Data Analysis

The aim of a data analysis can be

- descriptive

  presenting and summarising the data to get a feel for it.

- inferential

  estimate parameters and test hypotheses.

- predictive

  once we have estimated parameters, we can make predictions about future data.

Descriptive Statistics

It is difficult to draw meaningful conclusions from data in its raw state. Therefore we use plots and summary statistics. Summary statistics turn data into numbers. Important summary statistics are measures of central tendency (mean, mode, median). and measures of dispersion (standard deviation, interquartile range).

Inferential Analysis

In our analysis, we can either use data that has already been collected or data we collect ourselves. If we use previously collected data, then we should find out how it was collected.

If we collect our own data, then we should be careful to take a representative sample from the whole population

If the sample is chosen without prejudice and is large enough, we may assume that the statics of the sample represent the statistics of the whole population

Example

An opinion poll is conducted from 14:00 to 16:00 in London, why may it not reflect the views of all of ~~England?~~ England?

- London doesn't reflect all of England
- working people are not included.

Predictive Analysis

We make predictions about future events. We have a training set of data to find relations between attributes of the data and use the relationships to predict future behaviour.

Example (Linear Regression)

training set (X_i, y_i) $i=1,\ldots,n$

from which we infer a trend.

$$y = mx + b.$$

For example, if r is rainfall in a day
s is hours of Sunshine

Suppose $s = 9 - 0.1r$

If $r=2$, $s = 9 - (0.1)(2) = 8.8$

The Data Analysis Process

The following are nine steps an actuary could go through in performing a data analysis.

1. Develop a well-defined set of objectives to met by the data analysis.

   e.g.) summarise claims from medical insurance by age, gender, ...

   (descriptive)

   Predict the outcome of the next parliamentary election

   (predictive)

2. Identify the data needed for the analysis.

3. Collect the data. It may be availabe internally or gathered from an external Source (i.e. from a government office)

4. Processing and formatting the data for analysis (i.e. inputting the data into a spreadsheet or data base).

5. Cleaning the data (e.g. addressing unusual, missing or incomplete value)

6. Exploratory Data Analysis (Descriptive, inferential, predictive)

7. Make a model of the data.

8. Communicate the results

(e.g. ~~whata~~ what data was used
what analysis was performed
any assumptions made
the conclusions
any limitations of the analysis)

9. monitor the process
update the data
repeat the process as required.

Big Data

"Big data" refers to large data sets which can't be analysed using traditional methods.

Big Data has the following properties:

- large size
- speedy collection
- comes from many sources
- reliability may be hard to ascertain.

Examples

customer data, data from scientific sensors.

Reproducible Research

Reproducibility is when a statistical analysis is recorded and sufficient information is provided so that another party can repeat the analysis and ~~the same results.~~ get the same results.

For reproducibility, we need

- documentation of what we do
- including an audit of any decisions made when cleaning the data
- document the code, e.g. use lots of comments.

Doing things by hand creates problems with reproducibility.

Sam

Some things done by hand you should not do include

- editing tables and figures manually
- downloading data manually without documenting what you do.
- Pointing and clicking on a screen
- getting tables from others,