Displaying and describing the distributions of numerical variables

Friday, 1 February 2019 8:48 PM

By the end of the lesson I would hope that you have an understanding and be able to apply to questions the following concepts:

- Understand why we would need to have methods to display numerical data
- Understand what a grouped frequency distribution is and how to create one.
- Understand what a histogram is and how to create one (both by hand and using your CAS):
  - From a frequency table
  - From Raw data
- Know what to look for in a histogram
  - Shape
  - Outliers
  - Centre
  - Spread
  - Range

RECAP:

In the previous lesson we looked at displaying and describing categorical data.
This included:

Frequency tables
Bar charts
Stacked/Segmented bar charts

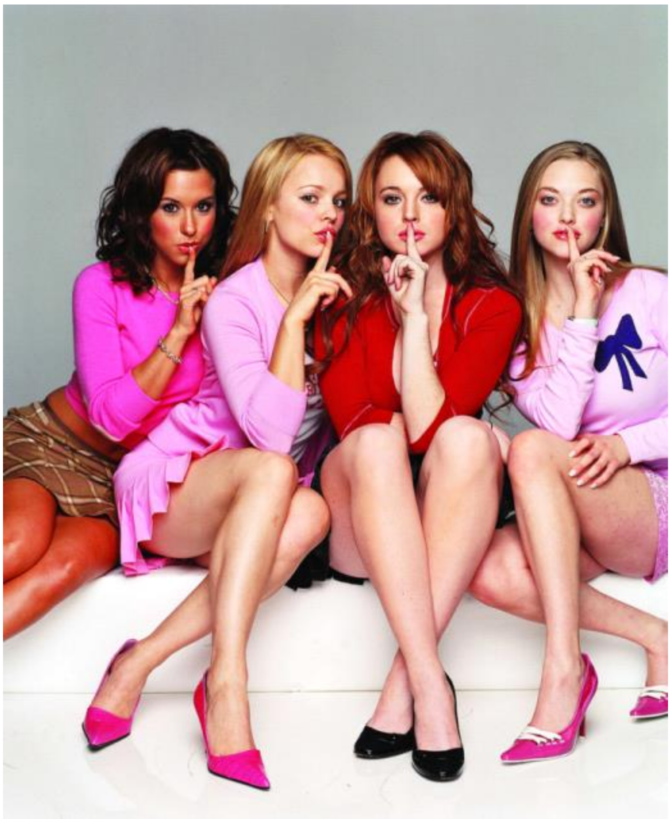
Categorical data might seem easy in comparison to numerical data as, categorical data deals with names. It's unlikely we are going to deal with data having hundreds of different names.

Sadly, this isn't true with numerical data.

Imagine doing a survey and asking people their ages. We might end up with 100 different answers. How would we even begin to display this information?

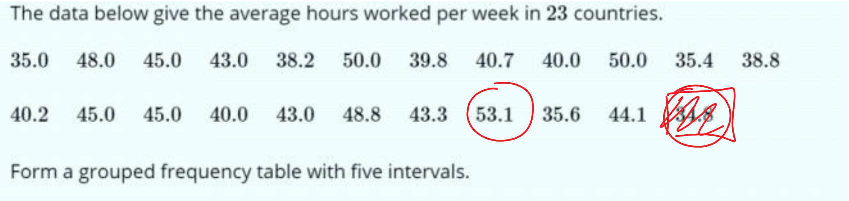
Grouping

There are bad groups ...



And then there are good groups:

And then there are groups of numbers!



The data below give the average hours worked per week in 23 countries.

35.0 48.0 45.0 43.0 38.2 50.0 39.8 40.7 40.0 50.0 35.4 38.8

40.2 45.0 45.0 40.0 43.0 48.8 43.3 53.1 35.6 44.1 34.8

Form a grouped frequency table with five intervals.

We can see from the above that we have taken a survey and asked people the average hours worked.
They have been asked to ensure their answers are correct to **1 decimal place**.
We can see a lot of numbers there.

We **could** leave the numbers like this, but it's better to display them using a **grouped frequency table.**

How to draw a grouped frequency table

The first thing we need to know is how many numbers we are going to include in an **interval**.
Wot is an interval?
It's a gap which includes a range of numbers.

30 – 34.9
35 – 39.9
40 – 45

The question says that it wants **5 intervals**.
These intervals are (normally) the same size.

Let's look at the highest number in the table: 53.1
Let's look at the lowest number in the table: 34.8

This would seem to suggest that intervals of 5 would make sense.

Now we construct the table.
All the tables can be set up with the same headings:

Average hours per week	Frequency	Column1
	Number	Percentage
30.0 – 34.9	1	4.3
35.0 – 39.9	6	26.1
40.0 – 44.9	8	34.8
45.0 – 49.9	5	21 7
50.0 – 54.9	3	13.0
Total	23	99.9

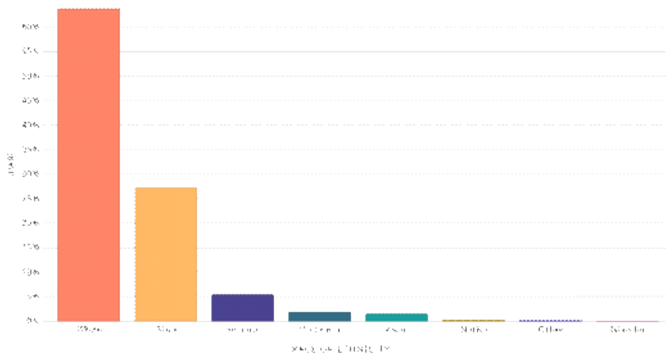
Be careful with the intervals!
Understanding which numbers form the intervals is really important

Not very pretty is it! So let's turn this into something much nicer.

The HISTOGRAM

Grouped frequency tables are nice, but they can be turned into something much nicer called a Histogram.

Basically (and sort of lying to you!), a Histogram is a bar chart with no gaps.



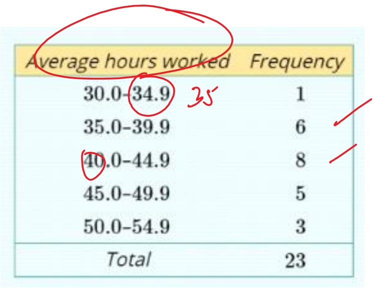
This is a bar chart!
It has gaps!
It's awesome for categorical data.

How to draw a histogram

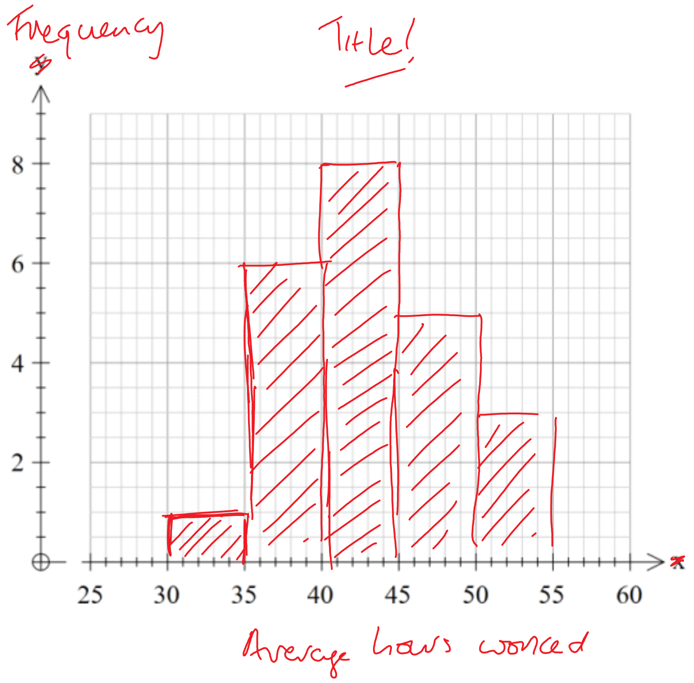
Firstly, we need to know which information to use on a histogram!

- Frequency
- Values of the variables

Awesome! We can extract that from the table above and get:



Average hours worked	Frequency
30.0-34.9 35	1
35.0-39.9	6
40.0-44.9	8
45.0-49.9	5
50.0-54.9	3
Total	23



Frequency always goes on the vertical axis.
The values of the variables goes on the horizontal axis.
Each interval has its own bars
There are no gaps between the bars.

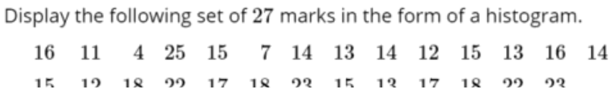
Can't I do this quicker?

Of course!
This is a CAS course! And, so, let's use the CAS to make this quicker.

The process:

- MENU -> Statistics
- Enter the data into a single list
- Ensure you give it a title
- Open the Set StatsGraph function
  - DRAW: ON
  - Type: Histogram
  - Xlist: main\marks
  - Freq: 1
  - SET
- Tap the graphing ICON
- Set Interval
  - Hstart: 2
  - Hstep: 4
  - OK

Let's try it with the data shown below:



Display the following set of 27 marks in the form of a histogram.

16 11 4 25 15 7 14 13 14 12 15 13 16 14
15 12 18 22 17 18 23 15 13 17 18 22 23

Display the following set of 27 marks in the form of a histogram.

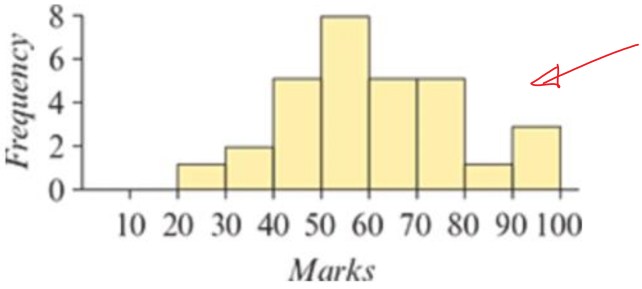
16 11 4 25 15 7 14 13 14 12 15 13 16 14
15 12 18 22 17 18 23 15 13 17 18 22 23

Describing a Histogram

Art is pretty ... but people keep wanting to describe it!
The same is true with Histograms.
It's lovely to see one, but we will be expected to describe the key features.

We can do this using three main talking points:

- Shape and outliers
- Centre
- Spread



Talking point 1: Shape and Outliers

We can talk about the shape of a histogram using an idea of Symmetry.

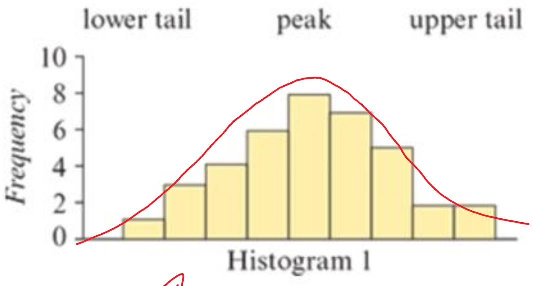
This face is symmetrical.
If we were to draw a line directly down the middle, we would see the mirror image on both sides.

Sadly, not all faces are symmetrical!
In fact, very rarely are they!

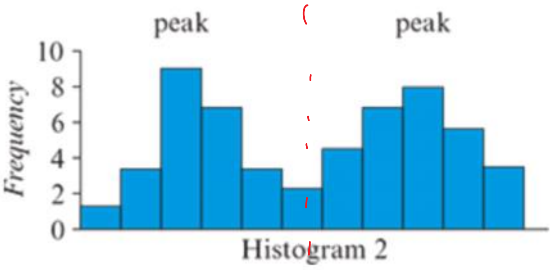
There are some awesome apps now which will show you how odd you would really look if your face was symmetrical.

If something is symmetrical it means we can (generally) draw a line down the middle and the left hand side of the line and right hand side of the line would have (roughly) the same shape.

What can we say about the following two histograms?



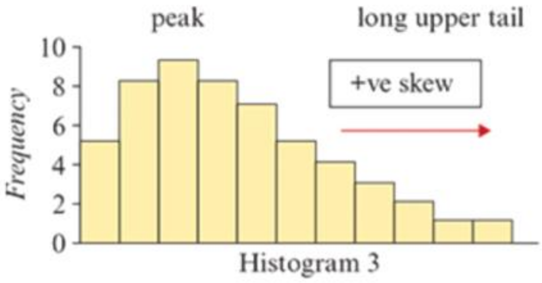
Single peaked distribution

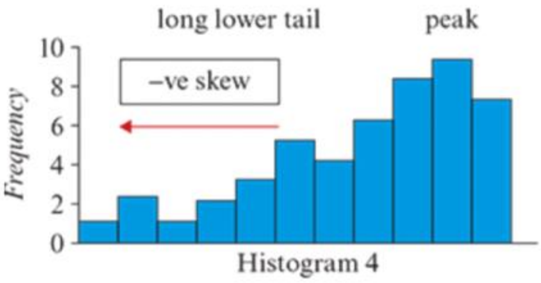


Twin-peaked .... NO!
Double-peaked distribution
Shows the data is **bi-modal**!

Like our faces ... not all data is symmetrical!

How might we describe the following data?



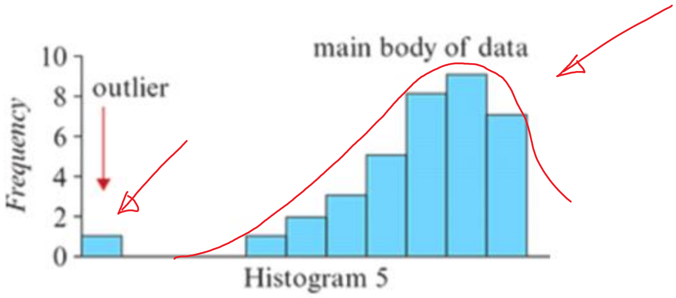


What about the loners (Outliers)?

Occasionally there might be one lonely piece of information stuck out on its own.
Much like me at a party!
This is called an outlier.

Generally they are data items which are too high or too low to fit into the rest of the data.
There are tests and calculations we can do to find the outliers, but that's beyond this part of the course.

Can you see the outlier below?

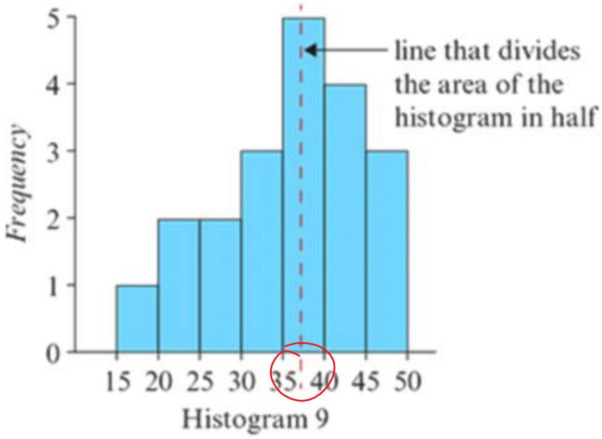


Keeping things in the centre?

Having described the shape of the histogram.
We need to let the viewer know where the data is centred around.
The centre is very important as it helps us **compare** histograms.

Generally, the centre of a histogram is where the middle value lies.
Hence we need to find where 50% of the data lies.

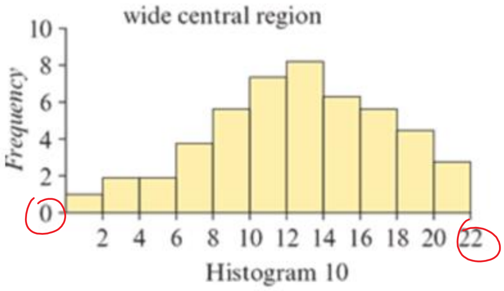
This is really easy if we have the frequencies as we can find out how many items there are in total, divide that by 2 and then locate which bar that data item lies in.



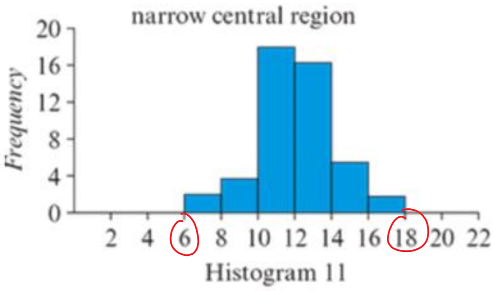
Finally we talk about Vegemite ... no .... Hold on ... Spread!

Are you a corner spreader?
Or do you just puddle the stuff in the middle?

The spread of the data is how widely it is distributed.
How would you describe the following?



Range = 22 - 0
= 22



Range = 18 - 6
= 12

We can describe the spread using the **range**.
You have already met the **range** in previous years!
Mean, Median, Mode and range.
The **Range** is the difference between the highest data item and the lowest data item.

Bringing it all together

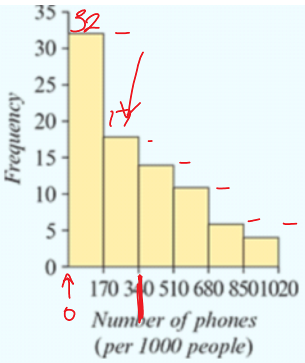
Let's bring it all together with a question from the *Cambridge Further Maths Textbook*

The histogram opposite shows the distribution of the number of phones per 1000 people in 85 countries.

a Describe its shape and note outliers (if any).

b Locate the centre of the distribution.

c Estimate the spread of the distribution.



The data is +vely skewed

No outliers

b) Centred around 170 - 340 phones per 1000 ppl

49

c) 1020 - 0 = 1020

Writing the answers in the form of a report

Report

For these 85 countries, the distribution of the number of phones per 1000 people is positively skewed. The centre of the distribution lies between 170 and 340 phones/1000 people. The spread of the distribution is 1020 phones/1000 people. There are no outliers.