Lecture 3

MORE GRAPHICAL METHODS

FOR PRESENTING DATA

Recap

We have already looked at some basic ways to present data graphically. These include:

- **Stem and leaf plots**
- **Bar charts** including **multiple** bar charts
- **Histograms**

We now look at some other methods...

Percentage relative frequency histograms

These are an extension to the **percentage relative frequency tables** we though about two weeks ago.

Recall the data on service time (in seconds) for calls to a credit card service centre:

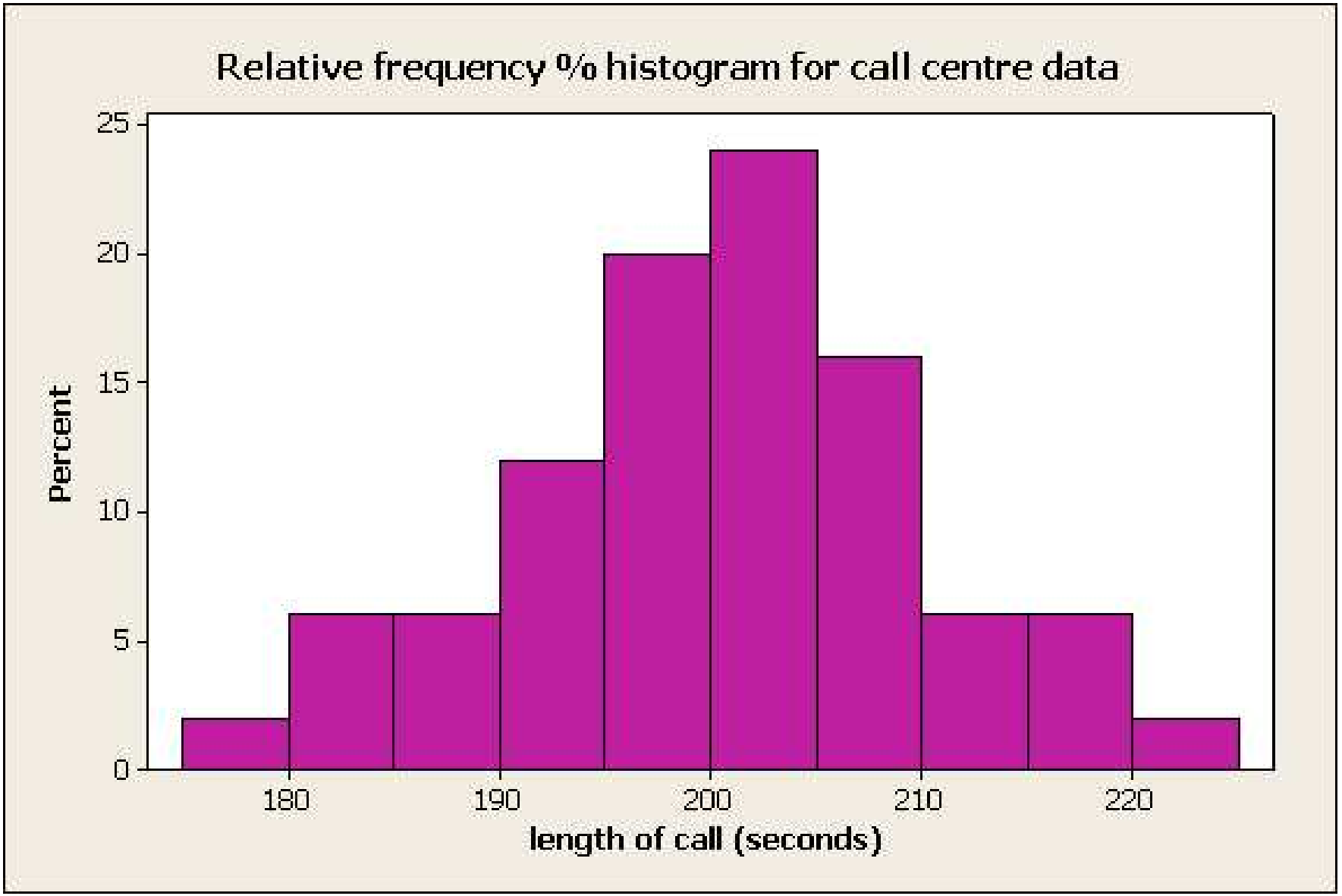
214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	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

A percentage relative frequency table for these data is:

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

You can plot these data like an ordinary histogram.

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

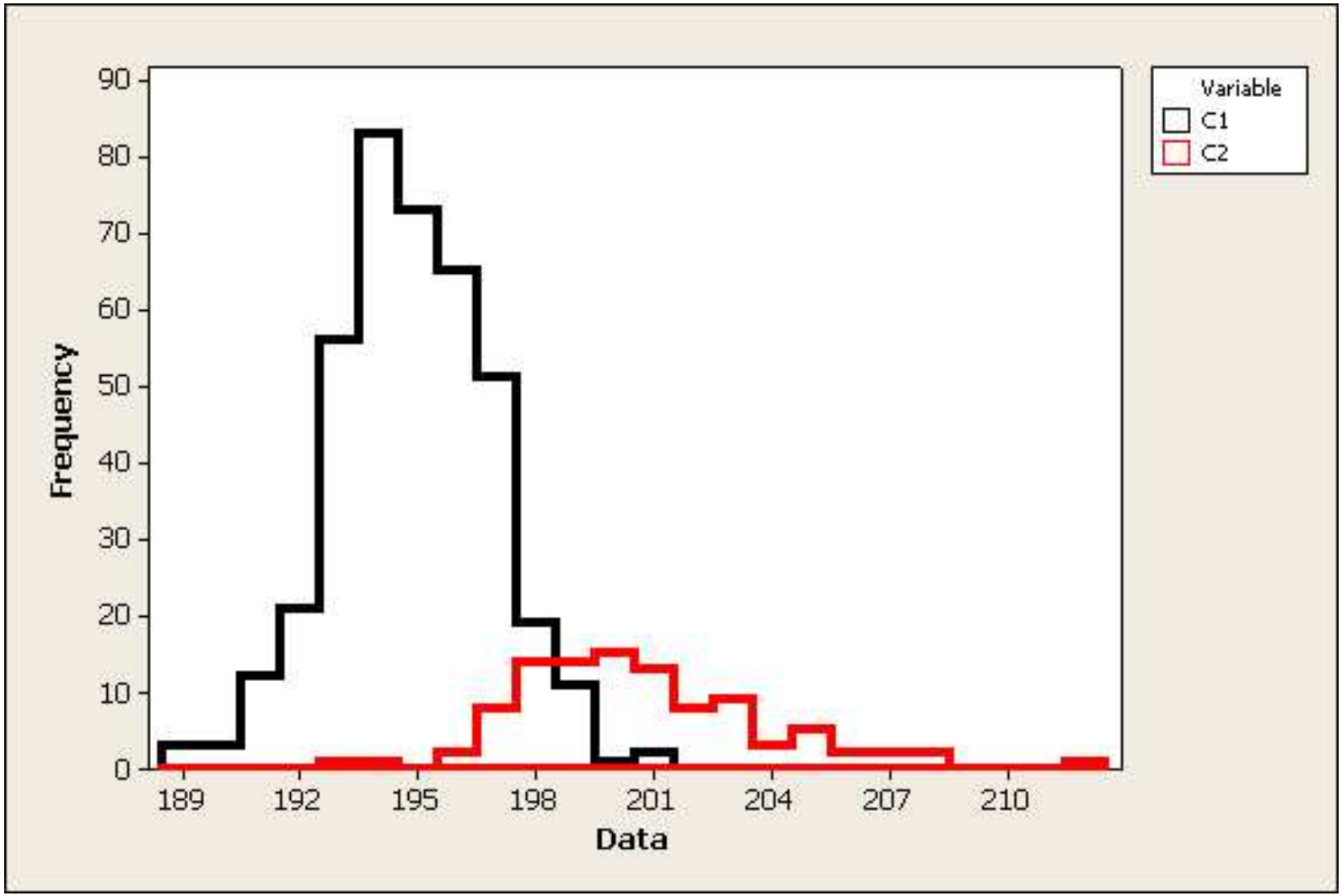

Relative frequency % histogram for call centre data
Percent
25
20
15
10
5
0
180
190
200
210
220
length of call (seconds)

Why use percentage relative frequency?

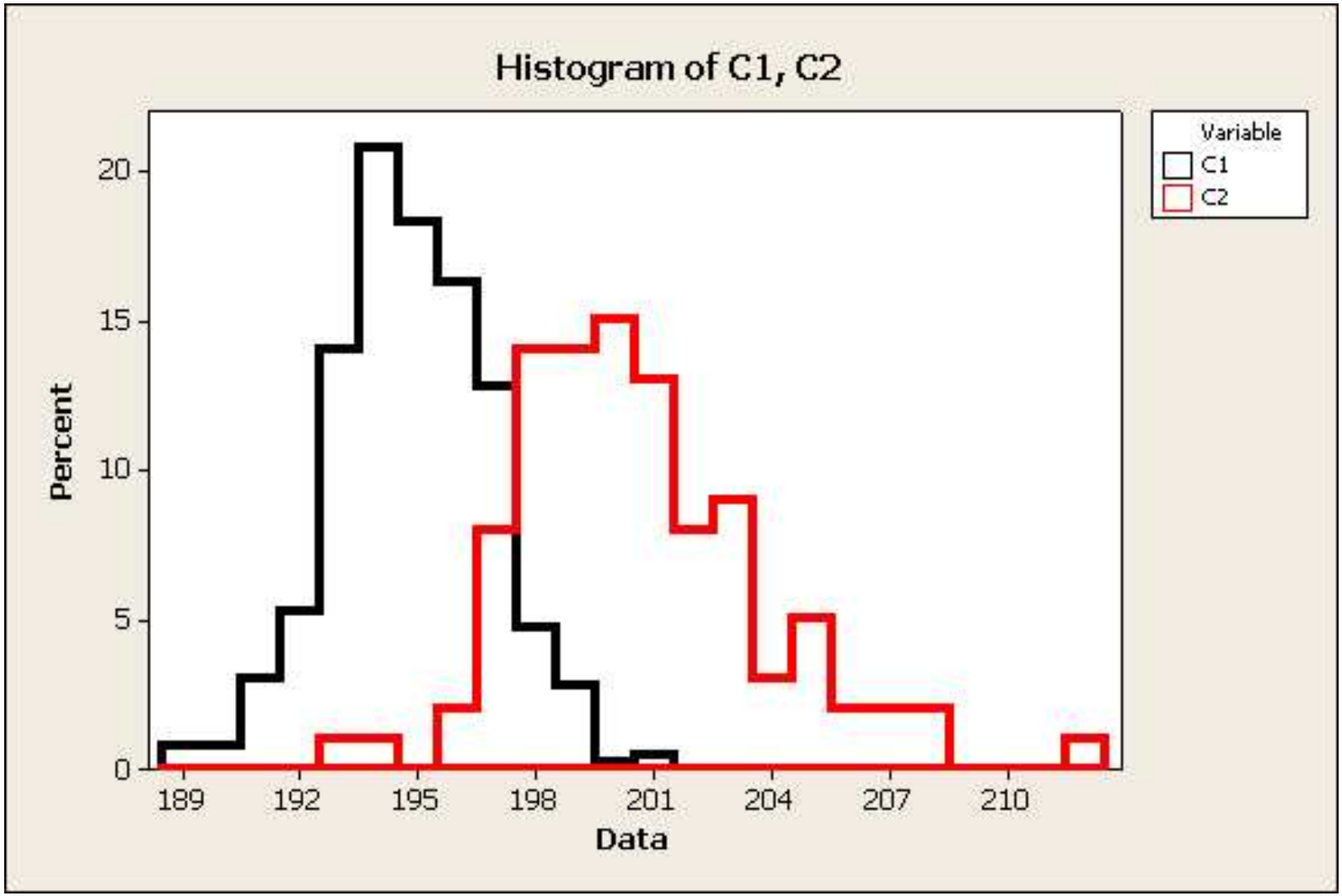
It is useful for comparing two or more histograms

- 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

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



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



This enables a more direct comparison between the two samples!

Relative frequency 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 **relative frequency polygons**.

Relative Frequency Polygons

These are a **natural extension** of the relative frequency histogram.

They differ in that, rather than drawing bars, each class is represented by one point and these are joined together by straight lines.

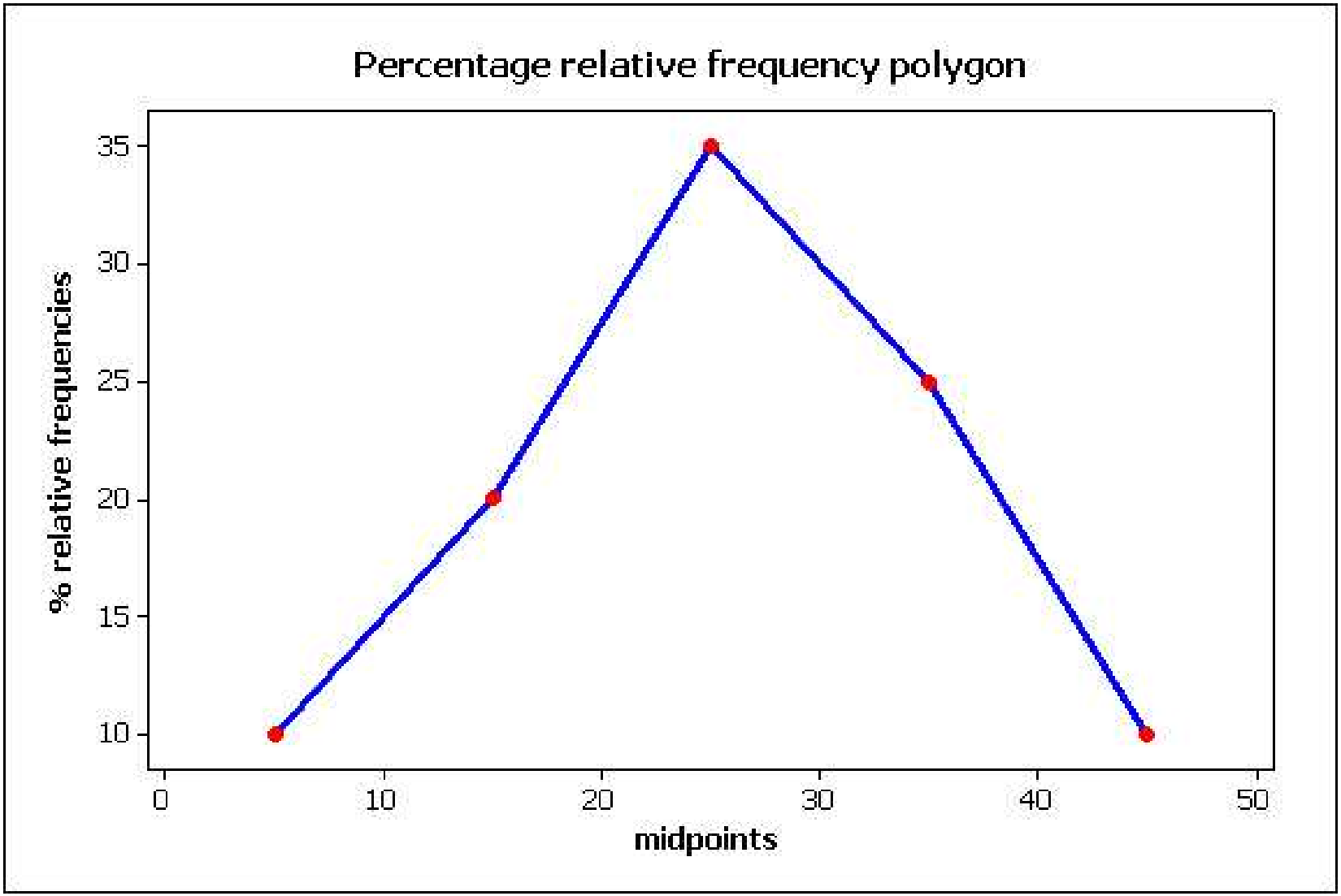
The method is similar to that for producing a histogram:

1. Produce a percentage relative frequency table
2. Draw the axes
3. Plot points: pick the mid point of the class interval on the x-axis and go up until you reach the appropriate percentage value on the y-axis and mark the point
4. Do this for each class
5. Join the points together with straight lines

Simple example

Produce a relative frequency polygon for the following data:

Class Interval	Mid Point	% Relative Frequency
$0 \leq x < 10$	5	10
$10 \leq x < 20$	15	20
$20 \leq x < 30$	25	35
$30 \leq x < 40$	35	25
$40 \leq x < 50$	45	10

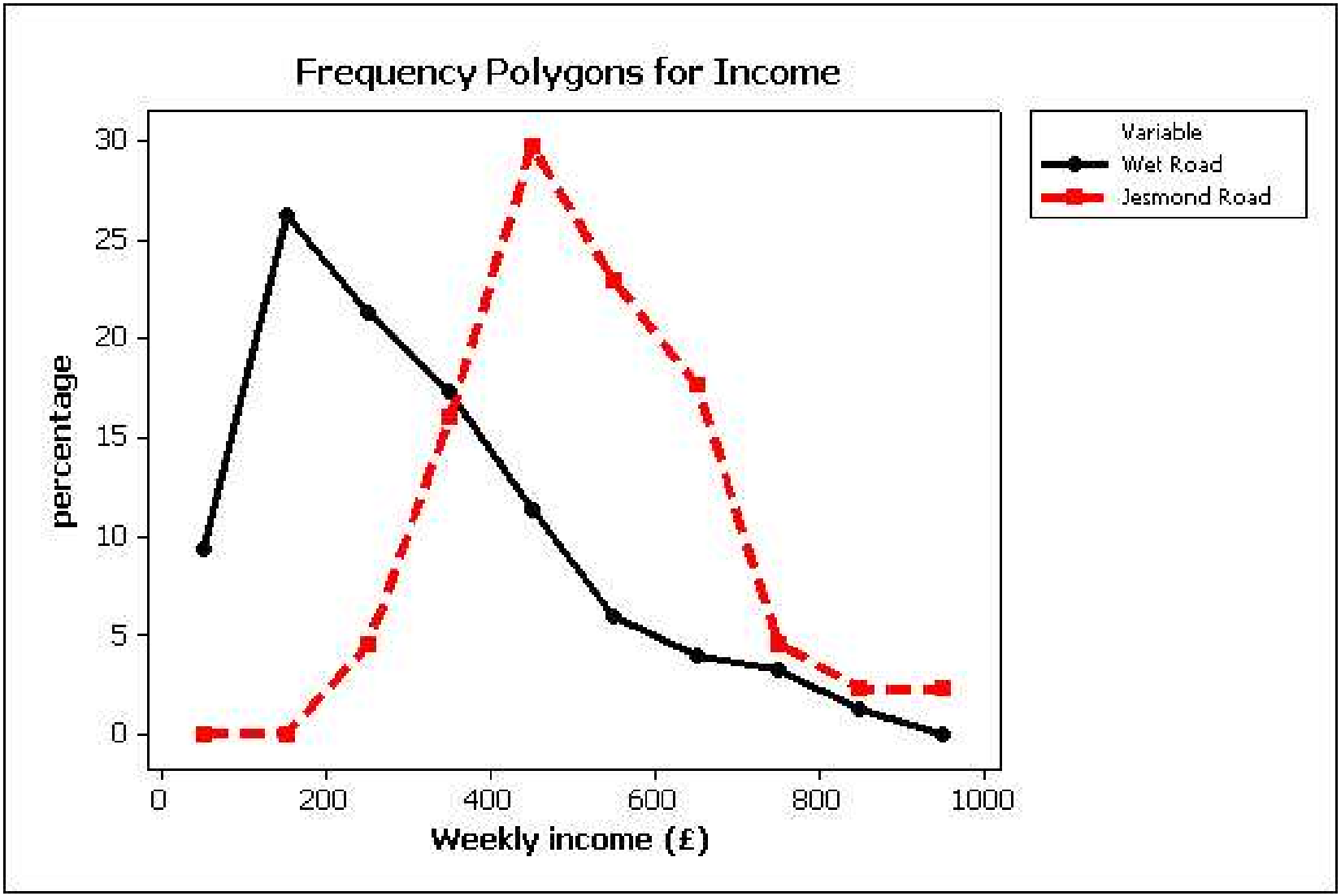

Percentage relative frequency polygon
% relative frequencies
35
30
25
20
15
10
0
10
20
30
40
50
midpoints

These percentage relative frequency polygons are very useful for comparing two or more samples

- We can easily overlay many polygons
- overlaying just *two* histograms gets messy!

Consider the following data on gross weekly income (in £) collected from two sites in Newcastle.

Weekly Income (£)	West Road (%)	Jesmond Road (%)
$0 \leq$ income < 100	9.3	0.0
$100 \leq$ income < 200	26.2	0.0
$200 \leq$ income < 300	21.3	4.5
$300 \leq$ income < 400	17.3	16.0
$400 \leq$ income < 500	11.3	29.7
$500 \leq$ income < 600	6.0	22.9
$600 \leq$ income < 700	4.0	17.7
$700 \leq$ income < 800	3.3	4.6
$800 \leq$ income < 900	1.3	2.3
$900 \leq$ income < 1000	0.0	2.3


Frequency Polygons for Income
percentage
30
25
20
15
10
5
0
0
200
400
600
800
1000
Weekly income (£)
Variable
Wet Road
Jesmond Road

Cumulative frequency polygons (Ogives)

Cumulative percentage relative frequency is also a useful tool.

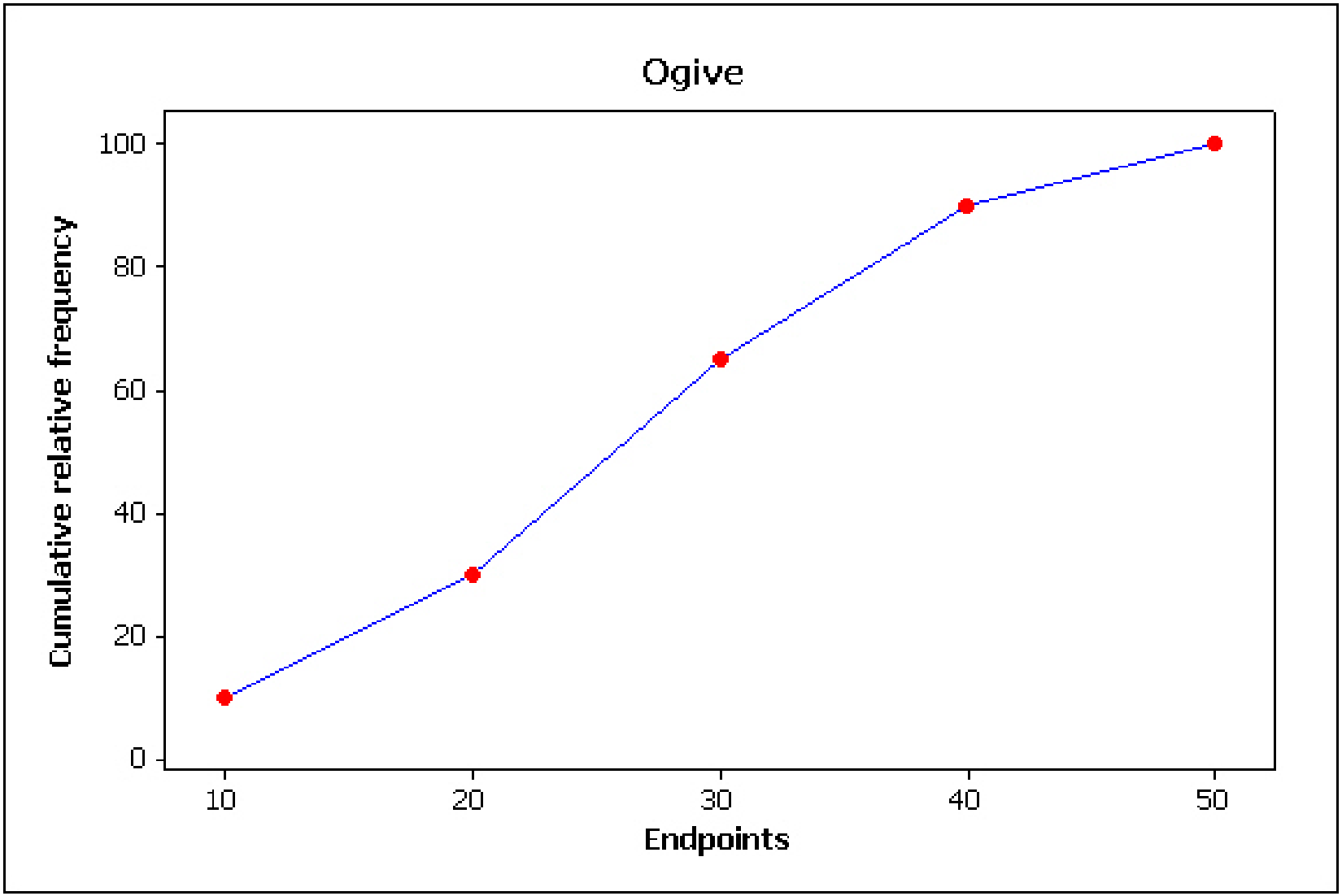
The cumulative percentage relative frequency is simply the sum of the percentage relative frequencies at the end of each class interval.

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

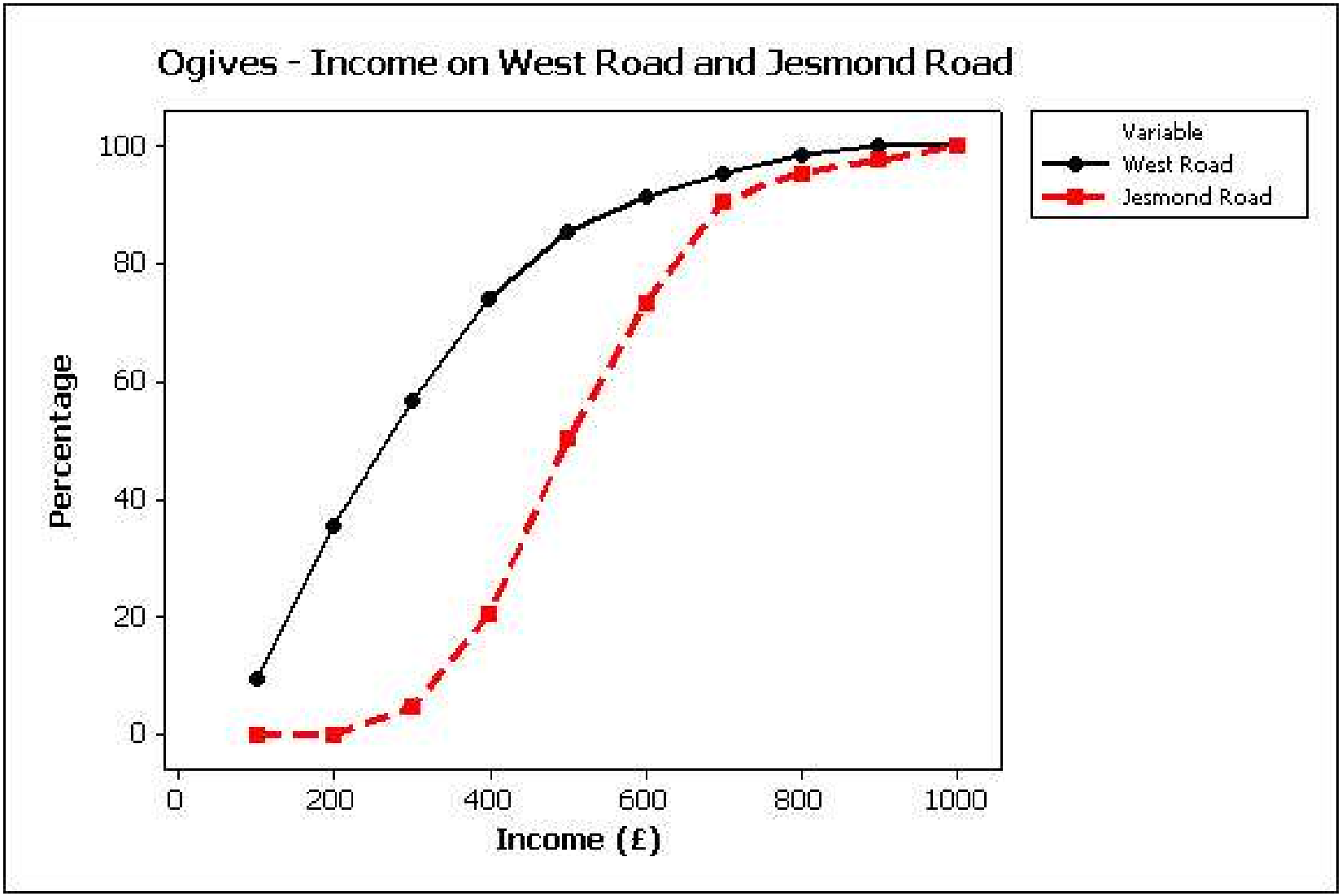
Consider the example from the previous section:

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

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



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



Pie charts

Pie charts are simple diagrams for displaying **categorical** or **grouped** data.

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

A pie chart consists of a circle divided into segments, one segment for each category.

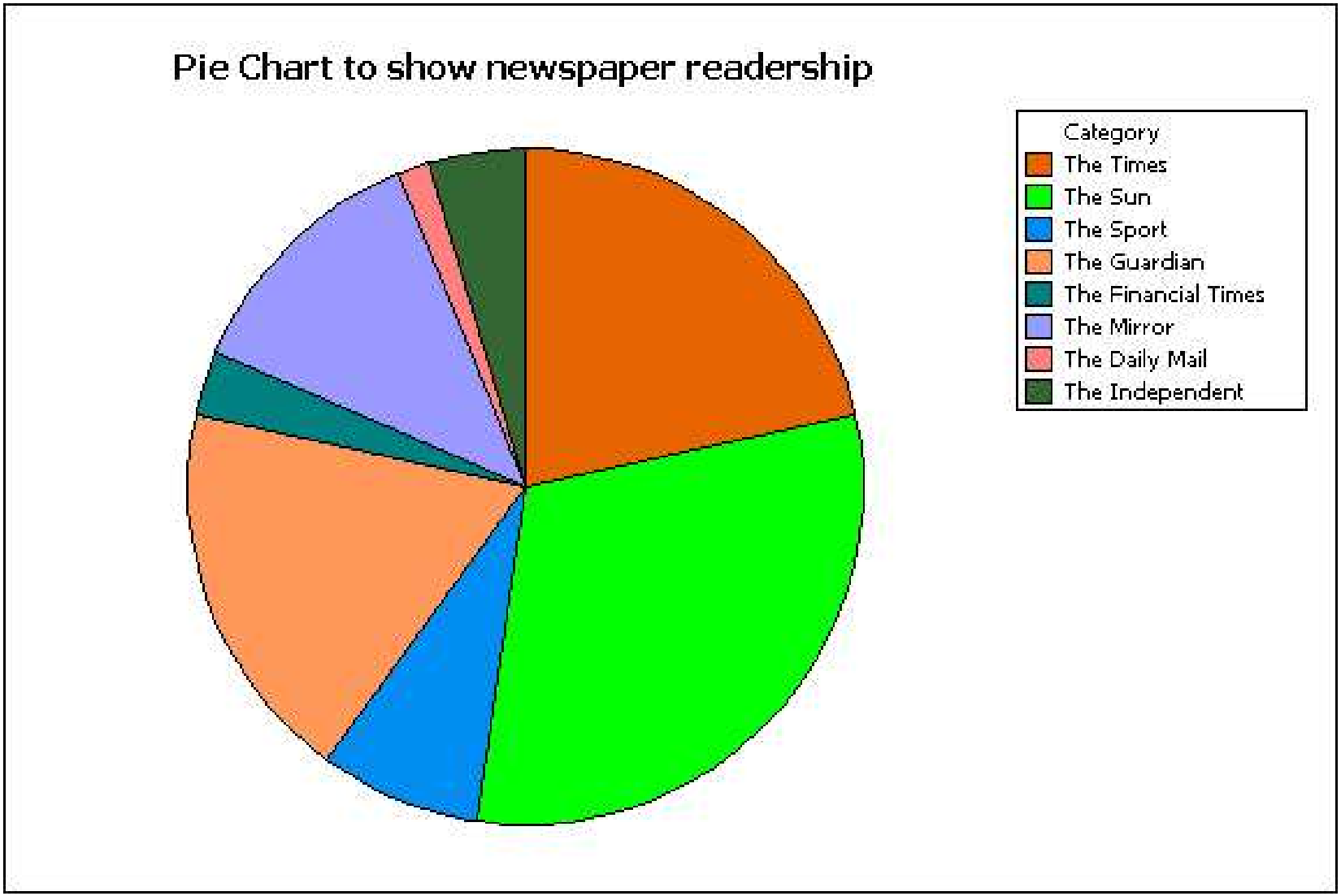
The **size** of each segment is determined by the **frequency** of the category and is measured by the **angle** of the segment.

As the total number of degrees in a circle is 360, the angle given to a segment is 360° times the fraction of the data in the category, that is

$$\text{angle} = \frac{\text{Number in category}}{\text{Total number in sample } (n)} \times 360.$$

Consider the data on newspaper sales to 650 students.

Paper	Frequency	Degrees
The Times	140	77.5
The Sun	200	110.8
The Sport	50	27.7
The Guardian	120	66.5
The Financial Times	20	11.1
The Mirror	80	44.3
The Daily Mail	10	5.5
The Independent	30	16.6
Totals	**650**	**360.0**


Pie Chart to show newspaper readership
Category
The Times
The Sun
The Sport
The Guardian
The Financial Times
The Mirror
The Daily Mail
The Independent

Time Series Plots

So far we have only considered data where we can **ignore the order** in which the data come.

Not all data are like this ...

... one exception is data which have been collected over **time**.

- Monthly sales of a product
- The price of a share at the end of each day
- The air temperature at midday each day

Such data can be plotted simply using **time** as the x-axis.

Example

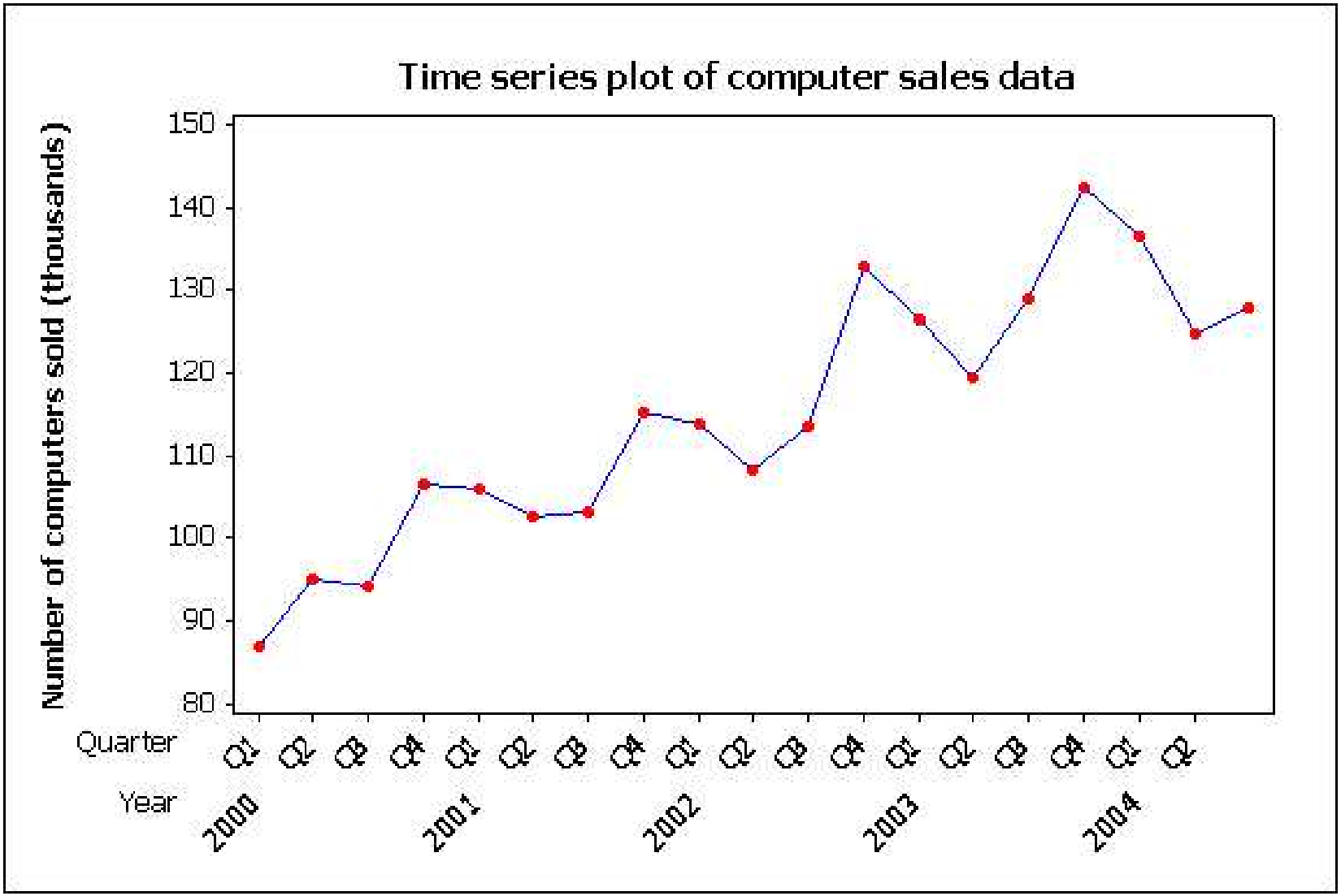
Consider the following data on the number of computers sold (in thousands) by quarter (January–March, April–June, July–September, October–December) at a large warehouse outlet.

Quarter	Units Sold	Quarter	Units Sold
Q1 2000	86.7	**Q1** 2003	126.3
Q2 2000	94.9	**Q2** 2003	119.4
Q3 2000	94.2	**Q3** 2003	128.9
Q4 2000	106.5	**Q4** 2003	142.3
Q1 2001	105.9	**Q1** 2004	136.4
Q2 2001	102.4	**Q2** 2004	124.6
Q3 2001	103.1	**Q3** 2004	127.9
Q4 2001	115.2		
Q1 2002	113.7		
Q2 2002	108.0		
Q3 2002	113.5		
Q4 2002	132.9		

By hand, a time series plot is constructed as follows:

1. Draw the x**–axis** and label over the time scale
2. Draw the y**–axis** and label with an appropriate scale
3. Plot each point according to time and value.
4. Draw lines connecting all points.

A time series plot of the computer sales data is shown on the next slide.


Time series plot of computer sales data
Number of computers sold (thousands)
150
140
130
120
110
100
90
80
Quarter
Year
Q1
Q2
Q3
Q4
2000
2001
2002
2003
2004

Scatter plots

The final type of graph we are going to look at is the **scatter plot**.

Such graphs are used to plot two variables which you believe might be **related** – for example:

- height and weight
- advertising expenditure and sales
- age of machinery and maintenance costs

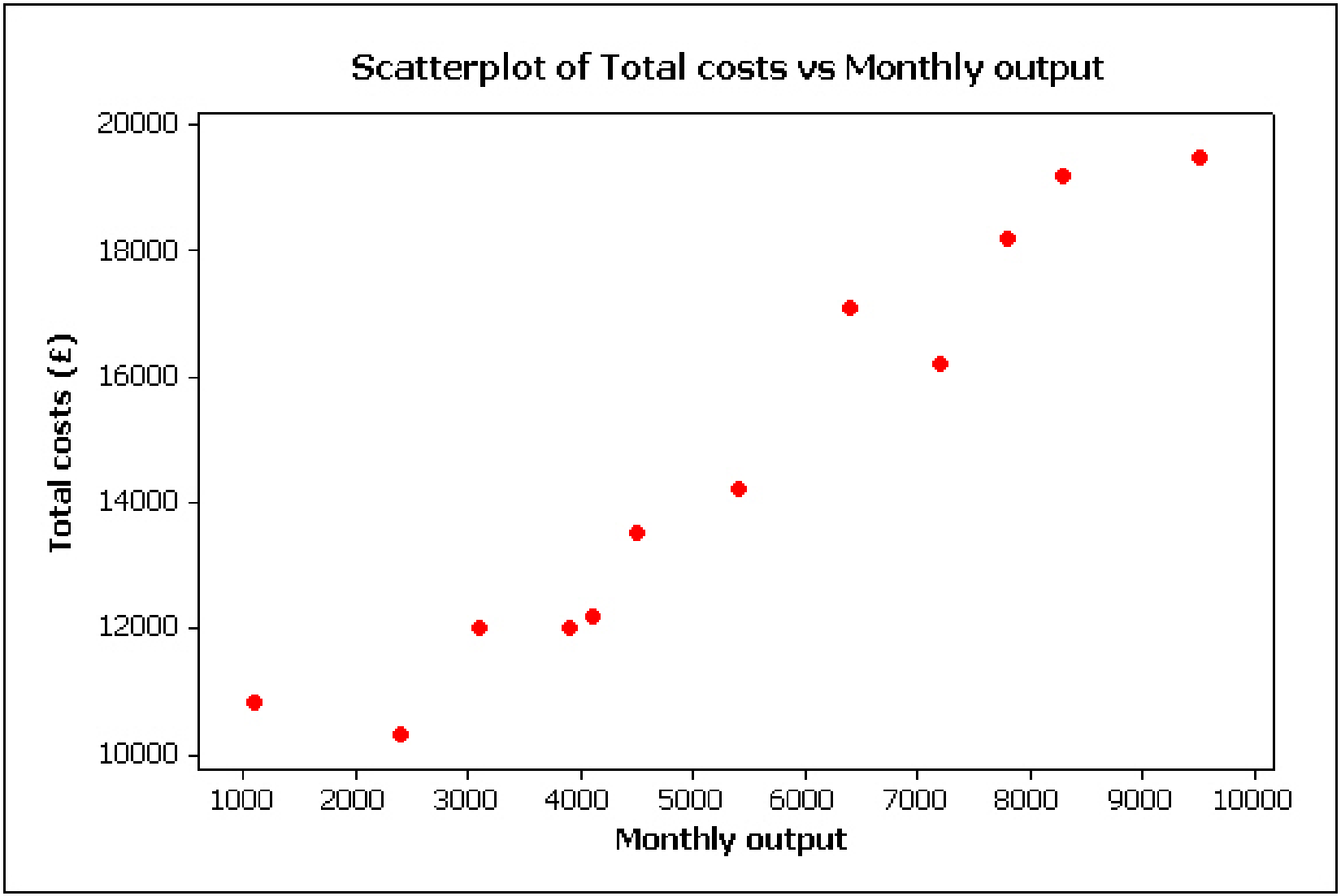
Example

Consider the following data for monthly output and total costs at a factory.

Total costs (£)	Monthly Output
10300	2400
12000	3900
12000	3100
13500	4500
12200	4100
14200	5400
10800	1100
18200	7800
16200	7200
19500	9500
17100	6400
19200	8300

If you were interested in the relationship between the cost of production and the number of units produced you could easily plot this by hand.

1. The "**response**" variable is placed on the y-axis. Here the response variable is "total costs"
2. The variable that is used to try to explain the response variable (here, monthly output) is placed on the x-axis – this is the **explanatory** variable
3. Plot the pairs of points on the graph


Scatterplot of Total costs vs Monthly output
Total costs (£)
20000
18000
16000
14000
12000
10000
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Monthly output