
FMSP
Let maths
take you further

Large Data Sets: Why?

Data are now being collected on a scale that was unimaginable even a few years ago. This is true not only in our personal lives but also in the workplace and in higher education; in academic disciplines where statistics was once peripheral it is now becoming central. It is critically important that we equip our young people with the skills they will need to live and work in this data-rich world.

A world full of data: Statistics opportunities across A-level subjects
Roger Porkess and Stella Dudzic, RSS 2013.

Purpose of this session

- Understand the types of large data sets that students might be given
- Understand how they might be expected to analyse it
- What challenges will this present for teachers and students?

What is a large data set?

‘Big’ data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.

A minimum size for a ‘large’ data set for AS and A level Mathematics should normally be 8-10 fields and data on at least 100 items for those fields.

Desirable characteristics

- The data set should consist of real data and, where possible, the source should be given (including URLs) so that students can find out how the data set was collected.
- Data may be reorganised into a standard format but the data should not be cleaned.
- It should contain a mixture of categorical and numerical data.
- Errors should not be deliberately introduced into the data set but may be present. Consideration of outliers might also be required.

New Specifications

- Become familiar with one or more specific large data set(s) in advance of the final assessment
- Use technology such as spreadsheets or specialist statistical packages to explore the data set(s)
- Interpret real data presented in summary or graphical form
- Use data to investigate questions arising in real contexts

New Specifications

- Students should explore the data set(s), and associated contexts, during their course of study to enable them to perform tasks that assume familiarity with the contexts, the main features of the data and the ways in which technology can help explore the data.
- Students should demonstrate the ability to analyse a subset or features of the data using a calculator with standard statistical functions.

Desirable characteristics

- One of the purposes of the data set is to ensure that students are familiar with the terminology and contexts relating to the data so that students can think about the data in advance and so be ready to engage in realistic interpretation.
- A short piece of text and/or a glossary may accompany the data set to help students understand the source of the data and associated terminology.

Desirable characteristics: Sample?

- It needs to be clear to learners whether the whole data set is (essentially) a population or whether it is a sample from a larger population.
- Working with a sample and making an inference about a population is an important feature of statistical work – if the data set is a whole population then samples associated with the data set might be introduced in examination questions.

Desirable characteristics: Models

- A data set for A level could include some data where the Normal distribution could be considered as a model.
- A data set for AS level should support questions related to the binomial – these could be based on proportions of the data that have some characteristic.

In the exam:

- Students should be expected to interpret output from a spreadsheet or statistical software.
- Students should be expected to interpret and explain terminology which has been introduced via the data set.
- Students should be asked to explain what effect missing data would have on a model that has been derived.

In the exam:

- Students should be asked to explain how they would collect data and to describe the drawbacks and advantages of particular sampling methods. They should understand the kinds of methods that were used to collect the data set(s) they have been working with.
- Students might be presented with a reasonably small selection of data within a question and be required to process these data, for example to produce and interpret summary statistics.

Types of Questions

- Short questions requiring brief comments/interpretation.
- Questions requiring deep interpretation of the data, using given graphs and summaries.
- Questions which require students to select from given graphs and summary data to solve a statistical problem.
- Modelling with trend lines for bivariate data (can include quadratic or exponential trend lines where appropriate).

Types of Questions

- Modelling with distributions.
- Hypothesis testing
- Describing a situation where data needed to be collected and how it might be done.

Sources of Data

- http://www.ons.gov.uk
- https://www.gov.uk/government/collections/maritime-and-shipping-statistics
- http://www.cdc.gov/nchs/nhanes/about_nhanes.htm
- http://www.gapminder.org/data/
- http://www.guardian.co.uk/data
- http://www.tsm-resources.com/useful-files.html
- http://www.coventry.ac.uk/ec/~styrrell/
- http://www.censusatschool.org.uk/
- http://mei.org.uk/data-sets

A data Set: Blackbird Data

- Open the excel file Blackbirds and the information PDF
- If you were using this with students for the first time, what sort of questions or issues would arise?

A data Set: Blackbird Data

- Look at the ideas for investigation at the end of the information sheet and try one of these!
- Or try using pivot tables and the ideas that follow that..
- Or make up your own

Be prepared to feedback your findings! Work in pairs.