Chapter 3

Collecting, presenting and summarsing data

3. Data: collection, presentation and summaries

Data are the key to many important management decisions.

- Is a new product selling well?
- Do potential customers like the new advertising campaign?

These and many other questions can be answered with **data**.

What we'll cover...

- The different **kinds of data** that exist
- How to **get** data
- How to **look at** the data we've got
- How to **summarise** our data numerically

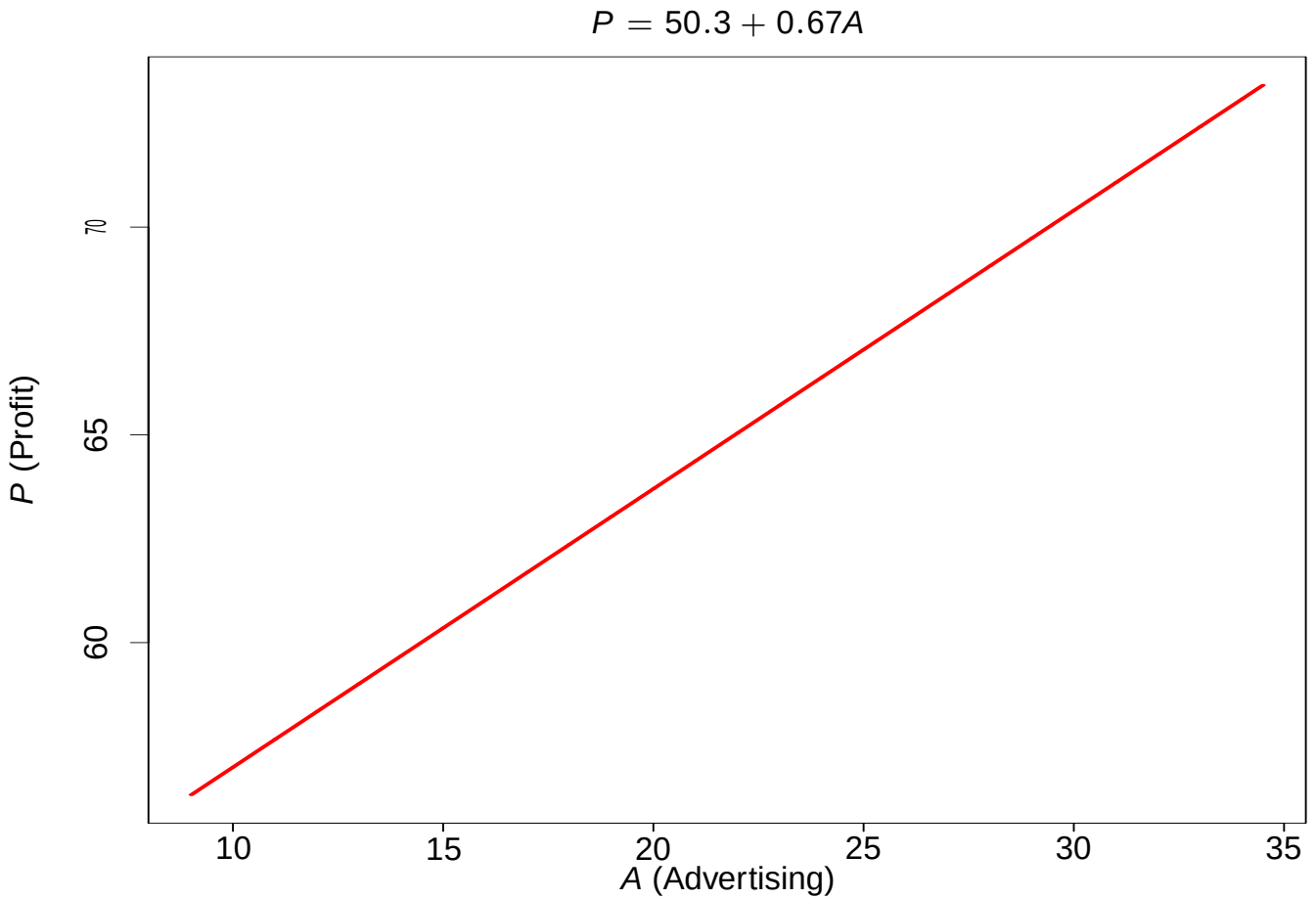
Example: Sizing clothes

- Most clothing now comes in essentially standard sizes . . .
- but where do these standards come from?
- By sampling from the population as a whole, standards can be set around the most common sizes.
- We cannot say that an individual is *exactly* a standard size;
- However, we *can* say that they will *probably* fall within a range either side of a standard size.

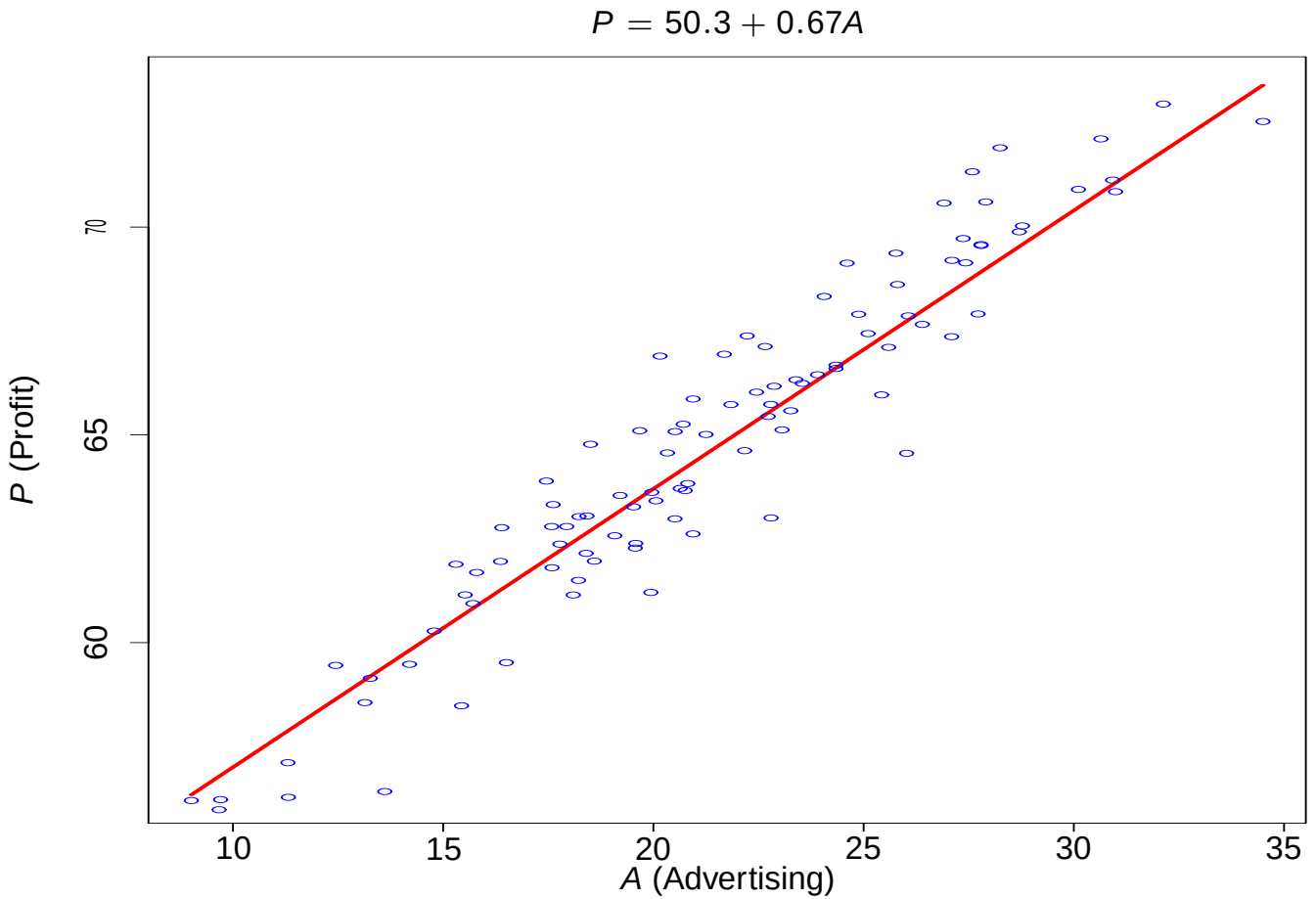
Example: Car maintenance

- Just imagine you're buying a new car . . .
- ...it might be useful to know how much it will cost to run over the next few years.
- Obviously this cannot be predicted exactly –
  - each car will be slightly different;
  - each user will be slightly different.
- Collecting data from people who have bought similar cars will give us some idea!

Example: Profit versus advertising



Example: Profit versus advertising



3.1 Important definitions

The quantities measured in a study are called **random variables**.

A particular outcome is called an **observation**.

A collection of such observations is the **data**.

The collection of all possible outcomes is the **population**.

3.1 Important definitions

Suppose we are interested in the height of students on A&F courses at Newcastle . . .

- Our **random variable** is “the height of students on A&F courses at Newcastle”.
- If Joe Bloggs is an A&F student, and we measured his height, then that value would be a single **observation**.
- If we measured the height of every first year A&F student, we would have a collection of such observations which would be our **data**.
- This would be a **sample** from the **population** which consists of all students registered on A&F degrees.

3.1 Important definitions

Ideally, to get a true idea of what is going on, we'd like to observe the whole population (take a **census**). However, this can be difficult:

- If the population is huge, then this would take ages!
- *And* it would be very costly!
- In reality, we usually observe a subset of the population . . . but how do we choose who to observe?

3.1 Important definitions (Section 2.1, summer revision booklet)

There are two types of variables: **qualitative** and **quantitative**.

- *Qualitative* variables have non–numerical outcomes, and are usually **categorical**
  - Sex of a person (categories: male or female);
  - Colour of a car (categories: red, black, silver, blue, ... – could code this to 1, 2, 3, 4, ...);
  - Mode of transport.
- *Quantitative* variables have numerical outcomes with a natural ordering
  - people's height
  - number of defective components in a batch

3.1 Important definitions (Section 2.1, summer revision booklet)

We are most interested in **quantitative** variables.

These variables can be subdivided into two types:

discrete and **continuous**.

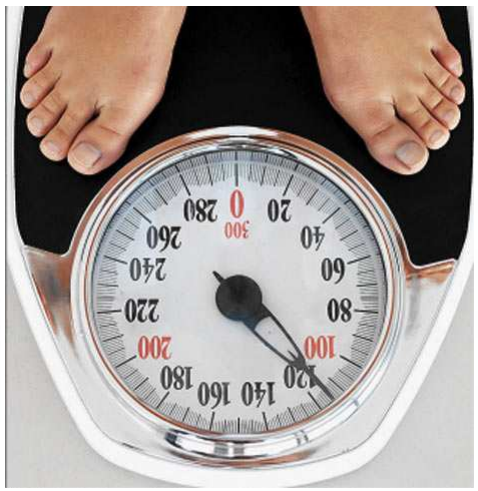
Discrete random variables

- can only take a sequence of distinct values (usually integers);
- are usually **countable** – e.g. the number of people attending a tutorial group;
- can be **ordinal** – where the outcomes are ordered.

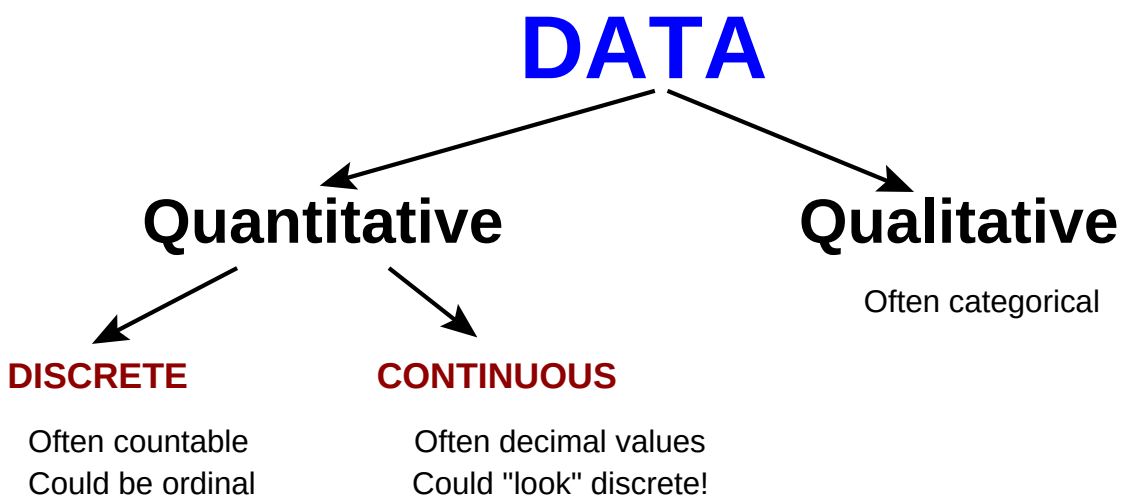
3.1 Important definitions (Section 2.1, summer revision booklet)

Continuous random variables

- can take any value over some continuous scale – e.g. height or weight.
- can be **measured** to a very high degree of accuracy (provided we have the equipment to do so) . . .
- . . . however, we can never say *precisely* how much someone weighs, for example,
- might be measured to the nearest whole number – and so could "look" discrete – be careful!



3.1 Important definitions



3.2 Sampling

We can rarely observe the whole population. Instead we observe some sub–set of this (called the **sample**).

The difficulty is in obtaining a **representative** sample.

For example, if you were to ask people leaving a gym if they took exercise this would produce a **biased** sample and would not be representative of the population as a whole.

The importance of obtaining a representative sample cannot be stressed too highly.

3.2 Sampling

There are three general forms of sampling techniques:

1. ***Random sampling*** – where the members of the sample are chosen by some random (i.e. unpredictable) mechanism.
2. ***Quasi–random sampling*** – where the mechanism for choosing the sample is only partly random.
3. ***Non–random sampling*** – where the sample is specifically selected rather than randomly selected.

3.2.1 Simple random sampling...

This method is the **simplest to understand**.

If we had a population of 200 students we could put all their names into a hat and draw out 20 names as our sample.

We cannot predict with certainty which 20 names will be drawn, and so the sample is unpredictable, and therefore completely **random**.

Each name has an **equally likely chance** of being drawn.

Furthermore each possible **sample of 20** has an equal chance of being selected.

3.2.1 Simple random sampling...

In reality the drawing of the names would be done by a computer and the population and samples would be considerably larger.

… and some disadvantages

We often don’t have a **complete list** of the population.

For example, if you were surveying the market for some new software, the population would be everybody with a compatible computer!

Not all elements of the population are **equally accessible**.

Purely by chance, you could pick an **unrepresentative sample**!

3.2.2 Stratified sampling...

This is a form of random sample where clearly defined groups, or **strata**, exist within the population.

If we know the overall proportion of the population that falls into each of these groups, we can take a simple random sample from each of the groups and then adjust the results according to the known proportions.

For example, if we assume that the population is 55% female and 45% male and we wanted a sample of 1000, we would

- decide to have 550 females and 450 males in our sample, and then
- pick the members of our sample from their respective groups randomly.

... and some disadvantages

We need clear information on the **size** and **composition** of each stratum which can be difficult to obtain.

We still need a **list of the entire population** to sample from.

3.2.3 Systematic sampling...

This is a form of **quasi–random sampling** which can be used when the population is clearly structured.

For example, if you were interested in obtaining a 10% sample from a batch of components being manufactured, you would

- select the first component at random, and then
- pick every 10th item to come off the production line after that

This scheme is **very easy to implement**!

... and some disadvantages

This method is **not entirely random**!

If there is a pattern in the process, it is very easy to obtain a **biased sample**.

Other types of sampling

Read through Sections 3.2.4 to 3.2.8 of the notes before your tutorial session next week.

- Multi-stage sampling
- Cluster sampling
- Judgemental sampling
- Accessibility sampling
- Quota sampling

3.2.9 Sample size

- Larger samples will generally give more precise information about the population.
- Expense and time tend to limit the size of the sample.
- E.g. national opinion polls often rely on samples in the region of just 1000.

3.3 Frequency tables

- After collecting data, the first stage of any analysis is to present them in a simple, easily understood way.
- **Tables** are perhaps the simplest means of presenting data.
- Tables can be informative, but some tables can be difficult to interpret, especially if they contain vast amounts of data.

3.3 Frequency tables

- **Frequency tables** are amongst the most commonly-used tables and are perhaps the easiest to understand.
- They can be used with continuous, discrete, categorical and ordinal data.
- Frequency tables have uses in some of the techniques we will see in the next lecture.

3.3.1 Frequency tables for categorical data

The following table presents the modes of transport used daily by 30 students to get to and from University.

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

3.3.1 Frequency tables for categorical data

The table obviously contains much information. However, it is difficult to see which method of transport is the most widely used.

Idea: count the number of students using each mode of transport!

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

This gives us a much clearer picture of the methods of transport used.

3.3.1 Frequency tables for categorical data

Also of interest might be the ***relative*** frequency of each of the modes of transport.

This is simply the frequency expressed as a proportion of the total number of students surveyed. If this is given as a percentage, as here, this is known as the **percentage relative frequency**.

Mode	Frequency	Relative Frequency (%)
Car	10	33.3
Walk	7	23.4
Bike	4	13.3
Bus	4	13.3
Metro	4	13.3
Train	1	3.4
Total	30	100

3.3.2 Frequency tables for count data

The following table shows the raw data for car sales at a new car showroom over a two week period in July.

Date	Cars Sold	Date	Cars Sold
1st July	9	8th July	10
2nd July	8	9th July	5
3rd July	6	10th July	8
4th July	7	11th July	4
5th July	7	12th July	6
6th July	10	13th July	8
7th July	11	14th July	9

Example 3.1

Present these data in a relative frequency table by number of days on which different numbers of cars were sold.

Solution to Example 3.1

Cars Sold	Tally	Frequency	Relative Frequency %
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
Totals			

Solution to Example 3.1

Cars Sold	Tally	Frequency	Relative Frequency %
1			
2			
3			
4	I		
5	I		
6	II		
7	II		
8	III		
9	II		
10	II		
11	I		
Totals	14		

Solution to Example 3.1

Cars Sold	Tally	Frequency	Relative Frequency %
1		0	
2		0	
3		0	
4	I	1	
5	I	1	
6	II	2	
7	II	2	
8	III	3	
9	II	2	
10	II	2	
11	I	1	
Totals	14	14	

Solution to Example 3.1

Cars Sold	Tally	Frequency	Relative Frequency %
1		0	0
2		0	0
3		0	0
4	I	1	7.14
5	I	1	7.14
6	II	2	14.29
7	II	2	14.29
8	III	3	21.43
9	II	2	14.29
10	II	2	14.29
11	I	1	7.14
Totals	14	14	100

3.3.3 Frequency tables for continuous data

With discrete data it is easy to count the quantities in the defined categories. With **continuous data** this is not possible.

For example, the following data set represents the service time in seconds for callers to a credit card call centre.

214.8	220.6	216.7	195.1	211.4
195.8	201.1	185.8	183.4	178.8
196.3	199.7	206.7	203.8	203.1
200.8	201.3	205.6	181.6	201.7
180.2	193.3	188.2	199.9	204.7
198.3	193.1	204.0	197.2	193.5
205.5	217.5	208.8	197.7	212.3
209.9	197.6	204.9	203.1	192.9
208.9	202.0	195.0	192.7	219.8
208.8	200.7	191.9	188.8	206.8

3.3.3 Frequency tables for continuous data

This is what we do...

1. Divide the range of the variable into smaller ranges called **class intervals**.
2. There should be **no gaps** between these intervals.
3. The class interval width should be a convenient number (e.g. 5, 10, 100, depending on the data).
4. You should aim for no more than about **ten to fifteen classes**.

Example 3.2

Create a frequency table for the call centre data.

Also, find the relative frequencies in each class interval.

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq \text{time} < 180$			
$180 \leq \text{time} < 185$			
$185 \leq \text{time} < 190$			
$190 \leq \text{time} < 195$			
$195 \leq \text{time} < 200$			
$200 \leq \text{time} < 205$			
$205 \leq \text{time} < 210$			
$210 \leq \text{time} < 215$			
$215 \leq \text{time} < 220$			
$220 \leq \text{time} < 225$			
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180			
$180 \leq$ time < 185			
$185 \leq$ time < 190			
$190 \leq$ time < 195			
$195 \leq$ time < 200			
$200 \leq$ time < 205			
$205 \leq$ time < 210			
$210 \leq$ time < 215	I		
$215 \leq$ time < 220			
$220 \leq$ time < 225			
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180			
$180 \leq$ time < 185			
$185 \leq$ time < 190			
$190 \leq$ time < 195			
$195 \leq$ time < 200			
$200 \leq$ time < 205			
$205 \leq$ time < 210			
$210 \leq$ time < 215	I		
$215 \leq$ time < 220			
$220 \leq$ time < 225	I		
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq \text{time} < 180$			
$180 \leq \text{time} < 185$			
$185 \leq \text{time} < 190$			
$190 \leq \text{time} < 195$			
$195 \leq \text{time} < 200$			
$200 \leq \text{time} < 205$			
$205 \leq \text{time} < 210$			
$210 \leq \text{time} < 215$	I		
$215 \leq \text{time} < 220$	I		
$220 \leq \text{time} < 225$	I		
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180			
$180 \leq$ time < 185			
$185 \leq$ time < 190			
$190 \leq$ time < 195			
$195 \leq$ time < 200	I		
$200 \leq$ time < 205			
$205 \leq$ time < 210			
$210 \leq$ time < 215	I		
$215 \leq$ time < 220	I		
$220 \leq$ time < 225	I		
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180			
$180 \leq$ time < 185			
$185 \leq$ time < 190			
$190 \leq$ time < 195			
$195 \leq$ time < 200	I		
$200 \leq$ time < 205			
$205 \leq$ time < 210			
$210 \leq$ time < 215	II		
$215 \leq$ time < 220	I		
$220 \leq$ time < 225	I		
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180	I		
$180 \leq$ time < 185	III		
$185 \leq$ time < 190	III		
$190 \leq$ time < 195	$\cancel{\text{IIII}}$ I		
$195 \leq$ time < 200	$\cancel{\text{IIII}}$ $\cancel{\text{IIII}}$		
$200 \leq$ time < 205	$\cancel{\text{IIII}}$ $\cancel{\text{IIII}}$ II		
$205 \leq$ time < 210	$\cancel{\text{IIII}}$ III		
$210 \leq$ time < 215	III		
$215 \leq$ time < 220	III		
$220 \leq$ time < 225	I		
Totals			

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180	I	1	
$180 \leq$ time < 185	III	3	
$185 \leq$ time < 190	III	3	
$190 \leq$ time < 195	𝍸 I	6	
$195 \leq$ time < 200	𝍸 𝍸	10	
$200 \leq$ time < 205	𝍸 𝍸 II	12	
$205 \leq$ time < 210	𝍸 III	8	
$210 \leq$ time < 215	III	3	
$215 \leq$ time < 220	III	3	
$220 \leq$ time < 225	I	1	
Totals		50	

Solution to Example 3.2

For the call centre data, we might get:

Class Interval	Tally	Frequency	Relative Frequency %
$175 \leq$ time < 180	I	1	2
$180 \leq$ time < 185	III	3	6
$185 \leq$ time < 190	III	3	6
$190 \leq$ time < 195	𝍸 I	6	12
$195 \leq$ time < 200	𝍸 𝍸	10	20
$200 \leq$ time < 205	𝍸 𝍸 II	12	24
$205 \leq$ time < 210	𝍸 III	8	16
$210 \leq$ time < 215	III	3	6
$215 \leq$ time < 220	III	3	6
$220 \leq$ time < 225	I	1	2
Totals		50	100

3.4 Graphical methods for presenting data

So far we have looked at

- different types of **data**,
- and how to summarise data in a **frequency table**

We now look at how to display data **graphically**.

Allows us to see the **main features** of a dataset quickly and clearly.

- Is the distribution symmetric or asymmetric?
- Are the data spread out or tightly concentrated?
- Do there appear to be groupings/clusters within the data?

3.4.1 Stem and leaf plots

Stem and leaf plots are a quick and easy way of representing data graphically.

They can be used with both **discrete** and **continuous** data.

A bit like constructing a frequency table – but carries visual information about the shape of the distribution.

Best seen via an example!

Example: Percentage returns on a share

The following numbers show the percentage returns on an ordinary share for 23 consecutive months:

0.2	**–2.1**	**1.0**	**0.1**	**–0.5**	**2.4**	**–2.3**	**1.5**
1.2	**–0.6**	**2.4**	**–1.2**	**1.7**	**–1.3**	**–1.2**	**0.9**
0.5	**0.1**	**–0.1**	**0.3**	**–0.4**	**0.5**	**0.9**	

- The largest value is 2.4 and the smallest –2.3, and we have lots of decimal values in between
- It seems sensible to have a **stem unit** of 1 and a **leaf unit** of 0.1

Example: Percentage returns on a share

A stem and leaf diagram for this set of returns might look like:

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

Stem **Leaf**

$n = 23$, stem unit $= 1$, leaf unit $= 0.1$.

Example: Production line data

If there is more than one significant figure in the data, the extra digits are **cut**, not **rounded** – e.g. 2.97 would be "cut" to 2.9.

Consider the following data on lengths of items on a production line (in cm):

2.97	**3.81**	**2.54**	**2.01**	**3.49**
3.09	**1.99**	**2.64**	**2.31**	**2.22**

The stem and leaf plot for this is as follows:

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

$n = 10$, stem unit $= 1$ cm, leaf unit $= 0.1$ cm.

Example: Production line data

Why do you think we **cut** the extra digits?

So values such as 2.97 do not end up in the wrong row!

Example 3.3

The observations in the table below are the recorded time it takes to get through to an operator at a telephone call centre (in seconds).

54	56	50	67	55	38	49	45	39	50
45	51	47	53	29	42	44	61	51	50
30	39	65	54	44	54	72	65	58	62

Stem	**Leaf**
2	9
3	0 8 9 9
4	2 4 4 5 5 7 9
5	0 0 0 1 1 3 4 4 4 5 6 8
6	1 2 5 5 7
7	2

$n = 30, \quad$ stem unit $= 10, \quad$ leaf unit $= 1$

Example 3.4

The stem and leaf plot below represents the marks on a test for 52 students. Comment on the distribution of these marks.

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

$n = 52$, stem unit $= 10$, leaf unit $= 1$.

Solution to Example 3.4

- **Asymmetric distribution** of marks – in fact, marks are **negatively skewed**
- Majority of students achieve marks in the 30s/40s
- The **modal class** is 35–39
- Large **spread** of marks

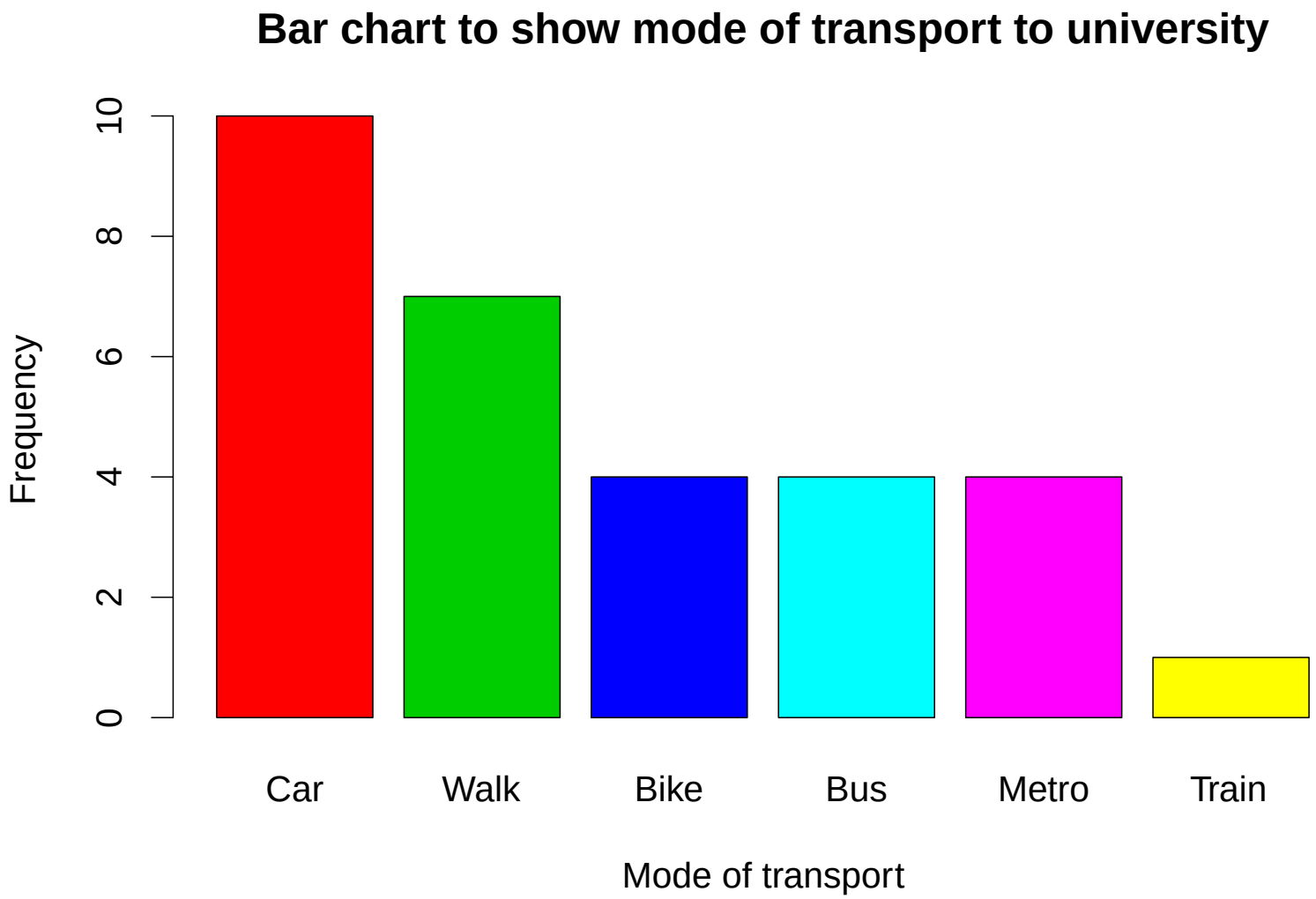
3.4.2 Bar charts

Bar charts are a commonly–used and clear way of presenting categorical data or ungrouped discrete frequency observations. See **section 2.3.2** of the **summer revision booklet**.

Recall the example on students' modes of transport:

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

3.4.2 Bar charts



3.4.3 Pie charts

These charts are commonly used within industry to communicate simple ideas, for example market share.

They are used to show the proportions of a whole.

They are best used when there are only a handful of categories to display.

See Section 2.3.3 of the summer revision booklet for more details.

3.4.4 Histograms

Bar charts have their limitations, one of which is that they cannot be used to present **continuous data**.

When dealing with continuous random variables a different kind of graph is required – one such graph is the **histogram**.

3.4.4 Histograms

Consider the following data, which show service times (in seconds) for a telephone call centre.

214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	212.3491
209.9000	197.6215	204.9101	203.1654	192.9706
208.9901	202.0090	195.0241	192.7098	219.8277
208.8920	200.7965	191.9784	188.8587	206.8912

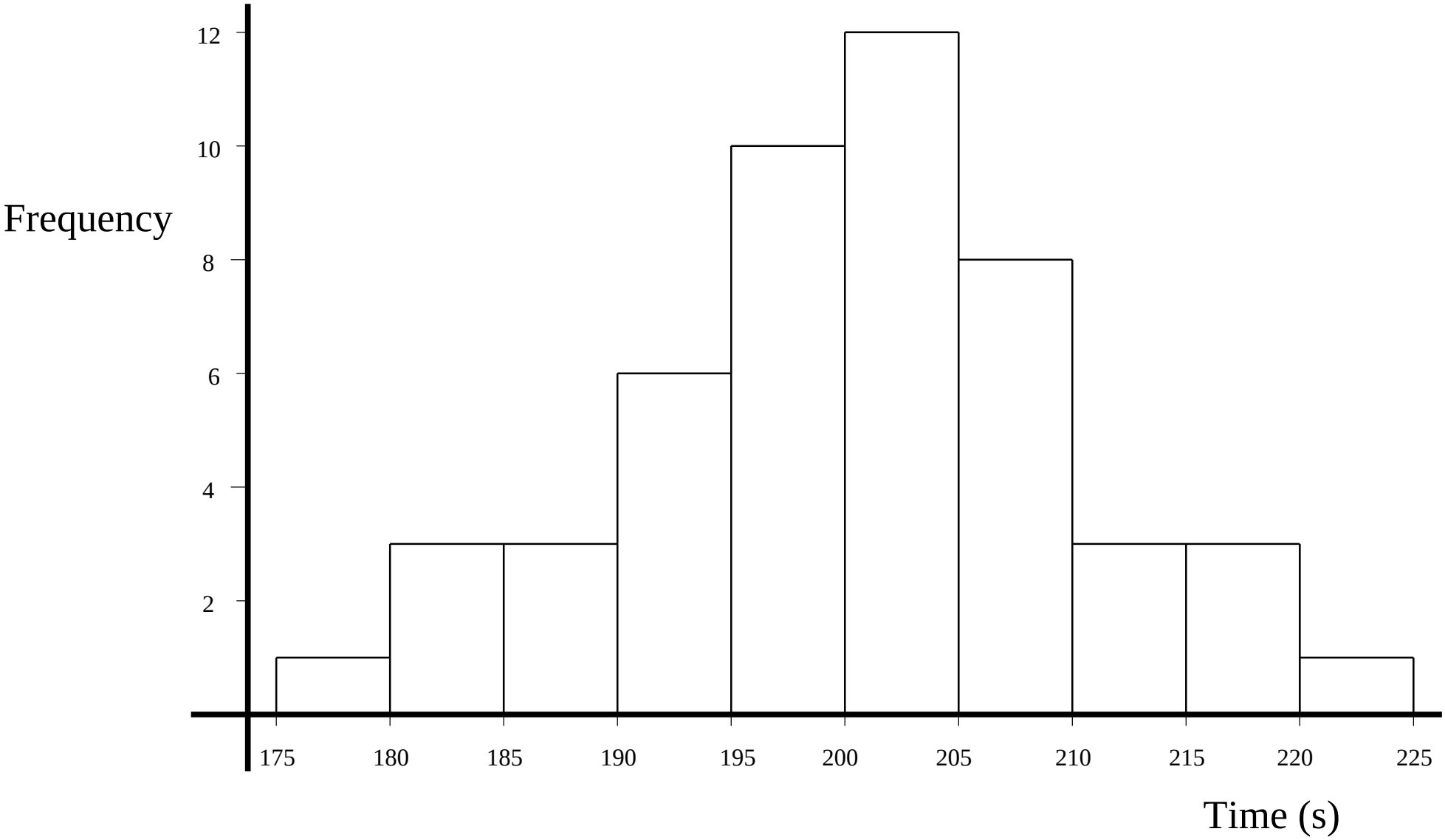
3.4.4 Histograms

Producing a histogram is much like producing a bar chart.

It is often best to produce a frequency table first which collects all the data together in an ordered format.

Service time	Frequency
$175 \leq \text{time} < 180$	1
$180 \leq \text{time} < 185$	3
$185 \leq \text{time} < 190$	3
$190 \leq \text{time} < 195$	6
$195 \leq \text{time} < 200$	10
$200 \leq \text{time} < 205$	12
$205 \leq \text{time} < 210$	8
$210 \leq \text{time} < 215$	3
$215 \leq \text{time} < 220$	3
$220 \leq \text{time} < 225$	1
Total	50

3.4.4 Histograms



3.4.4 Histograms

At first sight histograms look similar to bar charts. There are, however, some critical differences:

- The horizontal (x-axis) is a **continuous scale**
- As a result there are **no gaps between the bars**
- Strictly speaking, the **area** of the bar is proportional to the frequency.

Example 3.5

The *Holiday Hypermarket* travel agency received 64 telephone calls yesterday morning. The table below gives information of the lengths, in minutes, of these telephone calls.

Length (x) minutes	**Frequency**	**Frequency density**
$0 \leq x < 5$	4	0.8
$5 \leq x < 15$	10	1.0
$15 \leq x < 30$	24	
$30 \leq x < 40$	20	
$40 \leq x < 45$	6	

Complete this table, and construct a histogram for these data. What is the *modal class* here?

Solution to Example 3.5 (1/3)

First, the **frequency densities**, or bar heights.

Now we know that the area of the bar represents frequency.

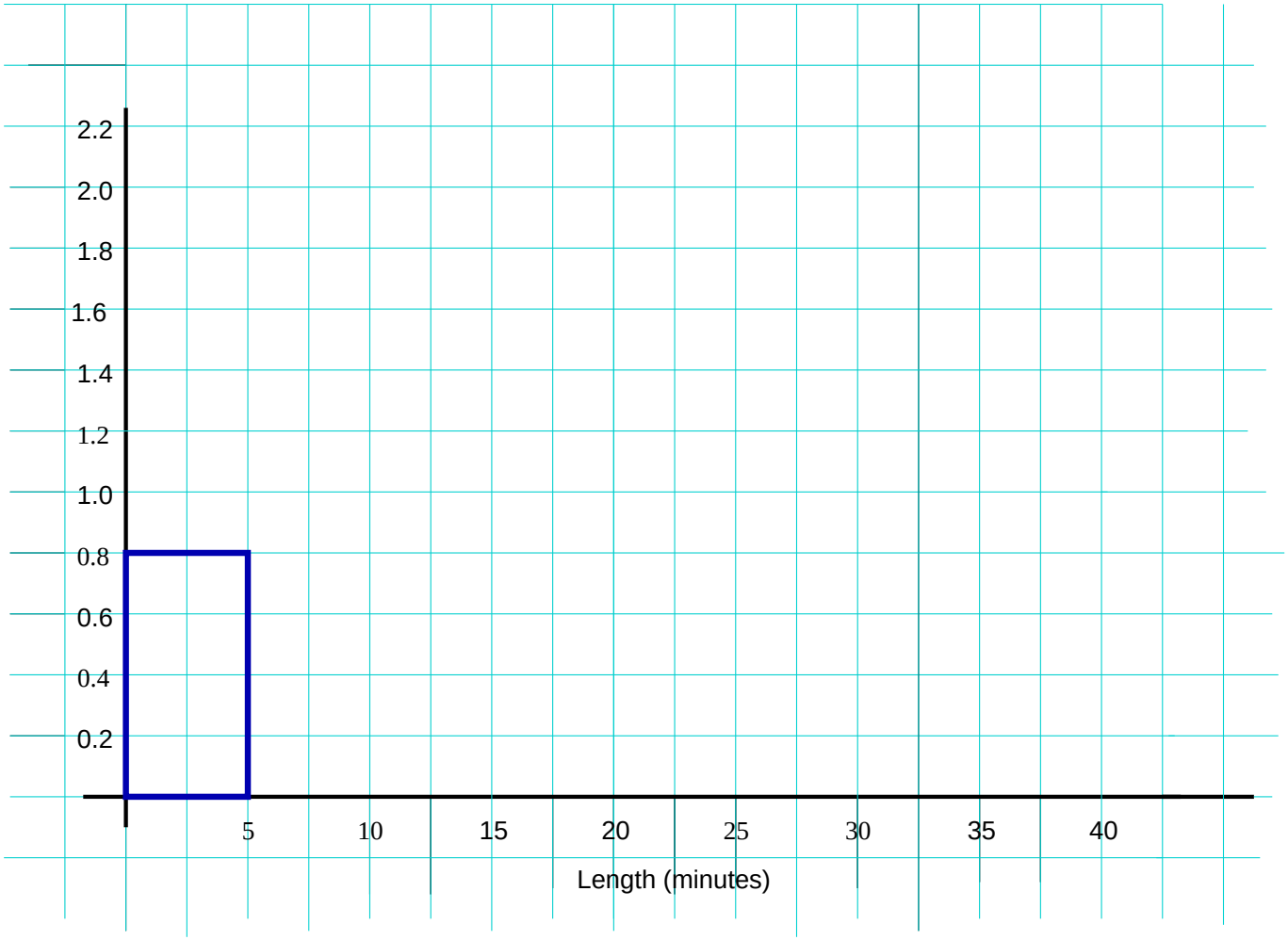
The width of the first is 5, and the area must be 4, so

$$\begin{aligned} \text{Area} &= \text{Base} \times \text{Height} \\ 4 &= 5 \times \text{Height} \end{aligned}$$

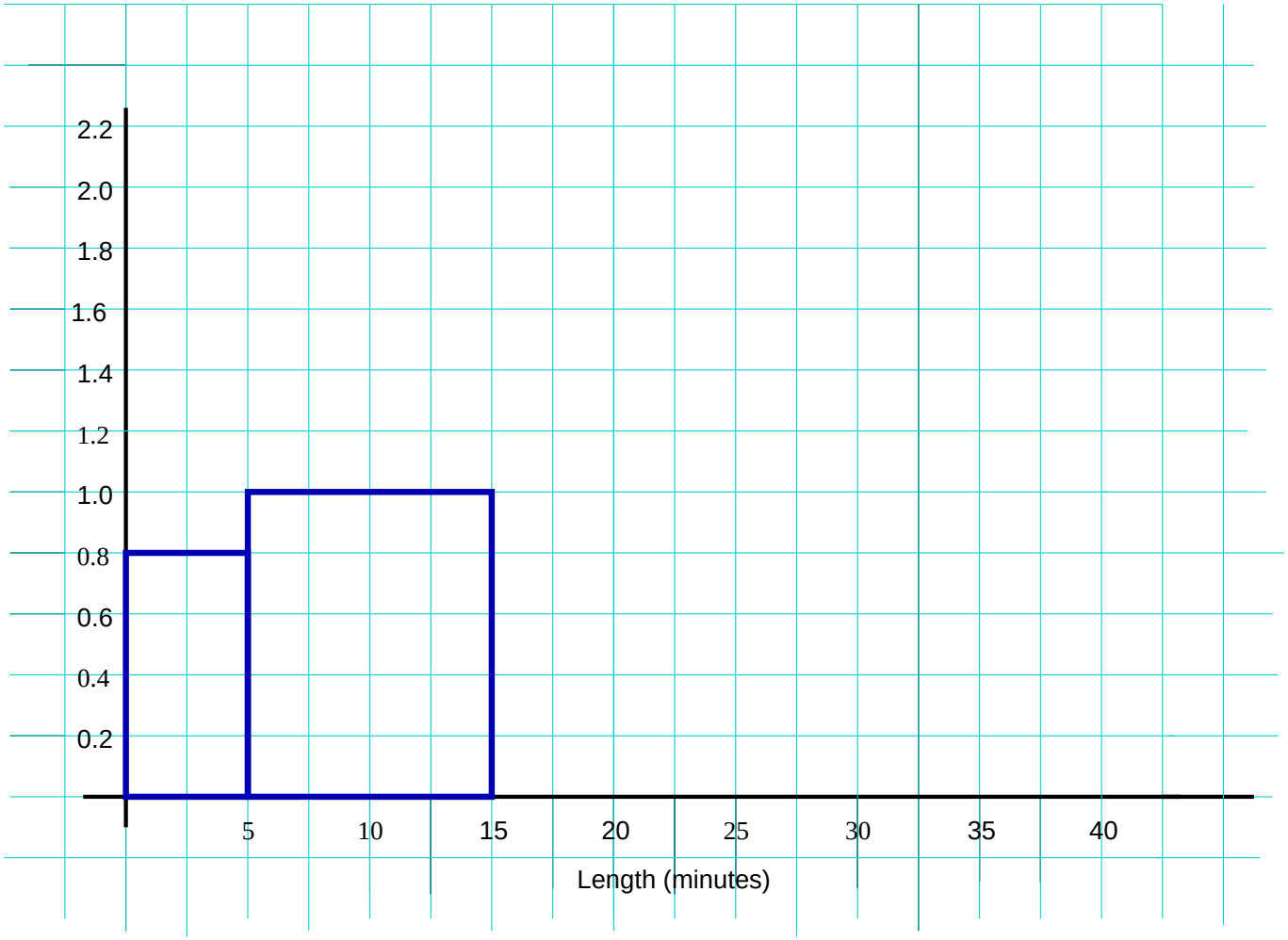
and so

$$\text{Height} = 4/5 = 0.8.$$

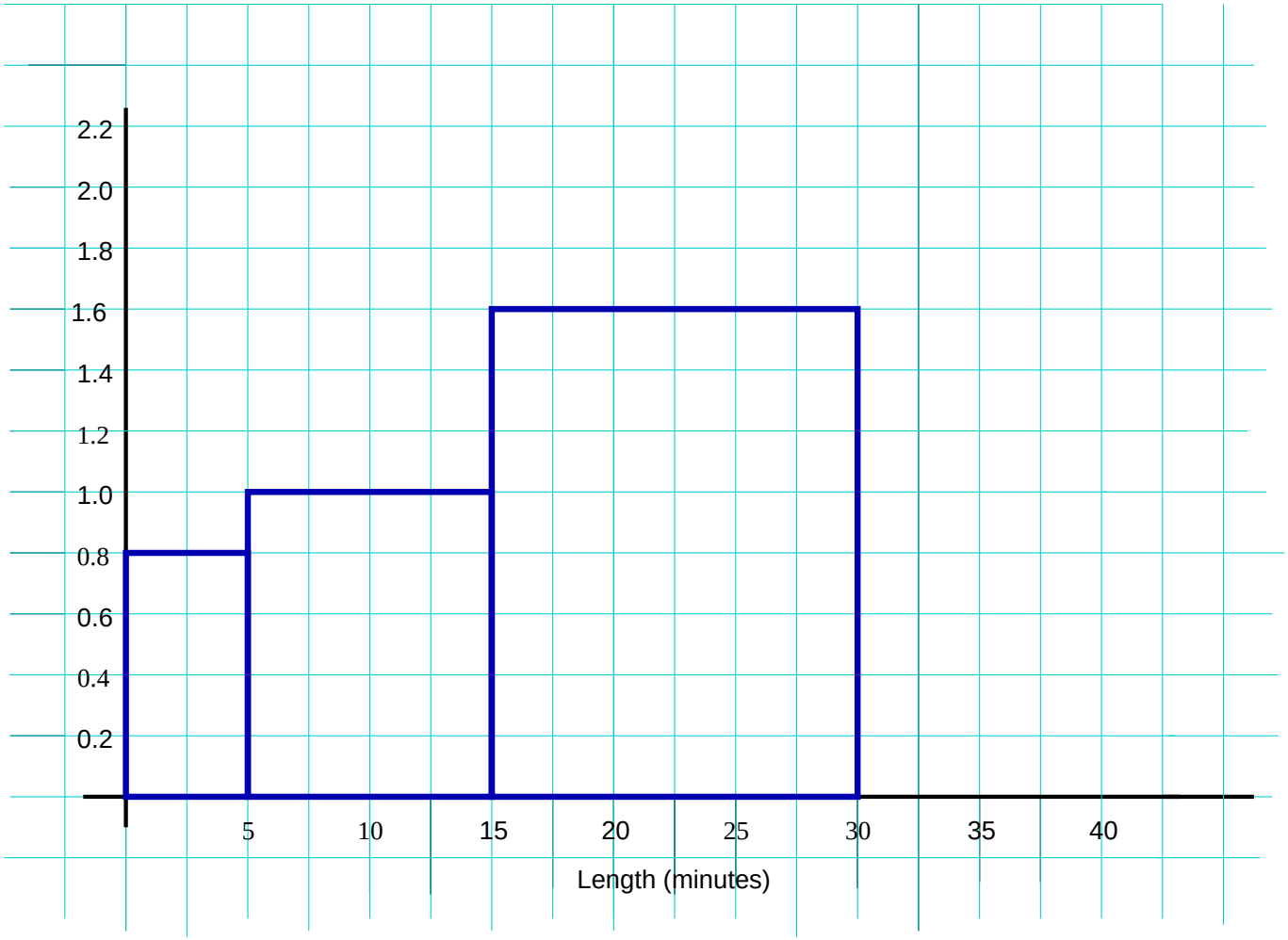
Solution to Example 3.5 (3/3)



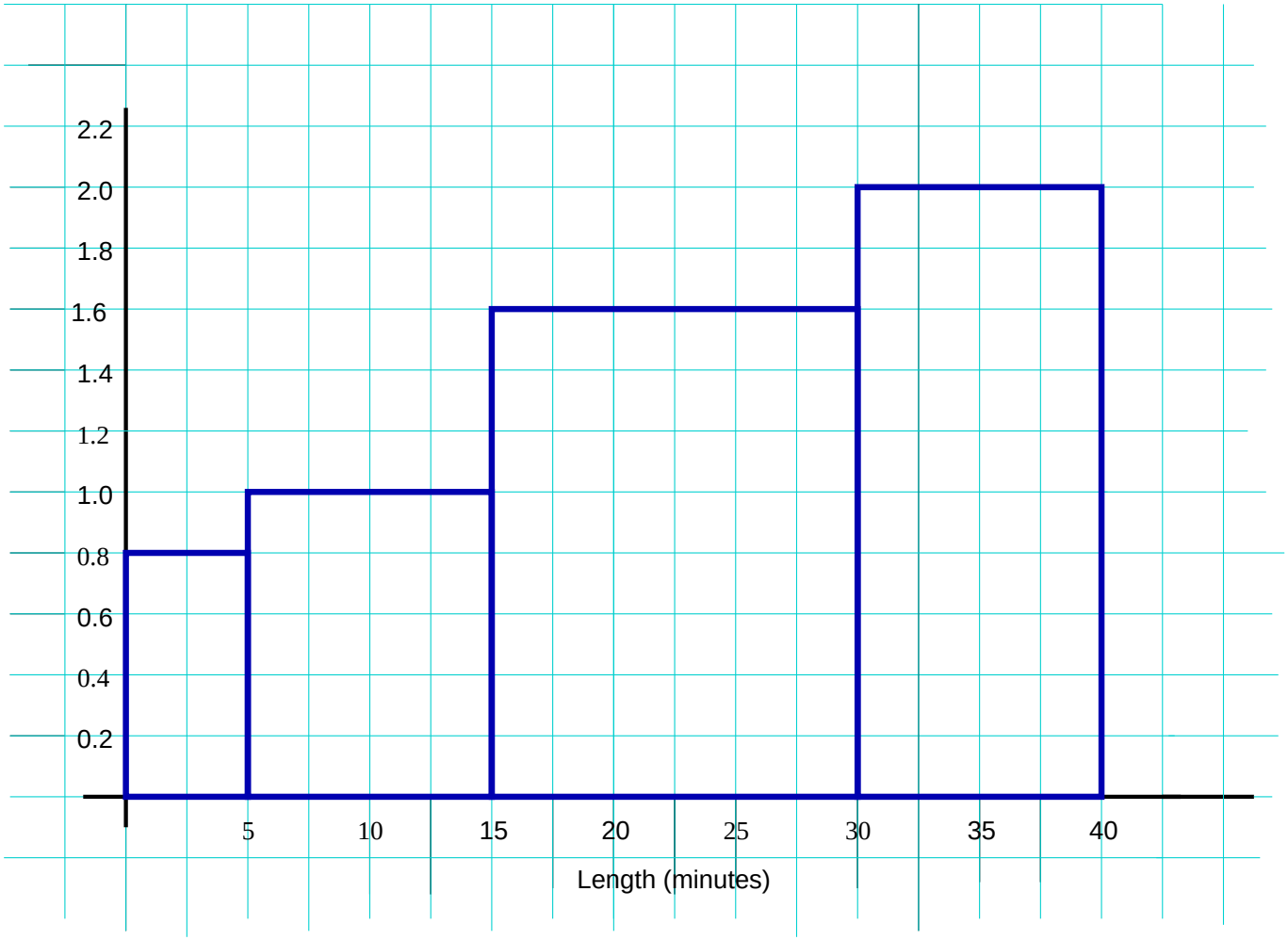
Solution to Example 3.5 (3/3)



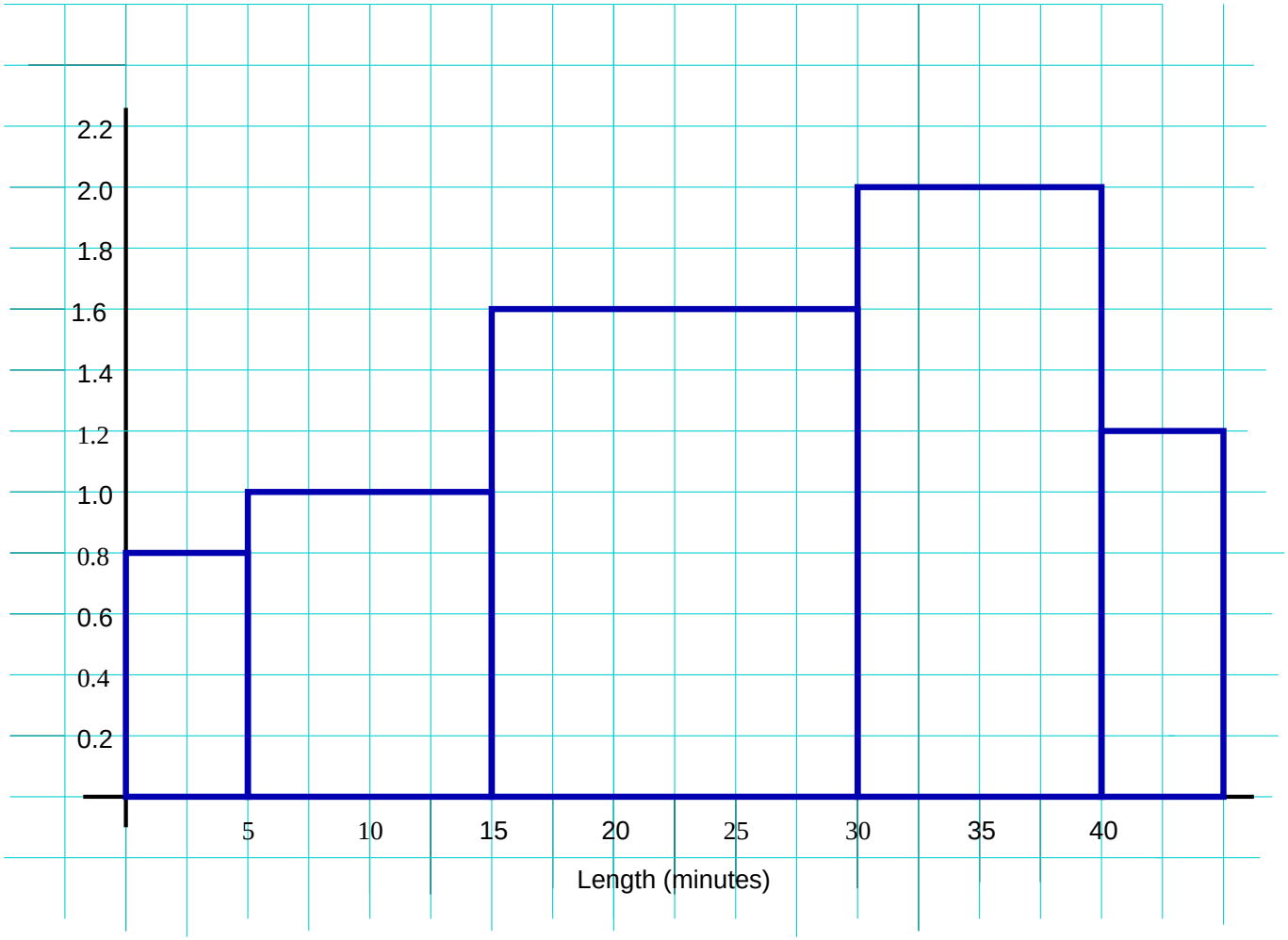
Solution to Example 3.5 (3/3)



Solution to Example 3.5 (3/3)



Solution to Example 3.5 (3/3)



3.4.4 Histograms

Summary

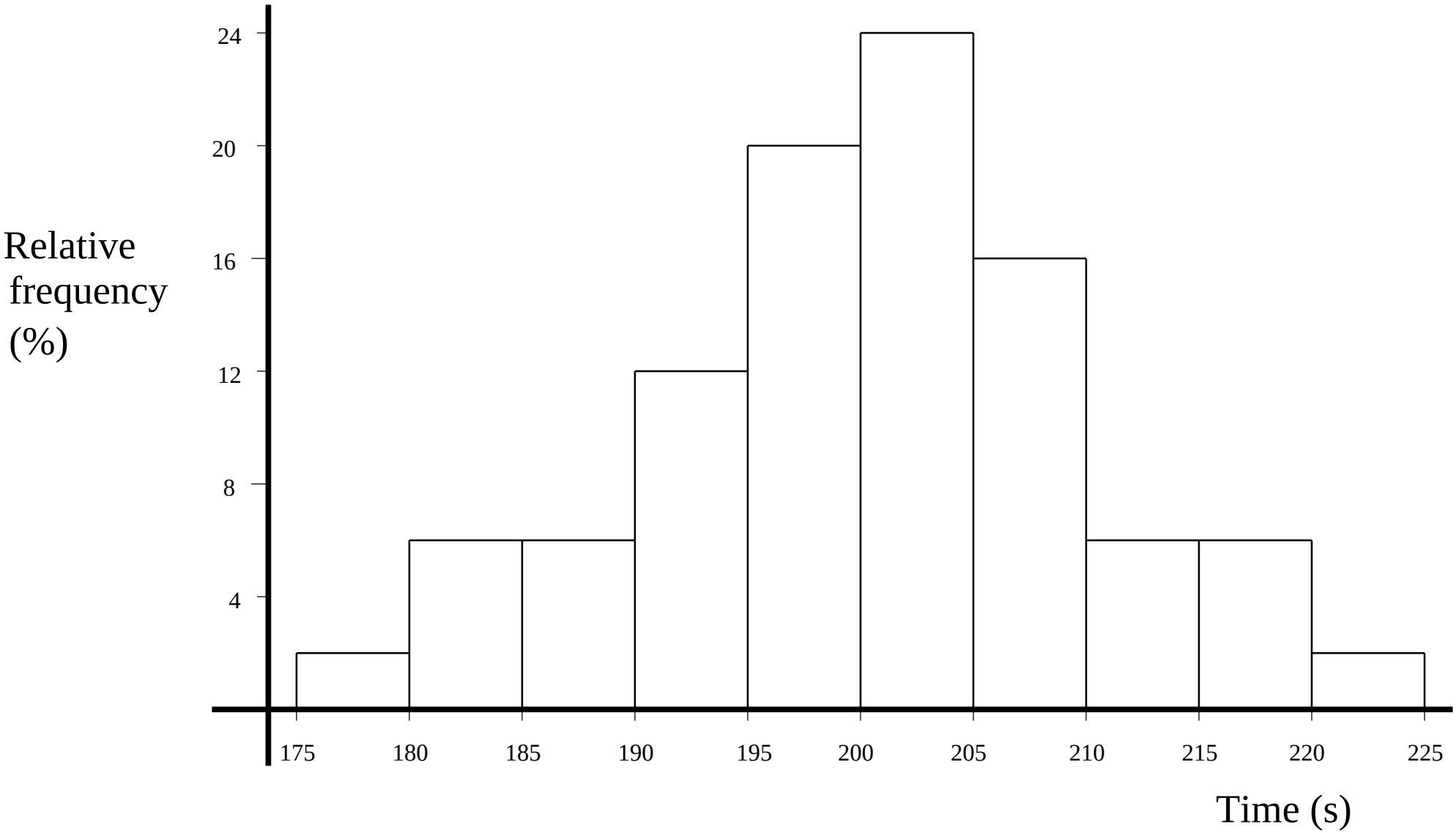
- Like a bar chart, but for continous data
- First step: Collect your data into a frequency table
- If the width of all class intervals is equal, just have "frequency" on the y–axis
- Otherwise, you will need to think about "frequency density"

3.4.5 Percentage relative frequency histograms

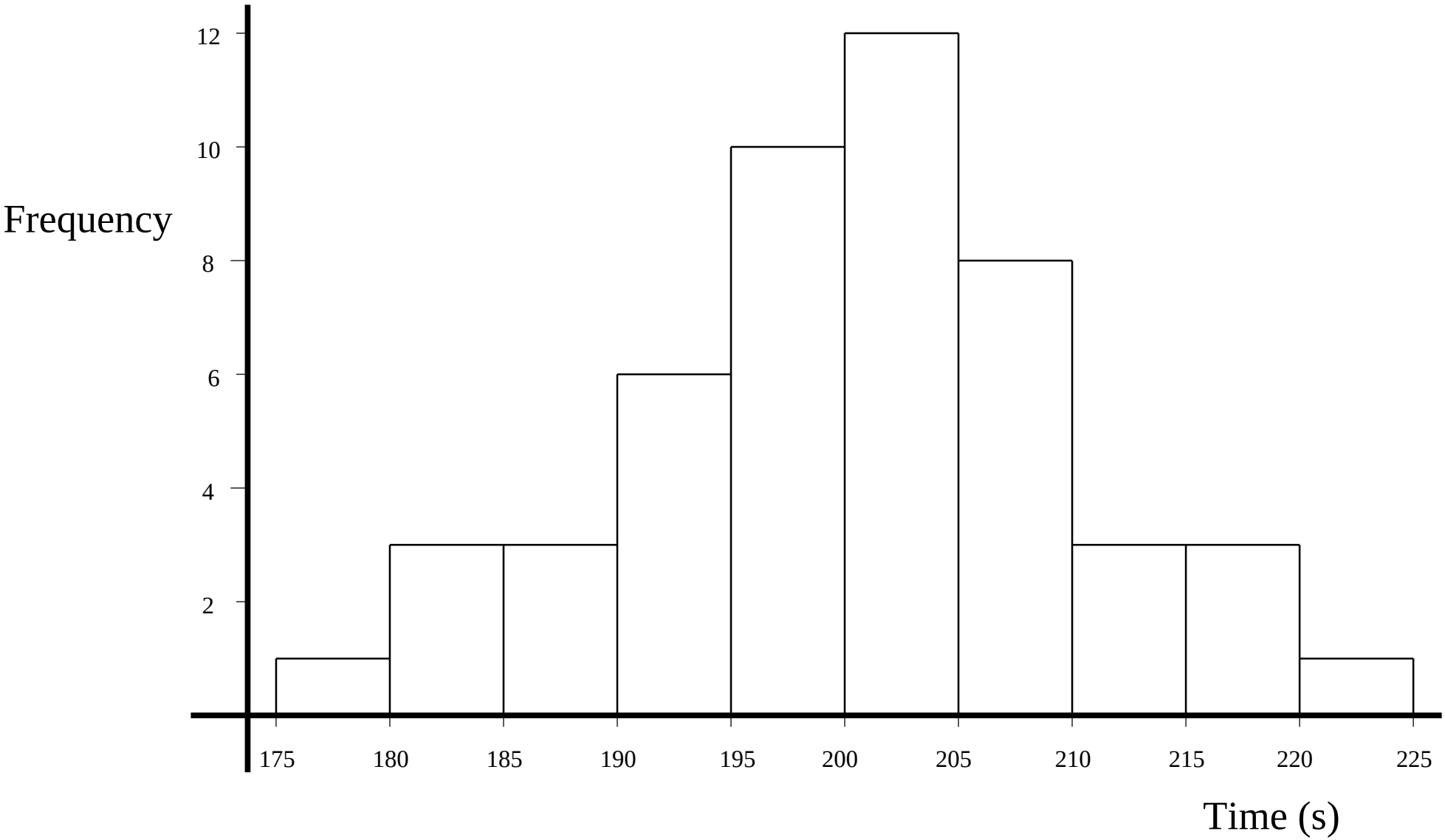
Instead of using **frequency** on the vertical axis (y-axis), you *could* use the **percentage relative frequency**. For example:

Service time	Frequency	Relative Frequency (%)
$175 \leq$ time < 180	1	2
$180 \leq$ time < 185	3	6
$185 \leq$ time < 190	3	6
$190 \leq$ time < 195	6	12
$195 \leq$ time < 200	10	20
$200 \leq$ time < 205	12	24
$205 \leq$ time < 210	8	16
$210 \leq$ time < 215	3	6
$215 \leq$ time < 220	3	6
$220 \leq$ time < 225	1	2
Totals	**50**	**100**

3.4.5 Percentage relative frequency histograms



3.4.5 Percentage relative frequency histograms



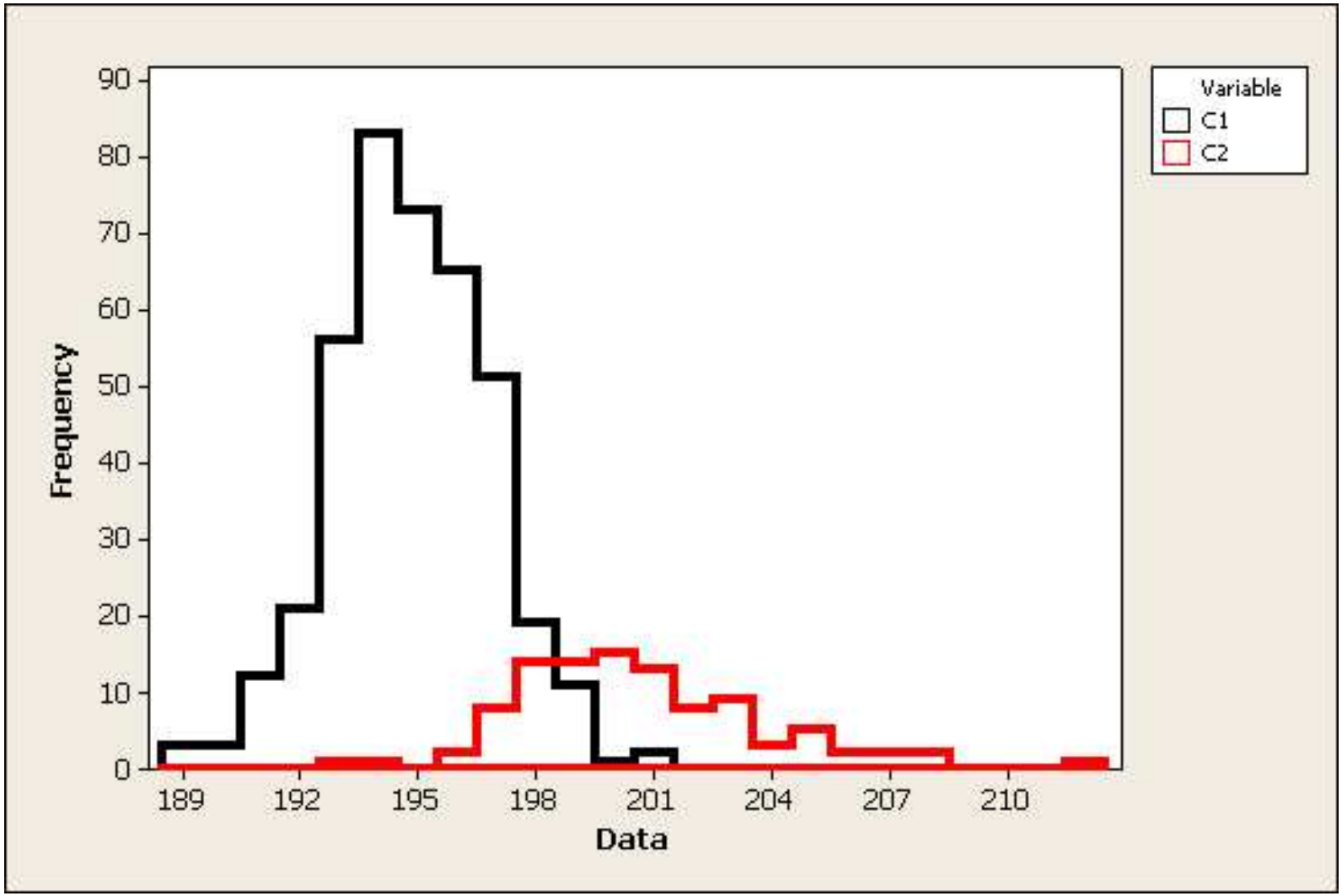
Why use percentage relative frequency?

It is useful for comparing two or more histograms

- If one sample were larger than the other, the standard histograms would look different just because of the different sample sizes.
- Looking at percentages "**puts both samples on the same scale**" and removes this difference.
- This enables us to look at **relative** differences.

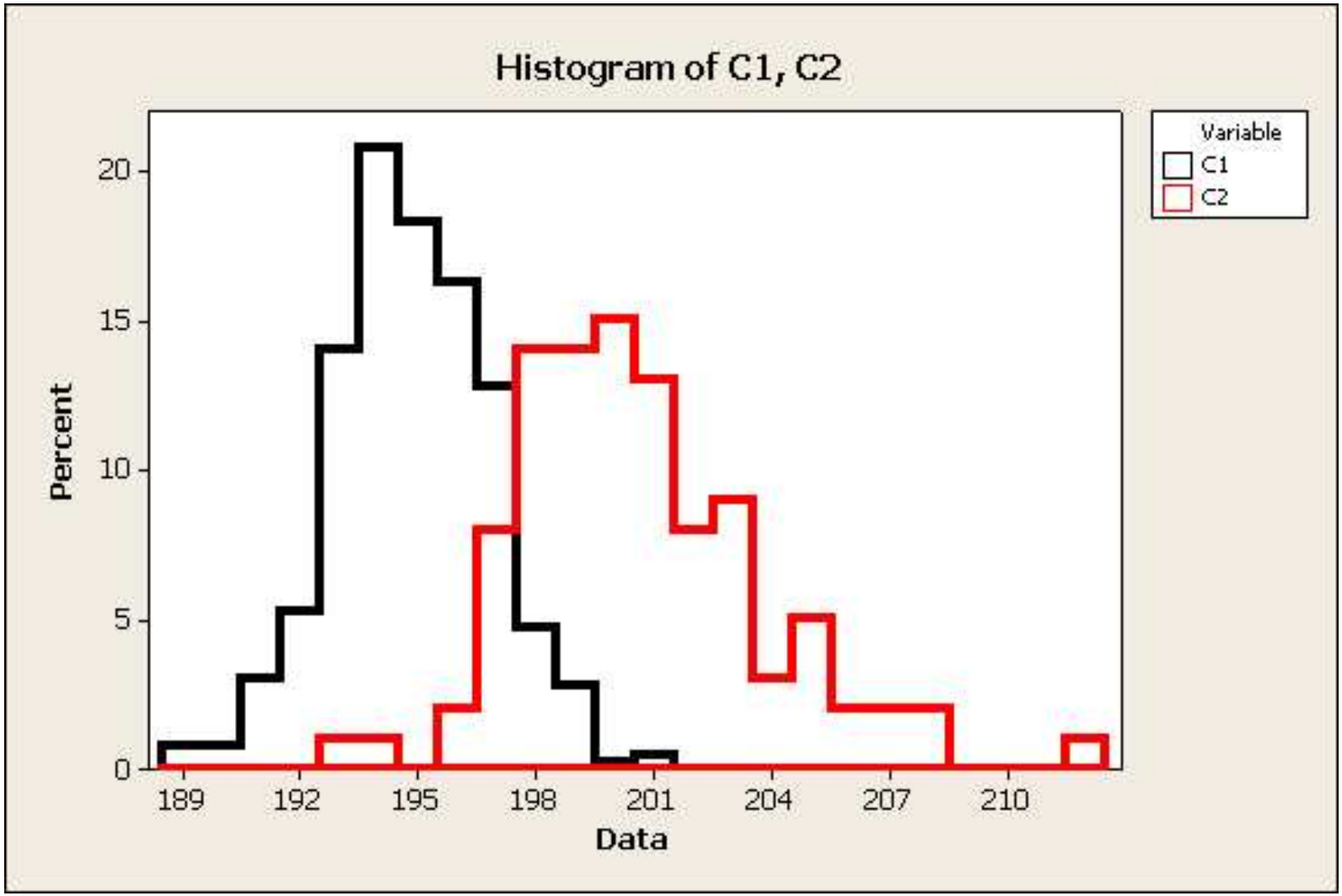
Why use percentage relative frequency?

The following (frequency) histograms were created from two samples, one of size 100 and one of size 400 ...



Why use percentage relative frequency?

... and this is what happens when we use **% relative frequency**



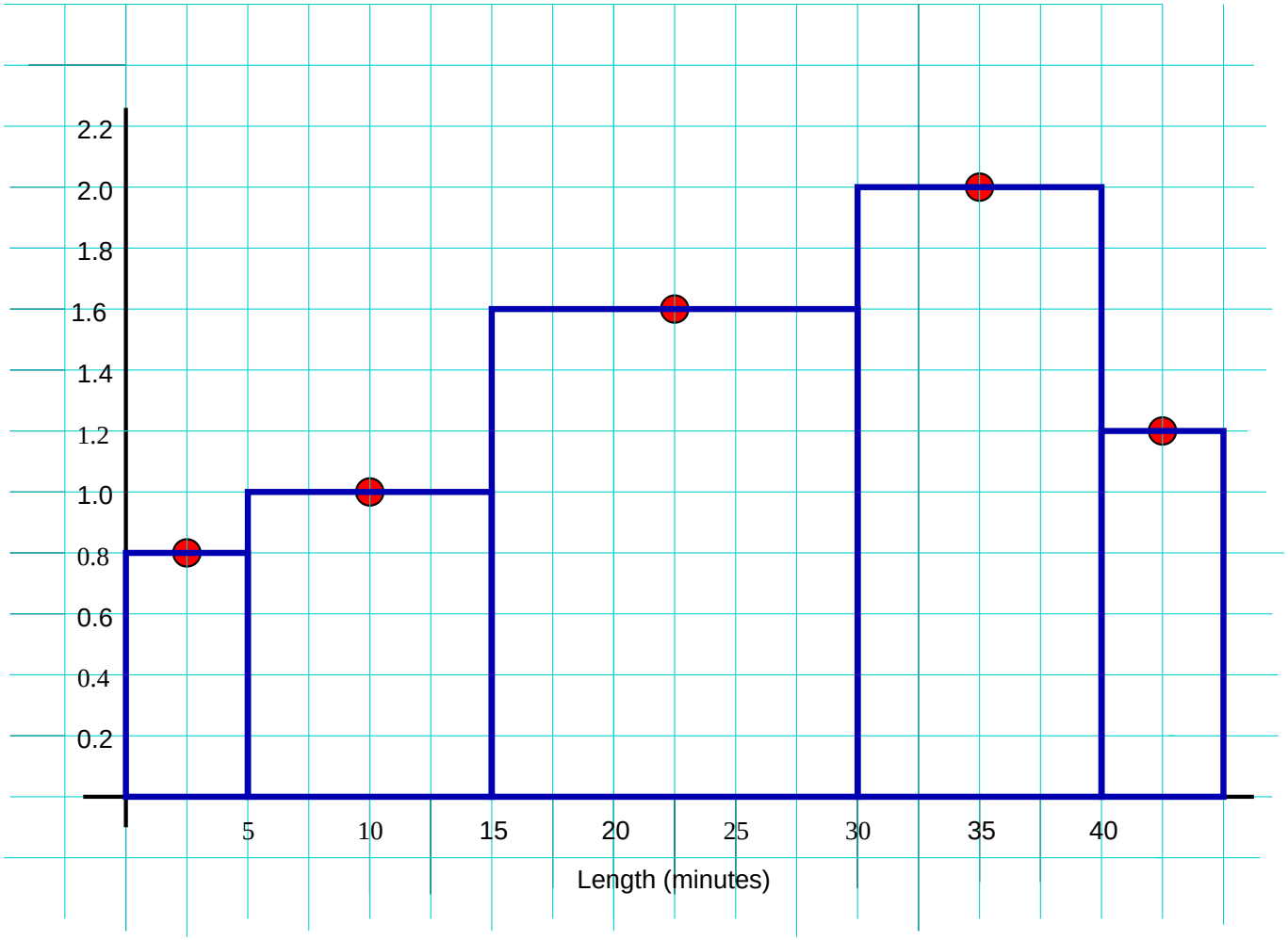
3.4.6 Polygons

So percentage relative frequency histograms are useful for comparing two groups.

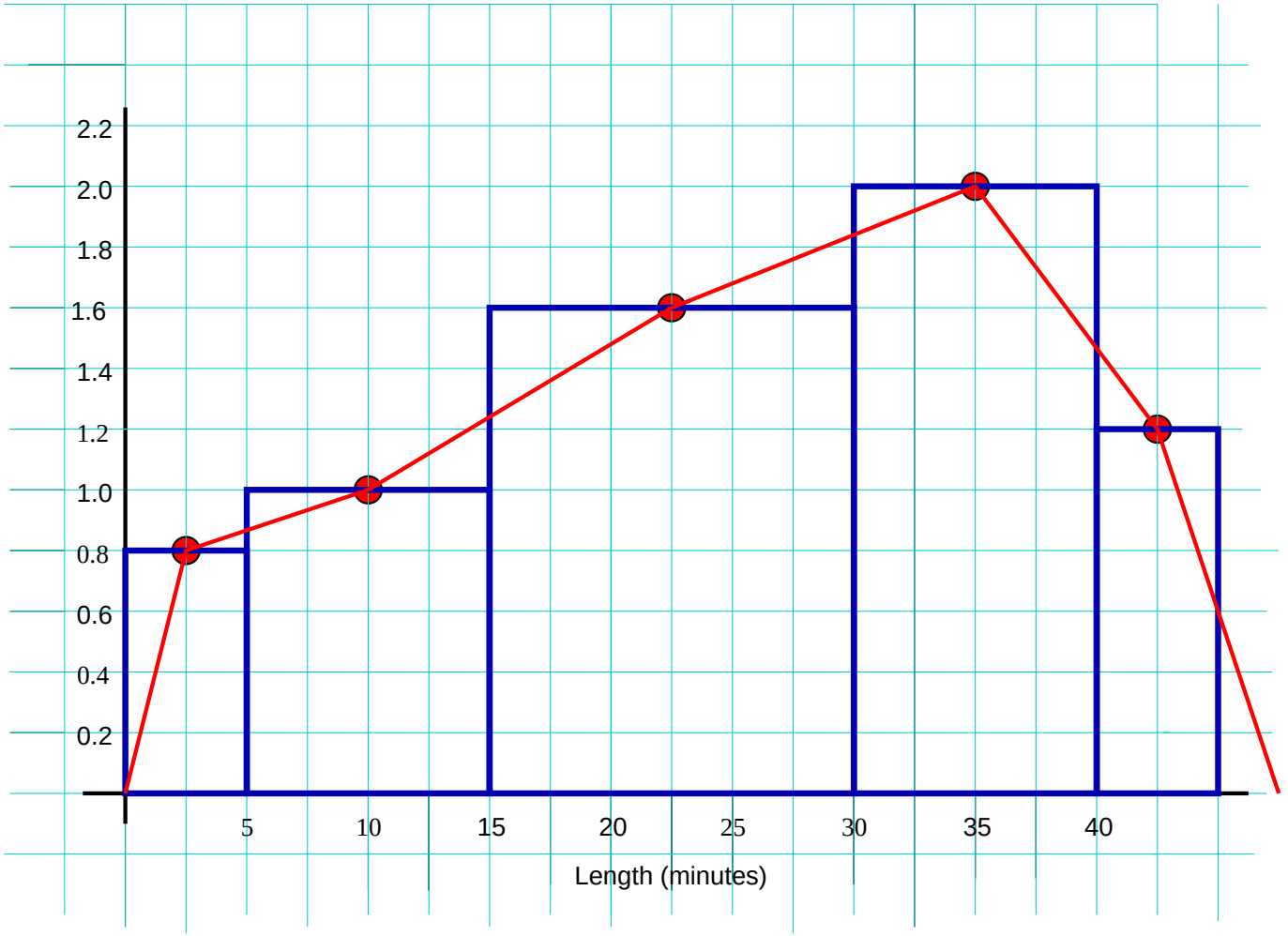
However, can you imagine how **messy** these “superimposed” histograms would get if we had three or more groups?

To get round this **messy problem** we use **polygons**.

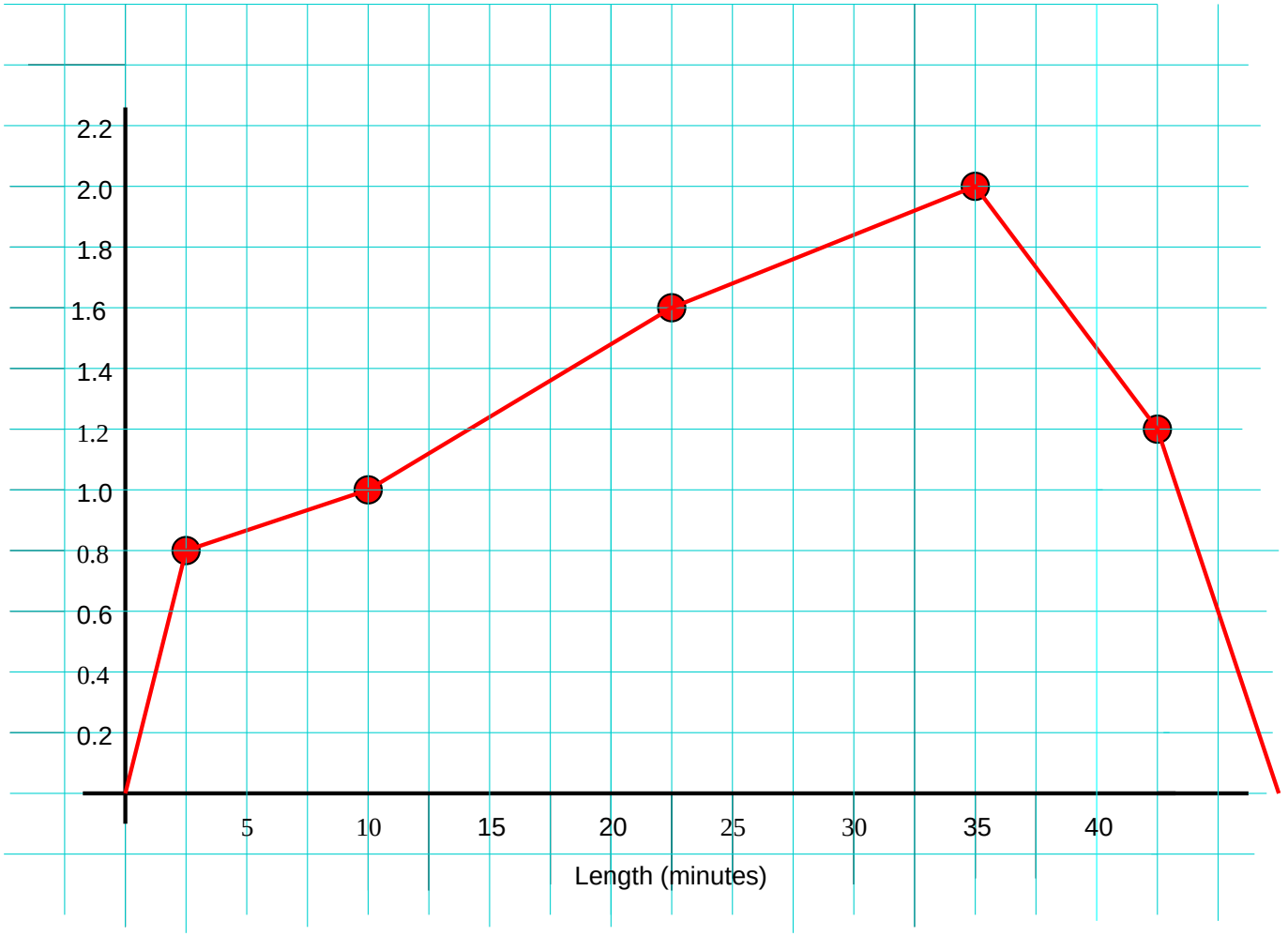
3.4.6 Polygons



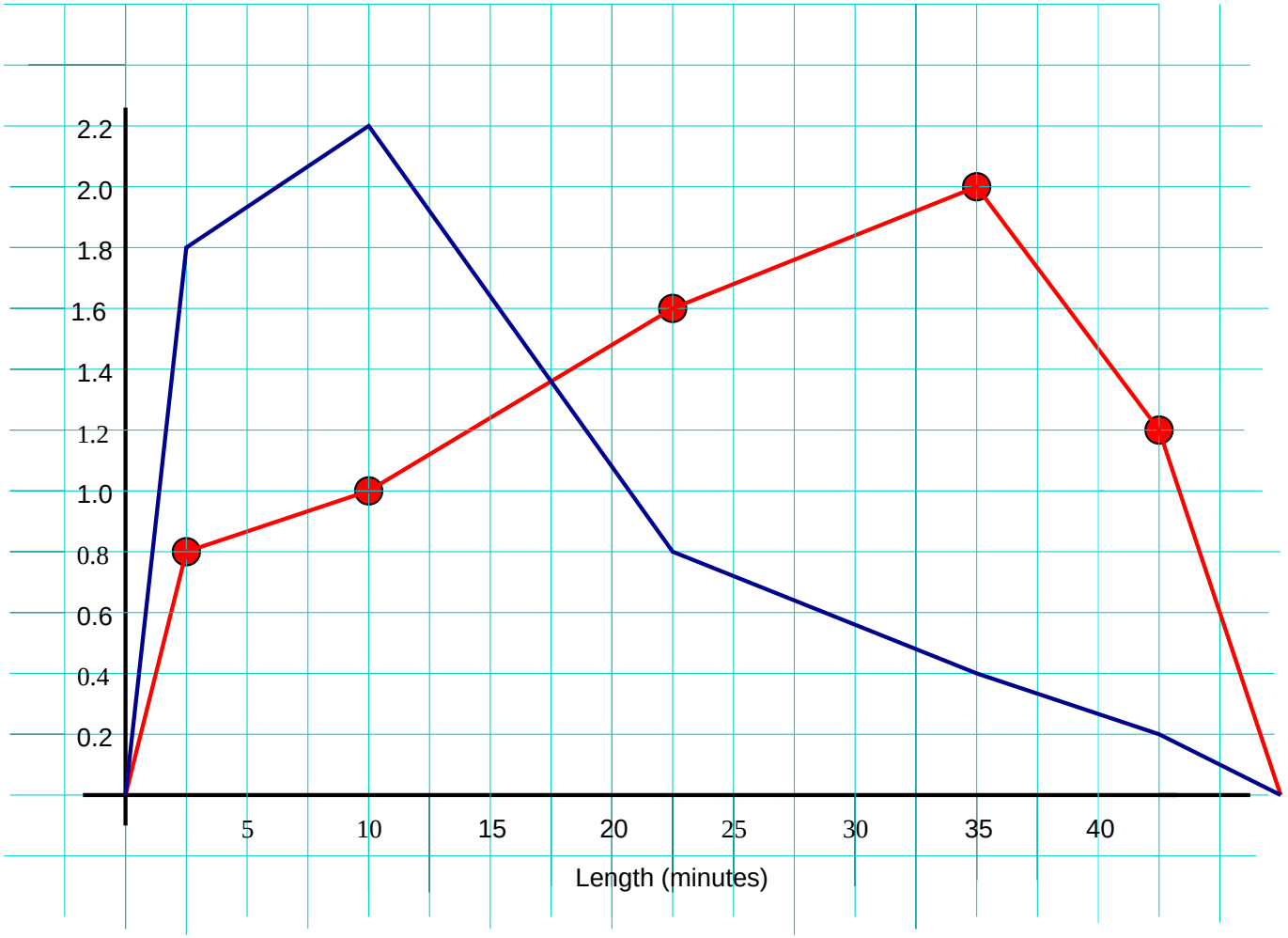
3.4.6 Polygons



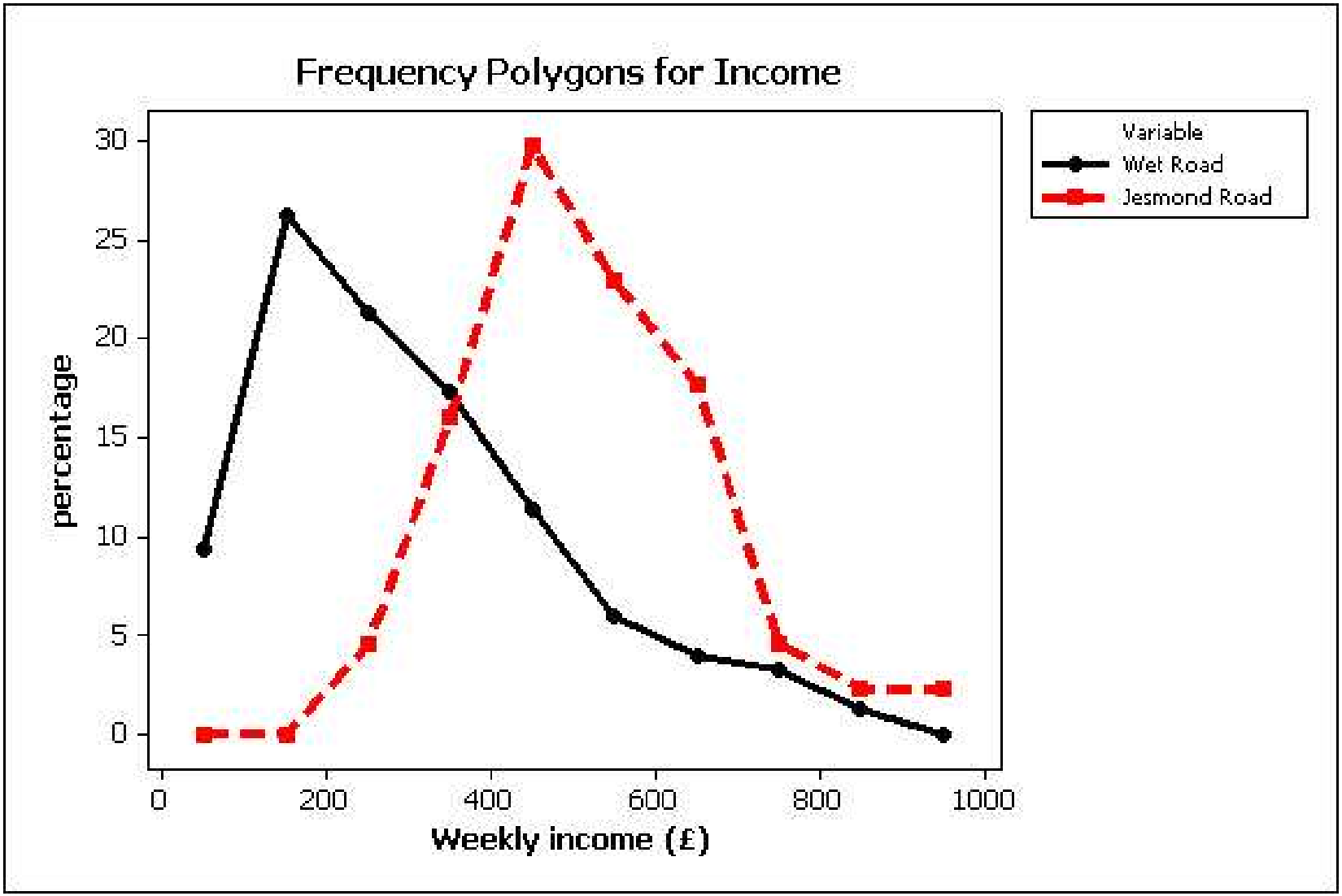
3.4.6 Polygons



3.4.6 Polygons



3.4.6 Polygons



3.4.7 Cumulative relative frequency polygons (Ogives)

The **cumulative** percentage relative frequency is simply the sum of the percentage relative frequencies at the end of each class interval.

In other words, we add the frequencies up as we go along!

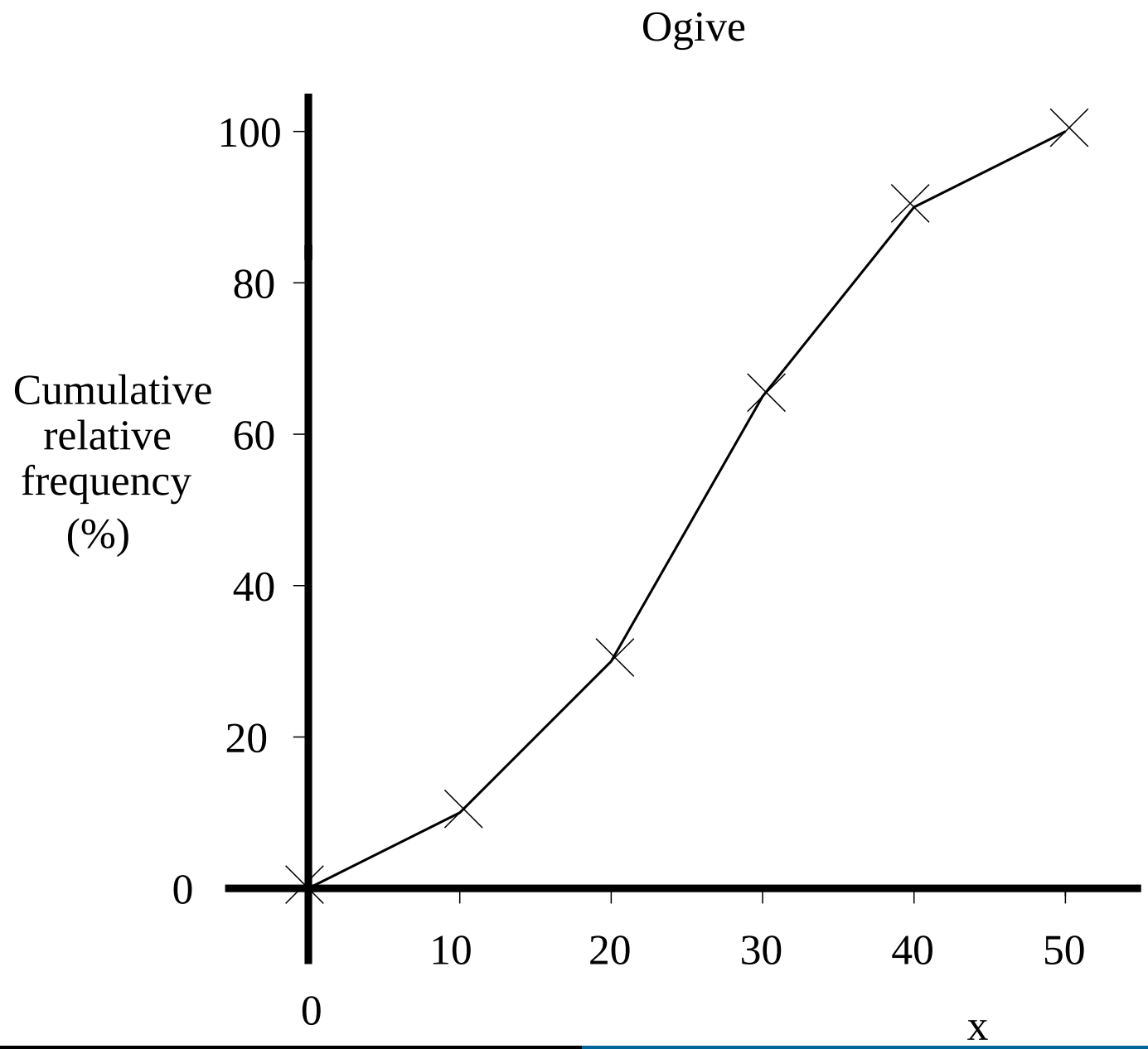
Consider the simple example:

Class Interval	% Relative Frequency	Cum. % Rel. Freq.
$0 \leq x < 10$	10	10
$10 \leq x < 20$	20	30
$20 \leq x < 30$	35	65
$30 \leq x < 40$	25	90
$40 \leq x < 50$	10	100

3.4.7 Cumulative relative frequency polygons (Ogives)

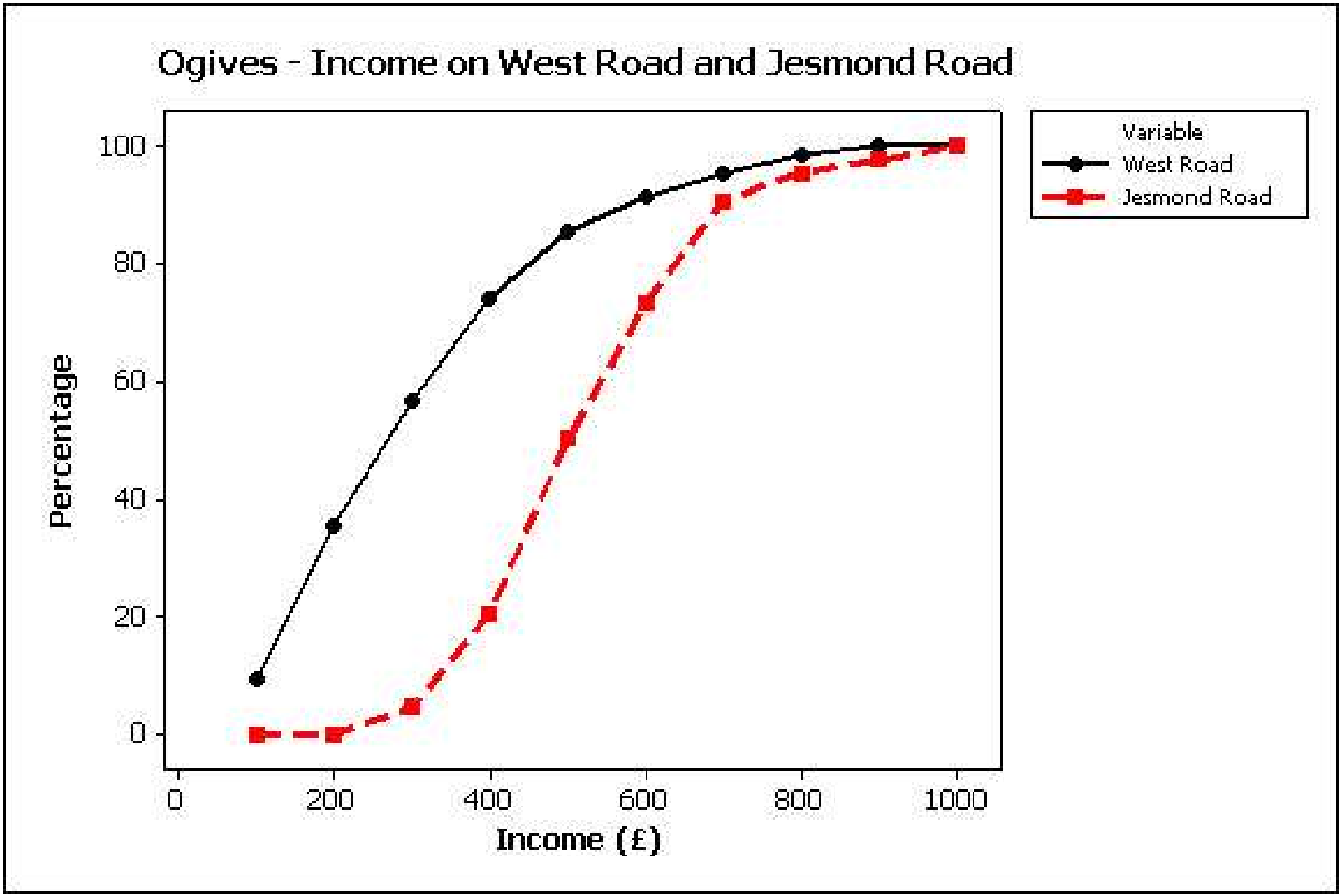
The corresponding graph, or **ogive**, is simple to produce by hand – watch out though! For these plots we use the **end–points** instead of the **mid–points**!

3.4.7 Cumulative relative frequency polygons (Ogives)



3.4.7 Cumulative relative frequency polygons (Ogives)

For the income data on West Road and Jesmond Road, we get:



Other graphical summaries

There are many other forms of graphical summary, including:

- Multiple Bar Charts
- Scatter plots
- Time series plots

These are widely used in business presentations.

3.5 Numerical summaries of data

So far we have only considered graphical methods for presenting data – useful starting points.

As we shall see, however, for many purposes we might also require **numerical** methods for summarising data.

Before we introduce some ways of summarising data numerically, let us first think about some notation.

3.5.1 Mathematical notation

Before we can talk more about numerical techniques we first need to define some basic notation.

This will allow us to generalise all situations with a **simple shorthand**.

3.5.1 Mathematical notation

Very often in statistics we replace actual numbers with letters in order to be able to write **general formulae**.

We generally use a single **upper case** letter to represent our random variable and the **lower case** to represent sample data, with subscripts to distinguish individual observations in the sample.

Amongst the most common letters to use is x, although y and z are frequently used as well.

3.5.1 Mathematical notation

For example, suppose we ask a random sample of three people how many mobile phone calls they made yesterday.

We might get the following data: **1**, **5**, **7**.

If we take another sample we will most likely get different data, say **2**, **0**, **3**.

3.5.1 Mathematical notation

Using algebra we can represent the **general case** as x_1, x_2, x_3:

1st sample	1	5	7
2nd sample	2	0	3
typical sample	x_1	x_2	x_3

3.5.1 Mathematical notation

1st sample	1	5	7
2nd sample	2	0	3
typical sample	x_1	x_2	x_3

This can be generalised further by referring to the random variable **as a whole** as X and the ith observation in the sample as x_i.

Hence, in the first sample above, the second observation is $x_2 = 5$ whilst in the second sample it is $x_2 = 0$.

The letters i and j are most commonly used as the index numbers for the subscripts.

The total number of observations in a sample is usually referred to by the letter n. Hence in our simple example above $n = 3$.

3.5.1 Mathematical notation

The next important piece of notation to introduce is the symbol

$$\sum$$

This is the upper case of the Greek letter **sigma**.

It is used to represent the phrase "**sum the values**".

3.5.1 Mathematical notation

This symbol is used as follows:

$$\sum_{i=1}^{n} x_i \quad = \quad x_1 + x_2 + x_3 + \cdots + x_n.$$

This notation is used to represent the sum of all the values in our data (from the first $i = 1$ to the last $i = n$).

This is often abbreviated to $\sum x$ when we sum over all the data in our sample.

3.5.2 Measures of Location

These are also referred to as measures of **centrality** or, more commonly, **averages**.

In general terms, they tell us the value of a "**typical**" observation.

There are three measures which are commonly used:

- the **mean**;
- the **median**, and
- the **mode**.

We will consider these in turn.

3.5.2 The Arithmetic Mean

The **arithmetic mean** is perhaps the most commonly used measure of location.

We often refer to it as the average or just the **mean**.

The arithmetic mean is calculated by simply adding all our data together and dividing by the number of data we have.

So if our data were **10**,**12**, and **14**, then our mean would be

$$\frac{10+12+14}{3} = \frac{36}{3} = 12.$$

3.5.2 The Arithmetic Mean

We denote the mean of our sample, or sample mean, using the notation $\bar{x}$ ("x bar").

In general, the mean is calculated using the formula

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

or equivalently as

$$\bar{x} = \frac{\sum x}{n}.$$

For small data sets this is easy to calculate **by hand**, though this is simplified by using the **statistics mode** on a University approved calculator (see next week's tutorials).

3.5.2 The Arithmetic Mean

Sometimes we might not have the raw data; instead, the data might be available in the form of a **table**.

It is still possible to calculate the mean from such data.

3.5.2 The Arithmetic Mean

Let us first consider the case where we have some **ungrouped discrete data**.

Previously we have seen the data:

Date	Cars Sold	Date	Cars Sold
01/07/12	9	08/07/12	10
02/07/12	8	09/07/12	5
03/07/12	6	10/07/12	8
04/07/12	7	11/07/12	4
05/07/12	7	12/07/12	6
06/07/12	10	13/07/12	8
07/07/12	11	14/07/12	9

The mean number of cars sold per day is

$$\bar{x} = \frac{9 + 8 + \ldots + 8 + 9}{14} = 7.71.$$

3.5.2 The Arithmetic Mean

These data can be presented as the **frequency table**:

Cars Sold $(x_{(j)})$	4	5	6	7	8	9	10	11
Frequency (f_j)	1	1	2	2	3	2	2	1

The sample mean can be calculated from these data as

$$\bar{x} = \frac{4 + 5 + \overbrace{6 + 6}^{\times 2} + \overbrace{7 + 7}^{\times 2} + \overbrace{8 + 8 + 8}^{\times 3} + \overbrace{9 + 9}^{\times 2} + \overbrace{10 + 10}^{\times 2} + 11}{14}$$

$$= \frac{(\mathbf{4} \times 1) + (\mathbf{5} \times 1) + (\mathbf{6} \times 2) + (\mathbf{7} \times 2) + \cdots + (\mathbf{11} \times 1)}{14}$$

$$= 7.71.$$

3.5.2 The Arithmetic Mean

We can express this calculation of the sample mean from discrete tabulated data as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} f_j x_{(j)}.$$

Here the different values of X which occur in the data are $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$.

In this example $x_{(1)} = 4,\ x_{(2)} = 5, \ldots, x_{(k)} = 11$ and $k = 8$.

3.5.2 The Arithmetic Mean

If we only have **grouped frequency data**, it is still possible to approximate the value of the sample mean.

Consider the following (ordered) data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

The sample mean of these data is **9.73**.

3.5.2 The Arithmetic Mean

Grouping these data into a frequency table gives

Class Interval	**mid–point** (m_j)	**Frequency** (f_j)
$8.0 \leq x < 8.5$	8.25	1
$8.5 \leq x < 9.0$	8.75	1
$9.0 \leq x < 9.5$	9.25	5
$9.5 \leq x < 10.0$	9.75	7
$10.0 \leq x < 10.5$	10.25	2
$10.5 \leq x < 11.0$	10.75	3
$11.0 \leq x < 11.5$	11.25	1
Total (n)		20

3.5.2 The Arithmetic Mean

Therefore, the (approximate) sample mean is calculated using the the frequencies (f_j) and the mid–points (m_j) as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} f_j m_j .$$

For the grouped data above, we obtain

$$\bar{x} = \frac{1}{20} (1 \times 8.25 + 1 \times 8.75 + \cdots + 3 \times 10.75 + 1 \times 11.25) = 9.775.$$

This value is fairly close to the correct sample mean and is a reasonable approximation given the partial information we have in the table.

3.5.2 The Median

The **median** is occasionally used instead of the mean, particularly when the data have an **asymmetric profile**.

The median is the **middle value** of the observations when they are listed in ascending order.

It is straightforward to determine the median for small data sets, particularly via a stem and leaf plot.

For larger data sets, the calculation is more easily done using `Minitab` (see Semester 2).

3.5.2 The Median

The median is that value that has half the observations above it and half below.

If the sample size (n) is an odd number, we have:

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{ largest observation.}$$

3.5.2 The Median

For example, if our data were

$$\mathbf{2}, \mathbf{3}, \mathbf{3}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{9},$$

then the sample size ($n = 7$) is an odd number and therefore the median is the

$$\frac{7+1}{2} = 4^{th} \text{ largest observation,}$$

that is, the median is the fourth largest (or smallest) ranked observation.

For these data the median $= 5$.

3.5.2 The Median

If the sample size (n) is an even number the process is slightly more complicated:

$$\text{median} = \text{average of the } \left(\frac{n}{2}\right)^{th}$$

$$\text{and the } \left(\frac{n}{2}+1\right)^{th} \text{ largest observations.}$$

3.5.2 The Median

For example, if our data were

$$\mathbf{2}, \mathbf{3}, \mathbf{3}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{9}, \mathbf{10},$$

then the sample size ($n = 8$) is an even number and therefore

$$\begin{aligned} \text{median} &= \text{average of the } \left(\frac{8}{2}\right)^{th} \\ &\qquad \text{and the } \left(\frac{8}{2}+1\right)^{th} \text{ largest observations} \\ &= \frac{5+6}{2} \\ &= 5.5. \end{aligned}$$

3.5.2 The Median

It is possible to estimate the median value from an **ogive** as it is half way through the ordered data and hence is at the 50% level of the cumulative frequency.

The accuracy of this estimate will depend on the accuracy of the ogive drawn.

3.5.2 The Mode

This is the final measure of location we will look at.

It is the value of the random variable in the sample which **occurs with the highest frequency**.

It is usually found by inspection (no formula).

3.5.2 The Mode

For discrete data this is easy – the mode is simply the most common value.

So, on a bar chart, it would be the category with the **highest bar**.

For example, consider the following data:

2, **2**, **2**, **3**, **3**, **4**, **5**.

Quite obviously the mode is 2 as it occurs most often.

3.5.2 The Mode

It is possible to refer to **modal classes** with grouped data.

This is simply the class with the greatest frequency of observations.

For example, the model class of

Class	Frequency
$10 \leq x < 20$	10
$20 \leq x < 30$	15
$30 \leq x < 40$	30

is obviously $30 \leq x < 40$.

3.5.3 Measures of Spread

A measure of location is insufficient in itself to summarise data as it only describes the value of a **typical outcome** and not how much **variation** there is in the data. For example:

Dataset 1	6	22	38
Dataset 2	21	22	23

For both of these datasets, $\bar{x} = \text{median} = 22$. However,

- dataset 1 ranges considerably from these average values;
- dataset 2 is very tightly concentrated around these averages.

3.5.3 Measures of Spread

The mean or the median does not fully represent the data.

There are three basic **measures of spread** which we will consider:

- the **range**
- the **inter–quartile range**, and
- the **sample variance**/**standard deviation**.

3.5.3 The Range

This is the **simplest measure of spread**.

It is simply the difference between the largest and smallest observations.

In our simple example above:

$$\text{Range for dataset 1} = 38 - 6 = 32$$

$$\text{Range for dataset 2} = 23 - 21 = 2$$

These clearly describe very different data sets. The first set has a much wider range (data more spread out) than the second.

3.5.3 The Range

There are two problems with the range as a measure of spread.

- Uses the two most extreme data points – value can be unduly influenced by one particularly large/small value (**outliers**)
- Only suitable for comparing (roughly) equally–sized samples

3.5.3 The Inter–Quartile Range

The inter–quartile range describes the range of the **middle half of the data** and so is less prone to the influence of the extreme values.

To calculate the inter–quartile range (IQR) we simply divide the the ordered data into **four quarters**.

The three values that split the data into these quarters are called the **quartiles**.

3.5.3 The Inter–Quartile Range

- The first quartile (**lower quartile**, $Q1$) has 25% of the data below it
- The second quartile (**median**, $Q2$) has 50% of the data below it
- The third quartile (**upper quartile**, $Q3$) has 75% of the data below it.

3.5.3 The Inter–Quartile Range

We already know how to find the median, and the other quartiles are calculated as follows:

$$Q1 = \frac{(n+1)}{4}\text{th smallest observation}$$

$$Q3 = \frac{3(n+1)}{4}\text{th smallest observation.}$$

Just as with the median, these quartiles might not correspond to actual observations.

For example, in a dataset with $n = 20$ values, the lower quartile is the

$$\frac{21}{4} = 5\frac{1}{4}^{\text{th}} \text{ largest observation,}$$

that is, a quarter of the way between the 5th and 6th largest observations.

3.5.3 The Inter–Quartile Range

Consider again the data ($n = 20$):

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Here the 5th and 6th smallest observations are 9.2 and 9.3 respectively. Therefore, the lower quartile is $Q1 = 9.225$.

Similarly, the upper quartile is the

$$\frac{3 \times 21}{4} = 15\frac{3}{4}^{\text{th}} \text{ smallest observation,}$$

that is, three quarters of the way between 10.3 and 10.4; so $Q3 = 10.375$.

3.5.3 The Inter–Quartile Range

The inter–quartile range is simply the difference between the upper and lower quartiles, that is

$$\begin{aligned} IQR &= Q3 - Q1 \\ &= 10.375 - 9.225 \\ &= 1.15. \end{aligned}$$

3.5.3 The Inter–Quartile Range

The interquartile range can also be estimated from the ogives in a similar manner to the median.

Simply draw the ogive and then read off the values for 75% and 25% and calculate the difference between them. This is especially useful if you only have grouped data.

Again the accuracy depends on the quality of your graph.

The inter–quartile range is useful as it allows us to make comparisons between the ranges of two data sets, without the problems caused by **outliers** or **unequal sample sizes**.

3.5.3 The Sample Variance and Standard Deviation

The **sample variance** is the standard measure of spread used in statistics.

It is usually denoted by s^2 and is simply the "average" of the squared distances of the observations from the sample mean.

We use the formula

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1};$$

the $(n - 1)$ divisor will be explained in semester 2.

3.5.3 The Sample Variance and Standard Deviation

We can rewrite this using more condensed mathematical notation:

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

or equivalently as

$$s^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^{n} x_i^2 - n (\bar{x})^2 \right\} .$$

Note that the notation x_i^2 represents the squared value of the observation x_i. That is, $x_i^2 = (x_i)^2$.

The **sample standard deviation** s is the positive square root of the sample variance.

3.5.3 The Sample Variance and Standard Deviation

A different calculation is needed when the data are given in the form of a grouped frequency table with frequencies (f_i) in intervals with mid–points (m_i).

First the sample mean $\bar{x}$ is approximated (as described earlier) and then the sample variance is approximated as

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

Example 3.6

Consider again the data

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Calculate the sample variance and hence the sample standard deviation.

Example 3.6: Solution (1/2)

The **sample mean** is

$$\bar{x} = 9.73.$$

If we use the second version of the formula

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2\right\},$$

we also need $\sum x^2$, which is:

$$8.4^2 + 8.7^2 + 9.0^2 + \ldots + 11.4^2 = 1904.38.$$

Example 3.6: Solution (2/2)

So

$$s^2 = \frac{1}{\mathbf{20} - 1}\left\{\mathbf{1904.38} - \mathbf{20}\left(\mathbf{9.73}\right)^2\right\}$$

$$= 0.5748.$$

The standard deviation is just the square root of this, i.e.

$$s = \sqrt{0.5748} = 0.7582.$$

3.5.4 Box–and–whisker plots

Box and whisker plots are another graphical method for displaying data and are particularly useful in highlighting differences between groups.

These plots use some of the key summary statistics we have looked at earlier – the **quartiles** – and also the **maximum** and **minimum** observations.

3.5.4 Box–and–whisker plots

The plot is constructed as follows:

- Lay out an x–axis for the full range of the data
- Draw a rectangle with ends at the the upper and lower quartiles
- Split the rectangle into two at the median: You now have the "**box**"
- Finally, draw lines from the box to the minimum and maximum values – these are the "**whiskers**"

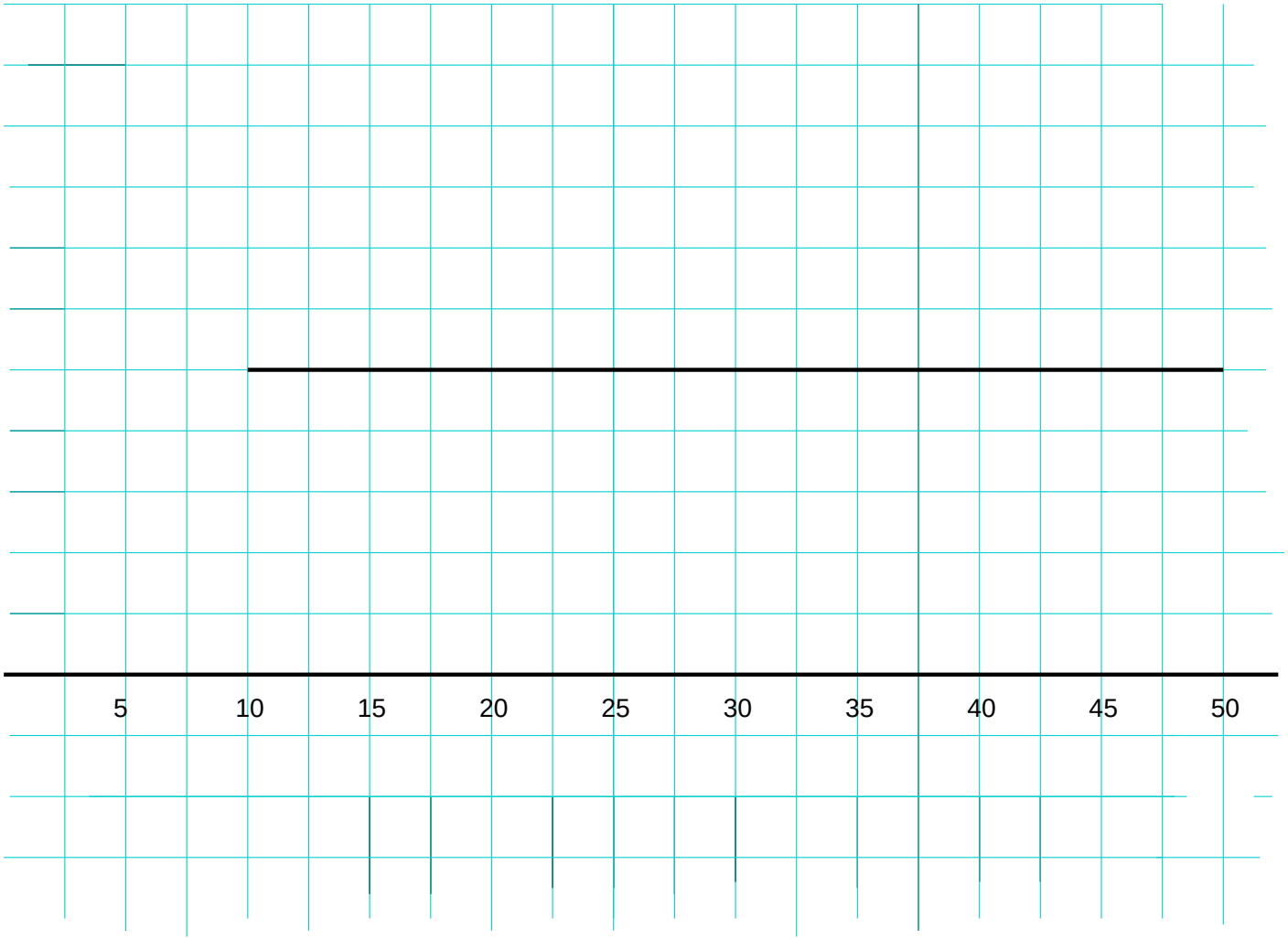
Example 3.7

Suppose that, from our data, we obtain the following summary statistics:

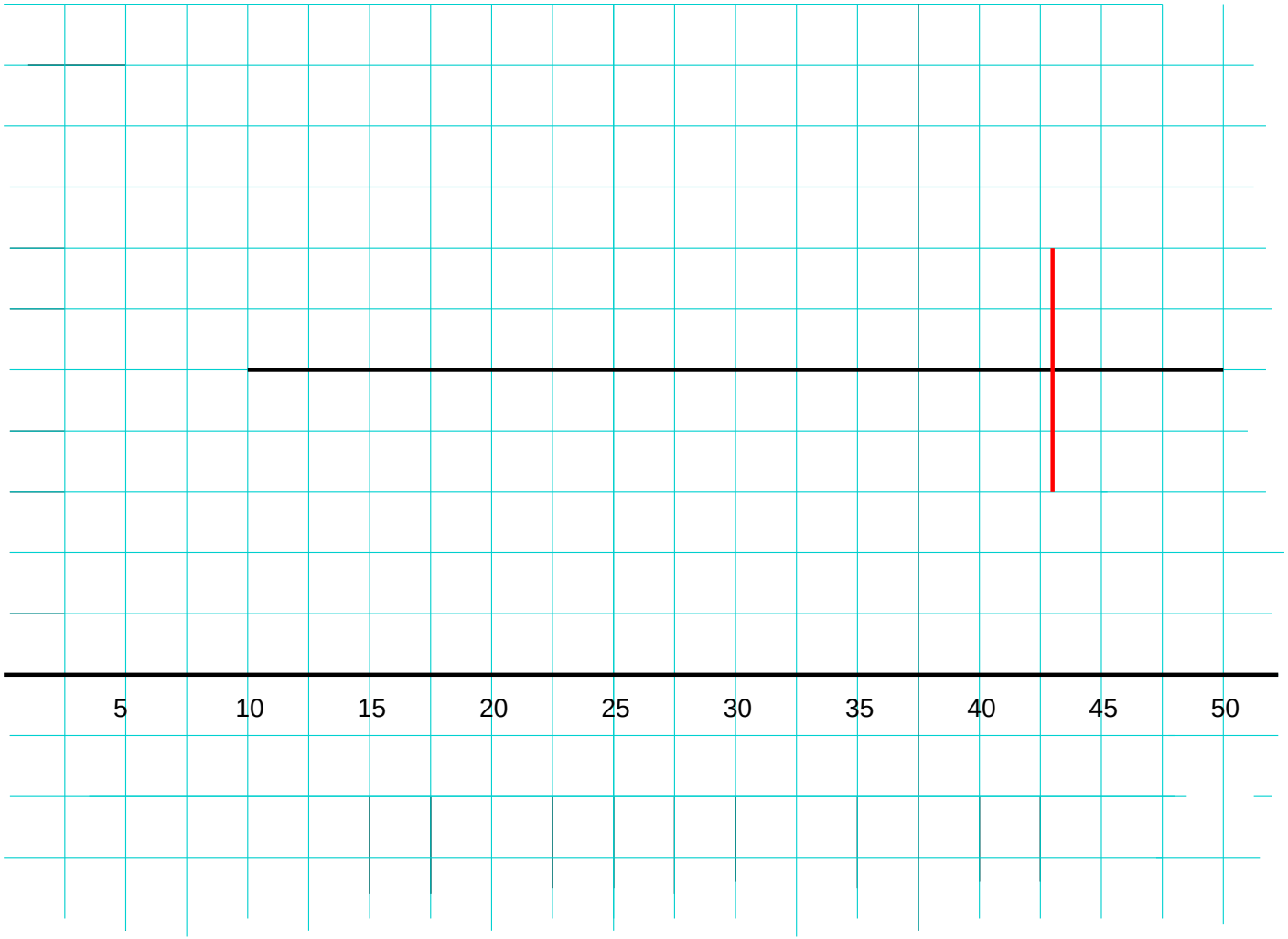
Minimum	$min = 10$
Lower quartile	$Q1 = 40$
Median	$Q2 = 43$
Upper quartile	$Q3 = 45$
Maximum	$max = 50$

Complete the associated box–and–whisker plot, and comment.

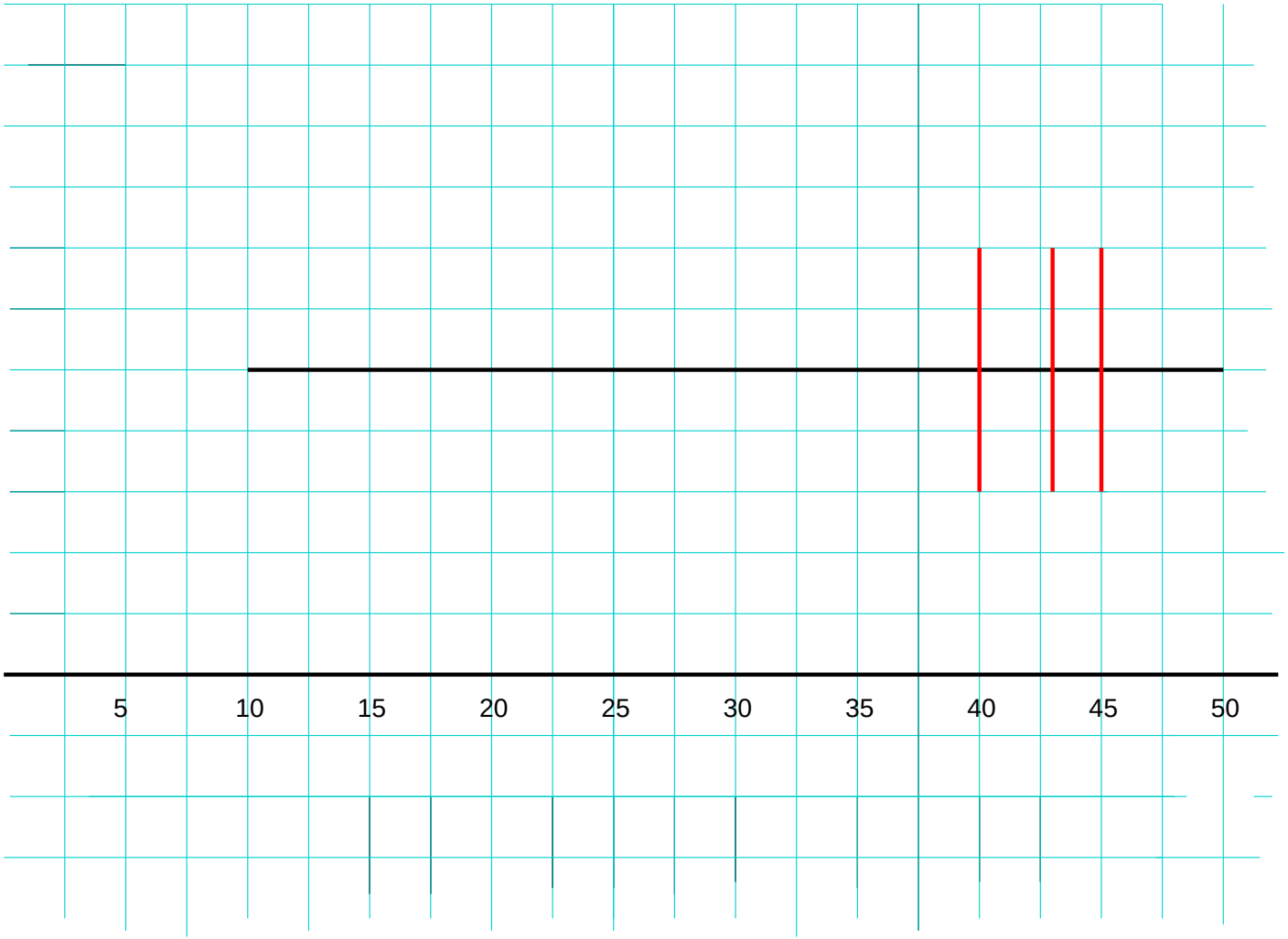
Example 3.7: Solution



Example 3.7: Solution



Example 3.7: Solution



Example 3.7: Solution