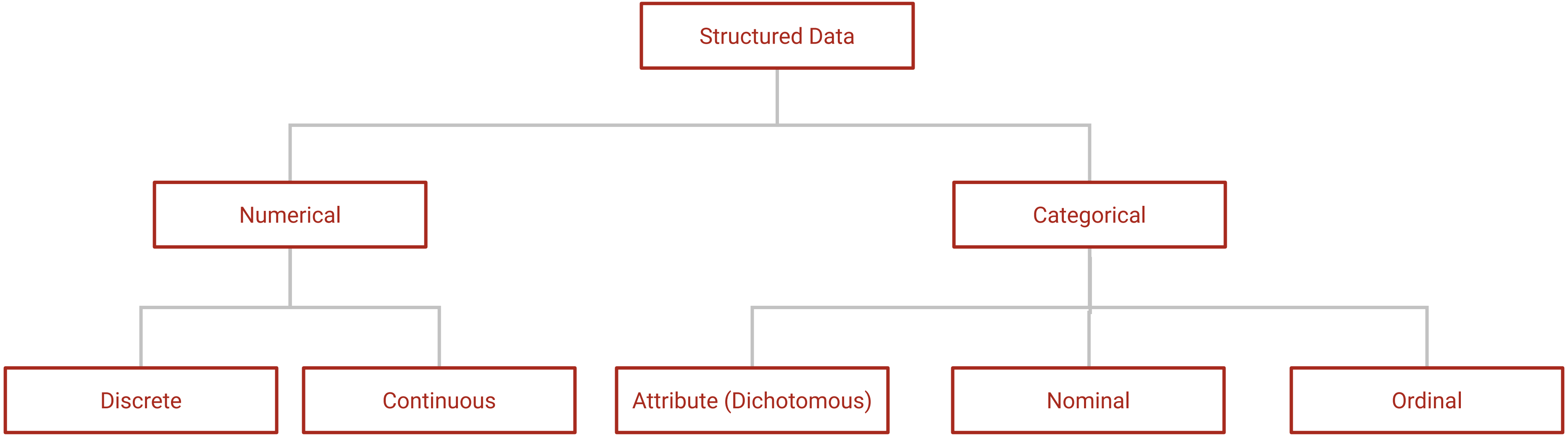
Summarizing Data

We will talk about...

- Types of Data
- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Spread
  - Range
  - Variance
  - Standard Deviation
  - Interquartile Range
- Measures of Symmetry
  - Coefficient of Skewness

Types of Data



Structured Data

Numerical

Categorical

Discrete

Continuous

Attribute (Dichotomous)

Nominal

Ordinal

Identify the type of data

1. How old are you? (Give your age as on your last birthday.)
2. How tall are you? (State as accurately as you can.)
3. Which gender do you identify with?
4. What colour are your eyes?
5. Do you smoke?
6. How would you rate your looks? (10 =Drop-dead gorgeous, 1= Seen better days)

Measures of

Central Tendency

Mean

By far the **most common measure** for describing the location of a set of data is the mean.

For a set of observations denoted by x_1, x_2, x_3, ,x_4, . . . x_n

or x_i where i = 1 , 2 , 3 ... n,

the mean is defined by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (read as “x bar”).

Mean (Continued)

For a frequency distribution with possible values x_1, x_2, x_3 … x_k with corresponding frequencies f_1 , f_2 ,f_3 ….. f_k ,

where $\Sigma f_i = n$, the mean is given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i \, x_i$.

Median

1) Consider placing the ‘n’ observations in order of magnitude. Sort the data in ascending order.

2) The median is a value, which splits the data set into two equal halves, so that **half the observations are less than the median and half are greater than the median**.

3) If n is odd,then the median is the middle observation. If n is even, then the median is the midpoint of the middle two observations.

One of the potential advantages of the median for certain data sets is that it is robust or resistant to the effects of extreme observations (outliers).

Mode

Mode is defined as the value which occurs with the **greatest frequency** or the **most typical value**.

- Its use in practice is limited
- Example usage: Which company is performing better in terms of number of users?

Properties of Mean, Median and Mode

Mean

1) Influenced by outliers
2) Not applicable to nominal data.
3) Only one mean

Median

1) Not influenced by outliers
2) Not applicable to nominal data.
3) Only one median

Mode

1) Not influenced by outliers
2) Applicable to all types of data
3) More than one mode

Measures of
Spread

Variance/Standard Deviation

For a set of observations denoted by $x_1, x_2, x_3, ,x_4, . . . x_n$ or x_i

where i = 1 , 2 , 3 ... n, the variance is defined by $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

The symbol s^2 is used to denote sample variance. The variance can be calculated more easily using the alternative formula: $s^2 = \frac{1}{n-1}\left[\sum x_i^2 - n\bar{x}^2\right]$ which is equivalent to $s^2 = \frac{1}{n-1}\left[\sum f_i \, x_i^2 - n\bar{x}^2\right]$

The **standard deviation** is the positive square root of the variance.

Range

The range is a very **simple measure of spread** defined as the difference between the largest and smallest observations in the data set.

Range = $\max(x_i) - \min(x_i)$

The range is a **poor measure** of the spread of the data as it relies on the extreme values, which aren’t necessarily representative of the data as a whole.

Interquartile Range

The interquartile range (IQR) is another measure of spread which is like the range but which is **not affected by the data extremes**.

The quartiles divide a set of data into four quarters. They are denoted by Q1, Q2 and Q3.

Note that Q2 is just the **median**, while Q1 is called the **lower quartile** and Q3 the **upper quartile**.

The lower quartile, the median and the upper quartile are also referred to as the 25th, 50th and 75th percentiles.

The interquartile range is defined as (Q3 - Q1).

Measures of

Symmetry

Symmetry and Skewness

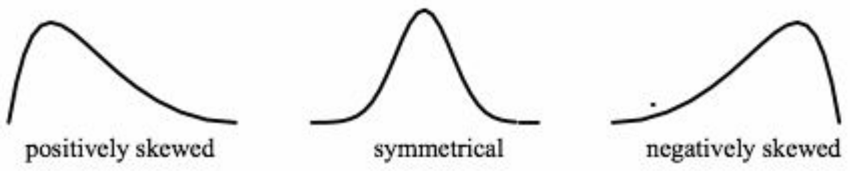
The next feature of interest is the shape of the distribution of a data set, that is, whether it is symmetric or skewed to one side or the other.

Here are the distributions for three data sets each of 200 observations:

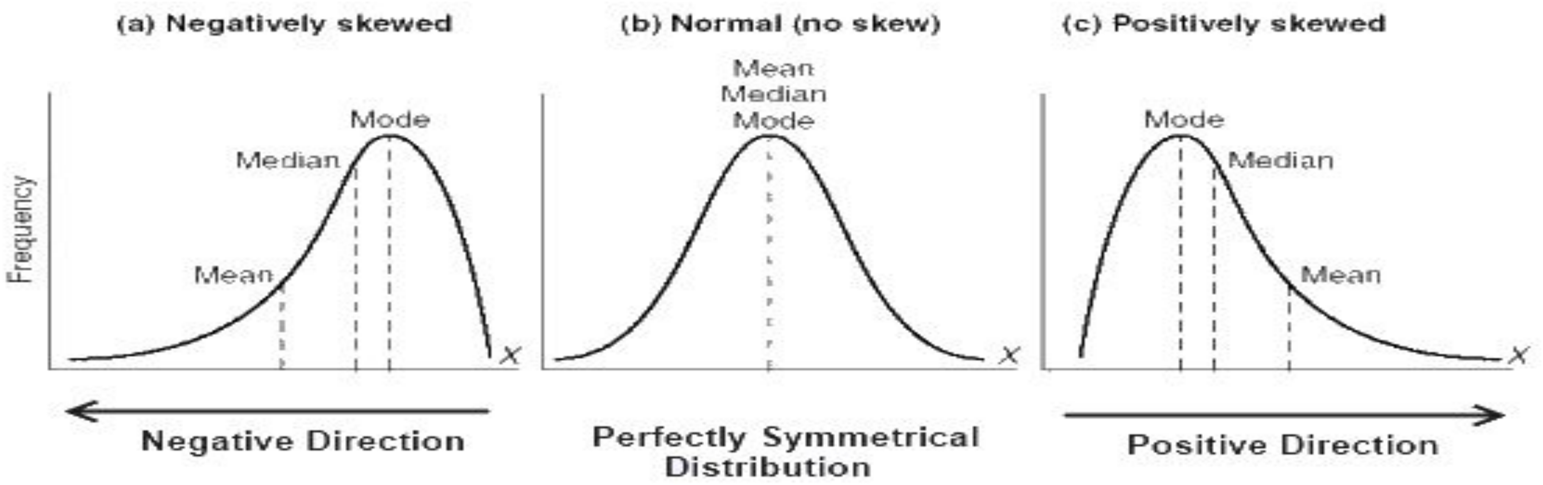
– the first is positively skewed

– the second is fairly symmetrical

– the third is negatively skewed.



Mean, median and mode for skewed data



Mean < Median < Mode

Mean = Median = Mode

Mean > Median > Mode

Coefficient of Skewness

$$\text{coeff of skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1.5}}$$

Outlier Detection

Rule of thumb:

Lower side outliers: Observations below (Q1 - 1.5 * IQR)

Upper side outliers: Observations above (Q3 + 1.5 * IQR)