Mahidol University
Faculty of Information
and Communication Technology

Curve Fitting

Rawesak Tanawongsuwan, Ph.D.
rawesak.tan@mahidol.ac.th

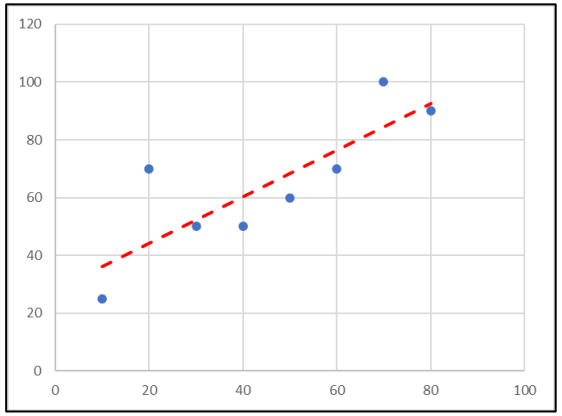
This slide is part of teaching materials for ITCS122 Numerical Methods
Semester 2/2023, Calendar year 2024

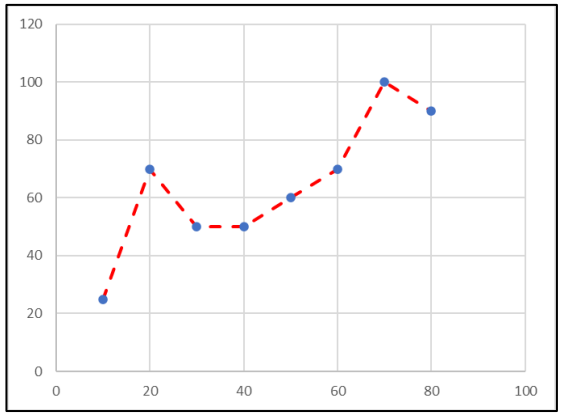
Curve Fitting

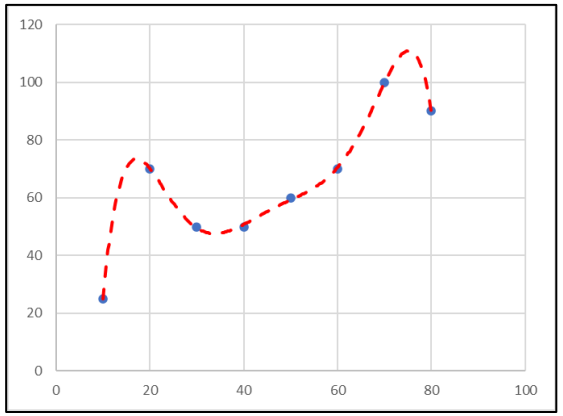
- Data are given for discrete values along a continuum
  - Estimation of points between the discrete values
    - Fit curves to data to obtain intermediate estimates
  - Simplification of a complicated function with a simpler function
    - Compute values of the complicated function at a number of discrete values along the range of interest, then a simpler function is computed to fit those discrete values

Two general approaches to curve fitting

1. Least-squares regression
   - Data contain error or scatter data → find a single curve to represent the general trend or pattern of the data
2. Interpolation
   - Data may be precise → find a curve that pass directly through each of the data points



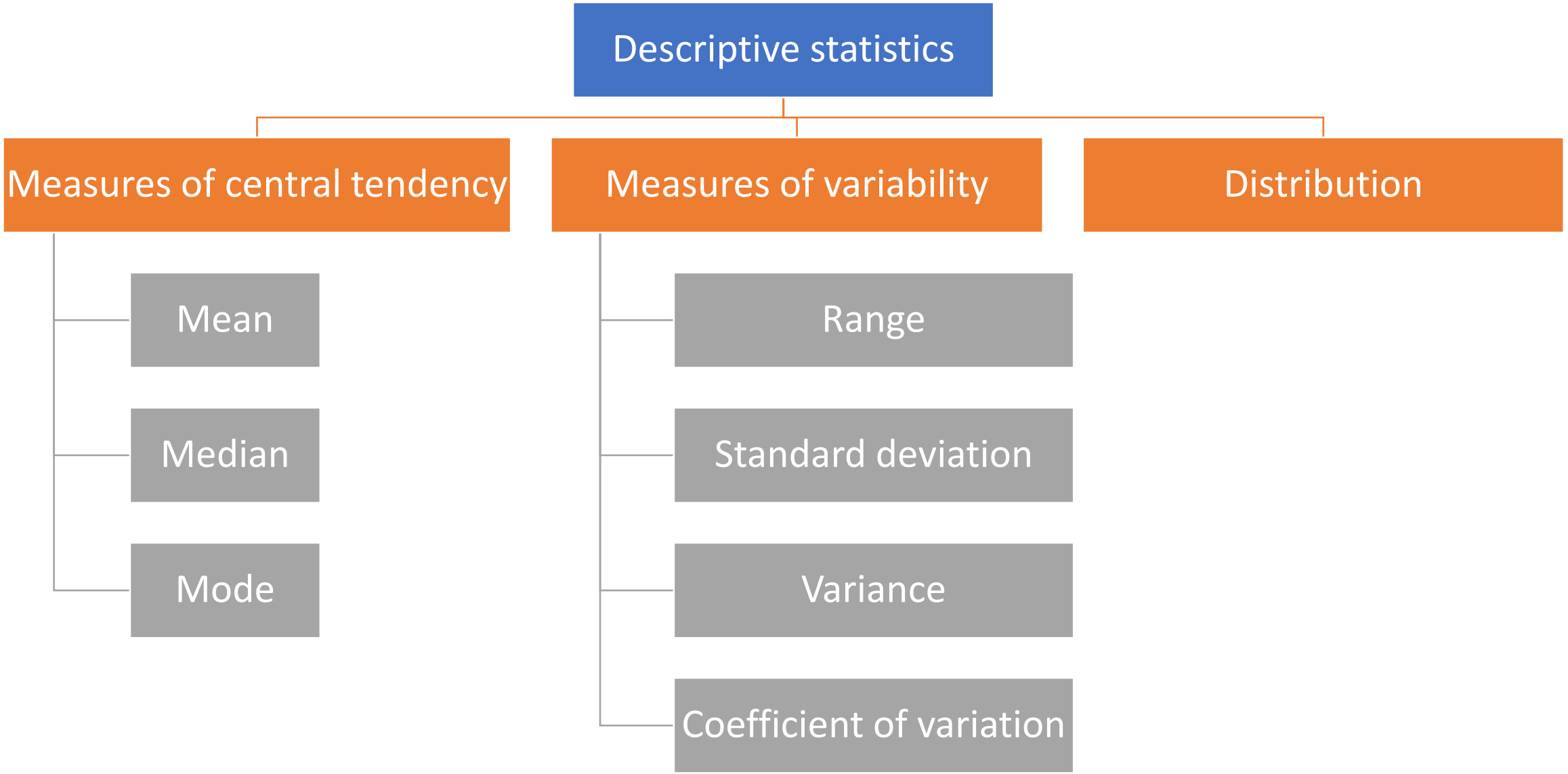




Descriptive statistics

- Descriptive statistics are summary statistics that quantitatively describe or summarize features from a collection of information

- There are 3 main types of descriptive statistics:
  1. The **central tendency** concerns the averages of the values.
  2. The **variability or dispersion** concerns how spread out the values are.
  3. The **distribution** concerns the frequency of each value.


Descriptive statistics
Measures of central tendency
Measures of variability
Distribution
Mean
Median
Mode
Range
Standard deviation
Variance
Coefficient of variation

Measures of central tendency

- Measures of central tendency estimate the center, or average, of a data set.

- Common measures are
  - Arithmetic mean
  - Median
  - Mode

Measures of central tendency

Arithmetic mean $(\mu, \bar{y})$ → The sum of the individual data points (y_i) divided by the number of points (n)

$$\bar{y} = \frac{\sum y_i}{n}$$

Median → The midpoint of a group of data (50th percentile).

- It is calculated by first putting the data in ascending order.
  - If n is odd, the median is the middle value.
  - If n is even, the median is the arithmetic mean of the two middle values.

Mode → The value that occurs most frequently.

Measures of variability (spread)

Range → the difference between the largest and the smallest values

Standard deviation (s_y) → the average amount of variability in data. It tells you, on average, how far each datapoint lies from the mean. The larger the standard deviation, the more variable the data set is.

$$s_y = \sqrt{\frac{S_t}{n-1}}$$

where S_t is the total sum of the squares of the residuals between the data points and the mean

$$S_t = \sum (y_i - \bar{y})^2$$

s_y or S_t is large → the data points are spread out widely around the mean

s_y or S_t is small → the data points are grouped tightly

Measures of variability (spread)

Variance → the square of the standard deviation

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^2 - (\sum y_i)^2/n}{n-1}$$

- The quantity $n-1$ is referred to as the degrees of freedom.
- S_t and s_y are based on $n-1$ degrees of freedom.
- Let's consider $S_t = \sum(y_i - \bar{y})^2$.
  - If $S_t = 0$ or $(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 = 0$, then knowing $\bar{y}$ and $n-1$ of the y_i values, then the remaining value of y will be known.
  - Thus, only $n-1$ of the values are said to be freely determined.
- Another justification for dividing by $n-1$ is the fact that there is no such thing as the spread of a single data point. For the case where $n = 1$, S_t and s_y yield a meaningless result of infinity.
- Note that the second formula does not require precomputation of $\bar{y}$

Measures of variability (spread)

Coefficient of variation (c.v.) → the ratio of the standard deviation to the mean

- It provides a normalized measure of the spread.
- It is often multiplied by 100 so that it can be expressed in the form of a percentage:

$$\text{c. v.} = \frac{s_y}{\bar{y}} \times 100\%$$

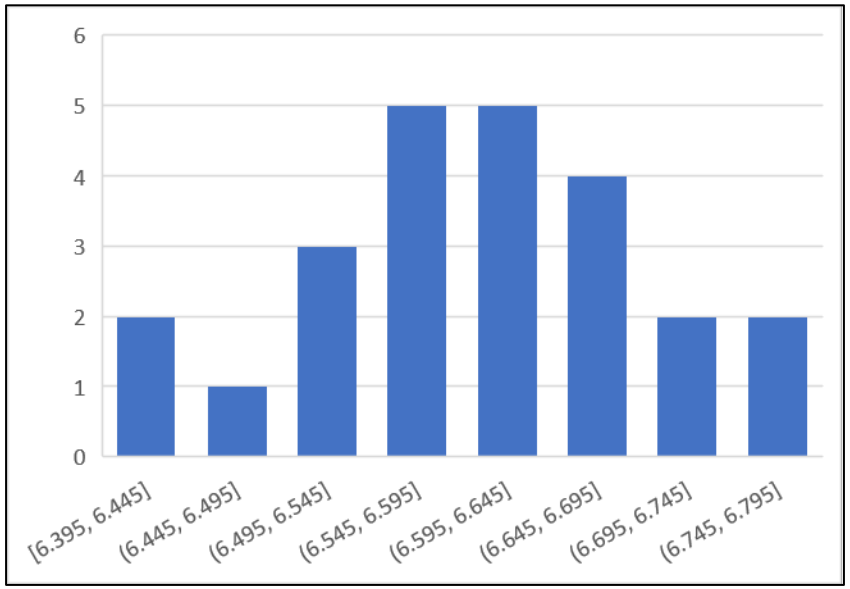
Simple Statistics Example

- Mean : $\bar{y} =$
- Standard variation : $s_y =$
- Variance : $s_y^2 =$
- Coefficient of variation : $\mathrm{c.v.} =$

i	y_i	$(y_i - \bar{y})^2$	y_i^2
1	6.395	0.04203	40.896
2	6.435	0.02723	41.409
3	6.485	0.01323	42.055
4	6.495	0.01103	42.185
5	6.505	0.00903	42.315
6	6.515	0.00723	42.445
7	6.555	0.00203	42.968
8	6.555	0.00203	42.968
9	6.565	0.00123	43.099
10	6.575	0.00063	43.231
11	6.595	0.00003	43.494
12	6.605	0.00002	43.626
13	6.615	0.00022	43.758
14	6.625	0.00062	43.891
15	6.625	0.00062	43.891
16	6.635	0.00122	44.023
17	6.655	0.00302	44.289
18	6.655	0.00302	44.289
19	6.665	0.00422	44.422
20	6.685	0.00722	44.689
21	6.715	0.01322	45.091
22	6.715	0.01322	45.091
23	6.755	0.02402	45.630
24	6.775	0.03062	45.901
sum	158.400	0.21700	1045.657

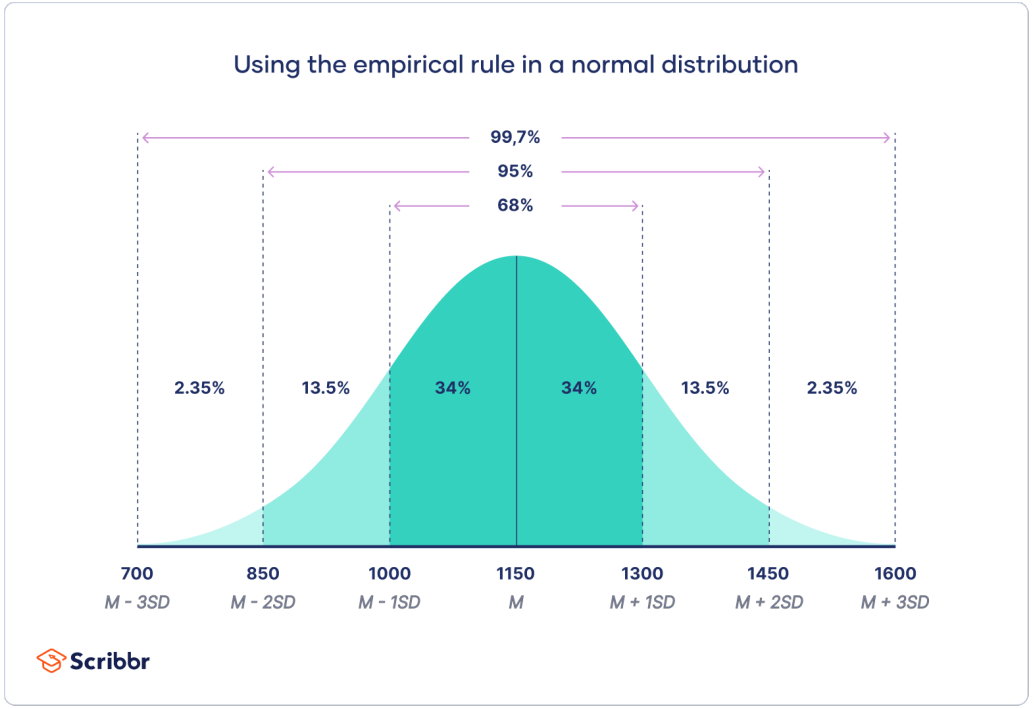
Distribution

- Data distribution – the shape with which the data are spread around the mean.
- A histogram is constructed by sorting the measurements into intervals or bins



Normal distribution

- Data is symmetrically distributed with no skew.
- When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.
- Normal distributions are also called Gaussian distributions or bell curves.
- The empirical rule, or the 68-95-99.7 rule



https://www.scribbr.com/statistics/normal-distribution/