



Chapter 1: Introduction to Statistics

The structure of presentation:

- A lot of definitions
- Main concepts of statistics

Be ready to learn what does variance, standard deviation and many other words mean)

- Things that you know
- A little bit of theorems

Variables

- A **variable** is a characteristic or condition that can change or take on different values.
- Most research begins with a general question about the relationship between two variables for a specific group of individuals.

Population

- The entire group of individuals is called the **population**.
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.

Sample

- Usually populations are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

Statistic vs Parameter

Describes sample

Describes population

\bar{x} ← mean → μ

S ← Standard deviation → σ

S^2 ← Variance → σ^2

\hat{p} ← proportion → P

n ← Size → N

Types of Variables

Variables can be classified as discrete or continuous.

- **Discrete variables** (such as class size) consist of indivisible categories (eg: 2 students , cannot be 2.5 students)
- **Continuous variables** (such as time or weight) are infinitely divisible into whatever units a researcher may choose. For example, time can be measured to the nearest minute, second, half-second, etc.

Measuring Variables

- To establish relationships between variables, researchers must observe the variables and record their observations. This requires that the variables be **measured**.
- The process of measuring a variable requires a set of categories called a **scale of measurement** and a process that classifies each individual into one category.

4 Types of Measurement Scales

1) A **nominal scale** is an unordered set of categories identified only by name (qualitative data).

- Nominal measurements only permit you to determine whether two individuals are the same or different.

- Order does not matter

Eg: Name, colors, labels, gender, etc.

2) An **ordinal scale** is an ordered set of categories. Ordinal measurements tell you the direction of difference between two individuals. Ranking/ placement

- The order matters

- Difference cannot be measured

Eg: 1st place with score 1.2s, 2nd place with score 2.7s and 3rd place with score 3.0s

4 Types of Measurement Scales

3) An **interval scale** is an ordered series of equal-sized categories. Interval measurements identify the direction and magnitude of a difference. The zero point is located arbitrarily on an interval scale.

- The order matters
 - The difference can be measured(except ratios)
 - No true “0” starting point
- Eg: 25°C, 50°C, 75°C

4 Types of Measurement Scales

4) A **ratio scale** is an interval scale where a value of zero indicates none of the variable. Ratio measurements identify the direction and magnitude of differences and allow ratio comparisons of measurements.

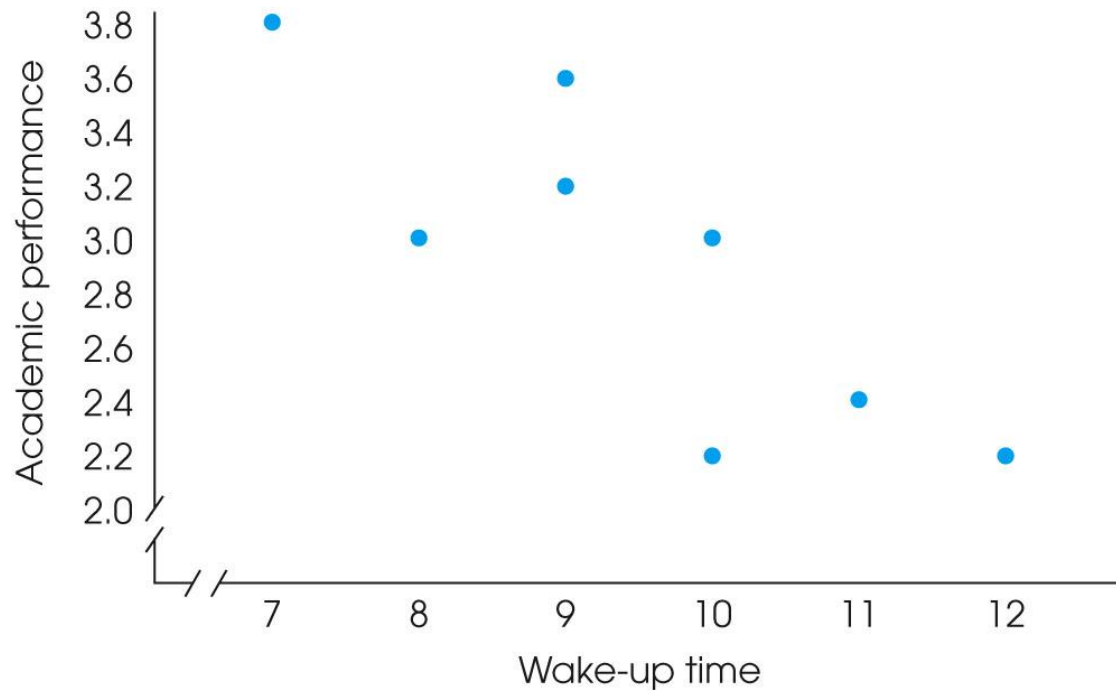
- The order matters
- Difference measurable(including ratios)
- Counts a “0” starting point

Eg: grades in the class, gpa

Correlational Studies

- The goal of a **correlational** study is to determine whether there is a relationship between two variables and to describe the relationship.
- A **correlational** study simply observes the two variables as they exist naturally.

Child	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0



Experiments

- The goal of an **experiment** is to demonstrate a cause-and-effect relationship between two variables; that is, to show that changing the value of one variable causes changes to occur in a second variable.

Experiments (cont.)

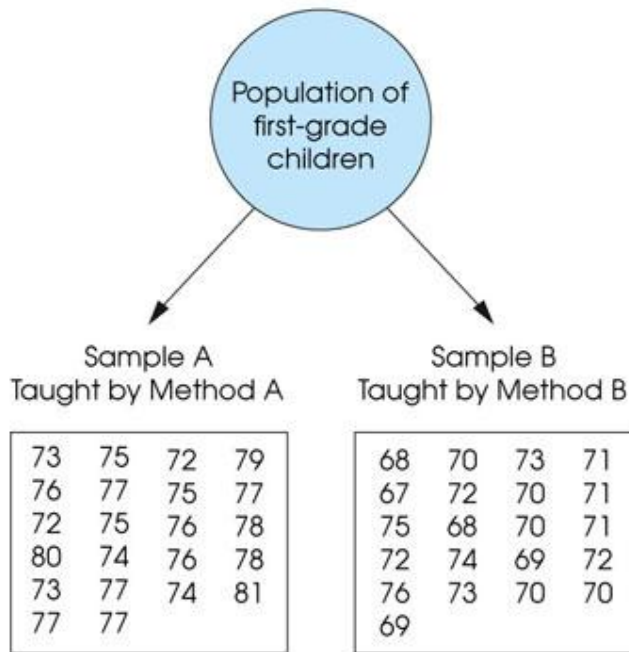
- In an **experiment**, one variable is manipulated to create treatment conditions. A second variable is observed and measured to obtain scores for a group of individuals in each of the treatment conditions. The measurements are then compared to see if there are differences between treatment conditions. All other variables are controlled to prevent them from influencing the results.
- In an experiment, the manipulated variable is called the **independent variable** and the observed variable is the **dependent variable**.
- Eg: $y=2x+3$ (variable y depends on x)

Step 1

Experiment:
Compare two
teaching methods

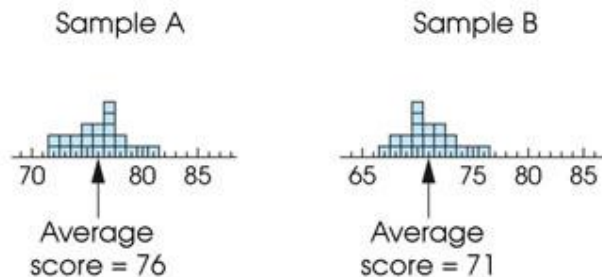
Data

Test scores for the
students in each
sample



Step 2

*Descriptive
statistics:*
Organize and simplify



Step 3

*Inferential
statistics:*
Interpret results

The sample data show a 5-point difference between the two teaching methods. However, there are two ways to interpret the results:

1. There actually is no difference between the two teaching methods, and the sample difference is due to chance (sampling error).
2. There really is a difference between the two methods, and the sample data accurately reflect this difference.

The goal of inferential statistics is to help researchers decide between the two interpretations.

Data

- The measurements obtained in a research study are called the **data**.
- The goal of statistics is to help researchers organize and interpret the data.

Descriptive Statistics

- **Descriptive statistics** are methods for organizing and summarizing data.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.

Inferential Statistics

- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Descriptive

- Organizing and summarizing data using numbers and graphs
- Data summary:
Bar graphs, histograms, Pie Charts, etc.
Shape of graph and skewness
- Measures of Central tendency:
Mean , Median and Mode
- Measures of variability:
Range, Variance and Standard Deviation

Inferential

- Using sample data to make an inference or draw a conclusion of the population
- Uses probability to determine how confident we can be that the conclusions we make are correct (Confident Intervals and Margins of Error)

Sampling Error

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics.

Ungrouped Data vs Grouped Data

Ungrouped Data – is a data with an individual value.

Grouped data - have no an individual value.

Says nothing ? Ok, let's see examples.

Frequency distribution. Ungrouped Data

- Eg: 2,3,3,5,7,7,7,7,8 □ ungrouped data

Number	f
2	1
3	2
5	1
7	4
8	1
total= 9	

□ Frequency table

Frequency distribution.

Grouped data

Eg. In the survey it has been observed that, there are 10 people with a weight between 60-79kg, 13 people between 80-99kg, 2 people between 100-119, and 1 between 120-140. Draw a frequency table.

Weight	f
60-79	10
80-99	13
100-119	2
120-140	1
	total= 26

The Mean

- **The mean** for ungrouped data, also known as the arithmetic average, is found by adding the values of the data and dividing by the total number of values. Thus,

Mean for population data:
$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Mean for sample data:
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where N – is the population size, n – is sample size, μ – (Greek letter mu) is the population mean, and \bar{x} – (read as “x-bar”) is the sample mean.

Taking a previous example.

Eg: 2,3,3,5,7,7,7,7,8

Number	f
2	1
3	2
5	1
7	4
8	1
total= 9	

□ Frequency table

sample mean =?

$$\begin{aligned}\text{sample mean} &= \text{sum} / n \text{ (or frequency)} = \\ &= [(2*1)+(3*2)+(5*1)+ (7*4)+(8*1)] / 9 = 5.44444\end{aligned}$$

The Median

- **The median** is the middle term in a data set.

The calculation of the median for ungrouped data consists of the following two steps:

1. Rank the given data set in increasing (or decreasing) order.

2. Find $\left(\frac{n+1}{2}\right)^{th}$ term in a ranked data set.

The value of $\left(\frac{n+1}{2}\right)^{th}$ term is the median.

- There are two possibilities
- 1) If n is odd, then the median is given by the value of the middle term in a ranked data.
- 2) If n is even, then the median is given by the average of the values of the two middle term.

The Mode

- The value that occurs most often in a data set is called **the mode**.

Measures of dispersion for ungrouped data

- Consider the following 2 examples:

Sample1: 66, 66, 66, 67, 67, 67, 68, 69

Sample2: 43, 44, 50, 54, 67, 90, 91, 97

The mean of sample1 is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{536}{8} = 67$$

The mean of sample2 is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{536}{8} = 67.$$

Each of these samples has a mean equal to 67. However, the dispersion of the observations in the two samples differs greatly. In the first sample all observations are grouped within 2 units of the mean. Only one observation (67) is closer than 13 units to the mean of the second sample, and some are as far away as 30 units.

Measures of dispersion

- The measures that help us to know about the spread of data set are called **the measures of dispersion**.
- The measures of central tendency and dispersion taken together give a better picture of a data set than measure of central tendency alone.
- Several quantities that are used as measures of dispersion are **the range, the mean absolute deviation, the variance, and the standard deviation**.

Range

- **The range** for a set of data is the difference between the largest and smallest values in the set.
- $\text{Range} = \text{Largest value} - \text{Smallest value}$

The mean absolute deviation

- The mean absolute deviation is defined exactly as the words indicate. The word “deviation” refers to the deviation of each member from the mean of the population.
- The term “absolute deviation” means the numerical (i.e. positive) value of the deviation, and the “mean absolute deviation” is simply the arithmetic mean of the absolute deviations.

Mean Absolute deviation (MAD)

Let $x_1, x_2, x_3, \dots, x_N$ denote the N members of a population, whose mean is μ . Their mean absolute deviation, denoted by *M.A.D.* is

$$M.A.D. = \frac{\sum_{i=1}^N |x_i - \mu|}{N}$$

For the sample of n observations, with mean \bar{x} , mean absolute deviation is defined analogously

$$M.A.D. = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

The variance and the standard deviation

- The average of the squared deviations for a data set representing a population or sample is given a special name in statistics. It is called **the variance**.
- The formula for population variance is

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

where x_i – population data

μ – population mean

N – population size .

The variance and the standard deviation

Thus, if we take the square root of the variance, we have the measure of dispersion that is known as the population standard deviation and denoted by σ . By definition we have

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

The variance and the standard deviation

The short-cut formulas for calculating variance are as follows:

$$\sigma^2 = \frac{1}{N} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{N} \right] = \frac{1}{N} \left[\sum x_i^2 - N \cdot \mu^2 \right]$$

and

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{n-1} \left[\sum x_i^2 - n \cdot \left(\bar{x} \right)^2 \right]$$

The variance and the standard deviation

Example: Find the variance and the standard deviation for the sample of 16, 19, 15, 15, and 14

Step1: Find the sum of values,

$$\sum x = 16 + 19 + 15 + 15 + 14 = 79$$

Step2: Square each value and find the sum

$$\sum x^2 = 16^2 + 19^2 + 15^2 + 15^2 + 14^2 = 1263$$

Step3: Substitute in the formula and calculate

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{4} \left[1263 - \frac{79^2}{5} \right] = 3.7$$

$$s = \sqrt{3.7} = 1.9$$

Hence the sample variance is 3.7 and sample standard deviation is 1.9.

Chebyshev's theorem

For any set of data and any $k \geq 1$, at least $100 \cdot \left(1 - \frac{1}{k^2}\right)\%$ of the values in the data set must be within plus or minus k standard deviations of the mean.

Remark:

In applying Chebyshev's theorem we treat every data set as if it were a population, and the formula for a population standard deviation is used.

Chebyshev's theorem

Example:

Let $\mu = 70$, $\sigma = 1.5$

If we let $k = 3$ from $100 \cdot \left(1 - \frac{1}{k^2}\right)\%$ we obtain that $100 \cdot \left(1 - \frac{1}{k^2}\right)\%$
 $= 100 \left(1 - \frac{1}{9}\right)\% = \frac{8}{9} \cdot 100\% = 88.89\%$.

The theorem states that at least 88.89% of data values will fall within 3 standard deviations of the mean. 88.89% of data falls within $(\mu \pm 3\sigma)$ or

$$70 + 3 \cdot 1.5 = 74.5 \text{ and}$$

$$70 - 3 \cdot 1.5 = 65.5$$

For $\mu = 70$, $\sigma = 1.5$, at 88.89% of the data values fall between 74.5, 65.5.

The interquartile range

Quartiles are the summary measures that divide a ranked data set into four equal parts. Three measures will divide any data set into four equal parts. These three measures are the first quartile (denoted by Q_1), the second quartile (denoted by Q_2), and the third quartile (denoted by Q_3). The data should be ranked in increasing order before the quartiles are determined. The quartiles are defined as follows:

$$Q_1 = \left[\frac{(n+1)}{4} \right]^{th} \quad \text{- ordered observation}$$

$$Q_3 = \left[\frac{3 \cdot (n+1)}{4} \right]^{th} \quad \text{- ordered observation.}$$

The difference between the third and the first quartiles gives the interquartile range. That is

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1.$$



Small revision

Statistic vs Parameter

Describes sample

Describes population

\bar{x} ← mean → μ

S ← Standard deviation → σ

S^2 ← Variance → σ^2

\hat{p} ← proportion → P

n ← Size → N

Mean for data with multiple-observation values

For Population:

Suppose that a data set contains values m_1, m_2, \dots, m_k occurring with frequencies, f_1, f_2, \dots, f_k respectively.

1. For a population of N observations, so that

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

Mean:

$$\mu = \frac{\sum_{i=1}^k f_i \cdot m_i}{N}$$

Mean for data with multiple-observation values

For Sample:

For a sample of n observations, so that

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

The mean is

$$\bar{x} = \frac{\sum_{i=1}^k f_i \cdot m_i}{n}$$

Mean for data with multiple-observation values

Example:

The score for the sample of 25 students on a 5-point quiz are shown below. Find the mean.

Score (m_i)	Frequency (f_i)
0	1
1	2
2	6
3	12
4	3
5	1
	$n = 25$

$$\bar{x} = \frac{\sum_{i=1}^6 f_i \cdot m_i}{n} = \frac{67}{25} = 2.68 \approx 2.7.$$

Median for data with multiple-observation values

As we know the median is $\left[\frac{n+1}{2} \right]^{th}$ observation.

Example:

Class	Number of sets sold	Frequency (month)	Cumulative frequency
1	1	3	3
2	2	8	11
3	3	5	16
4	4	4	20
5	5	2	22
6	6	1	23
7	7	1	24
		$n = 24$	

Since $n = 24$ then median = $\left[\frac{24+1}{2} \right]^{th} = \frac{12^{th} + 13^{th}}{2}$

Median for data with multiple-observation values

Class	Number of sets sold	Frequency (month)	Cumulative frequency
1	1	3	3
2	2	8	11
3	3	5	16
4	4	4	20
5	5	2	22
6	6	1	23
7	7	1	24
		$n = 24$	

- The 12th and 13th values fall in class 3. 12th value=3 ; 13th value=3.
- Therefore, Median $(3+3)/2=3$

Mode for data with multiple-observation values

Example:

The following data were collected on the number of blood tests a hospital conducted for a random sample of 50 days. Find the mode.

Number of tests per day	Frequency (days)
26	5
27	9
28	12
29	18
30	5
31	0
32	1

The mode is the most frequently occurring value. So it is 29.

Variance for data with multiple-observation values

Suppose that a data set contains values m_1, m_2, \dots, m_k occurring with frequencies, f_1, f_2, \dots, f_k respectively.

1. For a population of N observations, so that

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

The variance is

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (m_i - \mu)^2}{N} = \frac{\sum_{i=1}^k f_i \cdot m_i^2}{N} - \mu^2$$

The standard deviation is $\sigma = \sqrt{\sigma^2}$.

2. For a sample of n observations, so that

Variance for data with multiple-observation values

2. For a sample of n observations, so that

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

The variance is

$$s^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \bar{x}^2 \right]$$

The standard deviation is

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

A little bit of revision:

Ungrouped Data – is a data with an individual value.

Grouped data - have no an individual value.

Frequency distribution.

Grouped data

Eg. In the survey it has been observed that, there are 10 people with a weight between 60-79kg, 13 people between 80-99kg, 2 people between 100-119, and 1 between 120-140. Draw a frequency table.

Weight	f
60-79	10
80-99	13
100-119	2
120-140	1
	total= 26

Cumulative frequency

For any particular class, the cumulative frequency is the total number of observations in that and previous classes.

Table 1.8

Monthly earnings (in dollars)	Number of employees	Cumulative frequencies
301 to 400	4	4
401 to 500	8	12
501 to 600	16	28
601 to 700	10	38
701 to 800	7	45
801 to 900	5	50

Relative frequency

$$\text{Relative frequency of a class} = \frac{\text{frequency of that class}}{\text{sum of all frequencies}} = \frac{f_i}{\sum_{i=1}^n f_i}$$

Monthly earnings (in dollars)	Number of employees	Cumulative frequencies	Relative frequencies	Cumulative relative frequencies
301 but less than 400	4	4	4/50	4/50=0.08
401 but less than 500	8	12	8/50	12/50=0.24
501 but less than 600	16	28	16/50	28/50=0.56
601 but less than 700	10	38	10/50	38/50=0.76
701 but less than 800	7	45	7/50	45/50=0.9
800 but less than 900	5	50	5/50	50/50=1
	50		50/50=1	50/50=1

Histogram

A histogram is a graph in which classes are marked on a horizontal axis and either the frequencies are marked on the vertical axis. In a histogram, the bars are drawn adjacent to each other.

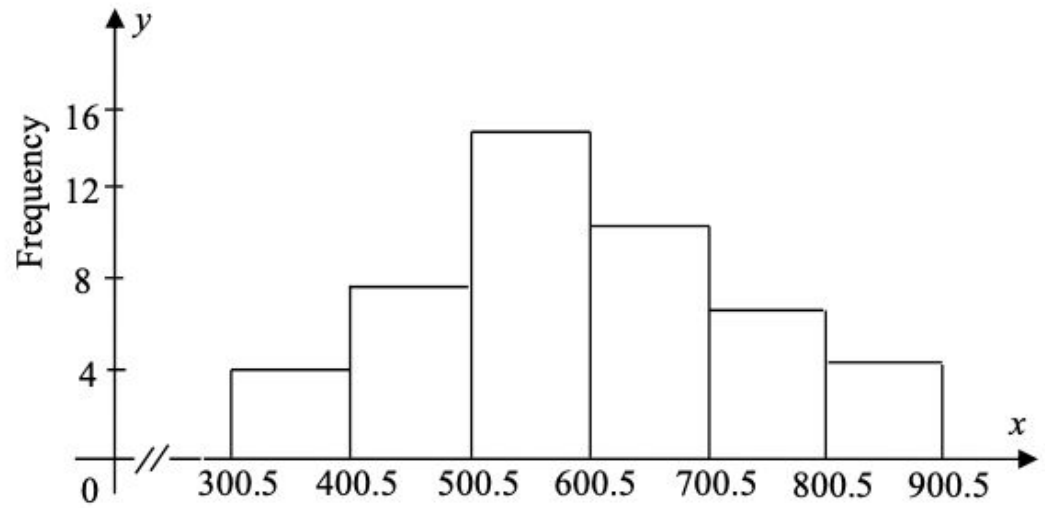


Fig. 1.1 Monthly earnings salaries frequency histogram

Mean for grouped data

Mean for the population data:

$$\mu = \frac{\sum_{i=1}^k m_i \cdot f_i}{N}$$

Mean for the sample data:

$$\bar{x} = \frac{\sum_{i=1}^k m_i \cdot f_i}{n}$$

Where m_i – is the midpoint of i^{th} class, f_i – is the frequency of i^{th} – class. k – is the total number of classes.

Example:

The following table gives the frequency distribution of daily commuting time (in minutes) from home to work for all 25 employees of a company

Daily commuting time (minutes)	Number of employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean of daily commuting time.

Solution:

Daily commuting time (minutes)	f_i	m_i	$m_i \cdot f_i$
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
	$N = 25$		$\sum_{i=1}^5 m_i \cdot f_i = 535$

$$\mu = \frac{\sum_{i=1}^5 m_i f_i}{N} = \frac{535}{25} = 21.40 \text{ minutes}$$

The Median for grouped data

$$\text{Median} = L_M + \frac{\frac{n}{2} - F_{M-1}}{f_M} \cdot C$$

L_M – lower boundary of the median class

n – number of observations

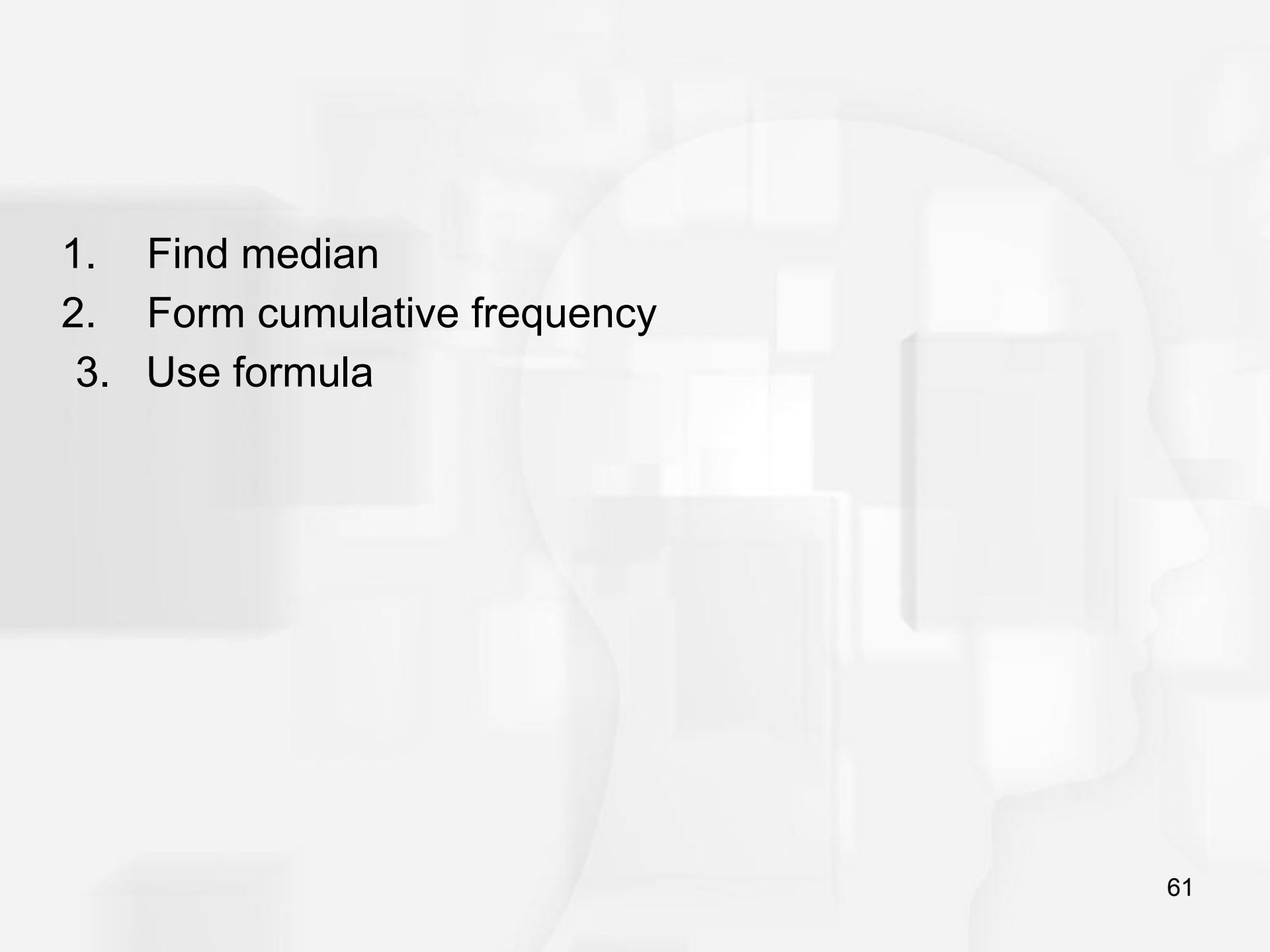
f_M – the number of observations in the median class

F_{M-1} – the number of observations in the $(M - 1)$ classes preceding the median class

C – width of the median class

Example: Find the median of the frequency distribution

Starting monthly salary(in dollars)	Frequency
900-1000	2
1000-1100	4
1100-1200	3
1200-1300	1
1300-1400	1
1400-1500	0
1500-1600	1
	$n=12$

- 
1. Find median
 2. Form cumulative frequency
 3. Use formula

1. Median = $12/2=6$
2. Cumulative frequency:

Starting monthly salary(in dollars)	Frequency	Cumulative frequency
900-1000	2	2
1000-1100	4	6
1100-1200	3	9
1200-1300	1	10
1300-1400	1	11
1400-1500	0	11
1500-1600	1	12

3. Substitute into the formula:

$$\text{Median} = L_M + \frac{\frac{n}{2} - F_{M-1}}{f_M} \cdot C$$

$$L_M = 1000; \quad n = 12; \quad F_{M-1} = 2; \quad f_M = 4; \\ C = 100$$

After substituting we get

$$\text{Median} = 1000 + \frac{6-2}{4} \cdot 100 = 1100$$

Modal class

Class	Frequency
5-10	1
10-15	2
15-20	3
20-25	7
25-30	4
30-35	3
	<i>n=20</i>

The modal class is 20-25, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence the mode could be given as 22.5.

Variance and standard deviation for grouped data

For a population of N observations, so that

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

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (m_i - \mu)^2}{N} = \frac{\sum_{i=1}^k f_i \cdot m_i^2}{N} - \mu^2$$

The standard deviation is $\sigma = \sqrt{\sigma^2}$.

Variance and standard deviation for grouped data

For a sample of n observations, so that $n = \sum_{i=1}^k f_i$

$$s^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \left(\bar{x} \right)^2 \right]$$

The standard deviation is $s = \sqrt{s^2}$