Displaying Grouped Frequency Distributions

For representing frequency (or relative frequency or percentage frequency) distributions, we may use one of the following graphs:

- ❖ Bar Chart
- ❖ The Histogram
- ❖ The Frequency Polygon

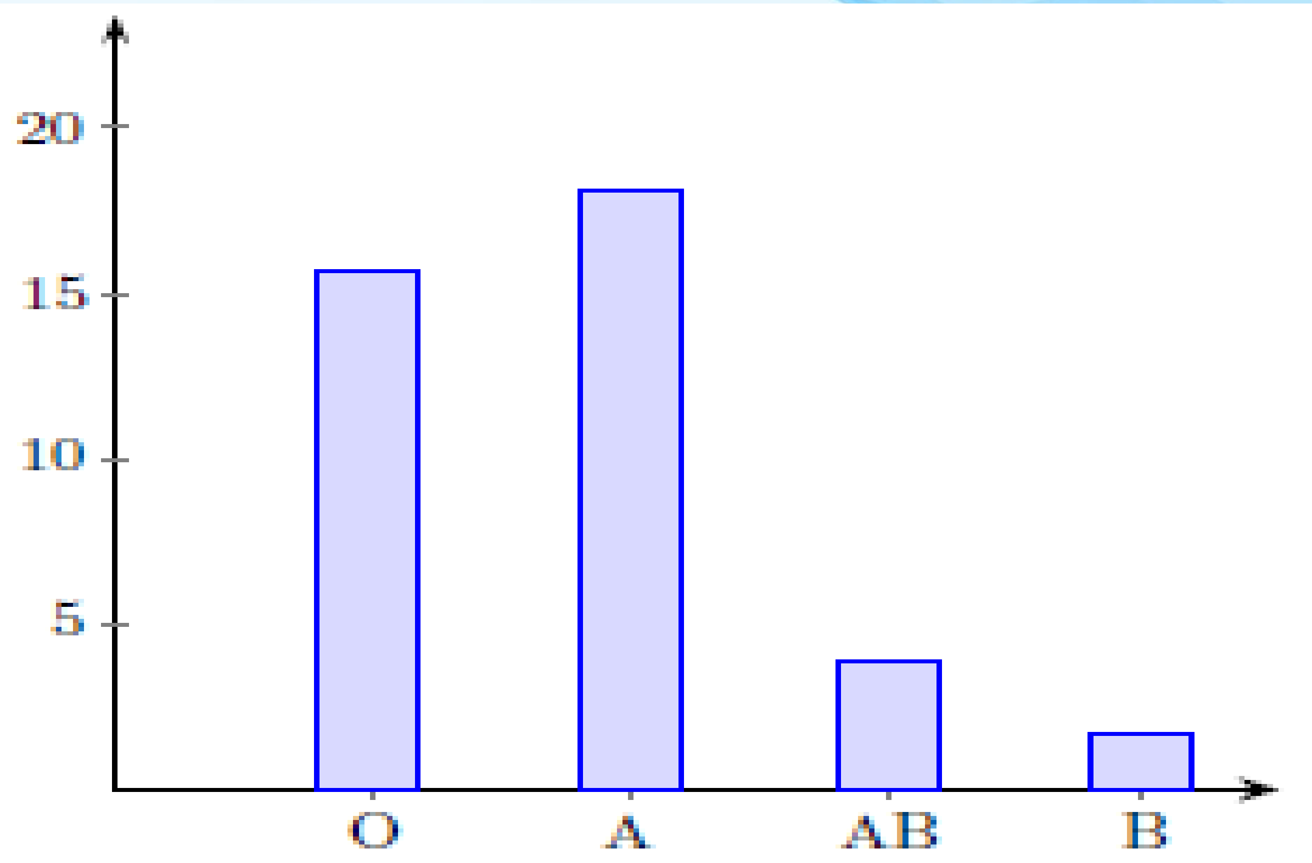
(1) Bar Charts:

In a bar chart, the frequency of each class is represented by a bar. The height of the bar corresponds to the frequency of the class. The width of the bar doesn’t matter.

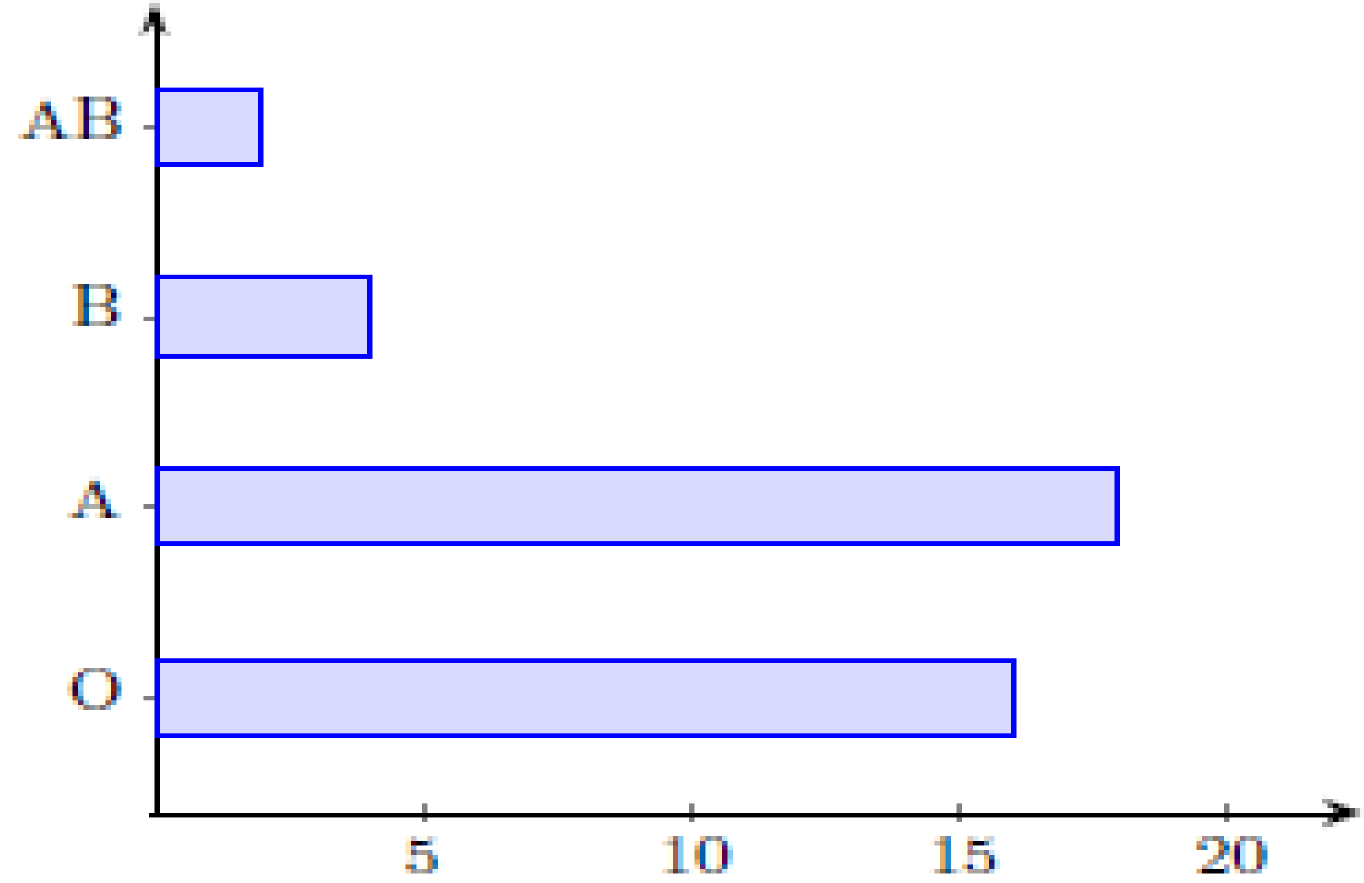
Example: Draw the bar chart for the following data

Blood group	Frequency
O	16
A	18
B	4
AB	2
Total	40

Answer:



Other presentation



EXERCISE

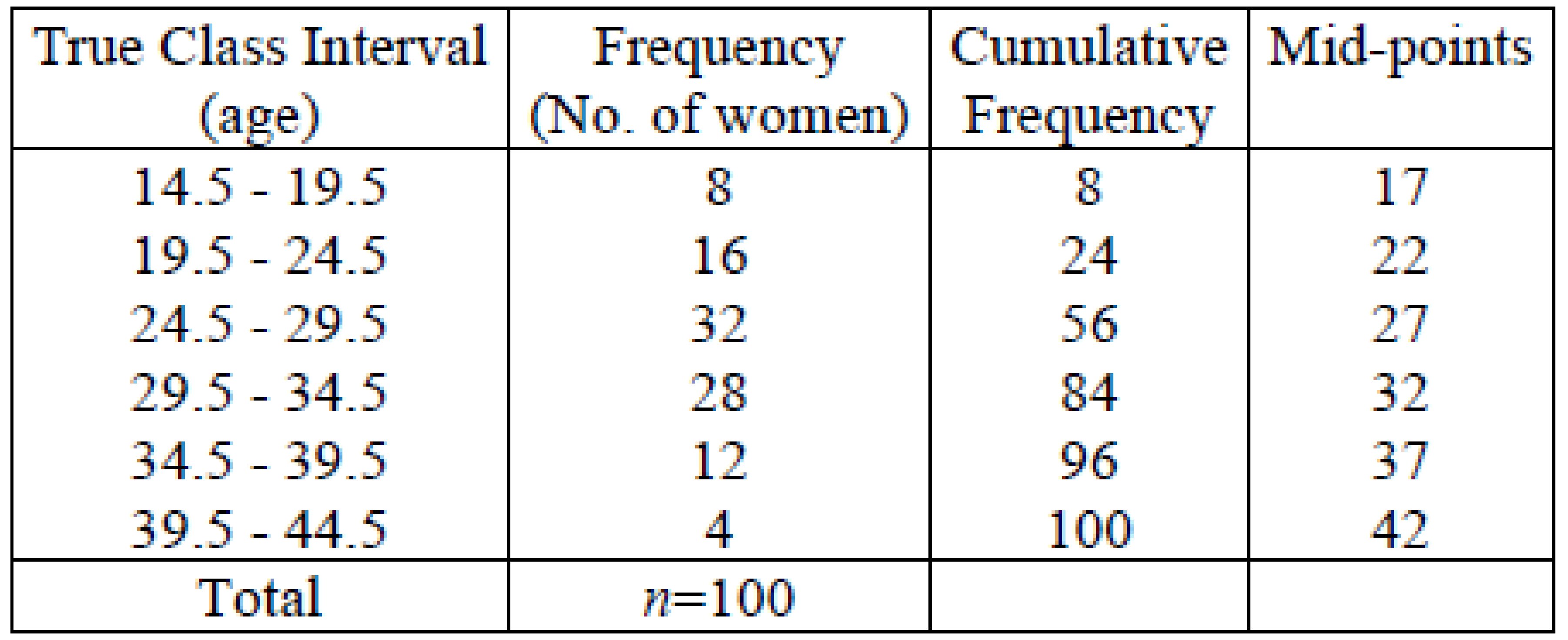
Present the previous data using pie chart?

(2) Histogram:

is similar to bar chart but they both have a basic difference that is in histograms, classes of the variable are adjacent to each other and the rectangular bars must touch each other. Histograms are generally used to represent quantitative data. The class intervals in a histogram are called as bins.

Example:

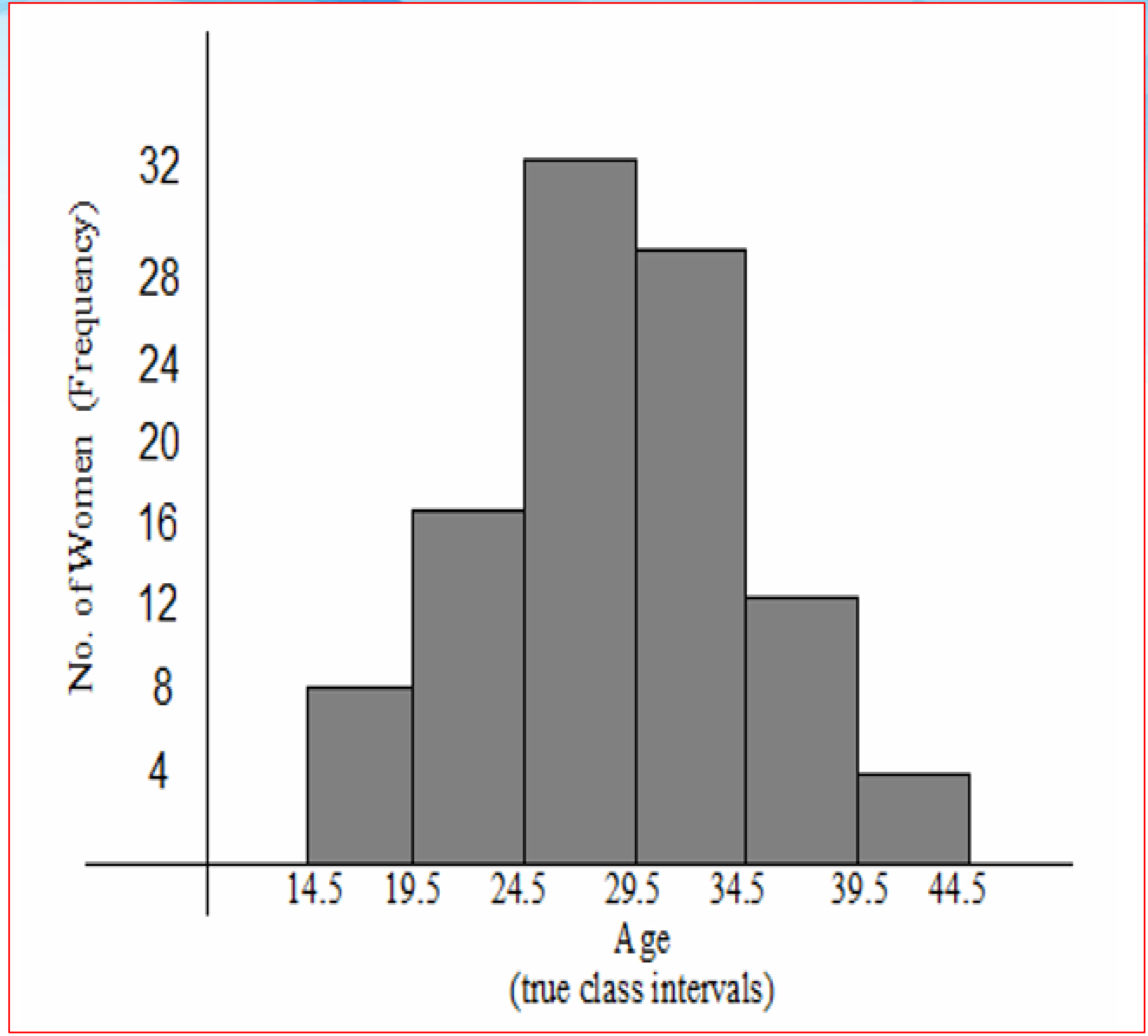
Consider the following frequency distribution of the ages of 100 women.



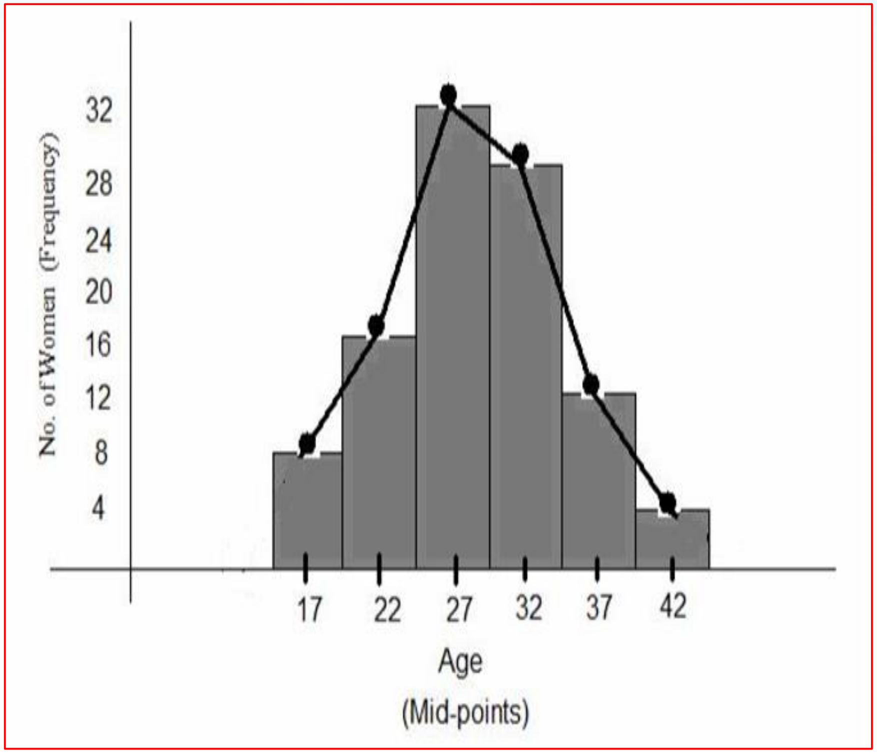
True Class Interval (age)	Frequency (No. of women)	Cumulative Frequency	Mid-points
14.5 - 19.5	8	8	17
19.5 - 24.5	16	24	22
24.5 - 29.5	32	56	27
29.5 - 34.5	28	84	32
34.5 - 39.5	12	96	37
39.5 - 44.5	4	100	42
Total	n=100		

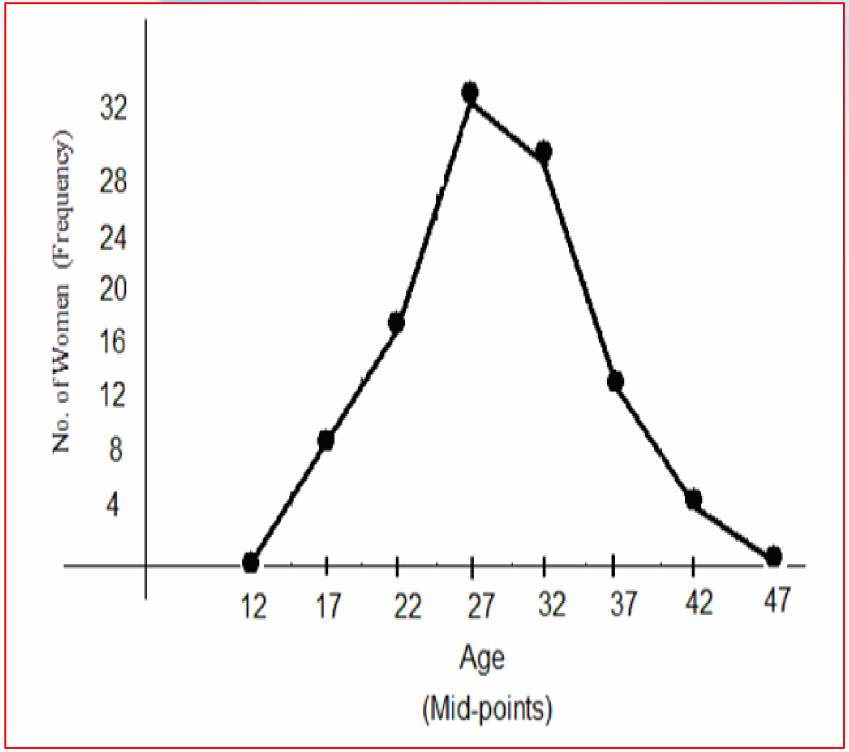
Width of the interval:

W =true upper limit – true lower limit = 19.5 – 14.5 = 5

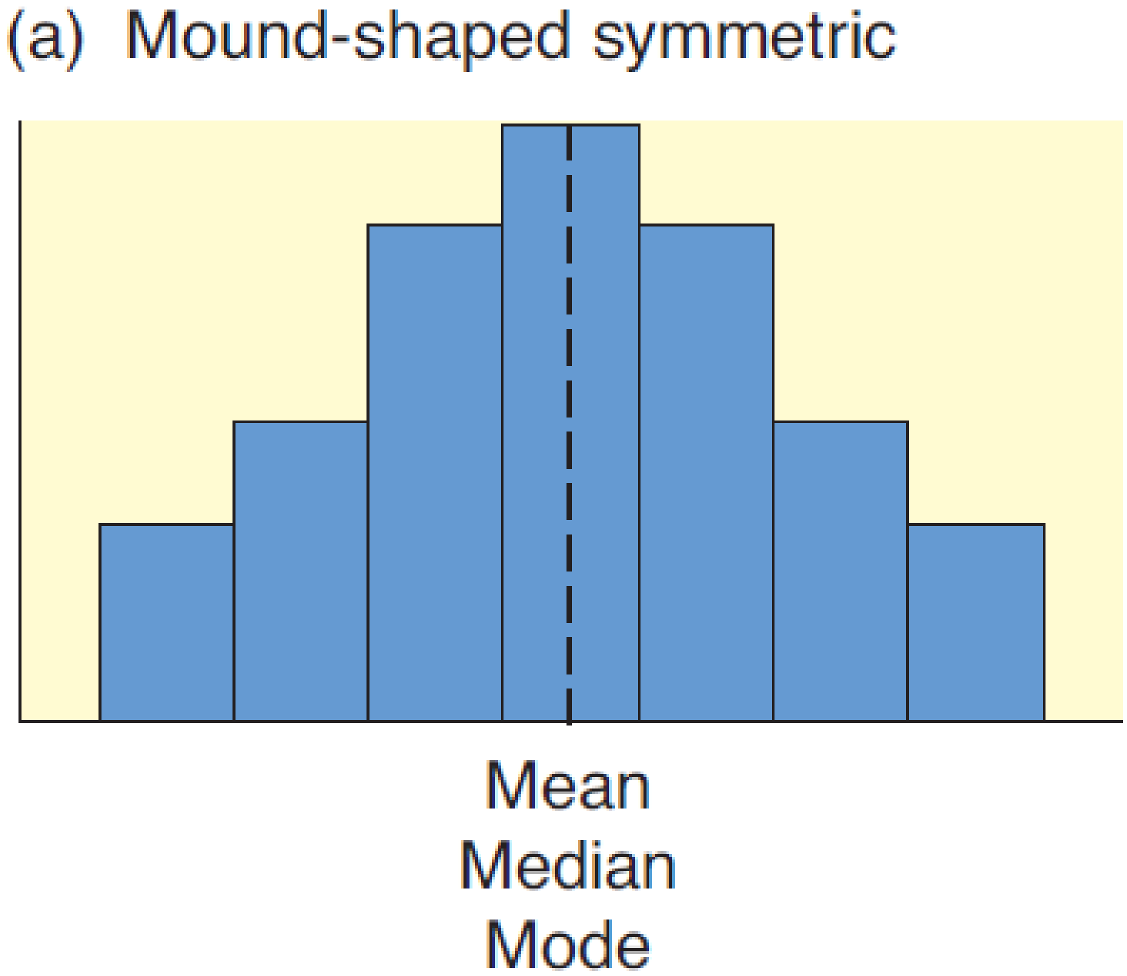

No. of Women (Frequency)
32
28
24
20
16
12
8
4
14.5
19.5
24.5
29.5
34.5
39.5
44.5
Age
(true class intervals)

The Frequency Polygon

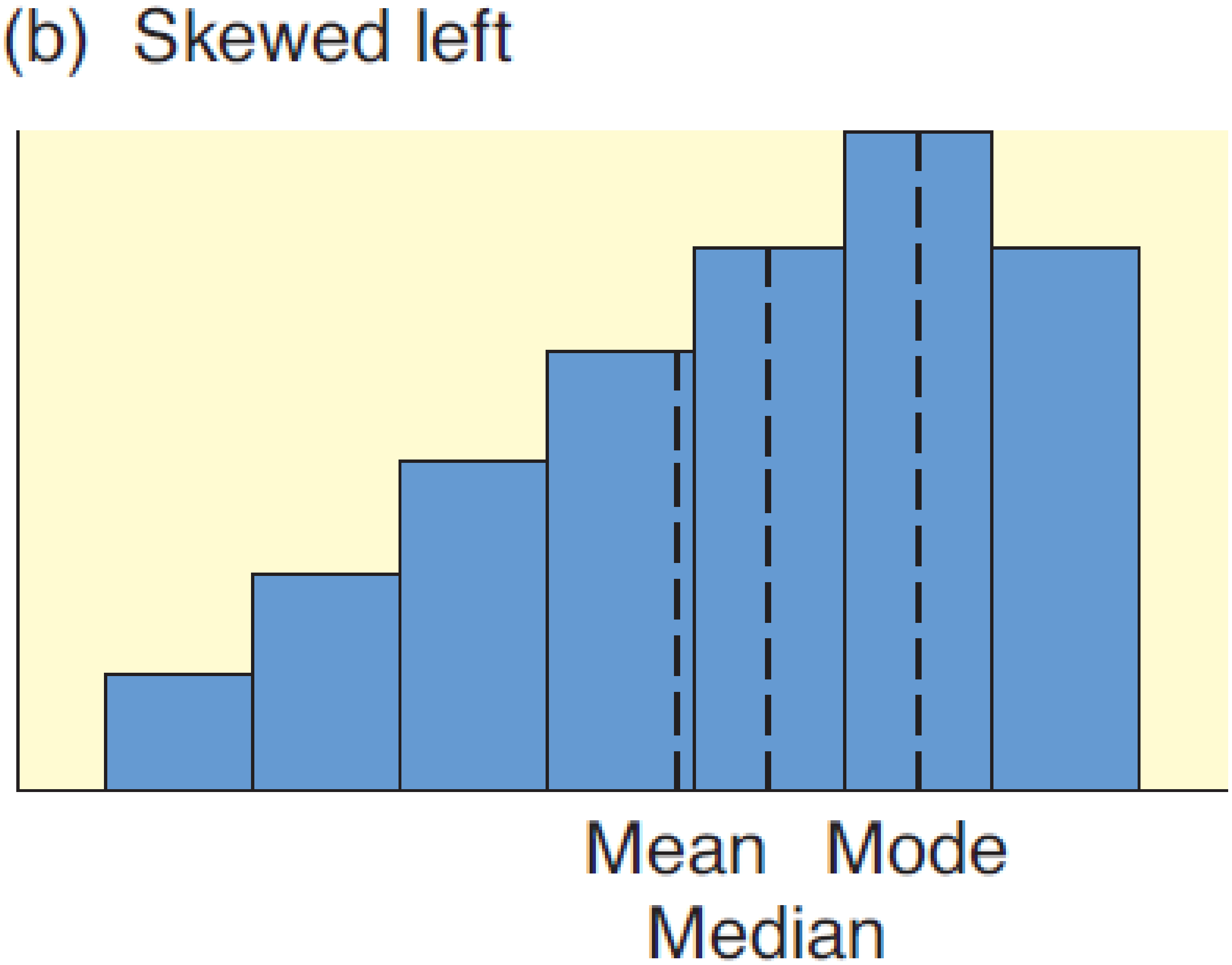


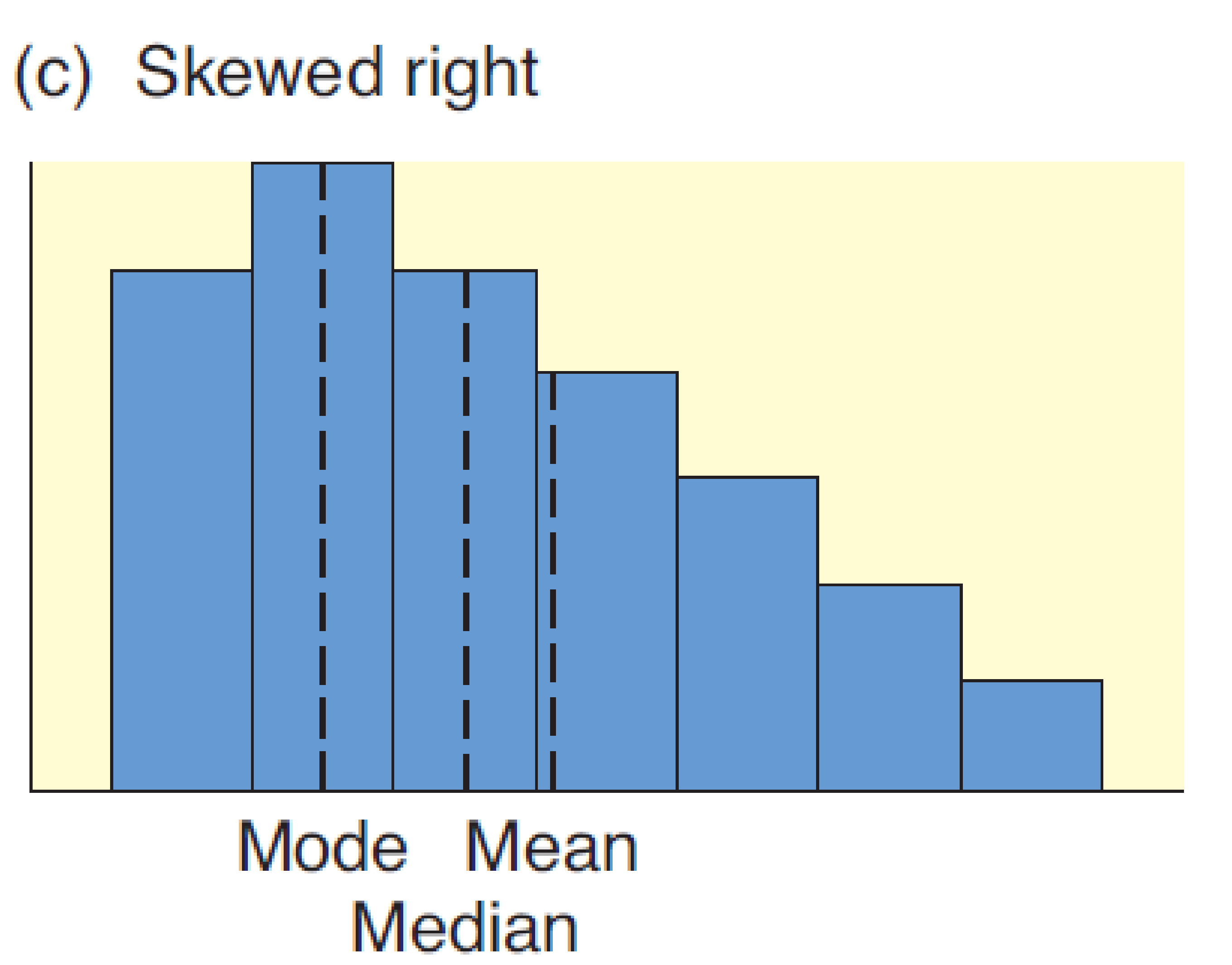


Distribution Types and Averages



(a) Mound-shaped symmetric


(b) Skewed left
Mean
Mode
Median


(c) Skewed right
Mode
Mean
Median

Measures of Location (Central Tendency)

In the last section we summarize the data using frequency distributions (tables and figures). In this section, we will introduce the concept of summarization of the data by means of a single number called "a descriptive measure".

Measures of Location (Central Tendency)

- The data (observations) often tend to be concentrated around the center of the data.
- Some measures of location are: the mean, mode, and median.
- These measures are considered as representatives (or typical values) of the data. They are designed to give some quantitative measures of where the center of the data is in the sample.

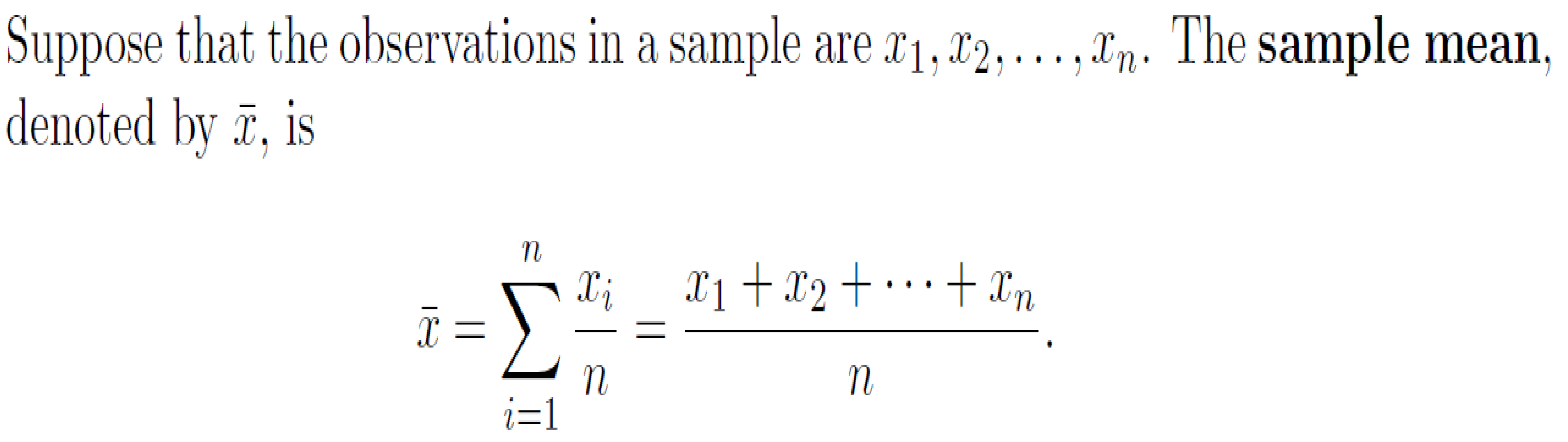
The most commonly used measures of central tendency are: the mean – the median – the mode.

1. Mean

The population mean (μ)	$\mu = \dfrac{X_1 + X_2 + \cdots + X_N}{N} = \dfrac{\sum_{i=1}^{N} X_i}{N}$ (unit)
The sample mean ($\bar{x}$)	$\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ (unit)

The sample mean $\bar{x}$ is a statistic and used to approximate (estimate) the population mean μ.

The Sample mean of the observations



Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Example:

Suppose that the following sample represents the ages (in year) of a sample of 3 men:

$$x_1 = 30, x_2 = 35, x_3 = 27$$

Then, the sample mean is:

$$\bar{x} = \frac{30+35+27}{3} = 30.67 \text{ (years)}$$

Note:

$$\sum_{i=1}^{n}(x_i-\bar{x})=0\,.$$

Prove that?

Advantages and disadvantages of the mean:

Advantages:

- Simplicity: The mean is easily understood and easy to compute.
- Uniqueness: There is one and only one mean for a given set of data.
- The mean takes into account all values of the data.

Disadvantages:

- Extreme values have an influence on the mean. Therefore, the mean may be distorted by extreme values.

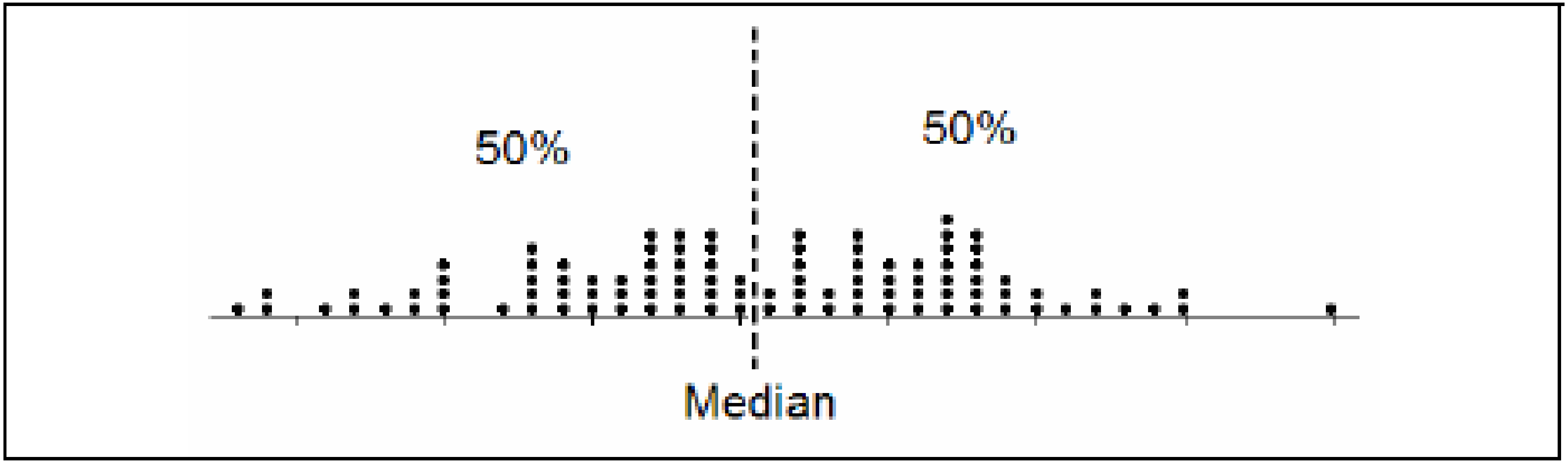
For example:

Sample	Data	mean
A	2 4 5 7 7 10	5.83
B	2 4 5 7 7 100	20.83

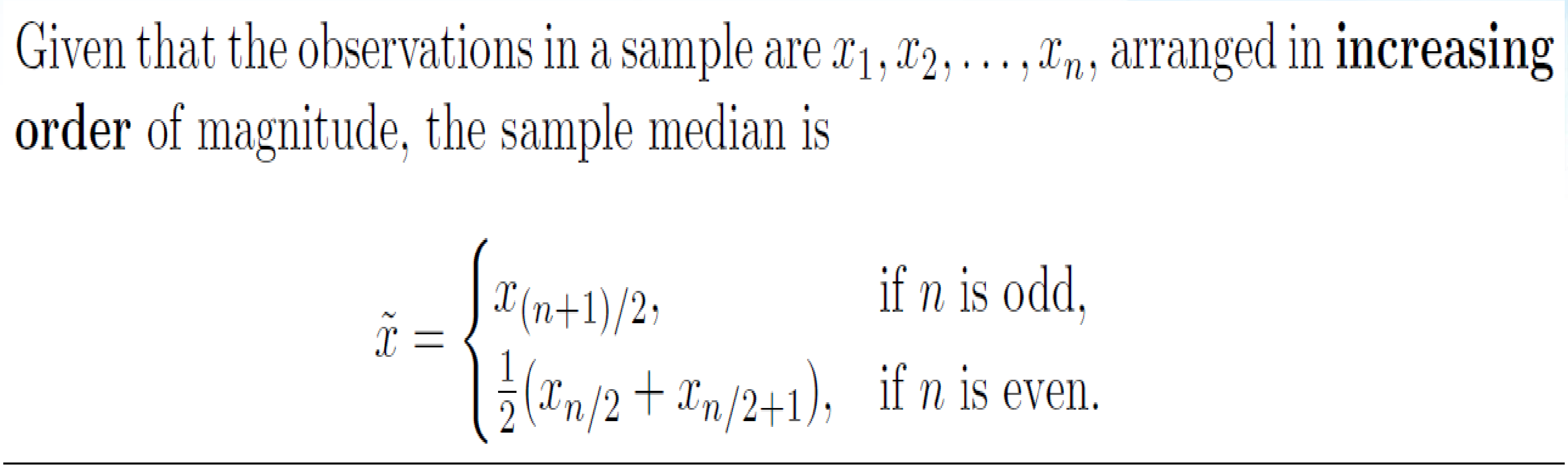
- The mean can only be found for quantitative variables.

Median

The median of a finite set of numbers is that value which divides the ordered array into two equal parts. The numbers in the first part are less than or equal to the median and the numbers in the second part are greater than or equal to the median



Median



Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd}, \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{if } n \text{ is even}. \end{cases}$$

Example:

suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

Ordered array: 1.7, 2.2, 3.11, 3.9, 14.7 **n=5 is odd**

$\bar{x} = 5.12$ $\tilde{x} = 3.11$

Note:

The mean is influenced by the extreme observations, whereas the median places emphasis on the true “center” of the data set.

See previous Example

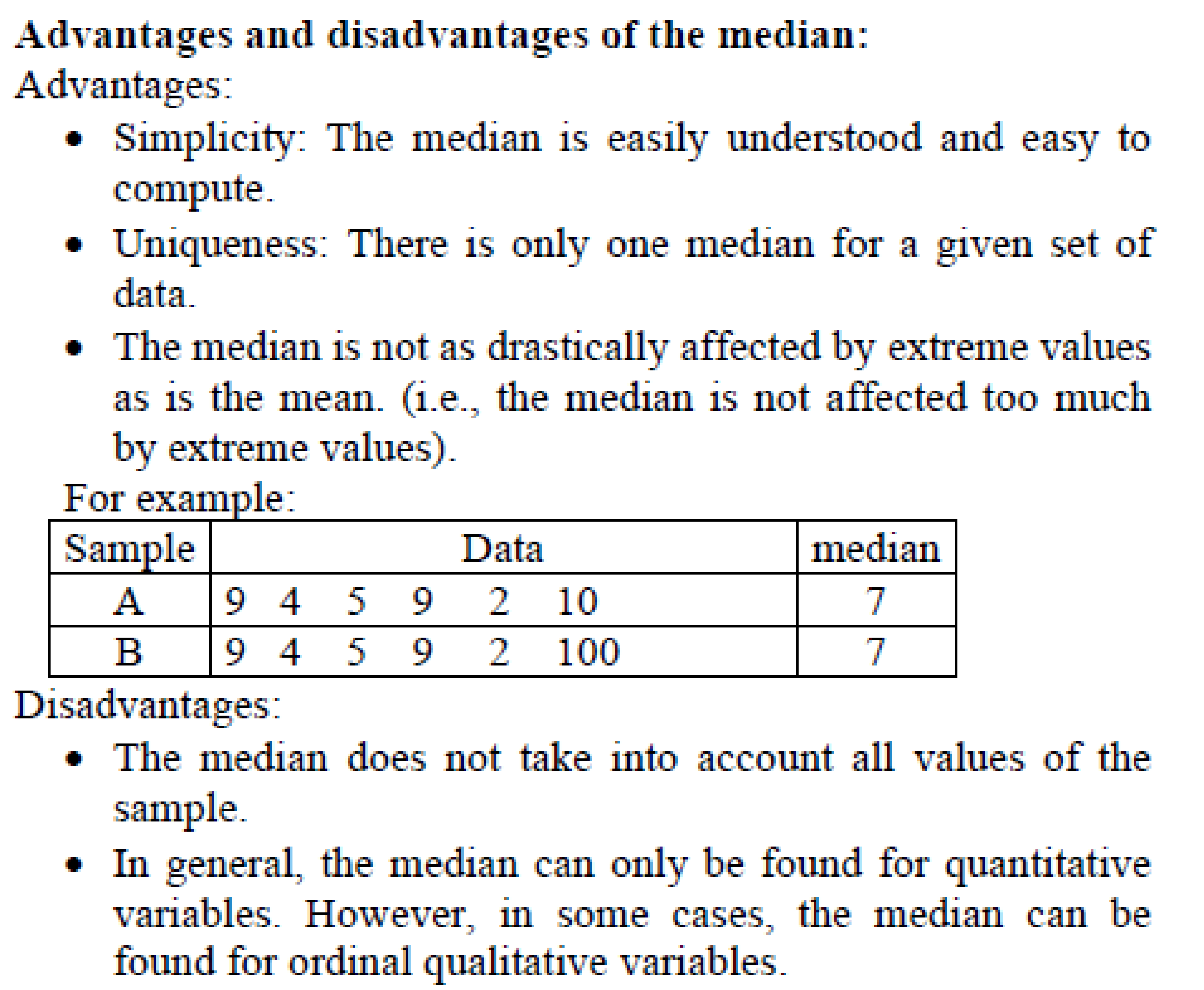
Example (even number):

Find the median for the sample values: 10, 35, 41, 16, 20, 32

Solution:

Ordered array: 10, 16, 20, 32, 35, 41   n=6 is even

$$\textbf{The median} = \frac{20+32}{2} = 26$$



Advantages and disadvantages of the median:

Advantages:

- Simplicity: The median is easily understood and easy to compute.
- Uniqueness: There is only one median for a given set of data.
- The median is not as drastically affected by extreme values as is the mean. (i.e., the median is not affected too much by extreme values).

For example:

Sample	Data	median
A	9 4 5 9 2 10	7
B	9 4 5 9 2 100	7

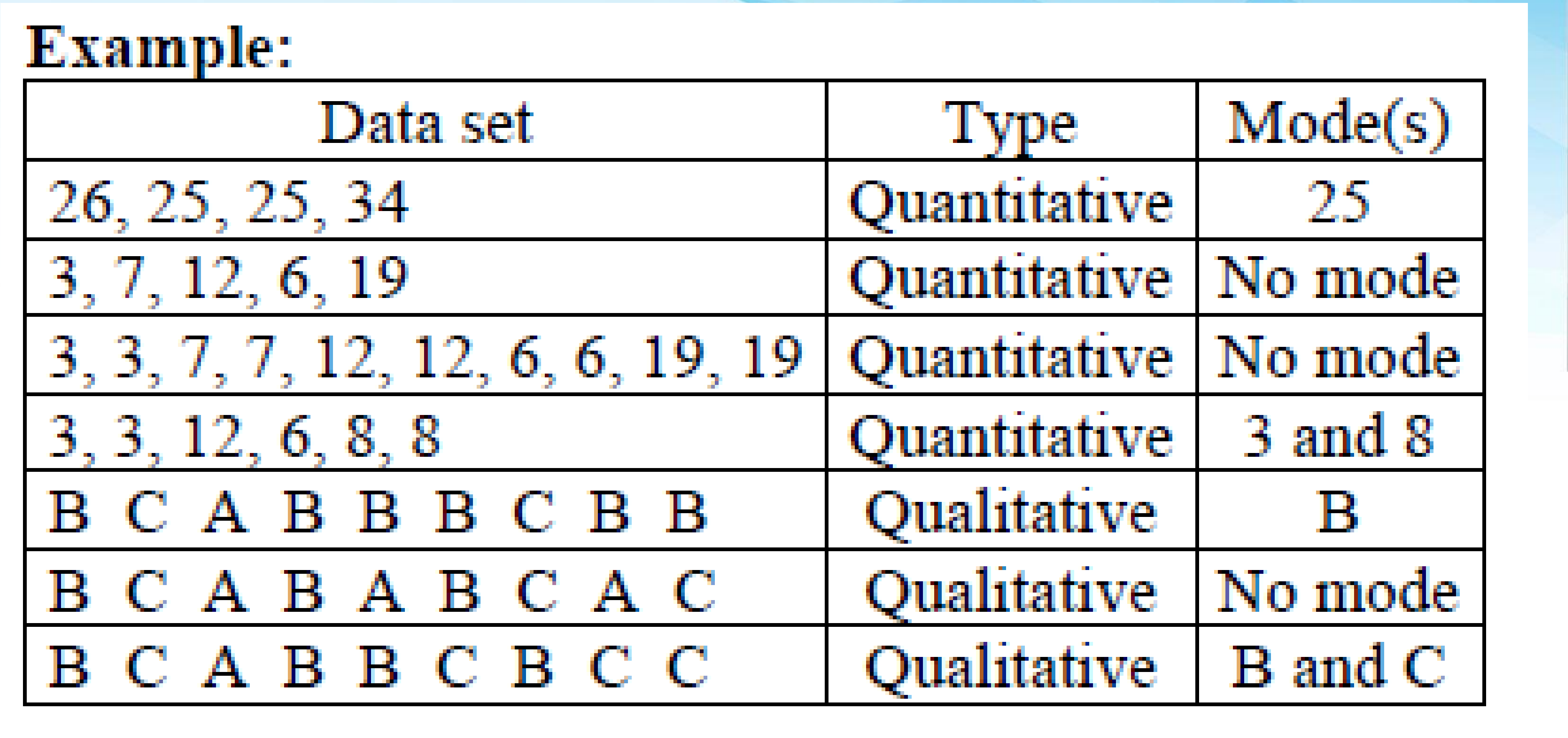
Disadvantages:

- The median does not take into account all values of the sample.
- In general, the median can only be found for quantitative variables. However, in some cases, the median can be found for ordinal qualitative variables.

Mode:

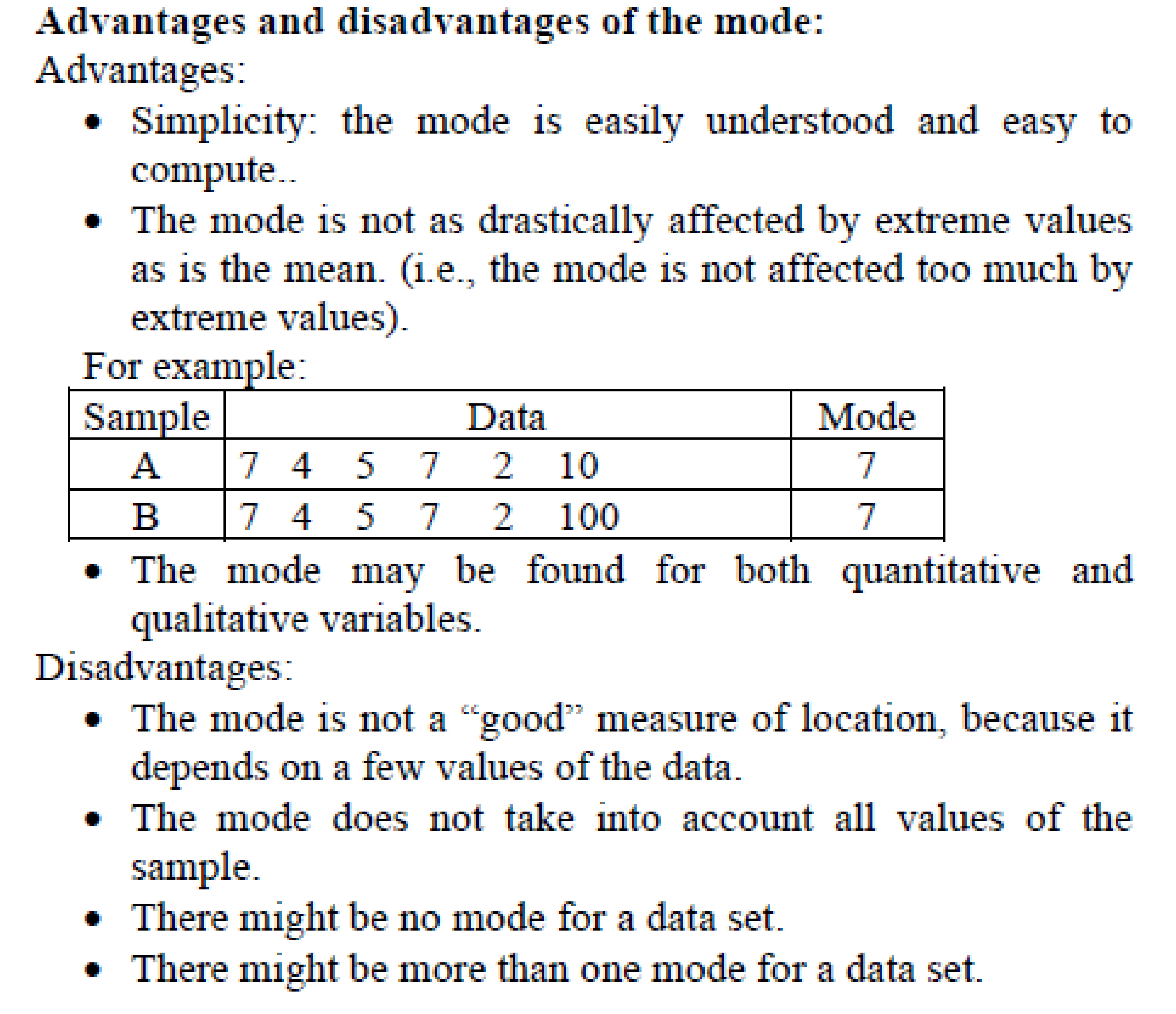
The mode of a set of values is that value which occurs most frequently. (i.e., with the highest frequency).

- If all values are different or have the same frequencies, there will be no mode.
- A set of data may have more than one mode.



Example:

Data set	Type	Mode(s)
26, 25, 25, 34	Quantitative	25
3, 7, 12, 6, 19	Quantitative	No mode
3, 3, 7, 7, 12, 12, 6, 6, 19, 19	Quantitative	No mode
3, 3, 12, 6, 8, 8	Quantitative	3 and 8
B C A B B B C B B	Qualitative	B
B C A B A B C A C	Qualitative	No mode
B C A B B C B C C	Qualitative	B and C



Advantages and disadvantages of the mode:

Advantages:

- Simplicity: the mode is easily understood and easy to compute..
- The mode is not as drastically affected by extreme values as is the mean. (i.e., the mode is not affected too much by extreme values).

For example:

Sample	Data	Mode
A	7 4 5 7 2 10	7
B	7 4 5 7 2 100	7

- The mode may be found for both quantitative and qualitative variables.

Disadvantages:

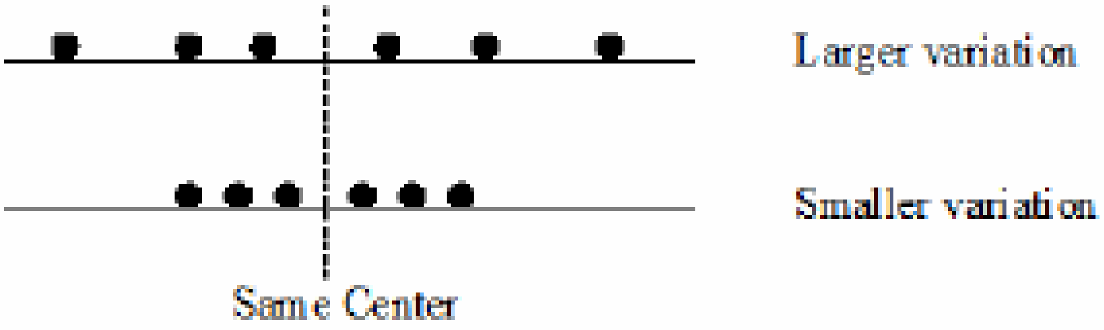
- The mode is not a "good" measure of location, because it depends on a few values of the data.
- The mode does not take into account all values of the sample.
- There might be no mode for a data set.
- There might be more than one mode for a data set.

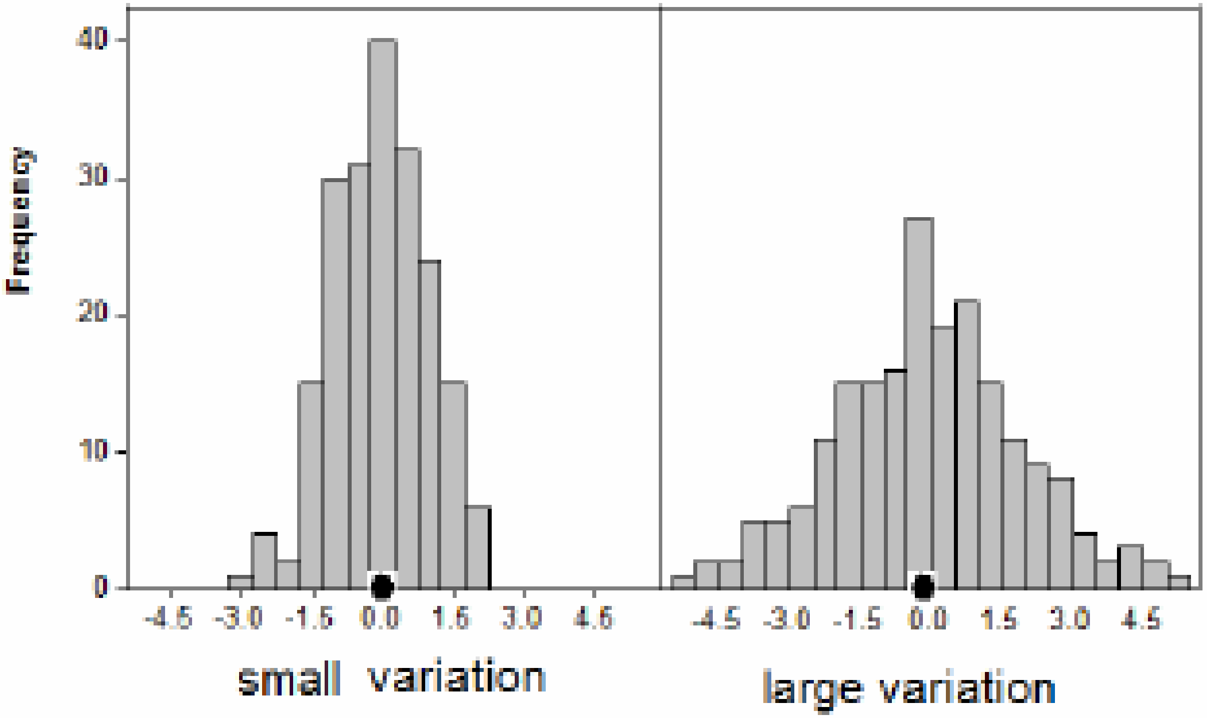
Example: suppose we have the following two data sets:

data 1: 59,60,60,61

data 2: 50,60,60,70

Calculate the Mean, Median and the Mode?
What do you see?


Larger variation
Smaller variation
Same Center


Frequency
40
30
20
10
0
-4.5
-3.0
-1.5
0.0
1.5
3.0
4.5
small variation
-4.5
-3.0
-1.5
0.0
1.5
3.0
4.5
large variation

Measures of Variability (Dispersion or Variation)

- The variation or dispersion in a set of data refers to how spread out the observations are from each other.
- The variation is small when the observations are close together. There is no variation if the observations are the same.
- Some measures of dispersion are range, variance, and standard deviation
- These measures are designed to give some quantitative measures of the variability in the data.

Range:

It is the simplest measure of variation and defined as

$$R = X_{max} - X_{\min}$$

Example:

What is the range of the following data set

42,55,47,41,57,50

Solution:

$$X_{\min} = 41$$

$$X_{\max} = 57$$

$$R = X_{\max} - X_{\min} = 57 - 41 = 16$$

The Sample Variance (S^2)

Let $x_1, x_2, \ldots, x_n$ be the observations of the sample. The sample variance is denoted by S^2 and is defined by:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} \quad (\text{unit})^2$$

where $\bar{x} = \sum_{i=1}^{n} x_i / n$ is the sample mean.

Note:

(n −1) is called the degrees of freedom (df) associated with the sample variance (S^2).

The Standard Deviation (S)

The standard deviation is another measure of variation. It is the square root of the variance

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{(unit)}$$

Example:

Compute the sample variance and standard deviation of the following observations (ages in year): 10, 21, 33, 53, 54.

Solution:

$n=5$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10+21+33+53+54}{5} = \frac{171}{5} = 34.2 \quad (\text{year})$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{5}(x_i - 34.2)^2}{5-1}$$

$$= \frac{(10-34.2)^2 + (21-34.2)^2 + (33-34.2)^2 + (53-34.2)^2 + (54-34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad (\text{year})^2$$

The sample standard deviation is

$$S = \sqrt{S^2} = \sqrt{376.7} = 19.41 \qquad \text{(year)}$$

Another Formula for Calculating S^2:

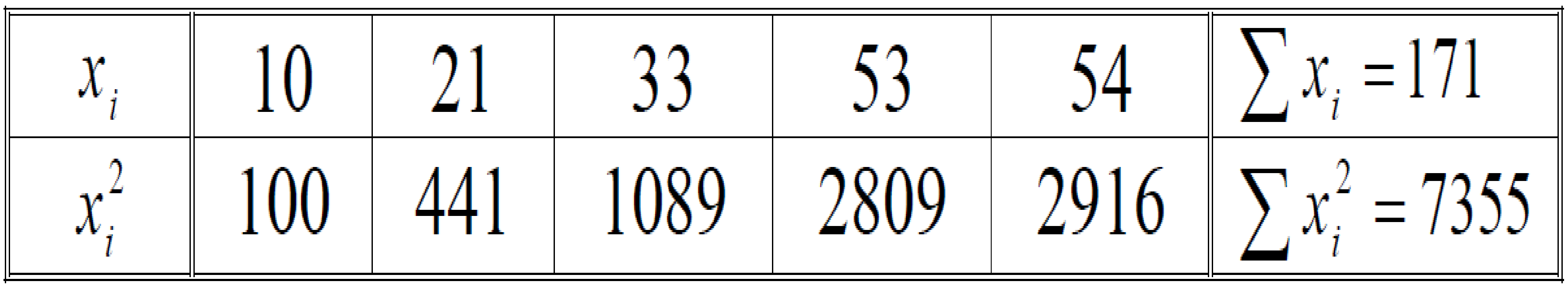
$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Note:

To calculate S^2 we need:

- n = sample size
- $\sum x_i$ =The sum of the values
- $\sum {x^2}_i$=The sum of the squared values

For the above example:



x_i	10	21	33	53	54	$\sum x_i = 171$
x_i^2	100	441	1089	2809	2916	$\sum x_i^2 = 7355$

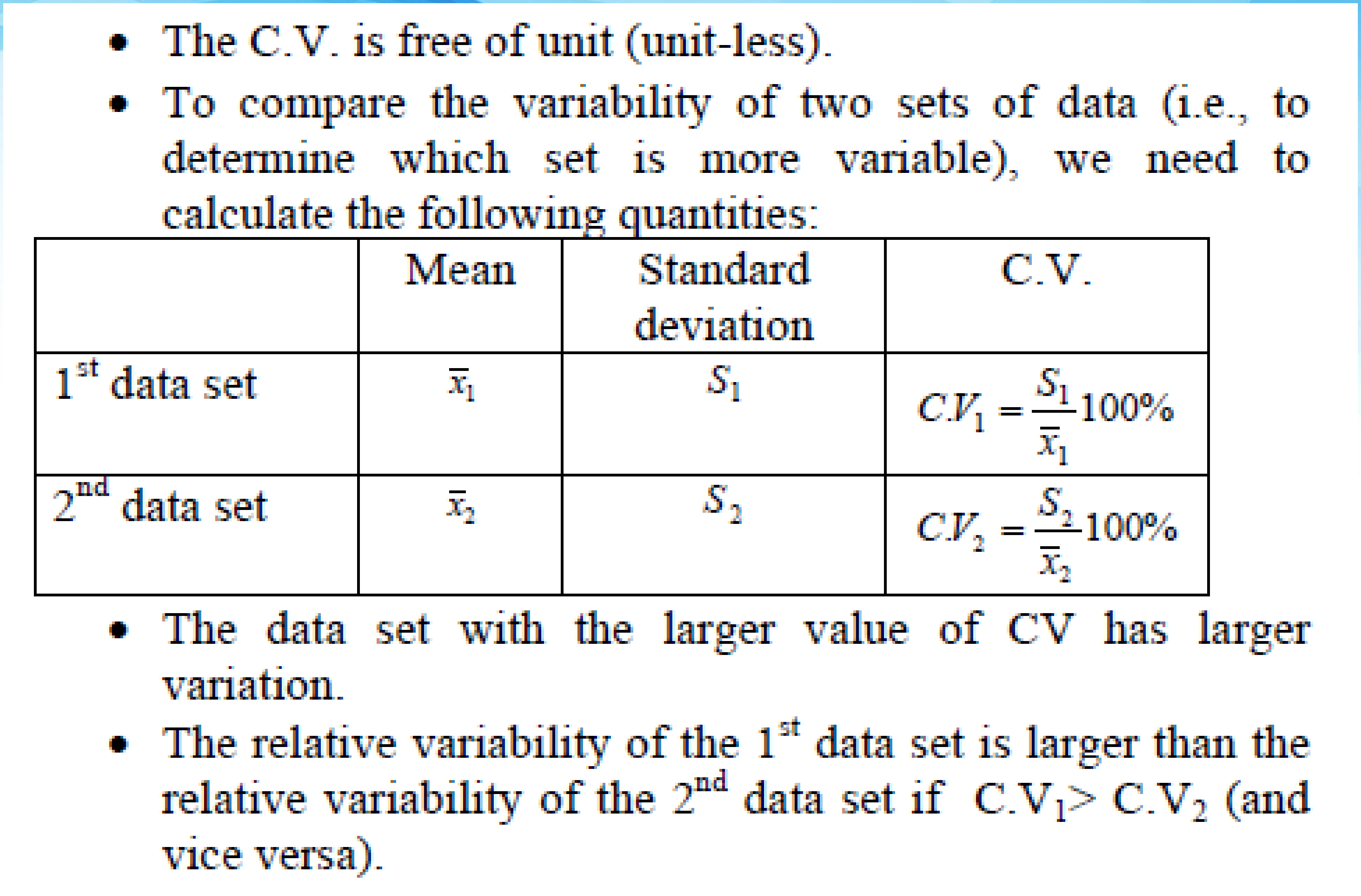
$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{7355-(5)(34.2)^2}{5-1} = \frac{1506.8}{4} = 376.7\ (\text{unit})^2$$

Coefficient of Variation (C.V.):

- The variance and the standard deviation are useful as measures of variation of the values of a single variable for a single population.
- If we want to compare the variation of two variables we cannot use the variance or the standard deviation because:

1. The variables might have different units.
2. The variables might have different means.

$$C.V. = \frac{S}{\bar{x}} \times 100$$



- The C.V. is free of unit (unit-less).
- To compare the variability of two sets of data (i.e., to determine which set is more variable), we need to calculate the following quantities:

	Mean	Standard deviation	C.V.
1st data set	$\bar{x}_1$	S_1	$C.V_1 = \frac{S_1}{\bar{x}_1} 100\%$
2nd data set	$\bar{x}_2$	S_2	$C.V_2 = \frac{S_2}{\bar{x}_2} 100\%$

- The data set with the larger value of CV has larger variation.
- The relative variability of the 1st data set is larger than the relative variability of the 2nd data set if $C.V_1 > C.V_2$ (and vice versa).

Example:

Suppose we have two data sets:

1st data set: $\bar{x}_1 = 66$ kg, $S_1 = 4.5$ kg

$$\Rightarrow CV_1 = \frac{4.5}{66} * 100\% = 6.8\%$$

2nd data set: $\bar{x}_2 = 36$ kg, $S_2 = 4.5$ kg

$$\Rightarrow CV_2 = \frac{4.5}{36} * 100\% = 12.5\%$$

Since $CV_2 > CV_1$, the relative variability of the 2nd data set is larger than the relative variability of the 1st data set.

If we use the standard deviation to compare the variability of the two data sets, we will wrongly conclude that the two data sets have the same variability because the standard deviation of both sets is 4.5 kg.