Data Characteristics

- Numerical Descriptive Measures
  - Location
    - Mean
    - Median
    - Mode
  - Dispersion
    - Range
    - Variation
  - Shape
    - Skewness
    - Kurtosis

Measure of Location

- Mean:
  - The arithmetic mean of a sample (or simple the sample mean) of n observations $x_1, x_2, \ldots, x_n$, denoted by $\overline{X}$ is computed as

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} x_i$$

143	141	140	139	142	141	142	143	139	142
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

$$\text{mean} = \overline{X} = \frac{\sum X_i}{n}$$

$$= \frac{X_1 + X_2 + X_3 + \ldots\ldots X_{10}}{10}$$

$$= \frac{143 + 141 + 140 + \ldots\ldots 142}{10}$$

$$= 141.2 \text{ mm}$$

Measure of Location (contd.)

- Median
  - The median is defined to be the value which divides the data into two equal parts, i.e. the 'middle' value. Half the values are above it and half are below it.
  - To obtain the median
    - Order the data in ascending order.
    - For n items in the data set, the median is the (n + 1)/2 observation

139	139	140	141	141	142	142	142	143	143
X_4	X_9	X_3	X_2	X_6	X_5	X_7	X_{10}	X_1	X_8

$\left(\frac{10+1}{2}\right)$ th observation = (11/2) = 5.5th observation.

So 5.5th observation = mean of 5th and 6th observations.
i.e. Median =141.5

Measure of Location (contd.)

- **Mode**
  - The mode of a set of n measurements X_1, X_2, ...X_n is the value of X that occurs with the greatest frequency. (i.e. the most popular or common value)

139	139	140	141	141	142	142	142	143	143
X_4	X_9	X_3	X_2	X_6	X_5	X_7	X_{10}	X_1	X_8

The mode is 142

Measures of Dispersion

- **Range**
  - The simplest measure of the spread or variation of a set of data is the range, R, which is defined as the difference between the largest and the smallest value.
  - Although the range is a simple-to-use measure of variation, it only uses the two extremes of the data set (i.e. the maximum and minimum values). Thus, it is significantly affected by extreme (outlying) values within the data.

143 141 140 139 142 141 142 143 139 142

Then Range = R = Maximum – Minimum
= 143 - 139
= 4 mm

Measures of Dispersion (contd.)

- **Variation/Standard Deviation**
  - It is based on the idea of measuring how far, on average, the observations are from the centre of the data. For technical reasons, to average the squared deviations we divide by n – 1 rather. The variance is measured in the square of the units of the original data

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

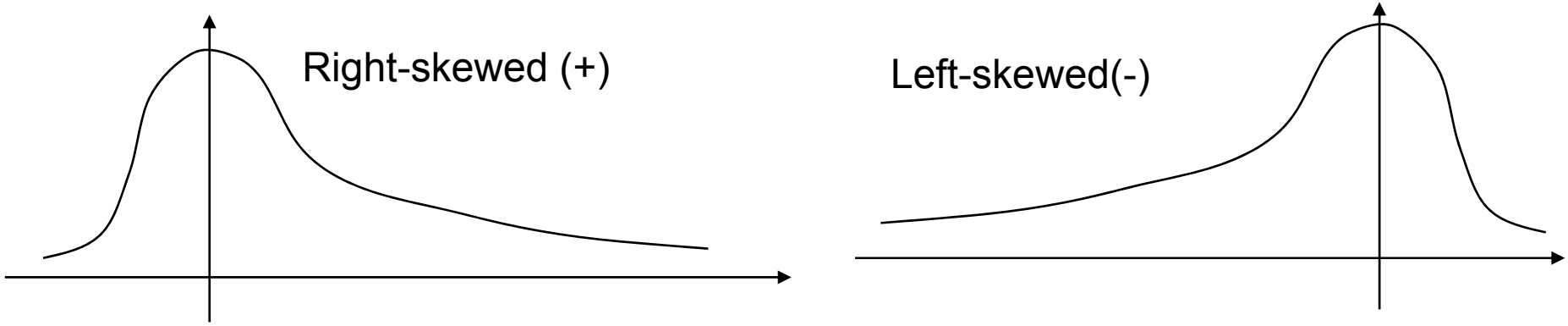
  - All the other summary statistics we have considered are in the same units as the data. Thus a better measure for practical use is the square root of the variance called the **standard deviation** defined by:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

Measure of Shape

- Skewness
  - If the distribution of the data is not symmetrical it is called asymmetrical or skewed
  - Skewness characterizes the degree of asymmetry of a distribution around its mean

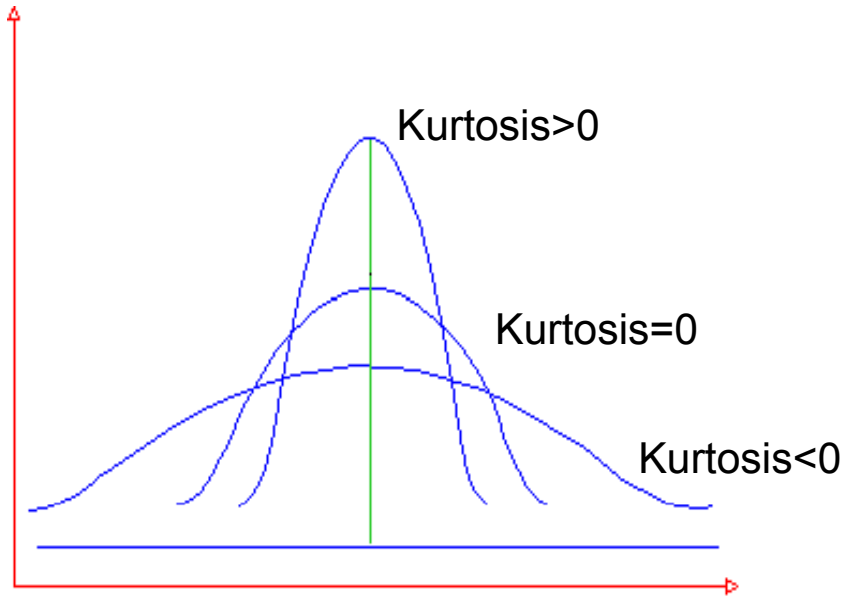
$$skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$



Measure of Shape (contd.)

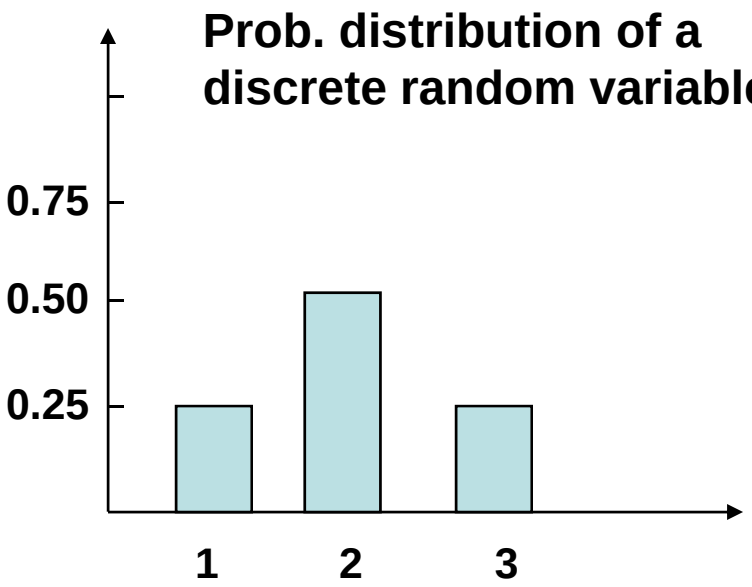
- Kurtosis
  - Kurtosis characterizes the relative peakedness or flatness of a distribution compared with the bell-shaped distribution (normal distribution)

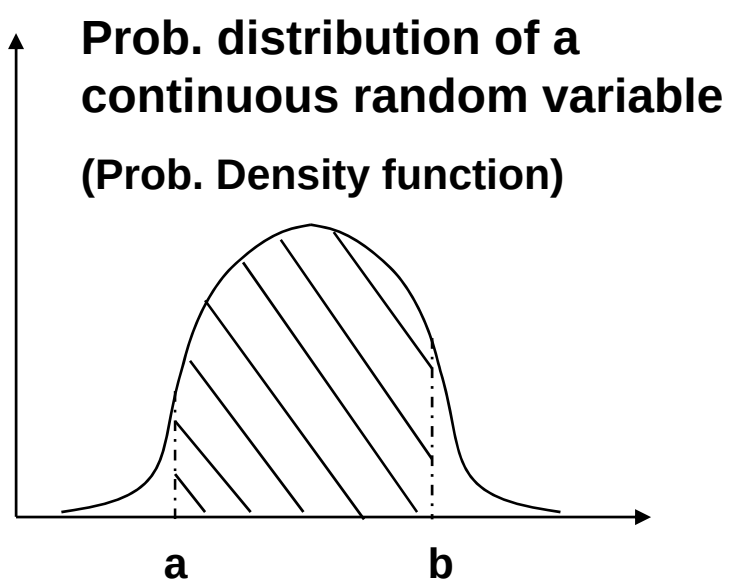
$$kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)}{(n-2)(n-3)}$$



Random Variables

- A variable whose values are random but whose statistical distribution is known
  - Discrete random variable is one that can assume only a countable number of values
  - Continuous random variables can assume any value in one or more intervals on a line





Normal Probability Distribution

- The probability density function
  - u is the mean
  - σ^2 is the variance

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-u)^2/2\sigma^2}$$