SECI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

CHAPTER 3

Descriptive Statistics

Measures of Central Tendency

Measurement of Central Tendency

- A measure of central tendency of a distribution is a numerical value that describes the central position of the data or how the data tend to build up in the center.

- Measurements:
  - Mean
  - Mode
  - Median

Mean

- Mean is the sum of the observations divided by the number of observations.
- It is the most common measure of central tendency,

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

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

Example

- Data: 13, 18, 13, 14, 13, 16, 14, 21, 13

Calculation:
$(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) \div 9 = 15$

- The mean is 15

- The mean is unique for every set of data.
- Meaningful for interval and ratio data.
- Can be affected by outliers – rare observations that are radically different from the rest.

- Example: 3, 4, 6, 4, 7, 3, 6, 5, **1500**
- Mean: **170.89**

Mean of Grouped Data

- The formula for the mean of grouped data,

$$\bar{x} = \frac{\sum_{i=1}^{h} f_i x_i}{n} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_h x_h}{f_1 + f_2 + \dots + f_h}$$

where,

f_i: frequency in a class or frequency of an observed value.

x_i: class midpoint or an observed value.

n : number of classes or number of observed values.

Example

Find the mean value for the following data:

Number of children (x_i)	1	2	3	4	5	6	7
Frequency (f_i)	5	12	8	3	0	0	1

Solution:

$\sum f_i x_i$ = 5(1) + 12(2) + 8(3) + 3(4) + 0(5)+ 0(6) + 1(7) = 72

$$\bar{x} = \frac{\sum_{i=1}^{h} f_i x_i}{n} = \frac{72}{29} = 2.5$$

Example

Find the mean value for the following data:

Class interval	Frequency
41 - 50	7
51 - 60	10
61 - 70	15
71 - 80	2
81 - 90	6
Total	50

Example

Class interval	Midpoint, x_i	Frequency, f_i	$f_i x_i$
41 - 50	(41 + 50) ÷ 2 = 45.5	7	(45.5 x 7) = 318.5
51 - 60	55.5	10	555
61 - 70	65.5	15	982.5
71 - 80	75.5	2	151
81 - 90	85.5	6	513
Total		50 (n)	2520

$$\bar{x} = \frac{\sum_{i=1}^{h} f_i x_i}{n} = \frac{2520}{50} = 50.4$$

Median

- The median is the **middle** value when the data are arranged from smallest to largest.
- To find the median, your numbers have to be listed in an order, so you may have to rewrite your list first.
- For an odd number of observations, the formula for the place to find the median is

  ([the number of data points] + 1) ÷ 2

Example

- Data: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Arrange in order: 13, 13, 13, 13, **14**, 14, 16, 18, 21
- There are nine numbers in the list, so the middle one will be the $(9 + 1) \div 2 = 10 \div 2 = 5^{th}$ number.
- So the median is 14.

- For an even number of observations, the median is the mean of the two middle numbers.

- Example:

2, 3, 3, **5, 6**, 7, 8, 9

$(5 + 6) \div 2 =$ **5.5 (median)**

- The median is meaningful for ratio, interval, and ordinal data.
- The median is not affected by outliers.

- Example: 3, 4, 6, 4, 7, 3, 6, 5, **1500**

3, 3, 4, 4, **5**, 6, 6, 7, 1500

- Median: **5**

Median of Grouped Data

- The median for the grouped data is given by

$$\boxed{\text{median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)}$$

where,

(Median class is the first class with the value of cumulative frequency equal at least $N/2$)

$\boldsymbol{L}$: lower limit of median class,

$\boldsymbol{N}$: total number of observations,

$\boldsymbol{cf_p}$: cumulative frequency of the class preceding the median class,

$\boldsymbol{f_{med}}$: frequency of the median class,

$\boldsymbol{W}$: median class size.

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example

Class	Class boundary	Frequency	Cumulative frequency
41 - 45	40.5 – 45.5	7	7
46 - 50	45.5 – 50.5	10	17
51 - 55	50.5 – 55.5	15	32
56 - 60	55.5 – 60.5	2	34
61 - 65	60.5 – 65.5	6	40
Total		40	

$N \div 2 = 40 \div 2 = 20$

.: median class = 51-55

$L = 50.5$

$N = 40$

$cf_p = 17$

$W = (50.5 - 55.5) = 5$

$f_{med} = 15$

$$\text{median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W) = 51.5$$

Mode

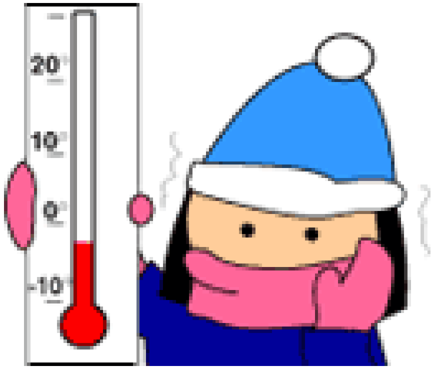
- The mode is the value that occurs most often.
- If no number is repeated, then there is no mode for the list.

Example

Case study: On a cold winter day in January, the temperature for 9 North American cities is recorded in Fahrenheit as follows:

-8, 0, -3, 4, 12, 0, 5, -1, 0

What is the mode of these temperatures?



Solution:

- Ordering the data from lowest to highest, we get:

  -8, -3, -1, 0, 0, 0, 4, 5, 12

- The mode of these temperatures is 0.

Example

Case study: A marathon race was completed by 5 participants. The time taken by each participant is recorded as follows:

2.7 hr, 8.3 hr, 3.5 hr, 5.1 hr, 4.9 hr

What is the mode of these times given in hours?

Solution:

Ordering the data from least to greatest, we get:

2.7, 3.5, 4.9, 5.1, 8.3

Since each value occurs only once in the data set, there is no mode for this set of data.

Example

Case study: In a crash test, 11 cars were tested to determine what impact speed was required to obtain minimal bumper damage. The collected data as shown below:

24, 15, 18, 20, 18, 22, 24, 26, 18, 26, 24

Find the mode of the speeds given in miles per hour.

Solution:

Ordering the data from least to greatest, we get:

15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26

Since both 18 and 24 occur three times, the modes are 18 and 24 miles per hour. This data set is **bimodal.**

Mode of Grouped Data

- The first step towards finding the mode of the grouped data is to locate the class interval with the maximum frequency.
- The class interval corresponding to the maximum frequency is called the modal class.

- The mode of grouped data is calculated using the formula:

The class interval corresponding to the maximum frequency is called the modal class.

$$\text{Mode} = l + h \times \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)}$$

Where,

$l \rightarrow$ lower class limit of the modal class.

$h \rightarrow$ class size.

$f_1 \rightarrow$ frequency of the modal class.

$f_0 \rightarrow$ frequency of the class preceding or just before the modal class.

$f_2 \rightarrow$ frequency of the class succeeding or just after the modal class.

Example

Compute the mode of the test score below.

Score (Class)	Frequency
41 - 45	1
36 - 40	5
31 - 35	7
26 - 30	16
21 - 25	8
16 - 20	2
Total	

.: modal class = 26 - 30

$$l = 26 - 0.5 = 25.5$$

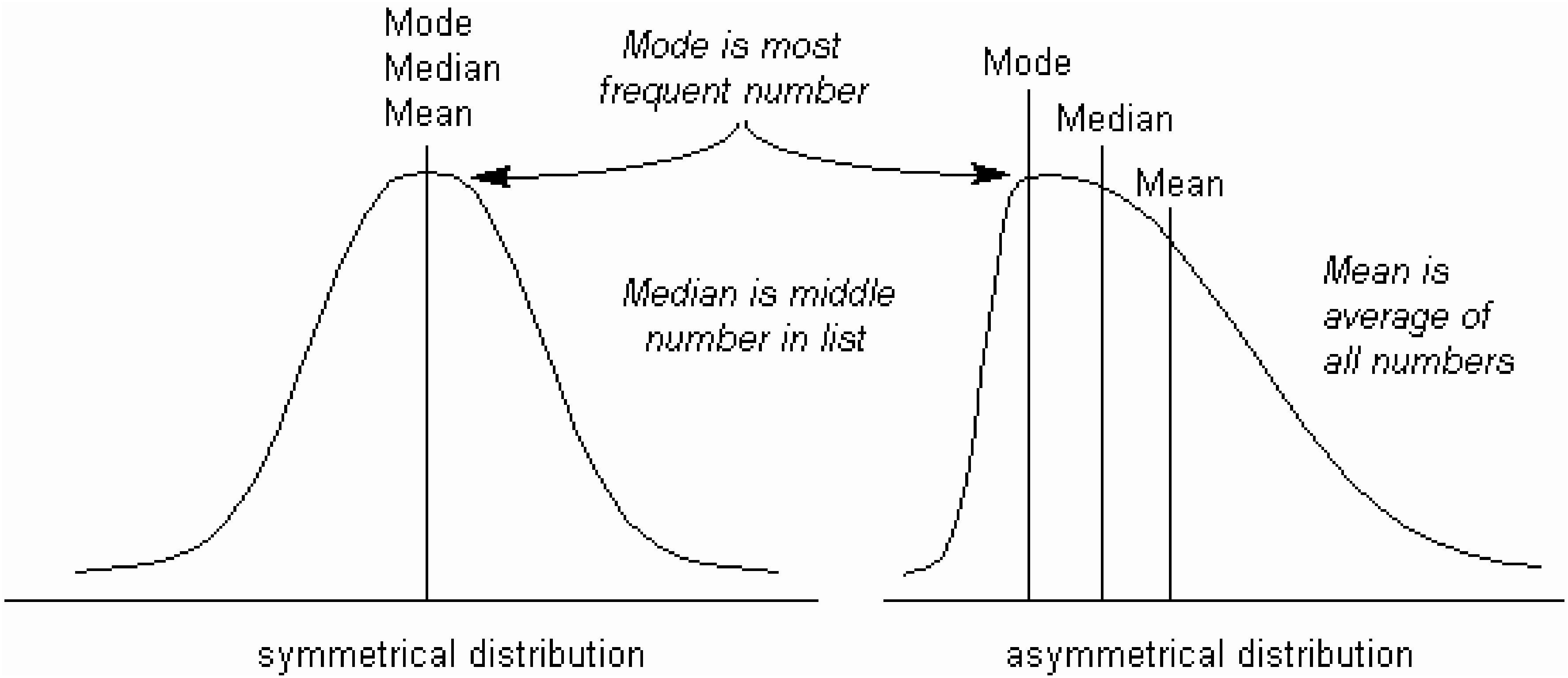
$$h = 5$$

$$f_1 = 14$$

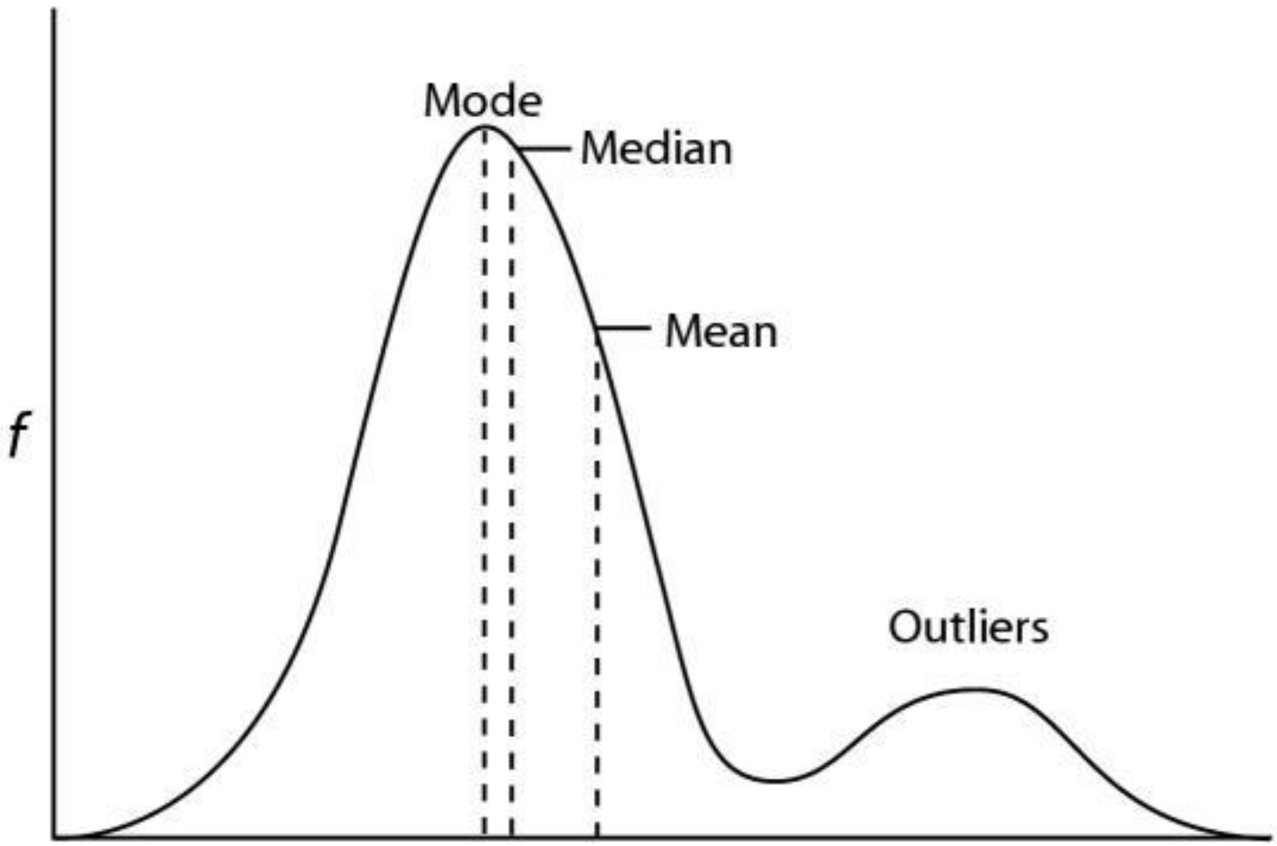
$$f_0 = 7$$

$$f_2 = 8$$

$$Mode = 25.5 + \left(\frac{14 - 7}{2 * 14 - 7 - 8}\right)5 = 28.2$$


Mode
Median
Mean
Mode is most
frequent number
Mode
Median
Mean
Median is middle
number in list
Mean is
average of
all numbers
symmetrical distribution
asymmetrical distribution

Example



Exercise #1

The owner of a shoe shop recorded the sizes of the feet of all the customers who bought shoes in his shop in one morning. These sizes are listed below:

8, 7, 4, 5, 9, 13, 10, 8, 8, 7, 6, 5, 3, 11,10, 8, 5, 4, 8, 6

(a)What is the mean of these values?

(b)What is the median of these values?

(c)What is the mode of these values?

Exercise #2

The table below gives the number of accidents each year at a particular road junction:

2001	2002	2003	2004	2005	2006	2007	2008
4	5	4	52	10	5	3	5

(a) Calculate the mean, median and mode for the values above.

(b) A road safety group want the council to do some improvement to make this junction safer. Which measure will they use to argue for this?

(c) The council don't want to spend money on the road junction. Which measure will they use to argue that safety work is not necessary?

Exercise #3

You grew fifty baby carrots using special soil. You dig them up and measure their lengths (to the nearest mm) and group the results:

Find the following:

(a) mean
(b) median
(c) mode

Length (mm) Class	Frequency
150 – 154	5
155 – 159	2
160 – 164	6
165 – 169	8
170 – 174	9
175 – 179	11
180 – 184	6
185 – 189	3

Exercise #4

Find mean, median and mode corresponding to the frequency table of samples of students cars and staff cars obtained from a college.

Age of vehicles	Students	Staff
1 – 3	23	30
4– 6	33	47
7 – 9	63	36
10 – 12	68	30
13 – 15	19	8
16 – 18	10	0
19 – 21	1	0
22 – 24	0	1

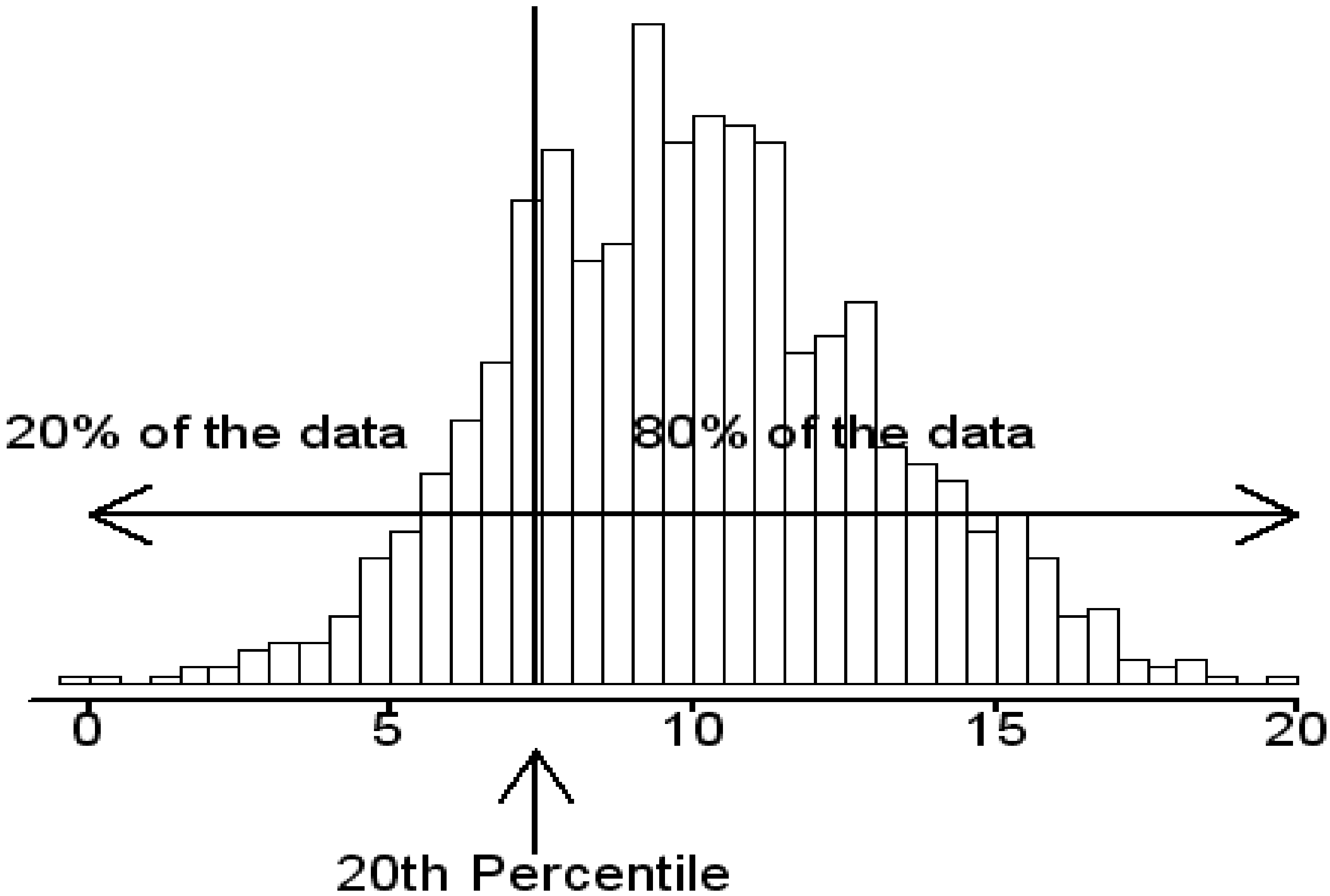
Data Profiles

- Percentile
- Quartile

Percentile

In a population or a sample, the P-th percentile is a value such that at least P percent of the values take on this value or less and at least (100-P) percent of the values take on this value or more.

Example



- Sort the data set so measurements are in order from lowest to highest,

  $Y[1], Y[2], \ldots., Y[N]$

- Calculate,

$$i = \frac{P}{100}(N)$$

- If i is not an integer, round up to the next highest integer k and use $Y[k]$ as the percentile estimate.
- If i is an integer, use $(Y[i] + Y[i+1]) \div 2$ as the percentile estimate.

Example

Given a set of data: 12, 4, 6, 11, 9,15, 20, 18, 25, 30

i) Calculate 80th percentile.

ii) Calculate 68th percentile.

Solution (i): 80th percentile

•Arrange in order:

4 6 9 11 12 15 18 20 25 30

•N = 10, P = 80 ; i = 80 x 10 ÷ 100 = 8

Y[8] = 20, Y[9] = 25, P_{80} = (20 + 25) ÷ 2 = 22.5

Example

Solution (ii): 68th percentile

$N = 10, \ P = 68$

4 6 9 11 12 15 18 20 25 30

$i = 68 \text{ x } 10 \div 100 = 6.8, \quad k = 7$

$\mathrm{Y}[7] = 18, \quad \mathrm{P}_{68} = 18$

- The process of finding the percentile that corresponds to a particular value x is:

$$\text{percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}}(100)$$

Example

Given a set of data as follows:

12 4 6 11 9 15 20 18 25 30

Find the percentile corresponding to the value,

$$Y[k] = 15$$

Solution:

Arrange the data in order,

4 6 9 11 12 15 18 20 25 30

$$\text{percentile of value } 15 = \frac{\text{number of values less than } 15}{\text{total number of values}}(100) = \frac{5}{10}(100) = 50$$

.: The value 15 is at the 50th percentile.

Quartile

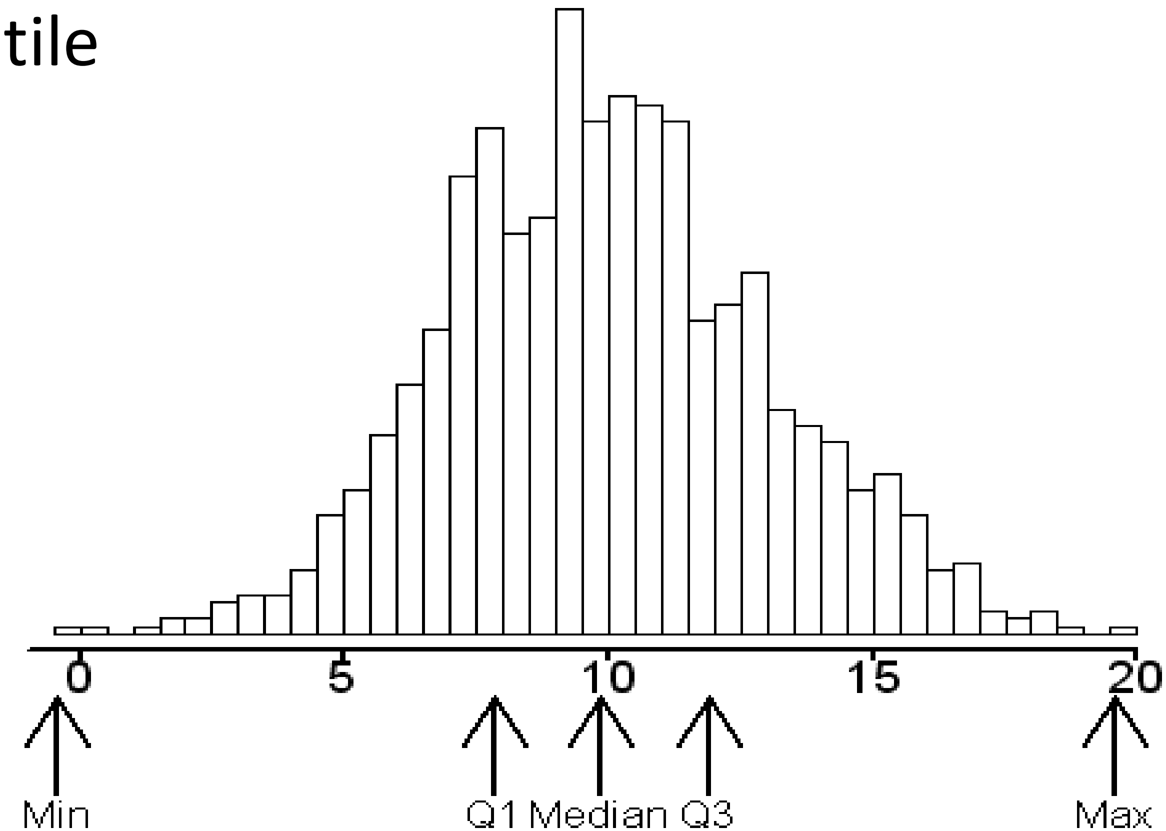
The 1st, 2nd, and 3d quartiles are the 25th, 50th, and 75th percentiles respectively.

Example

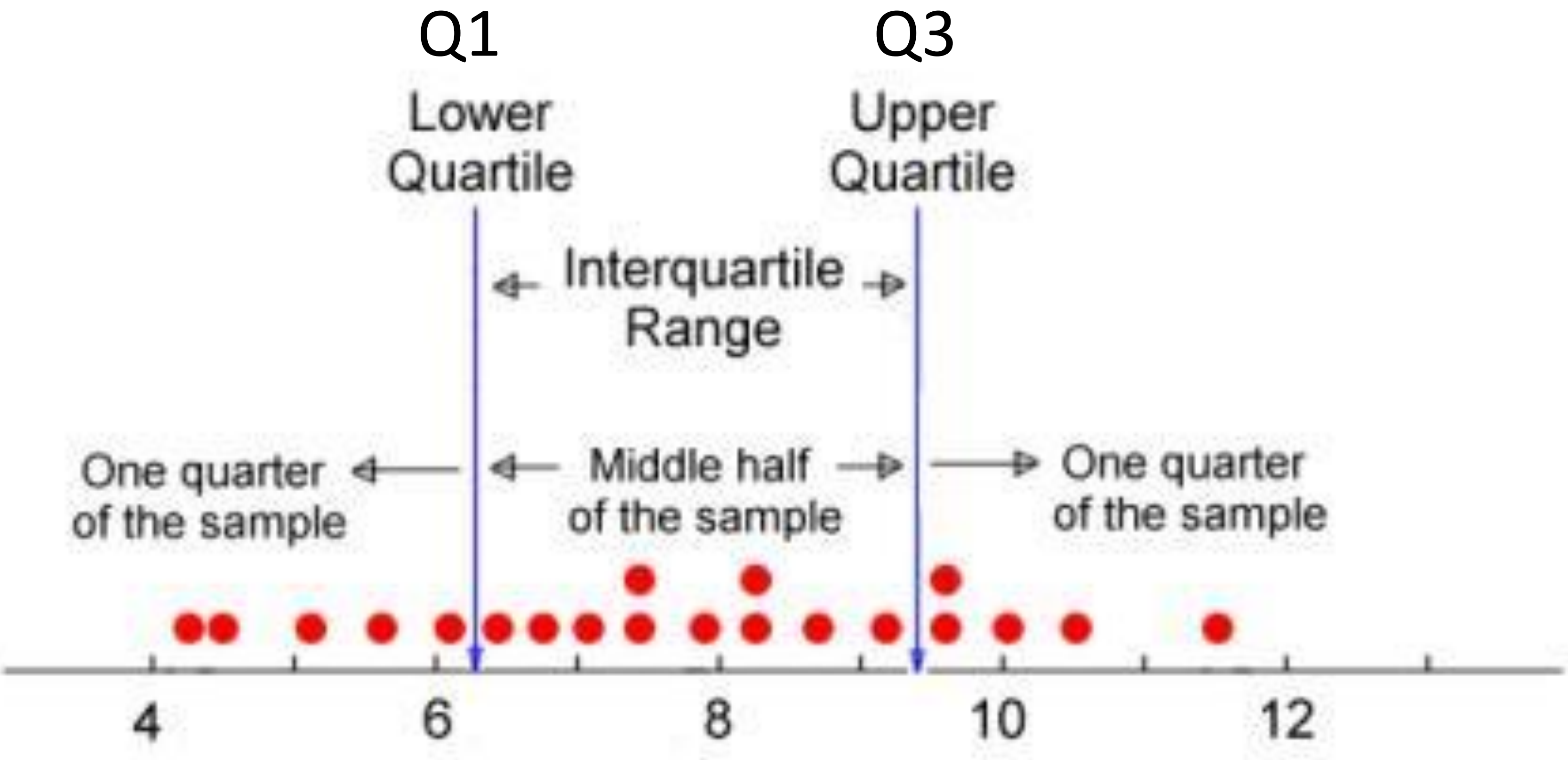
Q1 – 25th percentile

Q2 – 50th percentile (median)

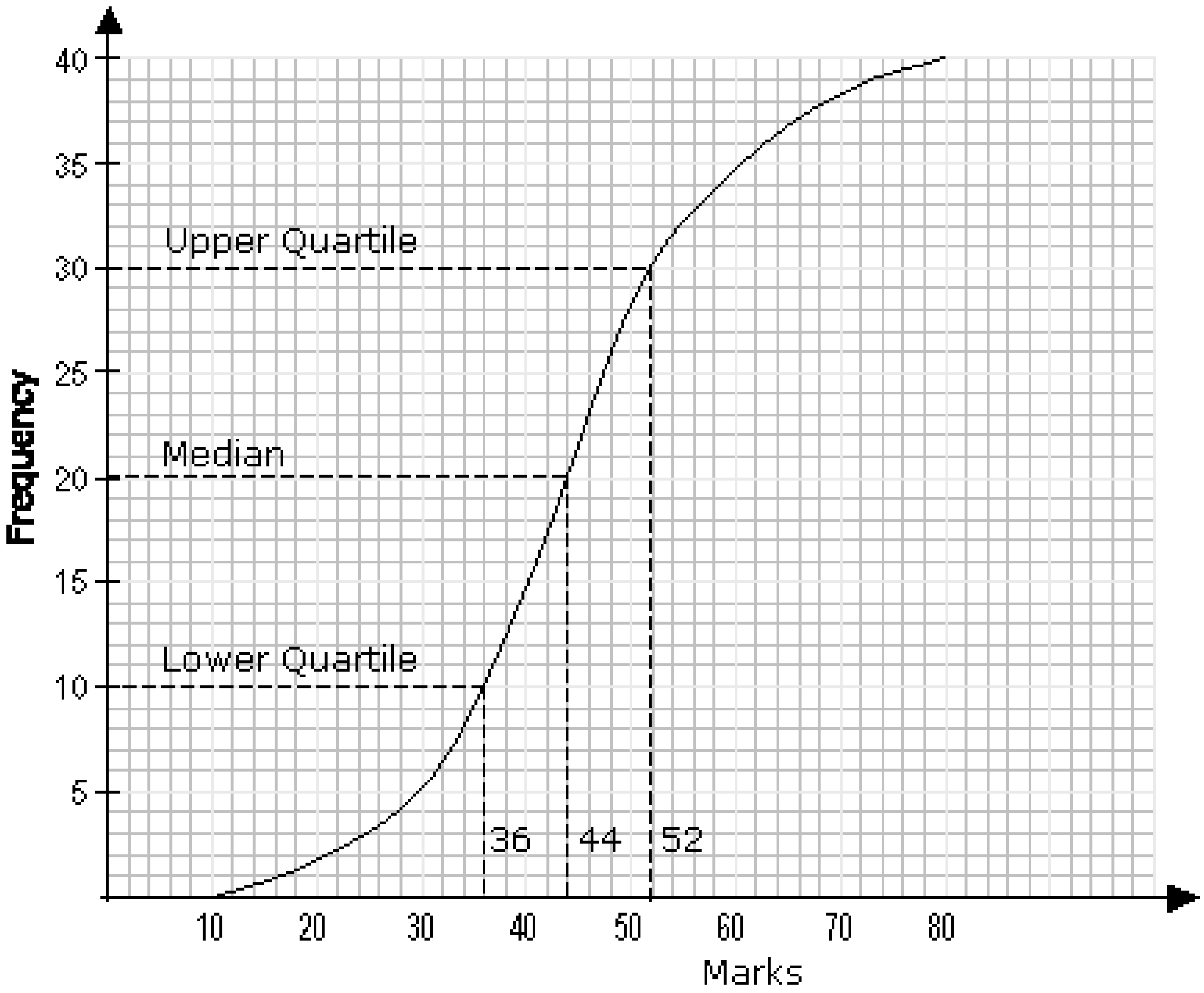
Q3 – 75th percentile



Example


Q1
Lower Quartile
Q3
Upper Quartile
Interquartile Range
One quarter of the sample
Middle half of the sample
One quarter of the sample
4
6
8
10
12

Example


Frequency
40
35
30
25
20
15
10
5
Upper Quartile
Median
Lower Quartile
36
44
52
10
20
30
40
50
60
70
80
Marks

Exercise #5

0.7901	0.8044	0.8062	0.8073	0.8079	0.8110
0.8126	0.8128	0.8143	0.8150	0.8150	0.8152
0.8152	0.8161	0.8161	0.8163	0.8165	0.8170

(a) Use the 18 sorted (left to right) weights of regular can drinks to find the percentile corresponding to the given value.

i. 0.8143

ii. 0.8062

(b) Find the indicated percentile and quartile.

i. P_{80}

ii. Q_3

iii. P_{33}

iv. Q_1

Measures of Dispersion & Shape

Measures of Dispersion

- Measures of dispersion measure how spread out a set of data is.

- Measurement elements used are **Range, variance, standard deviation.**

Range

- The range is the largest number in a set minus the smallest number.

- Example: 13, 18, 13, 14, 13, 16, 14, 21, 13

  The largest value in the list is 21, and the smallest is 13,

  so the range is 21 – 13 = 8.

Variance

- A measure of the dispersion of a set of data points around their mean value.

- It is a mathematical expectation of the average squared deviations from the mean.

- Sample:

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

- Population:

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

Standard Deviation

- A statistic used as a measure of the dispersion or variation in a distribution, equal to the square root of the arithmetic mean of the squares of the deviations from the arithmetic mean.

- It is the square root of the variance.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation
N = the size of the population
x_i = each value from the population
μ = the population mean

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

s = sample standard deviation
N = the number of observations
x_i = the observed values of a sample item
$\overline{x}$ = the mean value of the observations

Exercise #6

Adam has been playing golf on the weekends for the past three years. Recently, he started keeping track of his recorded scores. His scores for June and July at his favorite 9-hole (par 36) golf course are provided below:

45 49 42 56 41 36 34 38 41 45 40 42 41 39 38 40 39 36 41

Find the Range, Mean, Variance and Standard Deviation for the above data.

Measures of Shape

- Skewness
- Kurtosis

Skewness

- Occurs when a distribution is not symmetrical about its mean.
- A distribution is symmetrical when its median, mean, and mode are equal.
- A positively skewed (skewed to the right) distribution occurs when the mean exceeds the median.
- A negatively skewed (skewed to the left) distribution occurs when the mean is less than the median.

Measuring Skewness

- Formula to measure skewness for univariate data x_1, x_2, .., x_N :

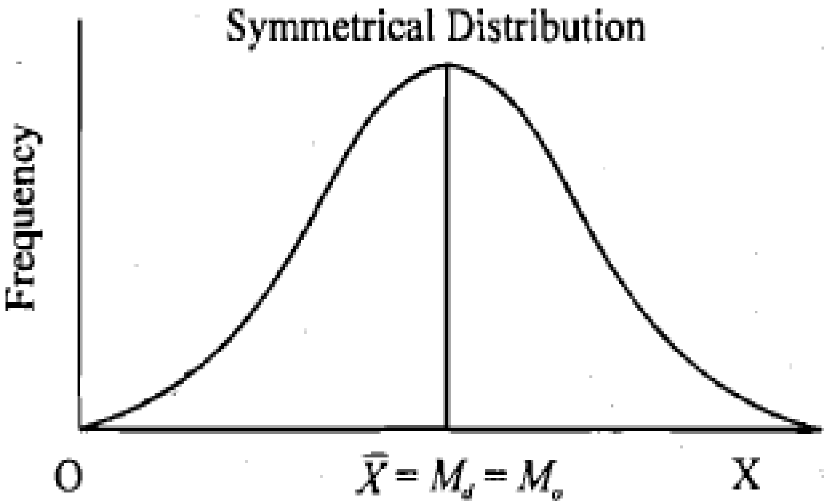
$$Skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

$\bar{x}$ = mean

Number of data points

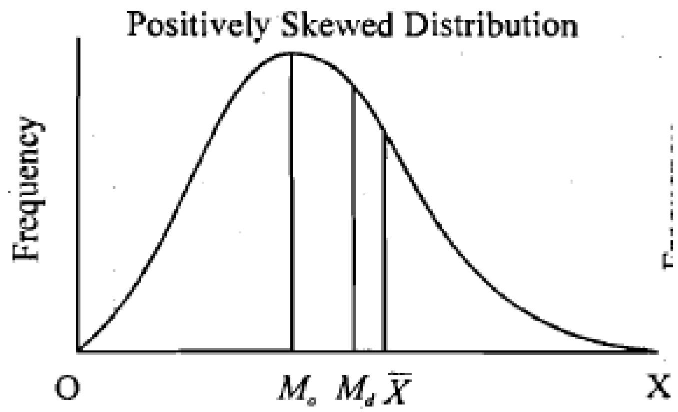
s = std deviation

- For normal distribution (symmetric distribution):
  - skewness = 0.
- Any symmetric data should have:
  - Skewness value near zero.
  - Distribution with mean, median and mode fall at the same point.

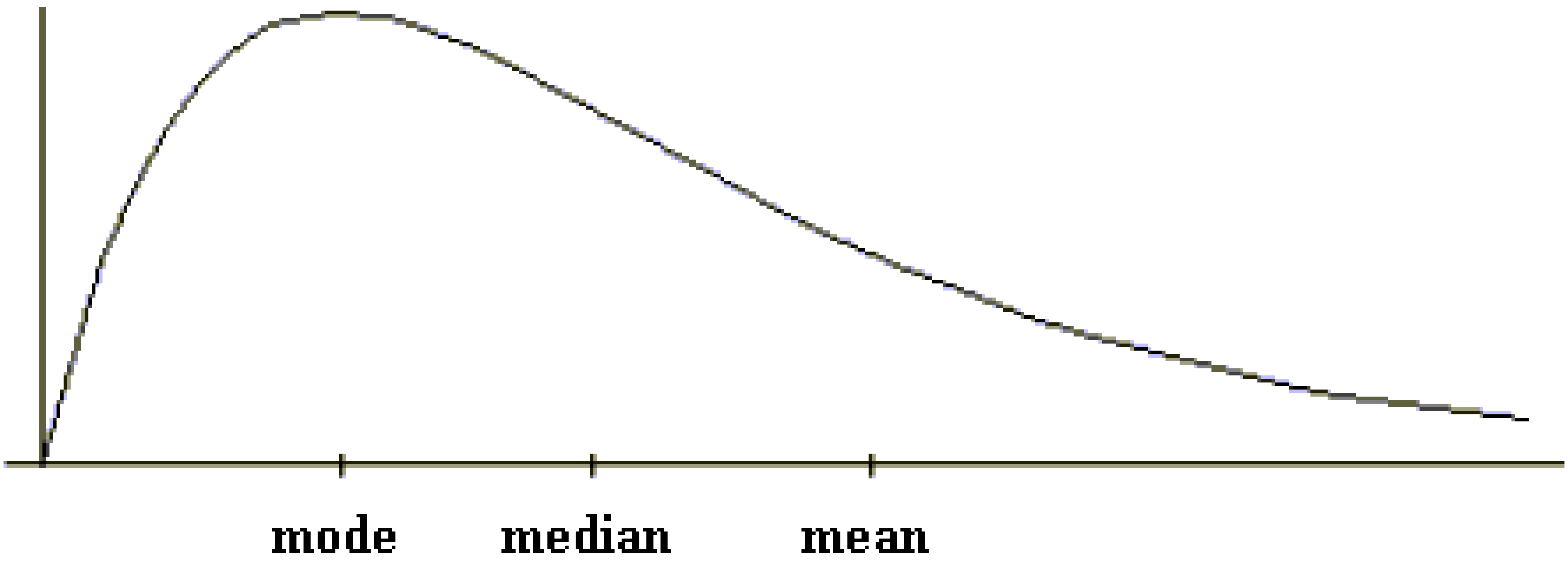


Positive/Right Skewed

- Skewness > 0
  - The distribution is asymmetrical and points in the positive direction.
  - Example: Test scores of difficult examination where almost everyone did poorly on it.
  - mode < median < mean

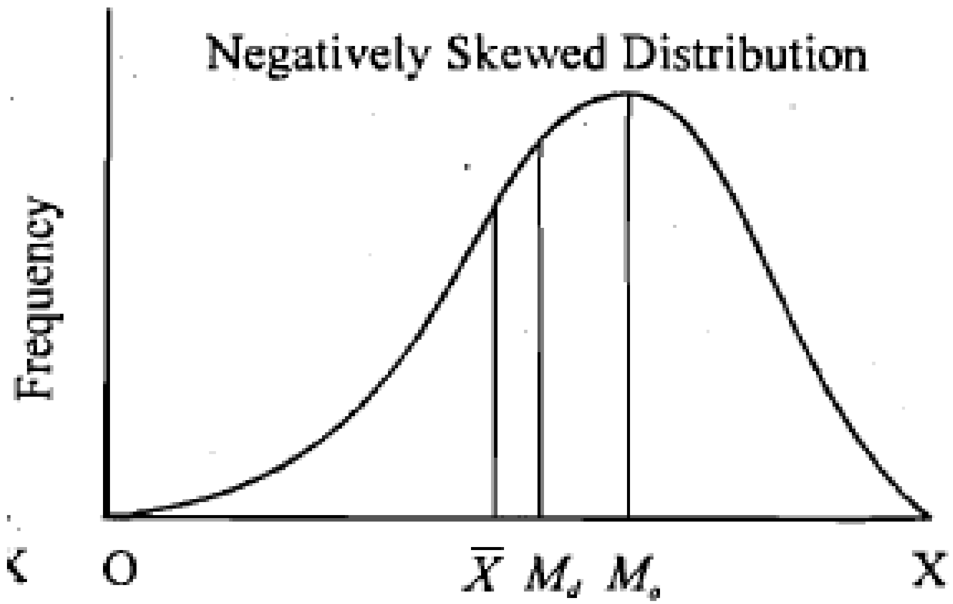


Positive/Right Skewed

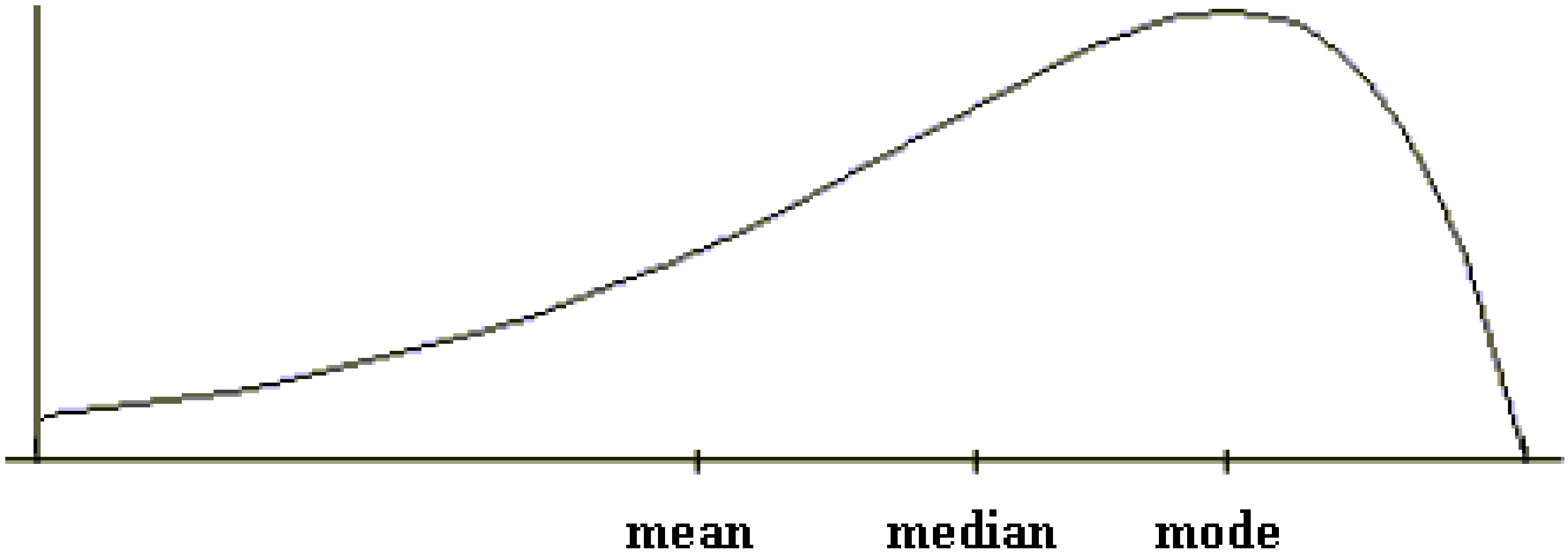


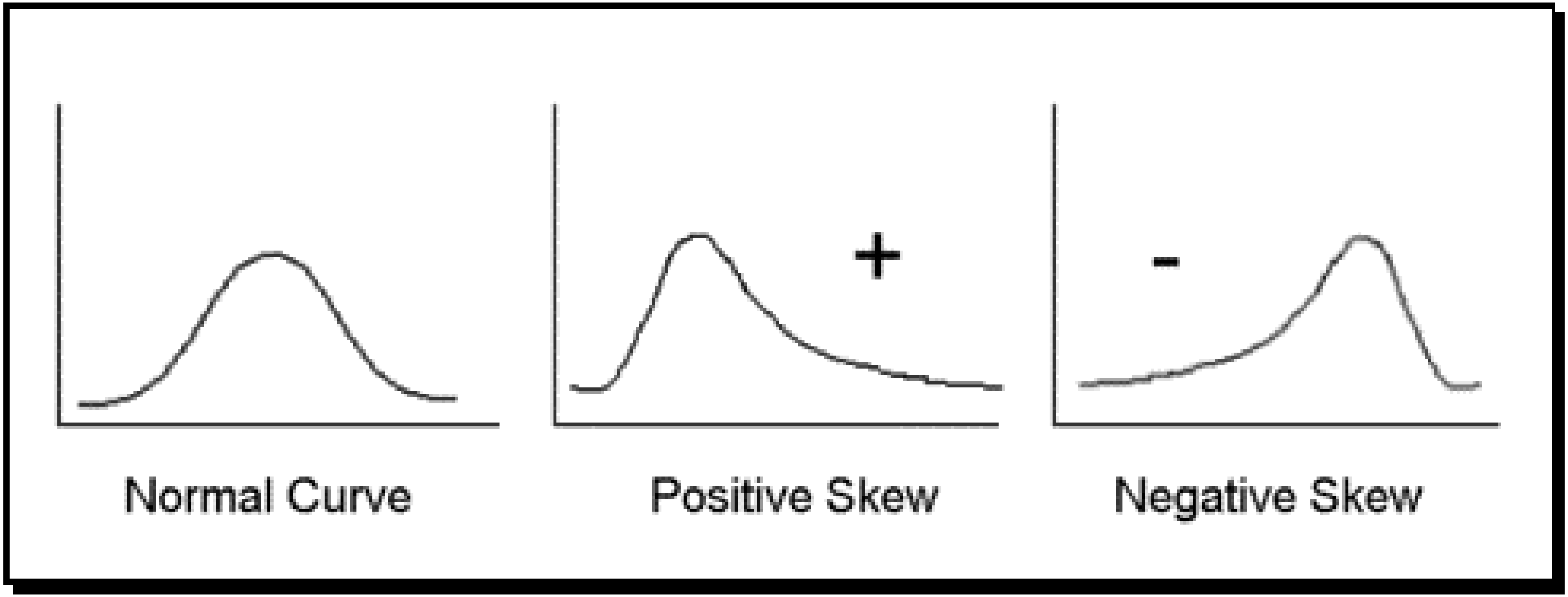
Negative/Left Skewed

- Skewness < 0
  - The distribution is asymmetrical and points in the negative direction.
  - Example: Test scores of difficult examination where almost everyone did good on it.
  - mode > median > mean



Negative/Left Skewed




+
-
Normal Curve
Positive Skew
Negative Skew

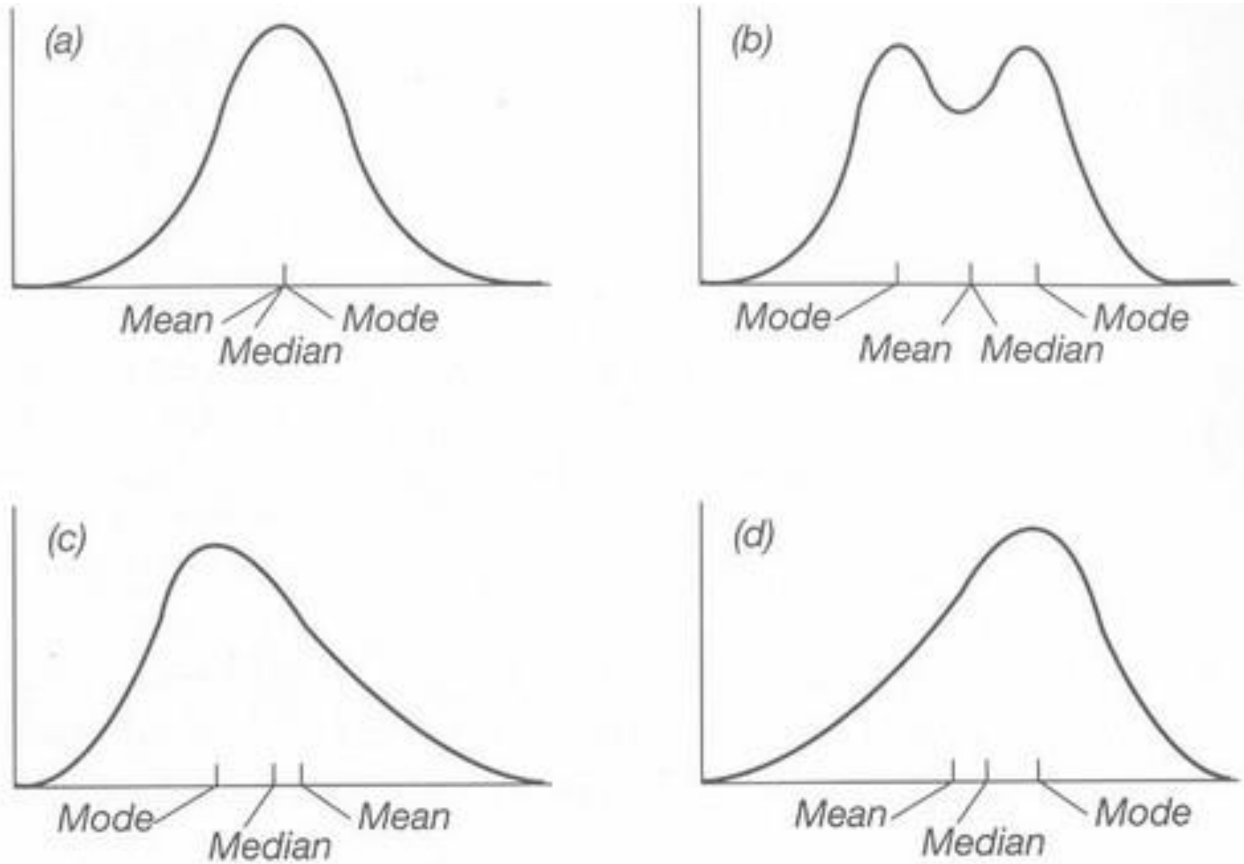


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution (b) is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.*

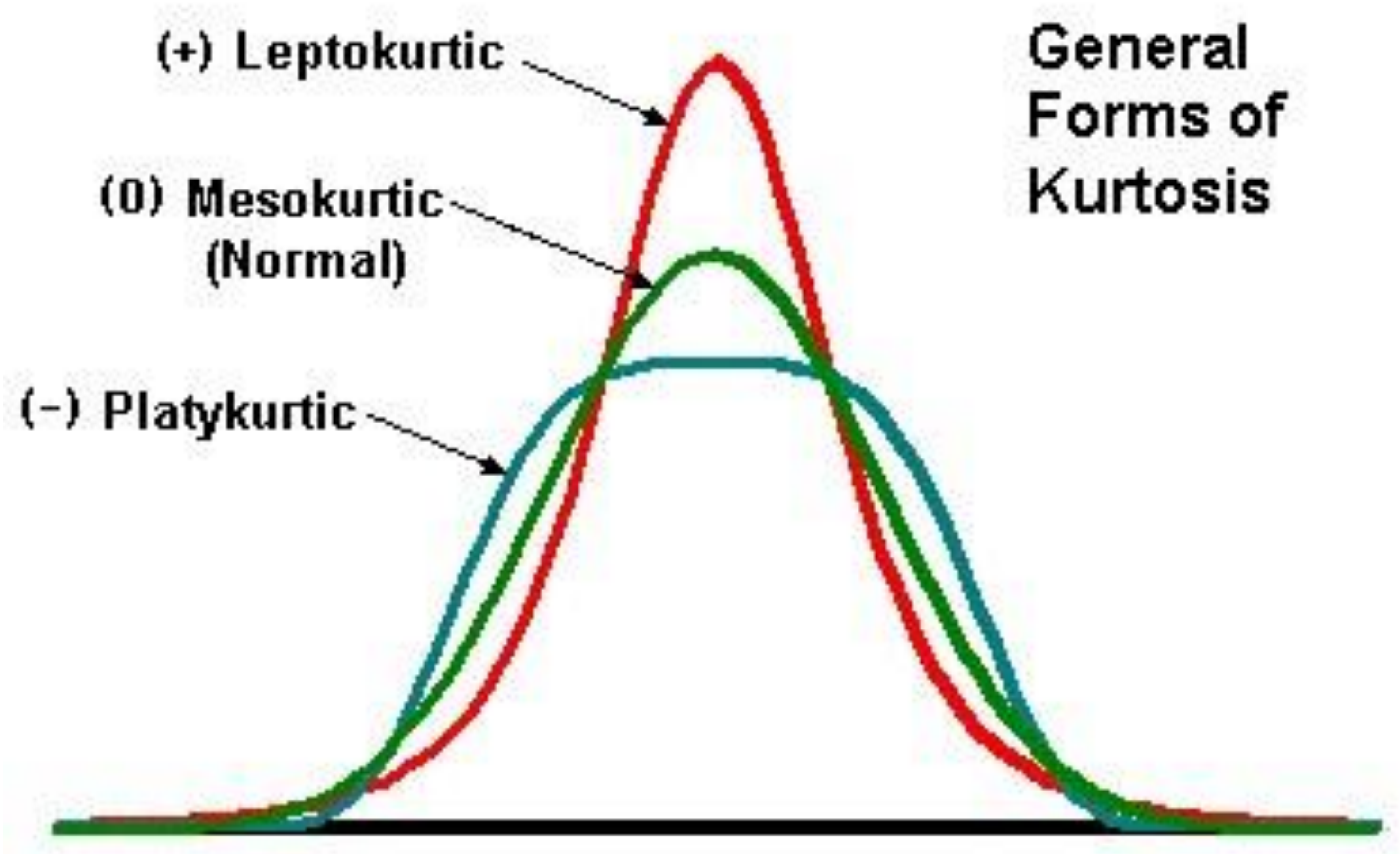
Kurtosis

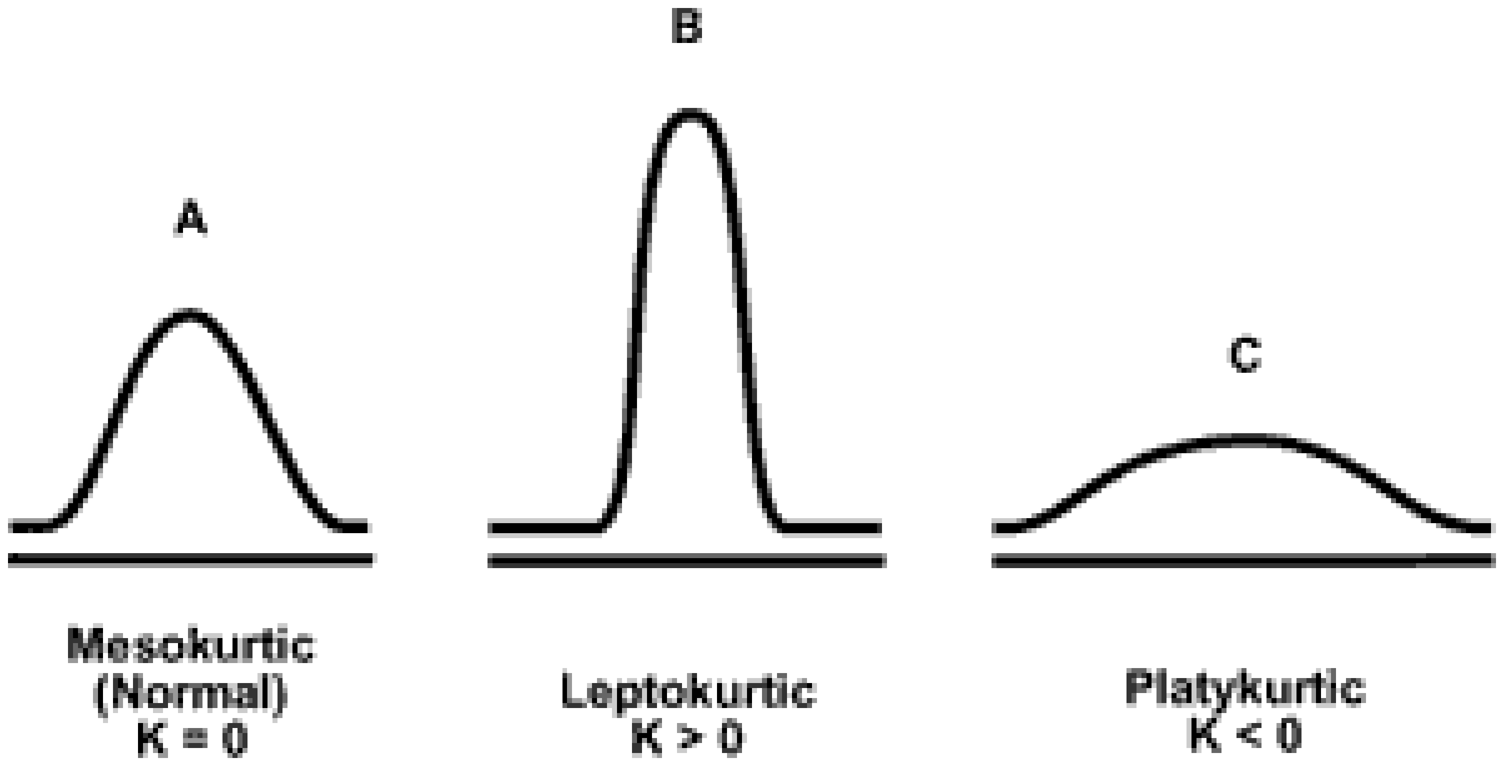
- **Kurtosis** is the statistic which describes the degree of peakedness or flatness of a probability distribution relative to the benchmark normal distribution.
- In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution
- Formula to measure kurtosis for univariate data x_1, x_2, .., x_N:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

Excess Kurtosis

- Excess kurtosis is simply **kurtosis – 3**.
- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈ 3 (excess ≈ 0) is called **mesokurtic**.
- A distribution with kurtosis < 3 (excess kurtosis < 0) is called **platykurtic**. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A distribution with kurtosis > 3 (excess kurtosis > 0) is called **leptokurtic**. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.


(+) Leptokurtic
(0) Mesokurtic
(Normal)
(-) Platykurtic
General
Forms of
Kurtosis


A
B
C
Mesokurtic
(Normal)
K = 0
Leptokurtic
K > 0
Platykurtic
K < 0

Example

Mira is interested in the elapse time (in minutes) she spends on riding a tricycle from home at Taman U to School of Computing, for three weeks (excluding weekends). She obtain the following data:

19.09, 19.55, 17.89, 17.73, 25.15, 27.27, 25.24, 21.05, 21.65, 20.92, 22.61, 15.71, 22.04, 22.60, 24.25.

Compute and interpret the skewness and kurtosis.

Example - Solution

$$Skewness = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{(N-1)s^3}$$

$$= \frac{\sum_{i=1}^{15} (x_i - 21.52)^3}{(15-1)s^3} = \frac{-8.245}{(14)(3.18)^3} = -0.0183$$

Interpretation: The skewness here is -0.0183. This value implies that the distribution of the data is *slightly skewed to the left* or *negatively skewed*. It is skewed to the left because the computed value is negative, and is slightly, because the value is close to zero.

Kurtosis = $$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

$$= \frac{\sum_{i=1}^{15}(x_i - 21.52)^4}{(15-1)s^4} = \frac{3086.1}{(14)(3.18)^4} = 2.15$$

Interpretation: For the kurtosis, we have 2.15 implying that the distribution of the data is *platykurtic*, since the computed value is less than 3.