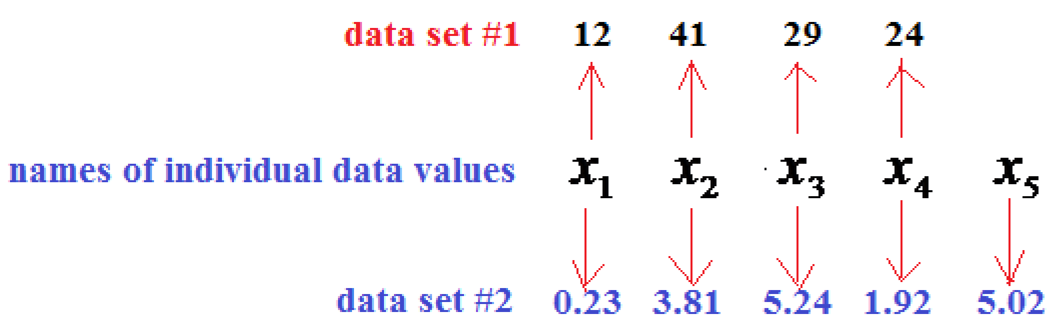
4.1 Notes

NUMERICAL SUMMARY OF DATA

Notations

- Since we will frequently use the sum of values in a data set, we want to use an "omnibus" expression that means the summation of all data values. For this purpose, we give a generic name to any given data set, say *x*, and attach a positive integer, say i, which represents the physical location of the i-th data value as subscript to the name *x*. In other words, x_i = the i-th data value in the data set called *x*.

  For example, we have a data sets with values:



  The sum of all data values in an arbitrarily given data set is given by:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + \cdots + x_n$$

The left hand side of the above formula is the compact notation of sum of all data values while the right hand side of the above formula *explicitly* denotes the sum of each individual value.

For the first data set, the sum is given by

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 12 + 41 + 29 + 44 = 126$$

Similarly, the sum of the second data set is

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$= 0.23 + 3.81 + 5.24 + 1.92 + 5.02 = 16.22$$

- **Population parameters vs. sample statistics**

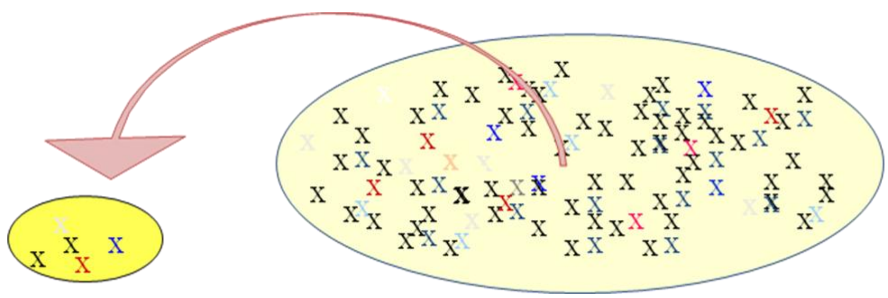
  Population parameter: A number that describes a population characteristic.

Sample statistics: A number that describes a sample characteristic.

Example: μ = population mean → population parameter
$\bar{x}$ = sample mean → sample statistic

- **Simple Random Sample (definition)**

Every possible sample of the same size has the same chance of being selected.

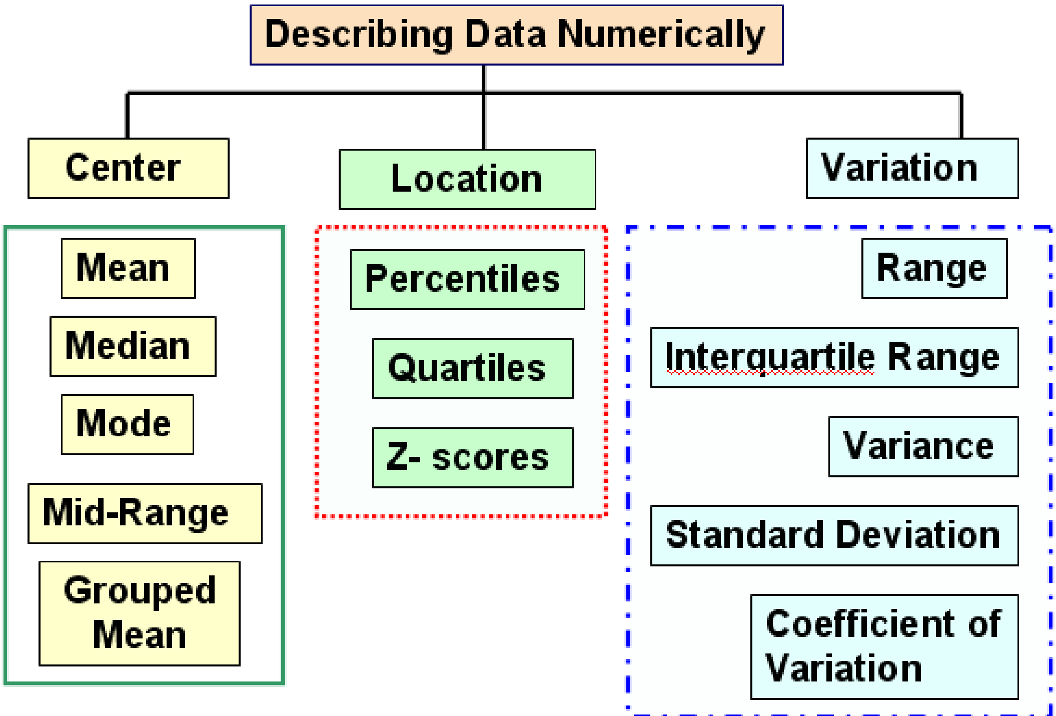


a). Random numbers can be generated by a random number table, a software program or a calculator.

b). Assign a number to each member of the population.

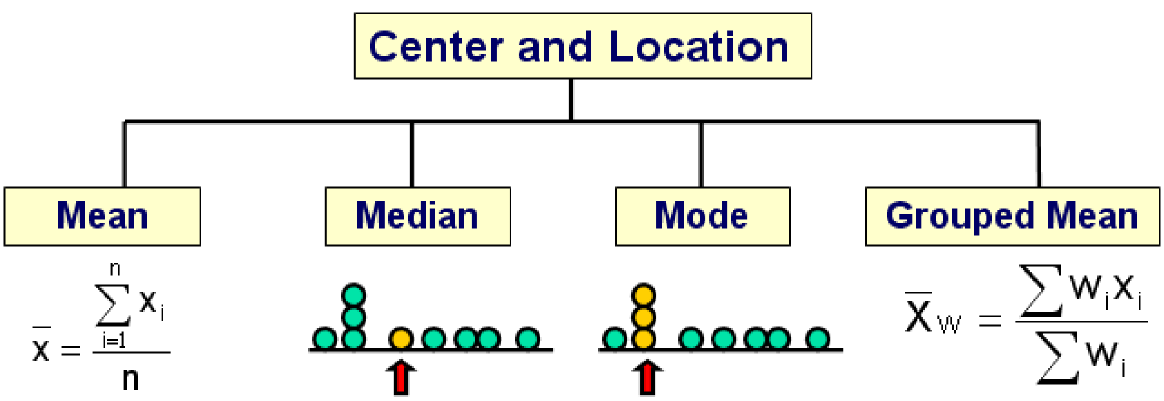
c). Members of the population that correspond to these numbers become members of the sample.

Preview of Numerical Measures



Measures of Center

Overview



Mean:

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

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Grouped Mean:

$$\bar{X}_W = \frac{\sum w_i x_i}{\sum w_i}$$

$$\mu_W = \frac{\sum w_i x_i}{\sum w_i}$$

- **The (Arithmetic) Mean**

The Mean is the arithmetic average of data values

Sample mean (n = Sample Size)

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

Population mean (N = Population Size)

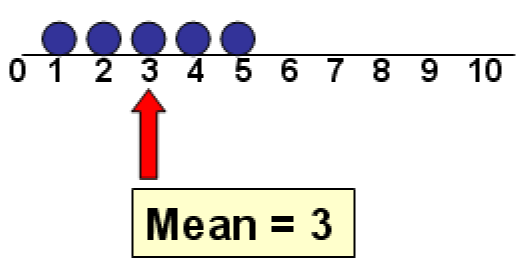
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Some comments about the mean:

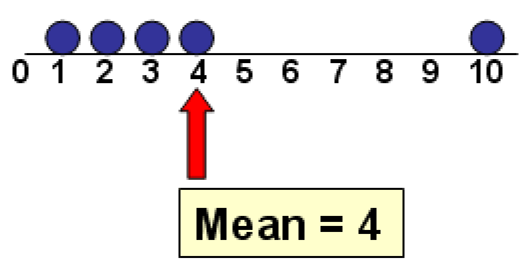
The most common measure of central tendency

Mean = sum of values divided by the number of values

Affected by extreme values (outliers)



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

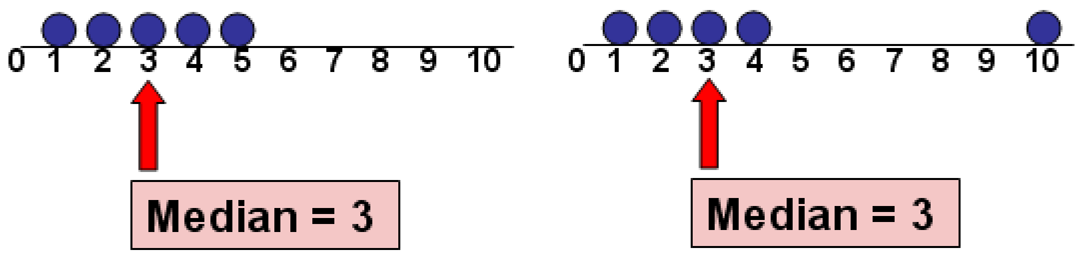


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

- **The Median**

In an ordered array, the median is the "middle" number

- If n or N is odd, the median is the middle number
- If n or N is even, the median is the average of the two middle numbers



Example: {2, 6, 7} → median = 6
Example: {1, 2, 6, 7} → median = (2 + 6) / 2 = 4.

Comments:

1. A median may not be a data value.
2. Question: Which measure of center is "better", the mean or the median?

   Answer: It depends on the situation and your purposes. The mean is sensitive to the extreme values. The median is robust to the extreme values.

   Example: Mean vs. median income for a highly skewed distribution.

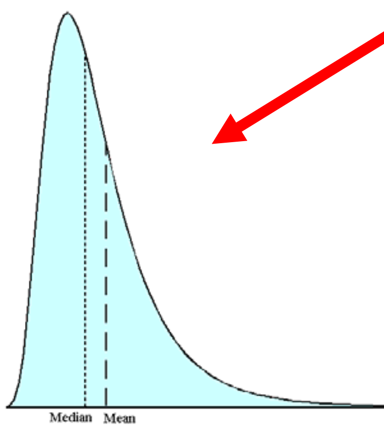
3. For a **symmetric** data distribution: mean = median.

Comparing the Sample Mean & Sample Median

Typically:

1. When a distribution is symmetric, the mean and the median are equal.

2. When a distribution is **skewed positively**, the mean is larger than the median. See the example below



Skewed positively

3. When a distribution is **skewed negatively**, the mean is smaller then the median.
   - Make a Sketch for this case like the one provided above:

- **The Mode**

A measure of central tendency

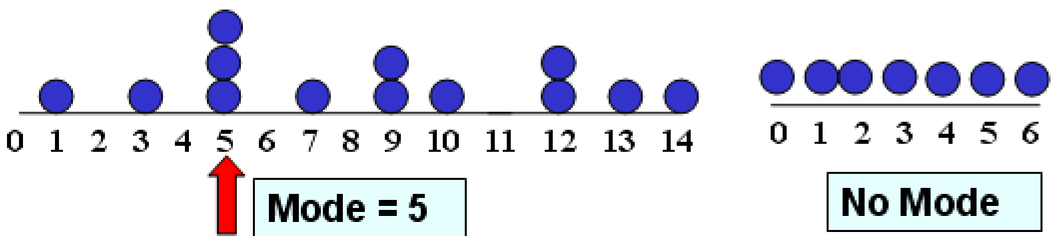
Value that occurs most often

Not affected by extreme values

Used for either numerical or categorical data

There may be no mode

There may be several modes



Comments

1. A data set can have more than one mode. A data set with one mode is unimodal data.
2. Use of the terms **bimodal** and **multimodal**.
3. Note the mode is not necessarily a "measure of center".

4) For categorical data, MODE is the only suitable measure of center