Bringing British Education to You

www.nccedu.com

Skills for Computing

Topic 10:

Accuracy and Correlation ; Presenting Results

Learning Outcomes for this Topic

By the end of this topic students should be able to:

- Use standard errors
- Represent and analyse paired data
- Recognise and interpret correlation
- Analyse and present results correctly
- Understand ways in which statistics are misused
- Learn to recognise mistakes in the way others present results

Uncertainty - 1

- A useful description of any statistical result must include the statistical error
- The error shows the uncertainty of a result.
- In this module we deal with random variations. Not all errors can be handled with statistics.

Uncertainty - 2

- Random variations are caused by changes in the environmental conditions, or in the measurement system.
- We use statistics to handle this type of error. We need a large number of values for accurate results.

Other Errors

- User error:
  - For example not entering a value correctly or omitting it, or not using the correct units.
- Systematic error:
  - This can affect all the measurements in the dataset.
  - For example “zero error”: not measuring length starting from the beginning of a ruler (the zero.)
  - It is an error in procedure; it does not vary randomly.

Example

- How much rain falls per month in England on average?
  - Answer 1: 82.1
    - This answer is totally meaningless (and would get zero marks in an examination!)
  - Answer 2: 82.1 cm
    - This answer is better, but it seems strange that it rains exactly the same amount every month of every year!
  - Answer 2: 82.1 cm ± 31.8 cm
    - This is correct: it tells us a value, and how much variation there is in that value

Sources of Error - 1

- Bias:
  - We can introduce errors before we start the measurement, if we make measurements in a way which changes the distribution of values.
  - This type of error is called “bias” and is usually caused by poor experimental design.
  - For example, when we obtain results from a survey, we are not obtaining information about the general population, only about people who have the time and patience to complete a survey.

Sources of Error - 2

- Noise:
  - No measuring device is perfect, there will be some noise in the measurements.
  - Noise can also be caused by changing conditions and interactions with the environment.

Sources of Error - 3

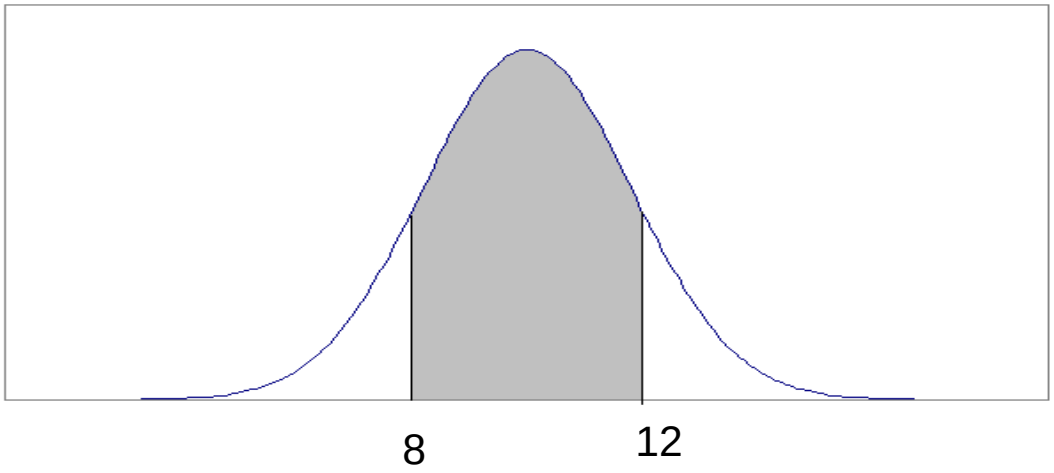
- Representation of values:
  - We can also introduce error when we have already made the measurement.
  - Rounding error: The online storage format uses fewer digits than the original data, changing its value.
  - Transfer error: This could be user error, or faulty communication between instrumentation and online storage. This type of error can be very large when it occurs.

Standard Error - 1

- Gives a measure of the distance of data values from the mean.
  - More accurate values, for example from more careful measurements, will have a smaller standard error.
  - The standard error is usually given as a range: the mean ± the standard error, for example 10 ±2.
- The standard error is an estimate of the standard deviation.

Standard Error - 2

- The shaded area shows the range of values that lies within a distance of one standard error from the mean.



Standard Error - 3

- The range of values which lie within one standard error either side of the mean is called the “spread”. About 70% of the values will be in this range.

- Another way of looking at this is that there is a probability of about 70% that a measurement will be close to the estimated mean (separated by at most one standard error.)

Standard Error - 4

- Sometimes scientists use a spread of ±two standard errors. In this case the spread covers about 95% of the range of values.

- We call this the 95% confidence interval. This means that about 95% of the time we will include the exact mean in this interval.

Combining Errors

- If measurements of different quantities are combined algebraically, then if they are uncorrelated we can usually combine the errors to obtain the error in the result.
- If we add or subtract quantities, the resulting standard error is the root of the sum of the individual standard errors.
- If we multiply or divide quantities, this rule applies to the relative standard error (the error divided by the quantity.)

Dependence

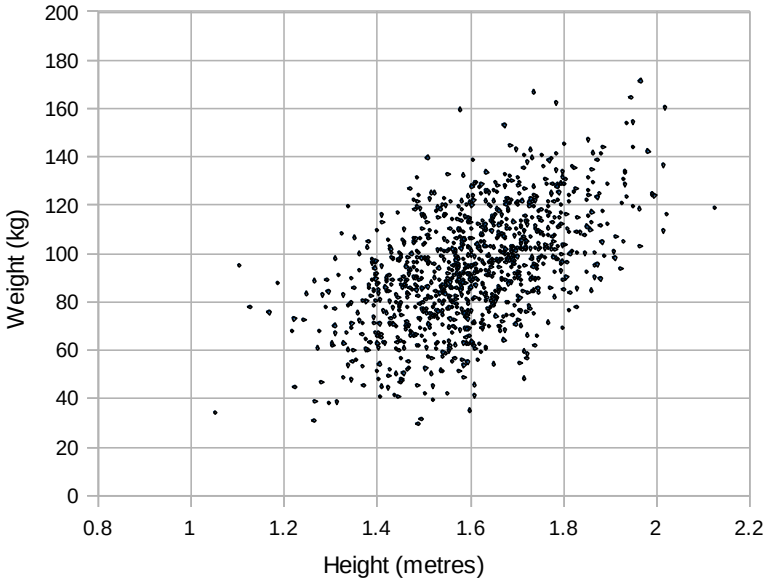
- We can combine data from two randomly varying quantities, to explore their relationship.
- If they are independent of each other, then analysing their combined values is the same as analysing them separately.
- We are often interested in whether one quantity directly changes another. We cannot determine this from statistical data analysis.
- However, even if all we know is that they vary in a similar way, this can be useful. For example, if one of the things is difficult to measure, we can deduce its value from the behaviour of the other.

Scatter Plots - 1

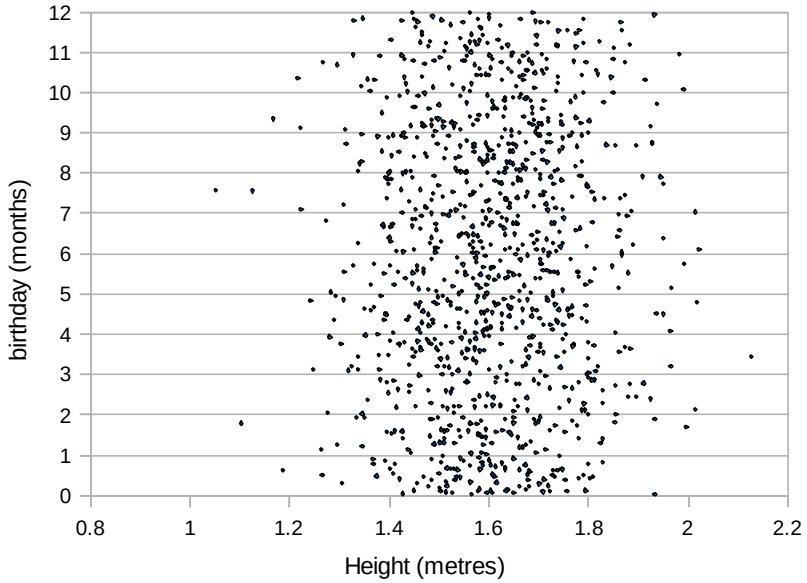
- We can visualise paired data values using a scatter plot.
  - A scatter plot gives us a 2-dimensional picture of the combined relative frequencies. Denser regions, where values are clumped together, are combinations of values which occur often.
  - Denser regions can also occur if one of the quantities has a higher relative frequency in that region, even if the data pairs are independent. In this case, the clustering will be along a horizontal and/or vertical direction.

Scatter Plots - 2

Correlated



Uncorrelated



Correlation - 1

- Correlated values
  - If the distribution of values of one quantity changes shape (for example, has a different mean) according to the values of the other, then the quantities are correlated.
  - In this case, if we know the value of one quantity, we have some partial information about the value of the other.
  - In the scatter plot shown previously, if you know that someone is very tall, then you can guess that they probably weigh a lot.

Correlation - 2

- Uncorrelated quantities
  - If the distribution of values of one quantity is the same for all values of the other, there is no statistical connection between them and we say they are “uncorrelated”.
  - In this case, if we know the value of one quantity, that gives us no information about the value of the other.
  - In the scatter plot shown previously, although some heights are more likely than others, knowing in which month someone was born does not help you to guess their height.

Lecture 2

Bringing British Education to You

www.nccedu.com

Using Statistics

- Statistical analysis of quantitative data is a powerful tool for summarising information.
- However, a summary of nonsense is still nonsense.
- Expressing information using numbers does not mean it is always true or meaningful.

Size of Dataset - 1

- Describing the typical value of a dataset by giving the mean, without giving the error, is meaningless

- Similarly, describing a proportion without giving the size of the dataset, is meaningless

Size of Dataset - 2

- For example:
  - 100% of people who replied to a survey have met aliens, (but only two people replied to the survey.)
  - During three years of testing of a new medicine in the United States and Japan, on average it helped reduce the patients' health problems and only 2 people had serious side effects,
  - (but there were only 5 people in the study so it only helped 3 patients.)

Choosing the Results - 1

- You can get the answer you want, by asking the right question
- You can prove almost anything, by choosing your data after the measurement

Choosing the Results - 2

- For example:
  - A washing powder manufacturer gave free samples to people and later asked them if they were happy with the results. The company was able to correctly claim that 90% of the households said their powder was good; But possibly 90% would be equally happy with another washing powder.
  - A common error for many researchers it to use only 'typical' measurement results. In fact, any random quantity will sometimes behave in a very non-typical way. Choosing which results you think should be the correct ones reduces the value of the data in the same way as the first example

No Reference to Source - 1

- To evaluate a result, we need to know how it was obtained.
- The data source is part of the description of the data.

No Reference to Source - 2

- For example:
  - This is even more important if there is any connection between the source and the quantities being analysed.
  - Mr X is a rich man who runs a large company, but he is also very kind and generous, according to 97% of the people who know him; (97% of the people who know him work for him and don't want to lose their jobs.)

Absence of a Control Set - 1

- Often, measurements can be affected by changing conditions.

- In that case, the significance of a change must be expressed with reference to the control dataset.

Absence of a Control Set - 2

- For example:
  - Boiled plastic was fed to 300 people who were ill with a cold, and 295 of them returned to full health within three weeks; (But if we had used a control group of people who had a cold and were instead fed boiled potatoes, we might have found that most of them had also recovered within three weeks.)
  - The control set technique is a powerful and underused way of cancelling out irrelevant factors without even knowing what they are.

Correlation Misuse - 1

- Correlation is one of the most misused statistical measures.

- Two quantities may be correlated because one of them depends on the other; or because they both depend on something else.

Correlation Misuse - 2

- For example:
  - If the population of a country is increasing, then any quantities which apply to some possibly small part of the population, will be correlated, even if they do not apply to the same people. Thus the number of televisions sold per year could be correlated with the rise in the number of blind people, but this does not mean that television makes you go blind.
  - Because causal relationships (when one thing makes another thing happen) are always correlated, it is common, but wrong, to believe that all correlations are causal.

Charting Results

- Data can be presented in a number of pictorial representations
- The appropriate representation should be chosen
- We will consider
  - Histograms
  - Bar Charts
  - Pie Charts
  - Time Series

Histogram

- Presents frequencies of items in bins
- Consider the following data, the maximum temperature for 30 consecutive days in degrees centigrade

19, 18, 17, 25, 26, 26, 27, 26, 30, 32, 25, 14, 15, 15, 14, 17, 19, 13, 25, 25, 26, 30, 24, 23, 21, 21, 19, 19, 21, 15

- Suppose we wish to show range of temperatures

Histogram

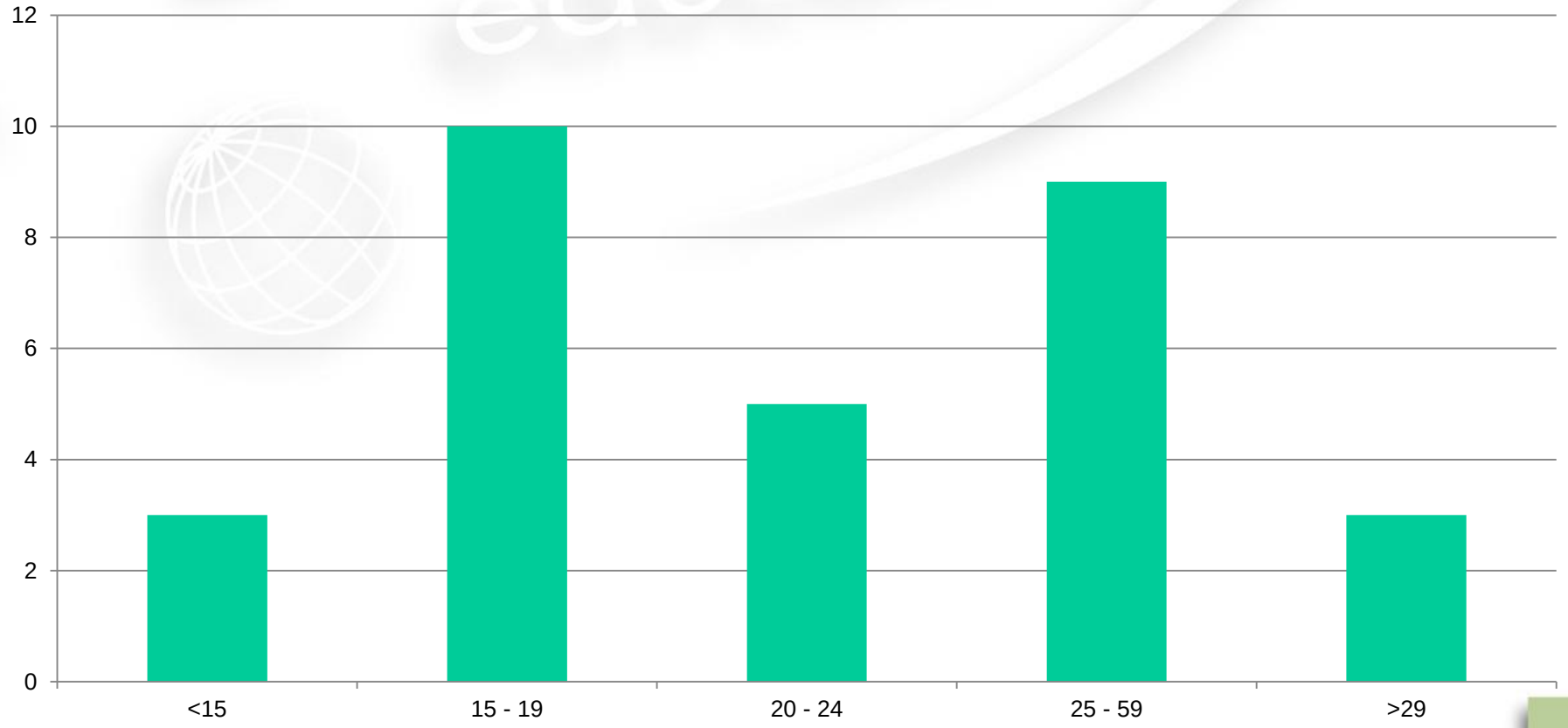
- We can create ***bins*** or ***classes*** to hold data and plot it
- Suppose we divide the range into 5 degree bands starting at 15 degrees and ending at 30
- We have a frequency table as shown on next slide

Histrogram - Table

Bin	Frequency
<15	3
15-19	10
20-24	5
25-29	9
>29	3

Now plot as shown on the next slide

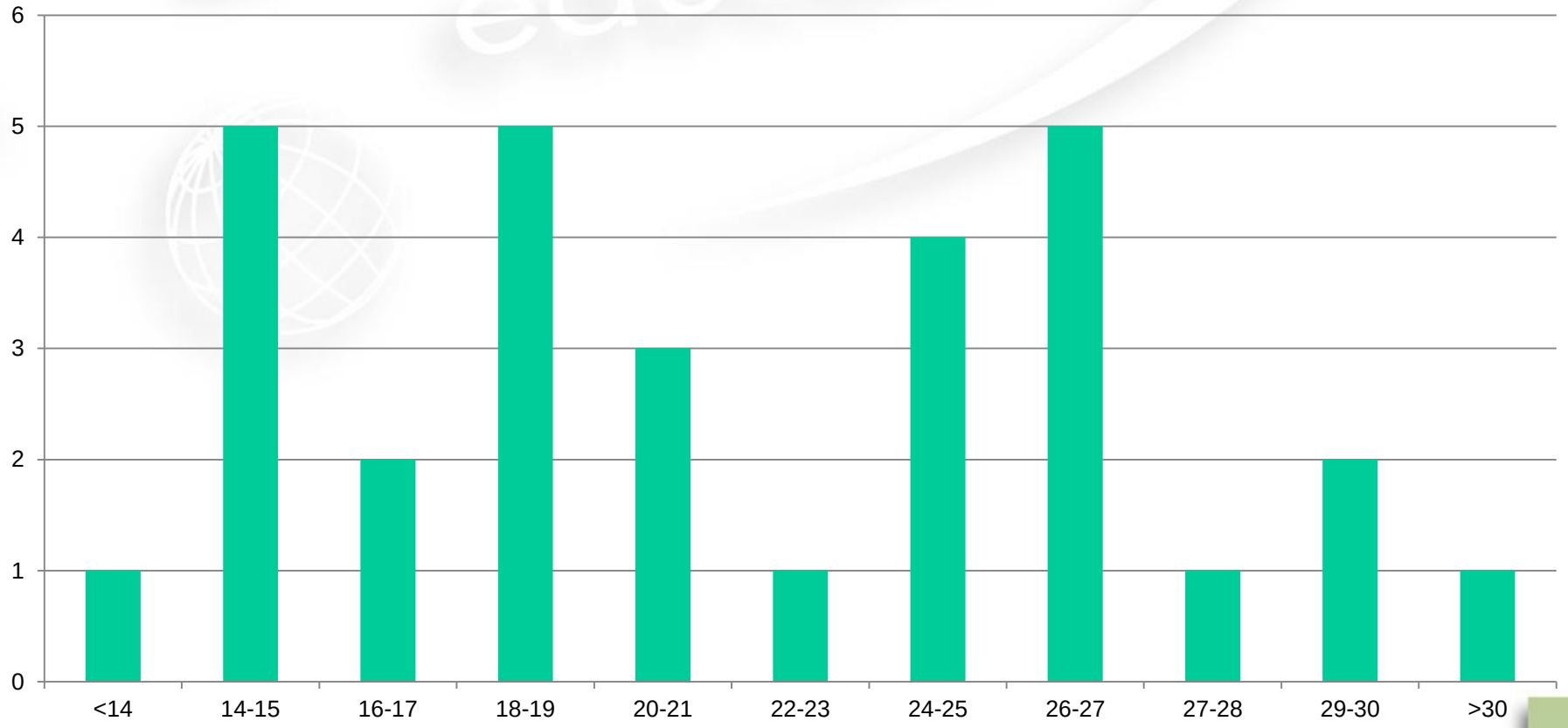
Histogram



Histogram

- Histograms are relatively easy to construct
- The size and location of the bins is key to a readable chart
- Consider the next slide which has more bins

Histogram – bins of size 2



Bar Charts and Pie Charts

- Used for categorical data (i.e. data placed in categories)
- Generally used to show relative frequencies

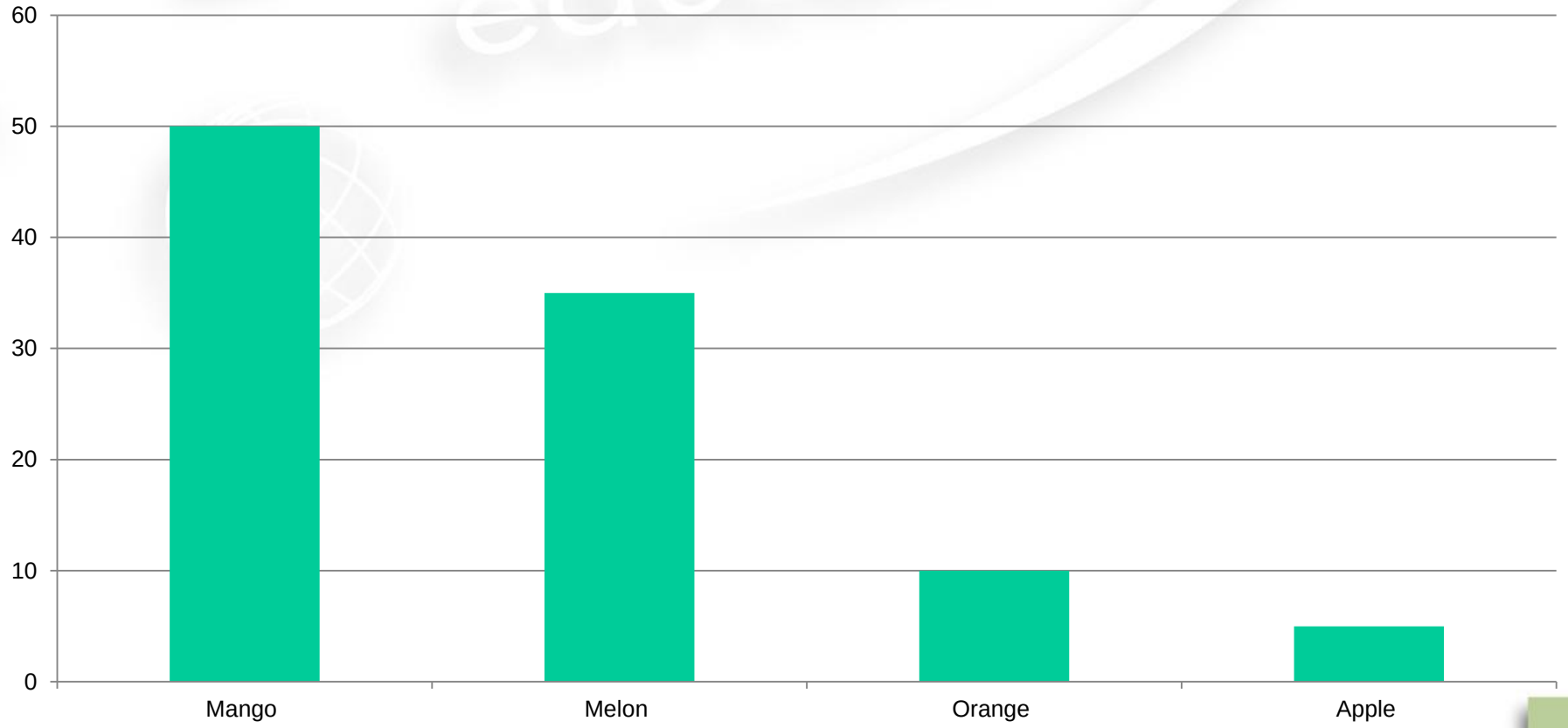
Bar Charts and Pie Charts

- A class of 40 students is asked to pick their favourite fruit from the list of 4 given below
  - Apple, Orange, Melon, Mango
- In the class
  - 20 chose Mango
  - 14 chose Melon
  - 4 chose Orange
  - 2 chose Apple

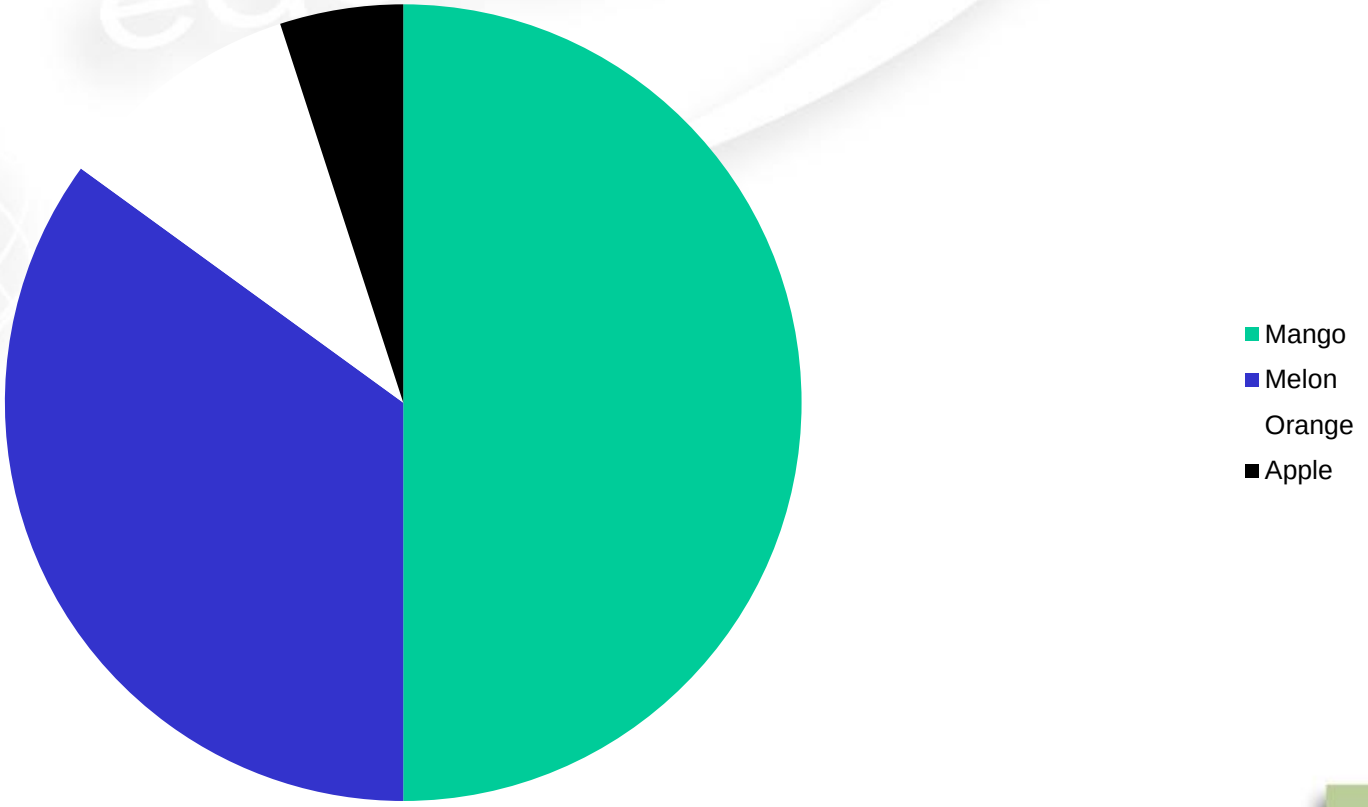
Bar Charts and Pie Charts

Category	Number	Relative Frequency (%)
Mango	20	50
Melon	14	35
Orange	4	10
Apple	2	5

Bar Chart



Pie Chart



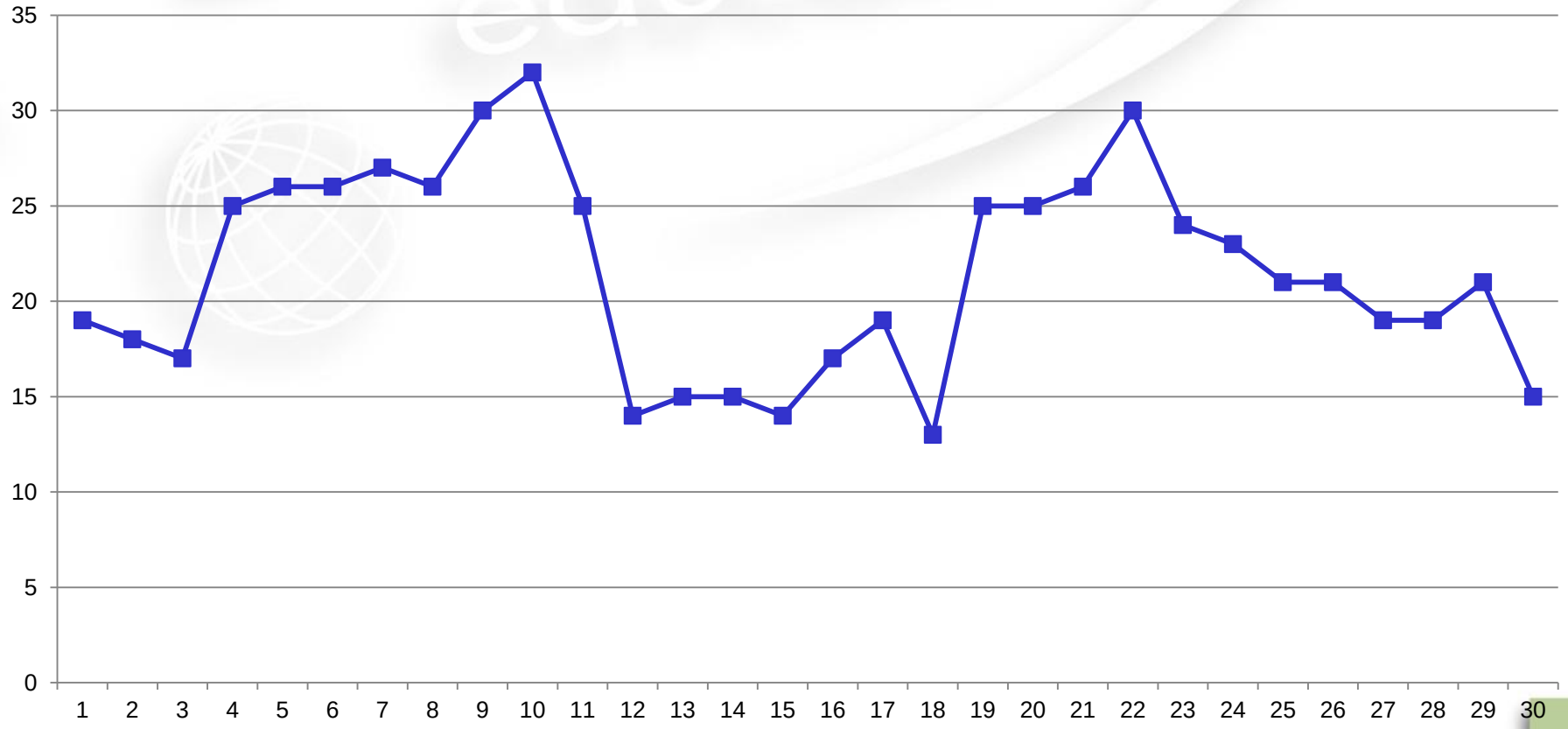
Time Series

- Show changing value with time
- Consider the variation of temperature given earlier

19, 18, 17, 25, 26, 26, 27, 26, 30, 32, 25, 14, 15, 15, 14, 17, 19, 13, 25, 25, 26, 30, 24, 23, 21, 21, 19, 19, 21, 15

- This can be plotted as a time series
- Assume that this is for days in June

Time Series Plot



Topic 10 – Presenting Results

Any questions?