Introduction to statistics

Preview

Introduction

In previous topics we have looked at ways of gathering data for research purposes and ways of organising and presenting it. In topics 11 and 12 we turn our attention to the art of turning raw quantitative data into useful information so as to extract from it answers to research questions.

The word 'statistics' sends shivers up the backs of many people—maybe a better title for the topics would be 'data analysis', just to stop them sounding so scary! And, anyway, 'data analysis' is what we are going to concentrate on in this topic. 'Statistics' usually implies all the underlying theory, mathematical formulas, looking up of tables, and so on, that we are going to downplay here. There'll be few formulas and messy technical details and we'll leave the number-crunching business to computers.

We start with a discussion of the scope of quantitative data analysis and how it fits into doing research. Then we discuss some common pitfalls and potential misuses of statistics. Next we present an overview of data analysis, covering the essential basic ideas. One of the problems in doing data analysis is that there are literally thousands of things you can do—you just have to look at the myriad charts and tables that appear in newspapers, journals, reports, and so on. So we will set the basic ideas into a framework that will guide you through the labyrinth. In the next topic we apply the basic ideas in a series of examples that together provide a look at most of the common types of analyses you may need.

Carr, R. 1999, *Statistics In A Day* (computer disk), XLent Works, Warrnambool.

Scope and limitations of data analysis

Data analysis is the analysis of quantitative data—data that arises from measurements of some kind. The measurements can come from many different sources (though we will see that there are only two important types of measurements)—daily records of sales or attendances, questionnaire data, financial records, government reports, and so on.

Quantitative data and its analysis obviously play an important role in many investigations. For research purposes, however, quantitative data is rarely used on its own. It is rare that we simply want to measure something and present the results. That might be how data appears in an annual report, but in research it is usually important to fit the work into some theoretical framework or model.

A model of a situation is a conceptual way of presenting answers to 'how' and 'why' questions. In sciences such as physics a model is often encapsulated in a 'law', such as Newton's 'for every action there is an equal and opposite reaction'. This law shows deep

understanding and can be applied to explain an extraordinary range of observations on everything from billiard balls to planets. In other sciences such as economics we still have quite definite models or laws such as 'people act to maximise their utility function'. These are perhaps not as strong as the laws in the physical sciences but are still essential. In management we have basic models explaining how organisations work that are essential for setting the framework. In general, if we didn't use models we would end up simply describing every single situation in isolation: we wouldn't be guided by what we find in other circumstances, we wouldn't be able to argue by analogy, we wouldn't be able to generalise. This is not a very satisfactory state of affairs—we need models.

How does data analysis fit in? Well, it is not especially good for 'model building'. Rather, it is good for refining or testing a model or for making predictions from it, but not for the actual building. For example, suppose we wish to determine how best to market a particular product. We'd need to know what factors are important in determining market share. We can't measure and analyse everything, so we'd have to be guided by experience (in other similar situations, maybe) or we might conduct focus groups or interviews comprising experts to elicit the important factors. We'd have to build a model first. The data we subsequently collect could be used to test or refine the model. This is very important. Data analysis may give us useful information about the important factors that will help in deciding precisely how to go about marketing the product. Maybe we'll find that some factors in the model are not important and can be dropped from consideration, for example. It might tell us to concentrate our efforts on one particular area that is most likely to lead to greater sales. Maybe it will tell us that the relationships predicted by the model are all wrong and that the model is entirely inappropriate. This is obviously important information! This same type of idea applies to the use of data analysis in research. Data analysis is good for model testing and refinement and for that reason is often an essential component of a research project.

Setting up problems for data analysis

Chapter 2 ('AM') of *Statistics In A Day* covers the essentials of data analysis and presents a simple six-step guide to doing it. Actually there are seven steps—the step we might call 'step zero' is not really part of data analysis, at least not part of the number-crunching normally associated with data analysis. But in some ways it is the most important step of all. It is the *formulation of the problem*.

Formulating the problem

Overall, step zero—'formulation'—is by far the hardest part of doing any quantitative research (or any research at all!). Statisticians find themselves spending about 90 per cent of their time helping people decide what they really want to know in the first place!

The ideas discussed in section 2.1 of *Statistics In A Day* have nearly all been discussed one way or another in other topics in this Study Guide, but some of the most important ideas should be mentioned again here.

Including important sources of variation

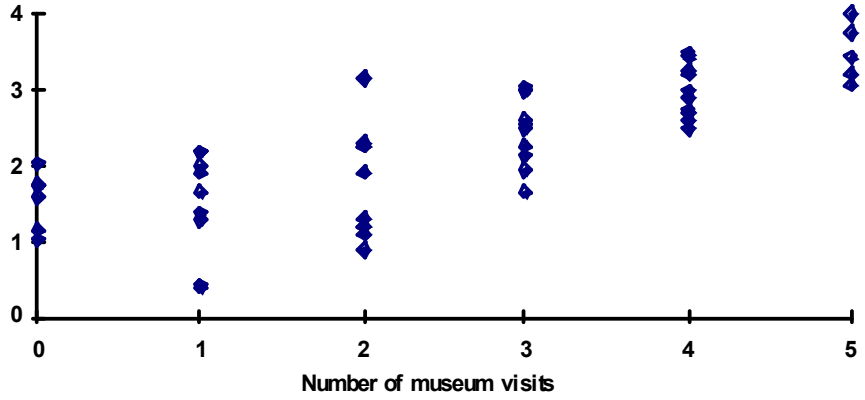
In section 2.1 of *Statistics In A Day*, point 1 mentions that data analysis is the analysis of variables. A variable is a quantity that we can measure (or a quantity derived from measurements). The number of visitors to a museum in a week, or the response to a particular

question ('Agree' or 'Disagree', for example) on a questionnaire are variables. Appropriate techniques for the subsequent analyses will be covered soon.

Point 2 is very important. In topic 2 of this Study Guide you encountered the notion of 'lying with statistics', and that 'statistics can be made to say whatever you want them to say'. Point 2 is about one great way of lying with or misusing statistics: if you ignore important factors (variables) in an investigation you can easily tell just about any story you like! Let me give you two examples:

EXAMPLE 11.1 Figure 11.1 is a plot showing two variables:
X = number of times a person visits an art museum in a year
and
Y = grade point average in their first year of university

FIGURE 11.1



There is an obvious trend. But it is incorrect to conclude that we should make high school students visit art museums because it will make them do better in university (that is, it is incorrect to claim that there is an effect on X on Y) because obviously there are other variables that cause both X and Y to vary—maybe 'interest in academic affairs', or even 'IQ', for example. This is an example of a very common misuse of statistics that can result from not including important sources of variation—you can 'show' results that are not real.

EXAMPLE 11.2 In table 11.1 is some data showing the total donations to museums of two types, E and A, over the last few years.

TABLE 11.1

E	91	56	3	27	63	88	24	4
A	74	5	93	35	67	7	29	95

There do not appear to be any major differences between the two types of museums. But wait on—there is another very important source of variation that I didn't tell you about. It's the size of the museum. If we include this extra variable it's now quite clear from table 11.2 that there *is* a difference between the two types of centres:

TABLE 11.2

		Size of museum							
		Tiny		Small		Medium		Large	
Type of museum	E	3	4	27	24	63	56	91	88
	A	5	7	35	29	67	74	95	93

This is an example of the other problem caused by not including important sources of variation—effects that are real can be easily hidden.

Keeping it simple

Point 3 (in section 2.1 of *Statistics In A Day*) is the KISS ('keep it simple, stupid!') principle, which for data analysis means that you must try to reduce the number of variables involved in a problem. You can rarely manage problems that involve more than a handful of variables if they are worked with all at once. This doesn't mean that you can't analyse data from a questionnaire with 100 questions (i.e. 100 variables), but it does mean that you can't work with the 100 variables all at once. You have to do something like this:

- Combine variables in some sensible way. This is often done on exams, for example, where instead of working with the marks for the separate questions or parts of questions, we give an overall score and work with that. There are many techniques that can be used for combining variables (and therefore reducing the number we have to work with)—it's called 'data reduction'.
- Break the problem down to a series of smaller problems. On a 100-question questionnaire we could, for example, analyse each question separately and present the results in a table. We'd then have 100 very simple one-variable problems. Or we could look for relationships between a person's response to the first question and each of the others, for example. Note that by doing this 'divide and conquer' we do run the risk of not including important sources of variation in a problem (point 2)—this can severely limit the types of questions that can actually be answered. If there are too many important sources of variation involved in a problem we will not be able to analyse it (KISS principle); but if we don't include these variables we may not get the right answer (identify important sources of variation). In the physical sciences (such as physics or chemistry) researchers are lucky because they can often control some variables and work with perhaps only one at a time and build up a picture that way; in the social sciences this is unfortunately not usually possible.

Generalising

Point 4 (of section 2.1 of *Statistics In A Day*) talks briefly about another great way of 'lying with statistics'. It applies when you are using sample data and trying to generalise. For example, maybe you just have a sample of data for the people who filled out the evaluation questionnaire you provided at the exit gate. And you want to make some general statements about all people who attended your function. Point 4 states that '[without further information] you can only generalise to the population from which your sample is randomly drawn'.

Here 'population' means the whole group (of people, say) that you wish to make a statement about. (Actually, strictly speaking, a population is the set of all possible measurements of the variables you are dealing with—it is the variables that you want to make the statements about, not the people themselves.) Your population of interest might be 'all people who attended

your function'. Point 4 makes the obvious point that you can only make a statement about this set *provided* your sample is randomly drawn from it. Which it probably is *not*, in this case! Probably all you can claim are statements about 'the group of people who are willing to fill out your questionnaire' (the population from which you drew your sample), which is almost certainly not the same as the set of all people who attended your function. And it is a serious (but really common) misuse of statistics to analyse your data and try to generalise too far.

The issue in point 4 only arises if you wish to generalise from a sample to a population. Luckily, you often have census data—*all* the data for a particular variable (maybe, for example, all the weekly attendance figures for the year)—and you just need to summarise it (e.g. give the total yearly attendance, or the attendance figures broken up into seasons).

Measuring values of variables

Point 5 in section 2.1 of *Statistics In A Day* is about measurement. It starts with the obvious statement that 'you [must] measure values of the variables that are used in the definition of your problem'. If you wish to see if a particular type of fundraising activity attracts more donations than another type of activity, it's not much use measuring just the amount raised under the first activity. We'd need to measure the amount raised under the second to compare. And any measurements must be unbiased (valid), too, so it wouldn't be much use measuring the amount of donations of only one particular type if what we wanted was an overall figure. That measurement would not be a valid measure of 'overall funds'.

Point 6 in section 2.1 of *Statistics In A Day* offers some bullet points that give some ideas for how to go about actually formulating a research problem that involves quantitative data. They are brief, but there is not really a lot that can be said in general, and you'll find that real-world research problems really need to be treated on a case-by-case basis.

Basic ideas of data analysis

Section 2.2 of *Statistics In A Day* takes you through two worked examples that together cover most of the basic ideas involved in data analysis. The examples are not specifically designed for this unit, but you should at least be able to relate to the problems discussed. The chapter breaks the work involved in doing data analysis into six steps that can be grouped into three phases discussed below. The steps are not really intended as a 'recipe' to be followed blindly—they are more like a guide. The analysis is carried out using Microsoft Excel together with XLStatistics (supplied on the Deakin Learning Toolkit). Before starting this section you might like to work through the example in the documentation for XLStatistics.

EXERCISE 11.1 **Work through section 2.2 of *Statistics In A Day*, up to 2.14. You should follow along on your computer. Go through the exercises in 2.13 using ShapeX.xls and ShapeXY.xls.**

Identifying, organising, entering and examining data

This part of section 2.2 covers phase 1 (steps 1, 2 and 3) of data analysis: identifying variables and organising data, entering the data into an analysis package, and examining the data descriptively.

The main points to note here are as follows:

- Analysis depends to a very large extent on the number and type of variables involved. You only have to consider two main types of variables:
  - — numerical—when you measure the variable you get numbers; and
  - — categorical—when you measure the variable you get labels.
- Data should always be organised with variables in columns, cases in rows. This is not as easy as it sounds—categorical variables are especially easy to hide! Look, for example, at the unstacked data in Cost.xls.
- After entering data into an analysis package, examine the data carefully looking for patterns and so on. The discussion in 2.10 (*Statistics In A Day*) explains what you need to look for. 'Strange' patterns are usually easy to spot!

In the XLStatistics workbooks most of the relevant tools for examining the data are provided in the Description box, but you will usually still need to vary scales, and so on. As appropriate, there are frequency charts, scatterplots and tables of numerical results. The summaries in the Description boxes are certainly not the only tools that you can use when examining the data, and often people will use results from other sheets in XLStatistics as well. For example, in a scatterplot, the addition of a trend line passing through the points often helps people see patterns. In the simplest case this might be just a straight line, but curves are also needed in many situations. These extra analyses can be found on other sheets in the XLStatistics workbooks so you should have a look around. The same tools are also available in most other analysis packages, of course. Use them.

Generalising and testing data

EXERCISE 11.2 **Work through section 2.2 of *Statistics In A Day*, 2.15–2.21. You should follow along on your computer. Do the exercises in 2.16, 2.18 and 2.19.**

This part of section 2.2 covers phase 2 (steps 4 and 5) of data analysis: generalising to the population and statistical testing.

This phase only needs to be done if you have sample data and wish to generalise to the population from which the sample was drawn. It is the 'statistical inference' phase of data analysis. The most important techniques are confidence intervals (2.18) and hypothesis tests (2.19).

These techniques are used in essentially all statistical testing—in different situations the underlying formulas change (but you don't need to know these anyway). The sections contain useful discussions of related analyses like power analysis and sample-size determination, as well as covering the basics of hypothesis testing and confidence intervals.

Writing up and presenting results

EXERCISE 11.3 **Work through section 2.2 of *Statistics In A Day*, 2.22–2.25.**

This part of section 2.2 covers phase 3 (step 6) of data analysis: writing up. There are no hard-and-fast rules here—it depends on for whom you are writing it up. But, as a rule, use pictures or tables when you can and keep technicalities to a minimum.

Review

Summary

In this topic we first discussed the role of data analysis in research and covered ideas for formulating problems. We then presented steps that can be used as a guide when doing data analysis and covered briefly the most important tools and techniques that are involved.