1

Analysing data

Syllabus objectives

1 Preparation and analysis of data

- 1.1 Use appropriate tools for cleaning, restructuring and transforming data to make it suitable for analysis.
- 1.2 Summarise data using appropriate analysis, descriptive statistics and graphical representation.
- 1.3 Select and carry out appropriate statistical tests of reasonableness.
- 1.4 Make appropriate assumptions about the data provided.
- 1.5 Repair corrupt or missing data.

0 Introduction

In this chapter we will cover the following topics:

- data in Subject CP2
- ‘cleaning’ the data
- summarising a data set
- extracting the relevant information from a data set.

1 Preparing the data

1.1 Data in Subject CP2

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

In some cases, however, 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.

Make sure you understand the data you've been given. If you don't take the time to do this, there is a danger that you will set off on the wrong track.

1.2 Cleaning the data

Types of data errors

Errors in numerical data in computer files usually consist of:

- wrong numbers
- outliers
- omissions or duplicates.

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' are extreme data values that don't appear to be consistent with the model – they don't fit the pattern. In real life, important data sets are sometimes input twice independently and then compared by computer. This will identify any inputting problems involving wrong numbers, omissions and duplicates, but will not pick up outliers.

Question

The data set provided for the worked example we will look at later consists of a list of the claim amounts in a portfolio of insurance claims.

Suppose that, in subsequent discussions with the client, you discover that the original data consisted of a set of paper-based lists taken from *international* records, which were then input by hand onto a spreadsheet and 'cleaned'.

What types of types of data errors do you think might have been present in this data set before it was cleaned?

Solution

There may have been some wrong numbers because some of the values may have been input incorrectly, *eg* omitted digits, anagrams or mistakes with repeated digits.

An inexperienced typist may have accidentally pressed a non-numeric key or included some letter O's instead of zeros.

Claims collated from different countries may be presented in different formats, which may result in errors. For example, in Europe different 'separators' are used, so that the number 12,345.67 would be written as 12.345,67.

Some claims may have been omitted completely or double counted.

Outliers would be present if the claims were given in a mixture of different currencies (*eg* £'s versus $'s) or different currency units (*eg* pounds versus pence).

The marks shown in the exam question will give an indication of how much data checking (if any) is expected. If the data set has not been cleaned and could contain errors, you will need to check the data and make corrections as necessary, then document the checks you have applied and any corrections you have made.

How can we identify data errors?

In the exam, you will not be able to look for the more subtle types of data error or to go back to the original documents to check things out. Usually, it will be sufficient to:

- scan through the data by eye to spot any obvious problems (*eg* missing entries)
- calculate a few summary statistics, such as the number of data values and the maximum/minimum values
- apply some automated Excel checks (see below)
- apply some reasonableness checks to the summary statistics
- reconcile the summary statistics with any additional information you have been given in the project specification.

As an initial check, scanning through the data by eye will usually allow you to spot problems such as missing values or repeated values. However, this may not pick up all the errors, especially if the data set is large. So you should also apply some automated checks using Excel formulae that will highlight any errors.

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.

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.

Another method of checking the data that can sometimes be useful is to plot a graph. Before you create the chart, consider what you are expecting it to show. This will help you explain why the results of the graph are reasonable, or not. This explanation should be included in your audit trail as a reasonableness check on the data.

Question

Can you think of a situation where this might be an effective way of identifying data errors?

Solution

Plotting a graph can be useful where we have a series of data values that we would expect to show a consistent progression – for example, a table of mortality rates plotted against age or the values of a financial index plotted against time. A graph would highlight any spikes or other irregularities that might be difficult to spot otherwise.

'Raw' data and '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'.

You should link the data in the 'clean' data worksheet to the 'raw' data worksheet so that if the original data changes, the 'clean' data will be updated. Only cells where an amendment to the value has been required following the data checks should be hard-coded. These cells should be clearly highlighted.

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. Make it clear in 'clean' data worksheet what data has changed. This can be achieved by shading the amended cells a different colour.

We will see an example of this in the Worked Example in Chapter 9, where we have to supply some missing data values that were not provided.

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.

1.3 Validating a data set

The purpose of validating a data set is to check that the distribution of the values is consistent with the statistical distribution that it is assumed to follow. This can involve checking:

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

It may also be necessary to carry out a chi-square test in order to check the goodness of fit of a particular distribution to a data set.

Question

Explain how you would carry out a chi-square test to check whether there is a statistical difference between a given data set and a given distribution.

Solution

The chi-square test aims to test the hypothesis that the data is from the given distribution. This is the null hypothesis.

1. Split the actual (observed) data into groups and determine the number of entries in each group.

2. Calculate the number of data entries that would be expected in each group if the data is a random sample from the given distribution.

3. Calculate the following for each group:

$$\frac{(O-E)^2}{E}$$

4. Sum the values calculated in stage 3 for all groups. This is the observed value of the test statistic.

5. Calculate or look up the critical value from the χ^2 distribution with degrees of freedom equal to the number of groups minus 1 at the given significance level *eg* 5%.

6. If the observed value of the test statistic is greater than the critical value, there is sufficient evidence to reject the null hypothesis at the given significance level and it is reasonable to conclude that the data does not come from the given distribution.

7. If the observed value of the test statistic is smaller than the critical value, there is insufficient evidence to reject the null hypothesis and therefore no significant evidence to suggest that the data does not come from the given distribution.

In Excel the function CHISQ.TEST can be used to carry out a chi-square test. This function takes the actual and expected frequencies as inputs and returns the *p*-value of the test, *ie* the probability that the observed and expected data are from the same distribution. For example, a result of 0.97 means there is a high probability that the data come from the same distribution. A lower value of 0.003 means there is a much lower probability that the data come from the same distribution.

If you wanted to perform the test at the 5% level, the *p*-value returned by the function would be compared to 5%. If the *p*-value is greater than 0.05, we accept the null hypothesis that the observed and expected data are from the same distribution. If it is less than 0.05, it is very unlikely the data are from the same distribution, and we reject the null hypothesis.

The CHISQ.TEST function assumes that the number of degrees of freedom is always equal to the number of groups minus one. Most of the time this will be appropriate, however in the unusual circumstance that a different number of degrees of freedom is required you can obtain the *p*-value using the CHISQ.DIST.RT function, which takes the test statistic and the required number of degrees of freedom as inputs.

We will look at the Excel functions that can be used for validating data in Chapter 6.

1.4 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.

Chapter 1 Summary

Key points regarding data analysis

- 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.
- If required, check the assumptions that the data comes from a particular statistical distribution.
- 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.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 1 Practice Questions

1.1 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:

- marital status (with M for married, W for widowed or S for single)
- 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.

The valuation involves applying a set of age-related mortality, retirement and withdrawal rates to each employee. These rates have been provided in a table in an Excel file.

(ii) List the checks that you could apply to ensure that these rates are correct.

1.2 Give three reasons why it is important to check the data used in an actuarial project to ensure that it is complete and accurate.

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 1 Solutions

1.1 (i) *Checks on the data*

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

For the 'marital status' column we could:

- scan by eye for any missing entries or ones that are not M's, W's or S's
- apply an automated Excel check to ensure that all the entries are either M, W or S
- use the filter feature in Excel to identify the different entries present in the column (which should only include M's, W's and S'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 marital status (or the ratios) with the corresponding figures from the previous valuation
- spot check some entries to check if there have been changes since the previous valuation, if there are confirm the record of the request for the change to be made.

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 Februarys or month 13s 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
- use the filter feature in Excel to quickly identify any invalid values or values not in a date format.

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.

(ii) *Checks on the decrement rates*

For the decrement rates we could:

- plot graphs of the three rates against age to look for any irregularities
- apply an automated Excel check to ensure that the rates at consecutive ages are similar, *eg* they don't change by more than 10% from one age to the next
- calculate the minimum and maximum rates to check that there are no outliers, *eg* rates that are 10 times too big
- calculate the average rates over all ages and check that these are roughly what we would expect.

If we had the necessary information, we could compare the rates with the decrement rates we have used for other pension schemes.

1.2 *Reasons for checking the data*

Any material errors in the data will invalidate the results of our calculations.

Some types of errors might prevent the software we are using from completing the calculations, giving us an error message instead.

It is a professional requirement to take appropriate steps to make sure that there are no material errors or omissions in the data we are using.

Errors in our results caused by data errors could lead to customers being unfairly treated, which could cause reputational damage, or to legal action being taken against us.