MAT 271E: PROBABILITY AND STATISTICS

PROF. DR. CANAN SARICAM

WEEK 3

DATA SUMMARY MEASURES – DESCRIPTIVE STATISTICS

DATA SUMMARY MEASURES-DESCRIPTIVE STATISTICS

Data summary measures – descriptive statistics are used for quantitatively describing the main features of a collection of information.

Summary measures fall into three broad categories:

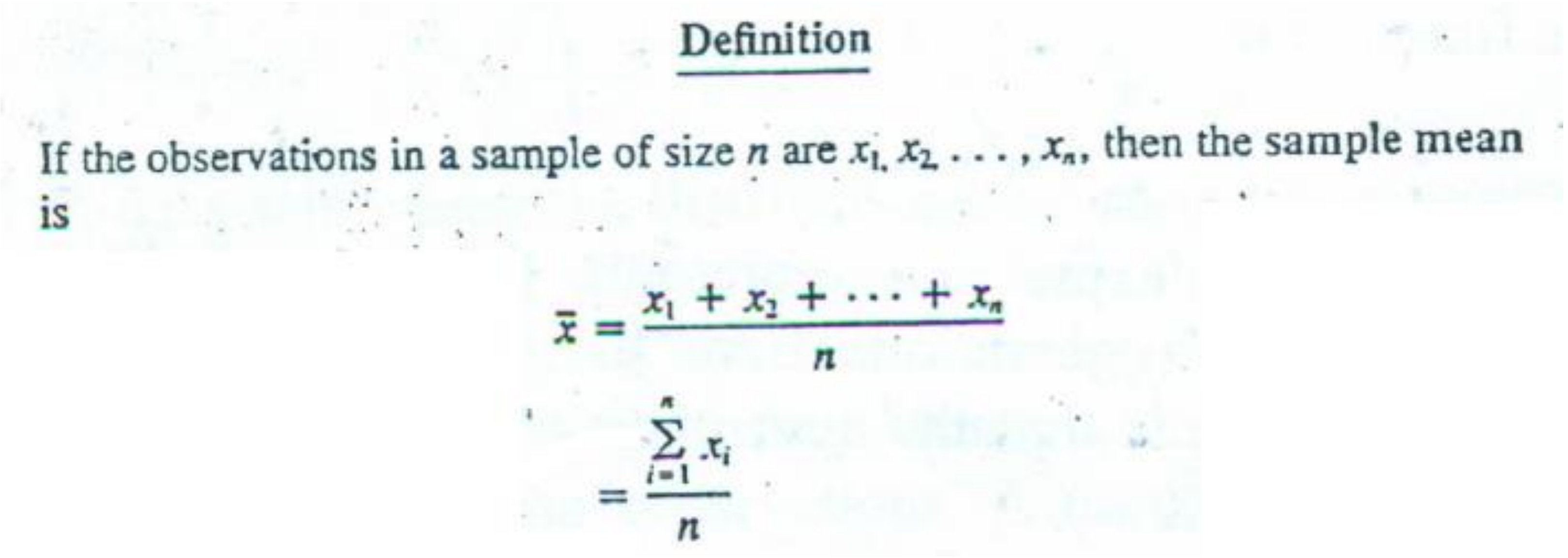
1. Measures of position
2. Measures of variability
3. Measures of skewness

1.MEASURES OF POSITION

- One of the important characteristics of a set of numbers is its **location**, or **central tendency**.
- A measure of position for a data set describes where the observations are concentrated.
- Mean
- Median
- Mode
- Percentiles and quartiles

Mean

- The most common measure of location or center of data is the ordinary arithmetic average or mean. Because we almost always think of the data as a sample, we will refer to the arithmetic mean as the sample mean.
- The mean of a set of observations x_1, x_2 ……….x_n denoted by X (read" X bar"), is the sum of the observations divided by the number observations.



Definition

If the observations in a sample of size n are $x_{1,}$ x_2 . . . , x_n, then the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$

Example

An engineer' has added polymer latex to portland cement mortar to determine its effects on tension bond strength (in kgf/cm2).
The data that result from this experiment are:
16.85, 16.40, 17.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, and 16.57.
The sample mean tension bond strength for the 10 observations is as follows:

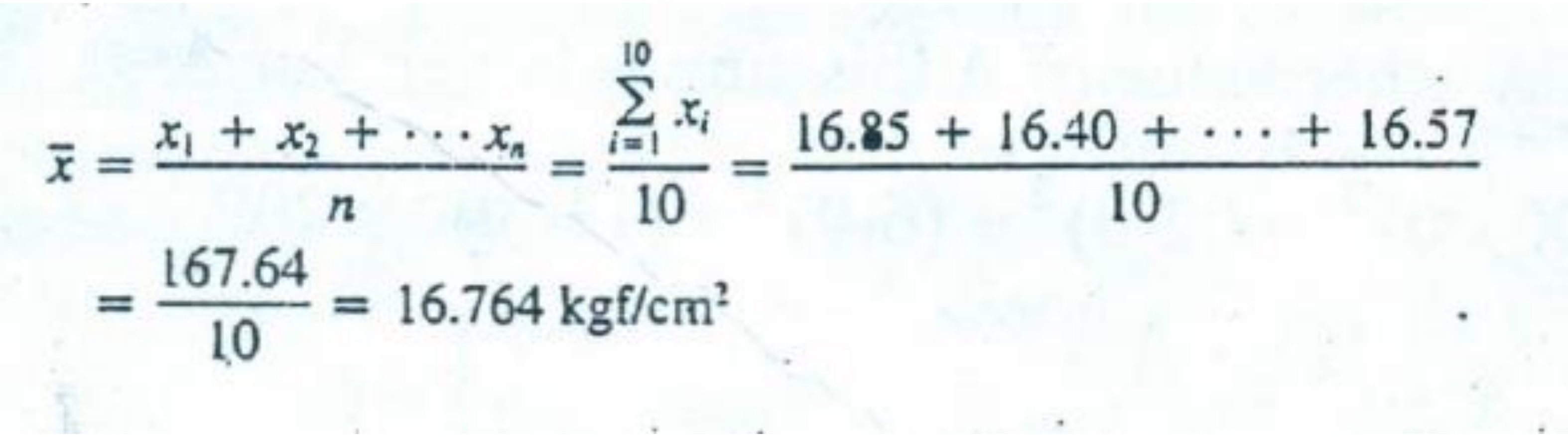
Example

An engineer' has added polymer latex to portland cement mortar to determine its effects on tension bond strength (in kgf/cm2).

The data that result from this experiment are:

16.85, 16.40, 17.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, and 16.57.

The sample mean tension bond strength for the 10 observations is as follows:



$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{16.85 + 16.40 + \cdots + 16.57}{10}$$

$$= \frac{167.64}{10} = 16.764 \text{ kgf/cm}^2$$

Properties of mean

1. The sum of the observations in a data set is equal to the mean multiplied by the number of observations.

$$\sum_{i=1}^{n} X_i = n \bar{X}$$

2. The sum of the deviations of the Xi observations from their mean is zero.

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

Trimmed mean

- A major disadvantage of the mean as a measure of position is that it is greatly influenced by extreme or outlying observations in a data set.

- A trimmed mean is a method of averaging that removes a small designated percentage of the largest and smallest values before calculating the mean.

- After removing the specified outlier observations, the trimmed mean is found using a standard arithmetic averaging formula. The use of a trimmed mean helps eliminate the influence of outliers or data points on the tails that may unfairly affect the traditional mean.

Example

Consider the following array of 12 observations:

1 1 2 4 5 7 8 9 12 12 13 130

for which the mean is 17. Find 50% trimmed mean.

Example

Consider the following array of 12 observations:

1 1 2 4 5 7 8 9 12 12 13 130

for which the mean is 17.

Note that 11 of the 12 observations in the data set fall below the mean. The mean lies away from the main concentration of data here because of the influence of the outlying observation 130.

To lessen the effect of the outlying observations on the mean, modified mean called trimmed mean is employed at times.

Example

The 50 percent trimmed mean is the mean of the central 50 percent of the arrayed observations of the data set.

Thus to calculate a 50 percent trimmed mean, we simply eliminate 25 percent of the observations at each end of the array and calculate the mean of the remaining middle observations:

~~1 1 2~~ **4 5 7 8 9 12** ~~12 13 130~~

(4+5+7+8+9+12) / 6 = 7.5

Example

Consider the following array of 12 observations:

1 1 2 4 5 7 8 9 12 12 13 130

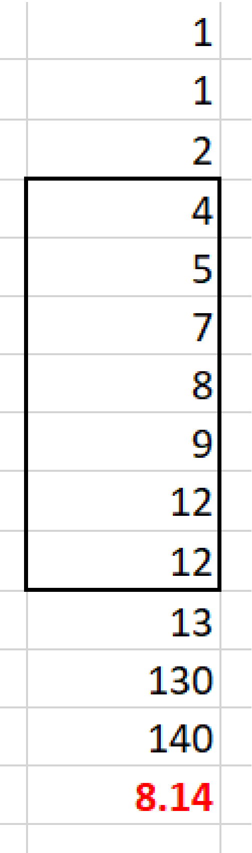
Find 40% trimmed mean.

Consider the following array of 13 observations:

1 1 2 4 5 7 8 9 12 12 13 130 140

Find 50% trimmed mean.

Example



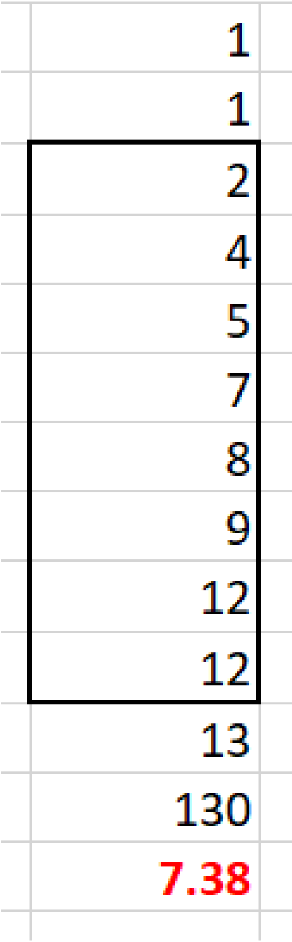
1
1
2
4
5
7
8
9
12
12
13
130
140
8.14

50% trimmed mean of 13 observations

13 x 0.5 = 6.5
An even integer (6) should be determined and half of this value should be removed from each end of the array (3 from the lowest observations, 3 from the greatest observations).

40% trimmed mean of 12 observations

12 x 0.4 = 4.8
An even integer (4) should be determined and half of this value should be removed from each end of the array (2 from the lowest observations, 2 from the greatest observations).



1
1
2
4
5
7
8
9
12
12
13
130
7.38

Median

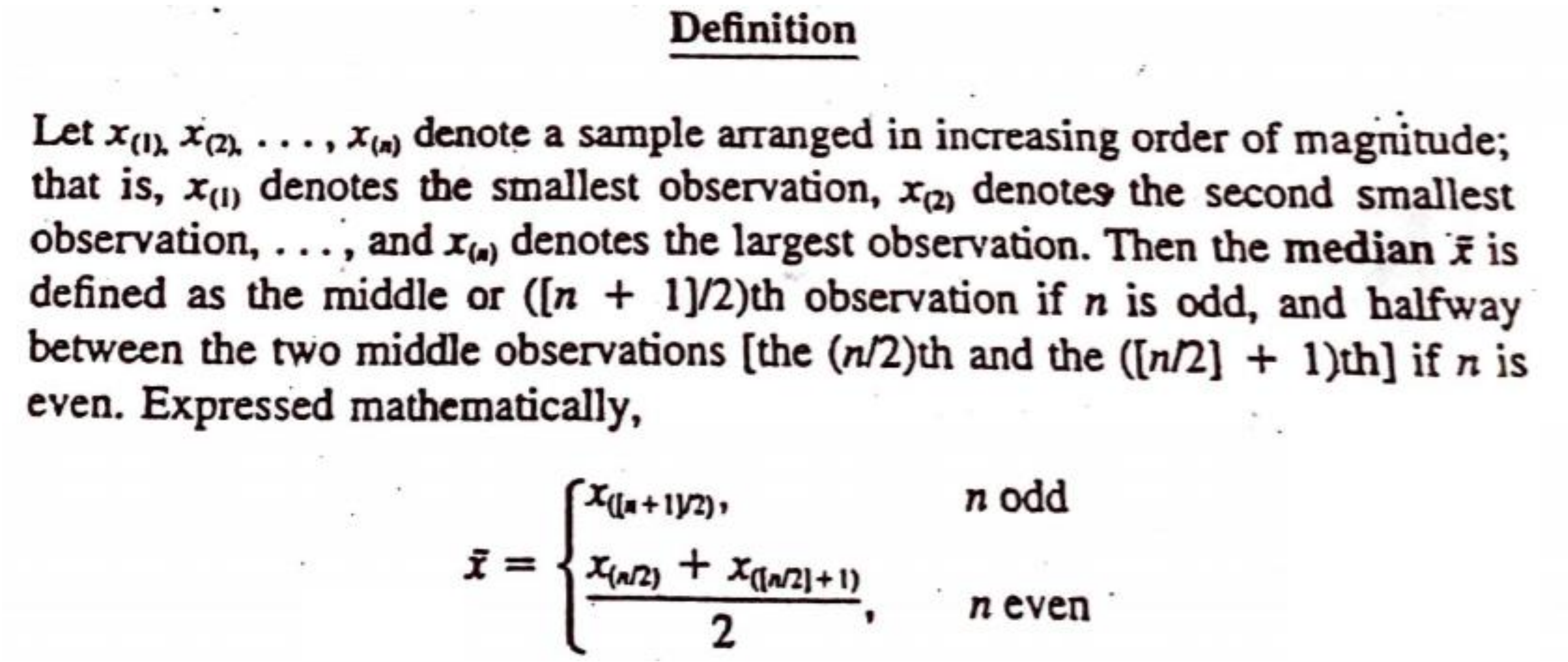
- In data sets where the distribution pattern is not symmetrical, the mean tends to be located somewhat away from the concentration of the observations.

- Also, when a data set contains one or more outlying observations, as in the trimmed mean example, the mean can be influenced by the outlying observations.

- A measure of position that is more central in the distribution, in the sense that it divides the arrayed observations into two equal parts, is the median.

- The **median** of a data set, denoted by Md, is the value of the middle observation in an array of the data set.

Median



Definition

Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote a sample arranged in increasing order of magnitude; that is, $x_{(1)}$ denotes the smallest observation, $x_{(2)}$ denotes the second smallest observation, . . . , and $x_{(n)}$ denotes the largest observation. Then the **median** $\tilde{x}$ is defined as the middle or $([n + 1]/2)$th observation if n is odd, and halfway between the two middle observations [the $(n/2)$th and the $([n/2] + 1)$th] if n is even. Expressed mathematically,

$$\tilde{x} = \begin{cases} x_{([n+1]/2)}, & n \text{ odd} \\ \dfrac{x_{(n/2)} + x_{([n/2]+1)}}{2}, & n \text{ even} \end{cases}$$

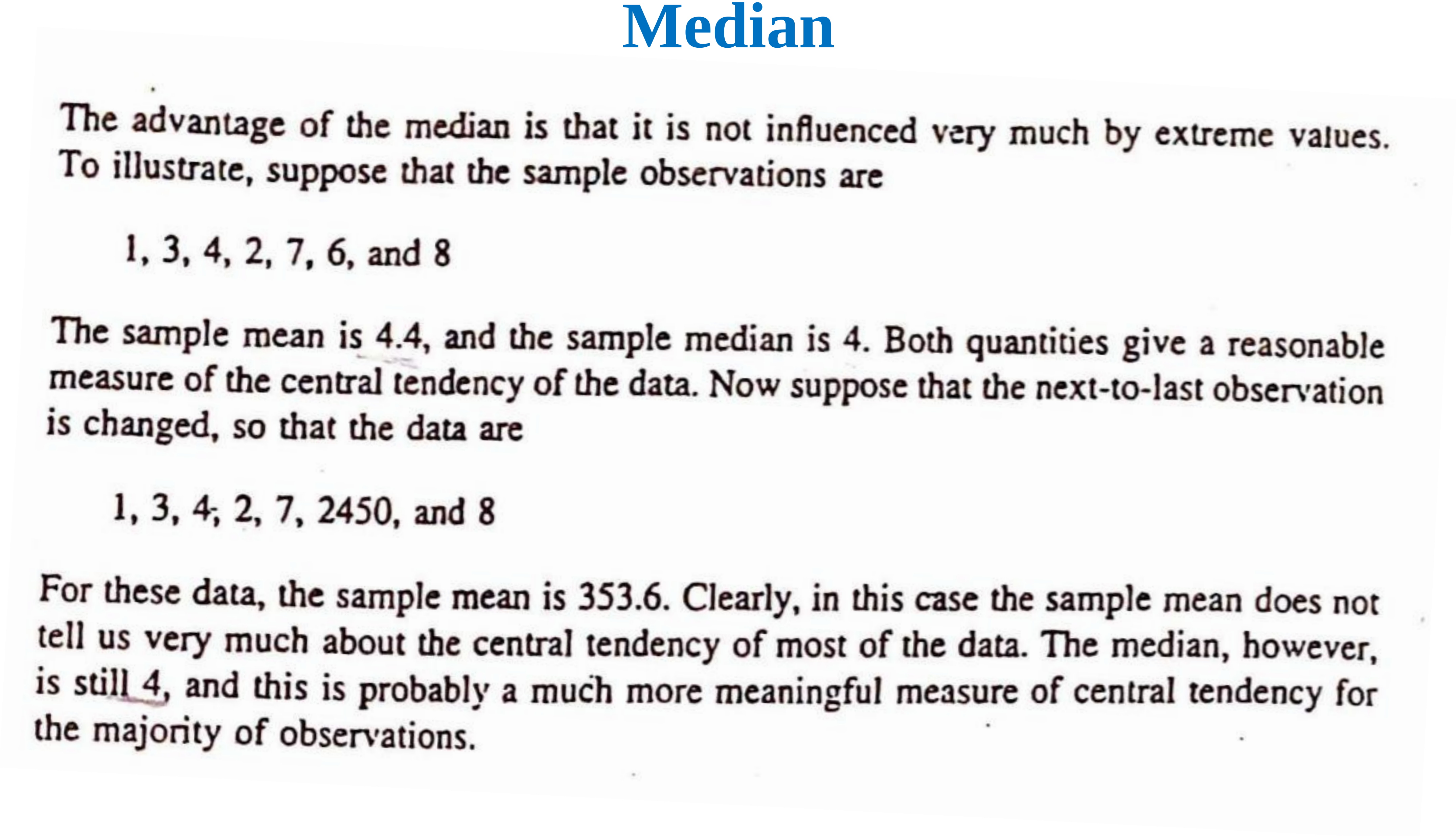
Example

Consider the following arrays of 7 observations:

1 3 4 2 7 6 8 (sample mean: 4.4)

1 3 4 2 7 2450 8 (sample mean: 353.6)

Median



The advantage of the median is that it is not influenced very much by extreme values. To illustrate, suppose that the sample observations are

1, 3, 4, 2, 7, 6, and 8

The sample mean is 4.4, and the sample median is 4. Both quantities give a reasonable measure of the central tendency of the data. Now suppose that the next-to-last observation is changed, so that the data are

1, 3, 4, 2, 7, 2450, and 8

For these data, the sample mean is 353.6. Clearly, in this case the sample mean does not tell us very much about the central tendency of most of the data. The median, however, is still 4, and this is probably a much more meaningful measure of central tendency for the majority of observations.

Mode

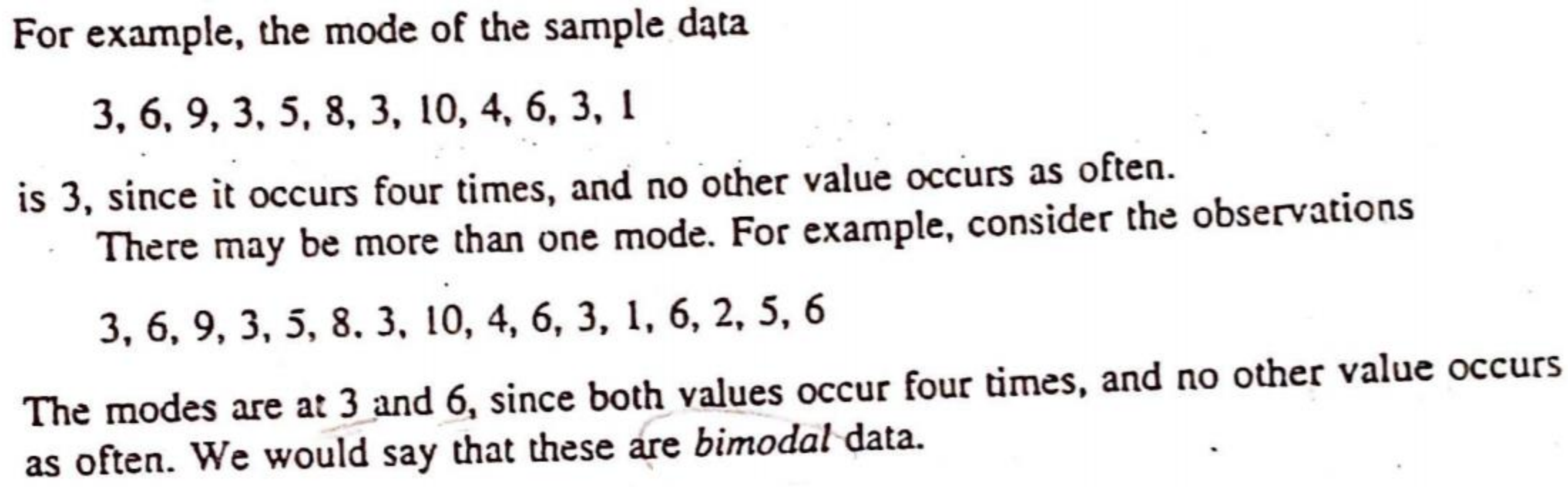
The mode is the observation that occurs most frequently in the sample.

3, 6, 9, 3, 5, 8, 3, 10, 4, 6, 3, 1

3, 6, 9, 3, 5, 8, 3, 10, 4, 6, 3, 1, 6, 2, 5, 6

Mode

The mode is the observation that occurs most frequently in the sample.



For example, the mode of the sample data

3, 6, 9, 3, 5, 8, 3, 10, 4, 6, 3, 1

is 3, since it occurs four times, and no other value occurs as often.

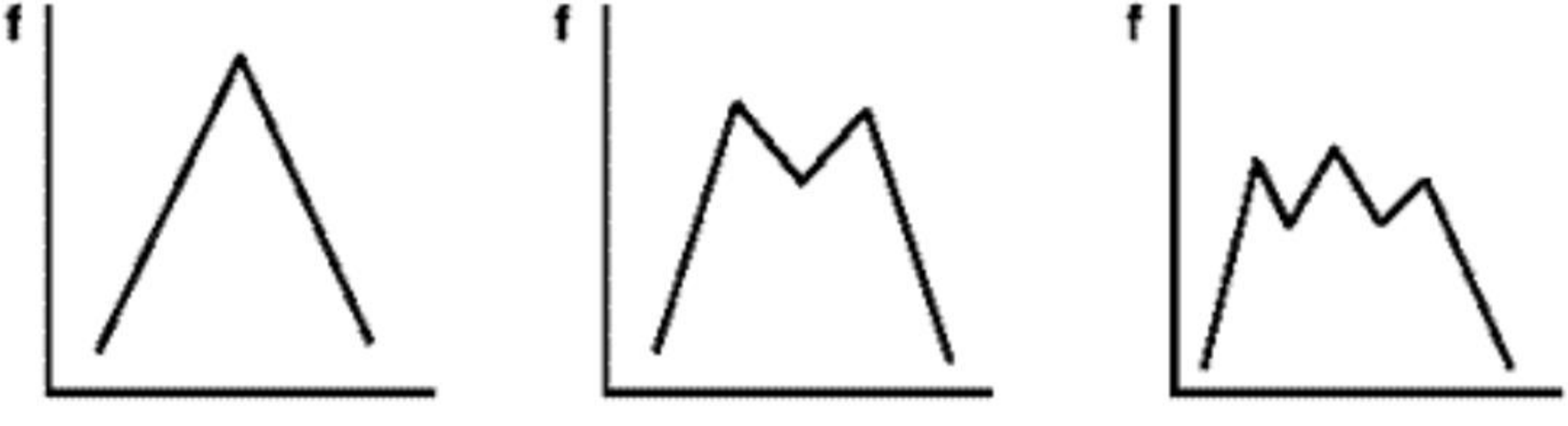
There may be more than one mode. For example, consider the observations

3, 6, 9, 3, 5, 8. 3, 10, 4, 6, 3, 1, 6, 2, 5, 6

The modes are at 3 and 6, since both values occur four times, and no other value occurs as often. We would say that these are *bimodal* data.

While most distributions that occur in statistical data have only one main peak (**unimodal**), other distributions may have two peaks (**bimodal**) or more than two peaks (**multimodal**).

Mode



Unimodal

Bimodal

Multimodal

Percentiles and quartiles

- Percentiles are important and widely used summary measures for the position of a data set.
- The median is actually a percentile. It is the 50th percentile, the value that divides the data array into two equal parts.
- Other widely used percentiles are the 25th and 75th percentiles and 10th and 90th percentiles.
- In general a **percentile** is a value such that a given percentage of a data set is at or below this value.
- Many percentiles are known by other names. When an ordered set of data is divided into four equal parts, the division points are **quartiles**.
- **25th percentile = 1st quartile**
- **50th percentile = 2nd quartile = median**
- **75th percentile = 3rd quartile**

Percentiles and quartiles

The *p*th percentile of a data set is a value such that at least *p* percent of the items take on this value or less and at least (100 - *p*) percent of the items take on this value or more.

Arrange the data in ascending order.

Compute index *i*, the position of the *p*th percentile.

$$i = (p/100)n$$

If *i* is not an integer, round up. The *p* th percentile is the value in the *i* th position.

If *i* is an integer, the *p* th percentile is the average of the values in positions *i* and *i* +1.

Example

Find 90th Percentile

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Example

90th Percentile

$i = (p/100)n = (90/100)70 = 63$

Averaging the 63rd and 64th data values:

90th Percentile = (580 + 590)/2 = 585

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Example

Find 50th Percentile = Median

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Example

50th Percentile = Median

Median = 50th percentile

$i = (p/100)n = (50/100)70 = 35$

Averaging the 35th and 36th data values:

Median = (475 + 475)/2 = 475

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

2.MEASURES OF VARIABILITY

Location or central tendency does not necessarily provide enough information to describe the data adequately.
The variability of the observations in a data set is often another feature of interest when data set is summarized. For example, in choosing vendor A or vendor B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

- Range
- Interquartile Range
- Variance
- Standard Deviation

Range

- The range of a data set is the difference between the largest and smallest data values.
- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Example

Find the range

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Example

Range = largest value - smallest value

Range = 615 - 425 = 190

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Interquartile range

- The interquartile range of a data set is the difference between the third quartile and the first quartile.

- It is the range for the middle 50% of the data.

- It overcomes the sensitivity to extreme data values.

Example

Find the interquartile range

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Example

3rd Quartile (Q3) = 525

1st Quartile (Q1) = 445

Interquartile Range = Q3 - Q1 = 525 - 445 = 80

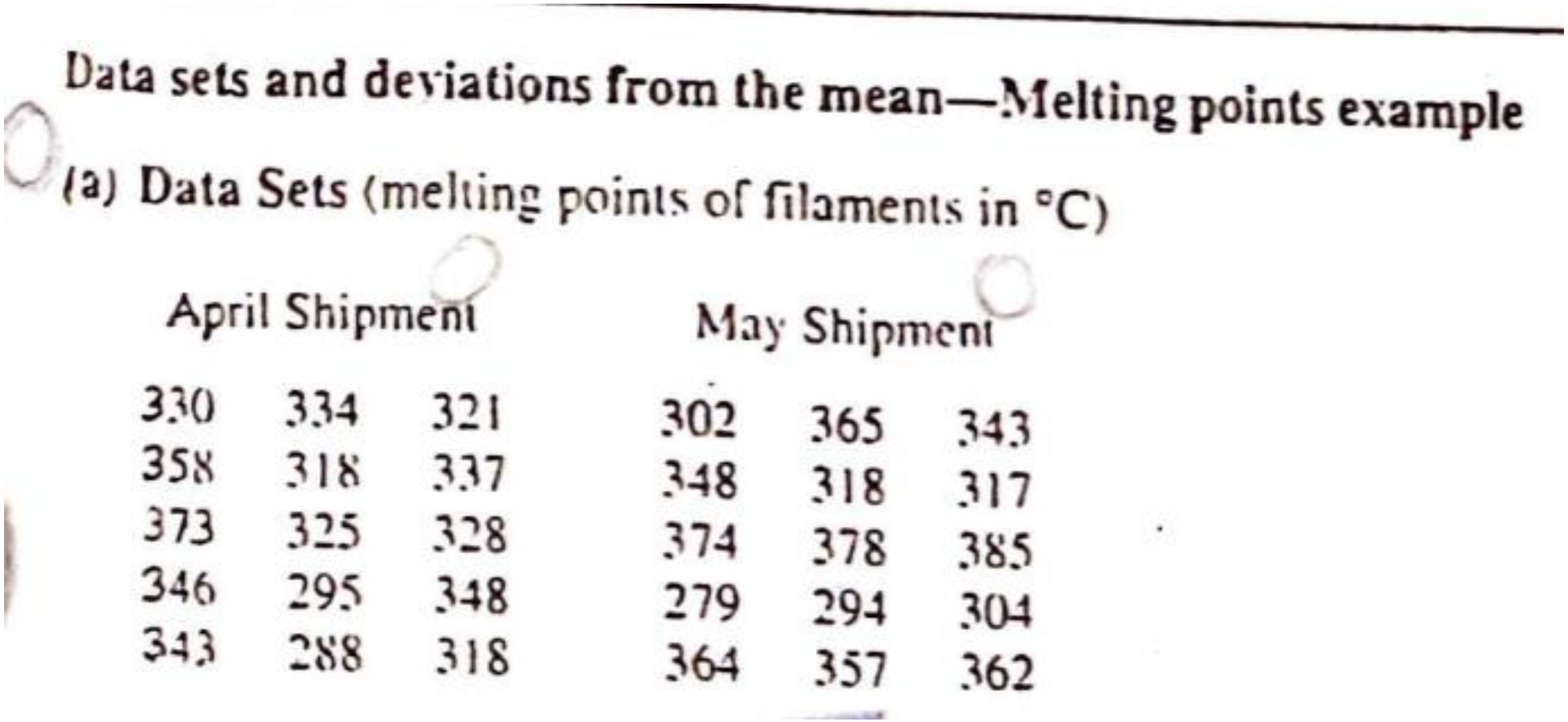
425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Variance

- The most commonly used measure of variability in statistical analysis is called the variance. It is a measure that takes into account all the observations in a data set.
- The variance is the average of the squared differences between each data value and the mean.
- The variance is denoted by s^2.

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

Example

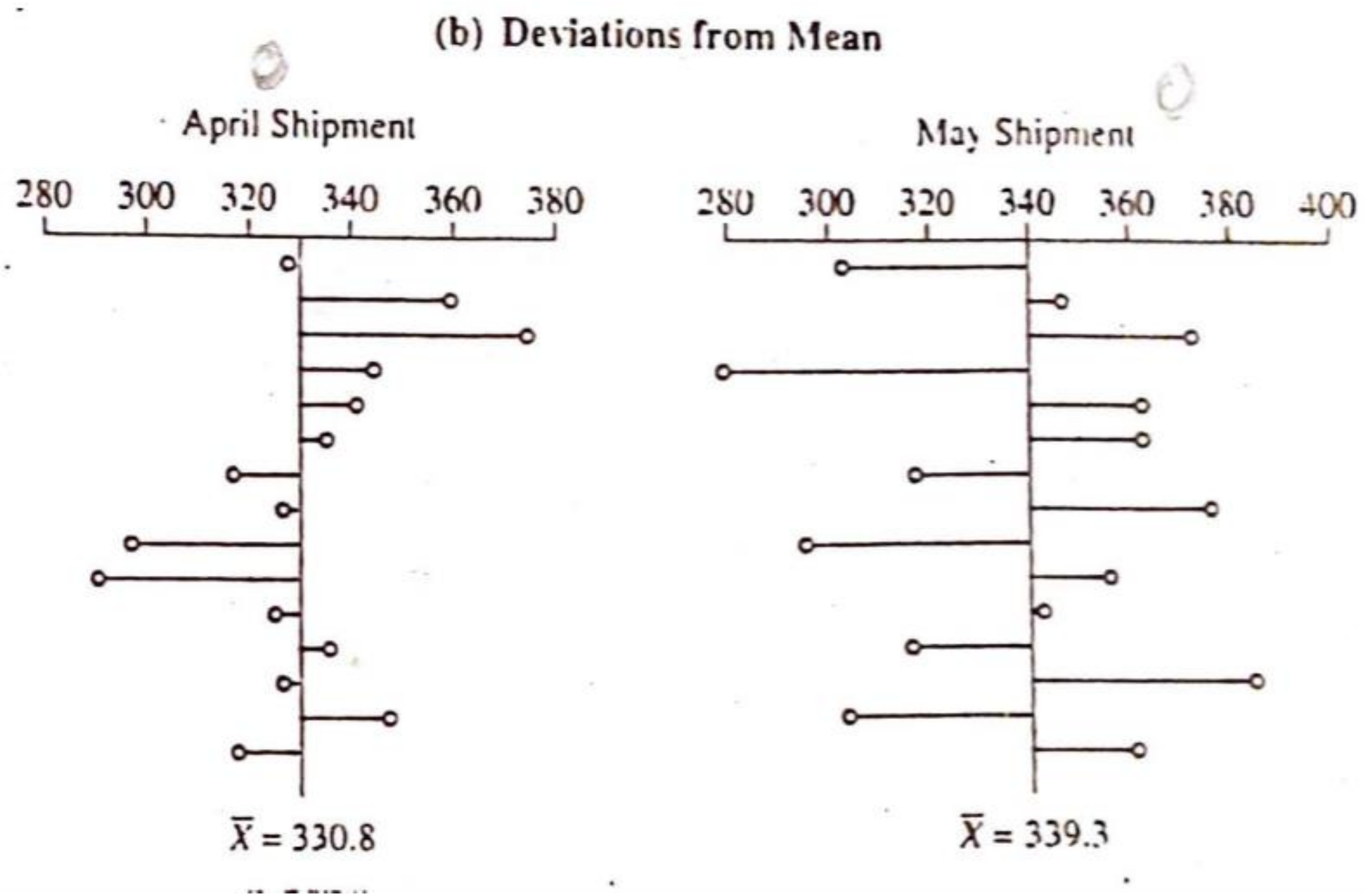


Data sets and deviations from the mean—Melting points example

(a) Data Sets (melting points of filaments in °C)

April Shipment			May Shipment		
330	334	321	302	365	343
358	318	337	348	318	317
373	325	328	374	378	385
346	295	348	279	294	304
343	288	318	364	357	362

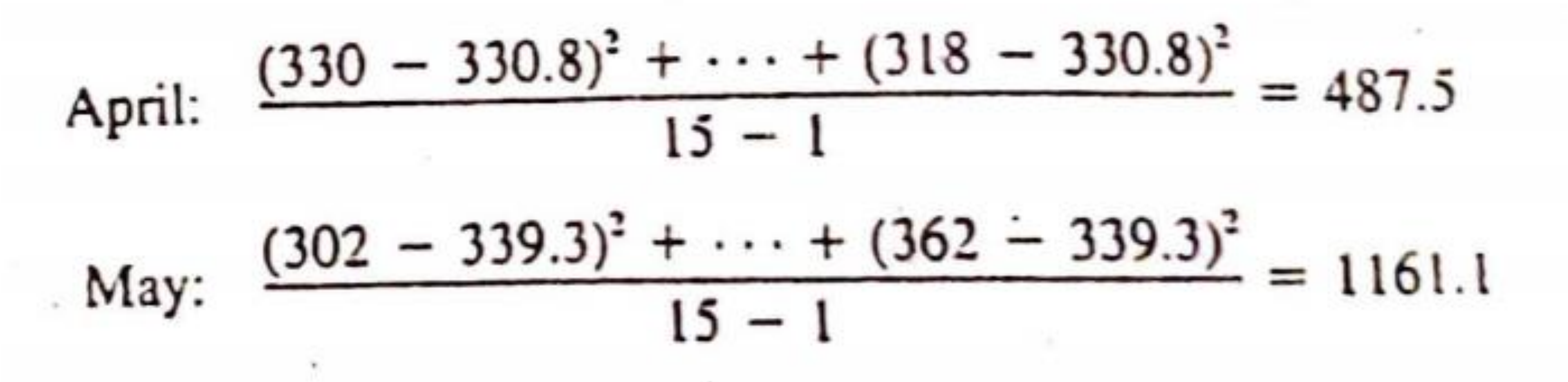
Example


(b) Deviations from Mean
April Shipment
280
300
320
340
360
380
$\bar{X}$ = 330.8
May Shipment
280
300
320
340
360
380
400
$\bar{X}$ = 339.3

Example

- To compare the variability of melting points far the two data sets, the customer considered the deviations of the observations· in each data set from their respective means.
- The means of the two data sets are 339.8 and 339.3, respectively. The two sets of deviations-are shown graphically in Figure b.
- This figure shows readily that there was greater variability far the May filaments than for the April filaments.
- The variance is a measure that provides quantified information about the variability in a data set.

Example



$$\text{April:}\quad \frac{(330-330.8)^2+\cdots+(318-330.8)^2}{15-1}=487.5$$

$$\text{May:}\quad \frac{(302-339.3)^2+\cdots+(362-339.3)^2}{15-1}=1161.1$$

The variance for the may filaments is more than twice that for the April filaments.

Standard deviation

- The variance is expressed in units that are the square of the units of the measure of the variable under study. For instance, in the melting points example, the variance is expressed in Celsius degrees squared.
- Often, it is desirable to return to the original units of measure. We obtain the original units by taking the positive square root of the variance, which is called as the standard deviation of a data set.
- It is measured in the same units as the data, making it more easily comparable, than the variance, to the mean.
- The standard deviation is denoted *s*.

$$s = \sqrt{s^2}$$

Standard deviation

- For the melting points example, the standard deviations are s = $\sqrt{487.5}$ = 22.1 Celsius degrees for the April filaments and s= $\sqrt{1161.1}$ = 34.1 Celsius degrees for the May filaments. The standard deviation again show that there was greater variability of melting points in the May filaments than in the April filaments.

3.Measures of skewness

- Measures of position are concerned with the location around which the observations are concentrated.

- Measures of variability consider the extent to which the observations vary.

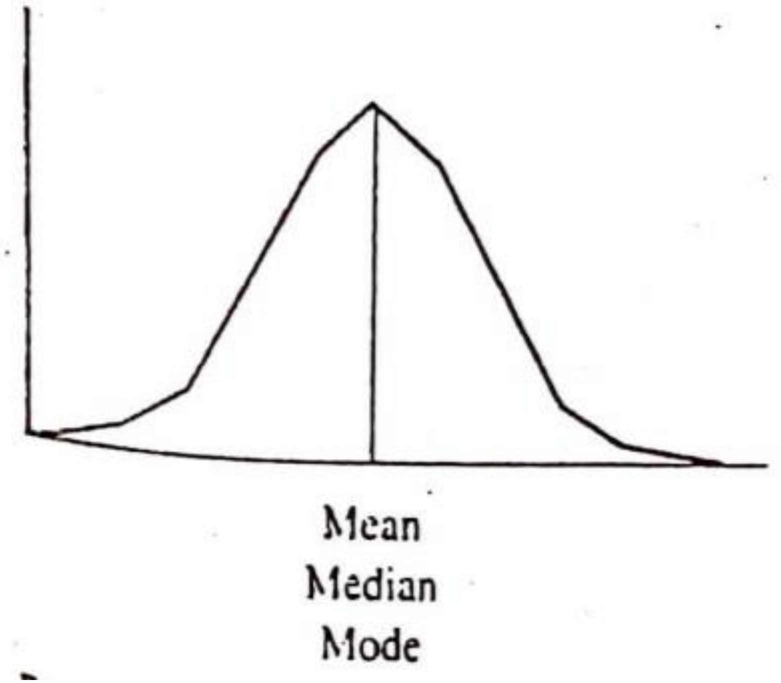
- Measures of skewness summarize the extent to which the observations are symmetrically distributed.

- Skewness
- Kurtosis

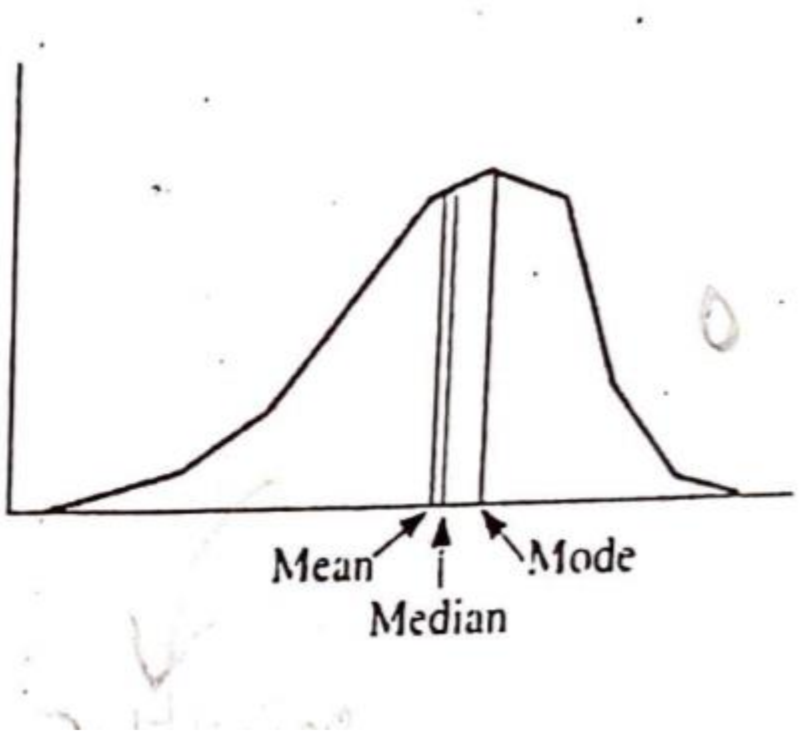
Skewness

- A data set with observations that are not symmetrically distributed is said to be **skewed**.
- The skewness of a data set is readily studied if the data are presented as a frequency distribution or a stem and leaf display.
- Another way of studying the skewness of a data set is to compare the values of the mode, the median, and the mean.
- When the mean differs substantially from the median and mode, we have evidence of skewness in the data set.

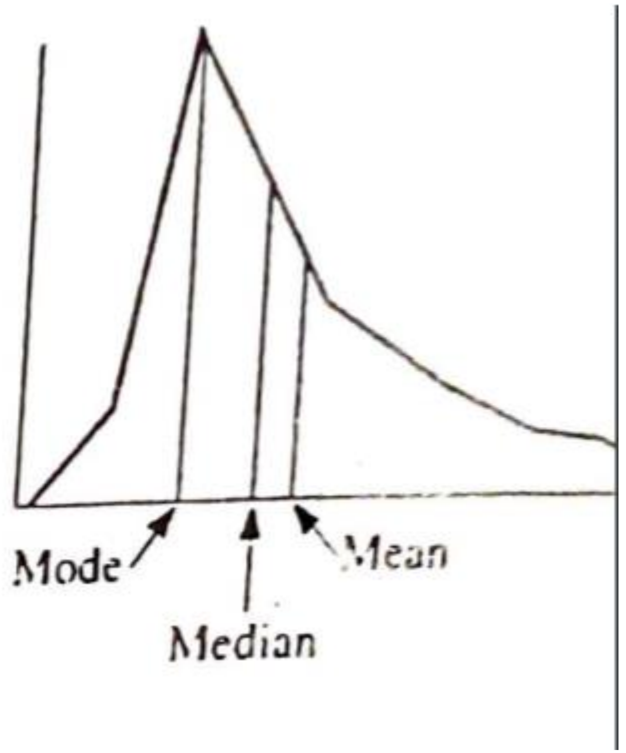
Skewness



Symmetrical



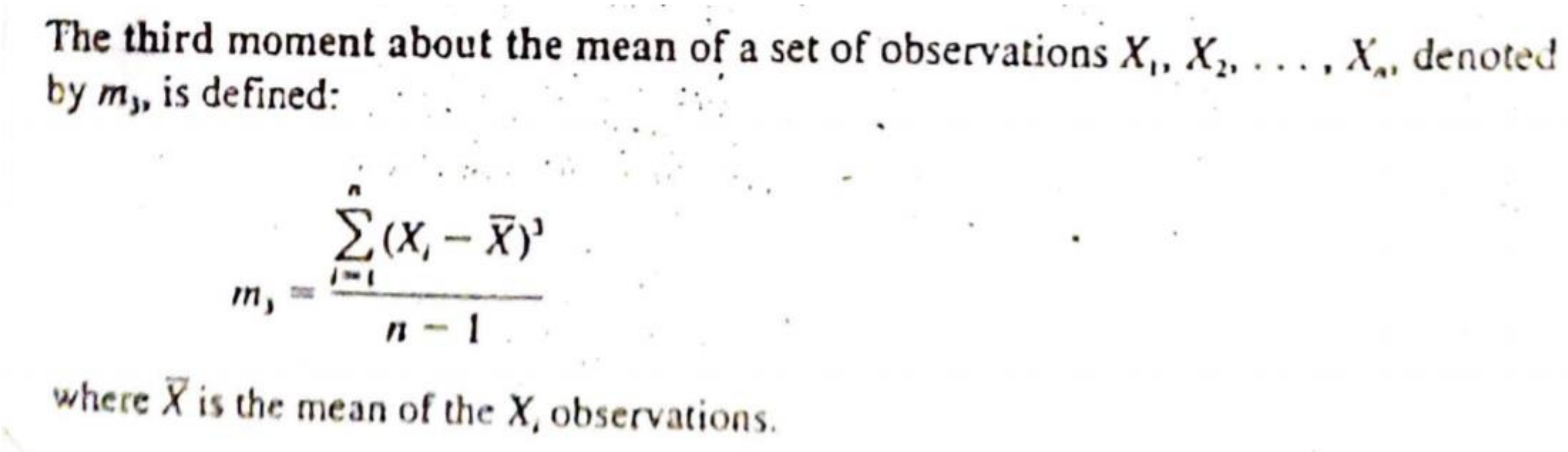
Skewed left (negatively)



Skewed right (positively)

Skewness

- In addition to informal analysis of the symmetry or lack of symmetry of a data set, we can also utilize some direct measures.
- **Skewness**:



The third moment about the mean of a set of observations $X_1, X_2, \ldots, X_n$, denoted by m_3, is defined:

$$m_3 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n - 1}$$

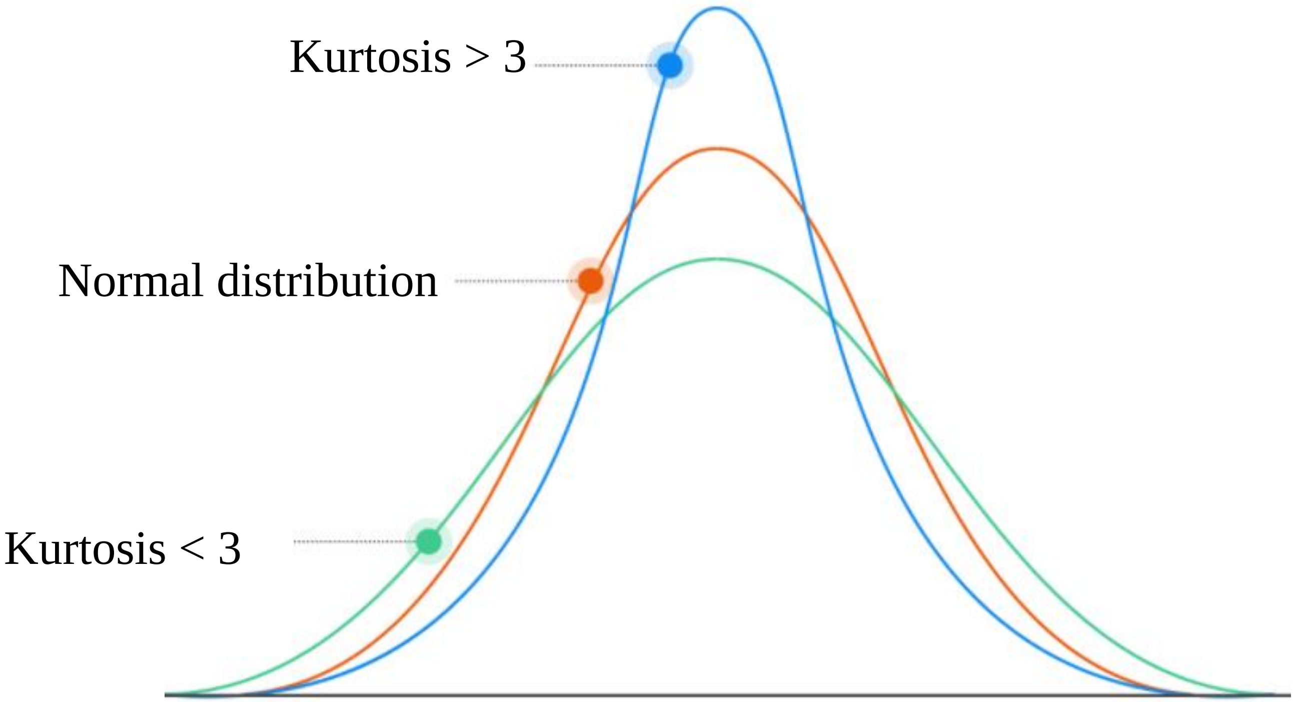
where $\bar{X}$ is the mean of the X_i observations.

- m3 will be positive or negative according to whether the direction of skewness is positive (right) or negative (left).
- If the observations in the data set are symmetrically distributed, about the mean, the third moment will be zero.

Kurtosis

The fourth moment about the mean is called as Kurtosis.

$$m_4 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n-1}$$



A normal distribution has a kurtosis of 3, so kurtosis values for a distribution are often compared to 3.

Basic excel formulations

=AVERAGE() in Turkish =ORTALAMA()

Returns the average (arithmetic mean) of the arguments. For example, if the range A1:A20 contains numbers, the formula **=AVERAGE(A1:A20)** returns the average of those numbers.

=MEDIAN() in Turkish =ORTANCA()

Returns the median of the given numbers. The median is the number in the middle of a set of numbers.

= MODE.SNGL() in Turkish =ENÇOK_OLAN()

Returns the most frequently occurring, or repetitive, value in an array or range of data.

Basic excel formulations

=VAR.S() in Turkish =VARS()

Estimates variance based on a sample (ignores logical values and text in the sample). In earlier versions, the formula was =VAR()

=STDEV.S() in Turkish =STDSAPMA.S()

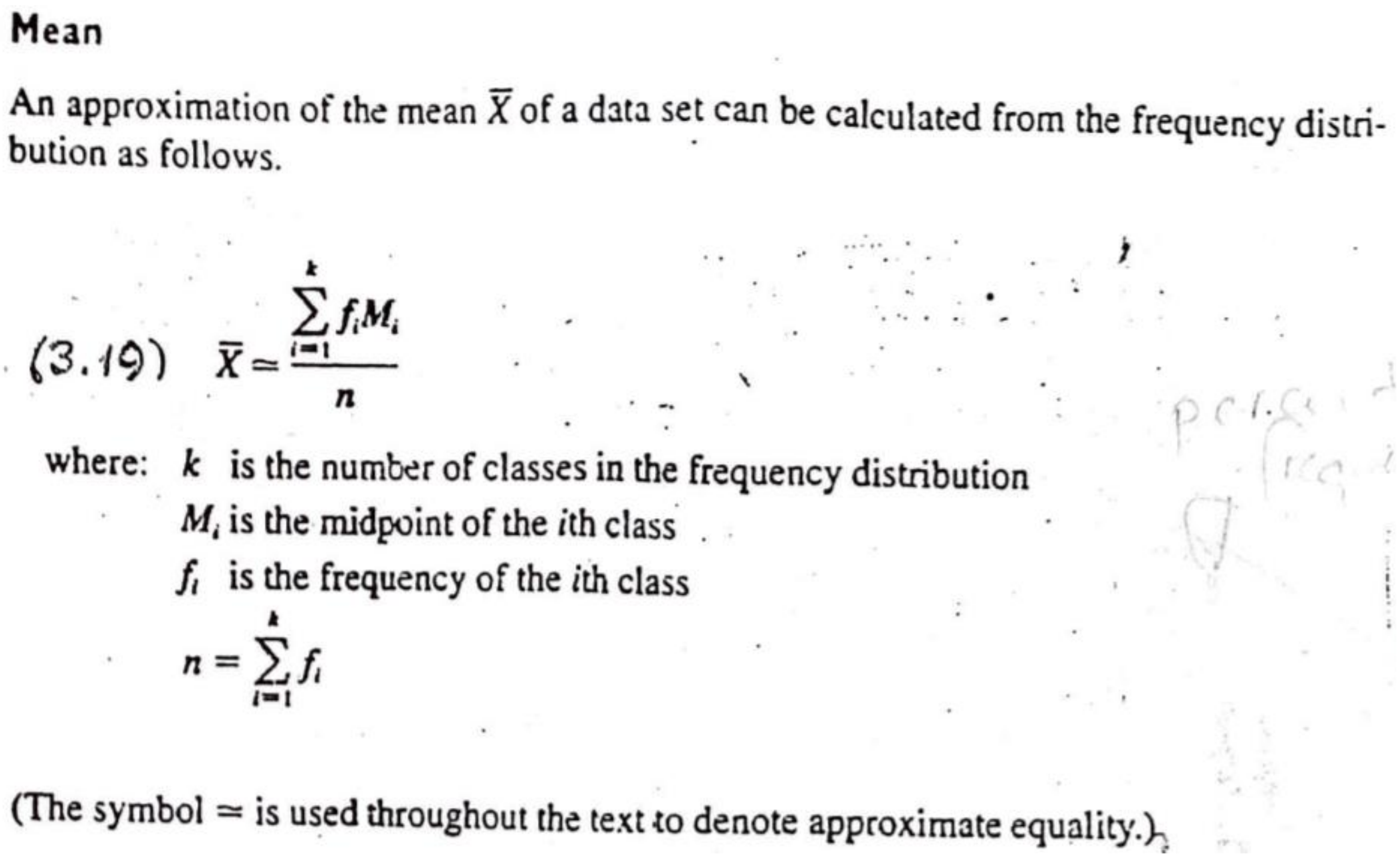
Estimates standard deviation based on a sample (ignores logical values and text in the sample). In earlier versions, the formula was =STDEV() OR =STDSAPMA

= PERCENTILE.INC() in Turkish =YÜZDEBİRLİK()

Returns the most frequently occurring, or repetitive, value in an array or range of data.

CALCULATIONS OF SUMMARY MEASURES – DESCRIPTIVE STATISTICS FROM FREQUENCY DISTRIBUTIONS

Mean



Mean

An approximation of the mean $\overline{X}$ of a data set can be calculated from the frequency distribution as follows.

$$(3.19) \quad \overline{X} \approx \frac{\sum_{i=1}^{k} f_i M_i}{n}$$

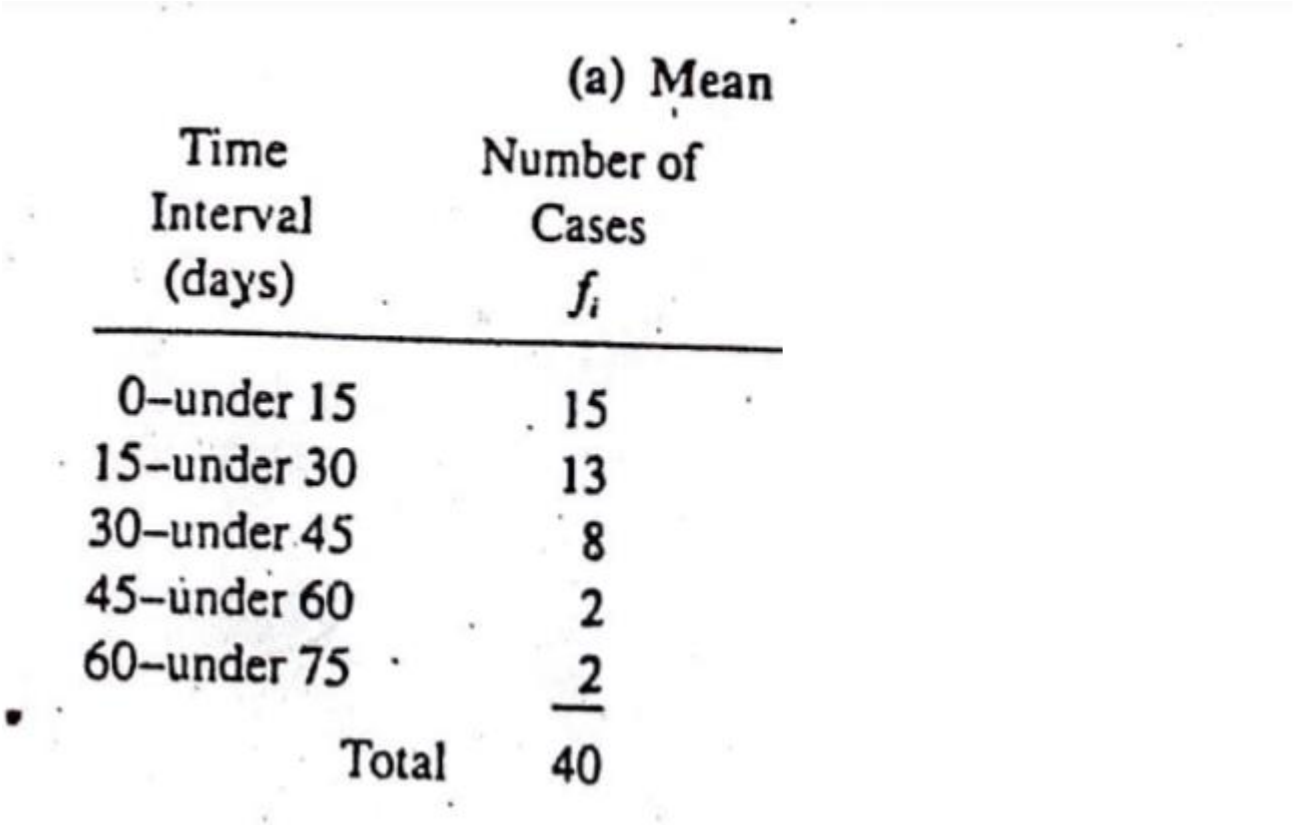
where: k is the number of classes in the frequency distribution
M_i is the midpoint of the ith class
f_i is the frequency of the ith class
$n = \sum_{i=1}^{k} f_i$

[illegible]
freq

(The symbol $\approx$ is used throughout the text to denote approximate equality.)

Example

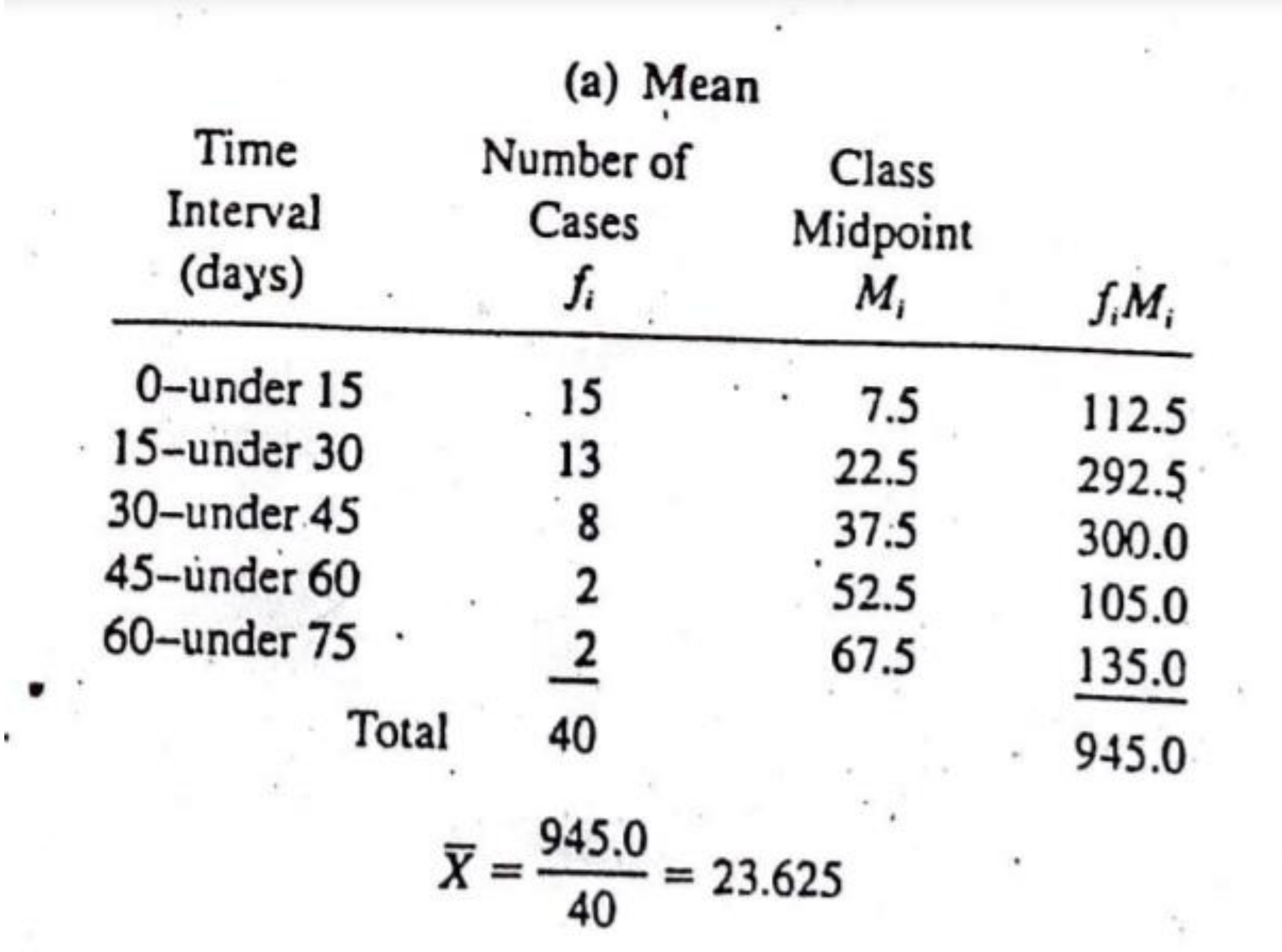
Find the mean of the data set from frequency distribution table



(a) Mean

Time Interval (days)	Number of Cases f_i
0–under 15	15
15–under 30	13
30–under 45	8
45–under 60	2
60–under 75	2
Total	40

Example

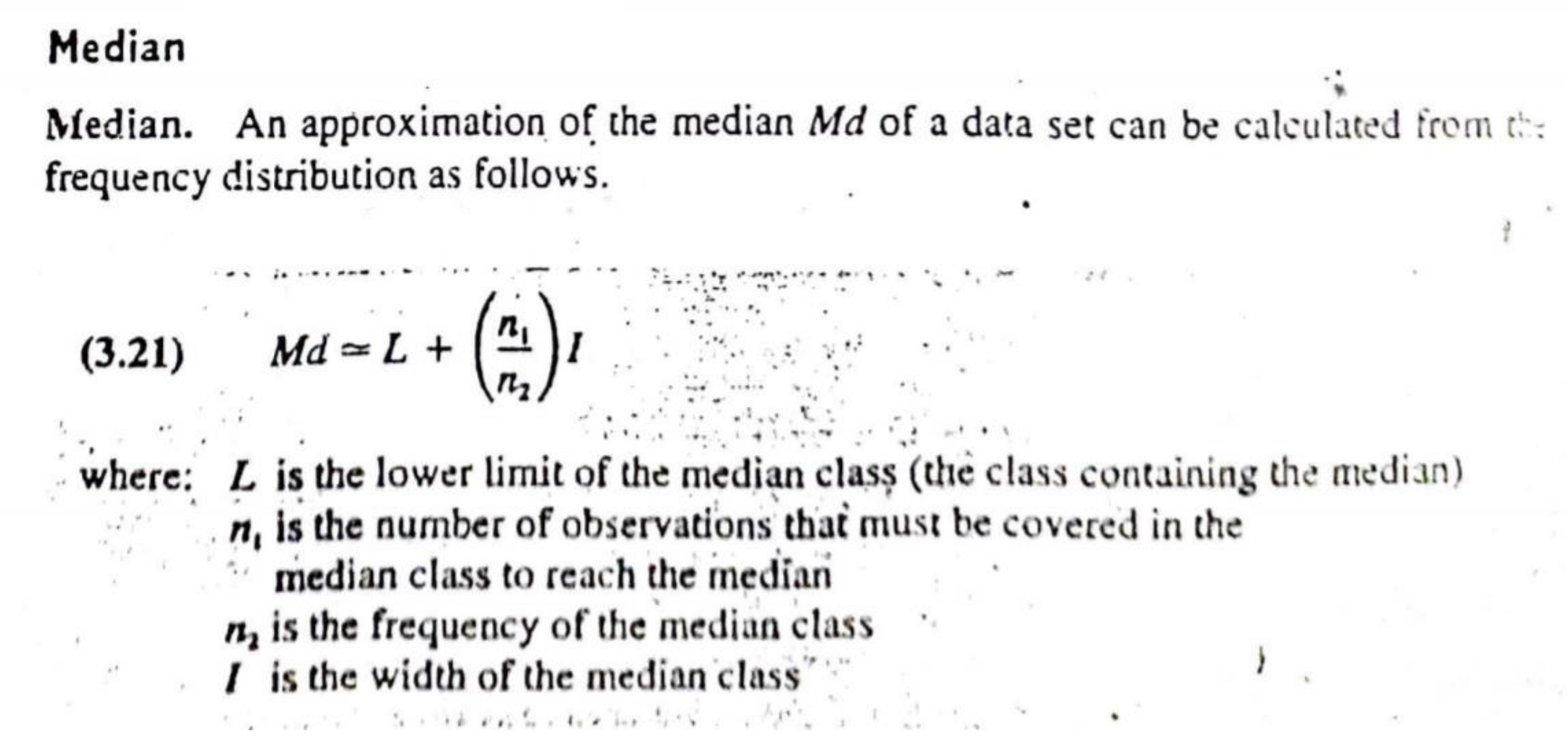


(a) Mean

Time Interval (days)	Number of Cases f_i	Class Midpoint M_i	f_iM_i
0–under 15	15	7.5	112.5
15–under 30	13	22.5	292.5
30–under 45	8	37.5	300.0
45–under 60	2	52.5	105.0
60–under 75	2	67.5	135.0
Total	40		945.0

$$\bar{X} = \frac{945.0}{40} = 23.625$$

Median



Median

Median. An approximation of the median *Md* of a data set can be calculated from the frequency distribution as follows.

$$Md \simeq L + \left(\frac{n_1}{n_2}\right) I \tag{3.21}$$

where: L is the lower limit of the median class (the class containing the median)

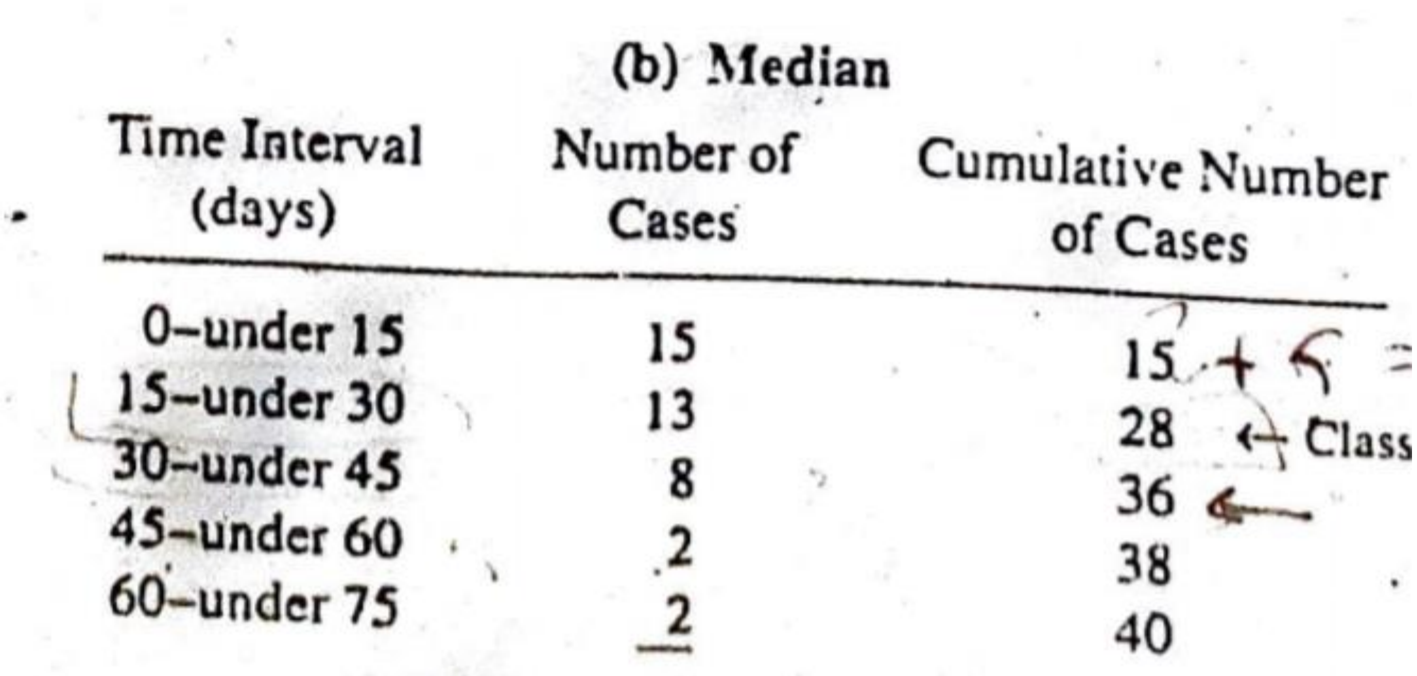
n_1 is the number of observations that must be covered in the median class to reach the median

n_2 is the frequency of the median class

I is the width of the median class

Example

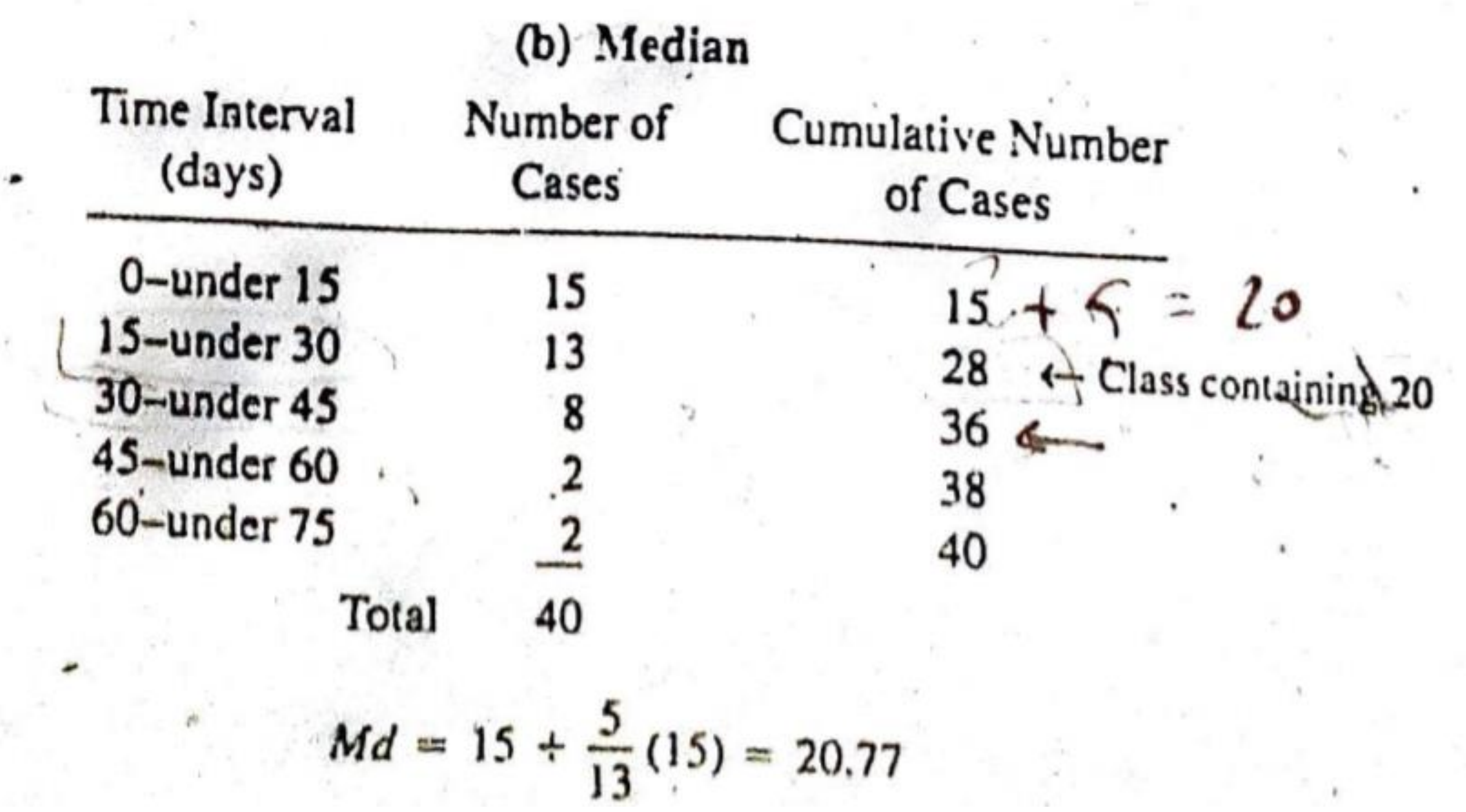
Find the median of the data set from frequency distribution table



(b) Median

Time Interval (days)	Number of Cases	Cumulative Number of Cases	
0–under 15	15	15	+ 5 =
15–under 30	13	28	← Class
30–under 45	8	36	←
45–under 60	2	38	
60–under 75	2	40	

Example



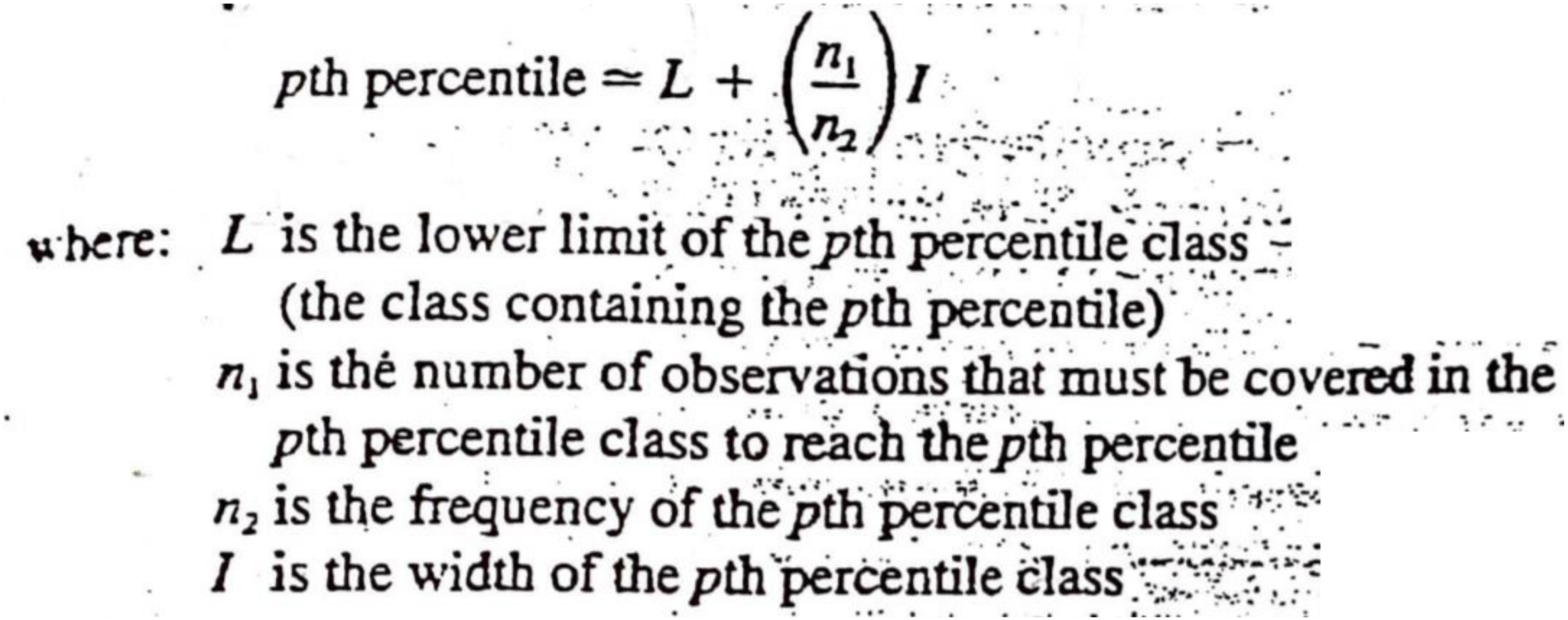
(b) Median

Time Interval (days)	Number of Cases	Cumulative Number of Cases
0–under 15	15	15 + 5 = 20
15–under 30	13	28 ← Class containing 20
30–under 45	8	36 ←
45–under 60	2	38
60–under 75	2	40
Total	40	

$$Md = 15 + \frac{5}{13}(15) = 20.77$$

n/2=40/2=20, L=15, n1=5, n2=13, I= 30-15=15

Percentiles



$$p\text{th percentile} \simeq L + \left(\frac{n_1}{n_2}\right) I$$

where: L is the lower limit of the pth percentile class (the class containing the pth percentile)

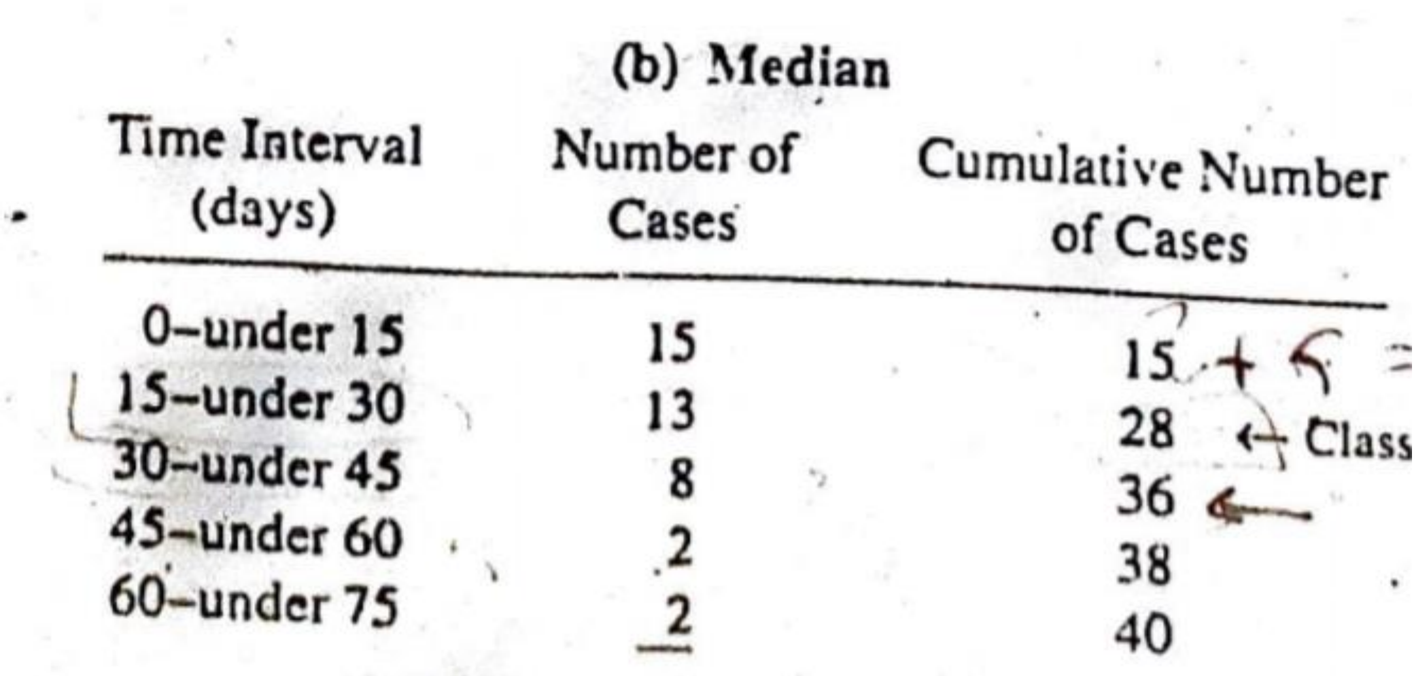
n_1 is the number of observations that must be covered in the pth percentile class to reach the pth percentile

n_2 is the frequency of the pth percentile class

I is the width of the pth percentile class

Example

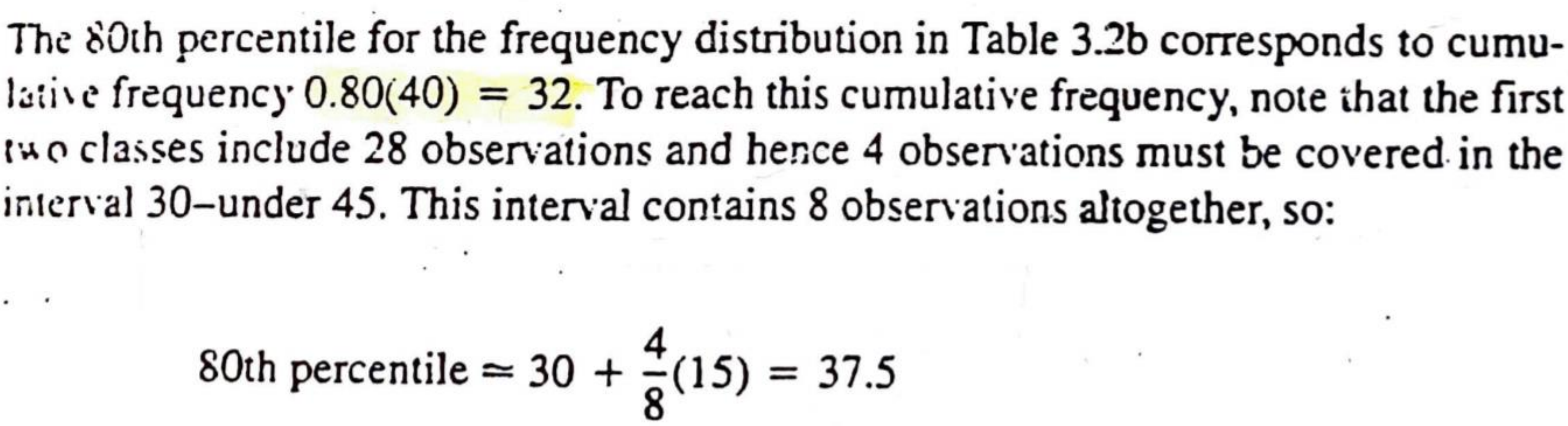
Find 80th percentile from frequency distribution table



(b) Median

Time Interval (days)	Number of Cases	Cumulative Number of Cases
0–under 15	15	15 + ৫ =
15–under 30	13	28 ← Class
30–under 45	8	36 ←
45–under 60	2	38
60–under 75	2	40

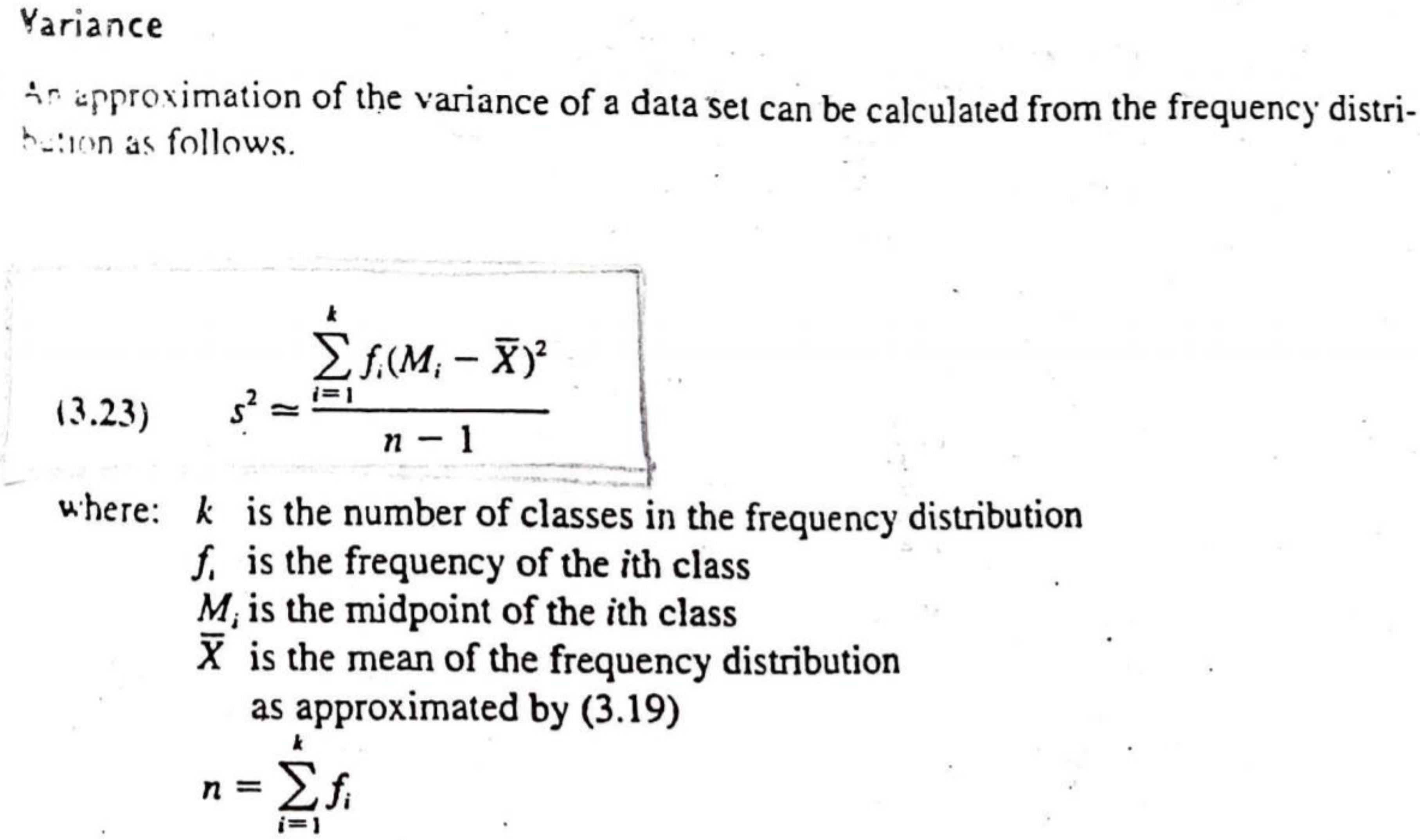
Example



The 80th percentile for the frequency distribution in Table 3.2b corresponds to cumulative frequency 0.80(40) = 32. To reach this cumulative frequency, note that the first two classes include 28 observations and hence 4 observations must be covered in the interval 30–under 45. This interval contains 8 observations altogether, so:

$$80\text{th percentile} \simeq 30 + \frac{4}{8}(15) = 37.5$$

Variance



Variance

An approximation of the variance of a data set can be calculated from the frequency distribution as follows.

$$(3.23) \qquad \boxed{s^2 \simeq \frac{\sum_{i=1}^{k} f_i (M_i - \bar{X})^2}{n - 1}}$$

where: k is the number of classes in the frequency distribution

f_i is the frequency of the ith class

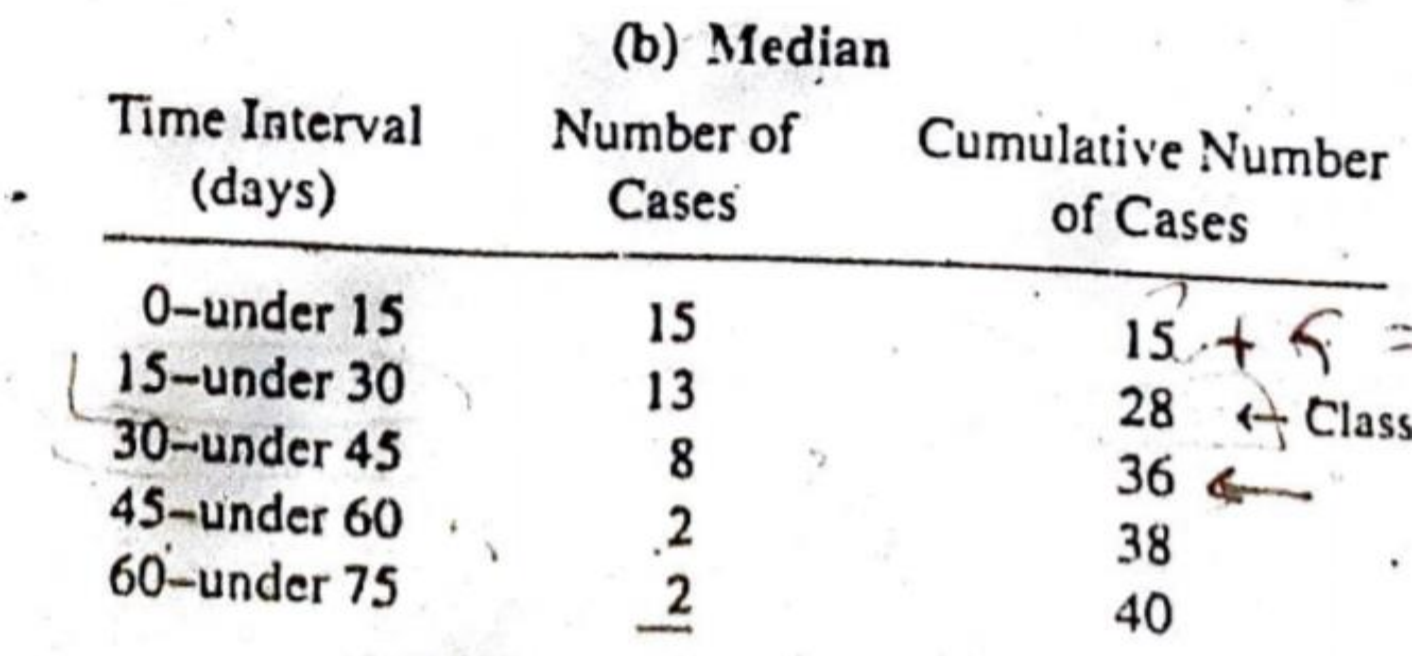
M_i is the midpoint of the ith class

$\bar{X}$ is the mean of the frequency distribution as approximated by (3.19)

$$n = \sum_{i=1}^{k} f_i$$

Example

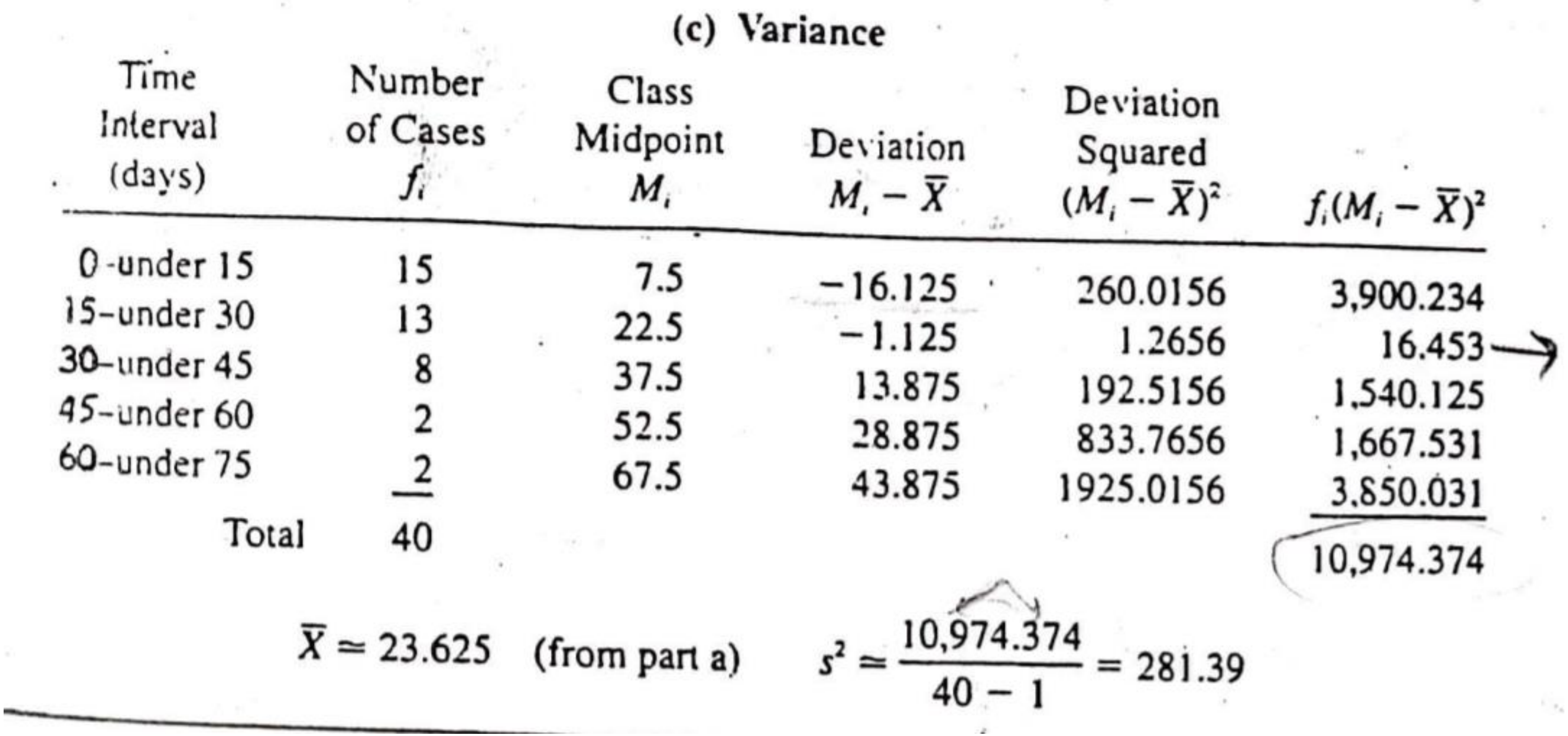
Find the variance and the standard deviation from frequency distribution table



(b) Median

Time Interval (days)	Number of Cases	Cumulative Number of Cases	
0–under 15	15	15	+ [illegible] =
15–under 30	13	28	← Class
30–under 45	8	36	←
45–under 60	2	38	
60–under 75	2	40	

Example



(c) Variance

Time Interval (days)	Number of Cases f_i	Class Midpoint M_i	Deviation $M_i - \bar{X}$	Deviation Squared $(M_i - \bar{X})^2$	$f_i(M_i - \bar{X})^2$
0-under 15	15	7.5	−16.125	260.0156	3,900.234
15–under 30	13	22.5	−1.125	1.2656	16.453
30–under 45	8	37.5	13.875	192.5156	1,540.125
45–under 60	2	52.5	28.875	833.7656	1,667.531
60–under 75	2	67.5	43.875	1925.0156	3,850.031
Total	40				10,974.374

$\bar{X} \simeq 23.625$ (from part a) $\qquad s^2 \simeq \dfrac{10{,}974.374}{40 - 1} = 281.39$