Bringing British Education to You

www.nccedu.com

Skills for Computing

Topic 8:

Data Acquisition

Learning Outcomes for this Topic

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

- Understand how data depends on the source
- Understand how to store data
- Understand different types of data

Data & Information

- ***Data*** is not the same as ***information.***
  - When we get values from measurements, we call this “data”.

- Why do we need data analysis?
  - We use data analysis to get as much information as possible from the data.
  - We organise this information so that we can use it easily.

Filtering Data

- An important part of data analysis is separating out the part of the data that we don't need.
- Sometimes we can decide which parts of the data to exclude by finding which values are extremely unusual. We will return to this later when we look at describing how often different values occur.
- One way to refer to the desirable and undesirable components of the data is as ***signal*** and ***noise*** (more on this in Topic 9).

Measurements

- Subjective values
  - We obtain data directly through our senses; but we do not always agree on what we see or hear.
  - We have direct access to only our own senses, so we do not use the same data.

- A scientific measurement is a way of obtaining values which does not depend on ***who*** is making the measurement

Surveys

- Sometimes what we want is not easily described by a number.
- For example: peoples' opinions about a product or service. We can ask their opinions in a list of questions called a ***survey***.
- A survey is often designed so that there is a small number of possible answers to each question (for example, 'yes' or 'no').
- We then obtain numbers which represent how many times each possible answer was chosen, and we can apply ***data analysis*** to these numbers.

Data Sources: Primary Sources

- If you have direct access to the original data, it is a ***primary source***. This is true even if the data was not originally obtained by you or your organisation.
- When you analyse data from a primary source, other peoples' interpretations do not affect your results.
- Example:
  - To find out how much a new mobile telephone costs, you could write down the prices in a shop, or a friend could do it for you.

Data Sources – Secondary Sources

- Sometimes you will further analyse data that has already been processed. It may be useful to use a ***secondary source***, to avoid repeating the same procedures.
- A description of the data includes a reference to the source.
- Example:
  - If your friend does not write down the prices in a shop, but gives you an estimate of a typical price, your friend is a secondary source. A review article in a magazine is also a secondary source.

Generating Data

- ***Simulated data*** is data that has been generated automatically, usually with a computer.
- We may want to see how well our theoretical description works, by comparing simulated data with ***experimental data***.
- We may use simulated data to test or explore a system. For example, we could generate data that simulates visits to a website, to see if the website is working before it is available online.
- Simulated data is also used for training, for example by pilots.

Data Integrity - 1

- Data storage enables data re-use
- Advantages of digital data storage:
  - The data can be processed directly.
  - The data can be compressed and backed up easily.
  - Backups (copies) of data prevent data loss.
  - Backups can be used to maintain different versions of data files.
  - Ideally, backups are maintained on different storage media in different locations.

Data Integrity - 2

- Data files may contain ***errors***.
  - Errors may be introduced by the user (for example by using the wrong file) or electronically (faulty copying).
  - Some types of storage include error-correcting code, which detects and fixes introduced errors. Examples include audio CDs and online peer-to-peer file exchange protocols.
  - Errors from copying may be reduced by making copies of the original file, not copies of copies.

Data Representation - 1

- Data may be stored in many ways, called ***formats***.
  - All datasets from all sources should be converted to the same format.
  - Simple formats are more widely compatible.
  - The CSV (comma-separated values) format is compatible with a range of software including common spreadsheets. Numbers are in plain text and separated by commas.

Data Representation - 2

- Data files start with a data description, called a ***header***.
  - The header gives additional information about the data values, for example the type of data, unit of measurement, number of values.
  - All the data that is combined in an analysis must use the same units of measurement, for example length in metres or currency in US dollars.

Quantitative Data

- ***Quantitative data*** consists of numbers.
  - A complete description of the dataset must include the type of quantity measured, including units if applicable.
  - The advantage of a quantitative description is that numbers are not vague nor a matter of personal opinion.
  - The same results can be reproduced by anyone with access to the primary source.

Qualitative Data

- ***Qualitative data*** consists of types, not values of the data
  - We describe the data by organising, rather than calculating.
  - Examples include nationality, address, and other information which is usually organised in a database.
  - Qualitative data can be misused when the categories are vague or a matter of personal opinion.
  - Sometimes qualitative data may be processed quantitatively, by using the relative frequency of each type of value.

Discrete Data

- ***Discrete data*** only takes specific values.
- Usually there is a small number of different possible values.
  - For example, the number of children in a family.
- We describe discrete data using the relative frequency of each value.
  - We will return to relative frequencies in Topic 9.

Continuous Data

- Continuous data takes any value, within a range.
  - For example, the amount of time you sleep each day in a week can be any value between 0 and 24 hours.
- Continuous data is anything measurable.
  - This includes all quantities, not only relative frequencies.
- Continuous data can be converted to discrete data by dividing it into categories. For example: the number of times every week that you sleep less than 8 hours in a day.

Comparing Data

- Measuring changes in behaviour:
  - We look for changes in data by comparing values to a reference dataset, called a ***control***.
  - If the control represents typical or desirable values, we call it ***baseline data***.
  - This is useful in factories, where we want to make sure that the behaviour of the equipment does not vary too much from the normal behaviour.

Topic 8 – Data Acquisition

Any questions?

Bringing British Education to You

www.nccedu.com