1 Averages

There are many types of **averages** /ˈæv.ᵊr.ɪdʒ/ . Two that we will meet are called the **mean** and the **median**. They are also known as measures of **central tendency**.

2 Mean

The arithmetic mean, in short the mean, is the proper name for what most people call the average. /miːn/

Definition : The mean of a set of values is the sum of all the values divided by the number of values.

The formula is written as $\bar{x} = \dfrac{\sum_{i=1}^{i=n} x_i}{n}$, where n is the number of values and $(x_i)_{1 \le i \le n}$ are the values,

in short $\bar{x} = \dfrac{\sum x}{n}$

read : « x bar is equal to sigma (the sum) of all the x-values divided by n »

3 Median

The values are arranged in ascending, or descending, order of size, then the median is a number dividing the set in two subsets of equal size : half the values are less than this median, half the values lie above.

If the number of values is even, then the median is the mean of the two middle values.

If the number of values is odd, then the median is the middle value. /ˈmiː.dɪ.ən/

Example : Let 3 ; 7 ; 10 ; 5 ; 4 ; 2 ; 2 ; 8 ; 4 be an array of numbers.
Find the mean and median.

Mean : $\bar{x} = \dfrac{\sum x}{9} = \dfrac{3+7+10+5+4+2+2+8+4}{9} = \dfrac{45}{9} = 5$

Median : first step, write in ascending order 2 ; 2 ; 3 ; 4 ; 4 ; 5 ; 7 ; 8 ; 10
there are 9 values, that is an odd number of values, the median is the middle value : 4

4 Frequency distribution

If the values in a distribution are arranged in ascending or descending order, showing their corresponding frequencies, the distribution is called a **frequency distribution**. /ˈfriː.kwənt.sɪ/

A table showing such a set of data is called a **frequency distribution table**.

Vocabulary average – frequency distribution (table) – mean – median – central tendency

4.1 Mean

To find the mean of a frequency distribution, do the following :

- ✔ Multiply each value by its corresponding frequency;
- ✔ Sum all these products;
- ✔ Divide this sum by the sum of all the frequencies.

That is $\bar{x} = \dfrac{\sum_{i=1}^{i=n} f_i x_i}{\sum_{i=1}^{i=n} f_i}$

where :

n is the number of measurements,

$(x_i)_{1 \le i \le n}$ are the values of the measurements,

f_i is the frequency of the measurement x_i ,

in short : $\bar{x} = \dfrac{\sum f x}{\sum f}$

4.2 Median

As the values are arranged in order of size in a frequency distribution table, the median can be read by looking at the row displaying the frequencies :

- ✔ add up all the frequencies : this is the total number of values in the set;
  - o if it is odd, add 1 and divide by 2 for computing the position of the median in the ordered set;
  - o if it is even, divide by 2, save and add 1 for computing the two positions of the values which mean is the median
- ✔ add up the frequencies from the left until you reach the said position or pair of positions;
- ✔ retrieve the corresponding value or pair of values;
- ✔ work out the median : the value itself if it is single, the mean if there is a pair of values.

<u>Example</u> : A test consisted of five questions. 1 mark was awarded per question for a correct solution and no marks for an incorrect solution. The following frequency distribution table shows how a class of students scored in the test.
Calculate the mean and median marks.

mark	0	1	2	3	4	5
Number of students	1	3	6	7	7	4

Mean : $\bar{x} = \dfrac{\sum f x}{\sum f} = \dfrac{1(0)+3(1)+6(2)+7(3)+7(4)+4(5)}{1+3+6+7+7+4} = \dfrac{0+3+12+21+28+20}{28} = \dfrac{84}{28} = 3$

Median : The sum of all the frequencies is $\sum f = 28$

This sum is even, then we compute two positions : $\dfrac{28}{2} = 14$ and $14+1 = 15$

We are looking for the 14th and the 15th marks, taken one by one. We work out a table showing the partial sums of frequencies :

Marks	0	1	2	3	4	5
Number of students scoring the mark	1	3	6	7	7	4
Number of students scoring **up to** the mark	1	1+3 = 4	4+6 = 10	10+7 = 17	17+7 = 24	24+4 = 28

Both 14th and 15th students scored 3, hence the median $\dfrac{3+3}{2} = 3$

4.3 Grouped frequency distribution

Sometimes the range of the values is very wide and it is not suitable to show all the values individually. When this happens we arrange the values into suitable groups called **class intervals** . When the information is arranged in class intervals, it is not possible to calculate the exact value of the mean. However it is possible to estimate it by using the **mid-interval value** of each class interval. The easiest way to find the mid-interval value is to add the two extreme values and divide by 2.

example :

Otherwise, the procedure for estimating the mean is the same as in the previous section.

4.4 Weighted mean

A weighted mean is one where the frequencies are replaced by **weights.** The weight /weɪt/ is a measure of importance of a particular value, *i.e.* each value has a statistical weight attached to it.

Calculating a weighted mean is exactly the same as finding the mean of a frequency distribution, except that the frequencies are replaced by weights, in short :

$$\bar{x} = \frac{\sum wx}{\sum w}$$ where $(w_i)_{1 \le i \le n}$ are the weights.

5 Graphical displays

5.1 Histogram

A **histogram** /ˈhɪs.tə.græm/ is constructed from a frequency table, when the scores are gathered into **class intervals**.

The intervals are shown on the ***x***-axis, and the frequencies of the scores in each interval are represented by the areas of rectangles located above each class interval : areas are proportional to frequencies.

Frequency or **relative frequency** may be used, the graph will have the same shape in either case.

Frequency is commonly used when the data set is small and relative frequency is used when the data set is large, or when we want to compare several distributions.

The main point to remember about histograms is that the area of the rectangles, not their heights, represent the frequencies : a class interval twice as wide will induce a rectangle twice lower than we might think.

Example :

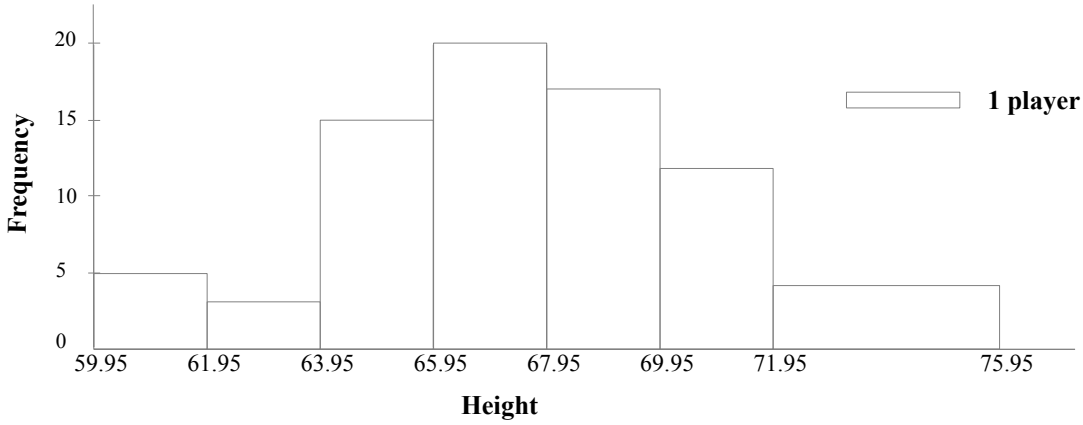
The following data are the heights (in inches to the nearest half of a tenth of an inch) of eighty male semi-professional soccer players.

n.b. : an inch is approximately 2.54 cm, and there are 12 inches in a foot.

height	[59.95 ; 61.95]	[61.95 ; 63.95]	[63.95 ; 65.95]	[65.95 ; 67.95]	[67.95 ; 69.95]	[69.95 ; 71.95]	[71.95 ; 75.95]
frequency	5	3	15	20	17	12	8

These data can be represented as the histogram drawn below.
Note *: the height of the last rectangle is not 8, but 4, as the last class interval is twice as wide as the others.*

Vocabulary class interval – histogram – mid-interval value – (frequency) polygon – relative frequency – weight



5.2 Frequency polygon

The **frequency polygon** is the simplest graphical display of a frequency table.

The scores are shown on the ***x***-axis and the frequencies are shown on the ***y***-axis : points which cartesian coordinates are these numbers are plotted in a cartesian system of reference. The points are then connected so that together with the ***x***-axis they form a polygon. The points of a frequency polygon corresponding to the lowest and highest values always lie on the ***x***-axis.

Absolute or relative frequencies may be used. The graph will have the same shape in either case. Frequency is commonly used when the data set is small and relative frequency is used when the data set is large, or when we need to compare several distributions.

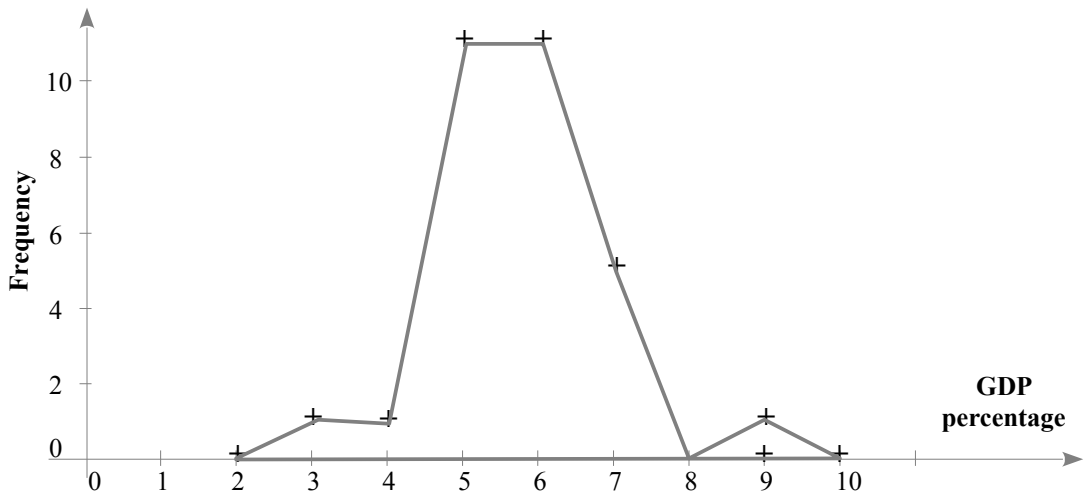
Frequency polygons are useful for comparing distributions and are also often used with cumulative frequencies.

Example :

Each country spends a certain percentage of its **Gross Domestic Product** on education. The table below shows, among 30 countries all over the world, how many spend a certain percentage. Values have been rounded to the nearest integer. For example, only one country spends 4 % of its GDP on education, while 5 countries spend approximately 7 %.

Percentage of the GDP	3	4	5	6	7	9
Number of countries	1	1	11	11	5	1

These data can be graphicaly represented as a frequency polygon.



Vocabulary frequency polygon – Gross Domestic Product – GDP

5.3 Bar graph

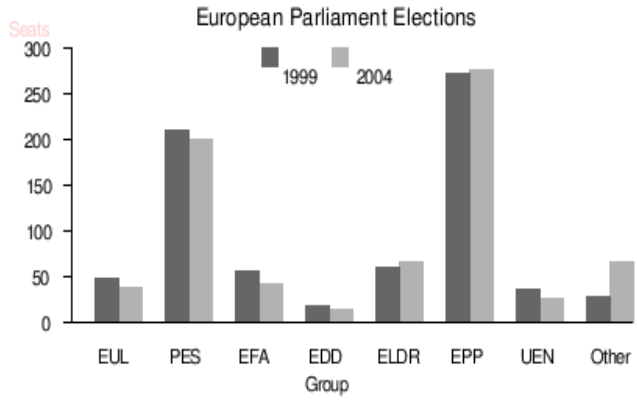
An alternative to the frequency polygon is the **bar graph** , where non-**contiguous** /kənˈtɪg.ju.əs/ bars (or thin rectangles) are drawn, with lengths proportional to the values that they represent. The bars can also be plotted horizontally.

This should mainly be used with **qualitative** or **discrete** /dɪˈskriːt/ data.

Example :

The following table lists the number of seats allocated to each party group in European elections in 1999 and 2004. The results of 1999 have been multiplied by 1.16933, to compensate for the change in number of seats between those years. Sometimes it can be horizontal.

This bar chart shows both the results of 2004, and those of 1999:



5.4 Pie chart

A **pie chart** (or a circle graph) is a circular chart divided into sectors, illustrating relative magnitudes or frequencies or percents. In a pie chart, the arc length of each sector (and consequently its area, and its central angle) is proportional to the quantity it represents.

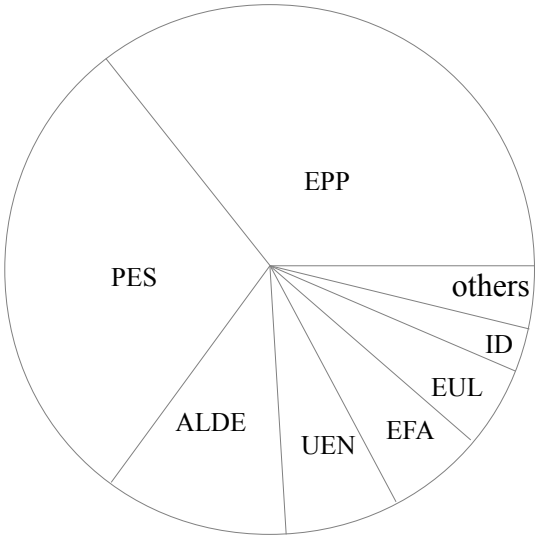
Together, the sectors create a full disk. It is named for its resemblance to a pie which has been sliced.

Example :

The following example chart is based on the results of the election for the European Parliament in 2009.

The following table lists the number of seats allocated to each party group, along with the derived percentage of the total that they each make up. The values in the last column are the derived central angles of each sector, an angle is the product of the percentage by 360°.

Group	Seats	Percent (%)	Angle (°)
EPP	284	36.2	130
PES	215	27.4	100
ALDE	103	13.1	47
UEN	44	5.6	20
EFA	42	5.4	19
EUL	41	5.2	18
ID	24	3.1	11
others	32	4.1	15
Total	785	100	360



The size of each central angle is proportional to the size of the corresponding quantity : the number of seats. Since the sum of the central angles has to be 360°, the central angle for a quantity that is the fraction $\boldsymbol{Q}$ of the total is 360 $\boldsymbol{Q}$ degrees. In the example, the central angle for the largest group (European People's Party, EPP) is 130° because $\mathbf{0.362 \times 360 \approx 130}$, rounded to the nearest integer.

Exercise 1

Vocabulary bar graph – (non-)contiguous – discrete – pie chart – qualitative

The pupils in a certain class were asked how they travelled to school. The results are shown in the table below :

Mode of transport	train	Bus	Bycicle	walking
Number of pupils	4	6	5	9

i. *Show the data in a bar graph*

ii. *Show the data in a pie chart*

5.5 Scatter plot

A **scatter plot** is a type of graph using cartesian coordinates to display sets of two variables data. The ***x***-axis is chosen for one of the variables, and the ***y***-axis is dedicated to the second variable. Each individual piece of data is represented by a point which abscissa is the value of the first variable, and ordinate is the second.

A scatter plot can suggest various kinds of **correlations** between variables. Correlations may be positive (rising), negative (falling), or null (**uncorrelated**).

If the pattern of dots **slopes** from lower left to upper right, it suggests a positive correlation between the variables being studied. This means that an increase in one of the variables generally yields an increase in the other variable.

If the pattern of dots slopes from upper left to lower right, it suggests a negative correlation. This means that

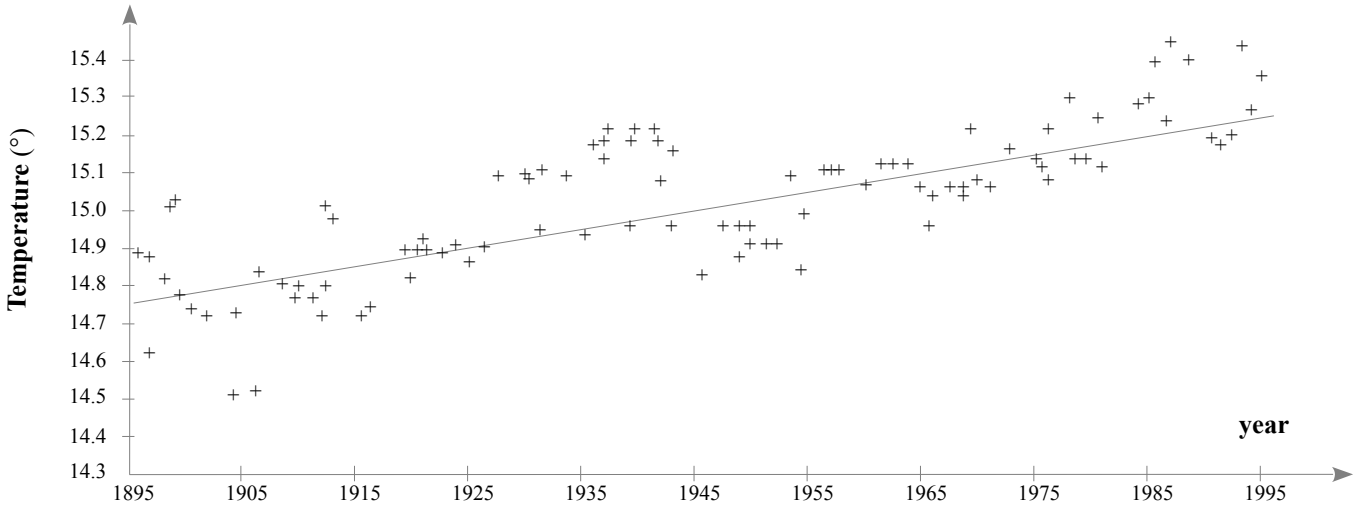
_____

If there is no obvious pattern in the scatter plot, it may be that the variables are uncorrelated.

Example :

A simple case of scatter plot is when one parameter is time, counted in years, days, seconds ...

Consider, for example, the global mean temperature on Earth, measured each year between 1896 and 1996. This data can be represented as a scatter plot, with year on the ***x***-axis and temperature on the ***y***-axis, as shown below.



When the scatter plot approximately follows a straight line, as is the case in our example, we say the there is a linear correlation between the variables. Many different methods exist to find a **line of best fit** (or **trendline**) for such a scatter plot.

The line of best fit computed with one of these methods has been drawn on our example scatter plot. It shows, without any possible doubt, that there has been an increase of global temperature between 1896 and 1996.

Vocabulary correlation – line of best fit – scatter plot – to slope – trendline – uncorrelated