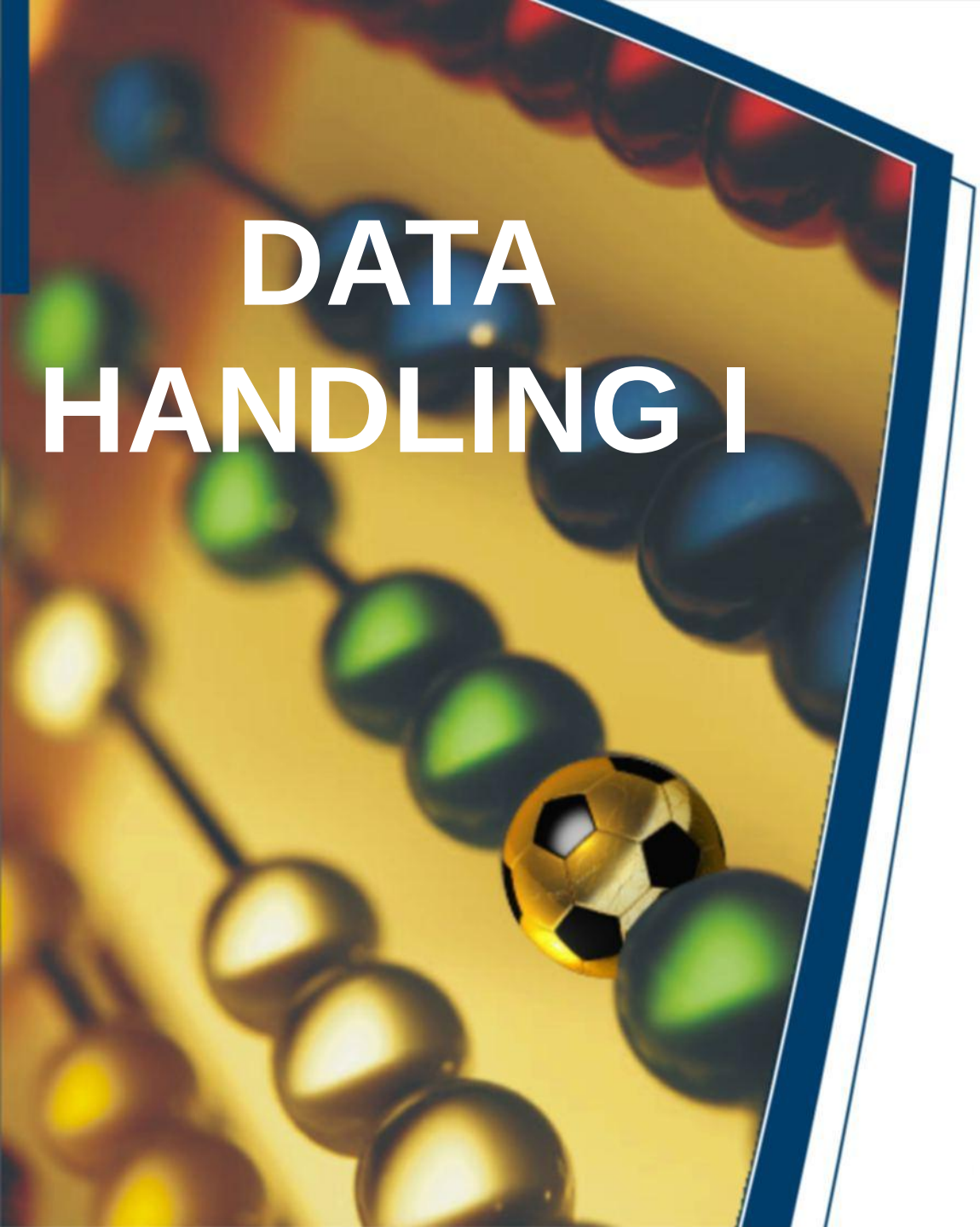


DATA HANDLING I

Nelson Mandela Metropolitan University
for tomorrow

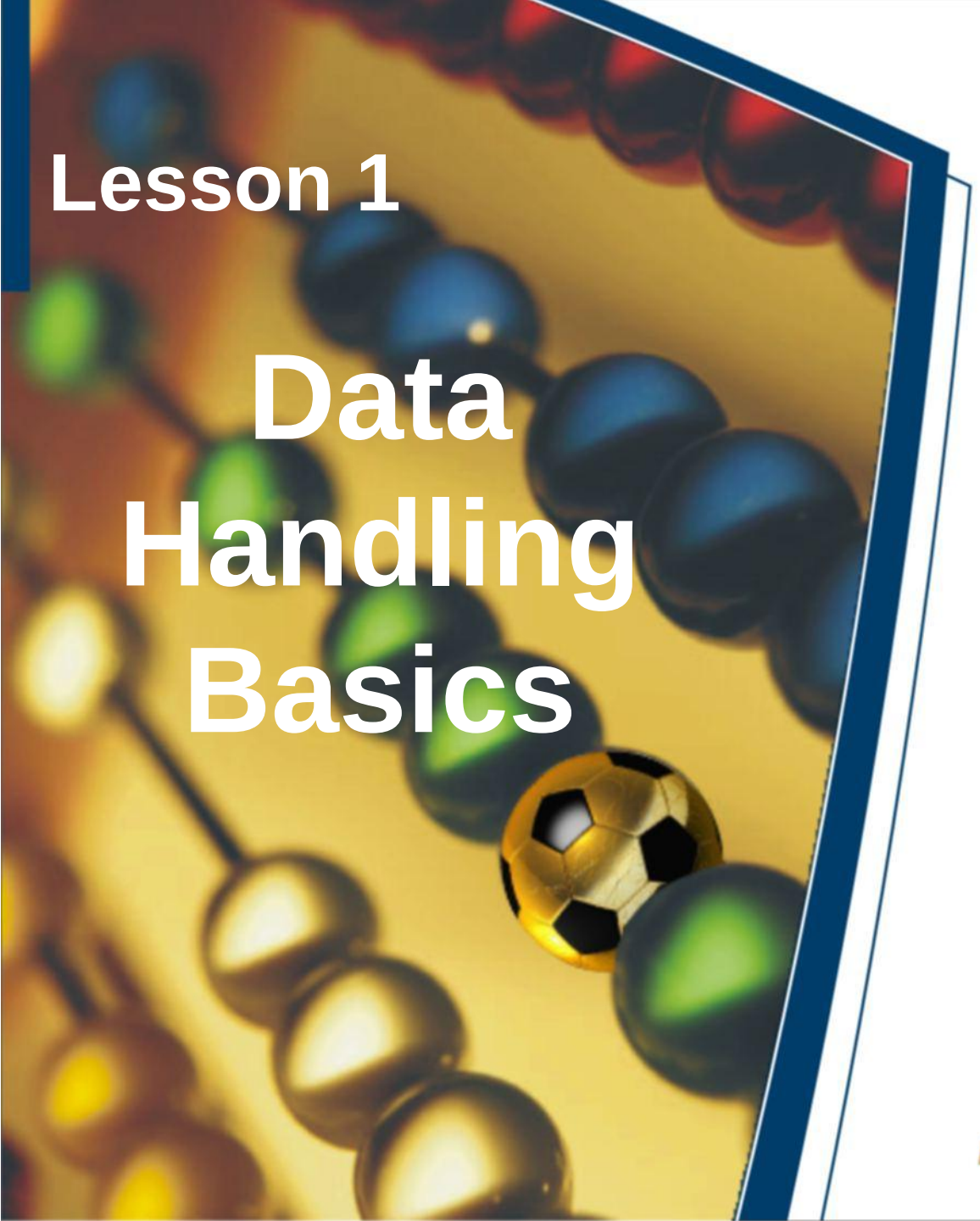
NCS Mathematics DVD Series

GOVAN MBEKI
MATHEMATICS DEVELOPMENT UNIT
Empowering young minds

Outcomes for this DVD

In this DVD we will:

- Review the basics of Data Handling (Statistics). Lesson 1
- Discuss ways to organise collected univariate numerical data. Lesson 2
- Discuss ways to represent organised univariate numerical data. Lesson 3
- Discuss ways to analyse and interpret univariate numerical data. Lesson 4



Lesson 1

Data Handling Basics

Nelson Mandela Metropolitan University
for tomorrow

NCS Mathematics DVD Series

GOVAN MBEKI MATHEMATICS DEVELOPMENT UNIT
Empowering young minds

What is Statistics?

Data Handling, more generally known as Statistics,
is one of the most widely-applied branches of mathematics.

Statistics refers to an action
that is performed on Data
in order to obtain information.

Important Stages in Statistics

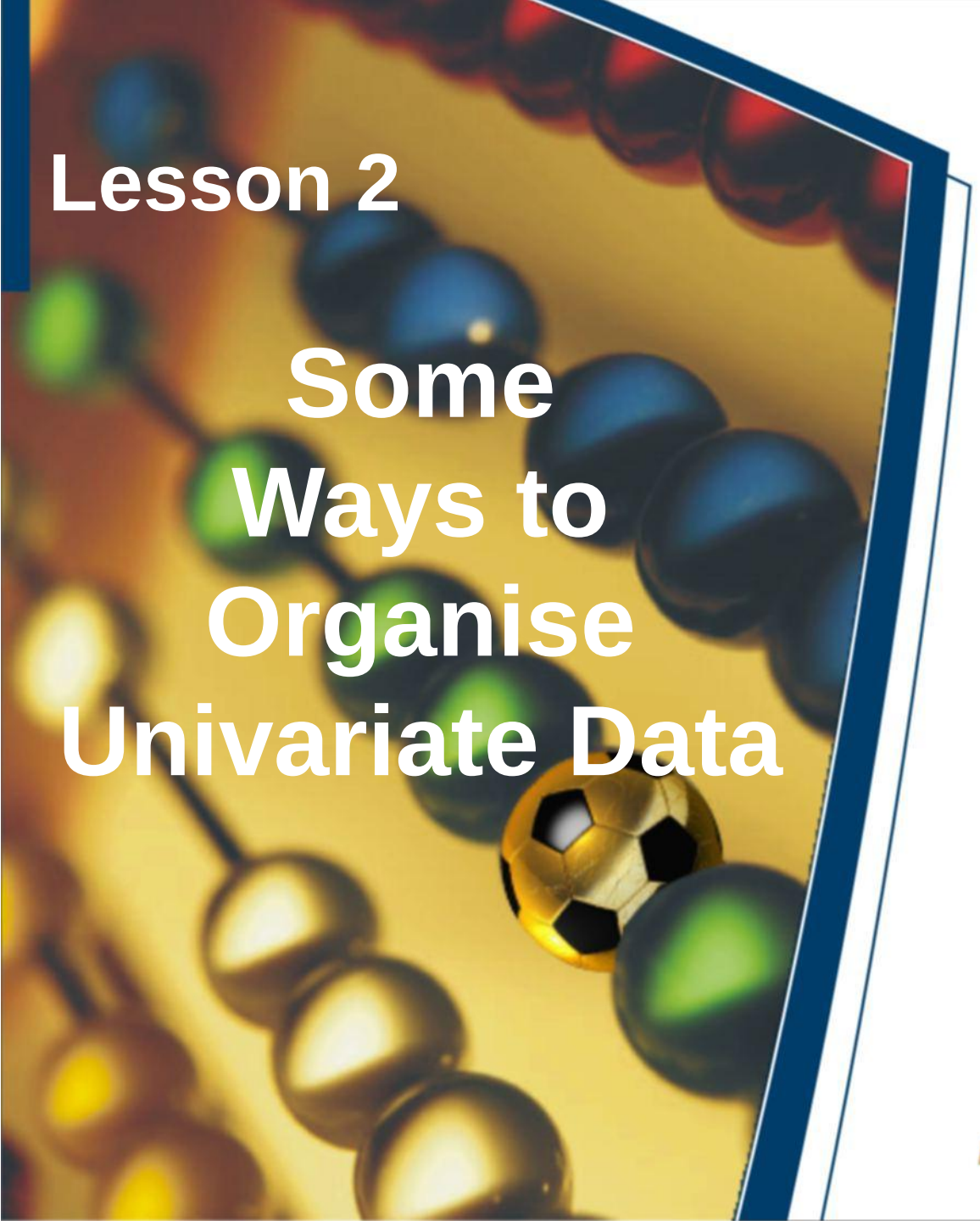
- Many important issues often confront the citizens of a country.
- When addressing these issues, appropriate data should be collected so that informed decisions could be made.
- Data collected should be organised and represented in such a way that it conveys a meaningful message.
- Data organised and represented in meaningful ways can be analysed to identify general trends and features.
- General trends and features can be be utilised to interpret data in order to draw conclusions and make predictions.

Local Issues and the collection of data

- Ordinary members of the public are often confronted by a range of issues.
- These may be human rights, social, economical, political or environmental issues.
- It is important to collect relevant data about the issues that affect them.
- This data may be used by the citizens to solve the issues themselves or can be taken to relevant authorities, such as the police or municipality, for them to address the issue.
- For anyone of these issues one should determine:
  The problem - Data to be collected - How to process data - How to display the data - How to interpret the data.

Ways to collect data

- Examining and interpreting figures, graphs, tables that appear in newspapers, magazines and journals.
- Collect information from radio, television or other media.
- Conduct surveys by means of questionnaires or interviews.
- Consultation with professional or other experts in the field.
- Conduct own or group-based experiments and make notes of results in tables and graphs.
- Use Internet search engines like Google, Yahoo or others.
- Use relevant books from the library or private collections.
- Gather information from government departments or agencies.
- Get information from non-governmental agencies or international agencies (e.g. United Nations).



Lesson 2

Some Ways to Organise Univariate Data

Nelson Mandela Metropolitan University
for tomorrow

NCS Mathematics DVD Series

GOVAN MBEKI MATHEMATICS DEVELOPMENT UNIT
Empowering young minds

Definitions of Basic Terms

Population: Collection of all potential observations that can be found in a given situation.

Sample: Collection of observations representing only a portion of the population.

Variable: A variable x has a value x_i for data item (observation) number i.

The Arrangement of Data

- Ungrouped Data or Raw Data is a collection of observations that has not been arranged in any meaningful fashion.

- Grouped Data is an arrangement whereby raw data have been grouped into a series of **equally spaced classes**.

The number of observations falling in each class interval is called the **class frequency**.

The resulting table is called the **frequency distribution table**.

How to organise Ungrouped or Raw Data

- Ungrouped Data or Raw Data
- Not arranged in any meaningful fashion

Example : The number of SMS calls received (variable *x*) in a certain day by 12 students may be recorded as: 0;3;6;5;2;5;4;8;3;5;5 and 7.

For further analysis by hand or PC the set of raw data is usually arranged in an **ascending** or **descending** order.

Organised in a descending order:

8 7 6 5 5 5 5 4 3 3 2 0

Organised in an ascending order:

0 2 3 3 4 5 5 5 5 6 7 8

How to organise Raw Data as Grouped Data

Example: Group the 12 SMS observations in a frequency table of 3 classes: 0; 2; 3; 3; 4; 5; 5; 5; 5; 6; 7; 8

Class No	No of SMSs	Tally	Frequency (f_i)
1	0 to less than 3	//	2
2	3 to less than 6	𝍸 //	7
3	6 to less than 9	///	3

Number of Classes: $\sqrt{n}$ and $\sqrt{12} \simeq 3,464$

$$\sum_{i=1}^{3} f_i = 12$$

- Limits describes the interval of each class. Lower limit of class 2 is 3 and the upper limit of class 3 is 9.
- Interval notation $[l, m)$ or $(l, m]$ can be used to describe each interval. l is the lower limit and m the upper limit.

Another example of a frequency table:

Class i	Mass (mg)	Frequency (f_i)
1	$[10;15)$	8
2	$[15;20)$	13
3	$[20;25)$	22
4	$[25;30)$	17
5	$[30;35)$	14
		$n=\sum_{i=1}^{5} f_i = 74$

Some questions:

1) How many classes?

2) What is the class width?

3) What is the lower limit of class 3?

4) What is the upper limit of class 4?

5) How many observations in class 3?

6) How many observations in the sample?

Tutorial 1: Organise Raw Data

Suppose that 35 students obtained the following marks out of 50 for a mathematics test:

46	40	12	10	47	23	26
8	29	34	37	17	40	50
18	23	33	23	24	15	35
23	19	22	28	35	27	42
29	26	46	33	27	19	28

PAUSE DVD

- Do Tutorial 1
- Then View the Solutions

1) Arrange the raw data in ascending order.

2) Group the raw data in a frequency table of 5 classes.

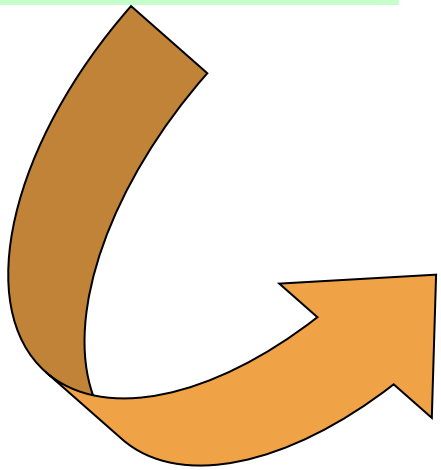
Number of Classes: $\sqrt{n} = \sqrt{35} \simeq 5,916$

Tutorial 1: Problem 1: Suggested Solution

Organise Raw Data in ascending order

1) Arrange the raw data in ascending order:

46	40	12	10	47	23	26
8	29	34	37	17	40	50
18	23	33	23	24	15	35
23	19	22	28	35	27	42
29	26	46	33	27	19	28



8	10	12	15	17	18	19
19	22	23	23	23	23	24
26	26	27	27	28	28	29
29	33	33	34	35	35	37
40	40	42	46	46	47	50

Tutorial 1: Problem 2: Suggested Solution

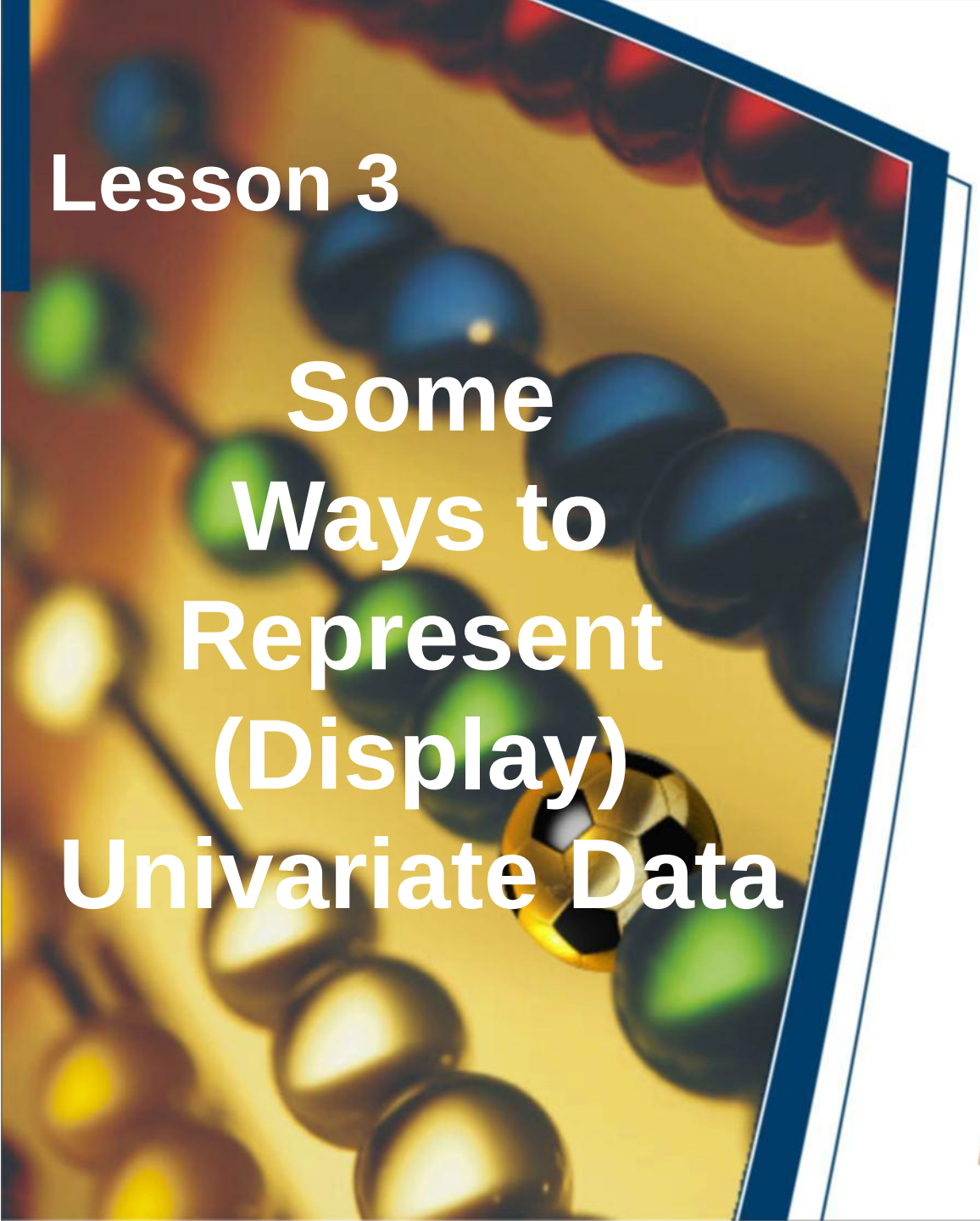
Organise Raw Data in a Frequency Table

2) Group the raw data in a frequency table of 5 classes:

Class i	Test Mark	Tally	Frequency (f_i)
1	$(0;10]$	//	2
2	$(10;20]$	///// /	6
3	$(20;30]$	///// ///// ////	14
4	$(30;40]$	///// ///	8
5	$(40;50]$	/////	5
			$\sum_{i=1}^{5} f_i = 35$

8 10 12 15 17 18 19 19 22 23 23 23 23 24 26 26 27 27

28 28 29 29 33 33 34 35 35 37 40 40 42 46 46 47 50



Lesson 3

Some Ways to Represent (Display) Univariate Data

Nelson Mandela Metropolitan University
for tomorrow

NCS Mathematics DVD Series

GOVAN MBEKI MATHEMATICS DEVELOPMENT UNIT
Empowering young minds

Display of Ungrouped Data by means of a Stem-and-leaf plot

- A stem-and-leaf plot is useful for displaying relative small sets of data.
- It is a quick and easy way to order and display a set of data.
- Leading digit(s) form the stem and trailing digit becomes the leaf.

Example : Transform the data into a stem-and-leaf plot.

3 17 1 29 7 35 10 37 2 25

0	3 1 7 2
1	7 0
2	9 5
3	5 7

Unordered

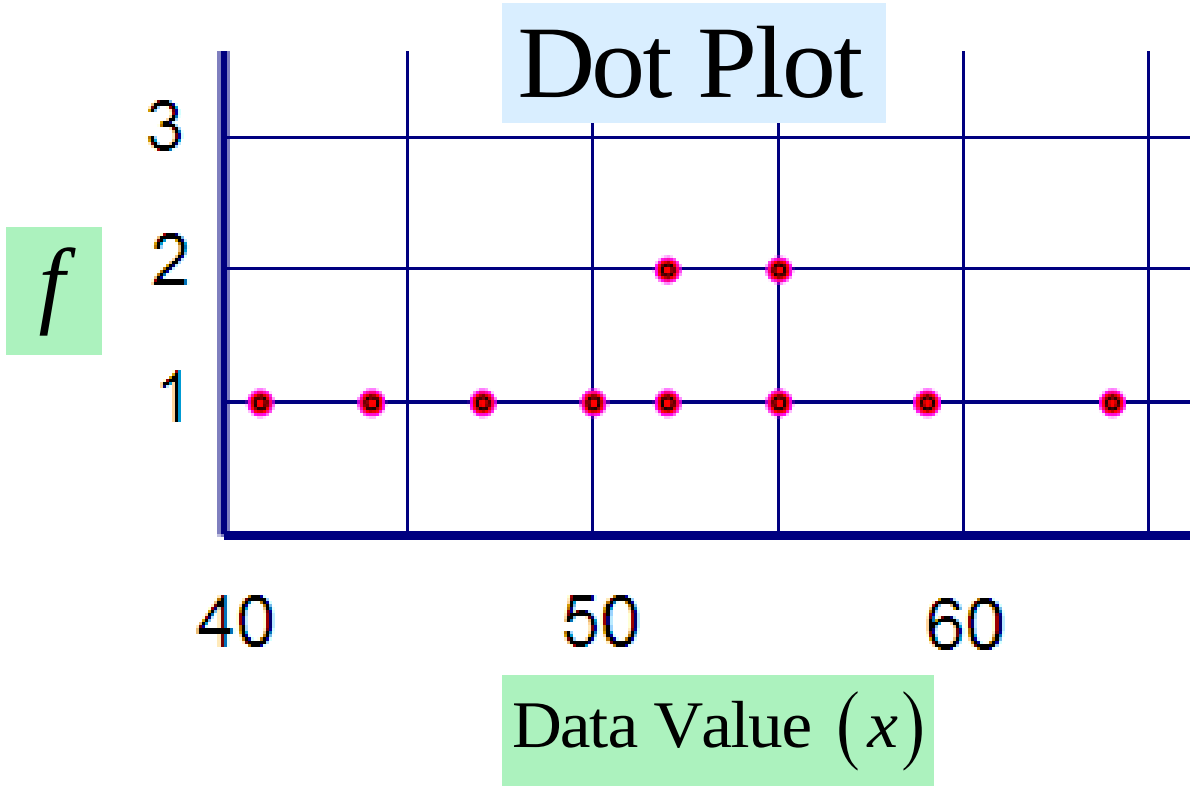
0	1 2 3 7
1	0 7
2	5 9
3	5 7

Ordered

Display of Ungrouped Data by means of a Dot Plot

Draw a dot plot for the values

42; 44; 47; 50; 52; 52; 55; 55; 59 and 64.



Tutorial 2: Display Ungrouped Data

Consider the test marks 35 students obtained for a mathematics test out of 50.

8	10	12	15	17	18	19
19	22	23	23	23	23	24
26	26	27	27	28	28	29
29	33	33	34	35	35	37
40	40	42	46	46	47	50

PAUSE DVD

- Do Tutorial 2
- Then View Solutions

1) Construct an ordered stem-and-leaf plot for the data.

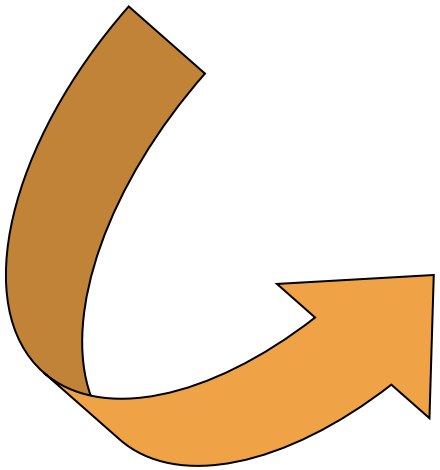
2) Construct a Dot Plot for the given data.

Tutorial 2: Problem 1: Suggested Solution

Construction of Stem-and-leaf plot

1) Construct an ordered stem-and-leaf plot for the data.

8	10	12	15	17	18	19
19	22	23	23	23	23	24
26	26	27	27	28	28	29
29	33	33	34	35	35	37
40	40	42	46	46	47	50



0	8												
1	0	2	5	7	8	9	9						
2	2	3	3	3	3	4	6	6	7	7	8	8	9	9
3	3	3	4	5	5	7							
4	0	0	2	6	6	7							
5	0												

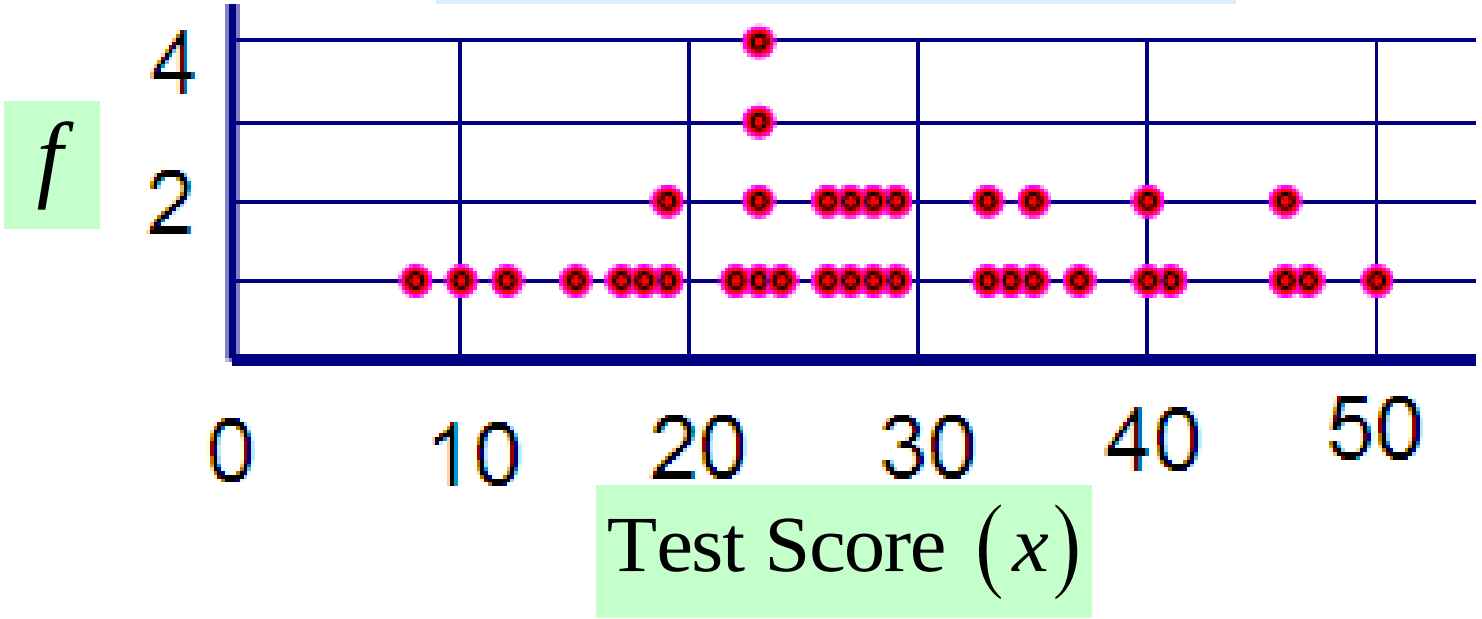
Tutorial 2: Problem 2: Suggested Solution

Construction of a Dot plot

2) Construct a dot plot for the data.

0	8													
1	0	2	5	7	8	9	9							
2	2	3	3	3	3	4	6	6	7	7	8	8	9	9
3	3	3	4	5	5	7								
4	0	0	2	6	6	7								
5	0													

Test Scores for 35 students



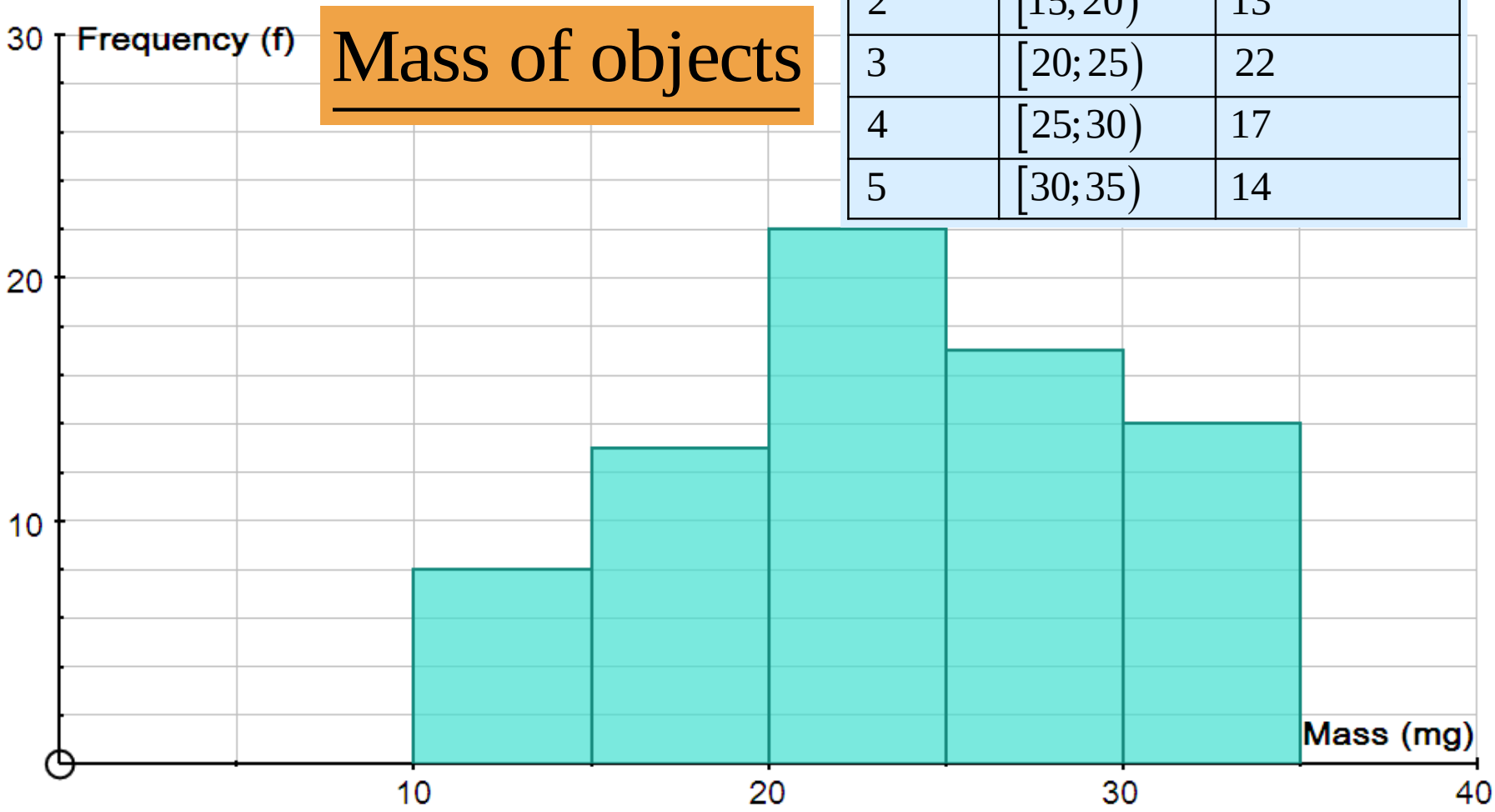
Display of Grouped Data

- **Histogram**
  - **Construct a column on each class interval.**
  - **Vertical height matching the frequency.**
- **Frequency polygon**
  - **Plot a dot above midpoint of each interval (Value called the class mark).**
  - **Dots are connected by straight lines.**
  - **Anchoring the polygon on the horizontal axis.**
- **Histogram and frequency polygon can both be displayed on the same graph.**

Histogram: Display of Grouped Data

- Construct a column on each class interval.
- Vertical height matching the frequency.
- Lable histogram.

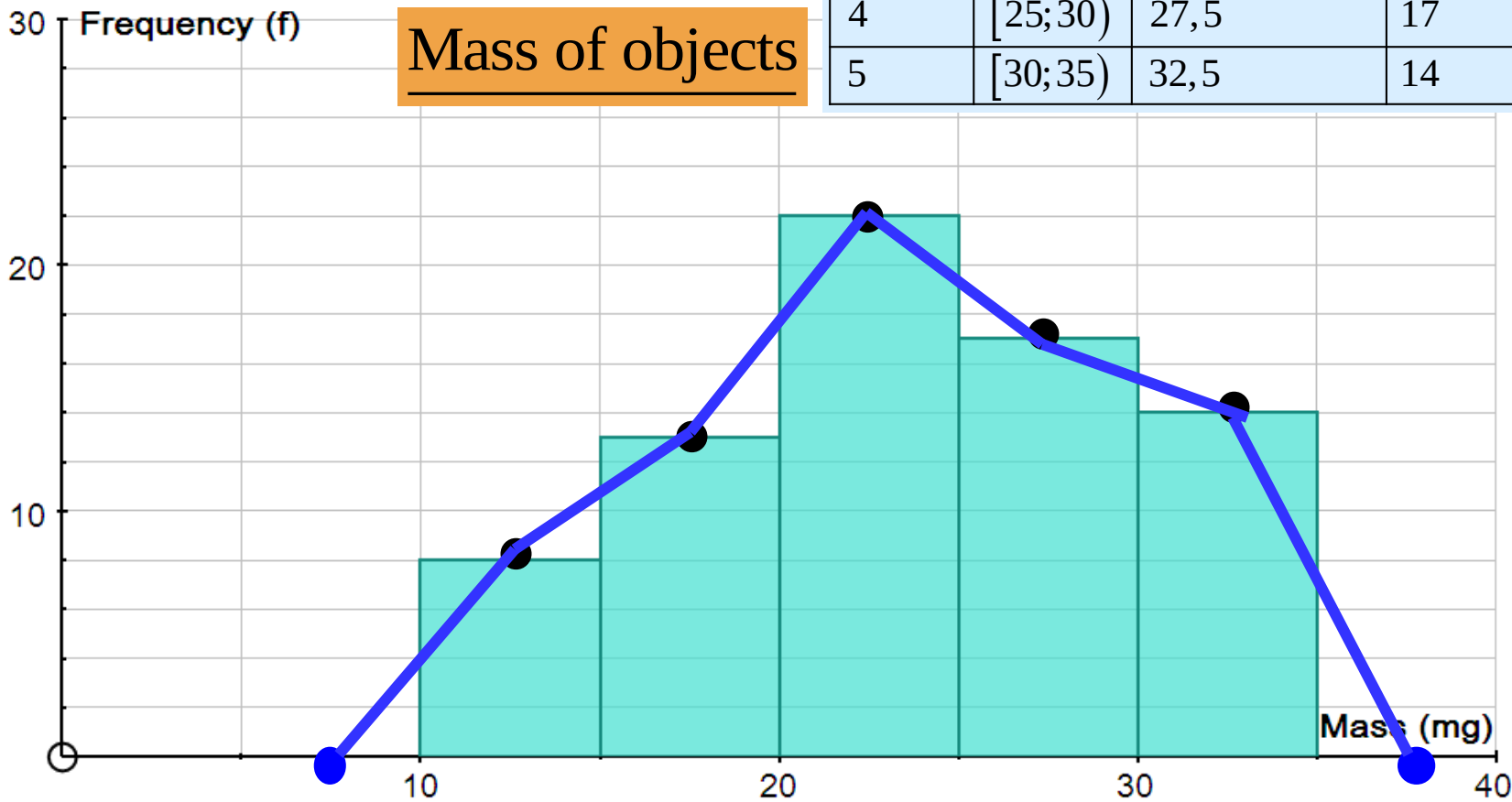
Class i	Interval Mass (mg)	Frequency f_i
1	$[10;15)$	8
2	$[15;20)$	13
3	$[20;25)$	22
4	$[25;30)$	17
5	$[30;35)$	14



Frequency Polygon and Histogram (Displayed on same graph: Grouped Data)

- Plot a dot above midpoint of each interval.
- Connect dots by straight lines.
- Anchor polygon at 7,5 and 37,5.
- Label frequency polygon.

Class i	Interval	Class Mark x_i	Frequency f_i
1	$[10;15)$	12,5	8
2	$[15;20)$	17,5	13
3	$[20;25)$	22,5	22
4	$[25;30)$	27,5	17
5	$[30;35)$	32,5	14



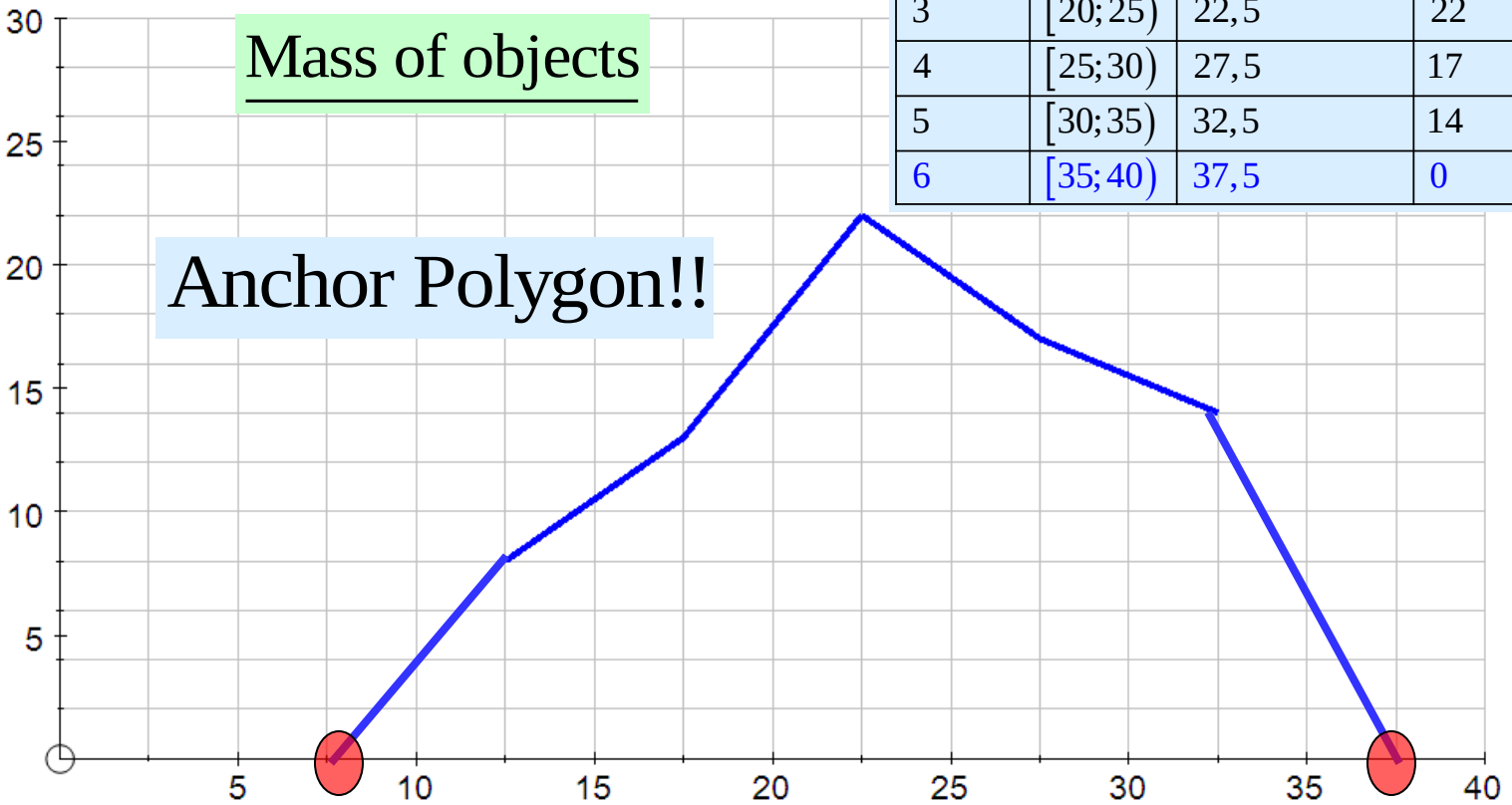
Frequency Polygon: Display of Grouped Data

Use of technology (Autograph)

Class i	Interval	Class Mark x_i	Frequency f_i
0	$[5;10)$	7,5	0
1	$[10;15)$	12,5	8
2	$[15;20)$	17,5	13
3	$[20;25)$	22,5	22
4	$[25;30)$	27,5	17
5	$[30;35)$	32,5	14
6	$[35;40)$	37,5	0

Mass of objects

Anchor Polygon!!



Tutorial 3: Histogram and Frequency Polygon

Test marks obtained by 35 students for a mathematics test out of 50 are summarised in this frequency table.

Class i	Interval	Class Mark (x_i)	Frequency (f_i)
1	$(0;10]$	5	2
2	$(10;20]$	15	6
3	$(20;30]$	25	14
4	$(30;40]$	35	8
5	$(40;50]$	45	5
			$\sum_{i=1}^{5} f_i = 35$

PAUSE DVD

- Do Tutorial 3
- Then View Solutions

1) Construct a Histogram for the data.

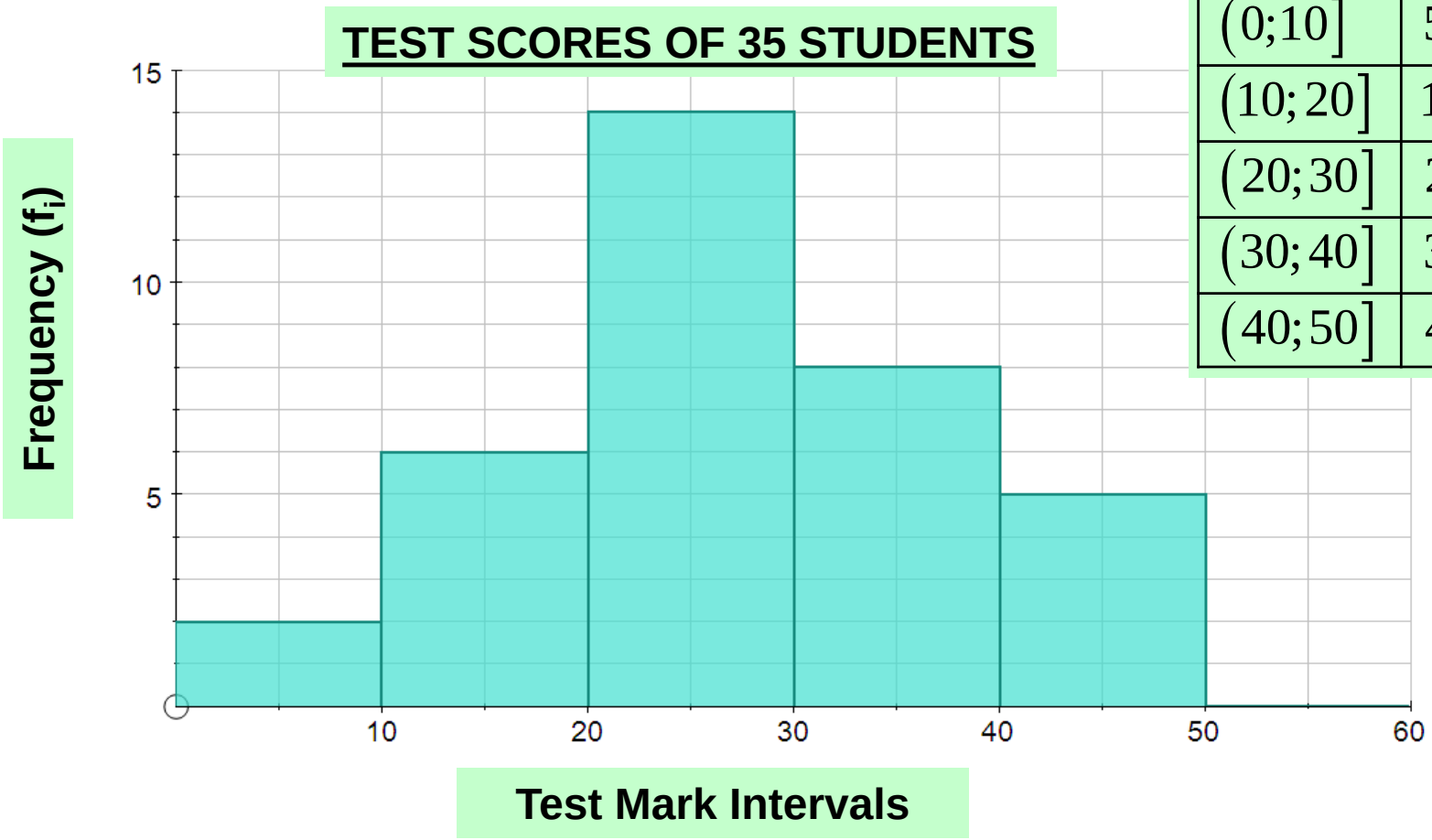
2) Construct a Frequency Polygon for the data.

Tutorial 3: Problem 1: Suggested Solution

Histogram of Grouped Data

1) Construct a Histogram.

Interval	x_i	f_i
$(0;10]$	5	2
$(10;20]$	15	6
$(20;30]$	25	14
$(30;40]$	35	8
$(40;50]$	45	5

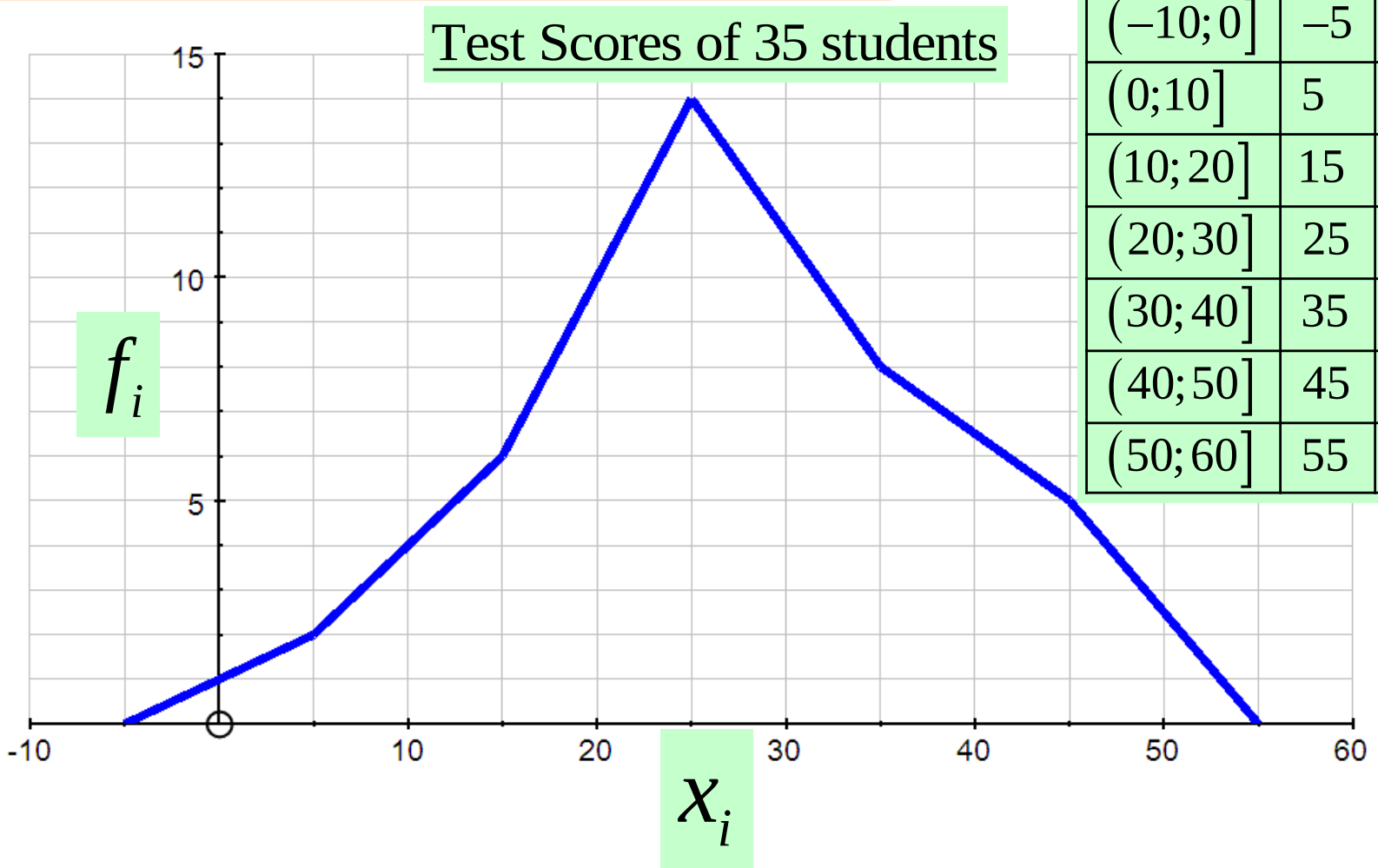


Tutorial 3: Problem 2: Suggested Solution

Frequency Polygon for Grouped Data

2) Construct a Frequency Polygon.

Interval	x_i	f_i
$(-10;0]$	-5	0
$(0;10]$	5	2
$(10;20]$	15	6
$(20;30]$	25	14
$(30;40]$	35	8
$(40;50]$	45	5
$(50;60]$	55	0



Relative and Cumulative Frequency Distributions

- Helpful to expand frequency distribution table by expressing frequency of each class as a relative frequency (%).
- Relative frequency (%) of a class $i = \frac{f_i}{\sum f_i} \times 100\%$.
- Cumulative frequency column is usually added.
- Relative cumulative frequency is formed in similar way.

Relative and Cumulative Frequency Distributions

A	B	C	D	E	F	G
Class i	Score	CM (x_i)	f_i	rel f_i (%)	cum f_i	rel cum f_i (%)
1	$[15;20)$	17,5	6	7,5	6	7,50
2	$[20;25)$	22,5	13	16,25	19	23,75
3	$[25;30)$	27,5	20	25	39	48,75
4	$[30;35)$	32,5	25	31,25	64	80,00
5	$[35;40)$	37,5	12	15	76	95,00
6	$[40;45)$	42,5	4	5	80	100,00
			$n = \sum f_i = 80$	$\sum = 100\%$		

Interpret Relative Cumulative Frequency values!

Note

1) Sample size $= n = \sum f_i$

2) Cumulative frequency of last class matches n

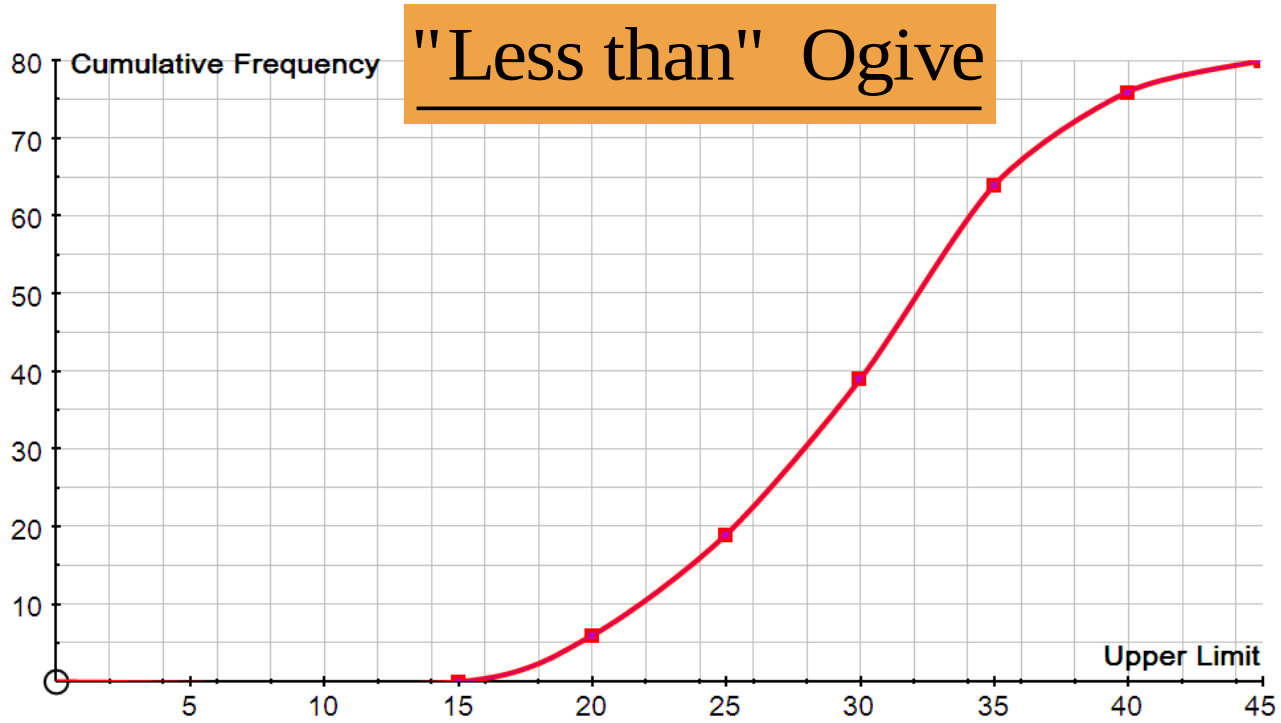
3) Sum of relative fequencies is always 100%

4) Cumulative relative frequency of last class is always 100%

“Less than” cumulative frequency polygon or “Less than” ogive

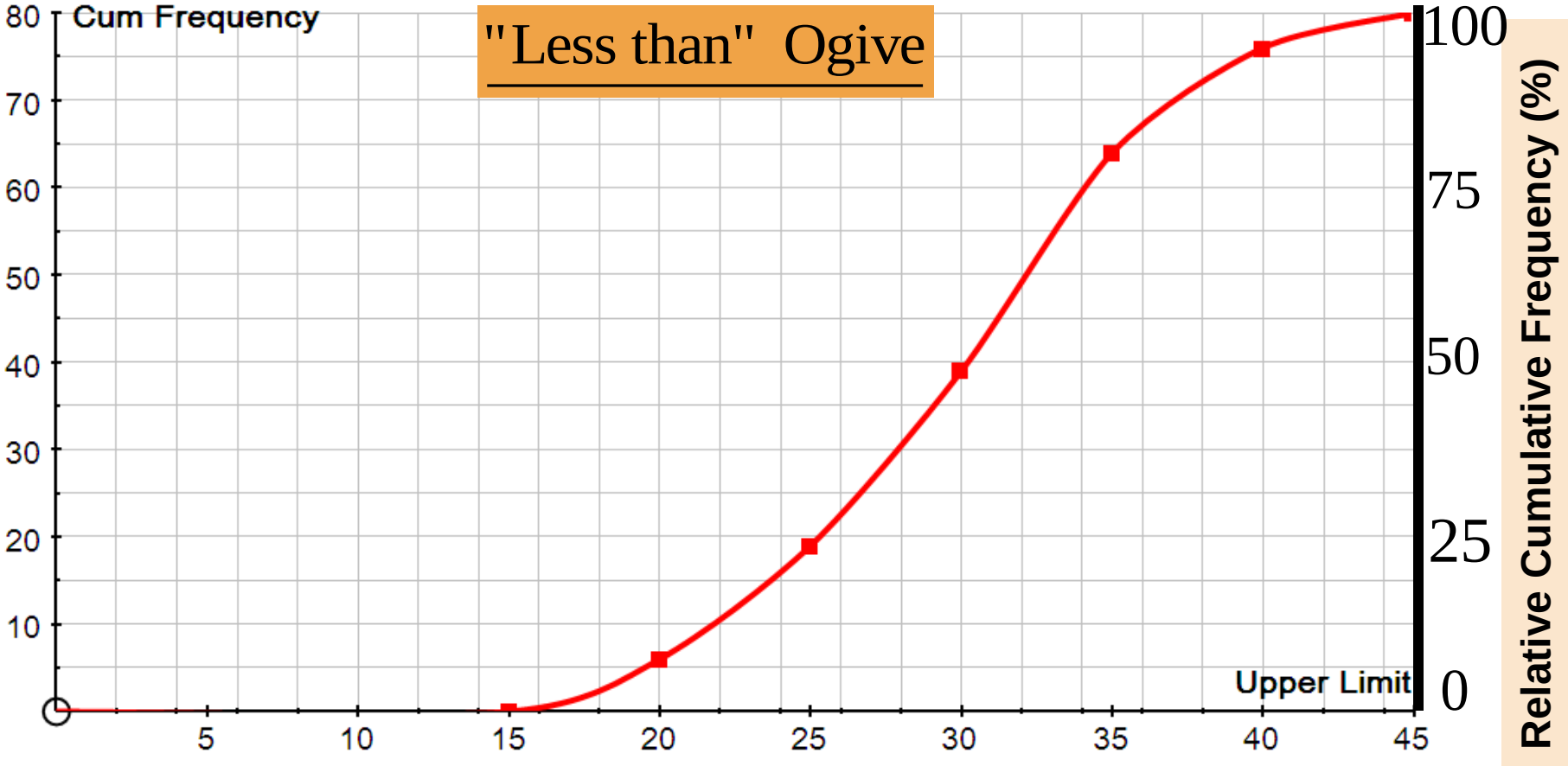
Plot **cumulative frequencies** (Vertical axis) against **upper class limits** (Horizontal axis)

Interval	[10,15)	[15;20)	[20;25)	[25;30)	[30;35)	[35;40)	[40;45)
Cum frequency	0	6	19	39	64	76	80
Upper Class Limit	15	20	25	30	35	40	45



“Less than” Cumulative and Relative Cumulative Frequency Polygons on the same system of axis

Interval	[10,15)	[15;20)	[20;25)	[25;30)	[30;35)	[35;40)	[40;45)
Cum frequency	0	6	19	39	64	76	80
Upper Class Limit	15	20	25	30	35	40	45
Relative Cumulative frequency	0,00	7,50	23,75	48,75	80,00	95,00	100,00



Tutorial 4 : Relative and Cumulative Frequency Distributions

"Less than" Cumulative and Relative Cumulative Frequency Ogives

Test marks obtained by 35 students for a mathematics test out of 50 are summarised in this frequency table.

Class i	Interval	Upper Class	f_i	Rel f_i	Cum f_i	Rel Cum f_i
0	$(-10;0]$	0	0			
1	$(0;10]$	10	2			
2	$(10;20]$	20	6			
3	$(20;30]$	30	14			
4	$(30;40]$	40	8			
5	$(40;50]$	50	5			
			$\sum f_i = 35$			

PAUSE DVD

- Do Tutorial 4
- Then View Solutions

1) Complete the table.

2) Construct the "Less than" Cumulative and Relative Cumulative Frequency Ogives on the same system of axis.

Tutorial 4: Problem 1: Suggested Solution
Relative and Cumulative Frequency Distributions

1) Complete the table.

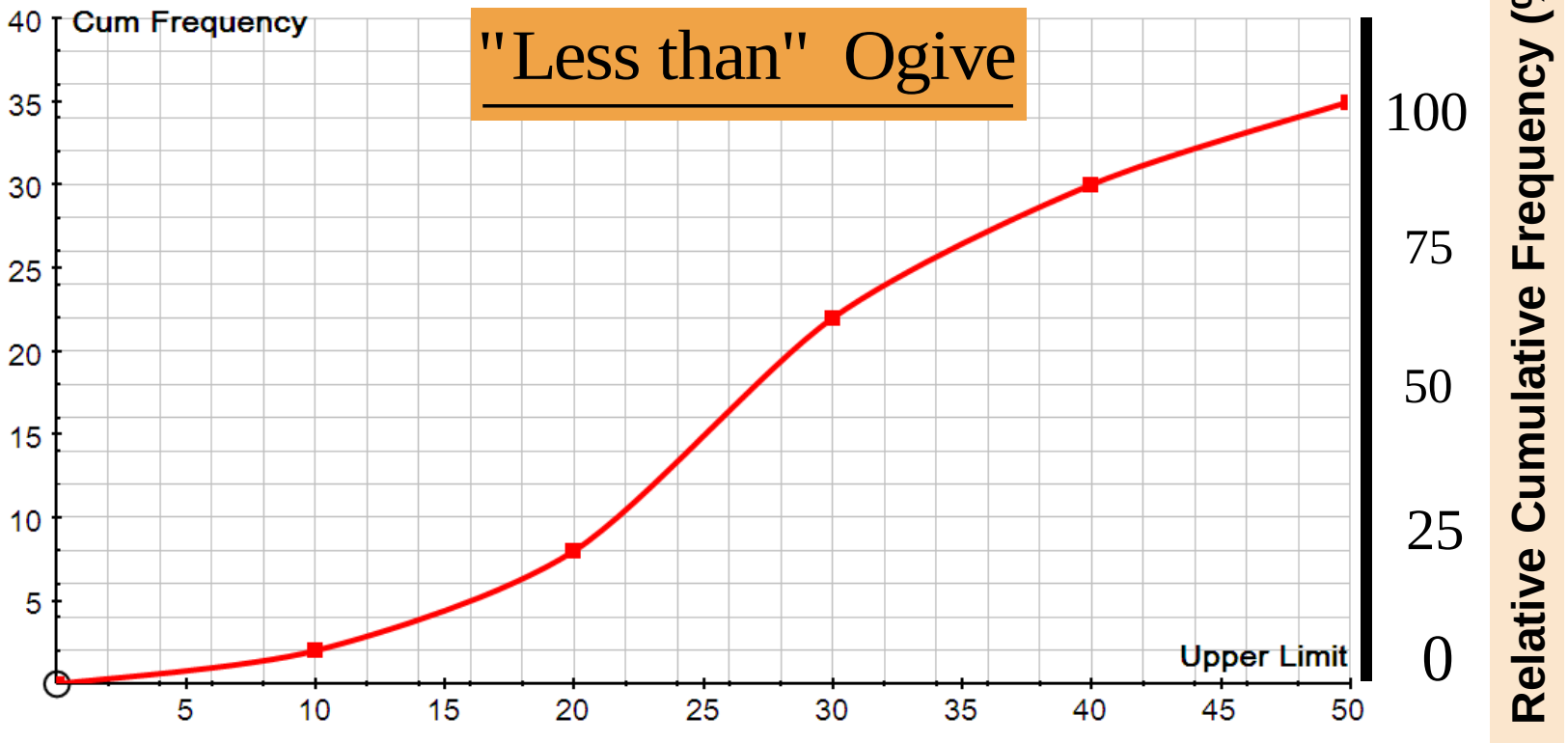
Class i	Interval	Upper Class	f_i	Rel f_i (%)	Cum f_i	Rel Cum f_i (%)
0	$(-10;0]$	0	0	0,00	0	0,00
1	$(0;10]$	10	2	5,71	2	5,71
2	$(10;20]$	20	6	17,14	8	22,85
3	$(20;30]$	30	14	40,00	22	62,85
4	$(30;40]$	40	8	22,86	30	85,71
5	$(40;50]$	50	5	14,29	35	100,00
			$n = 35$	$\sum = 100\%$		

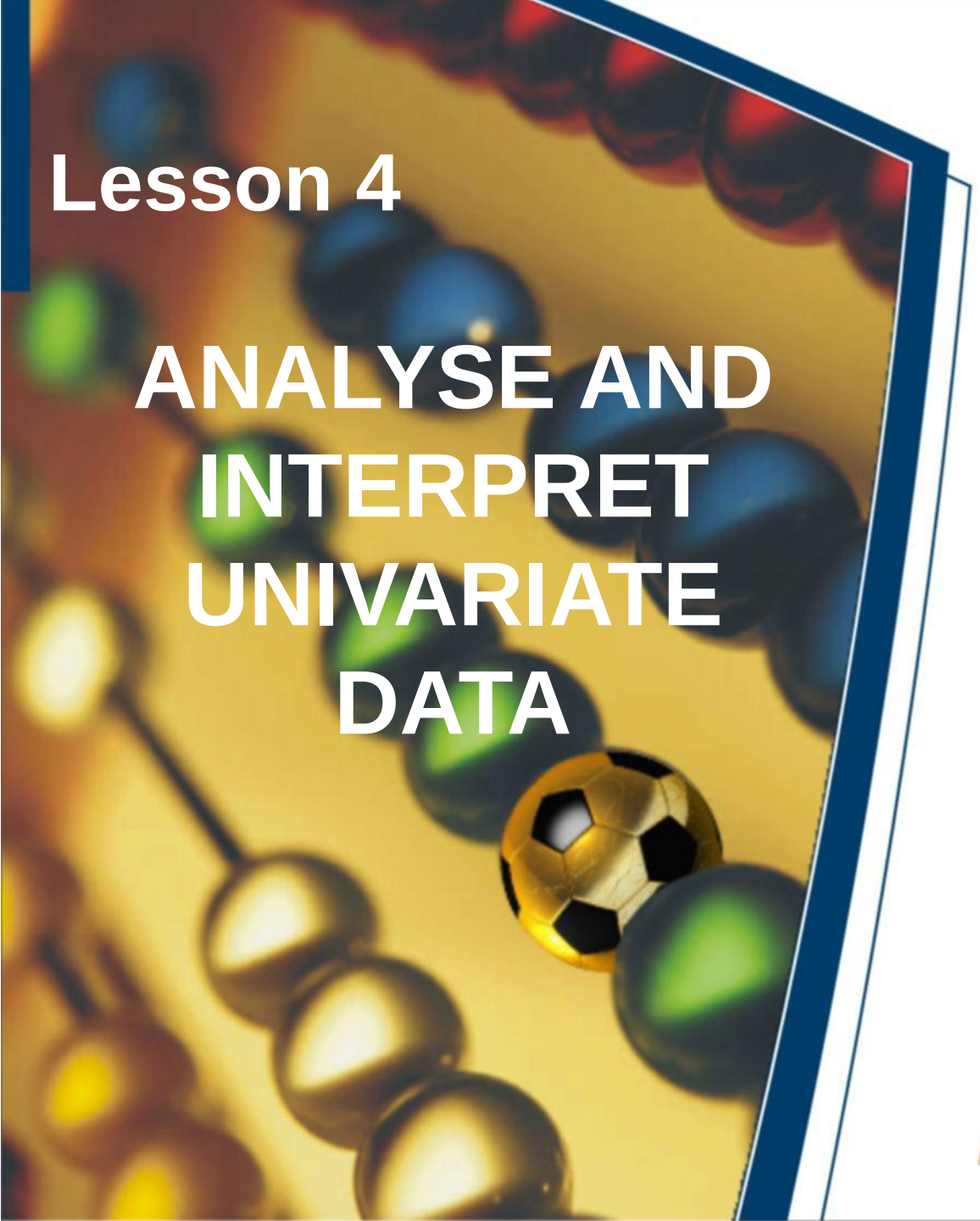
Interpret Relative Cumulative Frequency values!

Tutorial 4: Problem 2: Suggested Solution

"Less than" Cum and Rel Cum Freq Ogives

Interval	(–10;0]	(0;10]	(10;20]	(20;30]	(30;40]	(40;50]
Cum frequency	0	2	8	22	30	35
Upper Class Limit	0	10	20	30	40	50
Relative Cumulative frequency	0,00	5,71	22,85	62,85	85,71	100,00





Lesson 4

ANALYSE AND INTERPRET UNIVARIATE DATA

Nelson Mandela Metropolitan University
for tomorrow

NCS Mathematics DVD Series

GOVAN MBEKI
MATHEMATICS DEVELOPMENT UNIT
Empowering young minds

Measures of Central Tendency

Will consider the following three Measures of Central Tendency:

- Mean: Average of observations
- Median: Middle Value
- Mode: Most frequently occuring observation.

How to calculate the Mean: $\bar{x}$

Calculated by adding the observed values x_i and divide the sum by the number of observations n.

- Ungrouped data: $\bar{x} = \frac{\sum x_i}{n}$

- Grouped data: $\bar{x} = \frac{\sum (x_i \cdot f_i)}{n}$

Mean of Ungrouped Data

Calculate $\bar{x}$ for the observations 7; 13; 15; 17; 20 and 21.

- Ungrouped data: $\bar{x} = \frac{\sum x_i}{n}$

$$\therefore \bar{x} = \frac{7+13+15+17+20+21}{6}$$

$$= \frac{93}{6} = 15,5$$

Calculating the Median for Ungrouped Data

- Median (M_e) is the middle value.
- Location: $\left(\frac{n+1}{2}\right)^{\text{th}}$ position, provided data is ordered.

n odd: 7, 13, 14, 17, 20

Location of $M_e = \left(\frac{5+1}{2}\right)^{th} = 3^{rd}$ position

$\therefore M_e = 14$

n even: 7, 13, 14, 17, 20, 21

- Location of M_e ? $\frac{6+1}{2} = 3,5^{th}$ position
- Calculate $M_e = \frac{14+17}{2} = 15,5$

Calculating the Mode for Ungrouped Data

The mode is the **most frequently** occuring observation.

For ungrouped data:

M_o can be found by an inspection of the observations.

Consider the ordered ungrouped data

3; 5; 12; 12 and 13.

Mode: $M_o = 12$

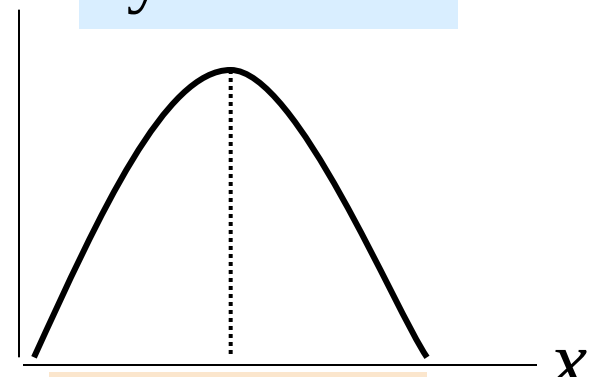
There can be more than one mode.

5 6 6 6 7 9 9 9 10

Comparison between mean $(\bar{x})$, median (M_e) and mode (M_o)

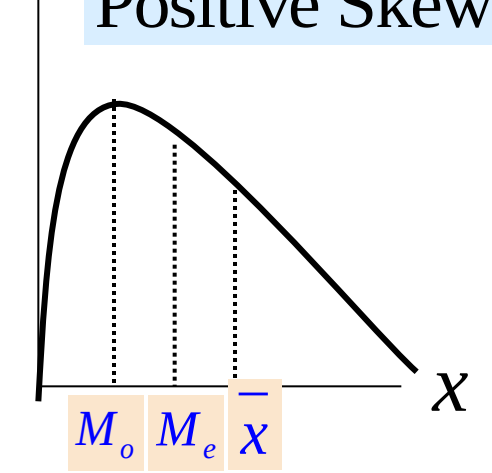
- The relative values of $\bar{x}$, M_e and M_o indicates the **shape of the distribution.**

Symmetrical



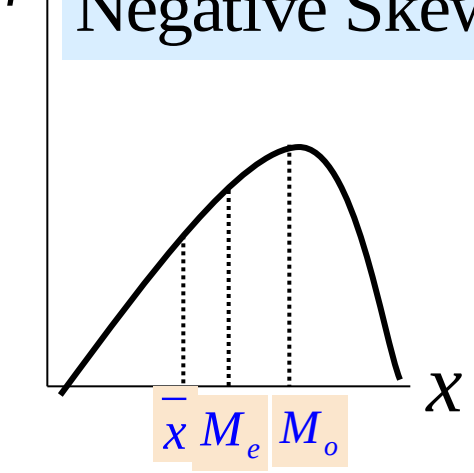
$\bar{x} = M_e = M_o$

Positive Skew



- A few very large values
- Tail to develop on right
- $\bar{x}$ and M_e dragged to right
- $M_o < M_e < \bar{x}$

Negative Skew



- More very large values
- Tail to develop on left
- $\bar{x}$ and M_e dragged to left
- $\bar{x} < M_e < M_o$

Tutorial 5: Measures of Central Tendency for Ungrouped Data

Below is a set of ordered ungrouped data.

$\{3;\ 3;\ 6;\ 9;\ 10;\ 13;\ 14;\ 14;\ 14;\ 15;\ 16\}$

1) Determine the Mode, Median and Mean.
2) Determine the skewness of the data.

Justify your answer.

PAUSE DVD

- Do Tutorial 5
- Then View Solutions

Tutorial 5: Problem 1: Suggested Solution

Mean, Median and Mode: Ungrouped Data

1) Determine the Mode, Median and Mean.

$\{3;\ 3;\ 6;\ 9;\ 10;\ 13;\ 14;\ 14;\ 14;\ 15;\ 16\}$

$$\text{Mode} = M_o = 14$$

$$\text{Median} = M_e = 13$$

$$\text{Mean} = \bar{x} = \frac{2\times 3+6+9+10+13+3\times 14+15+16}{11}$$

$$= \frac{117}{11} = 10,64$$

Calculator as a checking device!

Tutorial 5: Problem 2: Suggested Solution

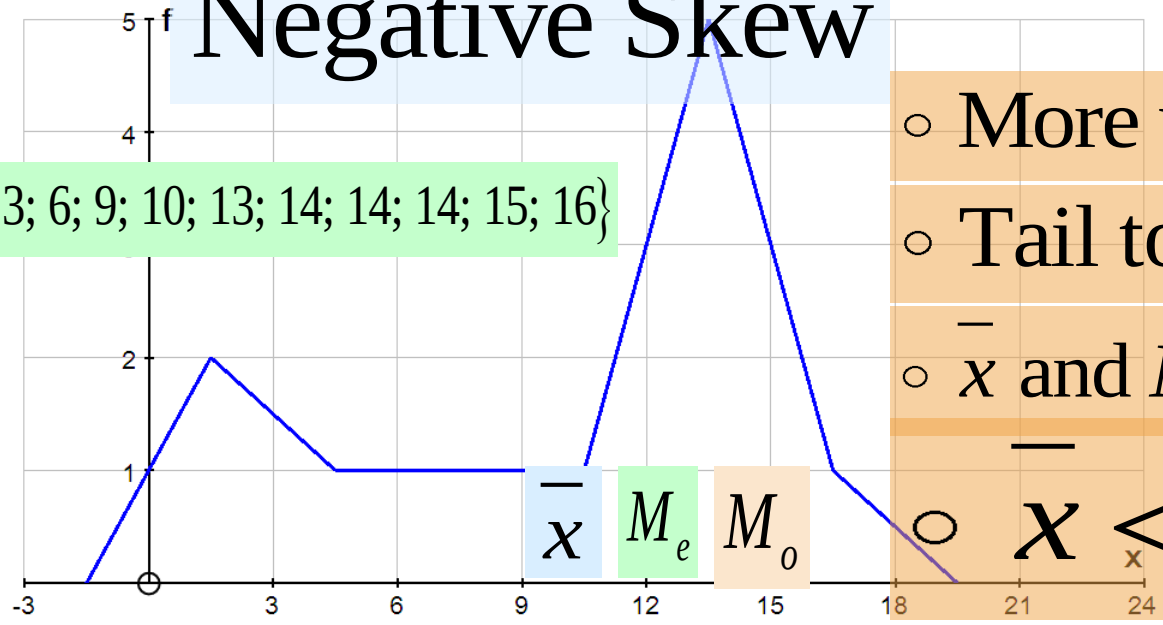
Skewness of the Data Distribution

2) Determine the skewness of the data.

$\bar{x} = 10,64$ $M_e = 13$ $M_o = 14$

Negative Skew

$\{3;\ 3;\ 6;\ 9;\ 10;\ 13;\ 14;\ 14;\ 14;\ 15;\ 16\}$



- More very large values
- Tail to develop on left
- $\bar{x}$ and M_e dragged to left
- $\bar{x} < M_e < M_o$

Mean of Grouped Data

Calculate $\bar{x}$ for the data in the table.

- Grouped data: $\bar{x} = \dfrac{\sum (x_i \cdot f_i)}{n} = \dfrac{2\ 380}{80} = 29,75$

A	B	C	D	E
Class i	Interval	CM (x_i)	f_i	$x_i \cdot f_i$
1	$[15;20)$	17,5	6	105,0
2	$[20;25)$	22,5	13	292,5
3	$[25;30)$	27,5	20	550,0
4	$[30;35)$	32,5	25	812,5
5	$[35;40)$	37,5	12	450,0
6	$[40;45)$	42,5	4	170,0
Calculator as a checking device!			$n = \sum f_i = 80$	$\sum x_i \cdot f =$ 2 380

Tutorial 6: Mean for Grouped Data

Test marks obtained by 35 students for a mathematics test out of 50 are summarised in this frequency table.

Class i	Interval	Class Mark (x_i)	Frequency (f_i)
1	$(0;10]$	5	2
2	$(10;20]$	15	6
3	$(20;30]$	25	14
4	$(30;40]$	35	8
5	$(40;50]$	45	5
			$\sum_{i=1}^{5} f_i = 35$

PAUSE DVD

- Do Tutorial 6
- Then View Solutions

1) Determine the Mean for the grouped data.

Tutorial 6: Problem 1: Suggested Solution

Calculate Mean: Grouped Data

Class i	Interval	x_i	f_i	$x_i \cdot f_i$
1	$(0;10]$	5	2	10
2	$(10;20]$	15	6	90
3	$(20;30]$	25	14	350
4	$(30;40]$	35	8	280
5	$(40;50]$	45	5	225
			$n = 35$	$\sum x_i \cdot f_i = 955$

Calculator as a checking device!

$$\bar{x} = \frac{\sum x_i \cdot f_i}{n} = \frac{955}{35} = 27,29$$

Measure of Variability of Data

- The measures of centrality $\left(\bar{x}, M_e, M_o\right)$ each expresses the centre of the observed data in a different way.
- Should the need arise to express the degree to which individual observations differ from each other, measures of variability are needed.
- The greater the variability between data values (observations) the more they will be spread out.

Example of Variability of Data

Samples A and B below indicate the volumes of 500 ml milk cartons, measued in a Quality Control laboratory:

Sample A (in ml)	485	492	503	504	511
Sample B (in ml)	461	490	503	523	

Variability of data in sample B: $523 - 461 = 62$

Variability of data in sample A: $511 - 485 = 26$

$\therefore$ Producer A should be the preferred provider to the purchaser.

Range of Data

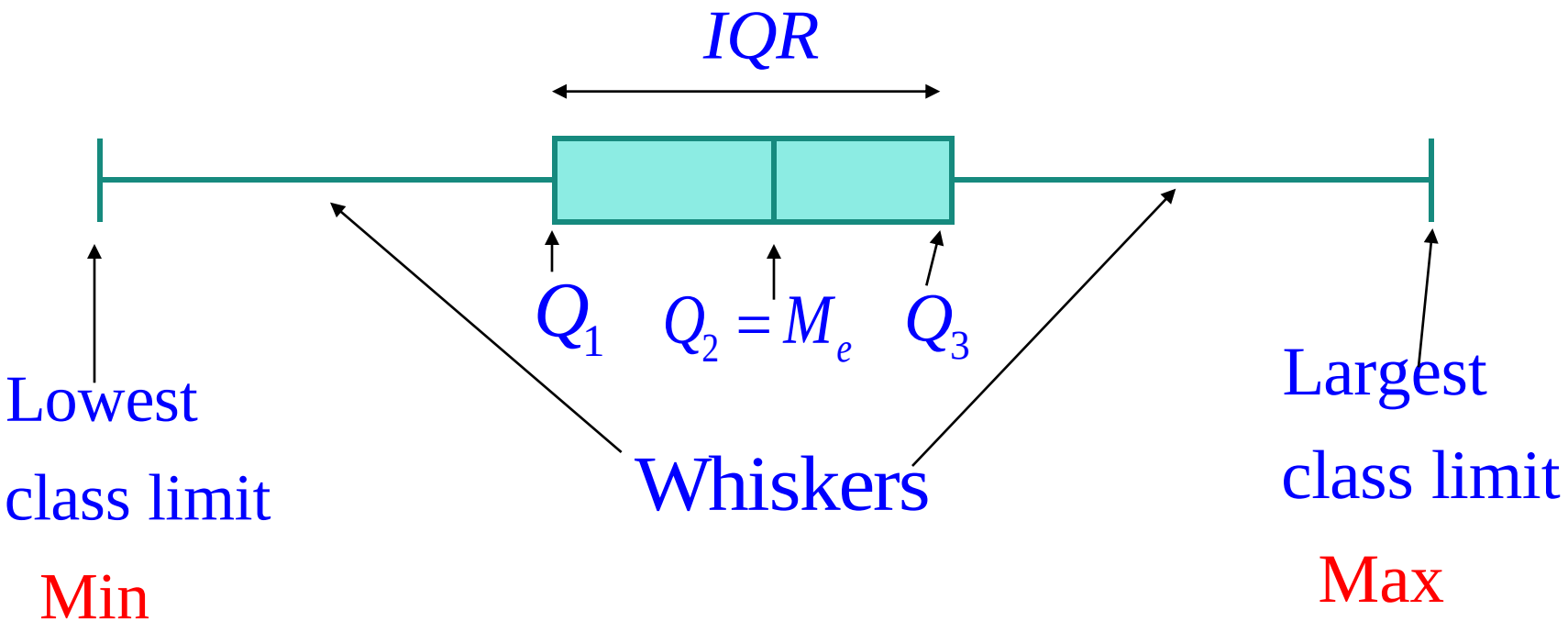
- The **sample range** for ungrouped data is calculated by subtracting the smallest data value from the largest.

- The **sample range** for grouped data is found by subtracting the lowest limit of the 1st class from the upper limit of the last class, if the grouping was done in an ascending manner.

Quartiles and Interquartile Range (IQR)

Box-and-Whisker Plot is a graph representing the sample distribution by a rectangular box covering the IQR and whiskers at the end of the box indicating central dispersion beyond the quartile range.

$$IQR = Q_3 - Q_1 \ \left(\text{middle 50\% of observations}\right)$$



Tutorial 7: Box-and Whisker Plot

A long distance runner recorded his last 15 practice runs.

Distances (in km) are given below:

24 19 21 27 20 17 32 22

26 18 13 23 30 10 13

1) Determine the five-number summary for data above.

2) Draw a box and whisker diagram for the data.

PAUSE DVD

- Do Tutorial 7
- Then View Solutions

(min; Q_1; M_e; Q_3; max) is known as **five-number summary** for data.

Tutorial 7: Problem 1: Suggested Solution
Five-Number Summary

Arrange data in increasing order :

10 13 13 17 18 19 20 21 22 23 24 26 27 30 32

$$\therefore \text{Median} = M_e = Q_2 = P_{50} = 21$$

$$\therefore \text{Lower Quartile} = Q_1 = P_{25} = 17$$

$$\therefore \text{Upper Quartile} = Q_3 = P_{75} = 26$$

$$\text{Min} = 10 \qquad \text{Max} = 32$$

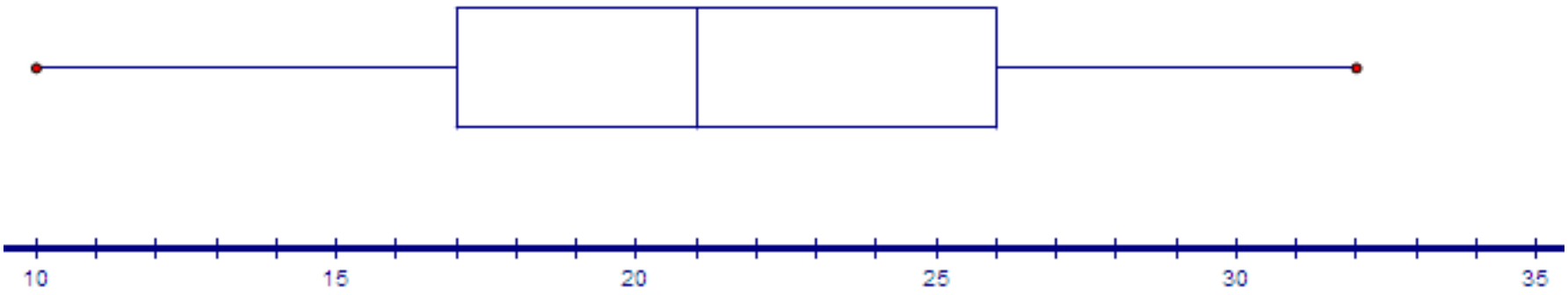
5 - Number Summary : $(10;\ 17;\ 21;\ 26;\ 32)$

Tutorial 7: Problem 2: Suggested Solution
Box and Whisker Plot

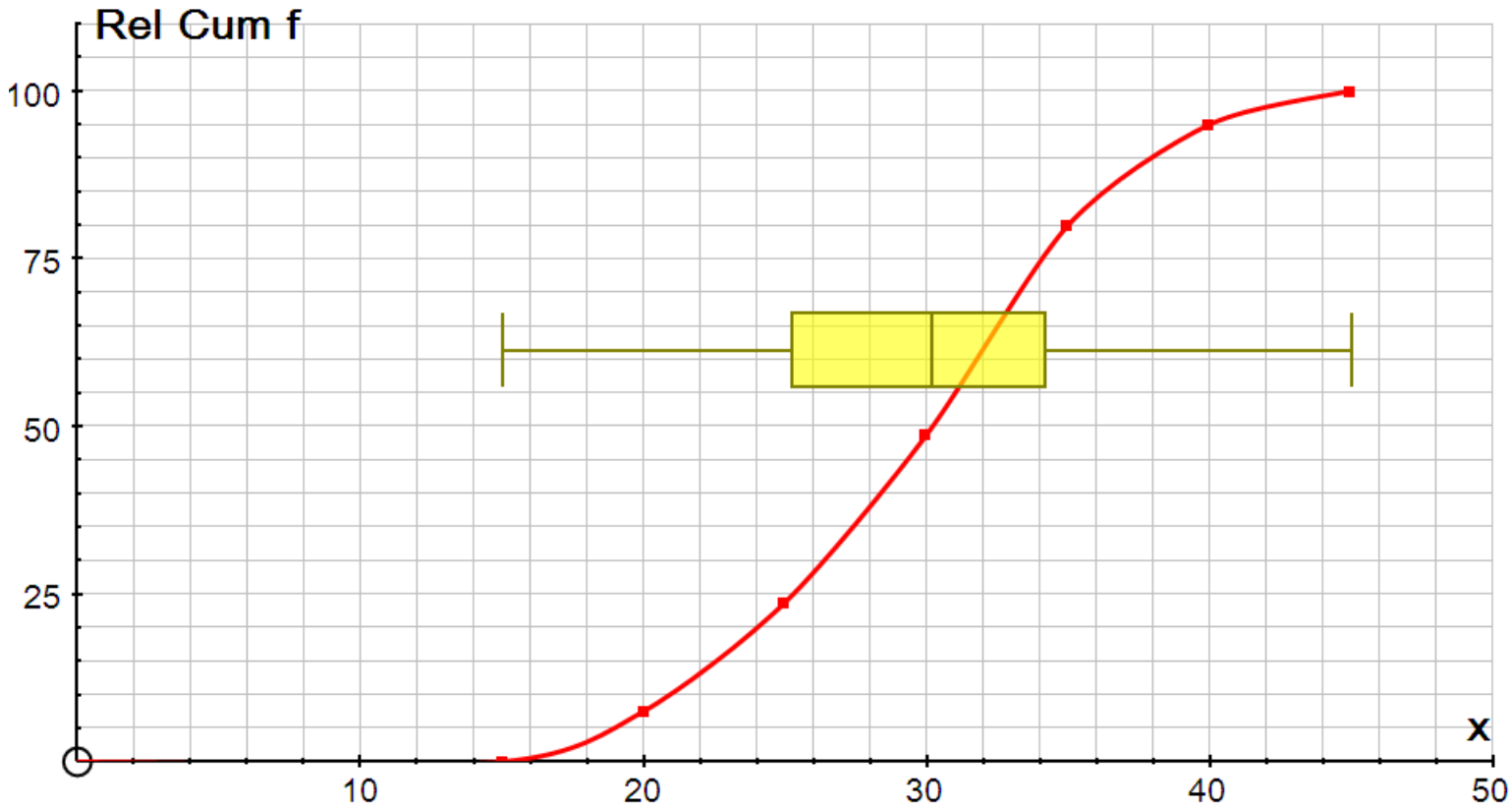
Athlete's data in increasing order:

10 13 13 17 18 19 20 21 22 23 24 26 27 30 32

$(10;17;21;26;32)$ is the **five number summary** of runner's data:

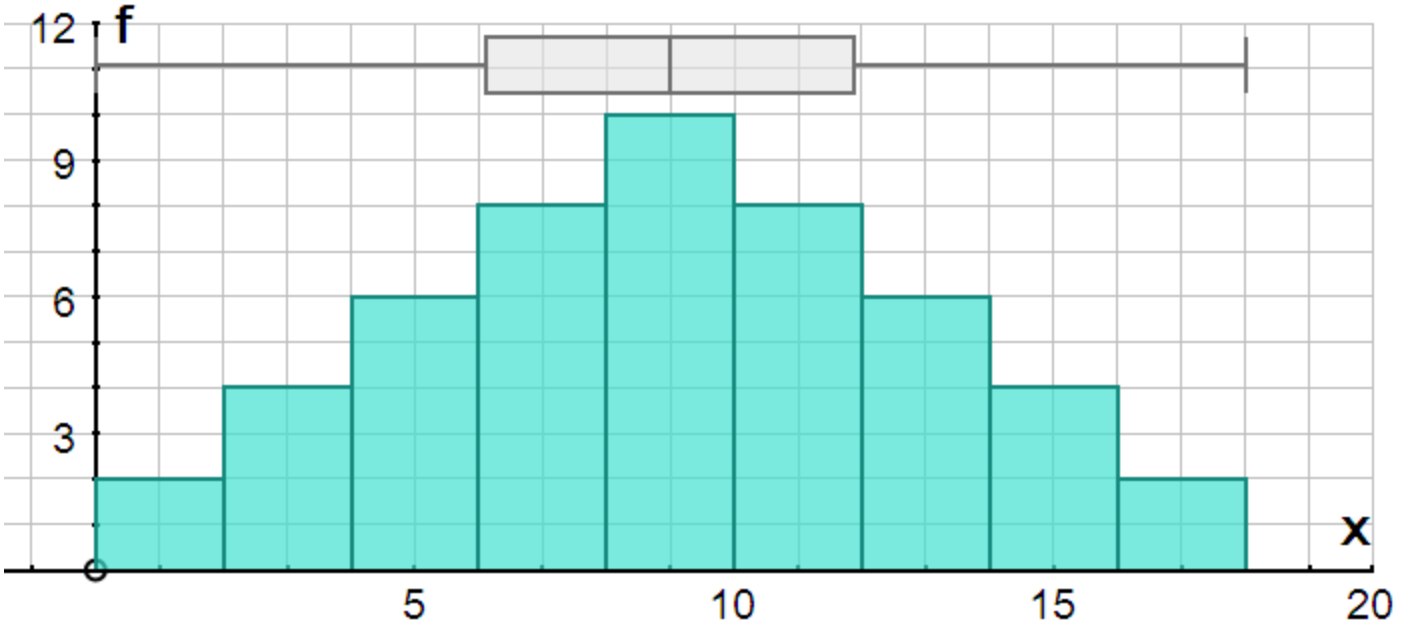


Box-and –Whisker Plot from “Less-than” Relative Cumulative Ogive



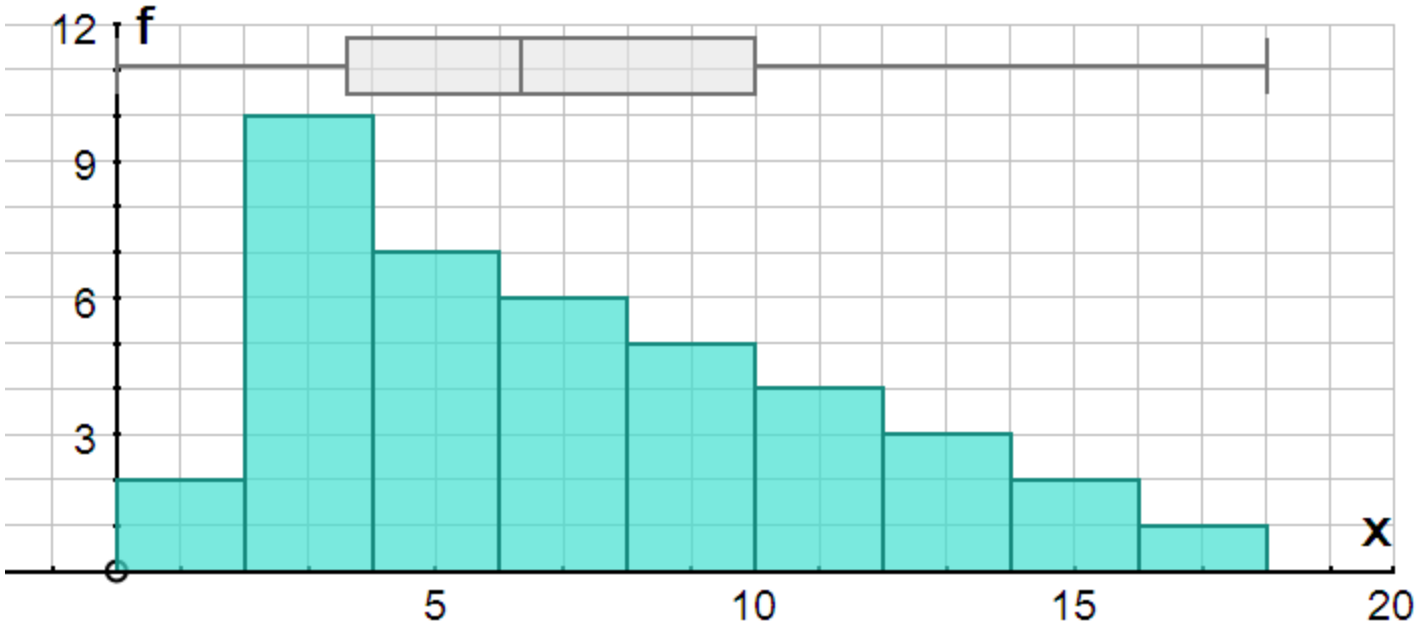
Relating a Box Plot to Symmetric Distribution

- Data is equally spread around the median.
- Box Plot is also symmetric.
- Median is generally in middle.
- Whiskers are approximately equal in length.



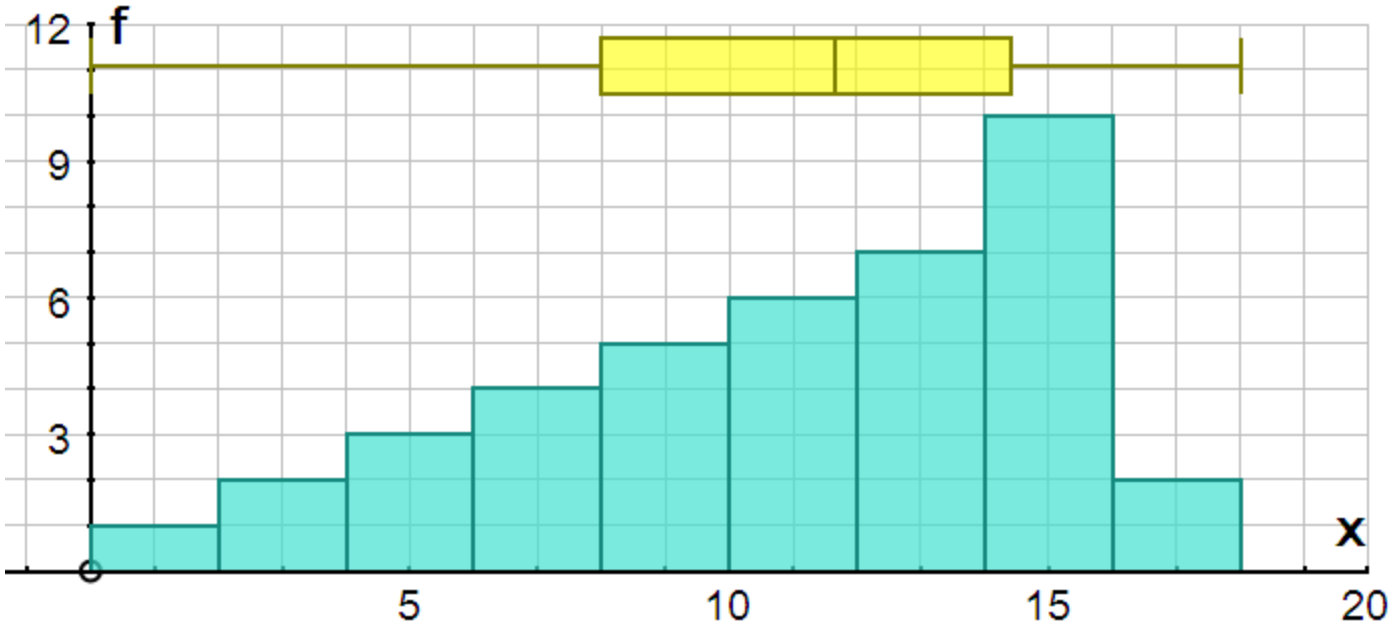
Relating a Box Plot to Positively Skewed Distribution

- Cluster of data values at the left-hand side of distribution.
- Box Plot is positively skewed.
- Median is off-centre and generally to the left.
- Left whisker short and right whisker long.



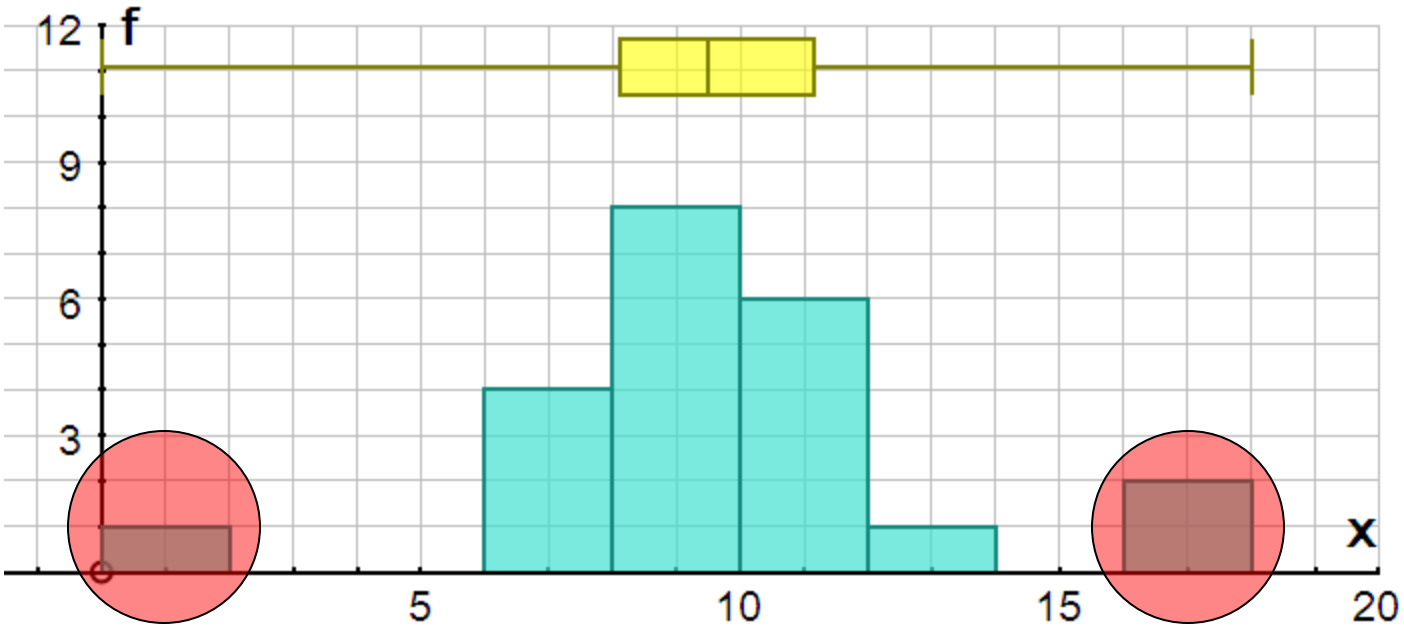
Relating a Box Plot to Negatively Skewed Distribution

- Cluster of data values at right-hand side of distribution.
- Box Plot is negatively skewed.
- Median is off-centre and generally to the right.
- Left whisker long and right whisker short.



Relating a Box Plot to Outliers in Distribution

- Distribution with outlier(s) is caused by values that clearly stand out from the main body of data.
- An outlier might signal an error in the data.



Standard Deviation and Variance for a Sample

The most important measure of variability is based upon the deviations $\left(x_i - \bar{x}\right)$ of the individual data values about the central value of the mean $\bar{x}$.

- For a sample the standard deviation $\left(s_x\right)$ is given by:

  - $s_x = \sqrt{\dfrac{\sum\left(x_i - \bar{x}\right)^2}{n-1}}$ (ungrouped data)

  - $s_x = \sqrt{\dfrac{\sum\left(x_i - \bar{x}\right)^2 \cdot f_i}{n-1}}$ (grouped data)

Sample variance is given by $s_x^{\,2}$

Standard Deviation and Variance for a Population

- For a population:

  μ : population mean and

  N : population size

  - Standard deviation for grouped data:

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

  - Variance for ungrouped data:

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

Tutorial 8: Variance and Standard Deviation

Given ungrouped data: 8,3; 9,0; 7,5; 9,5; 15,4

Calculate:

(1) Variance and

(2) Standard deviation

(3) Interpret spread of distribution in terms of the standard deviation.

Calculator as a checking device!

PAUSE DVD

- Do Tutorial 8
- Then View Solutions

Tutorial 8: Problem 1 & 2: Suggested Solution
Variance and Standard Deviation

Ungrouped data: 8,3; 9,0; 7,5; 9,5; 15,4

$$\bar{x} = \frac{8,3+9,0+7,5+9,5+15,4}{5} = \frac{49,7}{5} = 9,94$$

x_i	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$
8,3	–1,64	2,6896
9,0	–0,94	0,8836
7,5	–2,44	5,9536
9,5	–0,44	0,1936
15,4	5,46	29,8116
	$\sum\left(x_i - \bar{x}\right) = 0$	$\sum\left(x_i - \bar{x}\right)^2 = 39,532$

Calculator as a checking device!

$$\text{Variance} = s_x^2 = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}$$

$$= \frac{39,532}{4} = 9,88$$

$$\text{Standard Deviation} = s_x = \sqrt{9,88} = 3,14$$

Tutorial 8: Problem 3: Suggested Solution
Spread of Distribution

Ordered ungrouped data:

7,5 8,3 9,0 9,5 15,4

$s_x = 3,14$

95% Rule

$\bar{x} - s_x = 6,8$

$\bar{x} = 9,94$

$\bar{x} + s_x = 13,08$

Standard Deviation measures the spread of the distribution by using the mean as reference point.

$\therefore$ 4 of the 5 data values (or 80%) are within a standard deviation from the mean.

End of the First DVD on Data Handling

REMEMBER!

- Consult text-books for additional examples.
- Attempt as many as possible other similar examples on your own.
- Compare your methods with those that were discussed in the DVD.
- Repeat this procedure until you are confident.
- Do not forget:

Practice makes perfect!