Representation of Data

S1/1/1: Frequency distribution tables and histograms
S1/1/2: Stem-and-leaf diagrams
S1/1/3: Cumulative frequency diagrams
S1/1/4: Measures of central tendency
S1/1/5: Measures of central tendency
S1/1/6: Measures of variation
S1/1/7: Box-and-whisker plots
S1/1/8: Pass years' questions and solutions
S1/1/9: Measures of variation
S1/1/10: Use of calculators
S1/1/11: Mean and standard deviation of summarized data
S1/1/12: Evaluation

Prepared by

Tan Bee Hong

S1/1/1:

uency distribution tables and histograms

Learning Outcome

Students should be able to:

- **Understand what is statistics and the scope that it covers, as well as the different branches and uses.**
- **Construct frequency distribution tables and histograms for raw data, both ungrouped and grouped.**

Introduction

What is Statistics?

Statistics:

The science of collecting, describing, and interpreting data

Two areas of statistics:

Descriptive Statistics:

collection, presentation, and description of sample data

Inferential Statistics:

making decisions and drawing conclusions about populations

Introduction to Basic Terms

Population: A collection, or set, of individuals or objects or events whose properties are to be analyzed

- Two kinds of populations: *finite* or *infinite*

Sample: A subset of the population

Variable: **A characteristic about each individual element of a population or sample**

Two Kinds of Variables

Qualitative Variable:

A variable that categorizes or describes an element of a population

Quantitative Variable:

A variable that quantifies an element of a population

Subdividing Variables Further

Variable

Qualitative

Nominal

Categorizes/describes/names an element of a population

Ordinal

Incorporates an ordered position, or ranking

Quantitative

Discrete

can assume a countable number of values, usually associated with **counting**

Continuous

can assume an uncountable number of values, usually associated with **measurements.**

Frequency Distributions & Histograms

Frequency distributions and histograms are used to summarize large data sets

Frequency Distribution:

A listing, often expressed in chart form, that pairs each value of a variable with its frequency (number of times each value occurs)

Ungrouped Frequency Distribution:

Each value of *x* in the distribution stands alone

Example

Number of snails	0	1	2	3	4	5	
Frequency	3	5	11	8	2	1	Total 30

Grouped Frequency Distribution:

Group the values into a set of classes

Example

Mass in grams	101–110	111 – 120	121 – 130	131 – 140	141 – 150	
Frequency	1	4	2	7	2	Total 16

Ways of Grouping Data

(i) **Frequency distribution to show the speeds of 50 cars passing a checkpoint**

Speed (km/h)	20 – 30	30 – 40	40 – 60	60 – 80	over 80
Frequency	2	7	20	16	5
Class Boundaries	20 – 30	30 – 40	40 – 60	60 – 80	**80 – 120**
Class width	10	10	20	20	**2 x 20 = 40**

The interval 20 – 30 represents 20 km/h ≤ *speed* < 30 km/h.

Ways of Grouping Data

(ii) **Frequency distribution to show the lengths, to the nearest milimetre, of 30 rods.**

Length (mm)	27 – 31	32 – 36	37 – 46	47 – 51
Frequency	4	11	12	3
Class Boundaries	26.5 – 31.5	31.5 – 36.5	36.5 – 46.5	46.5 – 51.5
Class width	5	5	10	5

The interval 27 – 31 means 26.5 mm ≤ length < 31.5 mm.

Ways of Grouping Data

(iii) **Frequency distribution to show the lengths of 50 telephone calls.**

Length of call (min)	0 –	3 –	6 –	9 –	12 –	18 –	…
Frequency	9	12	15	10	4	0	…

The interval 3 – represents $3 \le \text{time} < 6$.

The class boundaries 0, 3, 6, 9, 12, 18

The class widths 3, 3, 3, 3. 6

Ways of Grouping Data

(iv) **Frequency distribution to show the masses of 40 packages brought to a particular counter art a post office.**

Mass (g)	–100	–250	– 500	– 800
Frequency	8	10	16	6

The interval –250 represents 100 g < *mass* ≤ 250 g.

The class boundaries 0, 100, 250, 500, 800

The class widths 100, 150, 250, 300

Ways of Grouping Data

(v) **Frequency distribution to show ages (in completed years) of applicants for a teaching post.**

Age (years)	21 – 24	25 – 28	29 – 32	33 – 40	41 – 52
Frequency	4	2	2	1	1

The interval 21 – 24 represents 21 ≤ age < 25.

The class boundaries 21, 25, 29, 33, 41, 53

The class widths 4, 4, 4, 8, 12

Histogram

<u>Histogram</u>: A bar graph representing a frequency distribution of a quantitative variable.

Notes:

- *x*-axis : class boundaries and *y*-axis : frequencies

In a histogram

- no gaps between the bars,
- The area of each bar is proportional to the frequency that it represents.

Histogram

Histograms often have bars of varying widths, so the height of the bar must be adjusted in accordance with the width of the bar.

The vertical axis is not labelled frequency but **frequency density**.

$$\boxed{\text{frequency density} = \frac{\text{frequency}}{\text{interval width}}}$$

Modal Class

In a grouped frequency distribution, the modal class is the interval with the greatest frequency density.

Example 1:

The haemoglobin test, a blood test given to diabetics during their periodic checkups, indicates the level of control of blood sugar during the past two to three months. The data in the table below was obtained for 40 different diabetics at a university clinic that treats diabetic patients:

6.5	5.0	5.6	7.6	4.8	8.0	7.5	7.9	8.0	9.2
6.4	6.0	5.6	6.0	5.7	9.2	8.1	8.0	6.5	6.6
5.0	8.0	6.5	6.1	6.4	6.6	7.2	5.9	4.0	5.7
7.9	6.0	5.6	6.0	6.2	7.7	6.7	7.7	8.2	9.0

1) Construct a grouped frequency distribution using the classes 3.7 – 4.7, 4.7 –5.7, 5.7 – 6.7, etc.

2) Which class is the modal class?

3) Construct a histogram for the blood test results.

Example 1:

6.5	5.0	5.6	7.6	4.8	8.0	7.5	7.9	8.0	9.2
6.4	6.0	5.6	6.0	5.7	9.2	8.1	8.0	6.5	6.6
5.0	8.0	6.5	6.1	6.4	6.6	7.2	5.9	4.0	5.7
7.9	6.0	5.6	6.0	6.2	7.7	6.7	7.7	8.2	9.0

1)

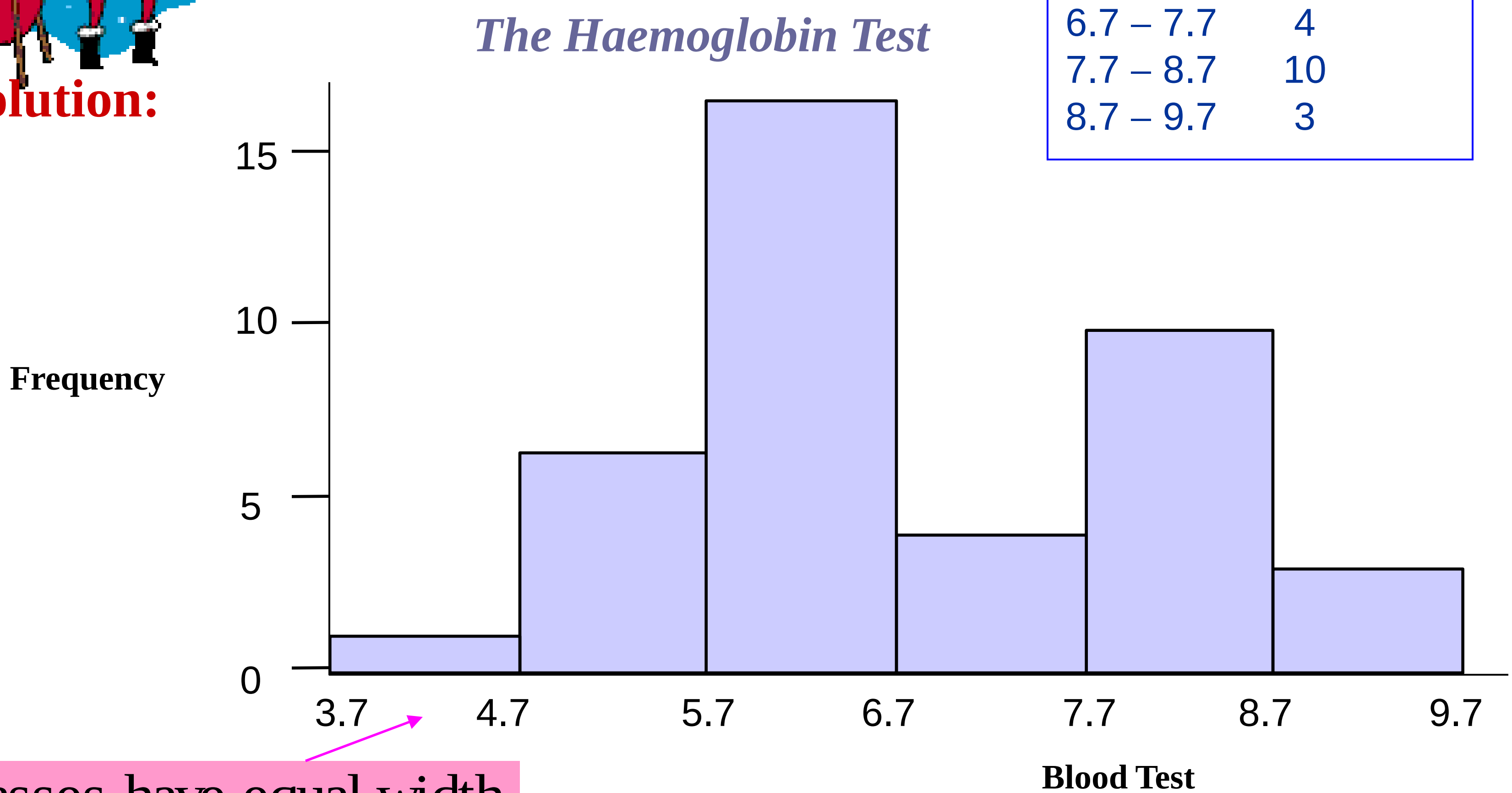
Class Boundaries	Frequency f
3.7 – 4.7	1
4.7 – 5.7	6
5.7 – 6.7	16
6.7 – 7.7	4
7.7 – 8.7	10
8.7 – 9.7	3

$$\sum f = 40$$

2) The class 5.7 – 6.7 is the modal class.

Solution:

Class	*f*
3.7 – 4.7	1
4.7 – 5.7	6
5.7 – 6.7	16
6.7 – 7.7	4
7.7 – 8.7	10
8.7 – 9.7	3



classes have equal width

can use frequency for the height of the bar

Example 2:

These are the examination marks for a group of 120 first year statistics students.

Marks	0 – 9	10 – 19	20 – 29	30 – 49	50 – 79
Frequency	8	21	53	28	10

Represent the data in a histogram, state the modal class and comment on the shape of the distribution.

Example 2:

$$\text{frequency density} = \frac{\text{frequency}}{\text{interval width}}$$

These are the examination marks for a group of 120 first year statistics students.

Marks	0 – 9	10 – 19	20 – 29	30 – 49	50 – 79
Frequency	8	21	53	28	10

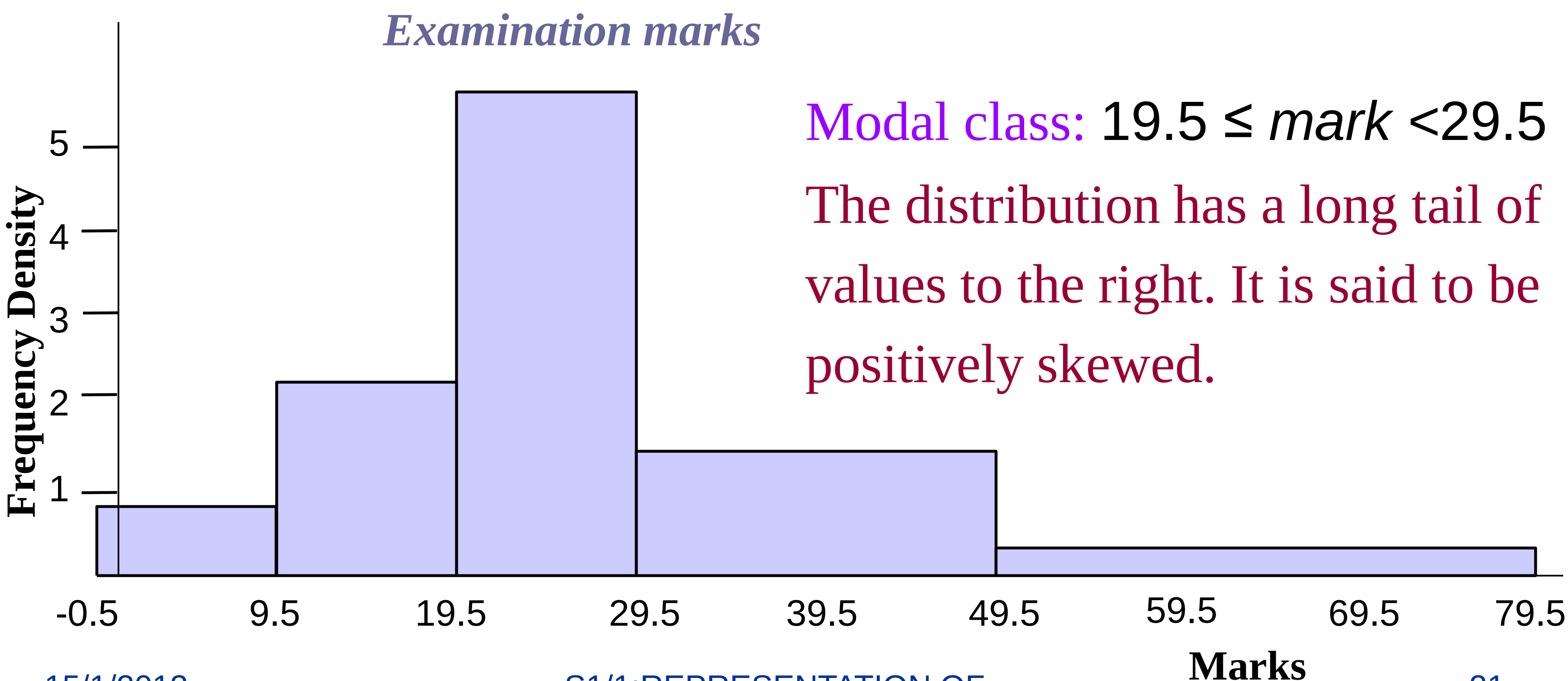
Represent the data in a histogram, state the modal class and comment on the shape of the distribution.

Mark	Class width	Frequency	Frequency Density
$-0.5 \le mark < 9.5$	10	8	0.8
$9.5 \le mark < 19.5$	10	21	2.1
$19.5 \le mark < 29.5$	10	53	5.3
$29.5 \le mark < 49.5$	20	28	1.4
$49.5 \le mark < 79.5$	30	10	0.3

Example 2:

Mark	Class width	Frequency	Frequency Density
$-0.5 \leq mark < 9.5$	10	8	0.8
$9.5 \leq mark < 19.5$	10	21	2.1
$19.5 \leq mark < 29.5$	10	53	5.3
$29.5 \leq mark < 49.5$	20	28	1.4
$49.5 \leq mark < 79.5$	30	10	0.3

Solution:



Modal class: $19.5 \leq mark < 29.5$

The distribution has a long tail of values to the right. It is said to be positively skewed.

Example 3:

The following frequency distribution summarizes in masses in grams (g), of a sample of 20 pebbles.

Mass (g)	101 – 110	111 – 120	121 – 130	131 – 140	141 – 150	over 150
Frequency	1	4	2	7	2	4

Represent the data in a histogram.

Practice Exercise

Advanced Level Mathematics

Statistics 1 Steve Dobbs and Jane Miller

Exercise 1B (Page 15)

1, 6, 8

S1/1/2:
Stem-and-leaf diagrams

Learning Outcome

Students should be able to:

- **Construct and interpret the stem-and-leaf diagram.**
- **State/list the advantage and disadvantage of using the stem-and-leaf diagram for presentation of data.**

Stem-and-leaf diagrams (Stemplots)

- A useful way of grouping data into classes while still retaining the original data.
- It is a combination of graphing and sorting.
- Gives a good idea at a glance of the shape of the distribution.
- Easy to pick out the smallest, largest values, mode and modal class.

Stem-and-Leaf Diagram:

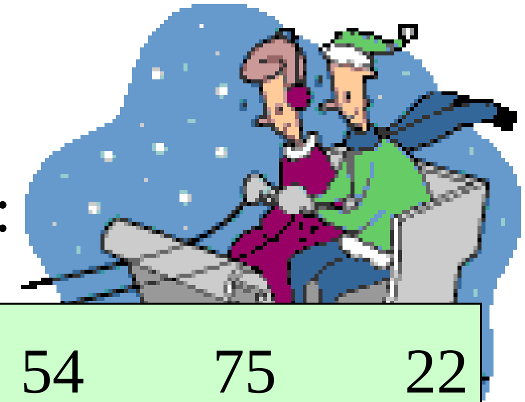
Each numerical data is divided into two parts:

The leading digit(s) becomes the ***stem***, and the trailing digit(s) becomes the ***leaf***.

The stems are located along the main axis, and a leaf for each piece of data is located so as to display the distribution of the data.

All the intervals must be equal width.

Example 4:



These are the marks of 20 students in an assignment:

84	17	38	45	47	53	76	54	75	22
66	65	55	54	51	33	39	19	54	72

The lowest mark is 17 and the highest mark is 84.

Choose intervals 10 – 19, 20 – 29, 30 – 39, ..., 80 – 89 for this data.

Stem	Leaf
1	7 9
2	2
3	8 3 9
4	5 7
5	3 4 5 4 1 4
6	6 5
7	6 5 2
8	4

Stem	Leaf
1	7 9
2	2
3	3 8 9
4	5 7
5	1 3 4 4 4 5
6	5 6
7	2 5 6
8	4

Key 1 | 7 means 17 marks

Slightly skewed to the right.

Back-to-back stemplots

A back-to-back stem-and-leaf display makes it possible to compare two distributions graphically

Example 5:

Use a stem and leaf diagram to compare the examination marks in French and English for a class of 20 pupils.

French	75	69	58	58	46	44	32	50	53	78
	81	61	61	45	31	44	53	66	47	57
English	52	58	68	77	38	85	43	44	56	65
	65	79	44	71	84	72	63	69	72	79

Key (French)
9 | 6 means 69

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

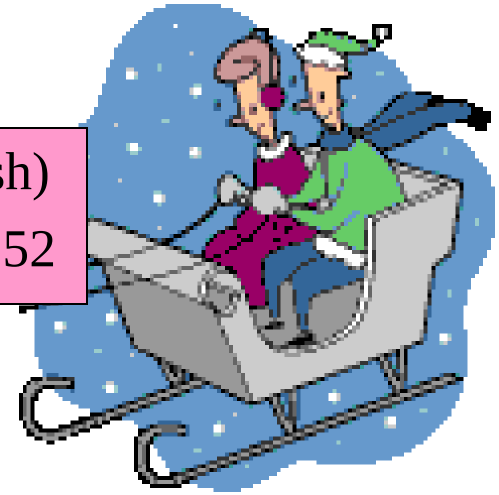
Key (English)
5 | 2 means 52

Example 5:

Key (French)
9 | 6 means 69

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

Key (English)
5 | 2 means 52



Arranged in order:

Key (French)
9 | 6 means 69

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

Key (English)
5 | 2 means 52

The class had higher marks in English than in French and it appears that they performed better in English.

Practice Exercise

Advanced Level Mathematics

Statistics 1 Steve Dobbs and Jane Miller

Exercise 1A (Page 8)

2, 4

S1/1/4 and S1/1/5:
Measures of central tendency

S1/1/6 and S1/1/9:
Measures of variation

S1/1/3:
Cumulative frequency diagrams

Learning Outcome

Students should be able to:

- Describe the meaning of measures of central tendency.
- Find the measure of central tendency, mean.
- Find the median and mode.
- Describe the meaning of measures of variation
- Find the measure of dispersion, range and quartiles, inter quartile range and semi-inter range
- Find the variance and standard deviation.
- Construct cumulative frequency table.
- Construct and interpret cumulative frequency graphs.

Measures of central tendency

<u>Introduction</u>:

Numerical values used to locate the middle of a set of data, or where the data is clustered.

Such value is called a **measure of location**, or a **measure of central tendency**,

The term *average* is often associated with all measures of central tendency.

Measures of central tendency

The mean

For discrete raw data $x_1, x_2, \ldots, x_n$

The mean, $\bar{x}$, of a data set of n values is given by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

Note: $\sum$ is read as 'sigma'. This notation for a sum is called $\sum$-**notation**.

Measures of central tendency

The mean

For data in an ungrouped frequency distribution

The mean, $\bar{x}$, of a data set in which the variable takes the value x_1 with frequency f_1, x_2 with frequency f_2 and so on is given by

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum x_i f_i}{\sum f_i}$$

When the data have been grouped into intervals,
we take the mid-interval values as representative of the interval.

mid-interval value = ½ (lower class boundary + upper class boundary)

Measures of central tendency

<u>Median</u>:

The value of the data that occupies the middle position when the data are ranked in order according to size

Notes:

Denoted by '*x* tilde' : $\tilde{x}$

<u>To find the *median for ungrouped raw data*</u>:

1. Rank the data
2. Determine the *depth* of the median: $d(\tilde{x}) = \frac{n+1}{2}$
3. Determine the value of the median

Measures of central tendency

<u>Mode / Modal value</u>:

The mode is the value of *x* that occurs most frequently

Note: A data set can have more than one mode if two or more values have the same maximum frequency.

A data set has no mode if all the values have the same frequency.

Measures of central tendency

<u>Mode</u>: For grouped frequency data

For grouped frequency data, it is only possible to estimate the mode.

Alternatively, we can give the modal class, which is the class with highest <u>frequency density</u>.

Measures of Position

- Measures of position are used to describe the relative location of an observation

- **Quartiles** is a one of the measures of position

<u>Quartiles</u>:

Values of the variable that divide the ranked data into quarters; each set of data has three quartiles

Ranked data, increasing order

25%	25%	25%	25%

L $\quad Q_1 \quad Q_2 \quad Q_3 \quad H$

Interquartile range

The difference between the quartiles, Q_1 and Q_3 is known as **interquartile range**.

It tells us the range of the middle 50% of the distribution,

Interquartile range = upper quartile – lower quartile

$$\boxed{\text{Interquartile range} = Q_3 - Q_1}$$

Measures of Variation

Measures of central tendency alone cannot completely characterize a set of data.
Two very different data sets may have similar measures of central tendency.

Measures of variation/dispersion are used to describe the spread, or variability, of a distribution

Common measures of dispersion:

range, variance, and standard deviation

Range

Range:

The difference in value between the highest-valued (H) and the lowest-valued (L) pieces of data:

$$\text{range} = H - L$$

Measures of variation

The standard deviation, *s*, is used as the measure of variability or spread.

The standard deviation, s, of a set of n numbers, with mean $\bar{x}$, is given by

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

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

$$\text{Variance} = (\text{standard deviation})^2 = s^2$$

Standard deviations are useful when comparing sets of data; the higher the standard deviation, the <u>greater the variability</u> in the data.

Measures of variation

When data are in the form of a frequency distribution, the formula for *s* is

$$s = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}$$

where $\bar{x}$ is the mean

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Example 6: Ungrouped raw data

The following data represents the pH levels of a random sample of swimming pools in a California town.

5.6	5.8	5.9	6.0	6.0
6.1	6.2	6.3	6.4	6.7
6.8	6.8	6.8	6.9	7.0
7.3	7.4	7.4	7.5	

Find the (i) mean, mode, median, standard deviation and variance.

(ii) range, first quartile, third quartile and interquartile range.

Example 7

The table shows the number of attempts needed to pass the driving test by 100 candidates at a particular test centre.

Number of attempts	1	2	3	4	5	6
Frequency (Number of candidates)	33	42	13	6	4	2

Find the (i) mean, mode, median, standard deviation and variance.

(ii) range, first quartile, third quartile and interquartile range.

S1/1/10:
Use of calculators

Learning Outcome

Students should be able to:

- **Use calculator to find**

  $\Sigma x, \Sigma x^2, \Sigma f x, \Sigma f x^2, \sigma, \sigma^2, x$.

Statistical Calculations

SD REG

Standard Deviation

SD

Use the MODE key to enter the SD Mode when you want to perform statistical calculations using standard deviation.

SD .... MODE 2 (fx-95MS)
MODE MODE 1 (Other Models)

- Always start data input with $\boxed{\text{SHIFT}}$ $\boxed{\text{CLR}}$ $\boxed{1}$ (Scl) $\boxed{=}$ to clear statistical memory.
- Input data using the key sequence shown below.
  <x-data> $\boxed{\text{DT}}$
- Input data is used to calculate values for n, Σx, Σx^2, $\bar{x}$, σn and $\sigma n\text{-}1$, which you can recall using the key operations noted nearby.

To recall this type of value:	Perform this key operation:
Σx^2	SHIFT S-SUM 1
Σx	SHIFT S-SUM 2
n	SHIFT S-SUM 3
$\bar{x}$	SHIFT S-VAR 1
σn	SHIFT S-VAR 2
σn-1	SHIFT S-VAR 3

- **Example:** To calculate $\sigma n-1$, σn, $\bar{x}$, n, Σx, and Σx^2 for the following data : 55, 54, 51, 55, 53, 53, 54, 52

In the SD Mode:

[SHIFT] [CLR] [1] (Scl) [=] (Stat clear)

55 [DT]

n =	SD
	1.

Each time you press [DT] to register your input, the number of data input up to that point is indicated on the display (n value).

54 [DT] 51 [DT] 55 [DT]

Edit Go To Favorites Help

Back Search Favorites

ess http://ftp.casio.co.jp/pub/world_manual/edu/en/fx115MS_991MS_E.pdf#search=%22casio%20570Ms%20manual%22 Go Lin

ogle casio 570Ms manual Go Bookmarks 0 blocked Check AutoLink Setti

53 DT DT 54 DT 52 DT

Sample Standard Deviation ($\sigma n-1$) = ***1.407885953*** SHIFT S-VAR 3 =

Population Standard Deviation (σn) = ***1.316956719*** SHIFT S-VAR 2 =

Arithmetic Mean ($\bar{x}$) = ***53.375*** SHIFT S-VAR 1 =

2.92 x 5.51 in

300% 28 of 42

Unknown Zone

Number of Data (n) = ***8*** SHIFT S-SUM 3 =

Sum of Values (Σx) = ***427*** SHIFT S-SUM 2 =

Sum of Squares of Values (Σx^2) = ***22805*** SHIFT S-SUM 1 =

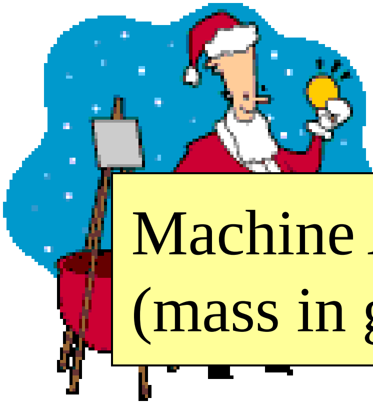
Data Input Precautions

- DT DT inputs the same data twice.
- You can also input multiple entries of the same data using SHIFT ;. To input the data 110 ten times, for example, press 110 SHIFT ; 10 DT.

Example 8:

Find the mean and standard deviation of the masses of the packets taken in the sample from each machine.

Machine A (mass in g)	196, 198, 198, 199, 200, 200, 201, 201, 202, 205
Machine B (mass in g)	192, 194, 195, 198, 200, 201, 203, 204, 206, 207

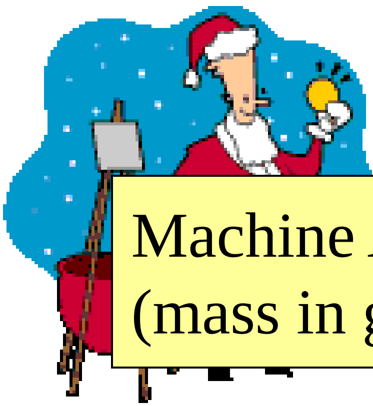


Example 8: Find mean

Machine A (mass in g)	196, 198, 198, 199, 200, 200, 201, 201, 202, 205

Machine A

Set SD mode		MODE	MODE	1						
Clear memories		SHIFT	Scl	=						
Input data		196	M+	198	M+	….	205	M+		
To obtain	$\bar{x} = 200$	SHIFT	2	1	=					
	$n = 10$	SHIFT	1	3	=					
	$\sum x = 2000$	SHIFT	1	2	=					
To clear SD mode		MODE	1							



Example 8: Find standard deviation

Machine A (mass in g)	196, 198, 198, 199, 200, 200, 201, 201, 202, 205

Machine A

Set SD mode — **MODE** **MODE** **1**

Clear memories — **SHIFT** **Scl** **=**

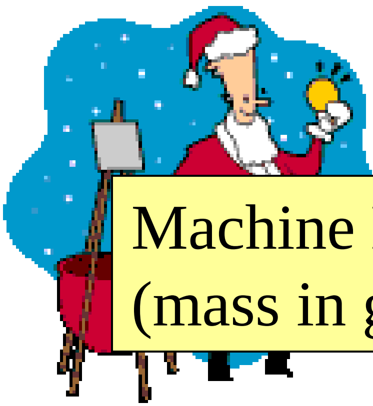
Input data — **196** **M+** **199** **M+** …. **205** **M+**

To obtain std. dev $= 2.36643191$3 — **SHIFT** **2** **2** **=**

$n = 10$ — **SHIFT** **1** **3** **=**

$\sum x^2 = 400056$ — **SHIFT** **1** **1** **=**

To clear SD mode — **MODE** **1**

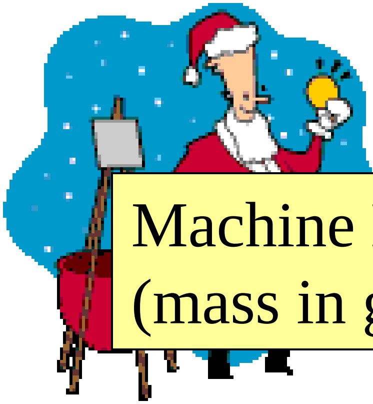


Example 8:

Machine B (mass in g) 192, 194, 195, 198, 200, 201, 203, 204, 206, 207

Machine B

Set SD mode		MODE MODE 1
Clear memories		SHIFT Scl =
Input data		192 M+ 194 M+ …. 207 M+
To obtain	$\bar{x} = 200$	SHIFT 2 1 =
	$n = 10$	SHIFT 1 3 =
	$\sum x = 2000$	SHIFT 1 2 =
To clear SD mode		MODE 1



Example 8: Find standard deviation

Machine B (mass in g)	192, 194, 195, 198, 200, 201, 203, 204, 206, 207

Machine B

Set SD mode	**MODE** **MODE** **1**
Clear memories	**SHIFT** **Scl** **=**
Input data	**192** **M+** **194** **M+** …. **207** **M+**
To obtain std. dev. $= 4.898979486$	**SHIFT** **2** **2** **=**
$n = 10$	**SHIFT** **1** **3** **=**
$\sum x^2 = 400240$	**SHIFT** **1** **1** **=**
To clear SD mode	**MODE** **1**

Practice Exercise

Advanced Level Mathematics

Statistics 1 Steve Dobbs and Jane Miller

Exercise 2A (Page 26)

1

Exercise 2B (Page 31)

7

Miscellaneous exercise 2 (Page 36)

2, 5

Cumulative Frequency

Cumulative Frequency Distribution:

- A frequency distribution that pairs cumulative frequencies with values of the variable

- The *cumulative frequency* for any given class is the sum of the frequency for that class and the frequencies of all classes of smaller values.

Cumulative Frequency

Cumulative frequency can be illustrated:

(a) When the data are **discrete and ungrouped**-
by drawing a **step diagram**,

(b) When the data are **continuous** or in the form of a
grouped discrete distribution-by drawing a
cumulative frequency polygon or **curve**.

(a) Cumulative Frequency

Step diagrams for discrete ungrouped data

The cumulative frequencies are plotted against the **upper class boundaries.**

Example 9

The table shows the number of attempts needed to pass the driving test by 100 candidates at a particular test centre.

Number of attempts	1	2	3	4	5	6
Frequency (Number of candidates)	33	42	13	6	4	2

Draw a step diagram for this data.

Example 9:

The table shows the number of attempts needed to pass the driving test by 100 candidates at a particular test centre.

Number of attempts	1	2	3	4	5	6
Frequency (Number of candidates)	33	42	13	6	4	2

The cumulative frequency distribution:

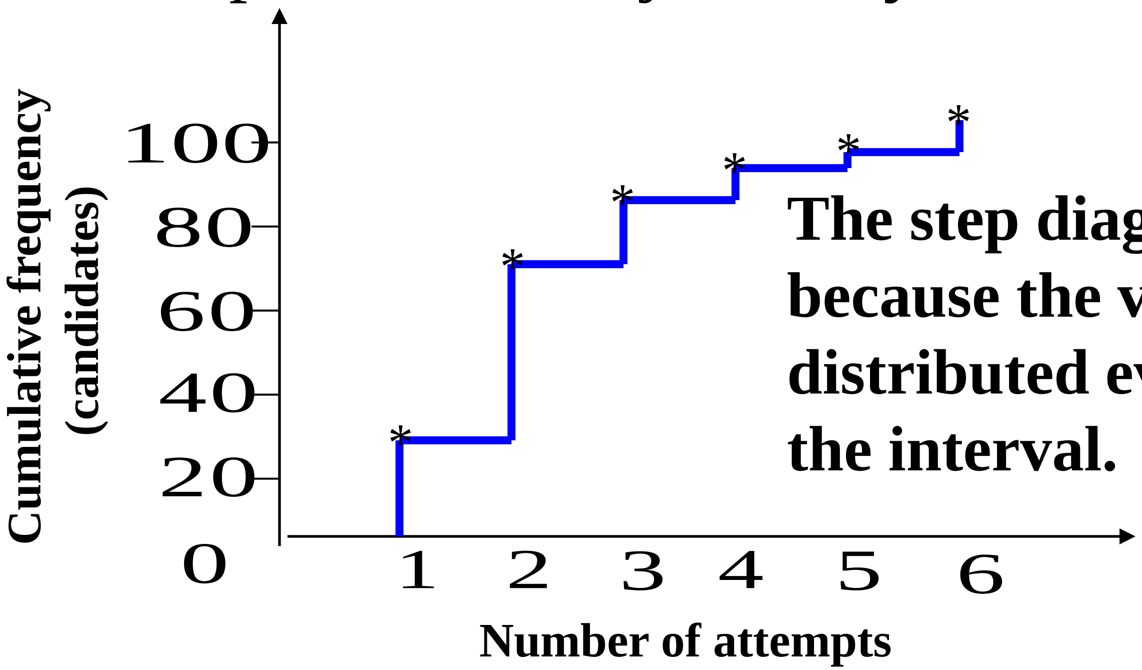
Number of attempts	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5	≤ 6
Cumulative frequency	33	75	88	94	98	100

Example 9:

The cumulative frequency distribution:

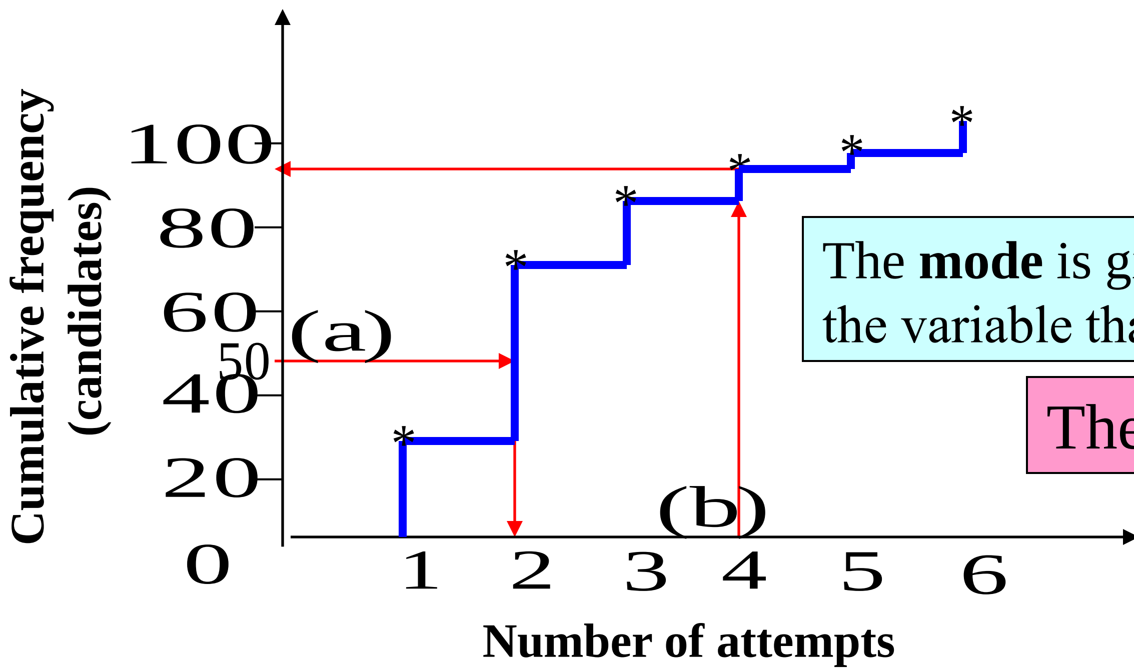
Number of attempts	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5	≤ 6
Cumulative frequency	33	75	88	94	98	100

When data are discrete (usually integer) values and also ungrouped, the points can be joined by a series of **steps** as shown:



The step diagram is necessary because the values are not distributed evenly throughout the interval.

Example 9:



The **mode** is given by the value of the variable that gives the 'steepest' step.

The mode is 2.

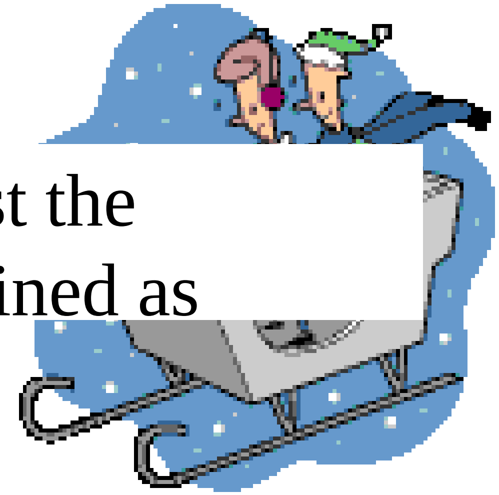
(a) From the graph, when the cumulative frequency is 50, the number of attempts is 2.

(b) To find how many took up to 4 attempts, go to the **top** of the step, this shows that there are 94 candidates.

If you go to the **bottom** of the step, this tells you the number of candidates who took fewer than 4 attempts.

(b) Cumulative Frequency

Cumulative frequency polygons and curves for grouped data



The cumulative frequencies are plotted against the **upper class boundaries** and the points are joined as follows:

(i) Cumulative frequency polygon,

Joint the points with straight lines, assuming that the readings are evenly distributed throughout the interval.

(ii) Cumulative frequency curve,

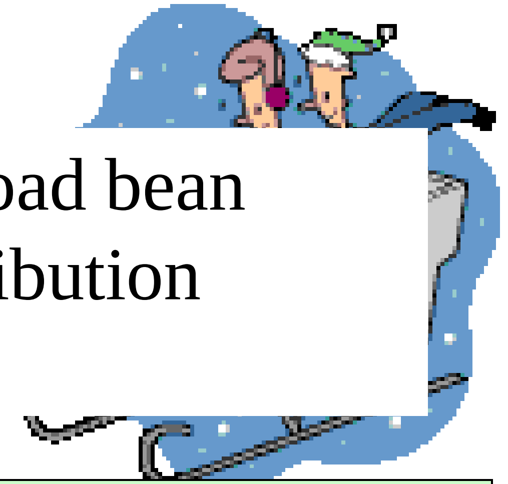
Joint the points with a smooth curve, assuming a distribution of readings throughout the interval which might not be even.

Cumulative frequency polygons and curves for grouped data

***Note*:**

Every ogive starts on the left with a cumulative frequency of zero at the lower class boundary of the first class and ends on the right with a c. frequency of 100% at the upper class boundary of the last class.

Example 10:



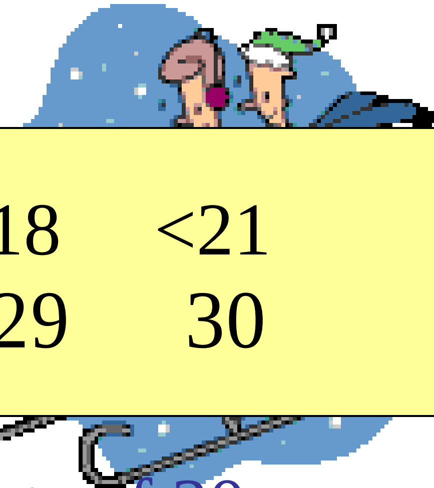
Six weeks after planting, the heights of 30 broad bean plants were measured and the frequency distribution formed as shown.

Height, x cm	$3 \le x < 6$	$6 \le x < 9$	$9 \le x < 12$	$12 \le x < 15$	$15 \le x < 18$	$18 \le x < 21$
Freq.	**1**	**2**	**11**	**10**	**5**	**1**

(a) Mean, modal class and variance.

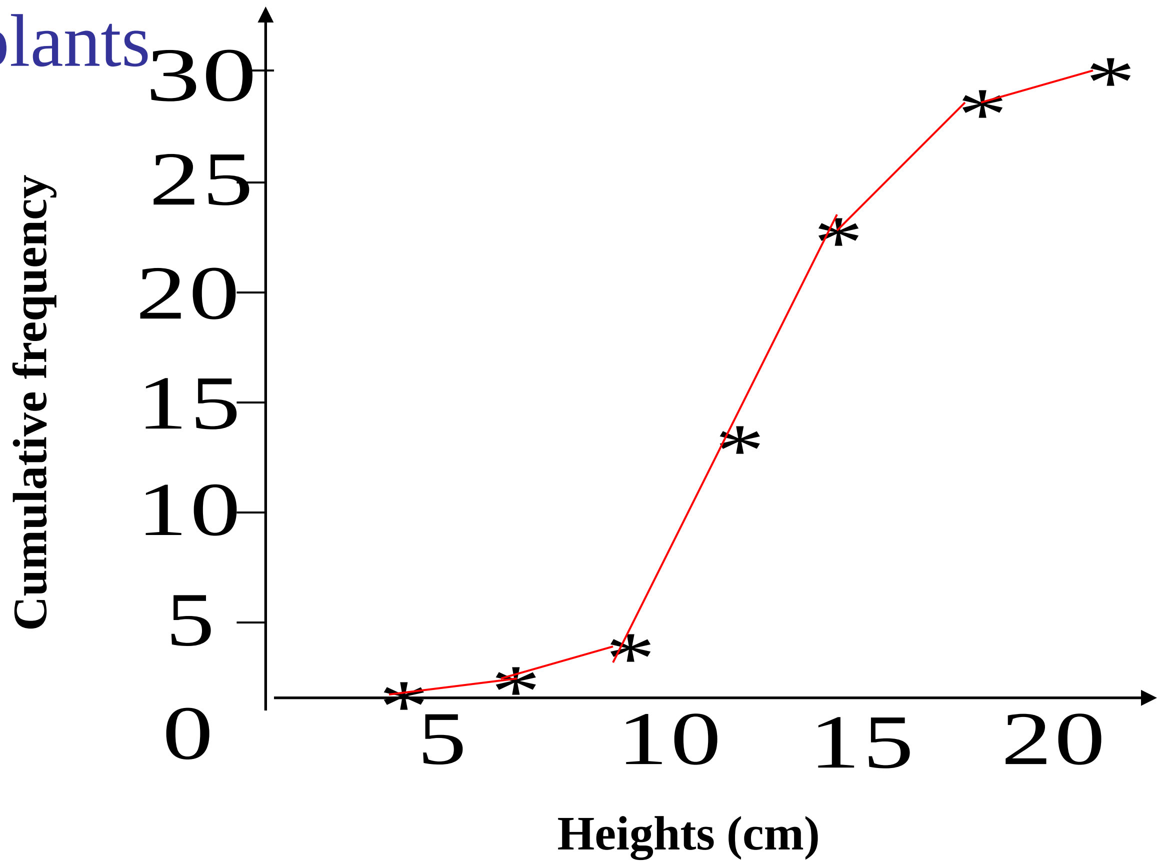
(b) Draw the cumulative frequency polygon and cumulative frequency curve.

Example 10:



Height,	<3	<6	< 9	< 12	< 15	< 18	<21
Cum. freq	0	1	3	14	24	29	30

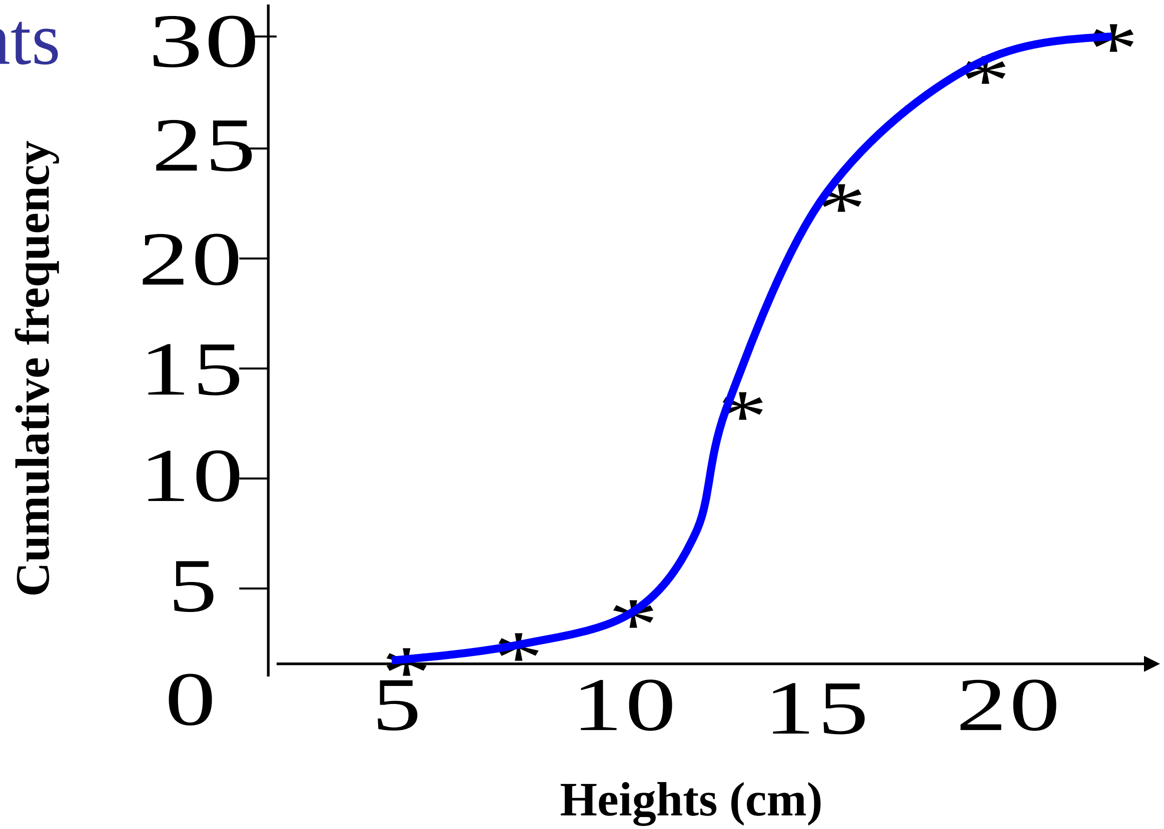
Cumulative frequency polygon to show the heights of 30 broad bean plants

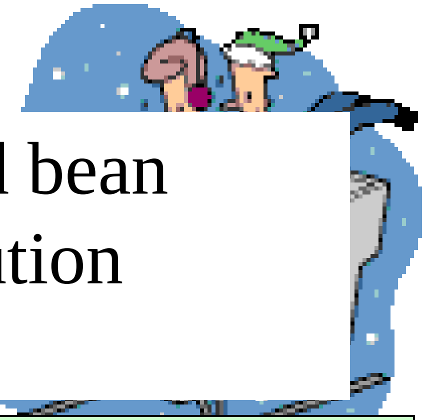


Example 10:

Height,	<3	<6	< 9	< 12	< 15	< 18	<21
Cum. freq	0	1	3	14	24	29	30

Cumulative frequency curve to show the heights of 30 broad bean plants





Example 10:

Six weeks after planting, the heights of 30 broad bean plants were measured and the frequency distribution formed as shown.

Height, *x* cm	**$3 \le x < 6$**	**$6 \le x < 9$**	**$9 \le x < 12$**	**$12 \le x < 15$**	**$15 \le x < 18$**	**$18 \le x < 21$**
Freq.	**1**	**2**	**11**	**10**	**5**	**1**

(c) From the cumulative frequency curve,

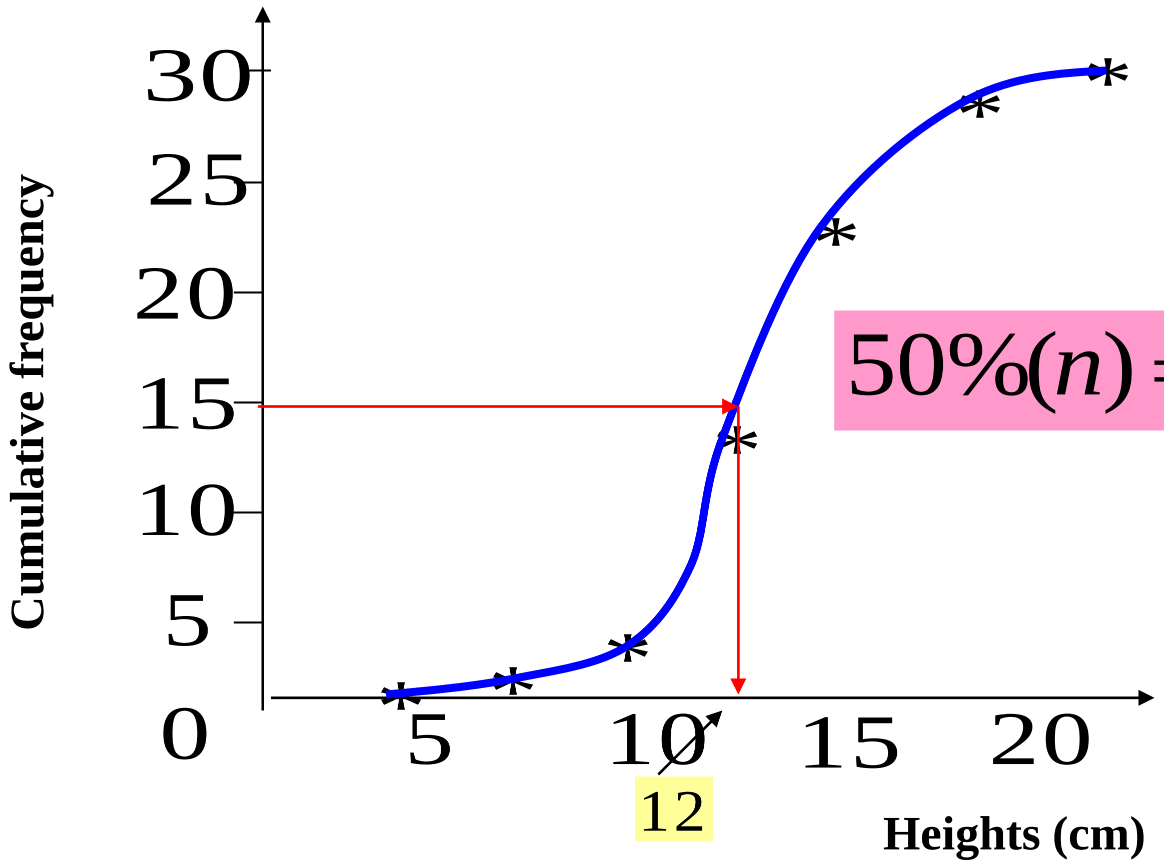
(i) Find the quartiles.

(ii) To find the number of plants that were less than 10.5 cm tall,

(iii) To find x where 90% of the plants were less than x cm tall.

Example 10:

To find the median height of the plant.

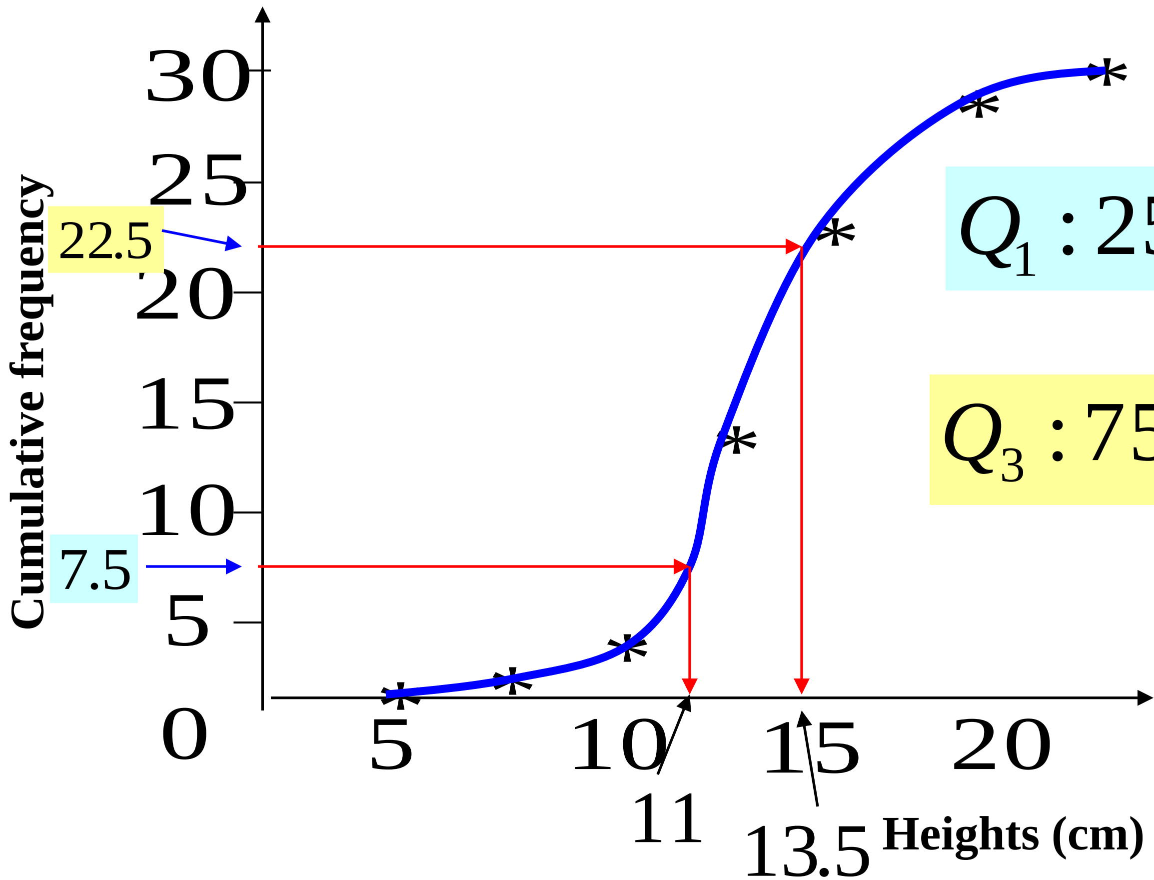


$$50\%(n) = 0.5(30) = 15$$

Median height for the plants is l2 cm, so $\tilde{x} = P_{50} = 12$

Example 10:

Find the Q_1 and Q_3 .



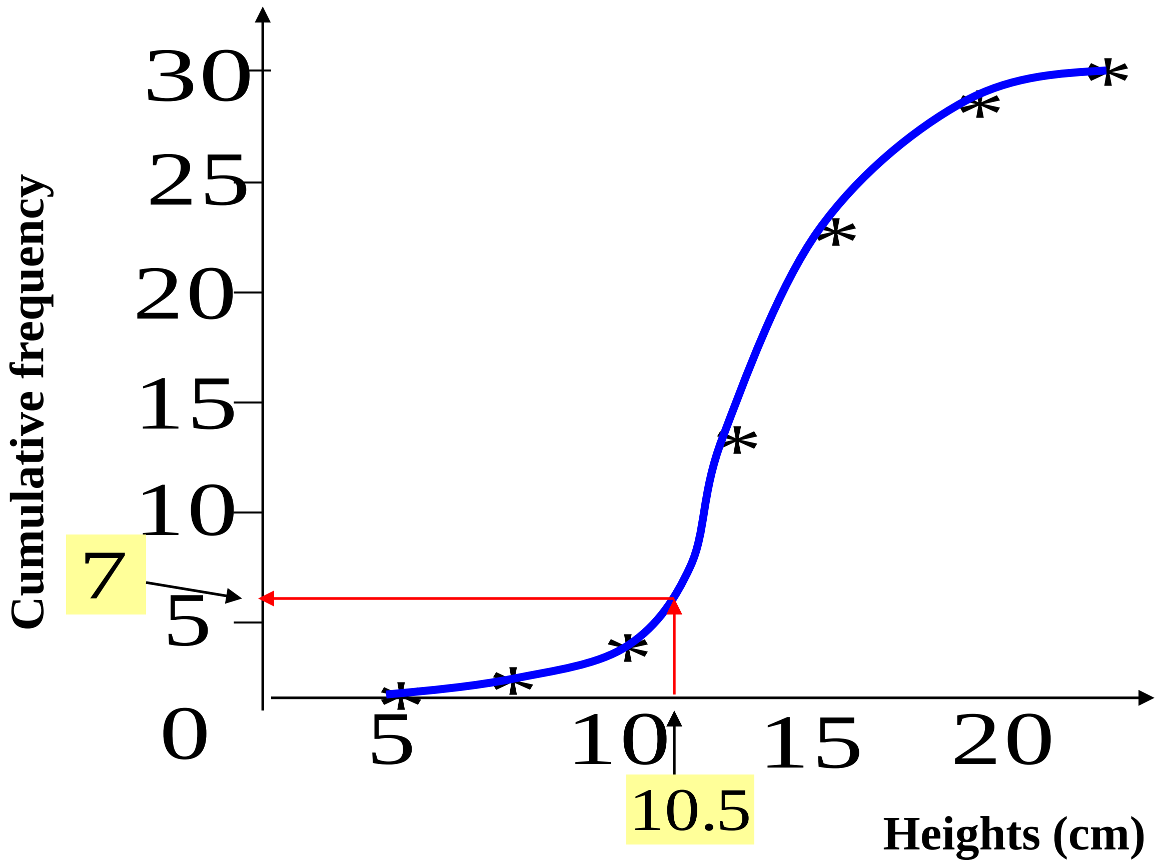
$$Q_1 : 25\% \text{ of } 30 = 7.5$$

$$Q_3 : 75\% \text{ of } 30 = 22.5$$

$$Q_1 = 11 \text{ and } Q_3 = 13.5$$

Example 10

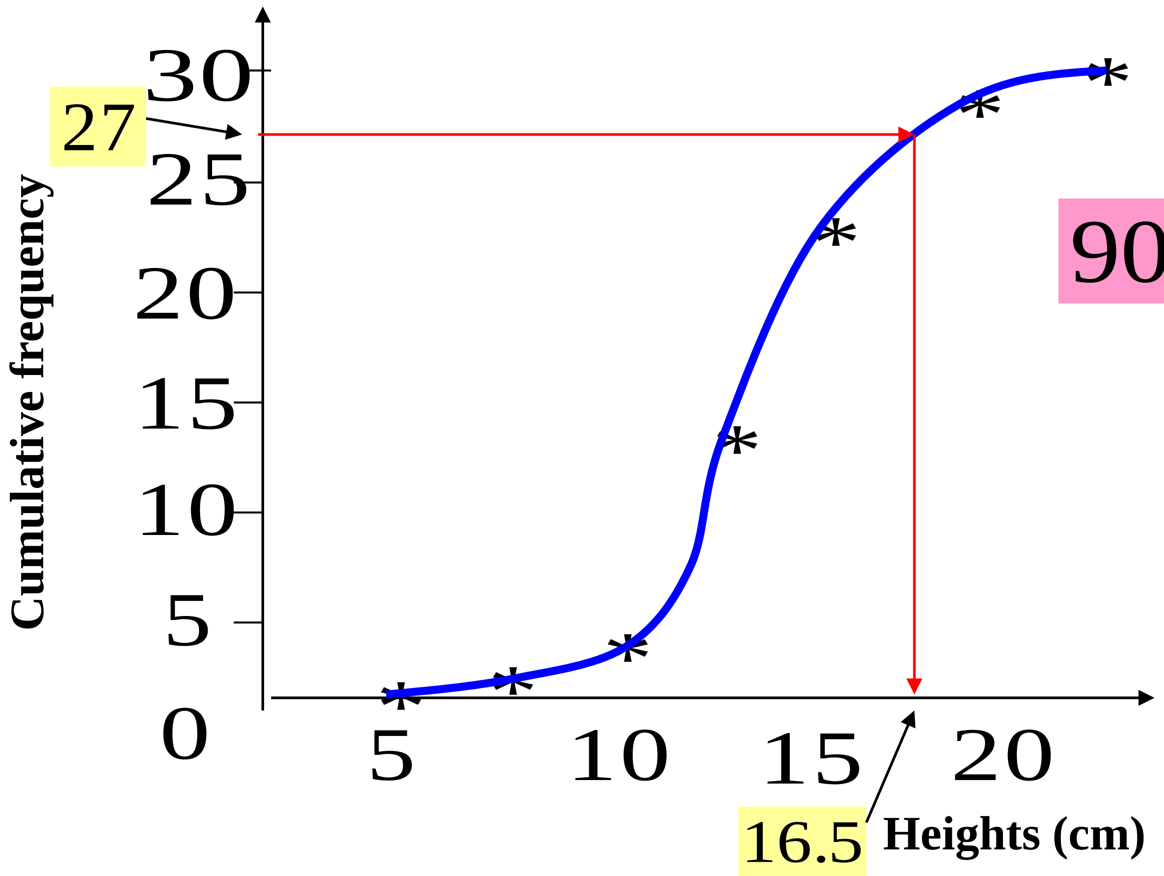
(i) To find the number of plants that were less than 10.5 cm tall:



From the graph, 7 plants were less than 10.5 cm tall.

Example 10:

(ii) To find x where 90% of the plants were less than x cm tall.



$90\% \text{ of } 30 = 27$

27 plants were less than 16.5 cm tall, so *x* = 16.5.

Practice Exercise

Advanced Level Mathematics

Statistics 1 Steve Dobbs and Jane Miller

Exercise 3A (Page 47)

1, 3, 5

Exercise 3B (Page 53)

6

Exercise 3C (Page 57)

2, 8

S1/1/7:
Box-and-whisker plots

Learning Outcome

Students should be able to:

- **Construct box-and-whisker plots**
- **Interpret box-and-whisker plots**

5-Number Summary

<u>5-Number Summary</u>:

The 5-number summary is composed of:

1. L, the smallest value in the data set
2. Q_1, the first quartile
3. $\tilde{x}$, the median (2nd quartile)
4. Q_3, the third quartile
5. H, the largest value in the data set

Notes:

- The 5-number summary indicates how much the data is spread out in each quarter.

Outliers

Sometimes unusually high or low values occur in a set of data.

Outlier are values > upper fence or values < lower fence.

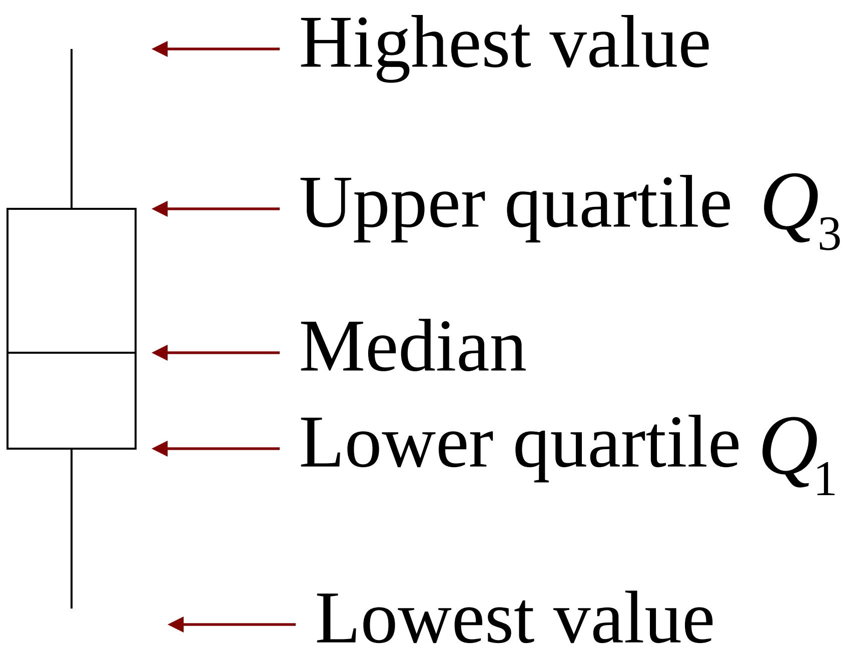
$$\text{upper fence} = Q_3 + 1.5(Q_3 - Q_1)$$

$$\text{lower fence} = Q_1 - 1.5(Q_3 - Q_1)$$

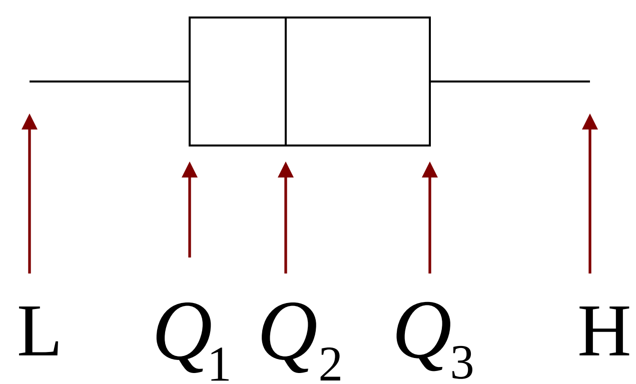
Box-and-Whisker Display

Box-and-Whisker Display illustrates the dispersion, or spread of the distribution.

Vertically

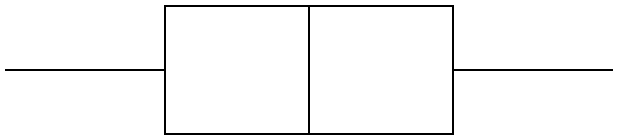


Horizontally



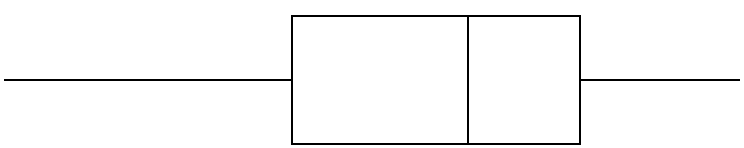
Box-and-Whisker Display

A box plot for a symmetrical distribution:



For a positively skewed distribution:

For a negatively skewed distribution:



Example 11:

A random sample of students in a sixth grade class was selected. Their weights are given in the table below. Find the 5-number summary for this data and construct a boxplot:

63	64	76	76	81	83	85	86	88
89	90	91	92	93	93	93	94	97
99	99	99	101	108	109	112		

Example 11:

63	64	76	76	81	83	85	86	88
89	90	91	$\boxed{92}$	93	93	93	94	97
99	99	99	101	108	109	112		

$$Q_2 = \frac{25+1}{2} th = 13th$$

$$Q_2 = 92$$

$$\text{Remaining } \frac{25-1}{2} = 12 \text{ numbers (first half)}$$

$$Q_1 = 84$$

$$Q_1 = \frac{12+1}{2} th = 6.5th$$

$$Q_3 = 99$$

Solution:

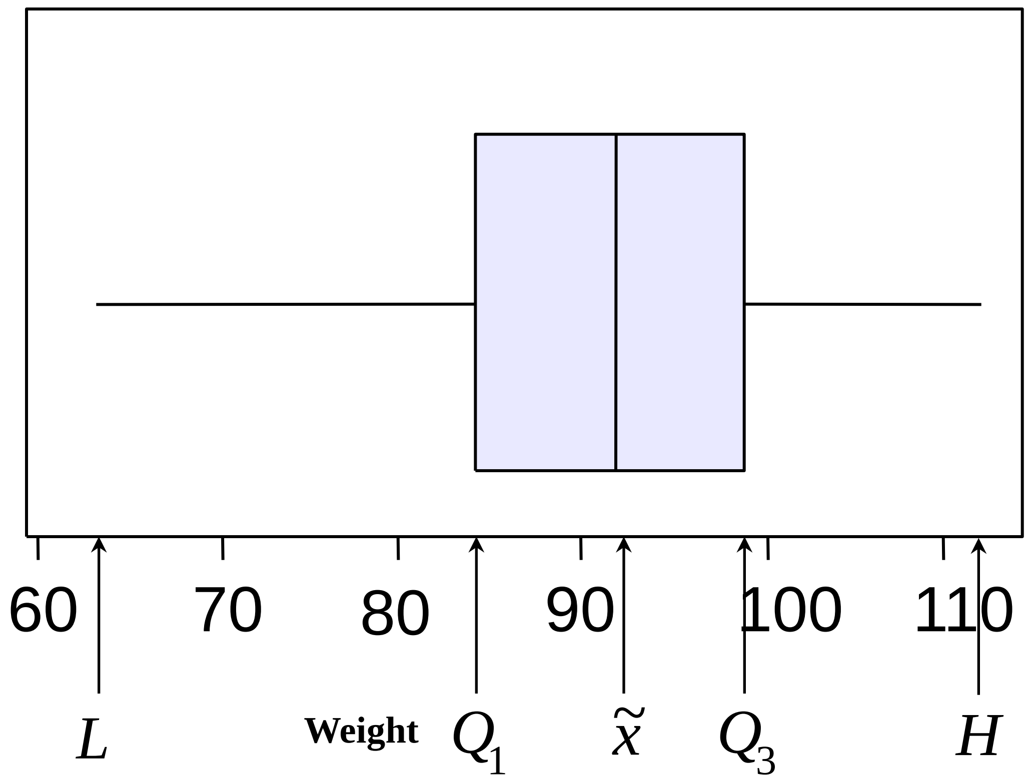
63	84	92	99	112
L	Q_1	$\tilde{x}$	Q_3	H

Example 11:

63	84	92	99	112
L	Q_1	$\tilde{x}$	Q_3	H

Boxplot for Weight Data

Weights from Sixth Grade Class



Example 12:

A class of pupils played a computer game which tested how quickly they reacted to a visual instruction to press a particular key. The computer measured their reaction times in tenths of a second and stored a record of the sex and reaction time of each pupil. Finally it displayed the following summary statistics for the whole class.

	Median	Lower quartile	Upper quartile	Min	Max
Girls	10	8	15	6	19
Boys	10	7	13	4	16

Example 12:

	Median	Lower quartile	Upper quartile	Min	Max
Girls	10	8	15	6	19
Boys	10	7	13	4	16

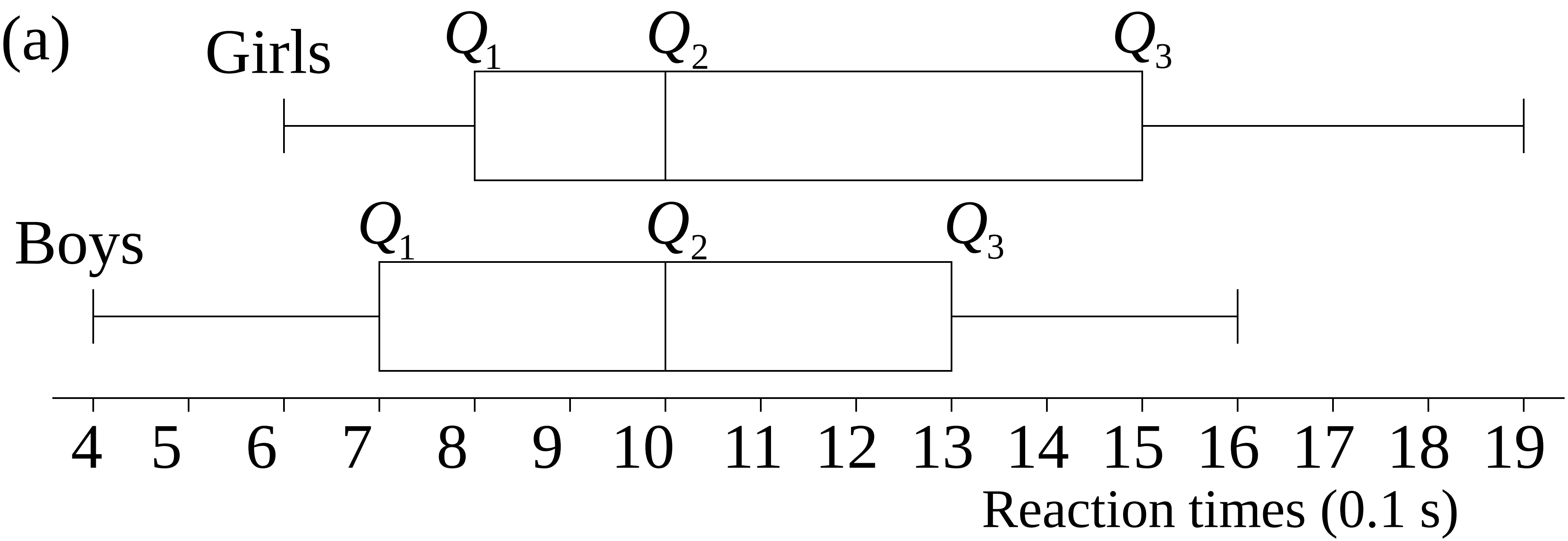
(a) Draw two box plots suitable for comparing the reaction times of boys and girls.

(b) Write a brief comparison of the performance of boys and girls in this game.

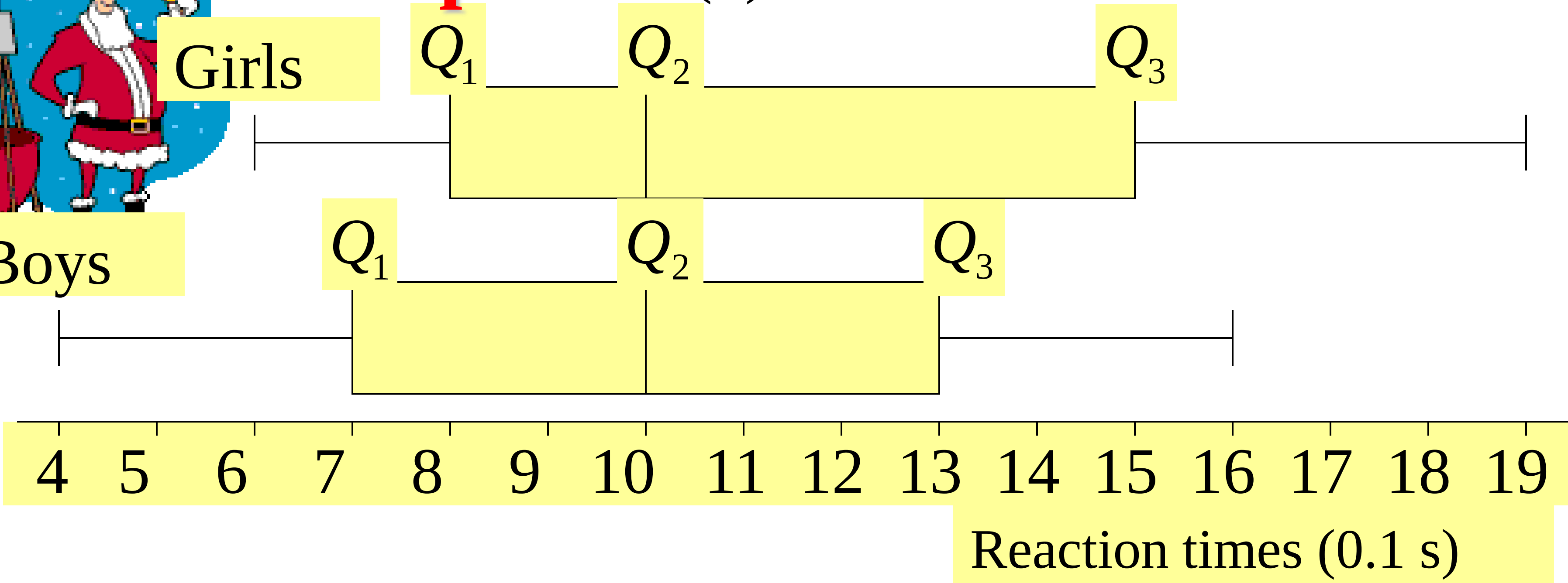
Example 12:

	Median	Lower quartile	Upper quartile	Min	Max
Girls	10	8	15	6	19
Boys	10	7	13	4	16

(a)



Example 12: (b)



The median reaction time for boys and girls is the same ($10 \times 0.1 = 1$ second).

The times for the boys are more evenly distributed, with a smaller range.

There are bigger spread of times for girls and their distribution is positively skewed.

→The boys have the faster reaction time.

Practice Exercise

Advanced Level Mathematics

Statistics 1 Steve Dobbs and Jane Miller

Exercise 3A (Page 49)

10, 12

S1/1/11:

Mean and standard deviation for summarized data

Learning Outcome

Students should be able to:

- Calculate mean and standard deviation from given totals.
- Using a method of coding to find the mean and standards deviation.

The method of coding

In general, if the set of numbers $x_1, x_2, \ldots, x_n$ is transformed to the set of numbers $y_1, y_2, \ldots, y_n$ by means of the coding

$$y = \frac{x - a}{b}$$

then $x = a + by$

so $\bar{x} = a + b\bar{y}$

and $s_x = bs_y$

Example 13:

Salt is packed in bags which the manufacturer claims contain 25 kg each. Eighty bags are examined and the mass, *x* kg, of each is found. The results are

$$\sum(x-25)=27.2, \qquad \sum(x-25)^2=85.1.$$

Find the mean and the standard deviation of the masses.

25.34; 0.9737

Example 14:

Using the coding

$$y = \frac{x - 200\,000}{25\,000}$$

to find the mean and the standard deviation of the following:

x	125 000	150 000	175 000	200 000	225 000	250 000	275 000
f	5	19	27	35	24	12	3

195400;34800

Practice Exercise

Advanced Level Mathematics

Statistics 1 Steve Dobbs and Jane Miller

Exercise 3C (Page 57)

6

Practice Examination 2 (Page 170)

6