Lecture 1

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: TY BSc

Subject : Model Documentation Analysis and Reporting

Subject Code: PUSASQF503

Chapter: Unit 1 Chapter 1

Chapter Name: Data Analysis

Today's Agenda

1. Data Analysis
    1. The given Data
    2. Raw data and Clean data
    3. Cleaning the data
    4. Identify data errors
    5. Fix the data errors
    6. Summarizing data
    7. Extracting information

1.1 The given Data

A typical assessment project will present you with a data set to work with.

What do you think – Can we use the given data as it is always or do we need to make some pre-processing of the data?

1.1 The given Data

- The data set you are given will not be in the precise form you require.
- For example, you might be provided with a set of values of the FTSE 100 index, which you must first convert into rates of return. As a result, some pre-processing of the data may be required.
- Also, the data you are given might contain errors. If these are not dealt with, the results of the analysis will be meaningless. So it is important to validate or 'clean' the data, ie identify and deal with any errors.
- You may also need to summarise the data set, ie calculate some key statistics that describe the 'shape' of the data.
- Note that you may be asked to create your own data! This is not as silly as it might sound – you may be asked to use a particular model to simulate a set of results, which will then form your data set.

1.2 Raw Data & Clean Data

It is good practice to keep the original data set intact and work from a 'copy' within Excel. This means that if a different, or corrected, data set is used later, it will be possible to compare the two data sets.

Ideally your spreadsheet should show separately the original 'raw' data with any warning messages from the Excel checks that were applied and the modified 'clean' data with the corresponding Excel checks now saying 'OK'.

1.3 Cleaning the Data

What do you think – What type of errors would the data contain?

1.3 Cleaning the Data

Types of data errors

Errors in numerical data in computer files usually consist of:

- wrong numbers
- outliers
- omissions or duplicates

1.3 Cleaning the Data

Types of data errors

Wrong numbers

- Wrong numbers can occur because of incorrect inputting. Particularly common are:
- omitted digits, eg 21553 instead of 215553
- anagrams, eg 2456 instead of 2546
- mistakes involving repeated digits, eg 1223 instead of 1233.

Outliers

- 'Outliers' are extreme data values that don't appear to be consistent with the model – they don't fit the pattern.
- They may distort the results.

Omissions or Duplicates

The data could have missing entries for certain values.
Also, there could be certain values that are double counted or entered twice (two times for same person, same stock etc.)

1.4 How can we identify data errors?

Usually, it will be sufficient to:

- scan through the data by eye to spot any obvious problems (*eg* missing entries). However, this may not pick up all the errors, especially if the data set is large.
- calculate a few summary statistics, such as the number of data values and the maximum/minimum values
- apply some automated Excel checks, using Excel formulae that will highlight any errors
- apply some reasonableness checks to the summary statistics. The summary statistics should highlight any outliers, as these may fall outside the normal range of values. The summary calculations will also throw up an error if you have applied an Excel function to data that contains invalid characters – for example, a letter O instead of a zero.
- reconcile the summary statistics with any additional information you have been given in the project specification.
- Plotting a graph can be useful where we have a series of data values that we would expect to show a consistent progression. A graph would highlight any spikes or other irregularities that might be difficult to spot otherwise.

1.4 How can we identify data errors?

- Try to incorporate some automated checks on the data and include a description of this in the audit trail. For large data sets, automated checks are more reliable than reviewing by eye.
- Document all the data checks you apply and any remedial action you take (even if no remedial action is required) and give reasons for your approach.
- However, don't spend too long working on the data. It is important to move on to develop the rest of the model.

Question

Two of the columns of data provided for a valuation of the benefits for employees of a large company who are members of the company's pension scheme are:

- sex (with M for male or F for female)
- date of birth (in the format DD/MM/YYYY).

(i) List the checks that you could apply to the data values in these two columns to 'clean' the data.

Solution

If we are told how many employees there 'should' be, we can start by counting the numbers of M's and F's to check that these match the numbers of males and females in the pension scheme on that date.

For the 'sex' column we could:

- scan by eye for any missing entries or ones that are not M's or F's
- apply an automated Excel check to ensure that all the entries are either M or F
- use the filter feature in Excel to identify the different entries present in the column (which should only include M's and F's)
- count the number of entries in the column and check that this is consistent with the number of employees in the pension scheme
- check with the company how many employees there should be.

More advanced checks we could apply (if we had the necessary information) would include:

- compare the numbers of each sex (or the gender ratio) with the corresponding figures from the previous valuation
- use the employees' names or employee numbers to check that the entries are consistent on an individual basis with the previous valuation.

Solution

For the ‘date of birth’ column we could:

- scan by eye for any missing entries or obvious errors, *eg* years containing 5 digits
- apply an automated Excel check to ensure that all the entries are valid dates, *eg* no 30th February’s or month 13’s or unpopulated entries such as 00/01/1900 or DD/MM/YYYY
- calculate the minimum and maximum age to check that there are no outliers, *eg* employees aged 12 or 105.

More advanced checks we could apply (if we had the necessary information) would include:

- calculate the average age and check that this is consistent with the employee profile
- compare the average age with the corresponding figure from the previous valuation
- use the employees’ names or employee numbers to check that the entries are consistent on an individual basis
- plot a graph of the number of employees born in each year (or each month) to look for any irregularities

1.5 How do we fix data errors?

- If you spot a data error that you think would significantly affect the results, you should modify the data as you think best, and document clearly in your audit trail what you have done and why.
- Try to set up your spreadsheet so that any changes made to the data at a later stage (possibly by someone else) will automatically be reflected in the subsequent calculations.
- Where possible avoid copying and pasting values since this means that changes to the data will not be reflected later in the calculations. Use cell references instead.
- If, for some reason, you cannot avoid pasting values, document very clearly what you have done so that someone else would be able to replicate your work

1.5 Summarizing a data set

The purpose of summarising a data set is to get an idea of the distribution of the values – the
'shape' of the data. This normally involves finding:

- the number of data values
- the highest and lowest values
- sample moments, such as the mean and standard deviation.

1.5 Extracting the relevant information from a data set

In some cases you may need to convert the data from its original form. For example, you might need to convert a history of market values of an asset into rates of return or you might need to convert dates of birth into ages before proceeding.

This conversion should be done after you have sorted out any problems with the original data.

Summary

- Make sure you understand the data you've been given.
- 'Clean' the data by identifying any obvious errors.
- Include some automated checks.
- Calculate summary statistics, such as totals and averages.
- Apply reasonableness checks to identify any outliers.
- Reconcile the summary statistics with any additional information given.
- Consider plotting a graph to highlight errors when checking a series of data values.
- 'Prepare' the data *eg* calculate any derived quantities and/or subdivide the data.
- Be prepared to create your own data set for a simulation.
- Document all the data checks you apply, even if no remedial action is required.
- Don't spend too long working on the data, especially if there don't appear to be any problems.
- Not all data sets will need all the steps outlined above. You will need to demonstrate that you can apply the appropriate steps.

Question

A colleague has mentioned that marks are awarded in Paper 1 of the CP2 exam for 'auto-checks', *ie* formulae in the spreadsheet that check the values in particular cells. One purpose of an autocheck is to check whether a value that has been entered by the user is an acceptable value for that cell, *eg* whether it is a valid date that falls within a permitted time period.

(i) State two other purposes that auto-checks can be used for and give an example of each. [2]

(ii) List four other types of check (other than auto-checks) that can be used to identify possible errors in the data in a spreadsheet. [2]

(iii) (a) Explain what is meant by a reasonableness check (also known as a 'sense check').

(b) Give an example of a reasonableness check that might be used in an actuarial context.

Solution

(i) ***Auto-checks***

Auto-checks can be also be used to check for:

- consistency, *eg* checking that the totals in a summary table are consistent with the totals in the data they are derived from or that a set of probabilities adds up to 1 [1]
- reasonableness, *eg* checking that a calculated value is similar to a rough estimate or lies within the range expected, *eg* probabilities lying in the range [0,1] .

(ii) ***Other types of checks***

Other types of checks that can be used include:

- checking for obvious errors 'by eye'
- calculating summary statistics (*eg* averages, minimum/maximum)
- using Excel's built in features such as auto-filter
- comparing against independent information or background information provided
- spot checks, *ie* calculating a few figures manually
- reasonableness checks, *eg* comparing a figure with a rough estimate.

Solution

(iii)(a) ***Reasonableness checks***

As the name suggests, a reasonableness check is intended to check whether a particular figure is believable, *ie* that it is not obviously wrong. This could involve:

- comparing it with a rough estimate
- checking that it lies in the range that would be expected.

(iii)(b) ***Example of a reasonableness check***

Examples in an actuarial context might include:

- a pension fund might check the total annual payroll for a large company by multiplying the number of employees by a published average wage for companies in that sector
- a non-life insurer might check the average claim size for a particular class of insurance by comparing it with the corresponding figure from the previous year, adjusted for inflation
- an investment manager might check the latest value of its investments by comparing the proportions held in each asset class with the proportions from the previous valuation.