SECI1143: PROBABILITY & STATISTICAL DATA ANALYSIS

CHAPTER 3

Descriptive Statistics

Prepared & updated by: Razana/Suhaila/Nzah

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: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

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) ÷ 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,

$$\overline{X} = \frac{\sum_{i=1}^{h} f_i X_i}{n} = \frac{f_1 X_1 + f_2 X_2 + \ldots. + f_h X_h}{f_1 + f_2 + \ldots. + f_h}$$

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:

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

$$\overline{X} = \frac{\sum_{i=1}^{h} f_i X_i}{n} = \frac{72}{29} = 2.5$$

Example

Find the mean value for the following data:

Class interval	Frequency
40.5 – 45.5	7
45.5 – 50.5	10
50.5 – 55.5	15
55.5 – 60.5	12
60.5 – 65.5	6
Total	50

Example

Class interval	Midpoint	Frequency	$f_i X_i$
40.5 – 45.5	(40.5 + 45.5) ÷ 2 = 43	7	43 x 7 = 301
45.5 – 50.5	48	10	480
50.5 – 55.5	53	15	795
55.5 – 60.5	58	12	696
60.5 – 65.5	63	6	378
Total		50	2650

$$\overline{X} = \frac{\sum_{i=1}^{h} f_i X_i}{n} = \frac{2650}{50} = 53$$

Median

- The median is the **middle** value when the data are arranged from smallest to largest.
- To find the median, your numbers must 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) ÷ 2 = 10 ÷ 2 = 5th** number.

- So, the **median is 14**.

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

- Example:

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

(5 + 6) ÷ 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**

  (sort) ↓

  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)*

L : lower class limit of median class,

N : total number of observations,

cf_p : cumulative frequency of the class preceding the median class,

f_{med} : frequency of the median class,

W : median class size.

Example

Class interval	Frequency	Cumulative frequency
40.5 – 45.5	7	7
45.5 – 50.5	10	**17**
50.5 – 55.5	**15**	**32**
55.5 – 60.5	2	34
60.5 – 65.5	6	40
Total	40	

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

.: median class = 50.5 - 55.5

$$L = 50.5$$

$$N = 40$$

$$cf_p = 17$$

$W = 5$ (55.5-50.5)

$$f_{med} = 15$$

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

Mode

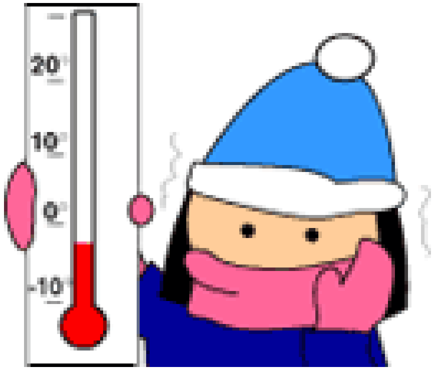
- The mode is the value that occurs **most often**.
- If **no number is repeated**, then there is **no mode** for the list.

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 least to greatest, we get:

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

- The **mode of these temperatures is 0 Fahrenheit**.

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 data set.

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.

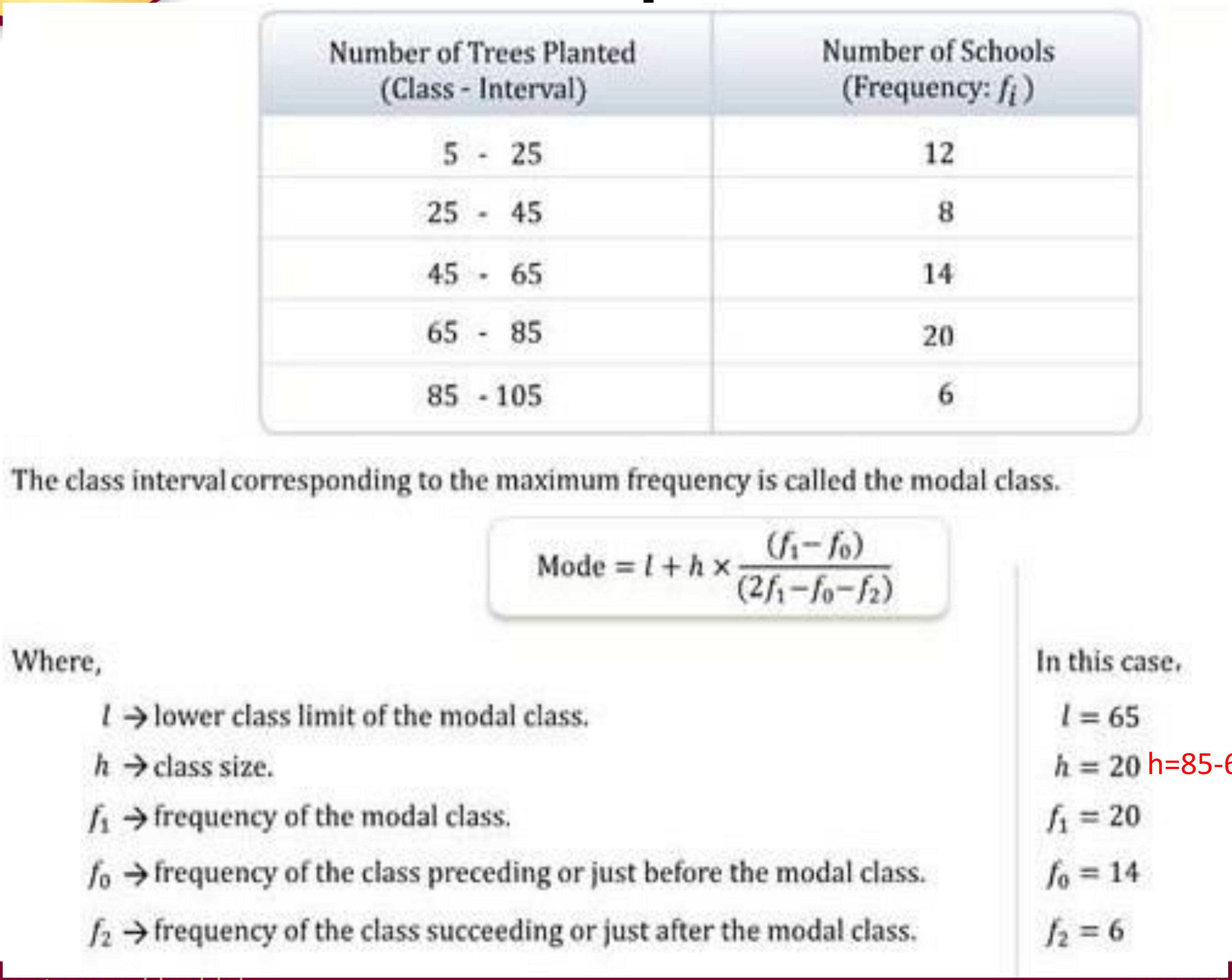
The mode of grouped data is calculated using the formula

$$\textbf{Mode} = l + h \times [(f_1 - f_0) \div (2f_1 - f_0 - f_2)]$$

where,

- $\boldsymbol{l}$: lower class limit of the modal class
- $\boldsymbol{h}$: class size if all class intervals have the same class size
- $\boldsymbol{f}_1$: frequency of the modal class
- $\boldsymbol{f}_0$: frequency of the class preceding or just before the modal class
- $\boldsymbol{f}_2$: frequency of the class succeeding or just after the modal class

Example



Number of Trees Planted (Class - Interval)	Number of Schools (Frequency: f_i)
5 - 25	12
25 - 45	8
45 - 65	14
65 - 85	20
85 - 105	6

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 → lower class limit of the modal class.

h → class size.

f_1 → frequency of the modal class.

f_0 → frequency of the class preceding or just before the modal class.

f_2 → frequency of the class succeeding or just after the modal class.

In this case,

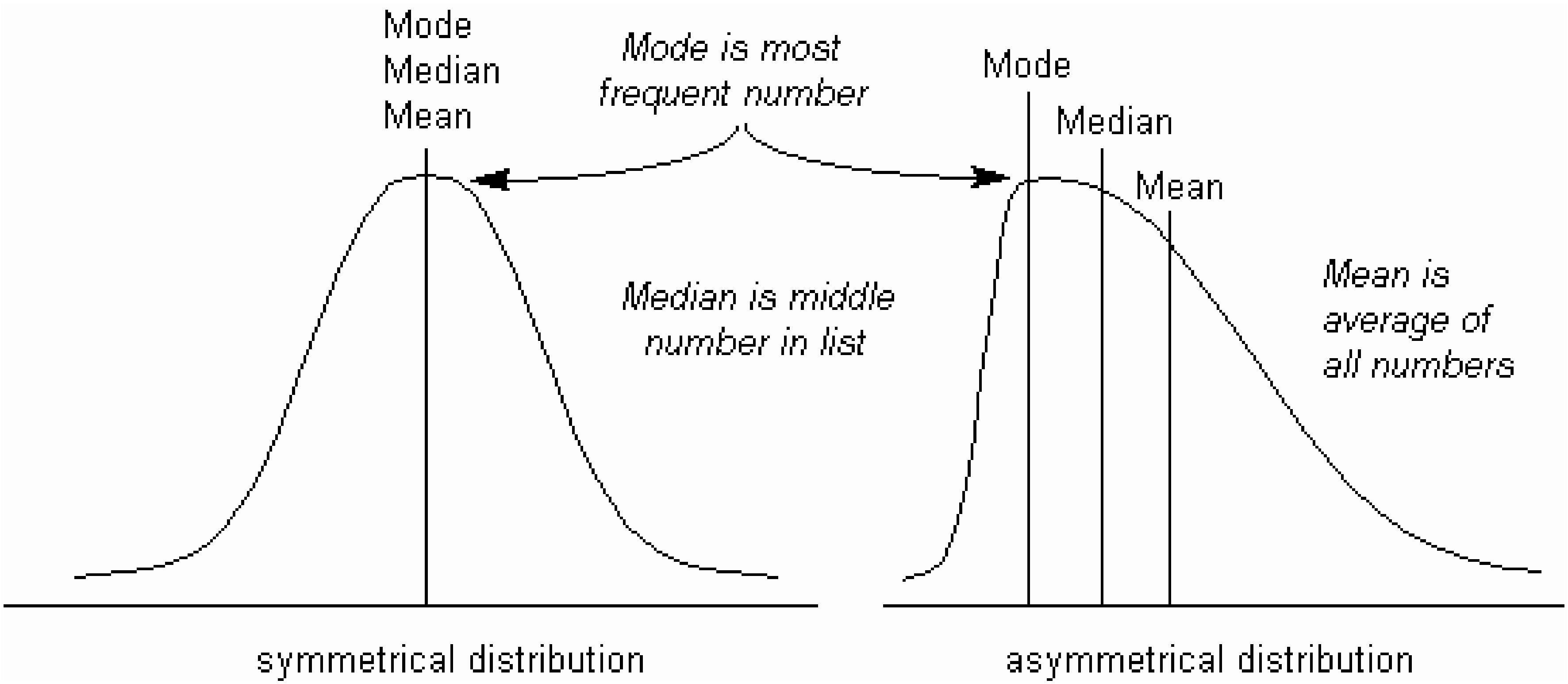
$l = 65$

$h = 20$ h=85-65=20

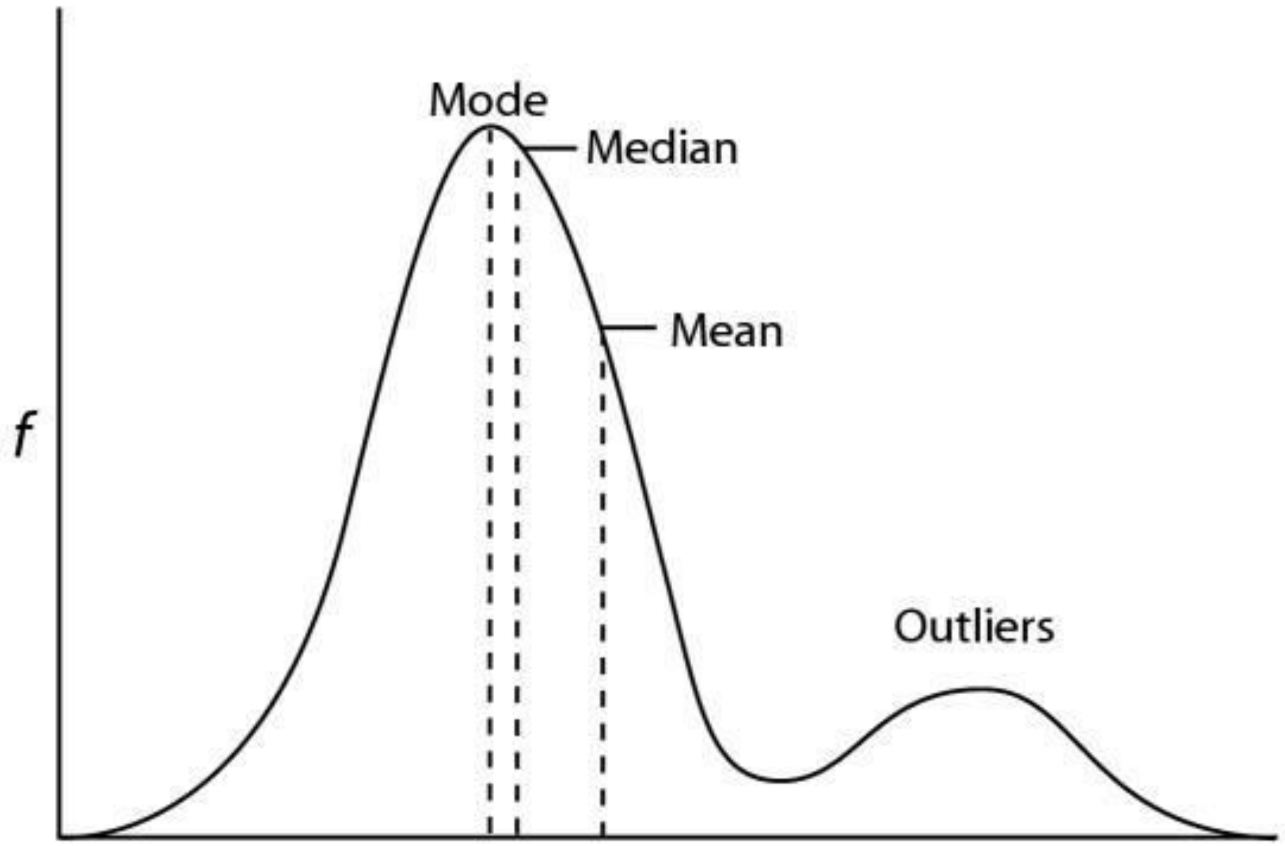
$f_1 = 20$

$f_0 = 14$

$f_2 = 6$


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	2	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

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

Length (mm)	Frequency
149.5 – 154.5	5
154.5 – 159.5	2
159.5 – 164.5	6
164.5 – 169.5	8
169.5 – 174.5	9
174.5 – 179.5	11
179.5 – 184.5	6
184.5 – 189.5	3

Exercise #4

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

Age	Students	Faculty/Staff
0.5 – 3.5	23	30
3.5– 6.5	33	47
6.5 – 9.5	63	36
9.5 – 12.5	68	30
12.5 – 15.5	19	8
15.5 – 18.5	10	0
18.5 – 21.5	1	0
21.5 – 24.5	0	1

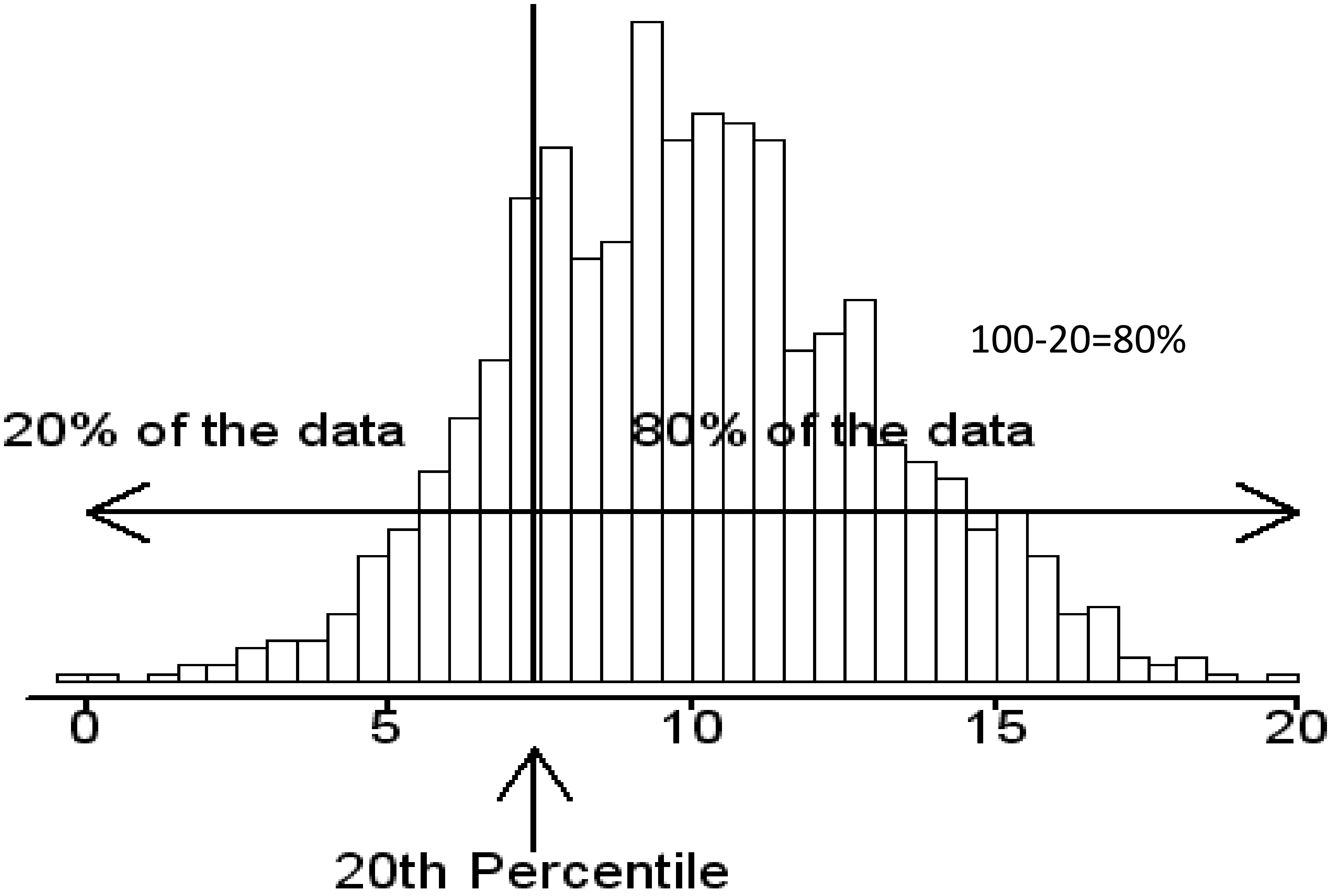
Data Profiles

- 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



100-20=80%

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

  Y[1], Y[2], …. , 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]) ÷ 2** as the percentile estimate.

Example

Given set of data:

12, 4, 6, 11, 9,15, 20, 18, 25, 30

i) Calculate 80th percentile.

ii) Calculate 68th percentile.

Solution (i):

- Arrange in order:

4 6 9 11 12 15 18 20 25 30

- $N = 10$, $P = 80$; $i = 80 \times 10 \div 100 = 8$ (integer)

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

Example

Solution (ii): Calculate 68th percentile

N = 10, P = 68

4 6 9 11 12 15 18 20 25 30

$i = 68 \times 10 \div 100 = 6.8$, $k = 7$

(not integer)

Y[7] = 18, $P_{68} = 18$

- 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 the 50th percentile.

Quartile

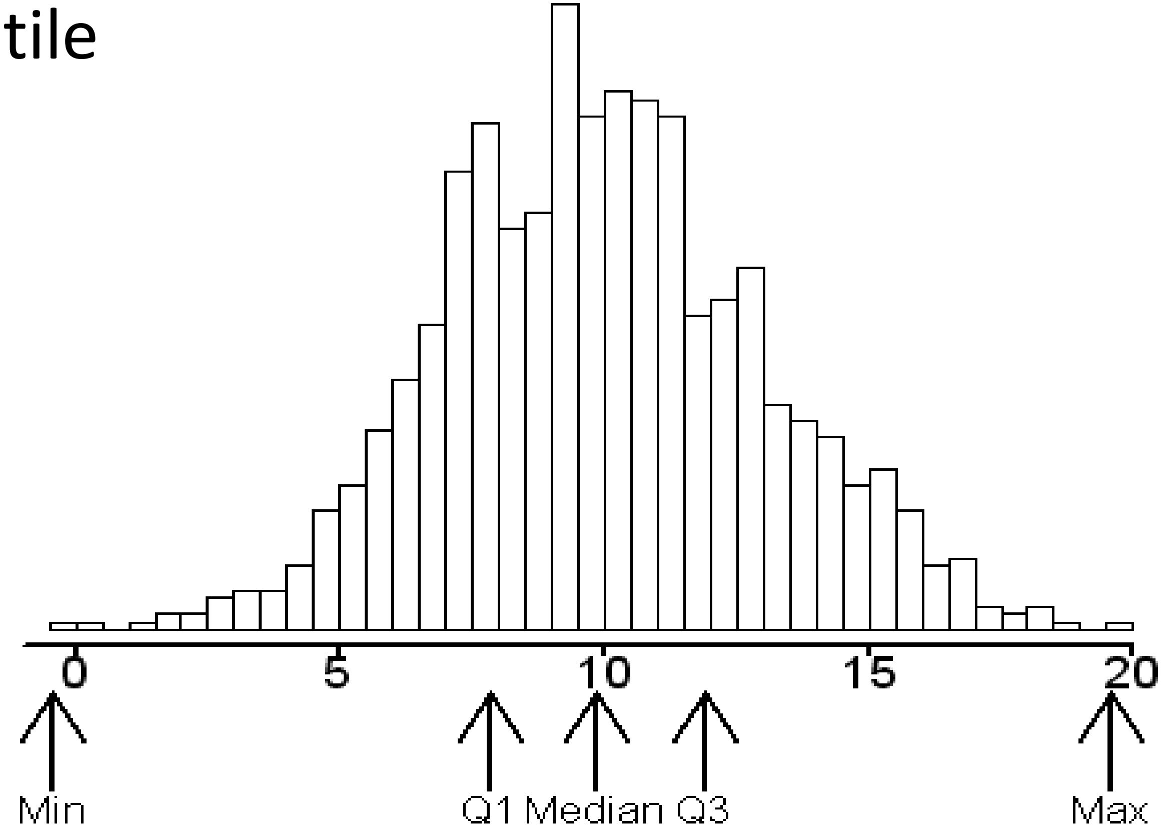
The 1st, 2nd, and 3d quartiles are the 25th, 50th, and 75th percentiles, respectively.

Example

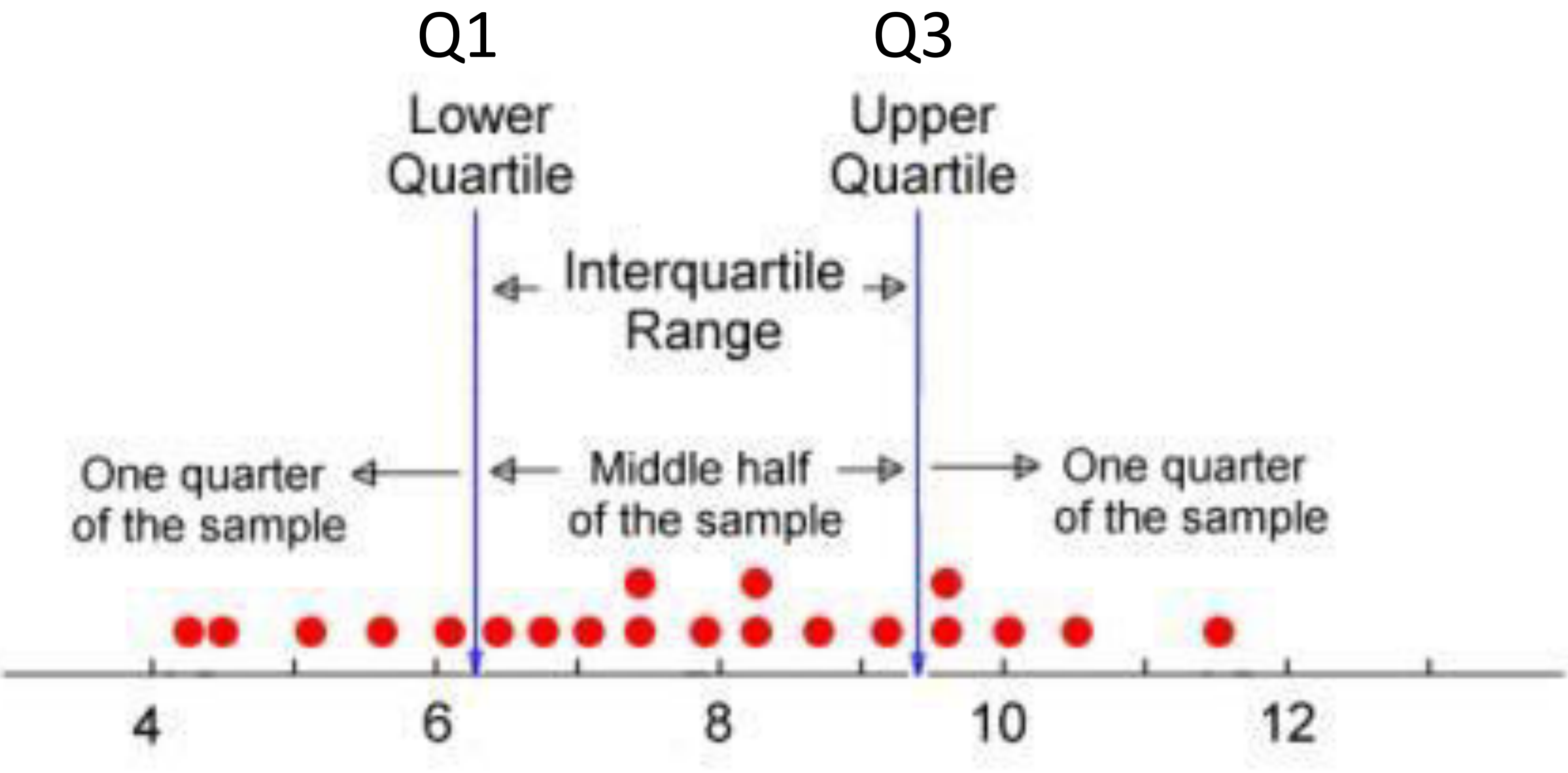
Q1 – 25th percentile

Q2 – 50th percentile (median)

Q3 – 75th percentile

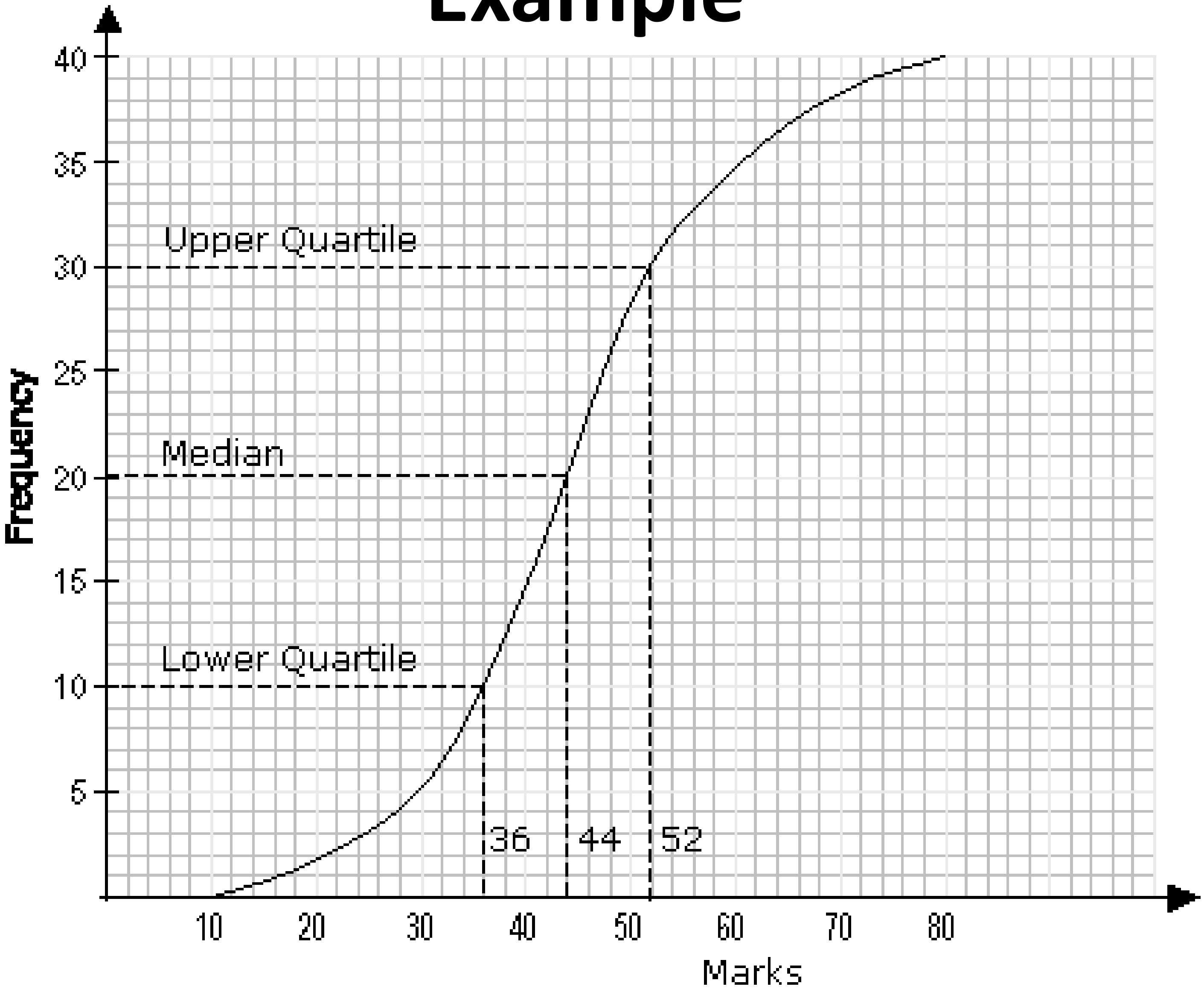


Example


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

Note:
Interquartile Range (IQR) = Q3 – Q1
Lower Limit: Q1 – 1.5 × IQR
Higher Limit: Q3 + 1.5 × IQR

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.

- **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.

- low variance: data points are close to mean (consistent)
- high variance: data points are far from mean and each other (inconsistent)

- Sample variance:

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

- Population variance:

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

Standard Deviation

- A statistic is 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**.

Exercise #6

Bailey has been playing golf on the weekends for the past three years. Recently, she started keeping track of her recorded scores. Her scores for June and July at her favorites 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, Variance, and Standard Deviation for the above data.

Measures of Shape

- **Skewness**:
  - **Positive skew**: Income distribution (few high earners drag mean > median).
  - **Negative skew**: Age at retirement (most retire early, few very late).
- **Kurtosis**:
  - **High kurtosis (Leptokurtic)**: Stock market crashes (extreme outliers).
  - **Low kurtosis (Platykurtic)**: Height of adults (most clustered around average).

Skewness

Purpose: Measures **asymmetry** in data.

Key Uses:

- **Identify Data Bias**:
  - Example: Income data is often **right-skewed** (mean > median). Knowing this helps avoid misinterpreting averages.
  - *Action*: Use median instead of mean for skewed data.
- **Improve Model Accuracy**:
  - Machine learning algorithms (e.g., linear regression) assume symmetric data. Skewed data can bias results.
  - *Action*: Apply log transformations to right-skewed data.
- **Risk Assessment**:
  - In finance, negative skewness indicates higher risk of extreme losses.

Real-World Example:

- **Insurance Claims**: Most claims are small (left-skewed), but a few are huge. Skewness helps price policies correctly.

Kurtosis

Purpose: Measures **tail heaviness** (outlier risk).

Key Uses:

- **Detect Outliers**:
  - High kurtosis (**leptokurtic**) means more outliers (e.g., stock market crashes).
  - Low kurtosis (**platykurtic**) suggests data is tightly clustered (e.g., heights of adults).
- **Quality Control**:
  - Manufacturing processes with high kurtosis may produce defective items sporadically.
- **Financial Modeling**:
  - Portfolio managers check kurtosis to assess crash risk (e.g., Bitcoin returns have high kurtosis).

Real-World Example:

- **Medical Trials**: Drug side-effect data with high kurtosis signals rare but severe reactions.

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:

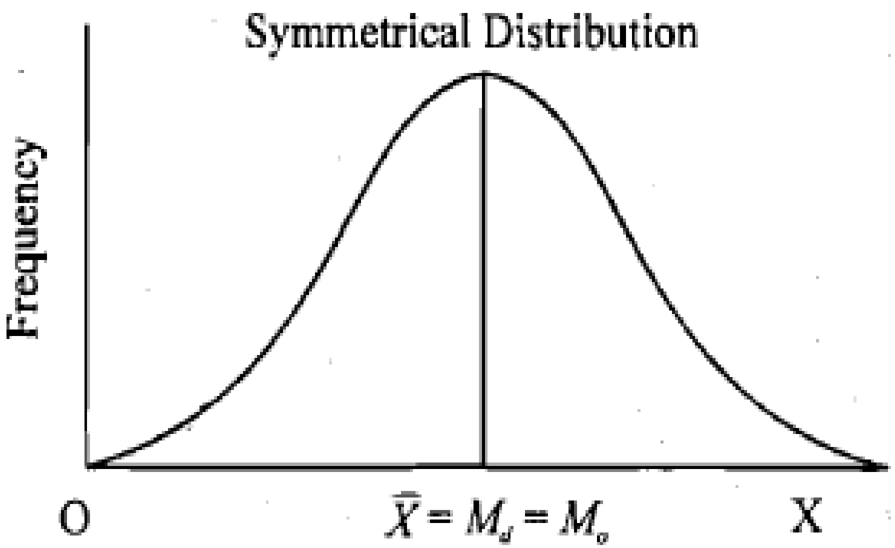
$$Skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

$\overline{X}$ = mean

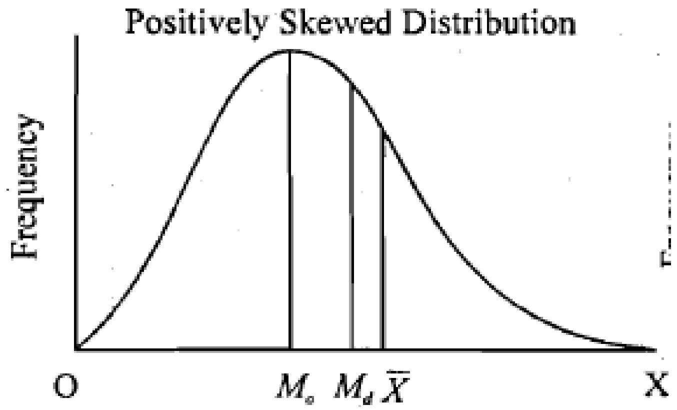
Number of data points

s = standard deviation

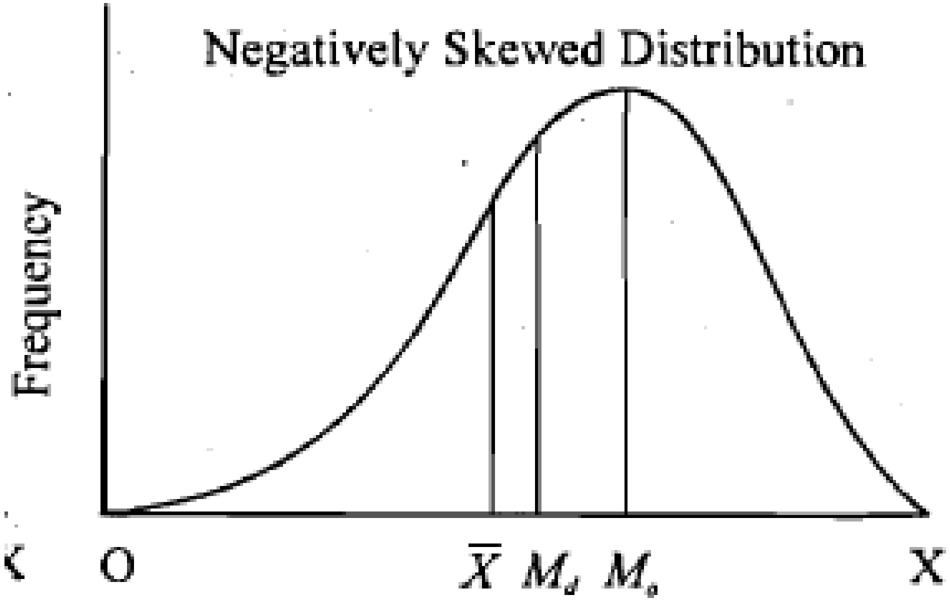
- For **normal distribution (symmetric distribution)**:
  - **skewness = 0**.
- Any symmetric data should have:
  - Skewness value near zero.
  - Distribution with **mean, median and mode fall at the same point**.



- **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**

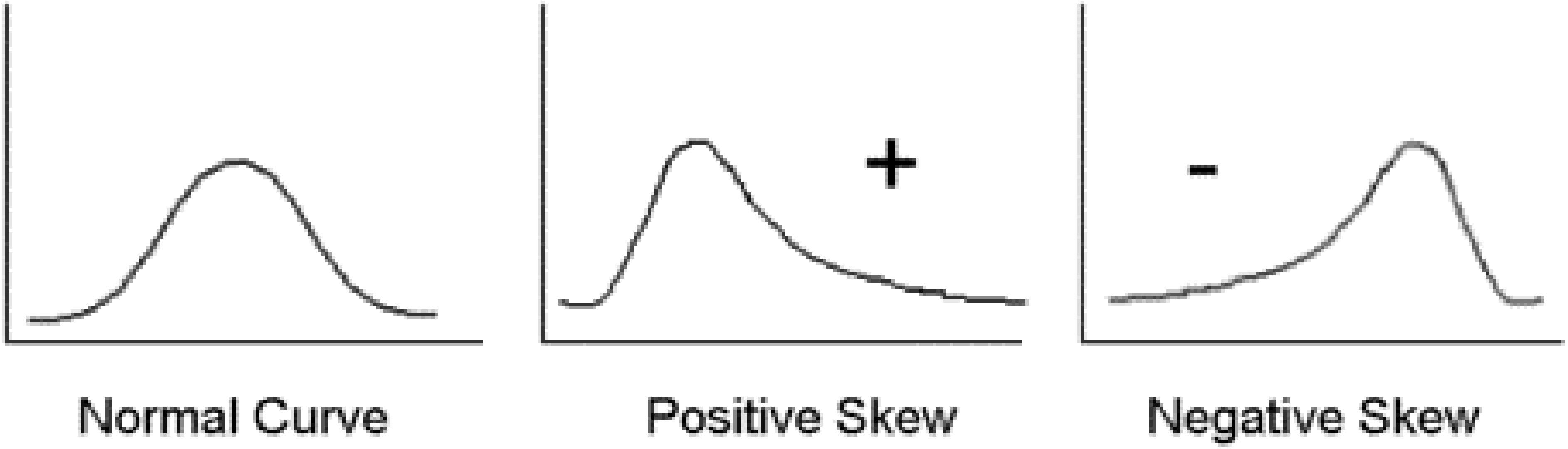


- **Skewness < 0**
  - The distribution is asymmetrical and points in the **negative direction.**
  - Example: Test scores of difficult examination where almost everyone did well on it
  - **mode > median > mean**

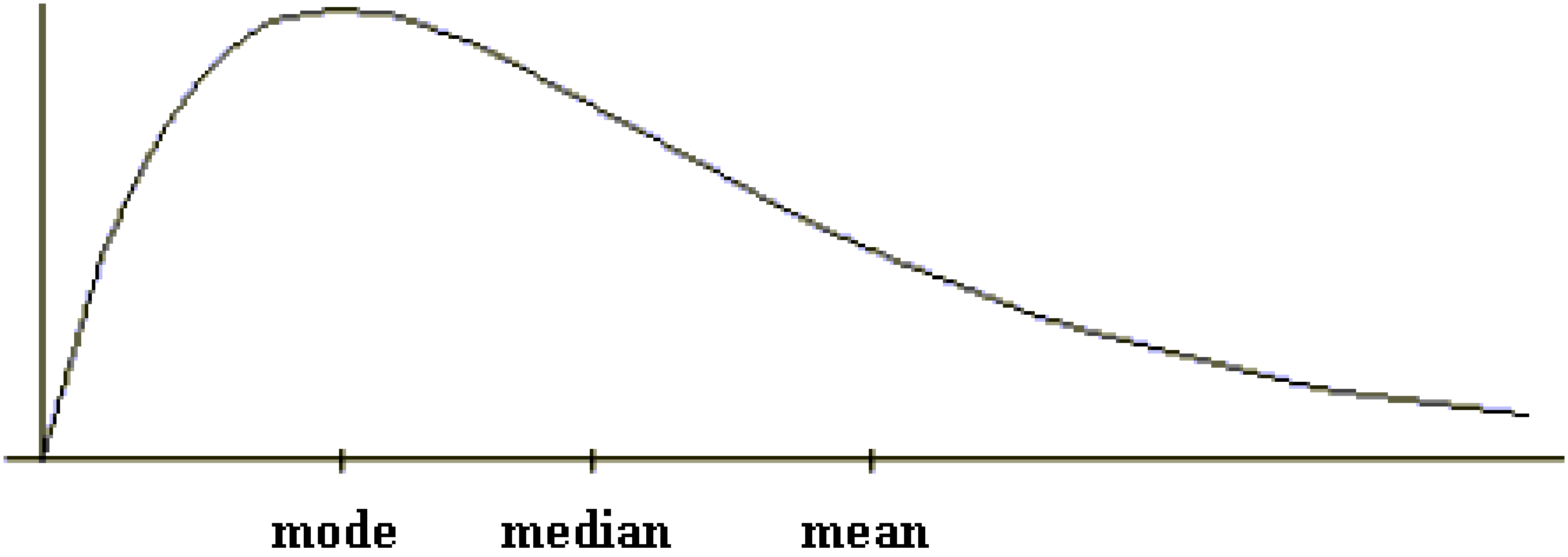


mode < median < mean

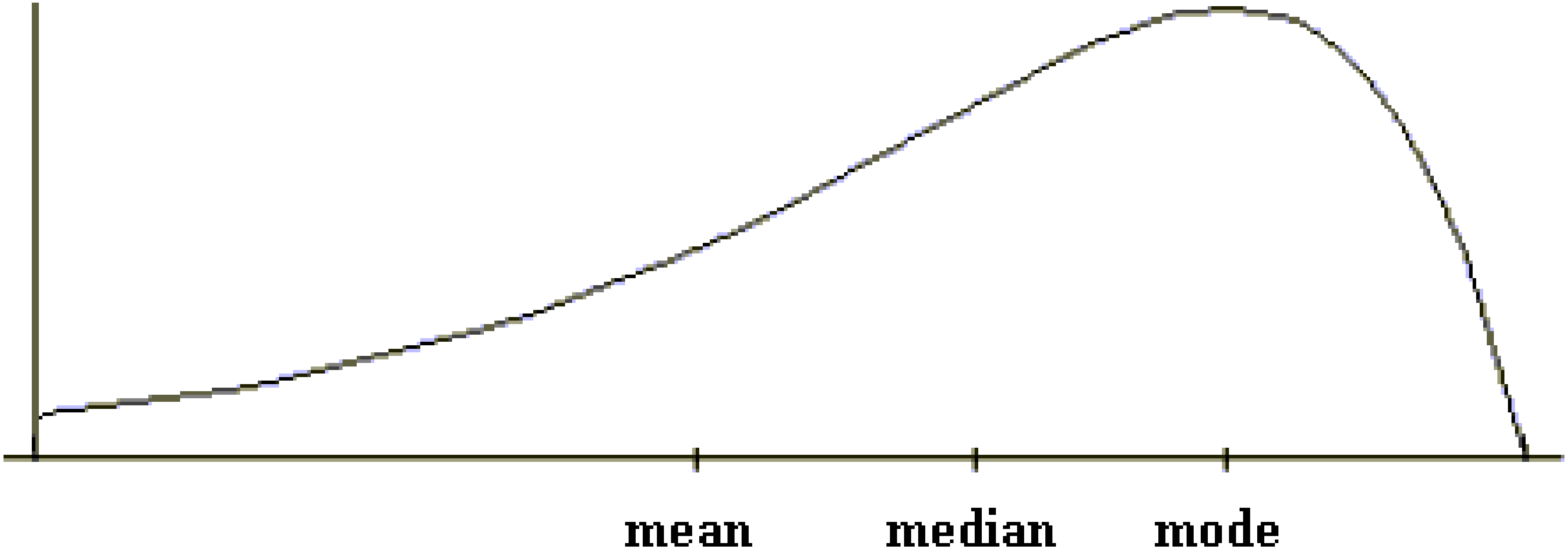
mode > median > mean



Right Skewed



Left Skewed



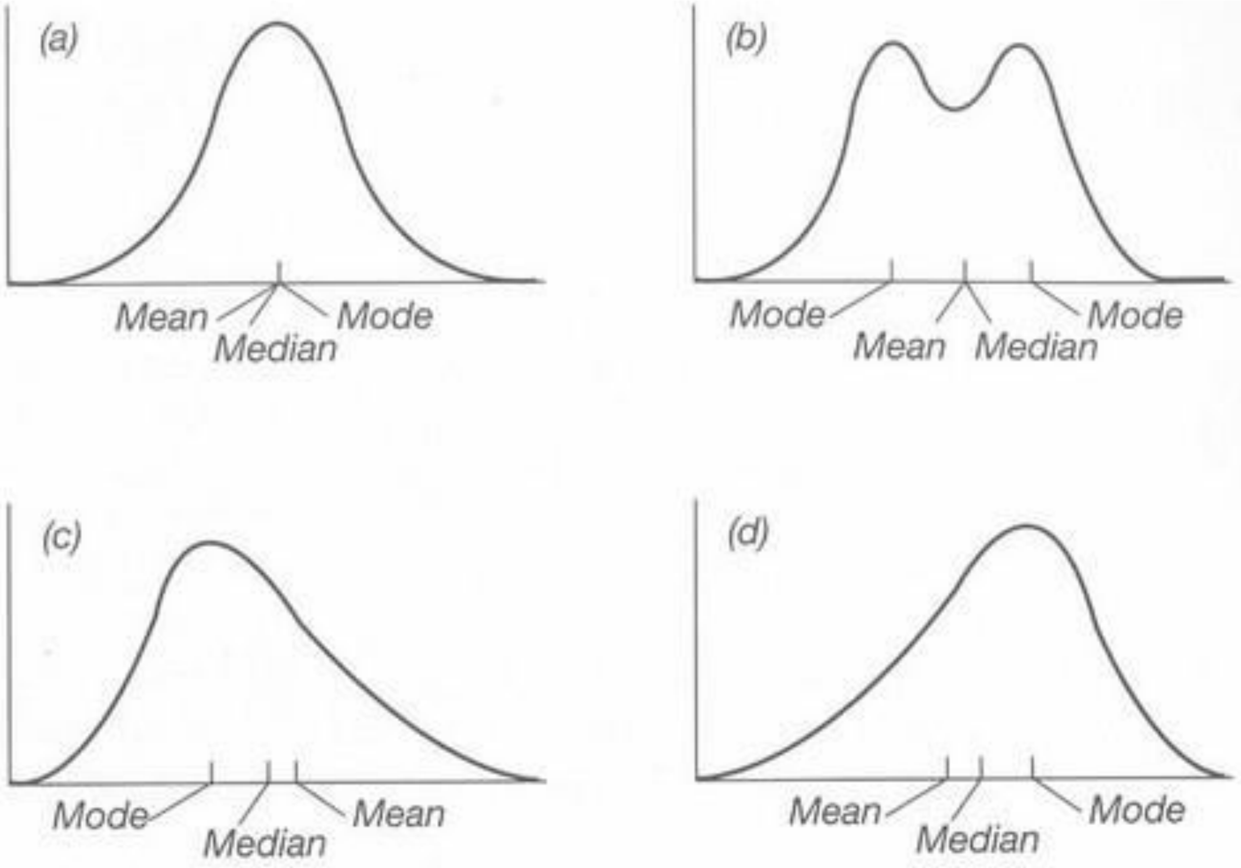


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (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.*

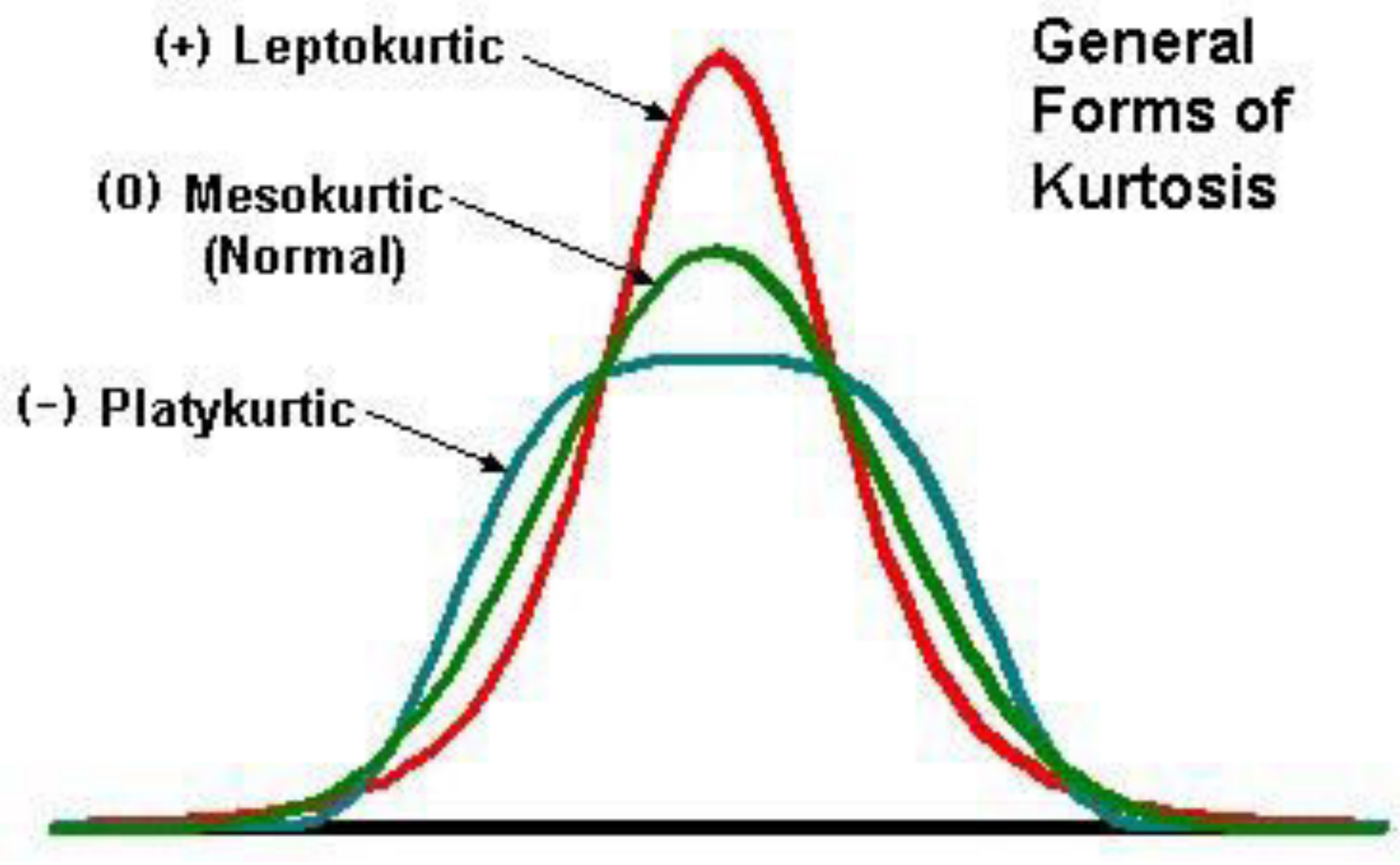
Kurtosis

- **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:

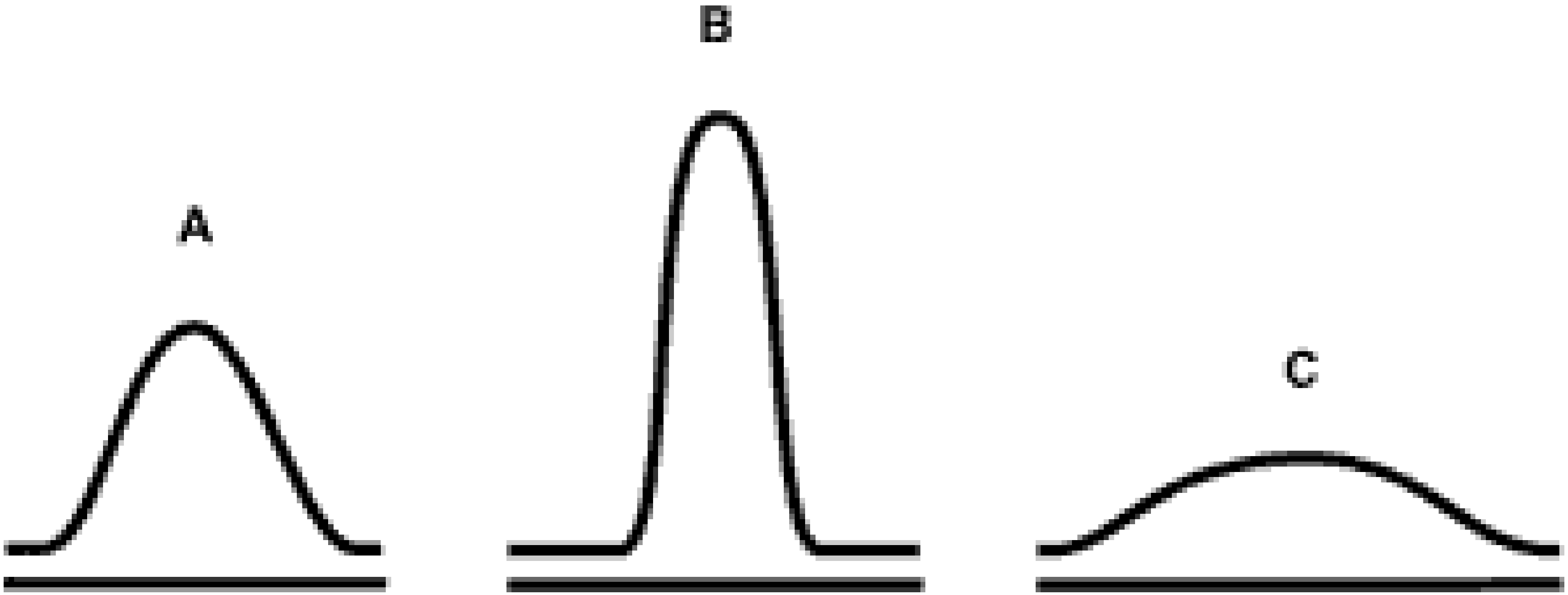
$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

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**.



General Forms of Kurtosis



Mesokurtic
(Normal)
K = 0

Leptokurtic
K > 0

Platykurtic
K < 0