

Lecture 2

GRAPHICAL METHODS FOR PRESENTING DATA

Tutorials

Remember,

- this week tutorials are replaced with **computer practicals**;
- All groups **except B and G** attend the same slot as last week but should go to the **Brig/Moss/Pond clusters** in the Herschel Building:

Group	Day	Time
A	Wed	9–10
B	Tues	1–2 (<i>Wed 12 last week</i>)
C	Thurs	1–2
D	Thurs	2–3
E	Thurs	3–4
F	Fri	9–10
G	Thurs	10–11 (<i>Fri 11 last week</i>)

- ... you will be taught how to use the “l-Assess” software to access the online **Computer Based Assessments** (CBAs)

Introduction

So far, we have looked at

- how **data** can differ
- how to take **surveys**
- how to **sample** from the **population**
- how to display data in a **tabular form**
 - frequency tables
 - *relative* frequency tables

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

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

The easiest way to describe how such a plot is constructed is via demonstration...

...but first, you need to think about the following...

- You need to decide on a reasonable number of **intervals** which span the range of the data
- These interval widths must be **equal**
- You should use **sensible** values for the interval widths

Construct a stem and leaf plot for the following data:

11 12 9 15 21 25 19 8

We need to decide on an **interval width** – **idea: try 10** – i.e.

- 0–9
- 10–19
- 20–29

This gives a **stem unit** of 10 and a **leaf unit** of 1:

0		8	9		
1		1	2	5	9
2		1	5		

Stem Leaf

$n = 8$ stem unit = 10, leaf unit = 1.

Some notes...

- “Stem” units are to the left of the line, “leaves” to the right
- For example, for the first observation – **11** – we put a “1” in the stem (one ten) and a “1” as the first leaf
- Each leaf must be equally spaced along the row
- It’s like a **bar chart** sideways! (see later)
- But better – it contains all the **raw observations**
- Put the data in **ascending order** first!

Example 1: 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

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

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

Stem **Leaf**

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

Example 2: Unemployment rates in the U.S.

It all looks pretty easy... so what can go wrong?

Consider the following data, which are the percentage unemployment rates for 10 U.S. states:

17 18 15 14 12 19 20 21 24 15

If you were to choose 10 as the interval width (i.e. go up in 10s), the stem and leaf plot would look like

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

Stem **Leaf**

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

The interval width is **too large**!

Two intervals is not enough to reveal any **patterns** in the data

Idea: Use an interval width of 5!

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

Stem **Leaf**

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

We can see the pattern more clearly now!

So, to summarise...

- If the interval size is too large, any **patterns** in the data will be **hidden**
- But! If the interval size is too small, there will be lots of **empty intervals!**
- Use your own judgement!

Example 3: Call centre data

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

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											

Stem Leaf

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

Example 4: 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.

Bar charts are a commonly-used and clear way of presenting categorical data

- As with stem and leaf plots, various computer packages allow you to produce these
- First, let us work through the process of producing these by *hand*.

Constructing a bar chart is a 5 step process:

- 1 Decide what goes on each **axis** of the chart. By convention the variable being measured goes on the x-axis and the frequency goes on the vertical y-axis
- 2 Decide on a **numeric scale** for the frequency axis. This
 - represents the frequency in each category by its height
 - must start at zero and include the largest frequency
- 3 Decide on a **suitable number scale** to label this axis
- 4 Draw the axes and label them appropriately
- 5 Draw a bar for each category. When drawing the bars it is essential to ensure the following:
 - the width of each bar is the same
 - the bars are separated from each other by equally sized gaps

Example

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

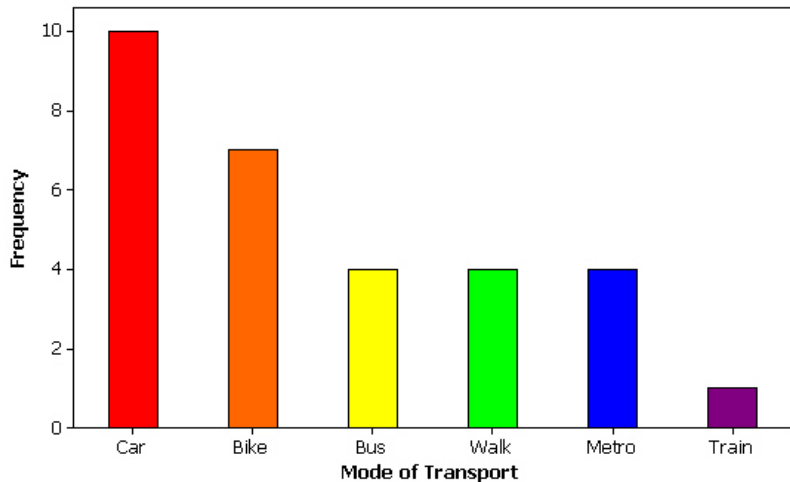
Draw a bar chart to represent these data

The first logical step is to put these into a frequency table, giving

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

We can then present this information as a bar chart, following the five steps on the previous slide.

Bar Chart to show Mode of Transport to University



Multiple Bar Charts

The data below gives the daily sales of CDs (in £) by music type for an independent retailer.

Day	Chart	Dance	Rest	Total
Monday	12000	10000	2700	24700
Tuesday	11000	8000	3000	22000
Wednesday	9000	6000	2000	17000
Thursday	10000	5000	2500	17500
Friday	12000	11000	3000	26000
Saturday	19000	12000	4000	35000
Sunday	10000	8000	2000	20000
Total	83000	60000	19200	162200

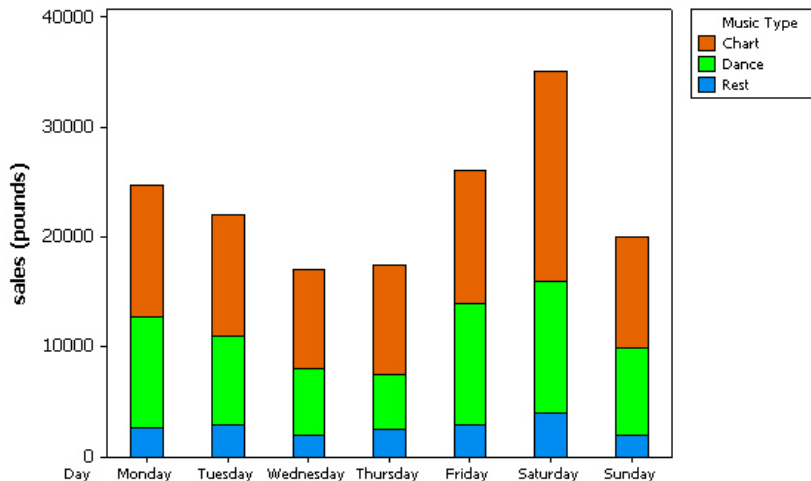
Bar charts could be drawn of **total sales per music type** in the week...

...or of **total daily sales**

It might be interesting to see **daily sales broken down into music types!**

This can be done using a **multiple bar chart**

Chart of Sales (pounds) vs Day and Music Type



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**.

At first sight these 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*
- The *area* of the rectangle is proportional to the frequency – not the height

Initially we will only consider histograms with equal class intervals.

Example

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

Produce a histogram to display these data graphically

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

Once we have the frequency table, the process is very similar to before...

- Find the maximum frequency and draw the **vertical** (y -axis) from zero to this value, including a sensible numeric scale
- The range of the **horizontal** (x -axis) needs to include not only the full range of observations but also the full range of the class intervals from the frequency table.
- Draw a bar for each group in your frequency table. These should be the same width and touch each other

Histogram to show length of service times for a call centre

