1 - Visualizing data

Types of data [ES 1.2, PS 1.1]

Favorite movie genres
Action
Comedy
Comedy
Drama
Drama
Horror

Qualitative

Top five U.S. jobs with most growth (projected 2024)
1. Personal care aides
2. Registered nurses
3. Home health aides
4. Food service & prep. workers
5. Retail salespersons

Qualitative

New York Yankees' World Series victories (years)
1923, 1927, 1928, 1932, 1936,
1937, 1938, 1939, 1941, 1943,
1947, 1949, 1950, 1951, 1952,
1953, 1956, 1958, 1961, 1962,
1977, 1978, 1996, 1998, 1999,
2000, 2009

Quantitative

Qualitative data are attributes, labels, or nonnumerical entries.

Quantitative data are numbers that are measurements or counts.

Levels of measurement	Description	Examples
Nominal level (Qualitative data)	No math operation	Movie genres, hair color, ZIP code, ID number, phone number, religion, race, blood type
Ordinal level (Qualitative or quantitative data)	Ordered, ranked	Top 5 jobs; worst 10 cities in US; grouped ranges $0-$19,999, $20,000-$39,999, $40,000-$59,999
Interval level (Quantitative data)	Subtraction is meaningful	Yankee wins by year, temperature
Ratio level (Quantitative data)	Division is meaningful; has "true" 0	Age, height, salary Non-example: temperature: 0 F ≠ 0 C, 2 C is not twice as warm as 1 C.

Visualize qualitative data with pie charts or bar graphs: (we skip these)

Favorite movie genres

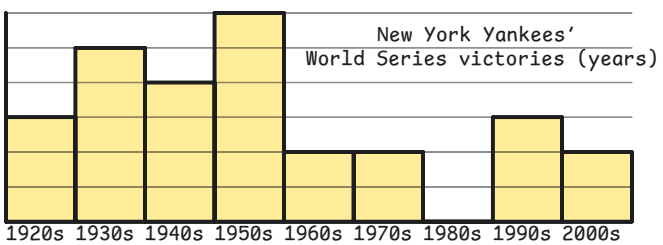
bar charts have gaps between bars; reordering categories reorders the bars

Visualize quantitative data with **frequency histogram**.

Guidelines:

- Use 5-20 classes or bins.
- Classes all have same width.
- Distribution of a histogram is described by shape, center, spread, outliers.

today



Central tendencies [ES 2.3, PS 1.3]

The **mean** of a dataset or data vector $x = (x_1, x_2, \ldots, x_n)$ is given by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

The **median** is the middle value of an ordered data set. If the data set has an even number of entries, the median is the mean of the two middle data entries.

Example 1. $\text{median}(1,1,3,3,7,9,11) =$ $\underline{\quad 3 \quad}$; $\text{median}(1,1,3,3,7,9,11,13) =$ $\underline{\quad 5 \quad}$

Example 2. $x = (1,2,3,6)$ has mean $\underline{(1+2+3+6)/4 = 3}$ and median $\underline{(2+3)/2 = 2.5}$

Example 3. $x = (1,2,3,14)$ has mean $\underline{(1+2+3+14)/4 = 5}$ and median $\underline{(2+3)/2 = 2.5}$

Moral: The median is less skewed by extreme values: the median is **robust**.

The **mode** of a data set is the data entry (or entries) that occur with the greatest frequency. A data set can have no mode, one mode, two modes ("bimodal"), etc.

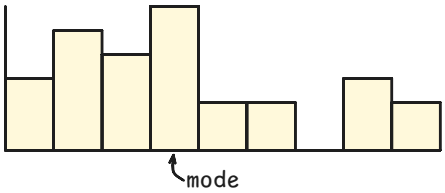
Example 4. Find the mode(s) of the data set and the distribution shown below.

(a)

$\{1,1,2,3,1,2,3,4,1,2,3,5,2,5,2\}$

2 occurs the most time (5 times) so it's the mode.

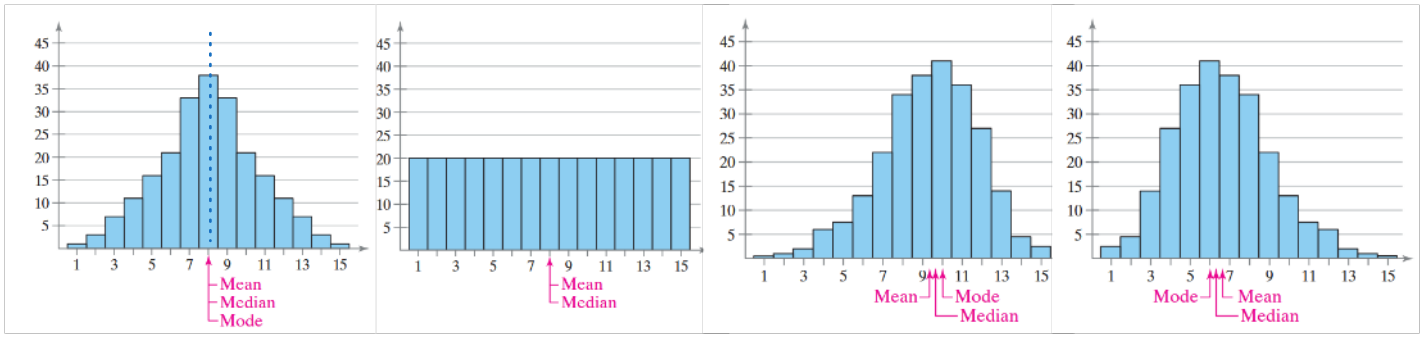
(b)



Shape of distributions [ES 2.3, PS 1.2]

Definitions. A frequency distribution is:

- **symmetric** if the distribution looks like roughly mirror images about some vertical line;
- **uniform** if all classes or bins have roughly the same frequencies.
- **skewed left** if the tail of the distribution extends to the left.
- **skewed right** if the tail of the distribution extends to the right.



Symmetric Distribution | Uniform Distribution | Skewed Left Distribution | Skewed Right Distribution

How to detect skewness: The tail of extreme values drags the mean more than the median, so:

- mean is much smaller than median suggests the distribution is skewed left,
- mean is much larger than median suggests the distribution is skewed right.