Statistics of data

Topics: Mean, Median, Mode, Variance, Std. Deviation, Normal distribution

Measuring the Central Tendency: (1) Mean

- <u>Mean (algebraic measure) (sample vs. population):</u>

  Note: *n* is sample size and *N* is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Trimmed mean:
  - Chopping extreme values (e.g., Olympics gymnastics score computation)

Measuring the Central Tendency: (2) Median

- <u>Median</u>:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Approximate median

Sum before the median interval

Interval width ($L_2 - L_1$)

$$median = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}}\right) width$$

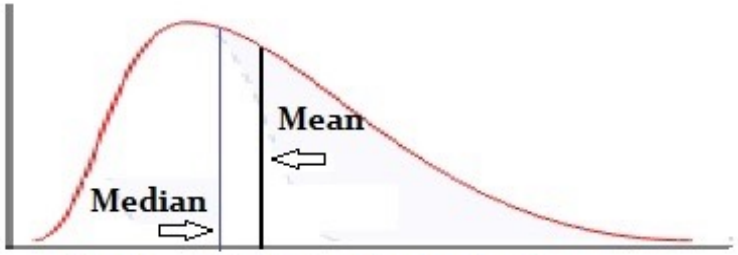
Low interval limit

Measuring the Central Tendency: (3) Mode

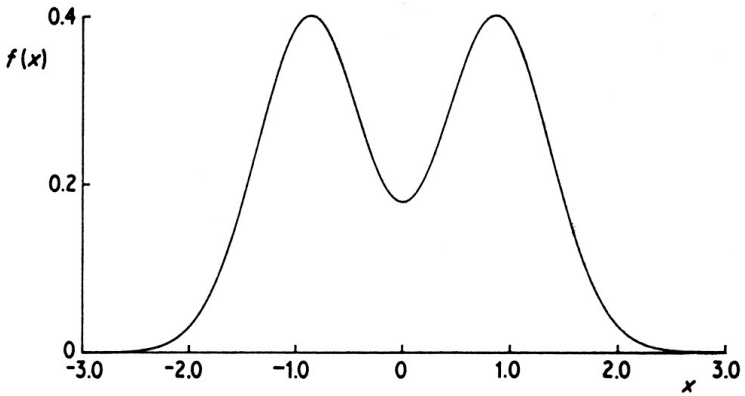
- Mode: Value that occurs most frequently in the data

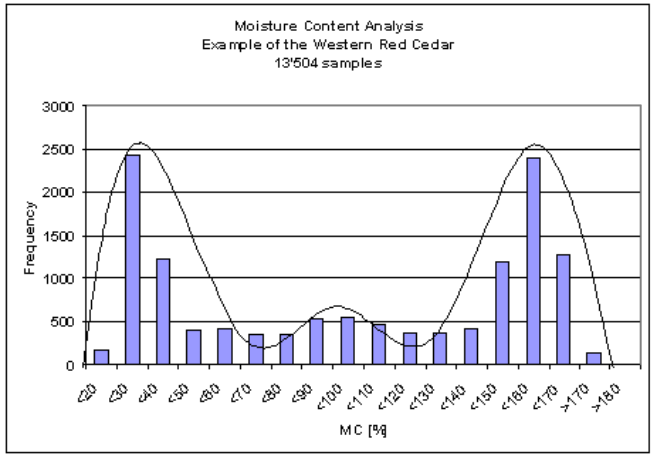
- Unimodal

- Multi-modal
  - Bimodal
  - Trimodal



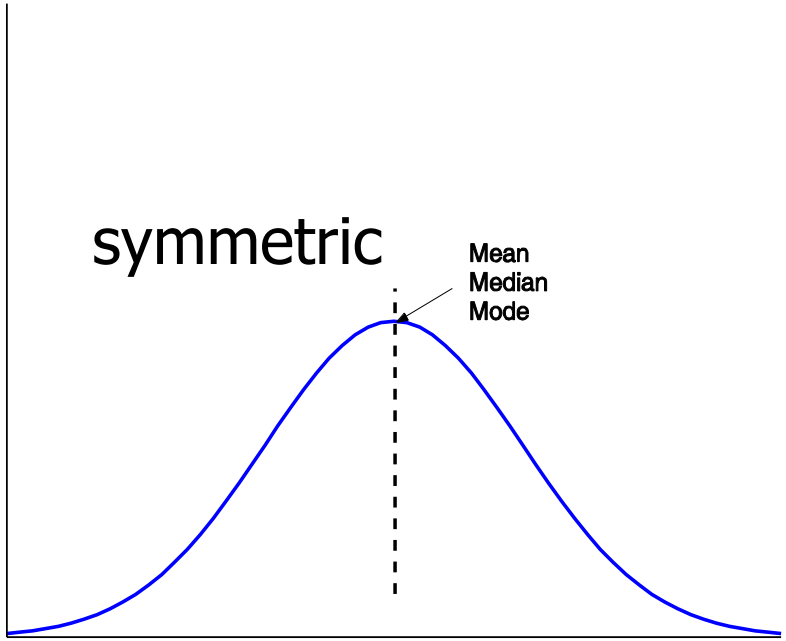
Right skewed distribution: Mean is to the right

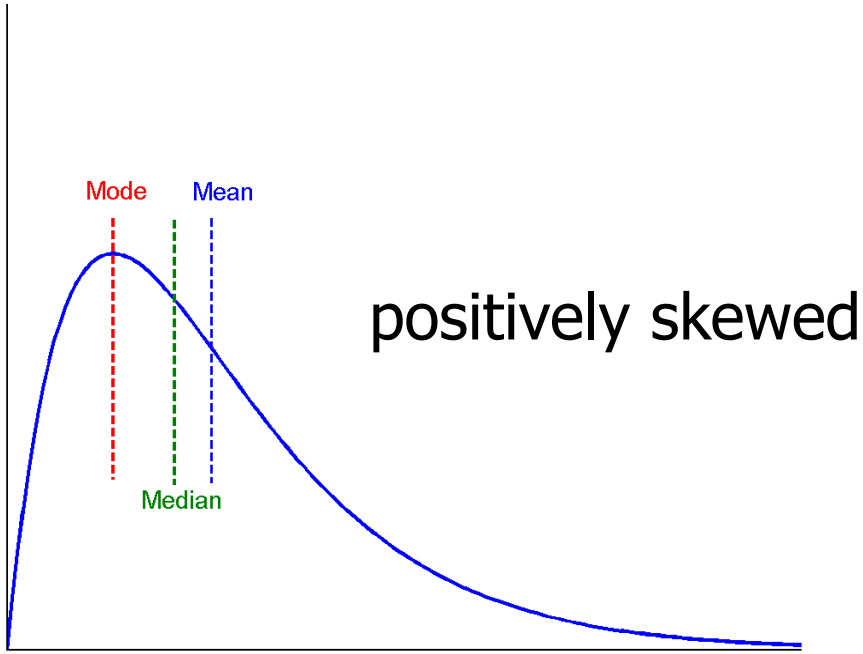


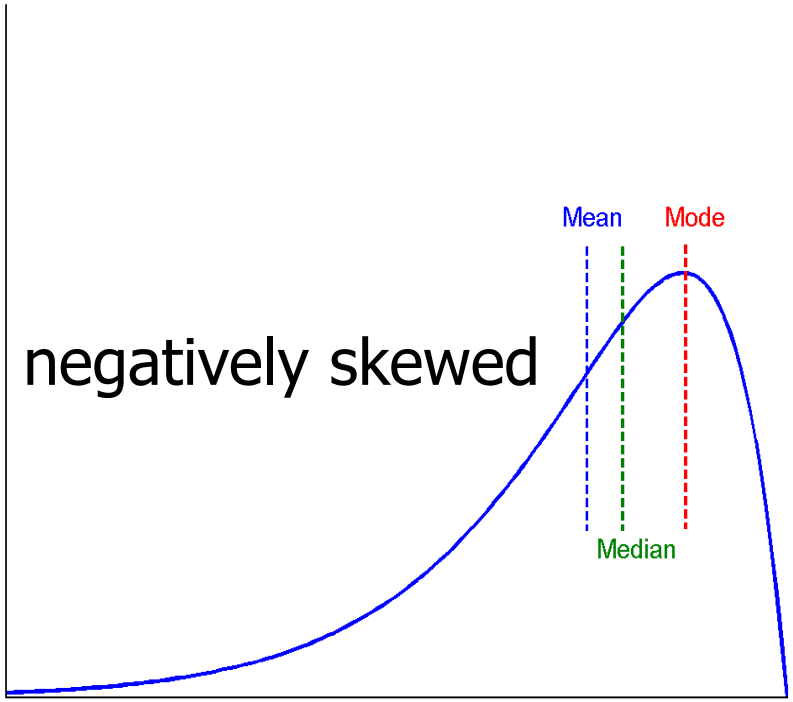


Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measures Data Distribution: Variance and Standard Deviation

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

Note: The subtle difference of formulae for sample vs. population
- n : the size of the sample
- N : the size of the population

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Properties of Normal Distribution Curve

Normal distribution is

- a distribution of data that has roughly the same amount of data on either side of the middle and
- has its most common values around the middle of the data