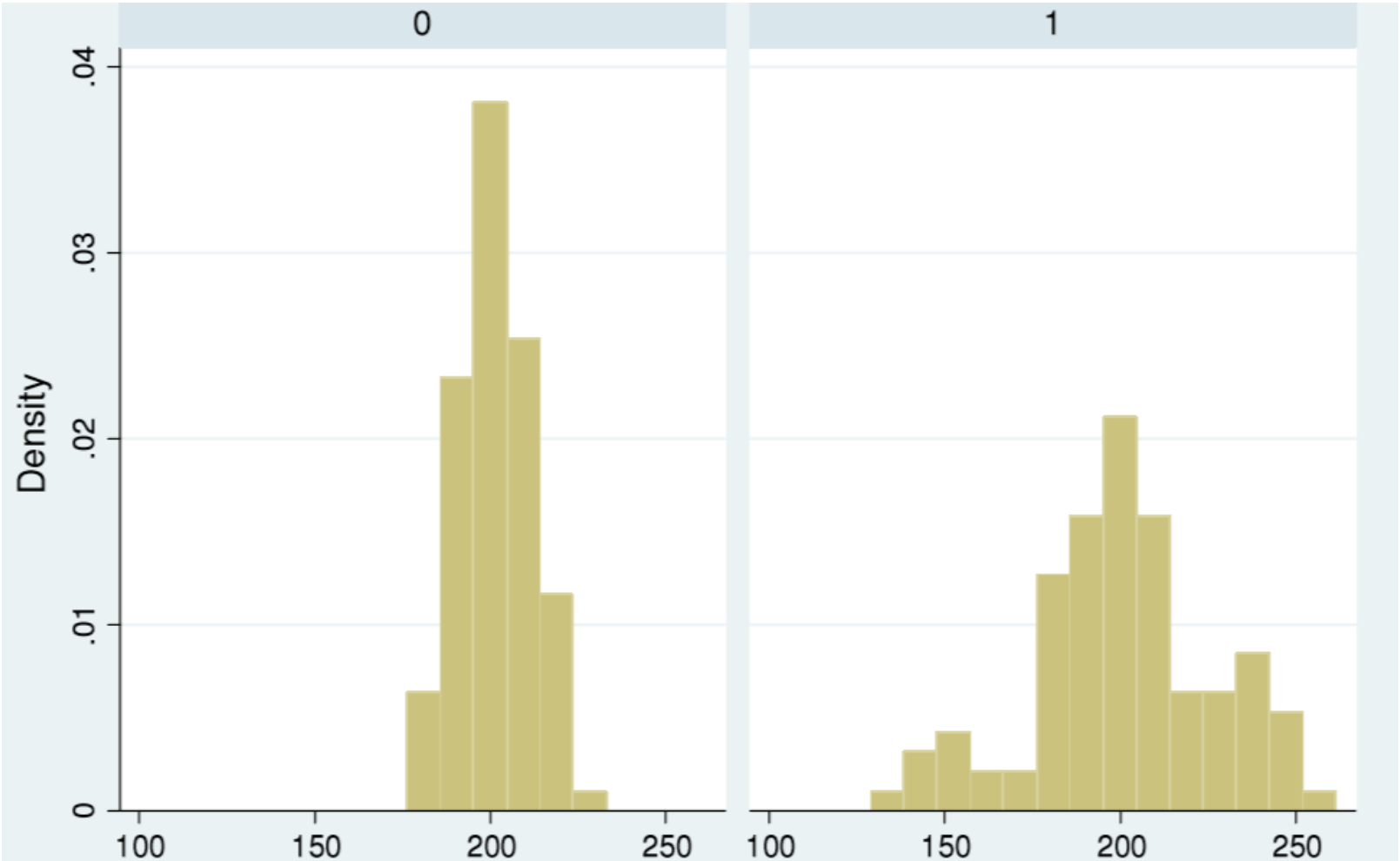
Review example: Five houses on a hill by the beach



Mean: ($3,000,000/5) = $600,000
Median: middle value of ranked data = $300,000
Mode: most frequent value = $100,000

Spread (variability) of Quantitative Data

Distribution of reselling home prices (1000$) in city 0 and city 1



Where is a greater dispersion (spread, variability, inequality)?

Range

One way to measure the spread is to calculate the range.

The range is the difference between the largest and smallest values in the data set:

$$\textbf{Range} = \textbf{max} - \textbf{min}$$

The range is simple to compute and easy to understand, but it uses **only** the extreme values and ignores the other values. Therefore, it's affected severely by outliers.

Calculate the standard deviation

- Each data value has an associated deviation from the mean, $x_i - \bar{x}$
- A deviation is positive if it falls above the mean and negative if it falls below the mean
- The sum of the deviations is always zero (i.e. the mean is the center)
- A measure of variation can be obtained by summarizing the *squared deviations* of each observation from the mean and calculating an *average* of these squared deviations. In formulas

$$\textbf{Variance}: \ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The original unit of measurement can be recovered by computing

$$\textbf{Standard Deviation}: \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Steps

1. Find the mean
2. Find the deviation of each value from the mean
3. Square the deviations
4. Sum the squared deviations
5. Divide the sum by n
6. Compute the square root

Example. Metabolic rates of 7 men (cal./24hr.) :

$$1792, 1666, 1362, 1614, 1460, 1867, 1439$$

$$\begin{aligned} \bar{x} &= \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} \\ &= \frac{1200}{7} = 1600\text{cal./24hr.} \end{aligned}$$

Observations	Deviations	Squared deviations
1792	$1792 - 1600 = 192$	$(192)^2 = 36864$
1666	$1666 - 1600 = 66$	$(66)^2 = 4356$
1362	$1362 - 1600 = -238$	$(-238)^2 = 56644$
1614	$1614 - 1600 = 14$	$(14)^2 = 196$
1460	$1460 - 1600 = -140$	$(-140)^2 = 19600$
1867	$1867 - 1600 = 267$	$(267)^2 = 71289$
1439	$1439 - 1600 = -161$	$(-161)^2 = 25921$
	sum $= 0$	sum $= 214870$

- $\sigma^2 = \dfrac{214870}{7} = 30695.71$
- $\sigma = \sqrt{30695.71} = 175.202$cal./24hr.

When the dataset is a sample, the two previous indices are adjusted dividing the sum of the squared deviations by $n-1$ instead of n. In formulas

- **Sample Variance**: $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Sample Standard Deviation**: $\sigma^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

The motivation of this modification will be explained in the third part.

- It is important to note that $s^2[s]$ has exactly the same properties as $\sigma^2[\sigma]$ even if its value is always smaller.
- σ^2 and σ are also called *population variance* and *population standard deviation*, respectively.

Example

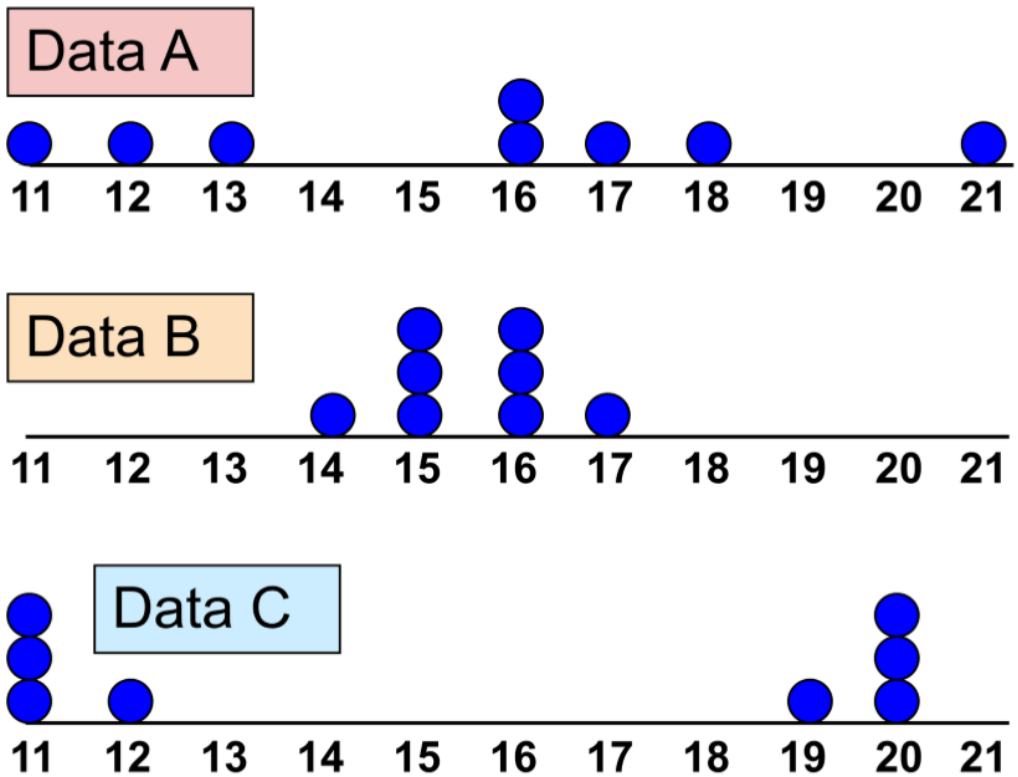
Data (x_i): $10, 12, 14, 15, 17, 18, 18, 24$

- $n = 8$;
- Mean$=\bar{x} = 16$

Observations	Deviations	Squared deviations
10	$10 - 16 = -6$	$(-6)^2 = 36$
12	$12 - 16 = -4$	$(-4)^2 = 16$
14	$14 - 16 = -2$	$(-2)^2 = 4$
15	$15 - 16 = -1$	$(-1)^2 = 1$
17	$17 - 16 = 1$	$(1)^2 = 1$
18	$18 - 16 = 2$	$(2)^2 = 4$
18	$18 - 16 = 2$	$(2)^2 = 4$
24	$24 - 16 = 8$	$(8)^2 = 64$
	sum $= 0$	sum $= 130$

- $s^2 = \dfrac{130}{7} = 18.5714$
- $s = \sqrt{18.5714} = 4.3095$: a measure of the "average" scatter around the mean.

Example



- **Data A**.
  $\bar{x} = 15.5$;
  $s = 3.338$
- **Data B**.
  $\bar{x} = 15.5$;
  $s = 0.926$
- **Data C**.
  $\bar{x} = 15.5$;
  $s = 4.567$

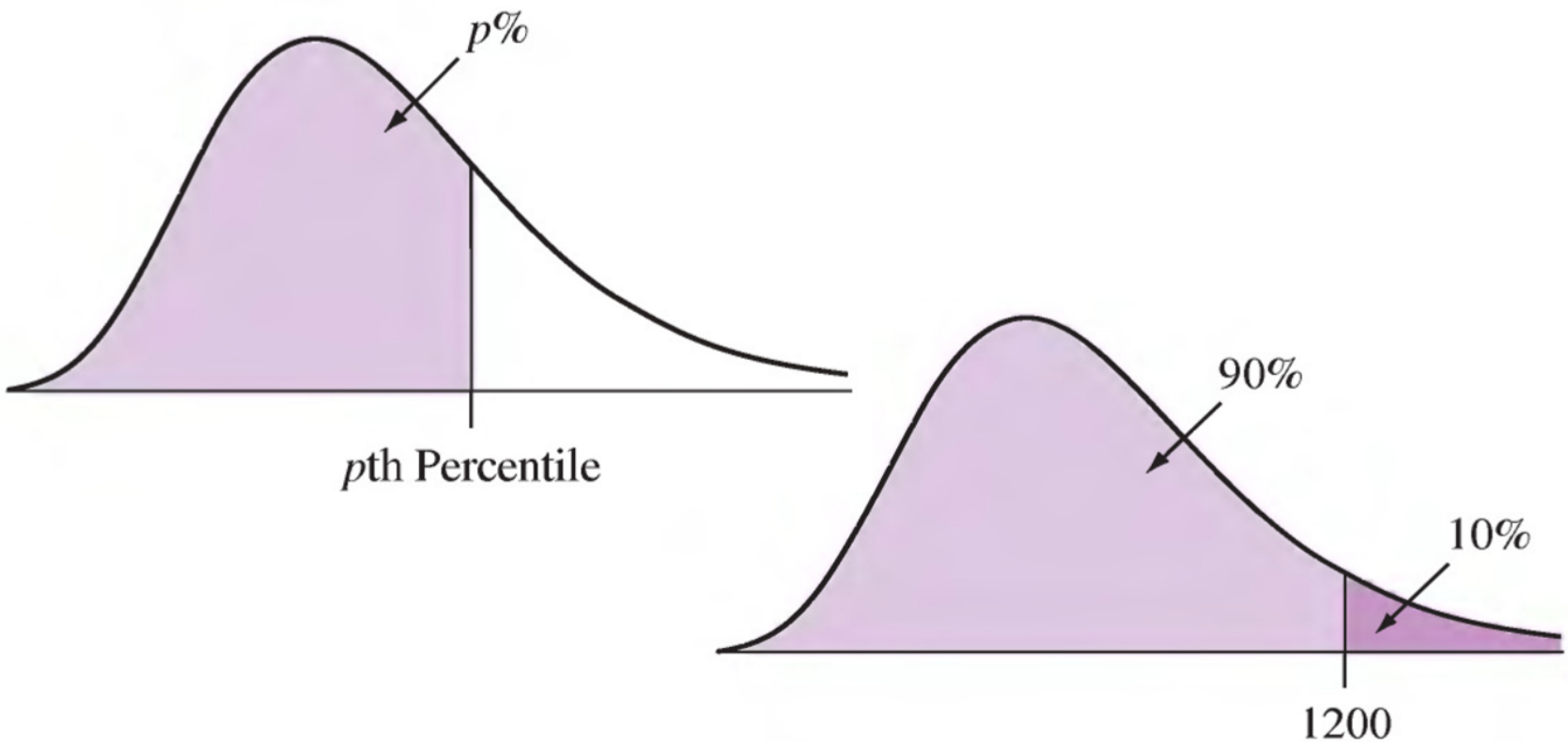
Properties of the standard deviation

- $\sigma[s]$ measures the spread (variability) of the data.
- $\sigma[s] = 0$ only when all observations have the same value, otherwise $\sigma[s] > 0$. As the spread of the data increases, $\sigma[s]$ gets larger.
- $\sigma[s]$ has the same units of measurement as the original observations. The variance $\sigma^2[s^2]$ has units that are squared.
- $\sigma[s]$is not resistant. A few extreme values can greatly increase its value.
- if $\sigma_x^2[s_x^2]$ is the variance of $x_1, x_2, \ldots, x_n$, and $\sigma_y^2[s_y^2]$ is the variance of $y_1, y_2, \ldots, y_n$, where $y_i = a + bx_i$, then $\sigma_y^2 = b^2\sigma_x^2[s_y^2 = b^2 s_x^2]$ and $\sigma_y =\mid b \mid \sigma_x[s_y =\mid b \mid s_x]$.

Measures of Position

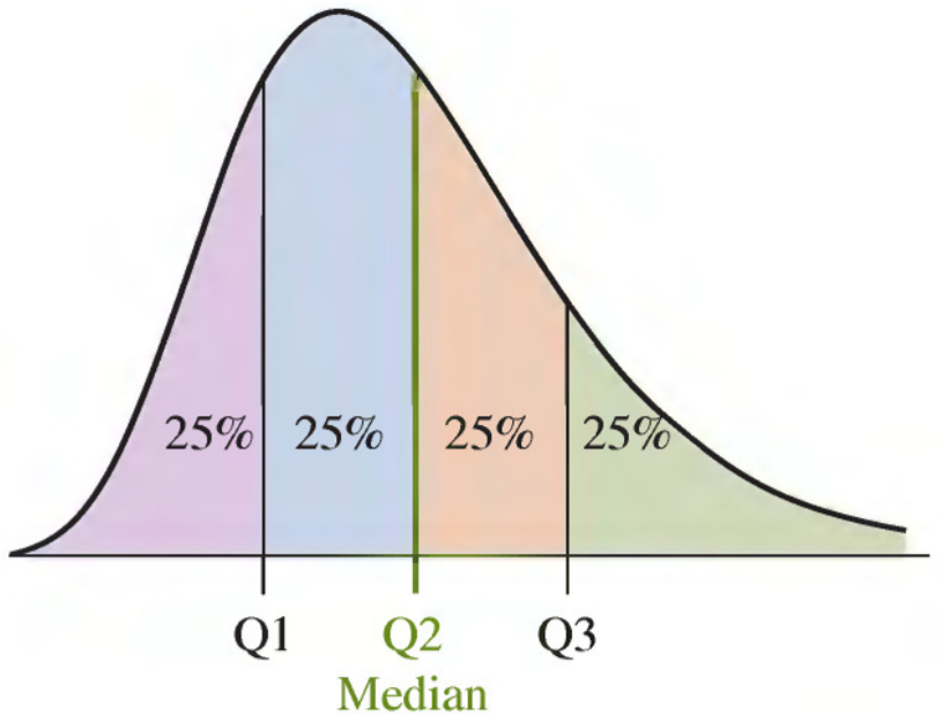
Percentile

The p^{th} percentile is a value such that p percent of the observations fall below or at that value.



Quartiles

The Quartiles Split the Distribution Into Four Parts. 25% is below the first quartile (Q1), 25% is between the first quartile and the second quartile (the median, Q2), 25% is between the second quartile and the third quartile (Q3), and 25% is above the third quartile.



SUMMARY: Finding Quartiles

- Arrange the data in order.
- Consider the median. This is the second quartile, Q2.
- Consider the lower half of the observations (excluding the median itself if n is odd).
- The median of these observations is the first quartile, Q1.
- Consider the upper half of the observations (excluding the median itself if n is odd).
- Their median is the third quartile, Q3.

Example: Cereal Sodium Data

Consider the sodium values for the 20 breakfast cereals. What are the quartiles for the 20 cereal sodium values?
The sodium values, in ascending order, are:

$$\overbrace{\begin{array}{cccccccccc} & & & & \mathbf{130} & \mathbf{140} & & & & \\ 0 & 50 & 70 & 100 & & & 140 & 150 & 160 & \mathbf{180} \end{array}}^{\mathbf{Q1 = 135}}$$

$$\underbrace{\begin{array}{cccccccccc} \mathbf{180} & 180 & 190 & 200 & \mathbf{200} & \mathbf{210} & 210 & 220 & 290 & 340 \end{array}}_{\mathbf{Q3 = 205}}$$

- The median of the 20 values is the average of the 10^{th} and 11^{th} observations, 180 and 180, which is Q2 = 180 mg.
- The first quartile Q1 is the median of the 10 smallest observations (in the top row), which is the average of 130 and 140, Q1 = 135mg.
- The third quartile Q3 is the median of the 10 largest observations (in the bottom row), which is the average of 200 and 210, Q3 = 205 mg.

Interquartile range (IQR)

It is the distance between the third quartile and first quartile:

$$\begin{aligned}\text{IQR} = Q_3 - Q_1 &= \text{upper quartile - lower quartile} \\ &= \text{75th percentile - 25th percentile}\end{aligned}$$

- IQR gives spread of middle 50% of the data and is not affected by extreme values. It is thus a resistant measure of variability.

Detecting Potential Outliers

Examining the data for unusual observations, such as outliers, is important in any statistical analysis. Is there a formula for flagging an observation as potentially being an outlier?

- **The $1.5\times$IQR Criterion for Identifying Potential Outliers**. An observation is a potential outlier if it falls more than $1.5\times$IQR below the first quartile or more than $1.5\times$IQR above the third quartile.

Example: Cereal Sodium Data

- For the breakfast cereal sodium data has Ql=135 and Q3=205. So, IQR = Q3-Ql=205-135=70.
- For those data

$$1.5 \times \text{IQR} = 1.5 \times 70 = 105.$$

  Ql-105=30 (lower boundary, potential outliers below), and Q3+105=310 (upper boundary, potential outliers above).
- By the 1.5×IQR criterion, observations below 30 or above 310 are potential outliers.
- The only observations below 30 or above 310 are the sodium values of 0 and 340 mg. These are the only potential outliers.

5-Number Summary of Positions (box-plot)

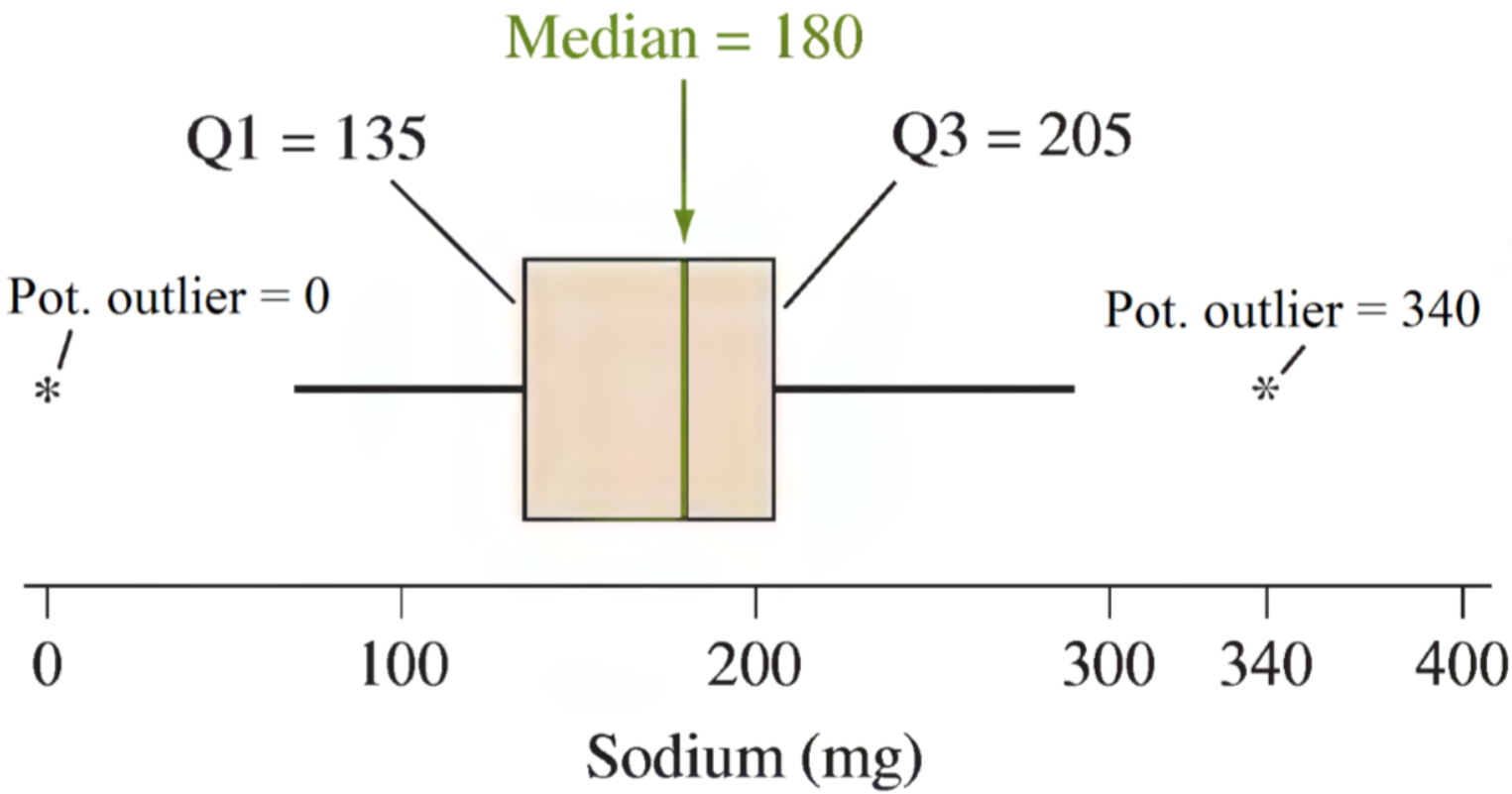
The 5-number summary is the basis of a graphical display called the box plot, and consists of

Minimum value, First Quartile, Median, Third Quartile, Maximum value

SUMMARY: Constructing a Box-plot

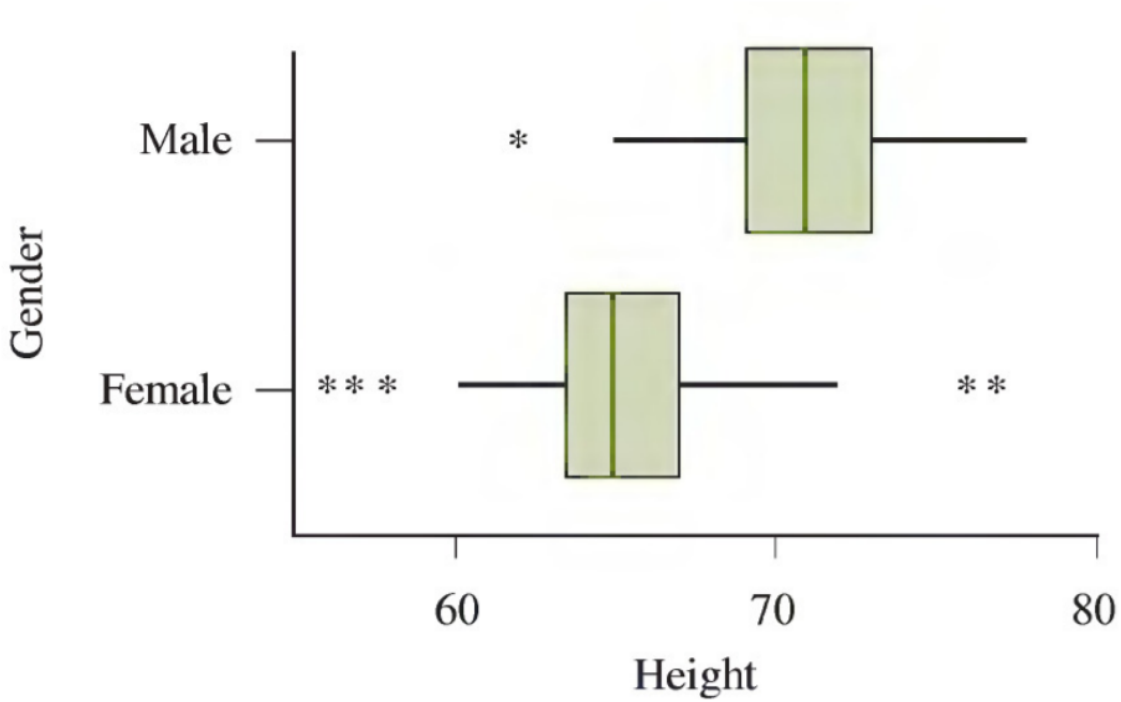
- A box goes from Q1 to Q3.
- A line is drawn inside the box at the median.
- A line goes from the lower end of the box to the smallest observation that is not a potential outlier and from the upper end of the box to the largest observation that is not a potential outlier.
- The potential outliers are shown separately (stars).

Example: Box-plot for Cereal Sodium Data



Comparing Distributions

- Quantitative data can be broken down by levels of a categorical variable.
- The side-by-side box-plot produces an excellent visual comparison for shape, outliers, variability, etc.
- The box-plot is less accurate than the histogram but very useful in comparing several distributions.



Z-score

- Z-value or Z-score or simply Z, represents **the number of standard deviations an observation is from the mean**. The z-score for a particular observation is calculated as

$$z_i = \frac{x_i - \bar{x}}{s}.$$

- The z-scores have **mean 0 and standard deviation 1**.
- **Positive z-score** indicates the observation is above the mean.
- **Negative z-score** indicates the observation is below the mean.
- The z-score also identifies position and potential outliers.
- An observation from a bell-shaped (or nearly symmetric) distribution is a **potential outlier if its z-score is beyond** ± 3, i.e. if its z-score < -3 or $> +3$ (**3 standard deviation criterion**).

Example

For a recent final exam the mean was 68.55 with a standard deviation of 15.45.

- Student A scored an 80%: $z_A = (80 - 68.55)/15.45 = \mathbf{0.74}$, which means the score of 80 was 0.74 standard deviation above the mean.
- Student B scored a 60%: $z_B = (60 - 68.55)/15.45 = \mathbf{-0.55}$, which means the score of 60 was 0.55 standard deviation below the mean.