The Design of Research Studies – A Statistical Perspective

John Maindonald,
Statistical Consulting Unit of the Graduate School
Australian National University.

john.maindonald@anu.edu.au

If we teach only the findings and products of science – no matter how useful and even inspiring they may be – without communicating its critical method, how can the average person possibly distinguish science from pseudoscience? . . . Many, perhaps most, textbooks for budding scientists tread lightly here. It is enormously easier to present in an appealing way the wisdom distilled from centuries of patient and collective interrogation of nature than to detail the messy distillation apparatus. The method of science, as stodgy and grumpy as it may seem, is far more important than the findings of science.

[Sagan 1997, The Demon-Haunted World, p. 26, Headline Book Publishing, London.]

The Design of Research Studies – A Statistical Perspective

Contents

*Sections that are asterisked are more technical.

Summary of Contents

Additional Material

Material that supplements these notes may be found on the web page:

http://wwwmaths.anu.edu.au/~johnm/planning

As of October 2000, the main addition is a set of notes on the design of experiments.

Introduction

> In this case I believe much more could be done than is, in fact, done to prepare for the future scientific career. For the logical principles of experimental design and of reasoning from experimental results are of great interest to post-graduate students, who would appreciate definite courses in this subject. In fact however, and at present, the majority of scientific workers enter their careers without this preparation, and learn as they go, by their own mistakes and those of their colleagues.
>
> [Fisher, R.A. in Bennett, J.H. (ed.) 1989, pp. 343-346. See chapter 9 references.]

These notes address, at a preliminary level, broad planning principles that apply to many different areas of research. Anyone who has a research degree should be aware of them, whether or not they arise in their own research. They give, also, pointers that may help in getting a clear view of where the researcher's project is headed. I will have been successful in my endeavour if I kindle in at least some readers interest both in the research process itself and in the examples.

There are several reasons why researchers should take an interest in broad-ranging issues in research planning:

1. The immediate research project may take twists and turns that are different from those for which earlier study has been a preparation. This is especially likely for highly applied projects, which typically demand a range of diverse skills.
2. Those who acquire a wide range of research skills are thereby better placed, after graduation, to turn their hand to tasks different from those for which their immediate research training has equipped them.
3. Broad-based research skills will best equip researchers to respond to changing demands, as they move from task to task and from job to job in the course of their careers.

Designing the instrument panel on a large aeroplane may appear like an engineering problem. It has, also, a large human engineering component. A layout that has the potential to confuse pilots may, in an emergency, be fatal[1].

I emphasize the critical and questioning role of scientific ways of thinking. It does not much matter where you start practicing scientific thinking. What is important is that you start. As Sagan (1997) notes[2]:

> Because its explanatory power is so great, once you get the hang of scientific reasoning you are bound to start applying it everywhere.

Criticism and questioning are in tension with the openness to imaginative insight that is equally important to the research process. Data may be in tension with the theoretical insights that generated their collection.

The issue of evidence is central. There must be an assessment of the evidence in the literature that is the starting point for the research project. There must be a research strategy that will bring together data that address the research question. Statistical analysis will extract from the data evidence that relates to the research question. Finally, the new research evidence must be integrated into the body of earlier knowledge, creating a coherent account that will appear as a report or paper or thesis.

---

[1] Thus if a warning indicator does not indicate clearly which engine has experienced problems, the pilot may shut down the wrong engine. An emergency may become a disaster.

[2] In *The Demon-Haunted World,* Headline Book Publishing, London, p. 279.

My examples range widely, from social science through to pure and applied biology and physical science, with medical and health examples strongly represented. Most people are interested in their own health. I am hopeful that such examples will be of wide interest to non-medical as well as medical researchers. I have tried to find examples that are not unduly technical. I have found it helpful, at various points, to draw on ideas from the approach to clinical medicine that has the name "Evidence-based Medicine (EBM)". For those who want to understand the practicalities of Evidence-Based Medicine, I recommend the book *Smart Health Choices*, subtitled *How to make informed health decisio*ns, by Judy Irwig and collaborators. These ideas may assist researchers both with their health needs and with their research planning!

The first drafts of this monograph were written for a course that introduced a series of short courses on statistical design and analysis. Any statistical analysis must have a context. Data collection and data analysis serve the wider aims of the research project. This requires a clear view of the project's aims. There are principles that should guide the design of data collection whenever this lies in the researcher's control. Where the researcher does not have this control, it is important to examine the processes that generated the data. Focusing attention back onto the contexts from which data have come is important both for use of the data that the researcher may already have, and for thinking about any future data collection. Data do not just happen!

I will be glad to receive comments or corrections, or examples that illustrate points that I have made. I am in debt to researchers from many different areas who over the years have brought me questions and data.

Dr Harold Henderson, from AgResearch (New Zealand), has given me extensive help in removing errors and obscurities from these notes, and in drawing interesting examples to my attention. Professor Susan Wilson, from the ANU Centre for Mathematics and its Applications, has made a number of useful suggestions. Dr Gail Craswall, from the ANU Study Skills and Learning Centre, has helped with proofreading. In no way are these individuals responsible for what I have made of their help!

John Maindonald
22 September 2000

1. The Research Enterprise

> . . . at the heart of science is an essential balance between two seemingly contradictory attitudes — an openness to new ideas, no matter how bizarre or counterintuitive, and the most ruthlessly sceptical scrutiny of all ideas, old and new. This is how deep truths are winnowed from deep nonsense. The collective enterprise of creative thinking *and* sceptical thinking, working together, keeps the field on track. Those two seemingly contradictory attitudes are, though, in some tension.
> [Sagan 1997, The Demon-Haunted World, p. 287. Headline Book Publishing, London.]

There is an inherent tension between openness to new ideas, and the ruthless criticism to which the scientific research process insists (or should insist) on exposing every new idea. As well as research principles and methodologies specific to particular disciplines, there are general principles and methodologies. These notes will focus on these general principles and methodologies, and particularly on statistical methodologies, though avoiding any attempt at rigid prescription of acceptable scientific procedure. In order to discuss research planning, we will establish a framework that is broad enough for most research projects. The plan should include examination of existing knowledge, a decision on a research question or questions, a plan to follow in seeking answers, an analysis of the research data, and an eventual report.

1.1 A Conflict that is at the Heart of Research

There are two key components to any research activity. Firstly, there must be generation of new ideas that may be worth investigation. This requires openness to new ideas. Secondly, there must be critical scrutiny of all ideas, whether they are an accepted part of knowledge or new. There will be an eventual rejection of ideas that cannot withstand criticism. These two components are in tension. Failure in either may spell doom for the scientific enterprise. If criticism comes on too strongly at too early a stage, good ideas may be squashed. If it appears too late, there may be a huge waste of time from pursuit of unfruitful paths. When ideas that have not received adequate critical evaluation become accepted knowledge, nonsense readily masquerades as science.

Different types of study call for different approaches. Unduly rigid prescription is undesirable. Any adequate account of scientific method must allow room for the exercise of imaginative insight. It must also pay regard to checks on the unconstrained use of the imagination. Unconstrained exercise of imagination leads to myth, fiction and to imaginative fiction that presents itself as science. It has led, at worst, to supposed science that has been little more than a vehicle for individual and cultural prejudices. Yet without productive forms of imaginative insight, science would stultify.

Ideas may come in many ways, from working out the implications of existing theory, in reverie, from one's reading, from brainstorming sessions, from dreams, as a by-product of the process of critical scrutiny and testing, and so on. What works for one person or for one research project may not work for another. The origins of creativity are a deep mystery, part of the mystery of our humanness. The study of creativity is itself a scientific study, one that has not yet advanced to the point where it can offer deep insights. Creativity has its best chance when the research enterprise has captured

the imagination. Researchers who find their task boring and uninspiring are unlikely to be very creative. A sense of wonder is important.

Generation of ideas is less the problem than the generation of ideas that have a good chance of withstanding scientific scrutiny. There is a huge traffic in the generation of ideas that have been scientifically fruitless — iridology, palmistry, crystal balls, the star signs, sympathetic magic, augury, UFOs, and so on. Ideas from these sources have been singularly unhelpful to the progress of science. When ideas appear, there must be mechanisms for deciding which are worth pursuing. Time and energy will not be well spent on the investigation of every crackpot idea. But how does one know which ideas really are totally crackpot, and which are worth pursuing? There can be no sure criteria. Typically the researcher will stay away from lines of research that have proved unfruitful in the past. There is a risk that in rejecting such sources out of hand, an important insight will sometime be missed. It is a risk that most researchers think justified by their assessment of the trade-offs.

Repeatability

In many (but not all) areas of knowledge, it is appropriate to ask whether results can and have been repeated, by different workers in different places. An effective way to silence would-be critics is to demonstrate that they can be repeated. Results that have been obtained in one time and place, and that others elsewhere are unable to reproduce, cannot contribute to science. To become part of the body of useful scientific knowledge, results must be repeatable. Thus Fisher (1935, §7) argued that

> . . . no isolated experiment, however significant in itself, can suffice for the experimental demonstration of a natural phenomenon. . . . In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

Tukey (1991) notes that:

> Long ago Fisher . . . recognised that . . . solid knowledge came from a demonstrated ability to repeat experiments . . . . This is unhappy for the investigator who would like to settle things once and for all, but consistent with the best accounts we have of the scientific method . . . .

Scherr (1983) uses more colourful language to make a similar point:

> The glorious endeavour that we know today as science has grown out of the murk of sorcery, religious ritual, and cooking. But while witches, priests and chefs were developing taller and taller hats, scientists worked out a method for determining the validity of their results: they learned to ask *Are they reproducible*?

The demand for repeatability applies with different force and in different ways in different areas of science.

Where it is not possible to demonstrate a claim experimentally, what recourses are available? There are other ways of gathering and using evidence, which however rarely give the secure knowledge that comes from a properly conducted experiment. The two examples in the next section will illustrate some of the issues.

1.2 The Merging of Different Insights and Skills

Planning should achieve a clear sense of where research is headed, and of how it will achieve its aims. How does one get the data and do the analyses needed for a convincing end result? We begin with two historically interesting examples from the

nineteenth century. The first is from the work of Florence Nightingale, and the second from the physician John Snow.

Florence Nightingale's Crimean War Data

Fig. 1 is similar to a graph, drawn by Florence Nightingale, that is reproduced in Cohen (1984).

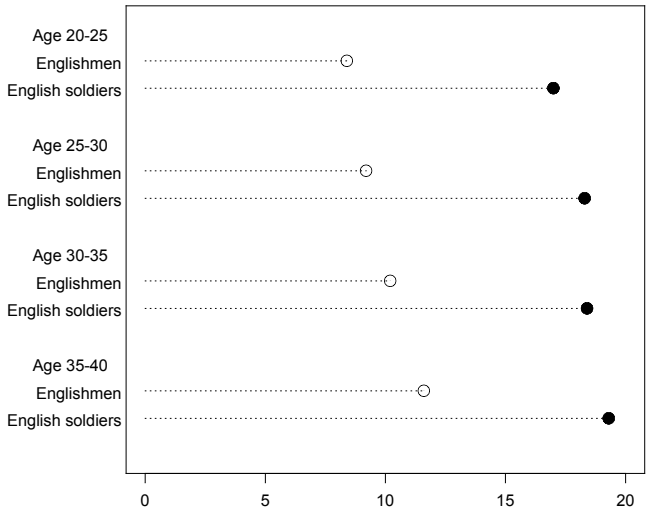


Fig. 1: Florence Nightingale's data showing deaths per 1000 per annum, for the general population and for soldiers living in barracks.

The clear message of Fig. 1 is that, at the time of the Crimean War, it was much more dangerous to be a soldier living in barracks in England than to be a male in the general population. Note that the pattern is the same for all four age groups. There were other important sources of evidence. Evidence about poor sanitation and hygiene at army barracks supported what the data seemed to say.
How much effort went into the collection of these data? Was it straightforward, just a matter of tallying up readily accessible official records? Or was it necessary to organise clerks to go out and collect it? What was Florence Nightingale's purpose in collecting it?

John Snow's Data on the London Cholera Epidemics

Our second famous set of historical data is from John Snow (1855). He presented data that showed that Londoners were much more likely to die of cholera if, after 1853, they took their water from the Southwark and Vauxhall company rather than from the Lambeth company:

Water Supply Company	Death Rate (per household)	Total Deaths
Lambeth	5 per 10,000	14
Southwark and Vauxhall	71 per 10,000	286

Often houses in the same street would get their water, some from one company and some from the other. So the source of the difference did seem to be the different sources of water. Snow noted that in 1853 the Lambeth Company had moved its

supply upstream to Thames Ditton, where the water was relatively uncontaminated. Snow wrote:

> It is extremely worthy of remark that whilst only 563 deaths occurred in the whole metropolis in the four weeks ending August 5th (1853), more than one half of them took place amongst the customers of the Southwalk and Vauxhall company and a great proportion of the remaining deaths were those of mariners and persons employed in the shipping on the Thames, who almost invariably drew their drinking water from the river.

Shoe Leather

Florence Nightingale and John Snow did much more than present data. Florence Nightingale's argument was of the kind: "Isn't this what you would expect from the conditions that prevail in British army barracks. For Snow the evidence from the 1854 epidemic clinched what he had begun to suspect on other grounds. Great cholera epidemics occurred in the British Isles between 1831 and 1866. There were competing theories as to the cause, with many blaming the air. Snow noted that cholera affected the intestines rather than the lungs, making it unlikely that it was spread as a poison in the air. He noted that when a ship went from a cholera-free country to a cholera-stricken port, the sailors would get the disease only after they had landed or taken on supplies. Exposure to the air was not enough. Snow engaged in scientific detective work. In one of the earliest epidemics he found the seaman who had been the first case, and noted that he had newly arrived from Hamburg, where the disease was active. Snow's book is a classic for the way he builds his case from the variety of evidence.

In a paper titled "Statistical Models and Shoe Leather", Freedman (1991) describes how Snow tramped around London gathering his information. Not just statistical analysis, but shoe leather, was crucial to the case that Snow finally made. It is always thus. The context from which the data come is crucial to their use and interpretation.

Statistical analysis, plus subject area insights

In the design of data collection, and in interpreting results, subject area insights should mesh with statistical and data analysis insights in ways that will vary from study to study. The researcher's challenge is to put together all the evidence – evidence from the literature, from the analysis of the researcher's own data, and less formal evidence that may not be amenable to statistical analysis, in a manner that presents a coherent story. This demand for coherence will appear repeatedly in these notes.

This monograph is written from the point of view of a practising statistician who has often been involved in the research of others. A key emphasis is that there must be a marriage of statistical insights with application area insights. There must be shoe leather as well as statistical analysis.

Careful planning will greatly increase the chances that, when your data analysis is complete, there will be a compelling story to tell. It is a fortunate researcher whose data tell a story that is as compelling as Florence Nightingale's data, or as John Snow's data. Good planning of the project, and of the data collection, can greatly increase the chances of such good fortune.

1.3 A Framework for a Research Project

The aim is to develop a framework that will be helpful in the later discussion of research projects. It is impossible to get started at all unless there is a research question, or at least the beginnings of a research question.

Asking the Right Question

An unfortunate choice of research question gets the research off to an unsatisfactory start. It gives an unsatisfactory basis for the planning of data collection. The question may at first be phrased in very general terms. A large part of the effort, initially, will then go into honing the research question, into giving it a clear focus. Often there will be some refining of the research question during the preliminary stages of the research.

Avoid questions that are unclear, or that do not give the research a clear focus, or that are too difficult to answer within the project's time and resource limitations. It is often possible to get a research degree by answering a question that is different from the one you set out to answer, but do not bank too much on this possibility! In government or industry, it may be pretty important to answer the question that was asked!

Clear research questions keep the research focused, and are a safeguard against diversion of undue energy into bypaths. One may have specific hypotheses, e.g. that two treatments for blood pressure are indistinguishable in their effect. Or one may wish to estimate the effect of a particular treatment. How does living at high altitudes affect the lung capacities of ten-year old children?

Good research planning and execution has multiple components. It should bring together relevant insights and skills from all contributing disciplines. This is a particular challenge for highly applied research, where there may be diverse multi-disciplinary demands.

Four Components of a Research Project

It will be convenient to group the different components of a research project under the headings:

1. assessment of the state of existing knowledge;
2. generation and honing of ideas;
3. the design and execution of research that will explore or test specific ideas;
4. analysis, interpretation and presentation of the resulting data.

In the next chapter, I will give a more detailed framework that has eight steps. While statistical ideas may not have much role in idea generation, they are certainly important in 1 (assessing existing knowledge), 3 (designing and executing research) and 4 (data analysis and interpretation). I will put particular emphasis on the review of existing knowledge, an area where the insights of experienced statisticians are sorely needed. Assessments of how effectively earlier workers have designed their study, and of how compelling their results are, may rely heavily on statistical insights. Even if the study design seems to stand up to critical scrutiny, the reader must ask whether the data interpretation is correct. Mistakes in the statistical analysis or in the interpretation of the analysis may lead to quite wrong conclusions, as in some of the examples that we give later.

What is the Current State of the Evidence?

Researchers will be wise to attend closely to the efforts of earlier researchers. That is why a literature review is often the starting point for new research. One wants to avoid re-inventing the wheel or pursuing what is already known to be a bypath. On the other hand, do not accept uncritically all claims made by earlier researchers. Their methodology may be inadequate, or they may have misinterpreted their data.
There are lessons both in the successes of earlier workers and in their mistakes. In order to learn from the mistakes, one needs to identify them. One aim of these notes is to sensitise readers to some of the mistakes that occur. What are the telltale signs that indicate that conclusions may not be securely based? It takes experience and maturity, and often involves issues of statistical design or interpretation.

What if the experts disagree?

The "experts" may not agree among themselves. If all sides agree that there is as yet no definitive evidence either way, and are taking different punts on what the future may hold, that is healthy. Where both sides consider that the evidence supports their judgment, the problem is clearly more fundamental. Underlying the disagreements are, as in the claimed link between salt consumption and blood pressure, differences of opinion on what is valid scientific evidence. It is then insightful to contrast the different sorts of evidence on which the different protagonists rely.

Examples

Here are examples where there is disagreement:

1. In members of the general population, does consumption of salt lead to an increase in blood pressure? The experts disagree. I have no doubt that any effect is small. Taubes (1998) was a useful summary of the evidence as it stood at that time. Since that time, there have been further results. Note in particular Sacks et al.(2001). After reading these notes you may want to look at the Sacks et al. paper, and perhaps at the Taubes article. You can then decide whether you agree with those who think that any effect is small, or with those experts who continue to believe that ingestion of salt, at levels that are typical in Western populations, leads to substantially heightened blood pressure.
2. What is the best way to teach reading? There are strongly conflicting opinions. You can read a careful summary of one set of opinions in McGuinness (1997). She has strong views, for which she presents evidence, on what works and what does not. Her arguments rely heavily on a detailed analysis of the processes involved in learning to read, in a manner that I find impressive. I would be surprised if the competing schools of thought to which she refers give up easily. I expect that they will stay around for a long time yet. I am impressed by McGuinness's approach to the issues that arise in teaching reading. This gives me confidence that the total case that she makes is, broadly, right. Some claims, e.g. her apparent complete dismissal of dyslexia as a recognisable condition, may be too extreme. Again, we will return later to a discussion of McGuinness's claims.
3. What are the long-term psychological effects of a sudden and unexpected death of a child or spouse in a motor vehicle crash for which they appeared to bear no blame. An important difference from the previous two questions is that the answer must rely on observational evidence. But is it possible to gather observational data that will closely mirror the data that one might get from an experiment? Lehman, Wortman and Williams (1987) identified 39 individuals who had lost a spouse, and 41 individuals who had lost a child in a crash over a period of four to seven years

prior to the study. They limited attention to crashes "which could happen to anyone", i.e. they had not happened because of drugs or alcoholism or errant driving. They matched exposed subjects with individuals who had not experienced a crash, based on gender, age, family income in 1976, educational level, number of children and ages of children. Their evidence seems to indicate that the major symptoms of bereavement continue much longer than earlier workers had acknowledged.

4. Classical economic arguments may view labour as a commodity for which demand will decrease as the price increases. It then follows that increasing the minimum wage will hurt the very individuals that it is designed to benefit, by reducing employment for those who are on low wages, other things "being equal". The theory relies on idealised assumptions that may or may not apply to real labour markets. There have been various attempts to test the theory against data. Card and Krueger (1994) used a case/control study approach. They compared employment in fast food restaurants in New Jersey, where there was a minimum wage increase in April 1992, with a control group of fast food restaurants in adjacent Eastern Pennsylvania. Card and Krueger found no reduction of employment in New Jersey, relative to Eastern Pennsylvania. Other researchers (e.g. Deere, Murphy and Welch 1995; Neumark and Wascher 1992) have used different methods, often relying on regression methods to partial out the effects of the various changes. Different researchers have obtained different results, some results seeming to support economic theory and some (such as Card & Krueger 1994) challenging it. The different groups of researchers hold differing views on what are legitimate methodologies. Who is right?

Statistical insights are important for all these questions. For the salt issue, there are a number of different types of study. Some of those types of study provide reliable evidence. Some do not. One of my aims is to convey a sense of the advantages and traps of the different types of study.

In discussing the teaching of reading, examination of data from studies that compare different methods is important, but not the only thing we ought to look at. We would like to know, not just that some methods work and others do not, but why they work. What impresses me about McGuinness's study is that she presents both a rationale for why her methods work, and data from studies that seem to show that her methods do indeed work better than other methods. We have a theory, supported at many of the crucial points by experimental data, that lends support to her claims. We do not always have a conjunction of scientific insight and statistical evidence that gives such coherence to the argument.

The Lehman, Wortman and Williams study of the effects of sudden and unexpected loss differed from many previous studies because of its use of a control group. It may therefore seem unsurprising that it reached different conclusions. How can one assess the effects of traumatic loss, unless there is an adequate standard for comparison?

The Deere, Murphy and Welch study of the employment consequences of minimum wage legislation does not directly contradict the Card and Krueger study. Card and Krueger examined employment in one industry only. The strength of their study is that they tried, by their choice of a control, to isolate all effects except that due to the change in minimum wage. They use a single instance to challenge a broad general theory. Deere et al. rely instead on regression adjustments. Their choice of explanatory variables is then open to question. Changing the explanatory variables, or using a transformed scale, may lead to quite different conclusions.

A Framework for Interpreting Results

Getting the scientific insight that will provide a framework within which to interpret the statistical results may be hard work. The data, and analyses, must be interpreted "in context". In a paper that makes this point with force, David Freedman (1991) calls the scientific insight component "shoe leather". He gives the example of John Snow, whose work we discussed above. Snow tramped around London over the course of the great cholera epidemics between 1831 and 1866, gathering evidence on the causes. "Remember that you also need shoe leather" is good advice for anyone who uses statistical methods.

We need to look for possible biases. When examining the work of other researchers, you may need to look in great detail at what they have done. This can be a problem if they are not very explicit about their methodology. When studies are designed to compare different reading methods, both the type of study and the design are important. The methods must be compared under conditions that are fair - it is no good using enthusiastic well-trained teachers for one method, and unenthusiastic poorly trained teachers for the other.

1.4 The Insights and Methods of Statistical Science

Here we will make a brief detour that looks at the role and nature of statistical science. Perhaps if we knew better what statistical science is, we would be better placed to comment on its role in research.

Statistical science is the science of collecting, organizing, analyzing and presenting data. This is a broad definition, much wider than the view of statistics that many first year statistics courses present. Actually, one needs a definition that is as broad as this in order to get to grips with the role of statistical science in research. I need a definition that is this wide in order to tell a coherent story! Details of this broad view of the nature of statistics will unfold as the discussion proceeds.

As data collection, analysis and interpretation are integral components of scientific research, it is scarcely surprising that statistical methodology often has a key role. Chapter 4 (pp.71-80) of JMP Start Statistics (1996) has a more colourful statement:

> The discipline of statistics provides the framework of balance sheets and income statements for scientific knowledge. Statistics is an accounting discipline, but instead of accounting for money, it is accounting for scientific credibility.
>
> …. Statistics is the science of uncertainty, credibility accounting, measurement science, truth-craft, the stain you apply to your data to reveal the hidden structure, the sleuthing tool of a scientific detective.

This is well said. A weakness is that it does not draw explicit attention to the large role of statistics in guiding data collection so that effort is directed where it will be most effective.

The Design of Data Collection

Faults in this department may be of many kinds. At worst, the design may be so fatally flawed that the data are incapable of answering the question that is asked. Or undue effort may go into getting information on features of the data that are irrelevant to the question asked. For comparing storage treatments for fruit, with treatments applied to whole trays, should effort go into getting a large number of fruit. Or is it the number of trays of fruit that are important? Experiments that are too small, or are otherwise incapable of providing answers to the questions that are asked, are in general a waste of resources. Experiments that are unnecessarily large, or that gather

large amounts of information at a level that makes little difference to the accuracy of the overall result, are also a waste of resources.

There is a great deal more to statistics than p-values

Discard any notion that statistics is all about hypothesis testing and p-values. These perhaps have their place, but they should not have pride of place. Researchers who are content with the calculation and presentation of an occasional p-value are setting their sights very low indeed. They have forgotten that the aim is to gain insight on questions that are of scientific interest. Often a reasonable aim is to develop a model that accurately describes the data, aids in understanding what the data say, and makes prediction possible. Compared to the insight that such a model may provide, the rejection (or acceptance) of a null hypothesis is a minor achievement.

Every study should address clear focused questions. One way to give a study focus is to choose a hypothesis that is to be tested. If there are many hypotheses, then focus is lost. The statistical testing of multiple hypotheses gives a similar lack of focus to the analysis. This point has especial force when there is an obvious good alternative, such as examination of a response curve. Researchers who find themselves presenting numerous p-values should rethink their analysis and/or their presentation.

The questions that statistical analyses are designed to answer can often be stated simply. This may encourage the layperson to believe that the answers are similarly simple. These notes will repeatedly make the point that effective statistical analysis requires appropriate skills. These skills are not acquired by taking one or two undergraduate courses. They are gained from professional training in the use of modern tools for data analysis, and from experience in using those tools with a wide range of data sets.

Influences on the modern practice of statistics

Statistics is a young discipline. Only in the 1920s and 1930s did modern ideas of hypothesis testing and estimation begin to take shape. Many recent advances have resulted from a dawning understanding of the new possibilities that result from the power of modern computers and modern computing tools. Different areas of statistical application have taken these ideas up in different ways, some of them starting their own streams of statistical tradition that have separated from the mainstream of development of statistical ideas. Gigerenzer et al. (1989) examine the historical origins of these different currents of ideas, commenting on how they have influenced practice in different research areas.

Both the statistical mainstream and many of these separate streams have placed an exaggerated emphasis on tests of hypotheses. Outside of the mainstream there has been a neglect of pattern, an all too common insistence on styles of analysis that are not insightful, a failure to take on board modern statistical analysis approaches and the policy of some editors of publishing only those studies that show a significant effect. Thus Nelder (1999) argues that

> … the practice of statistics has become encumbered with non-scientific procedures which perceptive scientists and experimenters are increasingly finding to be irrelevant to the making of scientific inferences. … The kernel of these non-scientific procedures is the obsession with significance tests as the endpoint of any analysis.

Why do these procedures continue in use, if they are in fact of such little help in making scientific inferences? Nelder has two targets of blame: (1) editors who will not accept papers unless they follow these procedures, and (2) his perception that

many scientists pass through their training without getting any real insight into the methods of science. Nelder is arguing that statistical science is a key component of scientific method.

1.5 The Data Analyst's Tools

Graphs

> One picture is worth ten thousand words
> [Frederick R. Barnard, Printer's Ink, 10 March 1927.]

There is no good substitute for close scrutiny of the data. Generally, graphs are the best way to do this. It does, though, make a lot of difference what form of graph you draw. Why is it so hard to detect, using numerical checks, features of data that are immediately obvious from examination of an appropriate graph?

Every statistical analysis should be accompanied by graphs. You can and should see the analysis both ways, statistical text and graphics. Tight linkage between statistical analysis and graphical presentation is the wave of the future. The aim is to combine the computer's ability to crunch numbers and present graphs with the ability of a trained human eye to detect pattern. It is a powerful combination.

Using and extending an analogy in the manual for JMP Start Statistics, statistical analysts require an attractive workshop, where you know just where to find each tool that you need, where the tools float back of their own accord into the right place after you've used them, and where going into the workshop to mend the rocking chair becomes a pleasure! In this workshop, graphs are pretty important tools.

There are some great books on the principles that should be followed in creating graphs. See especially Cleveland (1985, 1993), Tufte (1983, 1990 and 1997), Wainer (1997) and Wilkinson (1999).

Statistics and Mathematics

Statistics is not mathematics, in spite of the impression that some statistics textbooks give! Statistical methods rely heavily on mathematical theory. This is not a lot different from the way that quantum mechanics or relativity theory or other areas of theoretical physics have their own mathematical theory. While there is much that one can learn without getting deeply into this theory, there are limits, and any attempt to treat statistical methodology from an elementary point of view must hit against them. The big advantage of statistics over applications of theoretical physics is that the output from a statistical analysis can more often be summarised in a few readily intelligible graphs.

Statistical Software

The interplay between computing power and theoretical development has made a huge impact on statistical methodology, both for design of data collection and for analysis. These developments have taken advantage of the increased power of computers and of the programs that drive them. We can do a much better job on many analyses than was possible ten years ago. We have become much more aware of the benefits and traps of different analysis approaches. Both the teaching and the practice of statistics need to change to reflect these advances. Why continue to use makeshift methods that were necessary when statistical computing software was at a very early stage of development?

Influences from new research developments are obvious in the best of the statistical packages that have been designed or adapted for use in teaching statistics. Examples are Data Desk[3] and the more recent JMP (from SAS[4]). Both have a fresh and modern style, have great graphics, and link data analysis closely with graphics. The large packages that go back to the mainframe era of computers have often been slower to adapt.

SPSS[5] has been popular for the processing of data from large surveys. It has been slow to incorporate the modern abilities that one finds in S-PLUS[6], which I discuss below. Minitab[7], which at one time seemed the package of choice for use in teaching, now has a number of competitors in this market. Each package has its own areas of strength and weakness.

I have used S-PLUS, a system that is popular with professional statistical users, for the graphs that appear in this monograph. It has been a common test-bed for the implementation of new statistical methodology. It is strong on graphics, with a tight linkage between graphics and analysis. If an analysis is not already available, it is often straightforward to write a few lines of S-PLUS code that will do what is wanted. S-PLUS is built around an implementation of the S statistical language

R[8] implements a dialect of the same S language that is used in S-PLUS. An attraction of R is that it is free. Development of R is a substantial international co-operative effort. R has spawned a variety of associated projects. It is setting new directions for statistical software development, and will be highly important for the future of statistical computing.

The Statistical Consulting Unit has had a tradition of using GenStat[9]. GenStat handles hierarchical analysis of variance in a highly elegant manner. Its windows interface is superior to that in S-PLUS, especially for novices. Also it does better than S-PLUS at providing, by default, diagnostic output that users should examine as a matter of course.

Particularly for medical applications, Stata[10] is attractive. It has a high quality of technical documentation. Its web page is unusually helpful and careful in the documentation of known bugs and in the provision of fixes.

All of these packages have the potential to be generally good vehicles for initial analysis. None of them can be a substitute for expert knowledge or assistance. For anything that is non-trivial, decoding and understanding the output is usually, also, a non-trivial task.

Why not use Excel for data analysis?

Excel is a convenient tool for data entry, and possibly for simple data checking. Even for this purpose, there is need for care. Excel will not object if you have spaces or non-numeric values in columns of supposedly numeric data. You can use the sum icon, or the SUM function, to take the sum of such a column. Blank cells, or any cell that contains a non-numeric value, are ignored. Thus if you type 1O (one oh) instead

[3] http://www.datadesk.com
[4] http://www.sas.com
[5] http://www.spss.com
[6] http://www.mathsoft.com (in Australia http://www.cmis.csiro.au/S-PLUS)
[7] http://www.minitab.com
[8] To find out more about R, or to copy down the binaries (for the PC under Windows, for Unix or for Linux), go to the web site http://mirror.aarnet.edu.au . My document that describes the use of R for data analysis and graphics is available from http://wwwmaths.anu.edu.au/~johnm/r/r4dat-gr.pdf .
[9] http:// http://www.nag.co.uk/stats/tt/5thedition/new_5th.html
[10] http://www.stata.com

of 10, or 1l (one ell) instead of 11, the entry in that cell will be ignored[11]. There will be no warning.
Excel's statistical features have severe limitations and traps. To get the 2-sided 5% critical value for the normal distribution, one enters NORMINV(0.975). To get the 2-sided 5% critical value for the t-distribution with 20 d.f. one enters, inconsistently TINV(0.05, 20). What Excel calls a histogram is in fact a barchart. The function STEYX, which is supposed to return the "standard error of the predicted y-value in regression", in fact returns the square root of the error mean square. The data analysis toolkit has, for use in connection with regression, a so-called normal probability plot that is nothing of the sort. It gives a line if the y-values are evenly spaced.
Negotiating such traps may require professional statistical skills. Professionals usually opt for more appropriate tools, that allow better scope for their skills.
Anyone who wishes to work directly from an Excel spreadsheet to do simple analyses should consider ActivStats for Excel (Velleman 2000). This fixes most of Excel's errors and traps.

1.6 Practicalities

There are many important issues that are outside the scope of this monograph. These include: 1) funding; 2) the use of libraries and other information resources; 3) computing equipment requirements; 4) sources of help; 5) oral presentation of results; 6) intellectual property; and 7) job search. You will find brief comments on all of these, and useful references, in Greenfield (1997.)

References and Further Reading

Card, D. and Krueger, A. 1994. Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. American Economic Review 84: 772-793.

Cleveland, W. S. 1993. Visualizing Data. Hobart Press, Summit, New Jersey.

Cleveland, W. S. 1985. The Elements of Graphing Data. Wadsworth, Monterey, California.

Cohen, I. B. 1984. Florence Nightingale. Scientific American 250: 98-107.

Deere, D., Murray, A. and Welch, F. 1995. Employment and the 1990-1991 minimum-wage hike. American Economic Review 85: 232-237.

Fisher, R.A. 1935. The Design of Experiments. Oliver and Boyd.

Freedman, D. A. 1991. Statistical models and shoe leather, with discussion by R. Berk, H. M. Blalock and W. Mason. In Marsden, P., ed., Sociological Methodology 21: 291-358.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. 1989. The Empire of Chance. Cambridge University Press.

Greenfield, Tony, ed. 1996. Research Methods. Guidance for Postgraduates. Arnold, London.

Lehman, D., Wortman, C., and Williams, A. 1987. Long term effects of losing a spouse or a child in a motor vehicle crash. Journal of Personality and Social Psychology 52: 218-231.

McGuinness, D. 1997. Why our Children Can't Read. The Free Press, New York.

Nelder, J. A. 1999. From statistics to statistical science. Journal of the Royal Statistical Society, Series D, 48, 257-267.

Neumark, D. and Wascher, D. 1992. Employment effects of minimum and subminimum wages: panel data on state minimum wage laws. Industrial and Labor Relations Review 46: 55-81.

[11] Providing the column alignment (click on Format, then on Cells) is set to General, such illegal entries will appear left adjusted, whereas numbers are right adjusted. This allows a visual check.

[See also (1993) 47: 487-512 for a critique by Card and Krueger and a reply by Neumark and Wascher.]

Sacks, F.M., Svetkey, L.P., Vollmer, W.M., Appel, L.J., Bray, G.A., Harsha, D., Obarzenek, E., Conlin, P.R., Miller, E.R., Simons-Morton, D.G., Karanja, N., and Lin, P.-H. 2001. Effects of blood pressure on reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. New England Journal of Medicine 344: 3-10.

SAS Institute Inc. 1996. JMP Start Statistics.

Scherr, G. H. 1983. Irreproducible Science: Editor's Introduction. In *The Best of the Journal of Irreproducible Results*, Workman Publishing, New York.

Snow, John. (1855) 1965. On the mode of communication of cholera. Reprint ed., Hafner, New York.

Taubes, G. 1998. The (political) science of salt. Science 281: 898-907 (14 August).

Tufte, E. R. 1983. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, U.S.A.

Tufte, E. R. 1990. Envisioning Information. Graphics Press, Cheshire, Connecticut, U.S.A.

Tufte, E. R. 1997. Visual Explanations. Graphics Press, Cheshire, Connecticut, U.S.A.

Tukey, J. W. 1981. The philosophy of multiple comparisons. Statistical Science 6: 100-116.

Velleman, P. 2000. ActivStats for Excel. Data Description Inc., and Addison Wesley Longman.

Wainer, H. 1997. Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot. Copernicus Books.

Wilkinson, L. 1999. The Grammar of Graphics. Springer, New York.

2 The Structure of a Research Project

There are broad planning principles that apply across many different areas of research, and which are the subject of this monograph. In addition there are insights and approaches that are specific to particular areas of research.

Any effective research project must build on existing knowledge, and must ask pertinent and incisive questions. Where new data are needed, data collection methods should be designed to ensure that they are accurate, relevant and interpretable. The information in the data must be teased out in ways that will help answer the research question. Finally this information must be communicated.

Techniques for gathering, refining, systematising and interpreting information are a large part of research methodology. Some techniques are highly specific to individual subject areas. Others have a more general relevance that extends broadly to all research. Statistical techniques and insights may be needed at many different stages of a research project, starting with the overview of existing knowledge. Different research areas may have very different demands.

2.1 The Different Demands of Different Areas of Research

There are broad planning principles that apply to many different areas of research. The manner in which they apply varies. My examples will range widely, from social science through to pure and applied biology and physical science. The differences between the dominant demands of these different areas are in

1. the extent to which validity seems an issue. Are the data what they seem to be; do they really measure, for example, well-being? This is commonly a key issue in marketing or health social science. It is often an issue in biology. It is much less often an issue in physical science;
2. the signal to noise ratio – commonly low in marketing or health social science and high in physics, with biology somewhere in between;
3. the types of measurement instrument – whether questionnaires, visual assessment e.g. of a pattern on an agar plate, physical measurement, or a mixture.

One result of these differences in predominant emphasis is that researchers who have been trained in one area may find it difficult to make the necessary adjustments when they move to another area. For example, there are many areas of engineering where the signal to noise ratio is so low that it can, most of the time, be ignored. Those who have come from this background of experience may have difficulty making the necessary adjustment when they come to work on engineering aspects of experimentation with fruit, e.g. the mechanics of bruising.

Investigations that work very close to the limits of detectability require special care. Biases that are unimportant in more robust experiments can create havoc. The techniques used to detect a few molecules of a trace chemical must be far more rigorous than those that one would use to detect concentrations of a few milligrams per litre.

There are good reasons why you should be aware of the differing research demands of different areas of work. There are large areas of research that cross disciplinary boundaries. There may be components of your research that will call for a style of research different from that for which your undergraduate training has prepared you. Increasingly engineers who design new systems must worry about human engineering

issues, whether or not these have been part of the their training. Human engineering issues are for example crucially important in the design of aircraft instrument panels, in the design of aircraft fly-by-wire systems, and in the design of computerised systems for delivering precise doses of radiation. Biologists and anthropologists may, for their work, need to use measurement or chemical assay devices.
Many of those employed to do research on fruit storage or transport have been trained as engineers or chemists or physicists. They thus move from an area where variability is commonly not a major issue to an area where everything varies. The research demands are thus different. Food chemists may find it hard to adjust to working with the subjective judgements provided by taste panels. Engineers who move into management positions may be uncomfortable with market research methodology. Econometricians whose models of the total Australian economy cannot be rigorously tested may not be well attuned to the careful criticism and testing that is desirable in situations where this is a possibility. Models for use in hospital economics can and should be rigorously tested and criticised, in a manner that may not be possible for models of the Australian economy.
So even if some of the discussion seems remote from the current demands of your own research, bear in mind that you may at some point move into an area of work that requires this knowledge.

2.2 The Research Question

Suppose that you have decided to do study on the teaching of reading. There are several possible starting points

1. Your supervisor(s) may have a very specific study for you to undertake.
2. Your supervisor may tell you that he/she thinks that a specific topic requires attention, but you will need to make yourself familiar with the literature, decide just how much is already known, and come up with a research question that is reasonable within the resources and time that you have available.
3. You may be left pretty much on your own to search out a research question within the general area of the teaching of reading.

The likely extent to which the researcher will need to refine the research question varies from one area to another. In health social science the refining of the research question may be a large part of the exercise, while in biochemistry or physical science the research question may already be tightly prescribed.
Even if the research question seems to have been tightly determined, be prepared for surprises. It may turn out that the research question is not as clear, not as sharp as you had thought. Or you may find that it has already, largely, been answered. At the other extreme it may turn out to be impossibly difficult, so that you need to modify it to something less ambitious. You may find that, because of questions that arise as you proceed, there is a whole new area of literature that you need to explore.

2.3 Ways That Projects Differ

In addition to differences in the nature of research questions, projects may differ:

1. in the extent to which the researcher requires new knowledge, and in the extent to which that new knowledge is available from such `obvious' sources as books and journal articles;
2. in the methods that will be used for collecting data – experiment, published data, data archives, cross-sectional or longitudinal survey, etc.;
3. in the extent to which you will need to develop new methodology or new measuring instruments;

[It is possible to occupy a whole PhD with the development of methodology that other researchers can then use, perhaps a new method for estimating the amount of carbon in the soil, or perhaps a new health measurement scale.]

4. in the extent to which the research will be an individual effort, or part of a co-operative project.
5. in the range and extent of multi-disciplinary demands.

In all these areas, be prepared for surprises. Current measuring instruments may prove less adequate than you had expected, and you may have to develop your own. The skill demands may be different, and/or more diverse, than what you had initially expected.

Below, I will now set out steps that a research project might follow, and comment on the role of statistical insights and methods at each step.

Question: For each of the criteria 1-4 above, where in the spectrum does your project fall? Are there other special issues that arise for your research, that none of these criteria capture?

[The answers you give to this question may affect the importance you attach to the steps that I describe below for a `typical' research project.]

2.4 Eights Steps in a Research Project

My `typical' research project has eight steps in all. Some research projects will take the researcher right through the complete sequence. Others will focus on particular steps within this sequence. They may for example build heavily on the groundwork that other researchers have laid. Or they may set in place a foundation on which future researchers can build. The work involved in the earlier steps may be of such novelty or difficulty that all the effort goes into, for example, identifying the important issues that require study.

The eight steps fall under four broad headings:

1. assessment of the state of existing knowledge;
2. generation of ideas;
3. the design and execution of research that will explore or test specific ideas;
4. interpretation of the resulting data.

As you progress through to later steps, there is likely to be a fair amount of retracing of earlier steps. For example, all the later steps will help your understanding of the research context for the study, which is the focus of step 1. Following each step, you should review your progress to date, and revisit earlier steps as necessary.

Statistical insights have large implications for the design of data collection (step 3), for analysis (step 4), and for presentation of results. They may also have large implications for critical review of the existing literature (step 1).

Eight Steps

1. Search out the research context

There are several facets to this. It is necessary to know, as well as you can, the state of existing knowledge, what existing data may be available, etc.

What is the state of existing knowledge?

You will discover this by talking to any experts you can find, and by reviewing the literature. In a case where the experts disagree, or seem unable to give convincing reasons for their judgments, you should be careful about accepting at face value the opinion of any one expert. You may, finally, need to make your own judgment.

You may need to look critically at claims made in the literature. This may include

(i) looking critically at the experimental or sampling design that generated the data;

(ii) critical examination of the data analysis;

(iii) critical examination of the interpretation.

We will later give examples of the ways in which authors have got one or more of these wrong.

What Existing Data are Available?

There may be existing data that bears on the research question, but which have not been adequately analysed. You then must make a judgment call on how much effort it is worth putting in to analyse the data. The benefit is that it will cost you nothing to get the data. A risk is that the data may not be as useful as at first appeared. Notwithstanding any assurances that you receive about the relevance and usefulness of the data, it may turn out that the data are of poor quality, relevant to a different question, without the documentation that you need to make any use of them, or otherwise not useful. There may be good reasons why they have not been analysed and the results published. I've had this experience. So take care!

2. Canvass for ideas and formulate specific questions

Generation of a hypothesis, or of hypotheses, is not a statistical activity. It requires some of the elements of what social scientists now call qualitative research. In extreme cases, you may not, in the first project, get beyond the qualitative research stage. You may need to find ways to escape from current mindsets. Brainstorming techniques are often quite helpful. Many different people may have light to shed on the question at issue. So the idea is to get them together in a setting where they feed off and stimulate each other's thinking.

Questions of the "What is going on here?" type may not lend themselves, in the first instance, to quantitative investigation. The Ministry of Health in a developing country may be concerned to know why some medical services are used, and some are not. Or a particular service may be used in one centre, but not in others. It is necessary to talk to users, both of well-used and of under-used services, and to seek insight into what motivates people to use the services. An apple transport trial in which I participated would probably have benefited from the insights of horticultural producers on the causes of apple damage when fruit were transported. `Focus groups', on which there is an extensive literature, are a structured technique for seeking the insight that I have in mind.

3. Determine what type of study is needed

The study may be an experiment, or a quasi-experimental study, or a sample survey, or an observational study. You need to decide what kind of study is most likely to provide good answers to your research question. What is important is that you use a form of study that is in principle able to answer the questions that you ask. Here are some of the issues.

i. Properly designed experiments allow clear cut answers. If undertaken with proper care, there is often little room for argument.
ii. It is not always easy to design an experiment so that results are unequivocal. Thus human subjects know that their responses are being measured, and may change their behaviour. Doing double blind trials that compare a group who consume 6 gm of salt per day with a group who consume 10 gm per day has logistical problems. Is it possible to ensure that participants and clinicians do not know which diet subjects are on? How does one ensure that salt is the only difference?
iii. Many important questions do not lend themselves to experimentation. It is not ethical to expose different groups of human subjects to different levels of radiation, in order to develop a dose-response curve for the effects of radiation. No-one would agree to an experiment in which one group of school leavers was randomly assigned to go straight into the workforce, while another were assigned to go first to university, with the aim of seeing who gets the higher salary by age 30. An

imaginative government might however be able to mount an experiment in which different areas were randomly assigned to different approaches to tackling unemployment.

iv. In addition to the logistical problems of doing experiments, there are cost issues. Experiments in which large commercial buildings are randomly assigned to two different construction methods are, at the very least, unusual. They'd need a wealthy and enlightened backer. [Experiments of this kind have however been undertaken to compare the effects of different insulation regimes.] An experiment in which, after construction, there was a destruction test to determine the strength of the building, would require a very wealthy backer!

v. Observational or quasi-observational studies are typically much less expensive than experiments, and easier to mount. One way to make the comparison between the two types of construction method is to compare buildings that have been constructed using the two different methods. There will from time to time be earthquakes in one or other place that do an unplanned destruction test. Are the data from this just as good as data from a planned experiment? Are they even more useful?

vi. Governments and organisations, by the changes they make, are all the time carrying out experiments, though usually they describe them as "reforms". These changes might often be better run, in the first instance, as formal experiments. For example, Government might take five pairs of hospitals, with the two members of the pair carefully matched, then randomly assign one member of each pair to the current management regime and the other to the new management regime that is under trial.

vii. In what is really a case-control study, every motor-cyclist and every tenth car driver are stopped on a freeway and asked whether they have had a serious accident, requiring hospital admission, in the previous 12 months. The rate among car drivers is found to be twice that among motorists. Motor-cycle accidents may more often be fatal. Motor-cyclists who have serious accidents may give up motor-cycling and become car drivers. There are `confounding' effects at work here. (Christie et al., 1987.)

viii. In all studies that have an observational element, there is a potential for confounding. In a case-control study, the two groups may differ in more than the exposure.

Once again, what matters is that the study should in principle be able to answer the questions that are asked. This is an issue of statistical design.

4. Design the study

There are large statistical issues here. For experiments, what are the treatment units, how large should they be, and how many of them are needed? How can one avoid confounding? Would some form of blocking improve precision? How will information be collected on each experimental unit (e.g. measure all plants, or just a sample), and how should it be collected?

For sample surveys, what is your target population? What sample design will give the best precision for a given cost? How many primary sampling units are required, how many secondary sampling units, and so on? Will you design your own questionnaire, or will you adapt an existing questionnaire? How can you avoid questions that may puzzle respondents, loaded questions and/or ambiguous questions. How will you handle non-response?

Your design should include planning of the details of data recording. Will you enter data onto a sheet, or directly into a computer? If onto a sheet, do you need a specially

designed form or forms? If into a computer, do you need a computer entry form that can be displayed on the screen. How can you be sure that the data are entered correctly?
In experimental work, photographs and/or video recordings may be useful as records of information that you may want to check on later. (We found them invaluable when, in the apple transport experiment, we needed to check back afterwards on the original labelling on some of the wooden bins.)

5. Design and carry out a pilot study

This provides a check that your planning has been adequate, and should lead to refinement of your study design. The pilot study provides a check, of your general study planning, of the study design, of your measurement devices or instruments, and of practical aspects of data collection. In deciding whether you need a pilot study, consider whether you could afford to repeat the study should something go wrong. The `piloting' of a new form of questionnaire that is to be used as an `instrument' for measuring e.g. hospital patient satisfaction or general sense of well-being, may be a long and demanding process.

6. Carry out the study and collect the data

This is where the quality of your planning is, finally tested! Logistical, rather than statistical, skills are required at this point. Be sure, however, to keep your eyes and ears open for evidence of problems, or for the unexpected. A factor that you had not incorporated into your design may turn out to be important. There may be implications for your later interpretation of the data. Thus in the apple transport experiment that I mentioned earlier, the intention was to compare the effect of two truck suspension systems (mechanical and air bag). It turned out that the major source of damage was unstable bins! We became aware of this when we noticed that one bin that showed unusually serious damage was rickety.
An adjunct to the process of data collection must be careful checking and re-checking of data, to avoid errors. It is often helpful to do initial exploratory data summaries as data are collected. Any problems in the data can be investigated there and then.

7. Analyse the data

The data analysis has, broadly, two parts. There is an exploratory data analysis where you examine various forms of data summary, both in case they have a message that you need to consider and in order to check whether the assumptions of the intended formal analysis seem reasonable. Exploratory data analysis allows the data, as far as possible, to speak for themselves. I referred earlier to an apple transport experiment. In that experiment the exploratory data analysis started when fruit were examined for transport damage. Unusually heavy damage in a particular bin alerted us to the need to look for some major source of huge damage that had nothing to do with truck suspension.
The formal data analysis directly addresses the issues that the study was designed to examine. Following the formal analysis, there is further exploratory data analysis that one can and should do. There can be more carefully targeted checks on assumptions. (After the smooth has been removed, you can see the rough more clearly.) You can check whether there is anything that the analysis has missed.

8. Write the report(s) and/or the paper(s)

There are important issues here of statistical presentation. One can debate whether they are specifically statistical issues. They are issues where statisticians will have comments and insights. It is important to communicate results clearly and accurately. If those who need to assess or use the results cannot understand the exposition, the effort may have been largely wasted.

2.5 Effective Planning

Planning should find a balance between thoroughness and attention to detail on the one hand, and leaving room for learning as you go along. Here are points to keep in mind as you try to strike the right balance:

1. Plan for review and re-evaluation after finishing one phase of your study and before you move on to the next phase.
2. The results of the literature review may have big implications for planning. So do not set plans in concrete until you know what the literature says.
3. Wherever possible, use a pilot study to test the design, the techniques and logistics before proceeding with any major experimental or data collection exercise. Changes made to the design part way through an experiment or data collection exercise are a recipe for disaster.
4. If it becomes obvious part way through that changes really are needed, talk to a statistician about whether this is possible without invalidating the design. Ideally you should carry the current experiment through to conclusion, and then mount a new experiment with the changed plan.
5. Plan your general approach to data analysis, and ensure that you will have access to the resources and skills that you need. Unless you have been through the same type of analysis with the same type of data so many times that it has become routine, you should not try to plan the analysis in detail. The data may have a message for you about the details of the appropriate analysis.

The Literature Review

Books on statistics commonly focus on the role of statistics in the design of data collection and in analysis. They have little to say about the role of statistics in the review of existing knowledge. This is a deficiency.

If there are a small number of key papers that provide the information you need with complete clarity, you are fortunate! Questions that may arise are

1. Were there confounding factors; i.e. is it possible that the result is explained by something other than the factor assumed responsible for differences between groups?
2. Is the statistical analysis adequate? Is it correct?
3. Have the results of the statistical analysis been correctly interpreted?

Depending on the journal and on accidents of the refereeing process, published results are not always well analysed and/or presented. Your assessment of current literature may depend quite crucially on issues of statistical design and analysis. The standard of data analysis may vary from extremely cursory and inadequate to very careful. How does one tell the difference? There is a more extensive checklist in Appendix I.

Another issue is experimental precision. Did the experimenter use precise equipment, and/or a precise experimental design? You need this information both in order to make a good assessment of those papers, and because of the implications for your own design and data analysis. It is easier to be detached when you examine someone else's experimental design.

A further issue is bias. Results may be highly repeatable, but they may have consistent and unknown biases. The placebo[12] effect, and the tendency of many

[12] The placebo effect is an improvement that occurs merely because the patient is receiving the attention of medical staff. There may be an improvement from giving patients harmless and ineffectual tablets, e.g. made of glucose, to swallow.

medical conditions to improve over time, can operate in subtle ways to induce biases. It is necessary to ensure that the control group and the treatment group benefit equally from any placebo effect.
These issues become even more important when you examine reports, or documents copied down from the internet. Such material has often not been refereed at all, either by a subject area specialist or by a statistician.

Designing the Data Collection

Be sure to talk to a statistician! There are two key issues – getting a design that is valid, and getting efficient use of experimental resources. There can be a huge difference between a poor design and an efficient design in the amount of experimental material and/or effort. You should ensure that the experiment has sufficient accuracy that it will in principle be able to detect effects of the magnitude that are of interest.

Planning the Analysis

You are strongly recommended to see a statistician and plan out the broad details of your analysis. You should get a sense of what general style of analysis may be appropriate. At the same time, leave room for messages, found in the data themselves, about what analysis may be appropriate.

The Ethics of Planning, Execution and Analysis

Research must conform to accepted ethical principles. Fraud, involving the faking of results or the manipulation of data or results, is obviously a serious breach of ethical principles. When it happens, or is suspected, it creates serious ethical problems for fellow-workers. Indications of fraud are often evident in the data or in other forms of experimental evidence. In studies that have a high profile, it is almost inevitable that fraud will in due course be unmasked.
Researchers who work with animals or with human subjects must ordinarily seek ethical approval. The *Declaration of Helsinki*[13] sets out, in general terms, standards for medical research. The requirements are wide-ranging. They include:

- Research must conform to generally accepted scientific principles.
- There should be a careful assessment of the relative risks and benefits.
- Published results should "preserve the accuracy of the results".
- The protocol should include a statement of ethical considerations.
- There must be special caution where there may be environmental effects.

The quality of the science is an ethical issue. Flawed studies, if they carry any credence at all, may mislead. One should not put patients at risk or inconvenience, in order to carry out a study that brings no benefit or may mislead. For just these same reasons, there is a duty on researchers to fairly elicit and present the information that is in the data. These same issues arise, though perhaps less cogently, in other research. (Greenfield 1997, chapter 5.)
Silverman (1998) includes extensive discussion of issues that relate to the conduct of clinical trials. See especially chapter 13, pp.48-52.

References and Further Reading

Beveridge, W. I. B., 3rd edn 1957. The Art of Scientific Investigation. William Heinemann Ltd., London.

[13] The document may be found on the web site http://www.faseb.org/arvo/helsinki.htm

Christie, D., Gordon, I., and Heller, R. 1987. Epidemiology. An Introductory Text for Medical and Other Health Science Students. New South Wales University Press, Kensington NSW, Australia.

Greenfield, Tony, ed. 1996. Research Methods. Guidance for Postgraduates. Arnold, London.

Manly, B. F. J. 1992. The Design and Analysis of Research Studies. Cambridge University Press.

Silverman, W.A. 1998. Where's the Evidence. Debates in Modern Medicine. Oxford.

3 Alternative Types of Study Design

> It is better to light a candle than to curse the darkness.
> [Ancient Chinese proverb.]
>
> It is better to curse the darkness than to light the wrong candle.
> [Notice to workers in a fireworks factory.]
>
> *A first task must be to decide on a clear research question. The type of study design will depend on what it is hoped to achieve, on what information is already available, and on available resources. The study design will impose limits on the inferences that can be drawn from the data. Large studies may have components of two or more different types of study design.*

Structured methods for collecting data include **experiments**, **censuses** or **sample surveys**, prospective or retrospective **longitudinal studies**, **case-control studies, cross-sectional studies**, and various forms of structured **observational study**. Properly designed experiments or sample surveys are the most structured of all these approaches to data collection.

My focus is on quantitative studies. There are in addition various types of qualitative study. Often, some mix of qualitative and quantitative approaches will be appropriate.

3.1 The Question of Salt, Again!

Since the 1970s there has been a widespread expert medical view that salt consumption is unhealthily high in many industrialised countries. Official guidelines from the National Heart, Lung and Blood Institute and the National High Blood Pressure Education Program, both in the U. S. A., recommend a daily allowance of 6 grams, that compares with the current 10 gram American average. The issue is highly controversial. A huge amount of effort has been expended to determine what the effect of salt really is. Some answers have now, I believe emerged. Not everyone agrees.

An interesting aspect of the controversy is the variety of the approaches that have been used. The main studies have been of the following types:

1. Animal experiments, in the tradition of studies that Dahl (1972) conducted on rats;
2. *Inter-population studies*, often called *ecologic* studies, that compare different populations;
3. *Intra-population* studies that compare different individuals within the same population;
4. *Non-randomised Clinical Trials*;
5. *Randomised Clinical Trials,* but open to criticism because (i) they were not *double-blind*, and/or (ii) they did not use *placebo controls*, and/or (iii) the changes in salt intake were accompanied with changes in other aspects of the diet.
6. *Randomised Clinical Trials* that meet strict design requirements, i.e. `high quality' clinical trials.

Which of these sources of evidence do you consider reliable? A major aim of this chapter is to draw attention to the strengths and weaknesses of these and other study designs.

The Science of Salt – Background

Here is some further background to the salt controversy. The initial evidence was quite insecure. One type of argument came from blood chemistry. Increased salt consumption causes the kidneys to respond by excreting more salt. There will be a temporary increase in blood pressure. Might this not lead to a permanent increase? In 1972 Dahl bred a strain of rats that developed high blood pressure when fed large amounts of salt, suggesting that salt and blood pressure were somehow linked.
Dahl had earlier (1960) presented evidence that seems to link differences in hypertension (blood pressure) in different populations with differences in salt intake. The most convincing evidence seemed to come from studies that compared indigenous populations with people in industrialised societies. They found low salt and little hypertension in the indigenous societies, compared with high salt and much hypertension in industrialised societies.
Fig. 2, using data from Intersalt Cooperative Research Group (1988) shows a plot of median blood pressure against median sodium excretion, from 52 populations. (Each point is derived from 200 individuals; for each population researchers took a sample of 25 males and 25 females from each decade in the age range 20 - 50.)

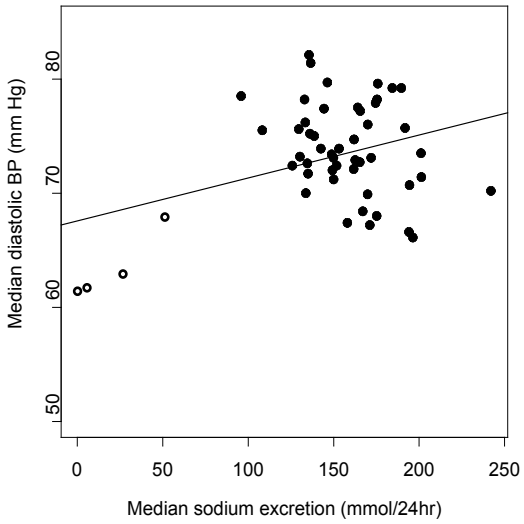


Fig. 2: Plot of median blood pressure versus salt (measured by sodium excretion) for 52 human populations. Four results (open circles) are for non-industrialised societies with very low salt intake, while other results are for industrialised societies.

There is a correlation of 0.43 between blood pressure and sodium. However the graph makes it clear that there are really two clusters of results, one for the industrialised societies, and one for the non-industrialised societies. For industrialised societies there is a slight negative correlation, that is not however statistically significant. So what is one to make of these results?
The arguments have always been controversial:

- The body needs salt for proper functioning. Too little salt may be dangerous. What is the optimum level? Dahl's rats developed hypertension only when fed huge amounts of salt. The human equivalent would be 500 grams per day.
- Indigenous societies differ from industrialised societies in many ways, not just in their consumption of salt.

It is now widely accepted that the most valid evidence comes from randomised controlled trials that meet strict protocols. Intra-population studies are commonly (but

not universally) thought to be more valid than inter-population studies. Here is a summary of the results from these different types of evidence:

1. Some ecologic studies have shown big differences between populations, correlating with their salt intake. [It makes a lot of difference which data one focuses on – witness the difference between non-industrialised and industrialised societies in Fig. 2.]
2. Intra-population studies have generally been unable to show a link between salt intake and blood pressure.
3. An overall analysis (meta-analysis) that included 30 randomised trials and 48 unrandomised trials found a substantial effect. This study has been criticised for failing to distinguish between randomised and unrandomised trials.
4. A 1998 (Graudal et al.) meta-analysis (overall analysis) of randomised controlled trials showed a small effect, possibly too small to be of clinical importance. If however attention is limited to cross-over trials, there is a more substantial effect.
5. The Sacks et al.(2001) results seem to clinch the issue, in favour of a modest effect, greatest for individuals who do not improve other aspects of their diet.

It is with good reason that ecologic studies are widely regarded as unreliable. Almost inevitably, the populations differ in many respects. Thus, in the salt studies, the populations almost certainly differ in the level of intake of fruit and vegetables. Why focus on salt? Pretending that one is seeing the effect of salt alone may just be wishful thinking. A number of other effects are at work. The effects are *confounded*, i.e. the data do not allow you to separate them. One might say that confounding is a confounded nuisance! Confounding is a very serious problem in observational studies.

Societies that have high salt intakes are typically those that consume highly salted preserved foods. They consume these foods because they do not have access to fruit and vegetables. Thus, in the inter-population studies, the effects of salt are confounded with the effects of low levels of fruit and vegetable consumption.

Recently the DASH (Dietary Approaches to Stop Hypertension) collaborative research group has reported on a series of trials that investigated the use of a diet rich in fruit, vegetables and low fat dairy products (Appel et al. 1997). The blood pressure was reduced both for normal subjects and for mild hypertensives, slightly more for the latter. There was no reduction in salt consumption.

The most reliable evidence is undoubtedly that from carefully conducted clinical trials, conducted with controls. Such trials, on diet more generally as well as on salt, are now providing insight on the superficially contradictory results that have been obtained from other types of studies. Questions remain. Should one limit attention to trials that are double-blind, i.e. neither the patients themselves nor the staff administering the trial know who is on which diet? Are cross-over designs more or less reliable than the completely randomised design? As often, one has to sift out the more directly relevant and reliable sources of information, and use them to interpret less reliable and/or relevant sources of information.

If you want to read up on the salt controversy, a good place to start is the Taubes (1998) article in the journal *Science*. Taubes draws attention to the major overview studies, and presents the views of the main protagonists. A fair summary of the evidence may be that in Sacks et al. (2001). Graudal et al.'s (1998) meta-analysis of 58 trials of persons with high blood pressure and 56 trials of persons with normal blood pressure had found that the effect is relatively small, and did not justify a recommendation, in the population generally, to reduce salt intake. However the Sacks et al. study broke new ground by controlling for other aspects of the diet,

notably the intake of fruit and vegetables, and thus carries greater conviction than these earlier studies. The effect of low sodium intake was, interestingly, greatest for those who did not control other aspects of their diet. This may be an illustration of Fisher's dictum that Nature will reveal her secrets only if we ask her more than one question at the same time. She may refuse any answer on the question of effects from salt unless we ask her also about the intake of fruit and vegetables.

In reviewing the literature you need to be aware of the strengths and weaknesses of different types of study. In planning your own study, you need to know the strengths and weaknesses of the alternative designs that are available to you. Additional issues arise when there are multiple studies.

3.2 Different Types of Study – Further Examples

The simplest kind of randomised experiment has a treatment and a control group, with a randomisation device used to make the assignment to treatment or control. Natural events can create the conditions of a randomised experiment. For example, in a local area of a city, buildings are constructed according to several different designs. Some survive an earthquake, while some do not. The only consistent difference between buildings that survive and those that do is in design.

In the earthquake example, it was after the earthquake (a natural intervention) that the different treatments were identified. In some instances it will be clear what aspects of building design or land features have favoured survival. In other cases, it may not be so clear. Is it the design of the foundations or of the superstructure that is crucial? Is the local geology an issue? There is rarely the same clarity of connection between effect and cause as in an experiment. Similar issues arise in studying the effects of a natural event or an accident on a wildlife habitat. Also, rather than a natural intervention, there may be a government intervention – perhaps a change in management regime.

Here are some of the possible types of study that investigate effects of an intervention on a wildlife habitat:

1. Gather observational data from a number of sites, spanning a range of management regimes. Use the data to determine conditions that lead to favourable outcomes.
2. Before/After studies of effects of management or natural changes (e.g. flooding) or accidents (e.g. oil spills).
3. Compare sites subject to natural changes (e.g. flooding) or accidents (e.g. oil spills), with comparable sites where they has been no intervention used as controls.
4. Study experimentally induced changes, with different management regimes applied (*by managerial choice*) to different sites.
5. Study experimentally induced changes, with different management regimes assigned (*at random*) to different sites.

We noted earlier the study that compared employment effects in one state where there had been a change in the minimum wage requirements, with those in a neighbouring state where there had been no change. The neighbouring state where there had been no change was used as a control. But what if we have an intervention (a change in minimum wage requirements), but no control? Can we mount a before/after argument? Here are summaries of the range of possible studies, for the study of minimum wage legislation:

1. Use U. S. national monthly data to study the effects of increases in the Federal minimum wage on April 1 1980 and April 1 1981. (Deere et al. 1995).
   [NB there was no control group.]
2. Panel study of state minimum wage changes, 1973 – 1989. (Neumark and Wascher 1992).
   [Horizontal comparisons across states, at one time, rely heavily on analytic models and numeric adjustments.]
3. Compare New Jersey (where there was a change in the minimum wage) with nearby Eastern Pennsylvania (where there was no change). (Card and Krueger 1994).
   [NB Control was chosen by the investigator. We have only one comparison between `treatment' and `control'.]

In an example (Freedman 1999) from the early history of investigations into the health effects of smoking, cases were persons admitted to hospital after diagnosis with lung cancer. Controls were patients admitted for other reasons. In such case-control studies, it is the outcome (lung cancer or not) that determines who will be in the study. The investigator then peeks to see what treatment the patient received.

3.3 The Eberhardt and Thomas Classification

Eberhardt and Thomas (1991) have a comprehensive classification that is intended for ecological studies. Their primary classification focuses on the level of control that the observer is able to exercise. This control is greatest in an experiment. Given that an experiment is planned, how will this control be exercised? The most secure results are from various forms of randomised experiment, such as we will consider in later chapters.

Where there has been a distinct perturbation, such as from a natural event (a flood, or an earthquake, or a volcanic eruption), this may sometimes closely mimic the conditions of a randomised experiment. Equally, it may not. Each case must be argued on its merits.

The following is a modification of Eberhardt and Thomas's classification:

- Events controlled by observer
  - ⇒ Randomised experiment
    (with/without controls) × (with/without replication) etc.
  - ⇒ Unrandomised experiment (includes haphazard assignment)
    (with/without controls) × (with/without replication) etc.
- Study of uncontrolled events
  - ⇒ Distinct Perturbation Occurs
    - ◊ Intervention analysis
  - ⇒ Distinct perturbation usually not evident
    - ◊ Domain of study restricted
      - o Assessment involving that restricted domain
        (i.e. not a random or other sample from the whole domain of interest; sampling frame is not the whole of the target population)
    - ◊ Sampling over entire domain of interest
      - o Analytical sampling
      - o Descriptive sampling
      - o Sampling for Pattern

There is a great deal more that might be said. Sampling issues arise in experimental as well as in non-experimental studies.

3.4 What Types of Study Should You Use?

Here is a list of research questions. What type of study would you use in each instance?

1. Compare consumer perceptions of 30 different chocolate formulations.
2. Assess the effectiveness of a method for 'cleaning' soil that is contaminated with heavy metals.
3. You are considering two advertising strategies for a new product. You want to determine which is likely to be more effective.
4. Assess the likely effect of proposed changes to plant quarantine requirements for produce imported to Australia.
5. A farm advisory service wishes to compare the relative effectiveness of two training programmes for farm staff involved in handling agricultural chemicals. What type of study is likely to give a good comparison?
6. A firm that offers metal turning services intends to mount a programme to improve the safety awareness of staff involved in handling lathes. It has a large number of widely scattered small manufacturing units. It wishes to determine the best strategy?
7. Assess the market, in the Canberra area, for a mass-produced small sailing craft.
8. Determine market niches that present supermarkets in the Canberra area are not filling.
9. You are a high school principal. What statistical information for the school's catchment area would be useful for your planning of the school's future development? How much of this information is available from school records or from official sources such as the Department of Education or the Department of Statistics? What information could you usefully get from a survey? Plan accordingly.
10. Since 1987 the British government has installed closed circuit TV cameras in a number of city and town centres throughout Britain. Set up a study that will determine whether these have been effective. [New Scientist, 23/30 Dec 1995, p.4].
11. Assess how the pattern of demand for hospital services is likely to be affected by a proposed change to services offered by public hospitals.
12. Set up a study to examine the implications of the varying prescription patterns of GPs for the quality of patient care and for medical costs.
13. You have been asked for advice on a study for determining whether calcium antogonists reduce the risk of stroke in patients with heart disease. How should you proceed?
14. Are male sperm counts declining in Australia? How might you set up an Australian study?
    [See New Scientist, May 11 1996, p. 10].
15. Are home births any more dangerous than hospital births?
    [See New Scientist, May 11 1996, p. 5].
16. Bricks are to be fabricated from waste plastic and wood chip. How would you determine the optimum particle size, baking temperature, baking time and % plastic?
17. You are asked for advice on what sorts of studies are needed to decide once and for all the dietary effects of salt. Is one individual study likely to be useful?

Should the focus be on careful evaluation of existing data, or on a new study? What advice would you give? Bear in mind that most research to date has focused on effects on blood pressure. Are there other effects of changes in dietary salt that ought to be a concern?
[The answers are not at all obvious. They are, though, good questions to think about.]

18. You are asked for advice on the validity of the evidence that Dianne McGuinness (1997) presents in her book *Why Our Children Can't Read*. What would be a good way to proceed? How long will you need? What help will you need?
19. A private health provider is responsible for 20 hospitals. It plans to move to a new funding and management regime. Before making the change, it wants to be sure that the changes will work and will be an improvement. Would you recommend moving some of the hospitals to the new regime on an experimental basis?
20. You have read the book *Smart Health Choices* (Irwig et al., 1999). You applaud the encouragement that it gives to patients to ask clinicians probing questions about their treatment choices. But will clinicians be able to respond well to such demands? Design a study to answer this question.
21. What are the pros and cons of screening for prostate cancer? [See e.g. Irwig et al. 1999; Moynihan 1998].
22. Consider the design of a study of the effects of changing sociological and political forces on taxation regimes in the Commonwealth of Australia since Federation?

References and Further Reading

Beveridge, W. I. B., 3rd edn 1957. The Art of Scientific Investigation. William Heinemann Ltd., London.

Eberhardt, L. L. and Thomas, J. M. 1991. Designing environmental field studies. Ecological Monographs 61: 53-73.

Irwig, J., Irwig, L., and Swift, M. 1999. *Smart Health Choices. How to make informed health decisions*. Allen and Unwin, Sydney.

Manly, B. F. J. 1992. The Design and Analysis of Research Studies. Cambridge University Press.

McGuinness, D. 1997. Why our Children Can't Read. The Free Press, New York.

Moynihan, R. 1998. Too Much Medicine. Australian Broadcasting Corporation.

Salt and Hypertension

Appel, L. J. et al. 1997. A Clinical Trial of the Effects of Dietary Patterns on Blood Pressure. The New England Journal of Medicine 336: 1117-1124.

Dahl, L. K. 1960. Possible role of salt intake in the development of hypertension. In Cottier, P., Bock, K. D., eds. Essential Hypertension – an International Symposium, pp. 53-65. Springer-Verlag, Berlin.

Dahl, L. K. 1970. Salt and Hypertension. American Journal of Clinical Nutrition 25: 231-244.

Graudal, N. A., Galloe, A. M., Garred, P. 1998. Effects of sodium restriction on blood pressure, renin, aldosterone, catecholamines, cholesterols, and trigliceride. Journal of the American Medical Association 279: 1383-1391.

Intersalt Cooperative Research Group. 1988. Intersalt: an international study of electrolyte excretion and blood pressure: results for 24 hour urinary sodium and potassium excretion. British Medical Journal 297: 319-328.

Sacks, F.M., Svetkey, L.P., Vollmer, W.M., Appel, L.J., Bray, G.A., Harsha, D., Obarzenek, E., Conlin, P.R., Miller, E.R., Simons-Morton, D.G., Karanja, N., and Lin, P.-H. 2001. Effects of

blood pressure on reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. New England Journal of Medicine 344: 3-10.

Taubes, G. 1998. The (political) science of salt. Science 281: 898-907 (14 August).

Smoking and Health

Freedman, D. 1999. From association to causation: some remarks on the history of statistics. Statistical Science 14: 243-258.

4. Experimental Design

> The statistical tools of experimental psychology were borrowed from agronomy, where they were invented to gauge the effects of different fertilizers on crop yields. The tools work just fine in psychology, even though, as one psychological statistician wrote, "we do not deal in manure, at least not knowingly." The power of these tools is that they can be applied to any problem – how color vision works, how to put a man on the moon, whether mitochondrial Eve was an African – no matter how ignorant one is at the outset.
> [Pinker, S. 1997. How the Mind Works, p.303. Norton, New York.]

> The methods of science, with all its imperfections, can be used to improve social, political and economic systems, and this is, I think, true no matter what criterion of improvement is adopted. How is this possible if science is based on experiment? Humans are not electrons or laboratory rats. But every act of Congress, every Supreme Court decision, every Presidential National Security Directive, every change in the Prime Rate is an experiment. Every shift in economic policy, every increase or decrease in funding for Head Start, every toughening of criminal sentences is an experiment. Exchanging needles, making condoms freely available, or decriminalizing marijuana are all experiments. . . . In almost all these cases, adequate control experiments are not performed, or variables are insufficiently separated. Nevertheless, to a certain and often useful degree, such ideas can be tested. The great waste would be to ignore the results of social experiments because they seem to be ideologically unpalatable.
> [Sagan 1997, The Demon-Haunted World, pp. 396-397. Headline Book Publishing, London.]

> *Experiments may answer questions you never thought to ask! Experiments teach by experience. Receptive and trained minds will learn more. Different applications have different needs.*
> *There is no more effective way to settle a disputed question than to do an experiment, when an experiment is possible. When fire-walkers walk across hot charcoal and emerge unharmed, it demonstrates that such a thing is possible. When one plant grows like crazy in a bed of compost, while its neighbour has no compost and wilts, it seems a convincing demonstration that compost helps growth. It seems convincing even though this is a rather poorly designed experiment.*
>
> *Not all questions lend themselves to experimentation. There is an accordingly greater challenge to design a study whose answers will be compelling. It will, usually, then be more difficult to reach firm conclusions. Thought experiments may often help understanding.*
>
> *The aim of experimental design is to ensure that the experiment can detect the treatment effects that are of interest, uses available resources to get the best precision possible. The choice of design can make a huge difference.*

The account that I give here will, as in the case of much else that this monograph touches on, be introductory. My aim is to give the flavour of experimental design, as it applies to a number of different application areas.
Francis Bacon (1561-1626) gives an early example of a controlled experiment. He applied five different treatments to wheat seeds – water mixed with cow dung, urine, and three different wines. The winner was urine, followed by the cow dung. By the standards of modern experimental design, Bacon's experiment was inadequate. It was not randomised, i.e. he did not use a random mechanism for assigning seeds to treatments.

Very simple experiments vary just one factor at a time. Indeed there are still experimenters who regard this as the proper strategy. Where there are multiple factors, the one-factor-at-a-time approach makes it very difficult to detect interactions. If there are no interactions, it may work reasonably well, but is inefficient.
Multi-factor experiments allow the detection of interactions. Degrees of freedom that are associated with any interactions that prove to be negligible are available for improving the precision of the standard deviation estimate. So the experimenter wins both ways. For purposes of estimating main effects, a single four-factor experiment is in general far more efficient than four single factor experiments. It will give the same accuracy with a much smaller use of resources.

4.1 Experimental Design Issues

We wish to compare two technicians who will use a pressure tester to compare apple firmness. How should we do the comparison? Should we give the testers separate samples of perhaps twenty apples? Or should we use one sample of twenty apples, with both technicians making firmness measurements on each apple?
In a clinical trial that compares two different therapies for treating arthritis, right and left hand grip strength will be among the outcome measurements. The measurements are highly variable. Is it useful to increase the precision by making repeated grip strength measurements? Or is the variation in measured grip strength for an individual patient of minor consequence relative to variation between patients? If it turns out to be useful to make repeated measurements on individual patients, should the repeat measurements be made at the same session, or at different sessions that are separated by a few days or weeks?
We plan an experiment in which trays of fruit are the experimental unit. In each of several cool-stores, different treatments will be applied to different trays. Should we opt for many trays with a small number of fruit on each, or for a small number of trays with a large number of fruit on each? Which is the better design? As the treatments are applied to whole trays, increasing the number of trays always increases the precision. Increasing the number of fruit per tray may or may not make a useful contribution to increasing precision. All depends on how fruit to fruit variation within a tray compares with between tray variation.

4.2 Randomised Controlled Trials

> What makes it possible to write a long article on controversies in controlled clinical trials without writing a much longer article on uncontrolled trials or uninvestigated therapies? Essentially this paradox arises because in controlled trials we have a model of perfection and we can discuss departures from it with ease. Without such a model, one tends to emphasise only major difficulties --- having swallowed a camel, why strain at a gnat?
> [Mosteller, Gilbert & Lewis, p. 14, in Shapiro & Lewis 1983.]

Randomised controlled trials are a good setting in which to consider a number of elementary aspects of experimental design. By contrast with agricultural experimentation, the design for a randomised controlled trial is often very simple in concept.
Where there are two treatment groups, subjects are randomly assigned to one or other treatment, and the result determined. Complications arise from the ethical and logistical difficulties of conducting a properly designed clinical trial.
A minor elaboration of the two-sample trial arises when subjects are matched, or when treatment comparisons can be made within subjects. In this case it may be possible to perform the analysis on the difference between the responses or on

log(ratio) of the reponses, or on some other measure of the difference. The analysis then reduces to a single sample analysis.
There are numerous examples of interventions that were introduced without first doing an experiment, and where the intervention was later shown to be harmful. Hormone injections in pregnancy were at one time thought to prevent miscarriage. A randomised controlled trial showed no effect, compared with placebo injections. Moreover this unproved therapy later proved to give an excess of cases of vaginal carcinoma and of breast cancer (Christie et al. 1987; Gehan & Lemak 1994, p.159). Section 5.1 gives initial data from this study.
Randomised controlled trials where there is matching provide a simple example of a block design. The individuals who are matched form a single block. Another form of matching arises when the different treatments are applied, in turn, to the one patient. The issue of whether there is treatment carry-over is then important. Also one has to design the trial so that changes over time can be distinguished from the treatment effect.

4.3 A Simple Taste Experiment

Consider a taste experiment, where a number of panellists assess the sweetness of two different milk products. They mark off their responses on a so-called Likert scale, thus:

Not sweet enough Too sweet
1 3 x 5 7 9

The investigator uses a ruler to read off the results. One way to make this easy is to place the 1 at 10mm, the 30 at 30mm, and so on. The `x' is at about 36mm. A reasonable way to do the experiment is to give each person both products. Here then is a set of results (shown as mm) from such an experiment:

```
   Person 1  2  3  4   5  6   7  8  9 10 11 12 13  14 15 16 17
4 units  72 74 70 72  46 60  50 42 38 61 37 39 25  44 42 46 56
1 unit   58 69 60 60  54 57  61 37 38 43 34 14 17  54 32 22 36
Diff.    14  5 10 12  -8  3 -11  5  0 18  3 25  8 -10 10 24 20
```

Fig. 3 shows the data graphically

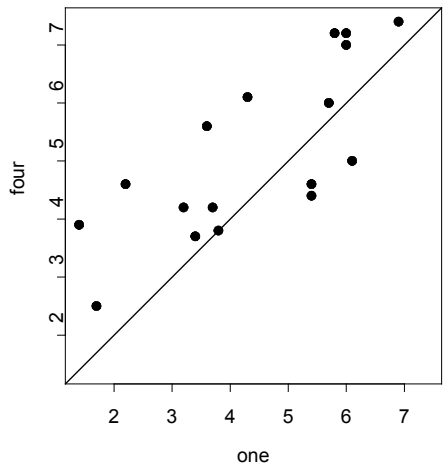


Fig. 3: Perceived sweetness for sample with four units of additive, versus perceived sweetness with one unit of additive

The diagonal line shows where the assessments for the two samples would be equal. Notice that tasters who give a higher assessment for the sample with one unit of additive also tend to give a higher assessment for the sample with four units of additive. The differences in the two assessments are relatively consistent.
The individual tasters have the role that blocks would have in an agricultural field design. Each taster compares the two treatments. In the field design, the two treatments are placed alongside in the one block. The design can easily be extended to allow a comparison with, for example, milk with no additive. There would then be three treatments per taster.
Experimental design questions that one might ask include:

1. What are the pros and cons of the above experiment, as against an experiment where 34 tasters were divided randomly into two groups of 17. Tasters in the first group all received the milk with one unit of additive, while those in the second group received the milk with four units of additive.
   [This alternative experiment would be a very imprecise experiment. Differences between tasters would introduce unwanted noise into the comparison between amounts of additive.]
2. What is the best way to improve accuracy? It is easier and cheaper to get each taster to repeat the comparison a number of times, rather than to bring in new tasters.
   [If individual tasters are highly consistent from one occasion to another, relatively to variation between tasters, it will not help much to get each taster to repeat the comparison a number of times. Increasing the number of tasters will always, in theory, improve the expected precision.]

4.4 The Principles of Experimental Design

The Three Rs

Randomisation, replication and blocking are often identified as the three chief principles. We will take them in the reverse order. Blocking, whereby within each block treatments are compared under conditions that are as similar as possible, is a device for reducing variability. Pairing – for example one treatment might be applied to one leg of a patient and the other to the other leg – is a simple form of blocking.
Replication reduces variability and ensures that there will be a valid estimate of experimental and other error. Note the contrast between replication of experimental units and repeated measurements on the same experimental unit. Repeated measurements on an experimental unit increase the accuracy for that unit. One still has only the one experimental unit.
Randomisation aims to balance out the effects of factors that are not amenable to experimental control. It does this by making chances equal. It does not ensure that treatment groups will be balanced with respect to these uncontrollable factors, only that the chances are equal.

Multiple Measurements on Each Experimental Unit

Consider an experiment where the individual apple is the experimental unit. Measurements of the amount of sugar (the "brix") may be inaccurate. So several measurements are taken on each apple. Note that while this increases the precision of the result for each apple, it does not increase the number of experimental units! There are no more apples than before!

The analysis can work with the means for each apple. Note that once the variability in the mean for an apple is negligible relative to variation between apples, there is no point in taking additional measurements on each apple.
This principle can be extended:

1. The experimental unit is a tray of apples (they all go into the store together), and the experimenter wonders how many apples to put on a tray. There's no point in getting results from further apples once the accuracy for the tray is small relative to variation between apples.
2. In a clinical trial, several clinicians may assess each patient. The experimental unit is the patient, not an assessment on a patient. Making more assessments on each patient is quite different from going out and finding further patients.

The Role of Experiments

Not all questions are suited to direct experimental investigation. No-one has yet devised a way to directly compare the effects of releasing different amounts of CO_2 into the earth's atmosphere. It is easy to imagine a fictional galactic empire where there are six earth-like planets that can, providing one books far enough in advance, be made available for climatological experiments. For better or for worse, all we can do is imagine and write science fiction about such an empire. In the world that we inhabit, we have just one planet at our disposal and controlled experiments are not a possibility.
Even though data have not come from an experiment, they may be analysed as though they had. Statistical models assume that data have been generated under ideal conditions that really only hold, if at all, in a very careful experiment. It is helpful to think about what sort of experiment might have generated the data, what the limitations of that experiment are, and where the potential for bias lies. Such thought experiments can help clarify assumptions.
Do not expect that one experiment will settle all outstanding issues. Experiments are a structured way to learn by experience. As Fisher (1960, §12.1) said

> . . . . in learning by experience, or by planned chains of experimentation, conclusions are always provisional and in the nature of progress reports, interpreting and embodying the evidence so far obtained.

Those who have a trained and receptive mind will learn more. Experimenters will do well to be receptive to the possibilities that

1. The experiment may challenge the assumptions that lay behind its design, perhaps even indicating that the research question was not entirely appropriate.
2. Having learned from your initial experience, it may be possible to do a better experiment next time. (So it is often unwise to blow all resources on one experiment!)

An advantage of a carefully designed experiment is that it is likely to teach the experimenter something, even if the experiment asked the wrong question! I have referred several times to an experiment that compared mechanical with air suspension on trucks used to transport apples. We had asked a question that related to truck suspensions. We learned instead about the damage due to unstable bins.
Note finally that different areas of application may require quite different styles of experiment, and may raise quite different issues.

The Language of Experimental Design

Important ideas and distinctions are:

- treatment units and measurement units. They may not be the same!

- randomisation, especially as opposed to haphazard assignment of treatments
- replication – genuine replication, effective replication and bogus replication
- blocking and other forms of local control
- levels of variation.

We begin with brief comments on ideas of treatment unit, measurement unit and blocking. A discussion of randomisation and replication will then follow.

Multiple Levels of Variation – Blocks

Multiple measurements on an experimental unit, e.g. multiple measurements on the one apple, increase the precision for the experimental unit. They do not increase the number of experimental units – additional apples if treatments are applied separately to each apple or additional patients if treatments are applied separately to each patient. Observations can be grouped within an experimental unit.

One can also group experimental units into blocks. Blocks then become another, now higher, level of variation. The simplest type of one-factor block design, the randomised complete block design, has one experimental unit from each of the treatment levels in each block, e.g.

	Block 1	Block 2	Block 3
Treatments	A, B, C	A, B, C	A, B, C

N. B. Treatments should be randomly allocated to experimental units, independently for each block.

Also possible are block designs where a subset of the treatments appear in each block. For example, we might have

	Block 1	Block 2	Block 3
Treatments	A, B	B, C	C, A

One treatment has been left out in each block, in a balanced way. This is a *balanced incomplete block* design. I have used this type of design for comparing the readings of different firmness testing devices on the same fruit. Each fruit was in effect a block. We did two sets of two readings, one pair with each of the devices, on the one fruit.

Block designs are widely used in agriculture, where the aim is to maximise the precision of treatment comparisons. Thus each block is chosen to be as uniform as possible. In the simplest form of randomised block design, all treatments occur once in each block. Blocks should be sampled from the wider population to which it is intended to generalise results, so that they might be on different sites.

In controlled climate chambers, each chamber may form a block, with one or more units from each treatment in each chamber. Or if there are differences between trays in a chamber, each tray may form a block.

In clinical trials blocks are more often used as a way of making it hard to predict treatment allocations for individual patients. Allocation of treatments to patients is random within blocks, subject to devices that achieve a roughly equal numbers in the different treatments. (ICH 1998, p.21). Where a surgical trial involves several different surgeons, blocking may be highly desirable as a mechanism for controlling variation. The patients that are allocated to a surgeon form a block, with random allocation to treatments within those blocks.

Randomisation

Randomisation prevents intentional or unintentional favouring of one treatment over another. It is also a way to ensure that observations are all drawn, independently, from the same distribution. Haphazard allocation, where the experimenter allocates treatments in any unsystematic way that seems right, is not randomisation.

Replication

Genuine replication increases the number of treatment units. Where there are blocks, there is a choice between increasing the number of blocks, and increasing the number of experimental units in each block. Increasing the number of observations on each experimental unit, while it is often a good idea, is not genuine replication.

4.5 Confounding

Experiments can and should be designed so that they are capable of revealing the effects of the factors and factor combinations that are of interest. In observational studies there may be such limited control over the design that it is impossible to separate effects out in this way. Some confounding is almost inevitable.

For example, two measuring instruments that are believed functionally identical may be set differently. If one instrument is used for measuring results from treatment A, and the other for measuring results from treatment B, the effect of the treatment is confounded with the effect of the instrument. Or if one technician assesses results from treatment A, and another technician assesses results from treatment B, there may be a technician effect that is confounded with the treatment effect.

The simplest form of experimental confounding occurs when two factors change together. High correlations between pairs of variables, common in observational studies, provides an indication that it will be difficult to separate their effects. Contrast this with the way that experiments vary factor levels under the control of the experimenter, to ensure that they do not change together.

Suppose we have two factors – level of lime, and level of phosphate. The following three designs illustrate the three different possibilities. An x indicates that a particular combination of factor levels is present.

	Phosphate(kg/ha)				Phosphate(kg/ha)				Phosphate(kg/ha)			
Lime(kg/ha)	0	10	40	400	0	10	40	400	0	10	40	400
0	x		x	x	x				x	x		
1000						x			x	x		
2500	x		x	x			x				x	x
8000	x		x	x				x			x	x
	No correlation				Factors confounded				Correlation			

Table 1: Three possible treatment allocations. An x is used to denote a treatment combination that is included in the experiment.

The first design is much preferable to the third. The same selection of levels of phosphate appears for each different level of lime. The second design does not allow any possibility for separating the effects of lime from those of phosphate. It is a hopeless design, unless one already knows the optimum ratio of phosphate to lime.

In clinical trials, age or sex may be a confounding factor. Suppose one has

	Treatment A	Treatment B
Females	7	15

Males	9	3

Then gender is a confounding factor for purposes of making treatment comparisons. The treatment A results will be slightly biased to the results for females, while the treatment B results will be relatively similar to the results for females[14].

Other examples of confounding

Why did doctors continue to practice bloodletting for so long? Most conditions will get better of their own accord, given time. In addition there was, for some patients at least, a placebo effect. The effects of the bloodletting, that were often harmful, were confounded with the effect of time and with the placebo effect.

We have already noted that confounding is the bane of observational data. It may also be the bane of studies where there is an intervention, but no control group (with random assignment) with which to compare the treatment group.

Was New Zealand's introduction of iodised salt really the cause of a dramatic reduction in goitre problems? Or would the problem have disappeared anyway because children had been getting iodine in school milk? As everyone received iodised salt, once it was introduced, it is impossible to be sure. Silverman (1985) gives numerous other such examples.

4.6 Experimental Design – Books for Further Study

The classical text is Fisher (1935), which has been through many editions. It is elementary in style, and remains one of a small number of books that can be recommended to the non-specialist. Other definitive texts are Cochran and Cox (1957), Cox (1958) and Box et al. (1978).

Different application areas differ in the types of design that find predominant use. In specific applications, there will be a range of practical issues that require attention. Robinson (2000) is attractive for the way that it combines attention to such practical issues with attention to the theory as and when it is necessary. Examples are drawn from many application areas, with a focus on industrial applications. For field experimentation, see Mead (1988), Petersen (1985), Pearce et al. (1988), and Williams and Matheson (1994). See also the very brief discussion of experimental design in Maindonald (1992). The manual for the statistical package Genstat (Payne et al. 1993) has helpful discussions of designs that are common in field experimentation. For clinical trials, Piantadosi (1997) and Silverman (1985) are particularly good. See also other books that are noted in the references.

References and Further Reading

Andersen, Bjorn 1990. Methodological errors in medical research: an incomplete catalogue. Blackwell Scientific.

Armitage, P. and Berry, G., 2nd edn. 1987. Statistical Methods in Medical Research. Blackwell Scientific Publications, Oxford.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639.

[14] These are the numbers in a trial that is reported in Gordon, N. C. et al.(1995): `Enhancement of Morphine Analgesia by the $GABA_B$ against Baclofen. Neuroscience 69: 345-349. Treatment A was Pantazocine plus placebo, while treatment B was Pantazocine plus Baclofen. When the data were analysed to take account of the gender effect, it turned out that the main effect was a gender effect, with a much smaller difference between treatments.

[The checklist that appeared as part of this statement can be found at: http://www.ama-assn.org/public/journals/jama/jlist.htm]

Box, G.E.P., Hunter, W.G., and Hunter, J.S. 1978. Statistics for Experimenters. Wiley, New York.

Cochran, W.G. and Cox, G.M. 2nd edn. 1957. Experimental Designs. Wiley, New York.

Cox, D.R. 1958. Planning of Experiments. Wiley, New York.

Fisher, R.A. [1935], 7th edn. 1960. The Design of Experiments. Oliver and Boyd.

Gehan, E. A. and Lemak, N. A. 1994. Statistics in Medical Research. Plenum Medical Book Company, New York.

ICH 1998. Statistical Principles for Clinical Trials. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Available from http://www.pharmweb.net/pwmirror/pw9/ifpma/ich1.html

Maindonald J H 1992. Statistical design, analysis and presentation issues. New Zealand Journal of Agricultural Research 35: 121-141.

Payne, R.W., Lane, P.W., Digby, P.G.N., Harding, S.A., Leech, P.K., Morgan, G.W., Todd, A.D., Thompson, R., Tunnicliffe Wilson, G., Welham, S.J. and White, R.P. 1993. Genstat 5 Release 3 Reference Manual. Clarendon Press, Oxford.

Pearce, S.C., Clarke, G.M., Dyke, G.V. and Kempson, R.E. 1988. Manual of Crop Experimentation. Griffin, London.

Peterson, R.G. 1985. Design and Analysis of Experiments. Marcel Dekker, New York.

Piantadosi, Steven. 1997. Clinical Trials: A Methodologic Perspective. Wiley 1997.

Robinson, G.K. 2000. Practical Strategies for Experimenting. Wiley, New York.

Schulz, K. F. 1996. Randomised Trials, Human Nature, and Reporting Guidelines. Lancet 348: 596 - 598.

Schulz, K. F., Chalmers, I., Hayes, R. J. and Altman, D. G. 1995. Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials. Journal of the American Medical Association 273: 408 - 412.

Shapiro, S. H. and Lewis, T. A. 1983. Clinical Trials. Issues and Approaches. Marcel Dekker 1983.

Silverman, W. A. 1985. Human Experimentation. A Guided Step into the Unknown. Oxford University Press, Oxford.

Williams, E.R. and Matheson, A.C. 1994. Experimental design and analysis for use in tree improvement. CSIRO Information Services, Melbourne.

5. Quasi-Experimental and Observational Studies

As noted in the previous chapter, the only sure way to know the effect of one or other change is to make the change and see what happens, i.e. do an experiment. However there are severe practical and ethical limits on what experiments are possible. Hence the various quasi-experimental methods that exercise non-experimental forms of control on the generation of data. Or the conditions of an experiment may be created by an accident of management or of nature.

Even though we do not have an experiment, is it sometimes (or often) possible to get data that we can treat, to a greater or limited extent, as though it had come from an experiment? This section will explore several types of study that aim to do just that. As we will see, there can be severe obstacles to reliable inference from such studies. Data from quasi-experimental studies are commonly analysed as though they had been gathered under experimental control. If the mechanisms that generated the data closely mimic those of a genuine experiment, this makes good sense. Where the data have few of the characteristics of experimental data, inferences that rely on statistical models are in general hazardous. The literature offers guidelines, arising from work such as I will discuss below, that careful researchers will study and use.

I will start with those types of study where there is the greatest potential to reproduce the conditions of an experiment, moving through to those farthest from the conditions of an experiment. The examples are all from clinical medicine. The following section discusses the use and limitations of regression modelling. It uses examples from the economics literature. Getting results that are credible and defendable, if this is possible, is not a simple matter of running data through a multiple regression program!

5.1 Some alternative types of non-experimental study

Accidents of Nature or Human Behaviour

Occasionally, an accident of nature or of human behaviour – an earthquake or an oil spill – creates conditions close to those of an experiment. There is a clear intervention. In the case of the earthquake we might be interested in comparing new building design with the old design, where the old design may be regarded as the control. For the oil spill, we will want to compare affected areas with comparable unaffected areas, preferably over a number of spills.

We noted how that in 1853 the Lambeth company in London had moved its water supply upstream. This was an intervention that closely mimicked a genuine experiment. One problem is that other changes were very likely made at the same time. Some pipes would have been replaced. So the related observations that Snow made were an important part of his evidence.

Cohort Studies (Longitudinal studies; retrospective studies; follow-up studies)

A key feature of experimentation is the control that the experimenter exercises over the way that levels of the different factors combine to affect the response. Retrospective longitudinal and case-control studies retain some elements of this control. At the same time, they have some of the features of an observational study. The health experience of one or more groups of people, often an exposed and a non-exposed group, is followed over some period of time. For example, the aim may be to

compare the health experience of doctors who smoked at the point of entry to the study with the health experience of those who did not. The doctors were not randomly assigned to a smoking and a non-smoking group! So there might be something different about the doctors who smoke, affecting both their health experience and their tendency to smoke. Much of the work on the health effects of smoking has been directed to ruling out such explanations.

Case-control studies

Again we wish to assess the effects of an exposure. Case-control studies aim, by the choice of 'cases', to exercise an 'after the event' control that as far as possible substitutes for direct experimental control. Those subjects who have the disease are 'cases', while the 'controls', chosen from the same population as the cases, do not have the disease. We classify both cases and controls as exposed or unexposed. The estimation of relative risk relies on cases and controls being representative of cases and controls in the community, with no regard to the likelihood of exposure or non-exposure. Depending on how subjects are selected, such associations are common. Persons known to have been exposed, and therefore thought more likely to be cases, may be more likely to find their way into hospital records.
Occasionally, case-controls involving quite small numbers of patients provide highly convincing evidence. Adenocarcinoma of the vagina in young women had been recorded rarely before it was diagnosed in eight patients treated in two Boston hospitals between 1966 and 1969. Each of the eight patients was matched with a female born nearest the time of the patient and from the same service. Seven of the eight mothers of patients with carcinoma had received diethylstilbestrol (DES), starting during the first trimester. No control mother had been given the synthetic estrogen. Thus we have

	With cancer	Without cancer
Mother had not taken DES	1	8
Mother had taken DES	7	0

In seven of the pairs, the mother of the daughter with carcinoma had taken DES, while the other mother had not. In one of the pairs, neither mother had taken DES. The probability that there will be this discordance in seven or more pairs out of 8, if the split between No DES and DES is equally likely to go either way, is 0.004. (Gehan & Lemak 1994, pp.158-159.[15])

Cross-sectional Studies

Essentially a cross-sectional study is a type of survey. It shows a current reality – the prevalence of smoking or the prevalence of lung cancer. It does not tell us incidence – the rate at which people are taking up smoking or getting lung cancer. There is no time dimension. Moreover there is a survivor effect – the only people who can be asked questions are those who are available to be asked. Christie et al. (1987) quote the (fictitious?) example of stopping all motor-cycle riders and every tenth car driver on a freeway and asking whether they have had a serious accident, requiring hospital admission, in the past 12 months. The rate among car drivers is found to be twice that among motor cyclists. Serious accidents may be more likely to kill motor-cyclists. Or

[15] It is not appropriate to apply a chi-squared test to the two-way table. Such an analysis ignores the pairing, and would be wrong.

perhaps, following a serious accident, many motor-cyclists give up their motor-cycles and become car drivers.

Case-Control versus Long-Term Follow-Up —An Example

Table 1 illustrates a limitation of case-control studies. It has data from a long-term follow-up study of patients who had undergone surgery for gastric cancer[16]. Patients whose cancer was detected by mass screening are compared with an unscreened group who presented at a hospital or doctor's surgery with gastric cancer.

	Number	5 year mortality
Unscreened Group	352	41.9%
Screened Group	308	28.2%

Table 1: Comparative 5-year mortality, between screened and unscreened groups, of patients who had undergone surgery for gastric cancer.

The data suggest a better prognosis for the group whose cancer is detected as a result of screening. However there are at least two differences between the screened and the unscreened group:

1. It is possible that some in the screened group would never have presented at a clinic; some of these cancers may stay dormant;
2. The screening will detect cancers at an earlier stage. Even without treatment these patients should survive longer than those whose cancer is detected, almost inevitably at a more advanced stage, when they present at a medical service.

Because the process that led to the detection of cancer was different between the screened and unscreened groups, the two groups are not comparable. The method of detection is a confounding factor. The screening may lead to surgery for some cancers that would otherwise lie dormant for long enough that they would never attract clinical attention.

One needs a longitudinal study that compares all patients in a screened group with all patients in an unscreened group. Table 2 presents results from such a study (c.f. Hisamuchi et al. 1991)

	Number	Mortality over 1960 - 1977
Unscreened Group	2683	95/100,000 p.a.
Screened Group	4325	45/100,000 p.a.

Table 2: Comparative 5-year mortality, between screened and unscreened groups, of patients who had undergone surgery for gastric cancer.

[16] The data appeared in Sugawara et al.(1984), in Japanese, in a paper of which I have no other details.

Evidence for Bias in Non-Experimental Studies

Earlier I drew attention to evidence that if clinical trials do not follow accepted standards for randomisation and concealment, then biases will result. Non-experimental studies offer even greater opportunities for bias. Petitti (1994, p.76, Fig. 6.1) refers to a study by Stampfer & Colditz that compared different types of study of post-menopausal estrogen use and coronary heart disease. Hospital case-control studies gave a higher relative risk than other types of study. The next highest risk estimates came from population case-control studies. Cross-sectional studies, and various prospective control studies, gave the lowest risks. See also Andersen (1990).

Experimental versus non-experimental studies

Non-experimental studies are useful in drawing attention to possible associations. Their results are in general compelling only when two or more of the following conditions are satisfied (1) the conditions closely mimic those of an experiment or (2) the effect is large and has no other plausible explanation (3) there are multiple confirmatory sources of evidence.

Smoking was blamed for lung cancer because most cases occurred among individuals who smoked. Many doctors, impressed by this evidence, then gave up smoking. This was followed by a large decrease in lung cancer rates among doctors, thus seeming to confirm that tobacco smoke was indeed the culprit. It is unusual to get such clear evidence from observational data. Various forms of corroborating evidence soon appeared. There is an excellent brief summary in Freedman (1999). Section 14.5 has further discussion of the evidence on health effects of smoking.

*5.2 Studies that rely on regression modelling

I have attached an asterisk to this section because it discusses difficult technical issues. These issues are however crucial for studies where conclusions rely on the interpretation of coefficients in a regression model that has several covariates.

Here one drops any pretence that there is a closely matching control group. All relevant variables are entered into a multiple regression equation. Consider Neumark and Wascher's (1992) investigation of the effect of minimum wage requirements in U. S. states. For 22 states, data covered the years 1973-1989, while for remaining states it covered the period 1977-1989. They derived a large number of equations. The estimated equation that they defend as an accurate model for teenagers is:

$$E = a - 0.17\ [SE\ 0.07] \times MW - 0.31\ [SE\ 0.07] \times PUE - 0.75\ [SE\ 0.03] \times PA + S + Y$$

Here E = estimated employment to population ratio, MW is a measure of the minimum wage, PUE is the prime-age male unemployment rate, PA is the proportion of the age group in school, S is a state effect and Y is a year effect.

The equation seems a fair representation of Neumark and Wascher's data. It predicts that if other variables are held constant, then increasing the minimum wage by 10% will reduce employment by about 1.7%. (The 95% confidence interval is 0.3% to 3.1%.)

There are various difficulties with this equation. Perhaps the most serious is that the proportion of the age group in school (PA) is directly correlated with E. If PA goes up and other variables are held constant, there are fewer young people available for employment. If one omits PA, the apparent effect of minimum wage changes disappears.

Earlier we noted the Card and Krueger study that compared the fast food industry in a state that introduced a minimum wage (New Jersey) with a neighbouring state

(Pennsylvania) that did not. The advantage of this approach is that it allows a direct comparison, without regression adjustments.

Lalonde's comparison between experimental and regression results

An important study, in any discussion of how far it is reasonable to press the use of regression methods, is Lalonde (1986), revisited more recently by Dehejia and Wahba (1999). Its point of departure was a randomised experiment that examined the effect of a US labour training program on post-intervention income levels. Individuals who had faced economic and social hardship prior to the program were randomly assigned, over a 2-year period, either to a treatment group that participated in the labour training program or to a control group. The results for males, because they highlight estimation problems more sharply, have been studied more extensively than the corresponding results for females. Male 1978 earnings increased, relative to those in the control group, by an average of $886 [SE $472].

Lalonde's idea was to replace the experimental control group with two non-experimental groups that had been studied extensively, then using regression methods to estimate the effect on earnings. The results are discouraging. The estimate depends strongly on the form of regression adjustment. Even more disturbingly, it was in every case negative, and different for the different comparison groups. The closest agreement was a decrease in earnings of $1844 [SE $762] when the analysis used one non-experimental control group, and a decrease of $987 [SE $452] when it used the other non-experimental control group. The figures improved slightly, i.e. became less negative, when comparisons were with subsets of the non-experimental control groups that more closely matched the characteristics of the treatment group.

Dehejia and Wahba (1999) revisited Lalonde's study, using his data. They used the propensity score methodology, as expounded e.g. in Rosenbaum and Rubin (1983). Here is a simplified description of the approach, as used by Dehejia and Wahba (1999). A propensity is a measure, determined by covariate values, of the probability that an observation will fall in the treatment rather than in the control group. Various forms of discriminant analysis may be used to determine scores. Comparison of treatment and control groups then uses only those observations whose propensity scores lie within the overlapping parts of the ranges of treatment and control groups. Comparison of treatment and control group then proceeds using the propensity score as the only covariate. Dehejia and Wahba (1999) used this methodology to reproduce, in comparisons using the non-experimental control groups, results that closely matched the experimental results. The task is not as hopeless as Lalonde's study seemed to indicate. It does however require a careful and subtle use of a methodology that is adapted for handling non-experimental comparisons. A straightforward use of regression methods will not work. In general Dehejia and Wahba's methods require extensive data. A key requirement is that the data must include information on all relevant covariates.

This work warns that coefficients in regression equations can be highly misleading. Regression modelling places two demands on the coefficients. They must model within group relationships acceptably well, and in addition they must model effects that relate to differences between groups. Even where the groups are reasonably well matched on relevant variables, the methodology may not be able to reconcile these perhaps conflicting demands. Where the ranges of some variables are widely different in the different groups, the task is even more impossible.

5.3 Knowledge Discovery in Databases (KDD)

The bringing together of different sources of data-based evidence may be highly useful. It may also present a confusing picture, as when the different claimed sources of evidence on the link between salt and blood pressure seem to tell a different story. We have noted that it is careful sifting and analysis of the different sources of evidence that seems needed.

Many different groups are now working to link data from museum collections into large databases. This raises interesting issues. There are extensive data on the locations of organisms that were collected for taxonomic purposes, but relatively little data on abundance. What use can we make, for estimating abundance, of information that a particular organism was collected in a taxonomic field excursion at a particular location on a particular day? What do we know about the collecting practices of the taxonomists who made the records? Did they lose interest in a species once they had seen more than two or three of them? Were they more interested in some species than in others? (Yes!) Efforts to use data from taxonomic field excursions to make inferences about species abundance seem fraught with hazards. There is no good way to calibrate across from the taxonomic field data to abundance estimates. Knowledge of the sources of the data, and of the purpose for which they were collected, will be crucial for making such use as is defensible of the data now being collected into databases. Often, as in the attempt to use taxonomic data to estimate abundance, any estimate must be hedged about with so many caveats that the usefulness of any inference must be questioned.

References and Further Reading

Causal Effects in Non-experimental Studies

Dehejia, R.H. and Wahba, S. 1999. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. Journal of the American Statistical Association 94: 1053-1062.

Freedman, D. 1999. From association to causation: some remarks on the history of statistics. Statistical Science 14: 243-258.

Lalonde, R. 1986. Evaluating the economic evaluations of training programs. American Economic Review 76: 604-620.

Rosenbaum, P. R. 1999. Choice as an alternative to control in observational studies. Statistical Science 14: 259-278, with following discussion, pp. 279-304.

Rosenbaum, P. and Rubin, D. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70: 41-55.

Quasi-Experimental and Observational Studies

Card, D. and Krueger, A. 1994. Minimum wages and employment: a case study of the fast food industry in New Jeresey and Pennsylvania. American Economic Review 84: 772-793.

Christie, D., Gordon, I., and Heller, R. 1987. Epidemiology. An Introductory Text for Medical and Other Health Science Students. New South Wales University Press, Kensington NSW, Australia.

Hisamuchi S, Fukao P, Sugawara N, et al. 1991. Evaluation of mass screening programme for stomach cancer in Japan. In: Miller AB, Chamberlain J, Day NE, et al., Eds.: Cancer Screening. Cambridge: Cambridge University Press, pp 357-372.

Neumark, D. and Wascher, D. 1992. Employment effects of minimum and subminimum wages: panel data on state minimum wage laws. Industrial and Labor Relations Review 46: 55-81.

[See also (1993) 47: 487-512 for a critique by Card and Krueger and a reply by Neumark and Wascher.]

Petitti, D. 1994. Meta-analysis, Decision Analysis and Cost-Effectiveness Analysis. Oxford University Press.

6. Sample Surveys, Questionnaires and Interviews

> It must be stressed that fact-collecting is no substitute for thought and desk research, and that the comparative ease with which survey techniques can be mastered is all the more reason why their limitations as well as their capabilities should be understood. Sound judgement in their use depends on this. It is no good, for instance, blindly applying the formal standardized methods generally used in survey or market research enquiries to many of the more complex problems in which sociologists are interested.
> [Moser and Kalton 1971, p.3]

Sampling is ubiquitous. A person buying a sack of potatoes will use a small sample of the potatoes as a basis for assessment of the contents of the sack. Auditors who are checking for mistakes or fraud will examine a sample of a firm's accounts. This chapter focuses on sample surveys that use samples to gain information on a human population.
Important concepts are target population and sampling frame. Probability based sampling schemes help avoid sampling bias and allow estimates of accuracy. Simple random sampling is the simplest such scheme. More complex schemes combine simple random sampling with cluster sampling and/or stratified sampling. Non-response is the bane of surveys of human populations. It may introduce serious bias.
Human sample surveys typically work with questionnaires. An inappropriate choice of questions, and/or poor overall design of the questionnaire, can bias responses. What strategies and checks can ensure that responses do genuinely answer the questions that were in the researcher's mind?
Qualitative approaches should often complement quantitative approaches. Qualitative investigation may help indicate what forms of quantitative investigation may be helpful and useful. It may shed light on what respondents intended by their answers.

A cook takes a spoonful of soup from the cooking pot to determine whether the amount of salt is right. From the taste of the spoonful, the cook generalizes to the whole pot of soup. Wine tasters taste a sample of the wine in a bottle, and on that basis make a judgment about the whole bottle. Auditors are not able to examine all transactions in the accounts that they scrutinise. Instead they take a sample of the accounts, and base conclusions on the sample. All the time we sample.
In an experiment, it may be necessary to take a sample from the experimental unit. If the experiment is a clinical trial that collects data on how the treatment affects the patients' blood, any measurement must be made on a sample of the blood! Results from the sample are taken as indicative of all the blood in the patient's body. In an experiment where trees are the experimental units, measurements of the amounts of calcium in the apples will be taken on a sample of the apples.
Survey data are widely used for decision and policy making. Unlike an experiment, the aim is not to study the effect of change, but to learn what is! While surveys may sometimes be used to gather data that will be used to evaluate the effects of contrived change, this is not a necessary or predominant survey context. Decisions on whether and how to market a new product, on the effects on government finances of changes in tax rates, or on priorities for new housing development, may rely crucially on information from surveys.

In this chapter the focus is on studies where samples are used to survey a human population, typically using questionnaires to elicit information. Many of the points carry over to surveys of organizations, or of animal or plant populations.

6.1 The Planning of Questionnaire Based Sample Surveys

Planning should be based around a clear idea of the purpose that the survey is intended to serve. There will be an initial set of steps that identify the research question or questions, identify any relevant information that is available from existing sources, and establish that a questionnaire-based sample survey really is the most appropriate way to go about getting the new information.

Respondents are likely to be more co-operative if they can be persuaded that the questionnaire addresses important concerns. There should be a preamble at the beginning of the questionnaire, or that goes out with the questionnaire, that sets out the purpose. It should explain how results will be used, and address confidentiality issues.

Survey planners become keenly aware of the demands that the conduct of the survey places on them. But remember that the survey is also an intrusion on those who respond. The survey planner has a duty to respondents to carry out the task in a way that makes their effort worthwhile.

Once the research question is clearly identified and it is agreed that a questionnaire-based sample survey will be effective in providing the answers, what then? There are logistical issues, there are sampling design issues, there are questionnaire design issues, and there are data analysis issues. I will make brief comments under each of these headings.

Logistical Issues

The logistics of carrying out the survey must be planned. Will responses be obtained by interview, post, telephone[17], or by some other method? Face to face interviewing can allow relatively subtle forms of questioning, and can give a good response rates. Effective conduct of interviews does however require skills that, for most individuals, take time and experience to develop. With postal and other forms of self-completion questionnaires, some form of motivation to respond is almost essential. There may be a reward. Even then, it will almost certainly be necessary to send reminders, or even phone or visit non-respondents, in order to get a reasonable response rate. Failure to follow up non-respondents can wreck an otherwise well-conducted survey.

Surveys of official agencies, or of organizations, will require the co-operation of the relevant officials or managers, people that Lynn (1996) calls "gatekeepers". Processes must be followed that may be specific to each organization. Negotiating a way through these processes can be frustrating and time-consuming.

Detailed logistics cannot be worked out until sampling design issues are resolved. What are the different tasks that are involved? Who will perform these various tasks? In a major survey, there are huge planning demands. For further discussion see e.g. Duoba and Maindonald (1988), Moser and Kalton (1971).

Sampling Design

1. What is the target population, i.e. the population about which information is required?

[17] Issues that arise in telephone surveys are discussed in Collins (1999).

2. What is the sampling frame, i.e. the population from which individuals will be sampled? While this should ideally be the same as the target population, some compromise is usually necessary. In a simple survey design, this may be a list of names and addresses, or names and phone numbers.
3. What method will be used for selecting the sample? Ideally the sample should be chosen using a probabilistic sampling scheme, of which the simplest is simple random sampling. Non-probabilistic methods, e.g. including in the sample whoever one can most readily find, have a serious risk of bias. Self-selected samples, e.g. where people who are interested ring in to give answers to questions that have appeared in a magazine, may be very seriously biased.
4. What steps will be taken to ensure high levels of response? What efforts will be made to follow-up non-respondents?

Developing the Questionnaire

Careful researchers will look carefully to see what they can learn from their predecessors. This may extend to using and adapting a well-tested questionnaire that was used by one or more earlier researchers. The survey researcher gets the benefit of the earlier testing, and of what can be been learned from previous use. Because the same or a very similar questionnaire was used, results bear direct comparison this those from previous research.

There are standard forms of questionnaire that have been developed to address particular types of question – mental health, feelings of physical health, and so on. These questionnaires have acquired the status of research instruments. Examples are the Beck Depression Inventory, the Minnesota Multiphasic Personality Inventory and the Personality Research Form[18]. Even with such existing and apparently well-tested questionnaires, users must check, to the extent that they can, that the questionnaire does its task well in the new setting. Be aware that the questionnaire may not live up to all the claims that have been made for it.

However good the questionnaire, responses will depend to an extent on the wording of the questions. Questions should be clear, not open to misinterpretation, and have the same meaning for respondents as for the designer of the questionnaire. This is an impossible ideal. There are however common and recognisable possibilities for misinterpretation that should be investigated and avoided.

Where no existing questionnaire is available, there are (at least) two different styles:

1. There are questionnaires that seek specific factual information. For example, the aim may be to discover how people spend their money, how much on food, how much on sport, how much on entertainment, and so on. There are surveys of what people eat.
2. There are questionnaires that investigate opinions, attitudes or feelings. What is the attitude of year seven students to science? The subject is complex and clearly has a number of different facets. The need is for questions that together capture something of the different and complex responses of the students to science.

For item 1, the general nature of the questions is clear. The problem is to express them in clear and unambiguous language. The questionnaires that are described in item 2 offer the greatest challenge. A good strategy is to identify a small number of themes, then center the questions around those themes. What are the appropriate themes? For each theme, what are appropriate questions? These will be

[18] These are discussed briefly, with references, in Streiner and Norman 1995.

supplemented with questions that provide any necessary background information on the respondents – age, sex, etc.
Here are suggested steps for developing the questionnaire. They will be explained in more detail below. Relative to common practice, they may seem unusually careful. But how, otherwise, can the questionnaire designer be confident that what respondents understand is similar to what he/she intended?

1. Make a draft of the questionnaire. Check that it has a clear coherent structure. Be sure to include a short preamble that explains the purpose of the questionnaire, what will happen to the results, what has been done to ensure confidentiality, and so on.
2. Get someone who is experienced with questionnaires to look over it with a critical eye. Make any necessary revisions.
3. Seek the co-operation of 10-15 potential respondents. Administer the questionnaire verbally. Note, using the headings in section 6.3, the behaviours that each question elicits. (Behaviour coding).
4. This is a follow-up to step 9. Once each set of results is complete, ask the respondent to explain their answers in a sentence or two. (Probing).
5. Make any necessary revisions.

For a large survey, the main survey should be preceded by a pilot survey with a substantial number (perhaps 30-100) respondents. After entering the data and carrying out a summary analysis, there should be a review both of the questionnaire and of the conduct of the survey.

The Analysis of Data from Sample Surveys

When there are a small number of questions that directly address points of interest, analysis is straightforward. Consider a neighbourhood sample survey directed to determining the extent of support for an intended beachfront development. If there is no quibble over the form of the question that was asked, if 70% of respondents oppose the development while 40% support it, if the sample size was several hundred, and if the response rate is more 90%, all that remains is to comment on the accuracy of the result.
Few surveys are so simple. Structuring questions around a small number of *themes*, in the manner that I suggested above, facilitates analysis. Individual summaries of data from 30 or 40 questions are rarely very insightful, especially if the sample is quite small. Summaries of what has been learned about each of 5 or 6 themes are much more comprehensible.
Another way to put structure into the summary is to classify questions according to the response they have elicited. For No/Yes questions, there may be questions that get very few "yes" responses, questions where "no" and "yes" are fairly evenly split, and questions where most responses are "yes". With 100 respondents anywhere between 40% and 60% will be consistent with a 50/50 split. With 400 respondents the range narrows to 45% - 55%[19].
Responses for each individual question will often be on a five or seven point Likert scale. An example is (in a survey of year seven students):

How interesting do you find science? Circle your choice:

[19] For random samples, these are the ranges that are consistent with a 50/50 split, as assessed by a 95% confidence interval for the population proportion. In practice, because simple random sampling has not been used, and because of non-response bias, these ranges may realistically be much wider than stated.

1: Not at all interesting	2: Not very interesting	3:Somewhat interesting	4:Quite interesting	5:Very interesting

This is a five-point Likert scale. "Not at all interesting" rates as 1, "Not very interesting" rates as 2, and so on. There may be four or five questions that focus around this same theme of attitudes to science, with high ratings indicating positive attitudes and low ratings indicating negative attitudes to science. A simple way to get an overall "attitudes to science" score may be to add the scores from the four or five individual questions[20]. If this seems inappropriate, one might use principal components analysis to determine scores[21]. It may even be possible to combine results from several themes into a single score.

6.2 The Language of Sample Surveys

Target Population and Sampling Frame

There must be a clearly defined target population. The target population is the population about which you would like information. Ideally your sample frame, i.e. the list of individuals from which you sample, should consist of all members of the target population. This is often difficult or impossible.

Suppose for example you want to conduct a survey of all residents in the ACT who have reached voting age. An attractive sampling frame is the electoral role. Use of this as the sampling frame will miss out on residents who are not Australian citizens and thus not registered to vote.

A famous historical example (Gallup 1976) illustrates the potential effect of an unfortunate choice of sampling frame. In 1936 the *Literary Digest* used around 2.4 million responses from lists of telephone owners, magazine subscribers and car owners to predict the result of the US Presidential election. It estimated that Roosevelt would get 43% of the vote, where in fact he received 62%. George Gallup's survey organization was then just starting up. Gallup made two estimates, which did not get the same publicity as the Literary Digest poll:

- Using a sample of 50,000 he predicted Roosevelt's victory, though with 56% of the vote rather than 62%
- Using a sample of 3000 from a sampling frame similar to that used by the Literary Digest, he predicted that the Digest poll would give Roosevelt 44% of the vote!

Even Gallup's sample of 50,000 was enormously larger than polling organizations would use today. Even in very well conducted sample surveys, non-sampling biases typically become more important than sampling error once the sample size is more than one or two thousand. In less well conducted surveys, or where the tradition of experience has been too short to allow the honing of the methodology, the cross-over point may be a few hundred or less.

[20] I am unconvinced by arguments that the ratings are not on an interval scale and should not be added. What is the alternative? The scores have to be combined somehow, formally or informally. The scale should have been chosen so the distance between "Not at all" and "Not very" is intuitively similar to that between "Somewhat" and "Not very". This is not to deny a need for caution.

[21] Principal components analysis determines a weighted combination of the scores, designed to account for as much of the variability as possible in the individual scores.

The Sample Selection Plan

Having decided on a sampling frame, there will need to be a sample selection plan, and a plan for handling non-response. We will discuss the non-response problem later, i.e. people who do not respond or cannot be found. For now, note that a low response rate, perhaps 50% or less, damages the credibility of results. The 50% who responded may give quite different responses to the 50% who did not respond. A difference in willingness to respond probably means that there are other differences also.

The simplest sample survey design uses simple random sampling. The sample frame is made up of all individuals who might potentially be in the sample. The sample surveyor takes a random sample from the sample frame. For example, a random sample might be taken from all names on the electoral role. We will discuss elaborations of this simple scheme in the next section.

Non-sampling Errors

Non-response is one of several types of non-sampling error. Other non-sampling errors may arise because the questions have been misunderstood, or have been interpreted differently from the way that the survey planners intended. The next section will examine implications for questionnaire design. Comments in Moser & Kalton (1971, p.482) are apt:

> There is incongruity in the present position. One part of the survey process (the sampling) is tackled by a tool of high precision that makes accurate estimates of errors possible, while in the other parts errors of generally unknown proportions subsist. This incongruity has a double implication. It means, first of all, that the survey designer is only partly able to plan towards his goal of getting the maximum precision for a given outlay of money, since the errors (and even costs) associated with the various non-sampling phases cannot be satisfactorily estimated in advance. And secondly, so long as these errors cannot be properly estimated from the results of a survey, the practitioner is in a position to give his client an estimate of the sampling error only, not of the total of *all* kinds of error. This is a weakness, and there is here a field of fertile research for students of research methodology. ... The operation of memory errors, the kinds of errors introduced in informal as opposed to formal interviewing, the effects of length of questionnaire on errors, the errors associated with different kinds of question, the influence of interviewer selection, training and supervision, the errors introduced in coding and tabulation --- these are but a few of the many fields in which ... there remains scope for research.

In a carefully conducted mail survey, there will be a second mail-out that will seek a response from those who did not respond to the first mail-out. In telephone surveys, it will often be necessary to make several calls in order to contact some of those in the sample. Respondents should then be classified according to the ease with which it was possible to contact them, and the response compared. If differences are greater than statistical error, this will suggest that non-respondents may be even more different.

Quota Sampling

Many commercial market research organizations use this as their preferred method. Its principal advantage is reduced cost, though technological change may now be changing the relative costs. There are serious, and usually unknown, risks of bias. Quota sampling is not usually carried out in a manner that allows a realistic estimate of error from any individual sample. This may perhaps be acceptable where the aim is to get ballpark indications only.

There are mechanisms that may help calibrate results from quota sampling. Error may be estimated by examining the results of repeated quota samples. Bias can be

estimated by making occasional comparisons with a probabilistic sample that is conducted in parallel.

Question: Do quota sampling and other non-probabilistic sampling methods have a role? If so, when are they appropriate?

Self-selected Samples

For example, readers of a magazine may be asked to write in and give their opinion. These are the most hazardous of all.

Question: What other planned ways are there to collect data, apart from experiments, sample surveys, longitudinal studies, and case-control studies? What are the different challenges of these other approaches for the statistical analyst?

*6.3 Sample Survey Design

In the discussion above, we introduced the terms

- target population
- sample frame
- non-response

In addition we introduced the idea of simple random sampling. An example was the choice of names at random from an electoral role.

Stratified random sampling compared with cluster sampling

In addition to simple random sampling, there are two further basic types of sampling systems:

1. In *stratified random sampling* the sample frame is divided up into relatively homogeneous strata. A random sample is then taken from within each stratum. For any given sample size, this should, if the strata are well chosen, improve precision.
2. In *cluster sampling*, the sampling frame is divided up into clusters, often clusters of people who live in the same general locality. The sampler then takes a random sample of clusters, though perhaps making the probability that a cluster will be chosen proportional to cluster size. For a given total sample size, cluster sampling generally gives reduced precision.

Clusters and strata both group together members of the population. In stratified sampling we sample from within all strata. In cluster sampling, we take only a sample of clusters. Stratification should improve precision. Cluster sampling usually results in lower precision for a given sample size, and we need to compensate by taking a larger sample.

We have noted that cluster sampling generally gives, for a fixed total sample size, reduced precision. The reason is that individuals in a cluster – in the same locality or in the same school – are likely to be relatively similar. Each new person in the same cluster contributes less additional information than someone newly taken at random. But even though one needs to increase the sample size in order to get the same precision, the cost may still be lower than for a simple random sample. It is often easier and less expensive, especially in remote areas, to contact a number of people who all live together in the same location, rather than to select the same number of individuals according to a totally random scheme.

The combining of stratified random sampling and cluster sampling in various ways leads to a huge variety of possible sampling designs. Dalenius (1985) distinguishes three basic *sampling systems* — (i) element sampling, (ii) cluster sampling, and (iii) multi-stage sampling. These may be used individually, or combined, to provide a

sampling system. The sample scheme determines how sample elements or clusters are chosen. Options are simple random sampling, stratified sampling, and various sampling schemes that give unequal probabilities of selection. In multi-stage sampling this choice is available at each stage.

Question: Compare experimental design with sampling. What are the points of contact between the theories that apply in the two cases? What are the differences? Does the idea of hierarchical strata of variation have a counterpart in survey design?

Question: Compared with a simple random sampling scheme, and assuming a fixed total sample size:
(i) How does cluster sampling typically affect the accuracy of the sample mean?
(ii) How does effective use of stratified random sampling affect the accuracy of the sample mean?

Multi-stage sampling

Cluster sampling can be mixed with stratified sampling to give stratified cluster sampling. Instead of using simple random sampling within each cluster, one uses cluster sampling. More generally, the sample procedure may be multi-layered, leading to multi-stage sampling. At each stage the method used may be stratified random sampling, or cluster sampling, or a mixture of the two.

Stratified Random Sampling – The Choice of Strata

Suppose the aim is to estimate the distribution of household expenditure on restaurant meals in the ACT, over a two-week period. The sampling frame might be a list of street addresses. It might be necessary to use the sample itself to estimate the number of households living at each address. Accuracy might be improved by stratifying regions of Canberra according to socioeconomic status. The argument is that expenditure will be higher in regions with high socioeconomic status. For stratification to be effective one needs a variable, positively correlated with the outcome that is of interest, that can be used to define the strata.

If we already had good information on where the patrons of restaurants lived, we would use that information. There might for example be an earlier survey that provides this information. Another way to proceed might be to conduct a preliminary survey of restaurants, asking patrons where they live.

Question: What might be good stratifying variables for surveys that

1. estimate the total number of wombats in New South Wales?
2. estimate the total dollar amount of accounting mistakes, over the course of a year, in the customer invoices of a sheet metal supplier?
   [The total amount of each invoice, the customer and the date, can be determined from computer records. Other information must be extracted manually.]
3. estimate the annual expenditure per household, in New South Wales, on overseas holidays?
4. estimate expenditure per household, in New South Wales, on holidays in Greece.
5. estimate amount spent per household, in New South Wales and the ACT, on boats and related pleasure craft?

Finally: Can you think of methods, better than surveying the whole population, for getting any of the above information?

6.4 Questionnaire Design

> *Research questions must translate into a set of questions, and into a questionnaire, that will provide answers to the questions to which you as a researcher want answers. What steps will help ensure responses that will give reliable and valid answers to the research question?*

Here we take up in more detail points that were raised in section 6.1. We discuss some recent ideas on approaches to checking and testing questions, and we list the types of problems that may occur.

Behaviour Coding

Where an interviewer administers the questionnaire, coding of respondent behaviour may be used to identify actual or potential problems. The problem behaviour code categories used in Oksenberg et al. (1991) were

1. Respondent interrupts initial question-reading with answer.
2. Respondent asks for repeat or clarification, or otherwise indicates uncertainty about the meaning of the question.
3. Respondent answers question as asked, but adds a qualification.
4. Answer is inadequate.
5. Respondent gives "don't know" or equivalent answer.
6. Respondent refuses to answer.

Questions that frequently elicit one of these behaviours are problem questions.

Probing

Respondents answer to the question as they understand it. This may differ from what the researcher intended. So a facet of the pre-testing is to follow administration of the questionnaire with probing designed to discover how the respondent understood the question. Oksenberg et al.(1991) quote as an example:

> "During the past twelve months, that is, since January 1 1987, *about* how many days did illness or injury keep you in bed for more than half the day."

Most respondents took this to mean not getting up in the morning and staying in bed till about noon or later. Others had in mind lengths of time, as little as 2-4 hours or as much as 12 or more hours. Another issue was whether staying in bed because they felt they were coming down with something would count as illness. About two thirds would have included this, while the other third would not.

What sorts of problems occur with questions?

The following classification of problem questions is adapted from Presser and Blair (1994, pp. 96-101)

1. Leading or loaded question
   "Did you spend at least 8 hours doing physical exercise last week?"
   [Should I have been doing a bit more exercise?]
2. Information overload (Question too long or intricate)
3. Unclear structuring of words or ideas
   "Before you got married, how long did you live in Canberra after you graduated from University?"
   [Marriage, living in Canberra and graduation are juxtaposed in a manner that will confuse many respondents! Do interrupted periods of residence count?]
4. Flow between questions
   "How satisfied are you with the schools in your neighbourhood", then
   "How satisfied are you with the grocery stores in the neighbourhood where you work?"
   [An alert is needed that the question will refer to a different neighbourhood.]
5. Confused Boundary Lines
   "How long have you lived in Canberra?"
   [Some who have moved in and out may count the most recent time; others the total time.]

6. Common term is not understood
   "Do you separate aluminium cans from your regular garbage?"
   [What is regular garbage? Is there an irregular kind?]
7. Double-barrelled question
   "Please indicate how you rate the job that the police and the courts do?"
   [The police are fine. I see a problem with the courts.]
8. Recall/response is difficult
   "How many times did you go to the movies in the past 12 months?"
   [I went a lot. It could have been 20 times or 50 times.]
9. Recall/response is impossible
   "How many kilometres did you drive in the last year?"
   [Few people will know this.]
10. Question seems a repeat of the previous question
    "How many times did you start your car's engine yesterday?"
    "How many times did you stop your car's engine yesterday?"
    [I've just told you!]
11. Inappropriate assumption
    "How many times did you drive over the speed limit on the way to work?"
    [I came by bus.]
12. Overlapping response categories
    "Which range is your salary in — $0-$30,000, $30,000-$60,000, or >$60,000?"
    [I get $30,000. Which box do I tick?]
13. None of these
    "Did you take this course for professional development or out of personal interest?"
    [Neither. My tutor told me I needed to come.]
14. Response categories too finely drawn
    "Please rate your tutor's ability to stimulate interest on a scale of 0 to 100, where < 50 is unfavourable and > 50 is favourable."
    [What does a 75 mean?]
15. Response categories not appropriate to question
    "Do you drive to work? NO YES CARPOOL"
    [What has carpooling got to do with it?]
16. Sensitive questions
    "How many sexual partners did you have in the past year?"
    [Some will refuse to answer. Others will be uncomfortable.]
17. Awkward syntax (an especial problem when an interviewer has to read the question out.)
    "The Department of Social Security has information in its files about census items like date of birth and sex for nearly everyone. Would you favour or oppose giving this information to the Bureau of Statistics for use in the Census?"]
    [You surely don't mean "sex for nearly everyone". You mean the DSS holds information on everyone's date of birth and sex.]
18. Open question
    "Did you have any special difficulties when you were a first-year student? If

you did, please describe them."
[Open questions have their place. They can be hard to code.]

"Cognitive laboratory methods" is a collective name for methods that try to tease out the thought processes that led to a particular response (Forsyth and Lessler 1991).

6.5 Questionnaires as Instruments

As noted in section 6.1, a particular form of questionnaire may be refined to the point where it becomes a recognised social science "instrument", widely used by different researchers. A key issue is: "What does the instrument actually measure?"

Content Validity

Does the statistical data connect strongly with the problem in which we are interested? Issues of content validity arise with particular force in psychometric testing. Do IQ tests really measure intelligence? Perhaps, if we knew what intelligence was, we could say. Note Nunnally's (1978, p. 94) comment:

> In spite of some efforts to settle every issue about psychological measurement by a flight into statistics, content validity is mainly settled in other ways. Although helpful hints are obtained from analyses of statistical findings, content validity primarily rests upon an appeal to the propriety of content and the way that it is presented.

Even apparently hard factual questions may measure something different from what we think they measure. Questions about sexual and other practices where there are strong social constraints are particularly difficult.

Face validity

Broadly, this has to do with the extent to which those who work in the area find the measure a credible instrument for its claimed purpose. Of course, researchers may be wrong.

New Glosses on Old Words

Surveyors use measuring tapes and theodolites. Social scientists use questionnaires as major measuring instruments. Is the analogy accurate and useful? I believe it is. The measuring instruments that social scientists propose do not have the obvious directness of a measuring tape. As Nunnally (1978, p.109) says:

> A construct is only a word, and although the word may suggest explorations of the internal structure of an interesting set of variables, there is no way to prove that any combination of those variables actually measures the word.
>
> ...
>
> New measurement methods, like most new ways of doing things, should not be trusted until they have proved themselves in many applications. If over the course of numerous investigations a measuring instrument produces interesting findings and tends to fit the construct name applied to the instrument, then investigators are encouraged to continue using the instrument in research and to use the name to refer to the instrument. On the other hand, if the evidence is dismal in this regard, it discourages scientists from investing in additional research with the instrument, and it makes them wonder if the instrument really fits the trait name that has been employed to describe it.

Streiner and Norman (1995) has a helpful review of literature on the design of questionnaires. Although they focus on health measurement scales, their critique has wider application.

Food Frequency Questionnaires

It has long been suspected that a high nutrient fat intake increases the risk of breast cancer. A number of large prospective cohort studies that have looked for such a link have found nothing. Other types of study, which however are open to objection for other reasons, do suggest an increased risk. See section 14.6 for further discussion. The favoured instrument for assessing fat intake has been a food frequency questionnaire (FFQ). Does failure to find a link mean there is no link, or is the problem with the measuring instrument? Kipnis et al. (1999) suggest that the problem may lie with the measuring instrument, at least to the extent that its properties require much more careful investigation than they have received to date. Specifically, they show that a person-specific bias in the recording of fat intake might explain the failure to find an association between nutrient fat intake and breast cancer risk.

There will now be studies that will allow estimation of the distribution of any person-specific bias. If the person-specific bias proves substantial, this will seriously undermine the use of the food frequency questionnaire as a measuring instrument in studies where relatively fine discrimination is required.

6.6 Qualitative Research

The structuring of information so that it can be collected by a questionnaire, or derived from an experiment, places severe constraints on what can be learned. There are large areas of knowledge which we can access only by allowing respondents opportunity both to determine the range of information and to structure its content in ways that make sense to them. This takes us outside of the bounds of the formal data collection approaches so far discussed. The term "qualitative study'' is used without prejudice to the possibility that it may later be possible to place a quantitative structure on some part of the information that is gathered.

Moreover quantitative studies start with qualitative judgements. There must be some judgment on which ideas are worth pursuing, on what the research question is to be. Where there is little previous research on which to rely for guidance, the over-riding initial demand may be for qualitative information that will provide clues on the questions that it is appropriate to ask.

Qualitative studies may be especially appropriate, as a first step, in getting started on studies where human interaction has a large role. For example, trained but often relatively inexperienced village midwives, intended to replace traditional birth attendants, were a major initiative of a former Indonesian government health minister. Why were some midwives accepted by villagers, and used in deliveries, while other midwives were not? What were the important considerations: attachment to traditional ways, medical competence, social standing, experience, knowledge of the local context, personal qualities, or what? Substantial insight into the likely social dynamics, which may well differ between villages, seems required before mounting a quantitative study.

The term `qualitative study' has been used by social scientists. Researchers in industry, or in the physical sciences, are more likely to speak of `idea generation', and the `refining and honing of ideas'. Thus the Scholtes (1988) monograph on industrial problem solving speaks of generating and honing ideas.

Qualitative studies may be treated as complete in themselves, or they may be explicitly intended to complement a quantitative study. In either case, there are often aspects of the study where it is helpful and appropriate to use quantitative methods. Thus graphical presentation, various forms of statistical summary, and clustering methods, have application to some quantitative studies. These approaches are used

for summary and the illumination of pattern, not for statistical inference. Anyone who conducts a qualitative study must however understand the different purposes and uses of these two different types of study. Results do not allow the same secure generalisation to a wider population that may be available from a carefully conducted quantitative study.

Qualitative studies may be used to generate questions that can then be addressed in follow-up quantitative studies. Sample selection issues, though less critical than in quantitative studies, are still important. Representativeness may be more important than the use of a sampling scheme that allows calculation of standard errors for any quantitative information.

When qualitative studies aim or claim to provide insights that stand on their own, it is important to know the extent to which results generalize to the relevant target group. Just as in quantitative studies, sample selection is a key issue. Any available checks on consistency with other evidence should be applied. The term `triangulation' has entered the social science jargon. Often an interpretative scheme or theory is imposed on the data. Are the data also consistent with other competing interpretative schemes?

References and Further Reading

General

Kipnis, V., Carroll, R.J., Freedman, L.S. and Li Li 1999. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. American Journal of Epidemiology 150: 642-651.

Sample Surveys

Biemer, P. B., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. and Sadman, S. (eds.) 1991. Measurement Error in Surveys. Wiley, New York.

Collins, M. 1999. Editorial: Sampling for UK telephone surveys. Journal of the Royal Statistical Society A, 162: 1-4.

Dalenius, Tore. 1985. Elements of Survey Sampling. Sarec, Stockholm.

Duoba, V. and Maindonald, J. H. 1988. Understanding Surveys. New Zealand Statistical Association, Wellington.

Gallup, G. [1972] rev. 1976. The Sophisticated Poll Watcher's Guide. Princeton Opinion Press.

Lynn, P. 1996. Sampling in human studies. In Greenfield, T., ed.: Research Methods. Guidance for Postgraduates, chapter 17.

Moser. C. A. and Kalton, G., 2nd edn. 1971. Survey Methods in Social Investigation. Heinemann Educational Books, London.

Questionnaire Design

Forsyth, B. H. and Lessler, J. T. 1991. Cognitive laboratory methods: A taxonomy. In Biemer, Groves, Lyberg, Mathiowetz and Sadman (eds.) 1991. Measurement Error in Surveys, pp. 393-418. Wiley, New York.

Judd, C. M., Smith, E. R. and Kidder, L. H. 1991. Measurement. From Abstract Concepts to Concrete Representaions. In Research Methods in Social Relations (sixth edition), Holt Rinehard and Winston Inc 1991, 42-61. (See also "Maximising Construct Validity", pp. 30-32.)

Nunnally, J. C., 2nd edn 1978. Psychometric Theory. McGraw-Hill, New York.

Oksenberg, L., Cannell, C., and Kalton, G. 1991. New strategies for pretesting survey questions. Journal of Official Statistics (Statistics Sweden) 7: 349-365.

Oppenheim, A. N. 1992. Questionnaire Design, Interviewing and Attitude Measurement. Pinter Publishers, London.

Presser, S. and Blair, J. 1994. Survey pretesting: Do different methods produce different results? Sociological Methodology 24: 73-104.

Streiner, D. L. and Norman, G. R., 2nd edn., 1995. Health measurement scales: a practical guide to their development and use. Oxford University Press.

Qualitative Research

Britten N, Jones R, Murphy E, Stacy R (1995). Qualitative Research Methods in General Practice and Primary Care. Family Practice 12: 104 - 114. Oxford University Press.

Greenhalgh, T. 1997. How to read a paper: the basics of evidence-based medicine. BMJ, London. [See the chapter on Qualitative Research.]

Kuzel A J (1992). Sampling in Qualitative Inquiry. In B K Crabtree & W L Miller (ed), Doing Qualitative Research (Vol. 3, pp 31-44). Newbury Park: Sage Publications.

Scholtes, P. R. 1988. The Team Handbook. Joiner Associates, Madison, Wisconsin.

7 Sample Size Calculations

Sample size calculations may be needed for many different types of study. Researchers should know roughly what precision they can expect from their study. How large a difference between treatments is detectable?
Sample size issues should be considered alongside, and be subordinate to, sample structure issues. It is good design that is needed, not necessarily a large sample size.

In a randomised controlled trial with control and treatment groups, a decision is needed on how many will be in the control group, and how many in the treatment group. Or if there is a limitation on the available numbers in the two groups, the researchers will want to know the implications for the accuracy of the result. A sample size calculation relies on various assumptions. A provisional model is needed for the data. Often it is possible to make a stab at the information that is needed. Where research breaks totally new ground, getting a good guesstimate may be more difficult.

7.1 Issues for sample size calculation

In the randomised controlled trial example, the *effect size* will be the difference between results for treatment and control. It is common to specify the effect size, and ask what size of experiment or sample is needed to detect an effect of that size.
At this point we focus on principles. There are details of how to do simple sample size calculations in the next section. Remember that sample size calculations, where they seem helpful, should be an adjunct to other aspects of planning. Do not allow preoccupation with sample size issues to distract attention from these other aspects. Here are reasons why sample size calculation may be helpful:

1. A sample size calculation requires either a clearly specified hypothesis or a clearly specified estimation problem. Insistence on a sample size calculation may help ensure a reasonably precise statement of the research question(s).
2. The attempt to specify large numbers of perhaps complicated hypotheses will create problems for sample size calculation. If the attempt at sample size calculation helps force this point on the researcher, all to the good. Numerous hypotheses, or hypotheses that are overly complicated, may indicate that the research does not yet have a clear focus. More work is needed in teasing out the main research questions.
3. The attempt to use results from the literature as a basis for sample size calculation may help draw attention to problems with the studies themselves, or with the reporting of results.

The importance and relevance of sample size calculations will vary from study to study. Here are points to note:

1. Researchers should certainly have a rationale for the size of their study. Size, i.e. number of replicates, is just one of several issues that call for attention. It is important that the research effort is used to maximum effect.
2. Where improvements in study design allow improved precision, this is usually preferable to increasing the sample size. They may help avoid the huge logistical problems that large or very large sample sizes can create. There may be a need to incorporate new factors into the design, e.g. individual operator effects when blood pressure measurements are taken.

3. Each new study should be seen as part of a total learning process. The key issue for the researcher is how the new study can best contribute to the total learning process, given the state of existing knowledge.
4. In highly exploratory studies, the effort put into trying to get high precision may be largely wasted. The initial study will often provide information that calls for substantial modification of the initial design. Such studies have the character of pilot studies. The priority should often be the collection of information that will assist in the design of later studies, rather than high precision.
5. Sample size calculations have received a huge amount of attention in such studies as medical case-control studies and clinical trials. Here, generally, they do have a useful role. However as with other studies, sample size is only one of a number of important design issues. It has too often been treated as the one issue of major importance.
6. If a study is to stand on its own, then sample size is highly important. If it is one in a series of studies that will finally be analysed together, then the sample size in that individual study may have more limited consequence.
7. There is an urgent need for mechanisms that will foster co-operation between different researchers who are working on similar questions, so that their work meets similar standards and can finally be evaluated in a single overall analysis. Questions of sample size in individual studies should be addressed in this wider context. This is a particular issue for clinical trials.
8. The aim should be accurate estimation of variability, and ensuring there are enough degrees of freedom to do this, rather than replication as such.
9. Once the study has been conducted, the initial sample size calculation has no relevance. The analysis will provide information on the accuracy of the estimated effects, and it is this that is of interest.

Strong assumptions may underlie sample size calculations. If the assumptions are not satisfied, then the answer may be seriously astray. If the same faulty assumptions underpin the eventual analysis, that will be wrong also.

Information Required for Sample Size Calculations

A useful side-effect of the demand for sample size calculations is that it forces a search for information that may be more widely relevant to understanding the scientific context and to the design of data collection. If it is impossible to find information based on sampling from a precisely similar population, then it will be necessary to canvass more widely, looking for a broadly similar population.

For comparing proportions, a conservative (i.e. erring on the large side) estimate is obtained by assuming that the population proportion is 0.5.

If no information on standard errors can be found, then an approach is to reformulate the comparison as a comparison of proportions. Comparisons based on comparing continuous variables typically have greater power than comparisons based on a comparison of proportions. So this is a conservative procedure, i.e. it will tend to over-estimate the sample size.

The Right Sample Structure

We have so far assumed that it is obvious what the sample units are. It is not always that simple. Suppose that two different devices for measuring fruit firmness are to be compared. A sample of fruit will be taken, with half then randomly assigned to one instrument and half to the other. The instrument used for any particular fruit will make two measurements. Note that even though two measurements are made on each

fruit, it is the number of fruit that is crucial. The experimental unit is an individual fruit. (A better design would of course be one where each instrument makes one or more assessments on the same fruit.)
This discussion is deliberately brief. Several computer programs that will handle straightforward types of sample size calculation are now available from the internet. See Brown et al. (1996), Thomas and Krebs (1997). Researchers are advised to use one of these programs for any sample size calculations or, better still, to consult a statistician who is knowledgeable about such matters.
Unless one is unusually fortunate in the information that is available from earlier trials, it will be necessary to guess at the standard deviation that should be plugged into the formula. There is often some arbitrariness in the choice of effect size. So the number that comes out at the end can be a rough guide only.

The Limitations of Power Size Calculations

Johnson (1998) argues against the use of power calculations in clinical pyschiatric trials. The aim should instead be to recruit at least 100 patients in each treatment group, and preferably 200. These are the numbers that are typically required to distinguish clinically significant effects. While smaller trials may sometimes be useful, they risk capturing the characteristics of an idiosyncratic subgroup of patients. Johnson's advice is specific to clinical trials, and perhaps to psychiatric trials. In other areas, there will be different norms.

*7.2 A Common Form of Sample Size Calculation

A wide class of sample size calculation formulae has the form

$$n = \left(\frac{(t_\alpha + t_\beta) \times SD}{\Delta} \right)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where Δ is the smallest difference that it is desired to detect, *SD* is the standard deviation of the difference for a sample of one in each group, and for a test at the 5% level t_α = 1.96. If all that is required is a 50:50 chance of finding a difference, then t_β = 0. For 80% power, t_β = 0.84, while for 90% power t_β = 1.28. The formula applies also to confidence intervals, where now `power' is the desired probability that the confidence interval for the difference of interest will have a half-width of less than Δ.
Here are some special cases:

1. For a one-sample t-test, *SD* is the standard deviation *s*. Thus for matched samples, *s* is the standard deviation of differences between sample pairs and *n* is the number of sample pairs.
2. For a two-sample t-test with *s* the pooled standard deviation, *SD* is $\sqrt{2}\, s$.
3. For a two-sample t-test, with different standard deviations s_1 and s_2 for the two samples, $SD = \sqrt{s_1^2 + s_2^2}$
4. For a comparison of a proportion with a given fixed proportion, $SD = \sqrt{2p(1-p)}$, where *p* is the proportion under the null hypothesis. More accurately, replace $(t_\alpha + t_\beta)SD$ by $t_\alpha \sqrt{2p_0(1-p_0)} + t_\beta \sqrt{p(1-p)}$, where p_0 is the proportion under the null hypothesis and *p* is the sample proportion.

The above formulae make more approximations than may often be desirable. However they are adequate for giving an indication of how the sample size formulae

function. In practice, you may prefer to use one of the sample size calculation programs that are available.

Clustering Effects

The formulae we have given assume that there is no clustering. In situations where there is clustering, the between cluster variance will often dominate the variance of estimated totals or means or differences of means. The number of clusters, not the total number of individuals, may be crucial. This is equivalent to the insight, in the experimental design context, that the number of experimental units is crucial.

A simple case, which however illustrates the general principle, arises when all clusters are the same size *m*. The variance (or its estimate) can be partitioned into a within cluster component, i.e. s_w^2 between individuals in the same cluster, and a between cluster component s_b^2, i.e. between individuals in different clusters. Then the variance of the mean of a sample of size *m* from a randomly chosen cluster is $s_b^2 + s_w^2 / m$. For a given cluster size *m*, one can estimate $SD^2 = s_b^2 + s_w^2 / m$, and use the square root of this, multiplied by $\sqrt{2}$ if the interest is in differences, as a standard deviation to plug into the formulae given above. The formula will give the number of clusters that are required.

Detecting Change

Given a statistic and a standard error estimate for it, one can adapt the above methodology. Thus in a straight line regression calculation that assumes independent and identically distributed errors with variance that we estimate to be s^2, the variance (= SE^2) of the slope estimate is:

$$\frac{s^2}{n s_x^2}, \text{ where we define } s_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$

Equation (1) above becomes

$$n = \left(\frac{(t_\alpha + t_\beta) \times s}{\Delta s_x} \right)^2 .$$

The minimum detectable effect size is

$$\Delta = \frac{(t_\alpha + t_\beta) \times s}{s_x \sqrt{n}}$$

7.3 Rules of Thumb

In a random sample of 100, proportions between 30% and 70% can be estimated with an accuracy, as measured by a 95% confidence interval, of ±10%. Thus if the sample proportion is 57%, the true proportion may lie between 47% and 67%. The difference between the proportions from two independent samples of size 100 can be estimated with an accuracy (here measured by the half width of the 95% confidence interval) of about 14%.

For proportions outside of the range 30% - 70%, accuracy will be better than indicated by the above formula.

Multiplying the sample size by a factor of 10 improves the accuracy, from ±10% to ±3%, approximately. The half width of the 95% confidence interval is reduced by a factor of √10, which is about 3.2, i.e. not all that different from 3.

In complex sample surveys it is customary to speak of the design effect. This is the number by which the size of a simple random sample must be multiplied in order to get the same accuracy in the complex sample survey. Thus a design effect of 1.5 implies that a sample size of 1500 will be required to give an accuracy, in an estimated proportion that is not too different from 50%, of ±3%. Design effects in the range of 0.75 to 2 are common.

References and Further Reading

Brown, B.W., Brauner, C., Chan, A., Gutierrez, D., Herson, J., Lovato, J., Polsey, J., and Russell, K. 1996. STPLAN. Calculations for sample sizes and related problems. Available from http://odin.mdacc.tmc.edu/anonftp/page_2.html

Johnson, T. 1998. Clinical trials in psychiatry: background and statistical perspective. Statistical Methods in Medical Research 7: 209-234.

Thomas, L. and Krebs, C.J. 1997. A review of statistical power analysis software. Bulletin of the Ecological Society of America 78: 128-139.

8 The Rationale of Scientific Research

> The aim of science is to seek the simplest explanation of complex facts . . . seek simplicity and distrust it.
> [A. N. Whitehead]

> Both scepticism and wonder are skills that need honing and practice. Their harmonious marriage within the mind of every schoolchild ought to be a principal goal of public education.
> [Sagan 1997, p. 289.]

> *Any adequate account of the scientific method must allow for the exercise of imaginative insight. It must also place checks on the unconstrained use of the imagination. There must be a mechanism for distinguishing claims that can be substantiated from claims that cannot be substantiated.*
> *It must allow a role both for data and for theory. Any collection of data pre-supposes some notion that these particular data are likely to be interesting and useful. In this sense, science is driven by theory. It is the genius of science that data may challenge and even destroy the theory that guided their collection. This is the means by which science places a check on unbridled exercise of the imagination.*
> *Theory works with models. Our special interest is in statistical models. A good model captures those aspects of a phenomenon that are relevant for the purpose in hand. A model is, inevitably, an incomplete account of the phenomenon. The reward for simplifying by ignoring what is irrelevant for present purposes is that the model is tractable – we can use it to make predictions.*

I use the word *science* in a broad sense, not much different from the word *knowledge*. Scientific research is directed to gaining new knowledge.

8.1 Balancing Scientific Scepticism with Openness to New Ideas

The methods of science stand in strong contrast to belief systems — religious systems, cults of every description, popular prejudices, political ideologies of both the left and right, those claiming magical or other powers of healing, the claims of much commercial advertising, faith healers, promoters of new therapies who resist the rigours of scientific testing, and so on. Scientific claims are open, at least in principle, to rigorous objective testing. Admittedly, science does not in practice always live up to these high ideals.

There is a strong contrast with systems of ideas that resist rigorous testing. These systems readily generate, or more often rehash, ideas that are away from current mainstreams of scientific knowledge. They have rarely shown much interest in rigorous testing. They typically spurn scientific standards, even as an ideal. Standards of evidence are weak.

Theory is a fruitful source of ideas. Ideas may come from methodically working through the implications of current theory. There may be a bold and imaginative extension or adaptation of existing theory. Or the challenge may come from a new theory that questions existing notions. Whatever their source, ideas should never have an automatic claim to credence. They must stand on their merits. There must be reality checks at key points along the way — does it happen as claimed? Occasionally a theoretical insight may seem so compelling that there is no need to check further.

Previously inexplicable facts now make perfect sense. Even here one has to proceed with caution, keeping in mind our capacity for mistake and self-deception, and our proneness to jump to conclusions. Scepticism, directed at current assumptions as well as at any new theory, must be the order of the day. There are many case-histories that demonstrate the need for caution.

There are by contrast well-known instances where the scientific community refused to take seriously, on the grounds that there was no mechanism, an idea that had strong empirical support. Or important and significant results may be dismissed out of hand. The examples that follow illustrate, in turn, these two possibilities.

Continental drift

My discussion pretty much follows the account the very readable account in Hallam (1989). Wegener (1880-1930) presented a range of evidence in support of his theory that the present continental land masses had formed from the splitting apart of older continental masses. He pointed out that the Western coast of Europe and Africa fits fairly well the contours of the Eastern seaboard of the Americas. He argued that former land bridges between continents explained important features of the present distribution of fauna and flora. But geologists had a long tradition of mechanistic explanation. Prominent and influential figures denounced Wegener's ideas, creating an intellectual climate where any young and bold spirit who took up these ideas thereby placed their career at risk.

Biologists were more sympathetic. They had rarely been lucky enough to find detailed mechanisms for the phenomena that they studied, and were more willing to live with the idea that an understanding of mechanisms would have to come later. At the same time, they respected the prevailing judgment of geologists that such splitting and moving of land masses was impossible. The opposition to Wegener's ideas remained strong through into the 1950s. The highly respected geophysicist and mathematician Harold Jeffreys (1891-1989) was especially vocal in his opposition to Wegener's ideas.

> A further impossible hypothesis has often been associated with hypotheses of continental drift and with other geological hypotheses based on the earth as devoid of strength. . . . In Wegener's theory, for instance . . . the assumption that the earth can be deformed indefinitely by small forces, provided only that they act long enough, is therefore a very dangerous one, and liable to lead to serious error.
> [Jeffreys 1926, p.261]

A group of younger researchers who revived Wegener's ideas, still without much idea of the mechanism involved, thereby risked their careers. One of those younger researchers – Edward Irving – took a position at the Australian National University. Australia provided, at that time, more fertile ground for his ideas. Far from leading geologists into serious error, the theory has been the point of departure for huge advances in the understanding of earth history. It is a cornerstone in a unified framework for the interpretation of data from biogeography, geophysics and geology.

Clues to the Functioning of the Immune System

The *bursa of Fabricus* is a small sac at the tail end of the digestive tract in birds. In the 1950s two graduate students, Glick and Chang, discovered that this organ has a vital role in the production of antibodies. Glick, who had been unable to find any effect from the removal of the organ, gave his chickens to Chang for a class

demonstration of the production of antibodies. The demonstration failed, a result of the surgical removal of the bursa while the chickens were still very young. A paper that described their finding was rejected by the journal *Science* as "uninteresting". It finally appeared in the journal *Poultry Science*, where it went unnoticed for many years. After it did finally come to attention, it became in due course the most quoted paper ever to appear in that journal (Clark 1995, p.42.) It marked the beginning of fundamental discoveries regarding the immune system.

There are many reasons why a good idea may be slow to gain acceptance. The forces of conservatism can act just as strongly in scientific communities as in other communities. The word of one dominating and influential figure may be enough to prevent a hearing. "How dare you challenge my authority?" While it is the force of the argument that should prevail, not the pronouncements of elder statesmen, this may not be what happens.

8.2 Data and Theory

> Science is different from many another human enterprise – not of course in its practitioners' being influenced by the culture they grew up in, nor in sometimes being right and sometimes wrong (which are common to every human activity), but in its passion for framing testable hypotheses, in its search for definitive experiments that confirm or deny ideas, in the vigour of its substantive debate, and in its willingness to abandon ideas that have been found wanting. If we were not aware of our own limitations, if we were not seeking further data, if we were unwilling to perform controlled experiments, if we did not respect the evidence, we would have very little leverage in our search for truth.
> [Sagan 1997, The Demon-Haunted World, p. 252. Headline Book Publishing, London.]

Data

Data are crucial to science. Up until the 20th century a prevailing view was that science was generalisation from data. The name given to this process of generalisation is *induction*, which contrasts with *deduction* as used in mathematics and logic.

The view of science that emphasised induction and generalisation from data was strongly influenced by Francis Bacon, who in 1620 published a book that argued for a new method of research that, as he claimed, gave 'True Directions Concerning the Interpretation of Nature'. In Bacon's 'improved' plan of discovery, laws were to be derived from collections of observations. (Silverman 1985.)

Theory

Scientists do not collect any old data. They collect the data that seem most useful. How do they get this sense that some data will be helpful, and other data of little use? For example a study of the effects of passive smoking is likely to look for specific effects, most likely effects that are known to be a result of active smoking. One would not expect to find that passive smokers have an unusually high number of ingrown toenails! So we will not waste effort on gathering data on ingrown toenails. We will examine the occurrence of lung cancer, bronchitis, heart disease, and so on, but not ingrown toenails. There's no theory to suggest that smoking of any kind might cause ingrown toenails.

For studying the health of children living in some area of New Guinea, one might collect data on age, sex, height and weight. Hair colour and eye colour are unlikely to be of interest, for this purpose. It seems obvious that height and weight are important indicators, but that hair and eye colour are unlikely to be relevant. It is assumed that some measures are useful and some are not. There is an extensive literature that provides guidance on what measures other workers have found useful, which sets out

"theory" that anyone who now undertakes collection of data on the health status of one or other human group will want to note[22]. Those who initiated work in this area had to make their own judgments on measures that seemed useful indicators of health status.

Any adequate understanding of science must have regard both to theory and to data. Researchers do not collect any data. Data collection is driven by a judgement of what is worth collecting. It is in this sense that theory drives scientific research. None of the great scientists have followed Bacon's prescription. Typically they showed unusual insight, aided sometimes by good fortune, in the data that they collected. Data may carry within themselves the power to challenge and perhaps destroy the theory that guided their collection. It is this that gives science its power. Statistical insights and approaches have a key role both in data collection and the extraction of information from data. They assist in the efficient choice of data, in teasing out pattern from the data, and in distinguishing genuine pattern from random variation. The pattern may be as simple as a difference between the means of two treatment groups, or a linear relationship between two variables.

This is a convenient place to introduce the idea of a 'model'. This is an important idea, both in science generally and in statistics.

8.3 Models

Consider the formula for the distance that a falling object, starting at rest above the earth's surface, moves under gravity in some stated time. The formula is:

$$d = \tfrac{1}{2} g t^2$$

where t is the time in seconds, g (≈ 9.8 m/sec/sec) is the acceleration due to gravity, and d is the distance in metres. Thus a freely falling object will fall 4.9 meters in the first second, 19.6 meters in the first two seconds, and so on. This formula describes the way that objects fall.

Observing the fall of a stone (especially if you happen to be underneath) is a different experience from encountering the formula on a piece of paper. There are important aspects of the fall about which the formula tells us nothing. It gives no indication of the likely damage if the stone were to strike one's foot. The formula can tell us only about the distance traversed in a given time, and other information that we can deduce from distance information.

Watching the stone fall and making measurements is different from doing calculations using the formula. The results will not be quite identical, if only because of the limits of accuracy of the measurements. The formula is a model, not the real thing. It is not totally accurate – it neglects the effects of air resistance. For the limited purpose of giving information about distance fallen it is, though, a pretty good formula. As Clarke (1968) says: "Models and hypotheses succeed in simplifying complex situations by ignoring information outside their frame and by accurate generalization within it."

A good model captures those aspects of a phenomenon that are relevant for the purpose in hand. A model is, inevitably, an incomplete account of the phenomenon. The reward for simplifying by ignoring what is irrelevant for present purposes is that the model is tractable – we can use it to make predictions.

There are also non-mathematical models. An engineer may build a scale model of a bridge or a building that is to be constructed. Medical researchers may speak of using some aspect of mouse physiology as a model for human physiology. The hope is that

[22] See for example chapters 7 and 8 in Little and Haas (1989).

results from experiments in the mouse will give a good idea of what to expect in humans. As those who know the history of such research understand all too well, animal medical models can be misleading. At best, they provide clues that must be tested out in direct investigation with human subjects.
The model captures important features of the object that it represents, enough features to be useful for the purpose in hand. An engineer can use a scale model of a building to show its visual appearance. The scale model might be useful for checking the routing of the plumbing. The model will be almost useless for assessing the acoustics of seminar rooms that are included in the building.

8.4 Regularities (Law-Like Behaviour)

Mathematical models describe law-like behaviour, i.e. one can use the model to describe or predict. The falling object formula predicts distances.
We take a variety of regularities for granted in our everyday lives. We expect that the sun will rise in the morning and set in the evening. We expect that fire will burn us, and so on. These expectations have nothing to do with logic. They are based on our experience of the world. We take such regularities for granted.
There is no logical reason why what has happened in the past will continue to happen in the future. There is no logical reason why the sun should continue to rise. Fortunately for humans, it does! Indeed, it is impossible to carry on our lives unless we do take such regularities for granted. We speak of law-like behaviour. The process by which we generalise from our experience of the world to rules that tell us what will happen in the future is called induction. Inductive science looks for regularities in phenomena.
The natural sciences look for very wide regularities. They have found a huge range of phenomena, many of them outside of the range of our everyday experience, that exhibit law-like behaviour. There has been more limited success in finding law-like regularities in the biological sciences. In the social sciences there has been very limited success in finding law-like behaviour.
The nature of the social sciences makes law-like behaviour hard to find. The phenomena are more complicated. Consider the complicated processes that are at work to make some people criminals, and some law-abiding citizens. The relatively simple falling object equation is a striking contrast with our incomplete understanding of the `forces' that work to make some people criminals. Typically there are many effects at work. It is impossible to do experiments or make observations that separate these effects out individually. The processes are almost certainly different for different individuals. While it is possible to say that children who suffer severe neglect or abuse are much more likely to become criminals, this is just one of many different factors that are at work. We cannot explain why criminal behaviour is a much greater problem in some societies than in others.

8.5 Statistical Regularities

Statistical regularities rely on probabilistic forms of description that have wide application over all areas of science. In studying how buildings respond to a demolition charge, there will be variation from one occasion to another, even for identical buildings and identically placed charges. There will be variation in which parts of the building break first, in what parts remain intact, and in the distance and direction of movement of fragments.
Deterministic models, i.e. models that do not use probabilistic or statistical forms of description, have a place, especially in the physical sciences. Statistical variability is

often so small that it can be ignored. In the natural sciences however, statistical variation is ubiquitous and statistical forms of description are generally essential. No two animals or plants or humans are identical.
Statistical models typically have at least two components. One component describes deterministic law-like behaviour. In engineering terms, that is the *signal*. The other component is *noise*, i.e. statistical variation. Here is an example. Different weights of roller are rolled over different parts of a lawn, and the depression noted[23]. What we find is:

	Weight (t)	Depression (mm)	Depression/Weigh t
1	1.9	2	1.1
2	3.1	1	0.3
3	3.3	5	1.5
4	4.8	5	1.0
5	5.3	20	3.8
6	6.1	20	3.3
7	6.4	23	3.6
8	7.6	10	1.3
9	9.8	30	3.1
10	12.4	25	2.0

Table 3: Depression, and Depression/Weight Ratio, for different weights of lawn roller.

We might expect that depression would be proportional to roller weight. That is the signal part. The values for Depression/Weight make it clear that this is not the whole story. Rather, we have

$$\text{Depression} = b \times \text{Weight} + \text{Noise}$$

Here b is a constant, which we do not know but can try to estimate. The Noise is different for each different part of the lawn. If there were no noise, all the points would lie exactly on a line, and we would know the line exactly. In Fig. 4 the points clearly do not lie on a line. We therefore explain deviations from the line as random "noise", at least until some more insightful explanation becomes available.

[23] Data are from Stewart, K.M., Van Toor, R.F., Crosbie, S.F. 1988. Control of grass grub (Coleoptera: Scarabaeidae) with rollers of different design. N.Z. Journal of Experimental Agriculture 16: 141-150.

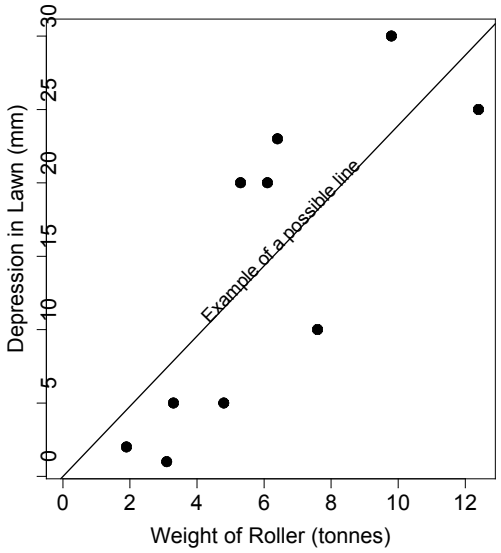


Fig. 4: Lawn Depression, for Various Weights of Roller, showing one possible line. The line is one of many that are consistent with the data.

We need a model for the noise also. We'll leave the details till later. Anyone who has done a first year course in statistics will expect to hear words such as *normal* and *independently distributed* used to describe the noise components. For now, let's call it a random term without spelling out the details.
It is a feature of statistical models that they have a signal component and a noise component. In some data the signal is strong and the noise small. In other data noise may dominate the signal. Fig. 5 illustrates the range of possibilities:

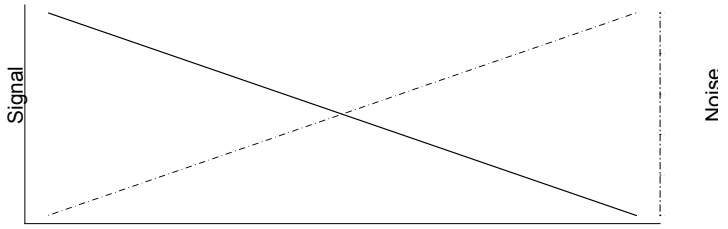


Fig. 5: Different positions along the horizontal axis correspond to different mixes of signal and noise. At the left extreme, there is only signal, while at the right extreme there is nothing except noise. Statistical models lie somewhere between these extremes.

We would prefer to get rid of the noise altogether. That is not a totally silly idea. While we cannot get rid of the noise altogether, we may be able to reduce it. There are several ways in which we might be able to do this:
1. By using more accurate measuring equipment.
2. By improving the design of the data collection.

A skilled experimenter will get as near as is reasonably possible to the extreme left in Fig. 5. That is where every experimenter would like to be.
Question: In the lawn roller experiment, how might one reduce the noise, i.e. reduce the scatter about the line or other response curve?

8.6 Imaginative Insight

How do radically new theories arise? No doubt generalisation from data, i.e. induction, has a role. At most it can be only part of the explanation. There is a large

element of imaginative insight – the recognition that looking at the phenomena in some new way will perhaps simplify the description, or explain former anomalies. Trying to understand imaginative insight may not be much different from investigating the psychology of scientists.

There are however styles of investigation that provide fruitful ground for the exercise of imaginative insight, and styles that are likely to confuse and derail it. Thus a carefully conducted experiment usually provides much better raw material for the exercise of imaginative insight than does unsystematic experimentation and poor design. In the former case anything that is unusual or unexpected will stand out as different and demand investigation, while in the latter case unexpectedly large or small values may have a multiplicity of explanations.

An apple transport trial in which I participated (Maindonald 1986) illustrates how careful design helps highlight anomalous results. The trial had sufficient elements of careful design that those few crates where there was heavy bruising stood out as anomalous. We found that they were unstable, shearing first to one side and then to the other as the truck negotiated bends in the road. Our design had neglected what turned out to be the most important factor affecting apple bruising. Nonetheless, because we had controlled for other factors such as the condition of the apples, the effects of bin instability stood out clearly.

8.7 Science as Hypothesis Testing

> . . . in learning by experience . . . conclusions are always provisional and in the nature of progress reports, interpreting and embodying the evidence so far accrued.
> [R. A. Fisher]

Imaginative insight readily creates worlds of its own that may have little connection with reality. There is a place for imaginative drama, fiction, legend and myth, but not as part of science. So there must be severe checks on the exercise of imaginative insight. How do we keep imaginative insight in check, ensuring that what we claim to find is real rather than the product of a fertile imagination. Why should we believe scientific explanations for patterns in the frost, rather than the claim that "the fairies did it"? The difference, according to Karl Popper, is that genuinely scientific theories can be tested. Instead of starting with data, Popper starts with a theory. Popper has little to say on where scientific theories come from.

There must be a motivation for collecting data. There must be a sense that some data are worth collecting and some are not. Researchers who are unclear why they are collecting data, and are not selective about what data they collect, typically end up with data that are of little use. Effective researchers are highly selective about the data they collect. They seek data that will address the questions that are of interest to them.

Any legitimate scientific theory will make predictions. For example, Newton's gravitational theory predicts that the earth and other planets will move around the sun in elliptical orbits. This prediction seems to be born out by the observed facts. So Newton's theory survives that particular test[24].

A scientific theory will not be rejected just because it cannot explain particular observations or results from a particular experiment. Kuhn (1970) argues that for a new theory to replace an old theory two conditions must be satisfied

[24] It almost survives it. Later work found small anomalies in the orbit of the planet Mercury. Einstein's theory of relativity is required to give a completely accurate description of the orbit of Mercury.

1. There must be serious cracks in the old theory, i.e. important facts that the old theory does not explain.
2. A new theory must be available.

Why replace a theory, even one that has evident flaws, unless something better is available with which to replace it?
There are further issues:

3. When observations or an experiment give results that are contrary to a well-established theory, is it the theory or the experiment that is mistaken? There may have been a flaw in the experimental procedure.
4. Flaws in experimental procedure are especially common when one is working at the limits of experimental technology. It may be at these limits that theory has its most extreme test.
5. Often, a small modification to the theory may be enough to accommodate a newly discovered anomaly.
6. Scientists may be so deeply wedded to the existing theory that they refuse to accept the new theory. This is particularly likely if the new theory is itself incomplete, i.e. many of the theoretical details have not been worked out. There are many examples of this.

8.8 Strategies for Managing Complexity

Complex systems defy ready understanding. Easily the most successful scientific strategy has been to restrict attention to limited aspects of a system where simple models may work. Once the subsystems are well enough understood, the hope is that it will be possible to bring the separate pieces of information together to give a useful account of the total system.
This reductionist approach has been spectacularly successful in physical science, biology and medicine. As Wilson (1998, p.58) says, "Reductionism is the search strategy used to find points of entry into otherwise impenetrably complex systems."
In the end however, the aim is to describe and explain the rich complexity of the systems under investigations. There is no virtue in naïve simplicity unless it leads, finally, to insights that enable us to get a handle on the complexity.
In practical applications of science, this complexity may extend far beyond the specific issues that motivated the scientific study. As an example of this complexity, consider the salinity that has affected or is threatening huge areas of Australian farmland. There are a large number of scientific issues that bear on this problem, some of which I list below. However none of the studies that one might conduct under these individual headings will, on their own, give the information needed to address the problem. Somehow the information from these various sources must be brought together.

An Example – The Desertification of Australian Land

Over large areas of Australia the destruction of forests has removed the trees that formerly soaked up water in the soil, leading to a rise in the water table. Salts are naturally present in the soil, in some places in substantial quantities. Irrigation brings in further dissolved minerals. These remain after the water has evaporated and build up slowly, adding to what is already in the soil. As long as the water table is well below the surface, rain will wash any salts down into the ground water, where they are not a problem. Once the water table rises to close to ground level, it brings the salts with it. Trees that have been left standing, and other vegetation, die off. In the end, the land becomes unusable. Coram (1998) quotes an estimate of 120,000

hectares of land in New South Wales that was affected by dryland salinity in 1996, with a further 5 million acres considered to be at risk.
There are many individual components to any study of this salinity problem.

1. Extent of the problem: What is the present and expected future size of the land areas that are affected?
2. Vegetation Effects: What is the extent of continuing damage from new clearing of vegetation? What is the potential remediation role of new tree plantings? Is it possible to find tree species that will grow and survive in saline soil?
3. Irrigation practices: How much of the problem is the result of past and current irrigation practices? How might changes in irrigation practices assist remediation? How effective (and cost-effective) would it be to use bores to replace the use of water from irrigation channels?
4. Groundwater draining and pumping: Is draining and/or pumping of groundwater a viable potential remediation strategy in some areas? Which areas?
5. Engineering of irrigation channels: What effects (e.g. damage to adjacent roads from the build-up of salt in the soil and/or from waterlogging) arise from loss of water from irrigation channels? What engineering solutions (e.g. better lining of channels) are available?
6. Land use strategies: What changes in patterns of land use might assist remediation. The replacement of agriculture by forestry can be highly effective. Those crops are preferable that do not require heavy irrigation.
7. Flow-on effects: How much of the problem in one or another area is the result of practices in other areas, perhaps more elevated or perhaps upstream?
8. Ecology: What are the effects on fauna and flora? How would alternative remediation strategies affect fauna and flora?
9. Social issues: What steps will ensure that remediation measures do not unduly disadvantage individual communities?

Also open to scientific study are political and economic consequences, flowing both from the present degradation of land and from proposed remedies.
There must be strategies for gathering whatever information is needed under each of these headings, and for creating from them an integrated plan of understanding and action. Questions worth considering are:

1. Are there changes that would be easy and cheap, and that would make substantial inroads on the problem?
2. What changes, ignoring for the moment their costs, would make the largest inroads?

Questions: Why is it hard to get action on the degradation of Australian land that is a result of salinity? Are there no good strategies? Or is the problem an inability to implement the strategies that are available? Is the needed co-operative action too difficult for our society's political and economic structures?

8.9 Cause and Effect

It is one thing to establish a correlation between two variables. It is another to establish a causal link. The direction of causation is sometimes obvious. It is rain that causes the wheat to grow, not growth of wheat that causes the rain. Heavy drinking causes the subsequent hangover. But what is the relationship between hard work and business success. Does success come first, leading people to work hard to maintain and improve their position? Or does hard work come first. Often, both variables are driven by a third variable. Weight and height are strongly correlated, but it makes no sense to claim that one causes the other. These issues have generated fierce continuing debate in the social science literature. References in Freedman

(1999, p.248) represent a range of perspectives. See also pp.78-80 of Greenhalgh (1997).

Cause and effect issues have appeared at several points earlier in these notes. Does salt in the diet cause high blood pressure? Does an increase in the minimum wage cause reduced employment? What long-term effects flow from sudden and unexpected traumatic loss?

Claims of causation are convincing when there is a cogent theory that establishes the causal chains of connection. Where the theory is complex, built from many individual components, those components must be open to testing. Complex theories must often rely on computer modelling to link the separate components. One example is the extensive body of theory that is designed to predict the global climatic impacts of human activity. Some might argue that it is a complex of theories rather than a single theory. This is a matter of definition.

8.10 Computer Modelling

> Many of the new biological challenges are of the "how do we put the pieces back together" type. Those problems are horrendously difficult for our current approaches.
> [Wilson, 1998, pp.91-92.]

Human impacts on climate change are a serious issue for our time. For science it is a huge problem of the "how do we put the pieces back together type". Many different sources of information and evidence must come together. Computer modelling seems the only viable approach.

Increased atmospheric levels of carbon dioxide and other implicated gases[25] increase the effectiveness of the earth's atmosphere as a heat shield. Much of the focus has been on increases in carbon dioxide levels that have resulted from increased use of fossil fuels. A 0.5°C average global increase in the temperature of the earth over the past century seems in part due to this and other human activities. Schneider (1996) reports an assessment of tree-ring and other evidence for temperature change in the past ten thousand years that suggests that such a large 100-year change has been unusual over this time, occurring no more than once in a thousand years. See also Crowley (2000).

Projections drawn up by the Intergovernmental Panel on Climate Change predict an average global warming of between 1.0°C and 3.5°C over the next century, a greater rate of climate change than at any earlier time in the past 10,000 years. Predictions are that sea levels will rise, some low-lying areas will be covered by sea, there will be loss of vegetation, farmers may need to change to new crops that are viable in the new climatic conditions, weather patterns will be less stable, and tropical diseases will affect many sub-tropical regions.

How were these figures obtained? It is not sensible to try to project current temperature trends into the future. The world's climate has changed continuously over time, making short-term trends a poor guide to what may happen in the future. Rather the evidence comes from computer modelling. The predictions from this modelling are unequivocal – present rates of release of CO_2 into the earth's atmosphere will lead to a temperature increase. If these rates continue to increase at about 1.5% per annum as in the recent past, the temperature increase over the next 100 years will be correspondingly larger.

Atmospheric and ocean currents are the moving parts of a huge engine that is driven by the sun's heat. The blanketing effect of the atmosphere, itself affected by life processes on land and in the sea and by human activities that include the use of fossil

[25] Other gases that are implicated are methane, nitrous oxide and hydrofluorocarbons.

fuel, are a part of the engine's control mechanisms. Understanding of the functioning of the individual components seems adequate for the building of computer models that make gross predictions, always assuming that ocean (and air) currents continue to follow pretty much their current patterns of movement. A worrying aspect of potential large temperature changes is that they may cause the engine to reconfigure itself. Changes in the flow of major ocean currents, such as have happened in past geological times, would bring changes in climate patterns that would be even more traumatic.

Computer models must accommodate, as best they can, all these different effects. Statistical methodology has a clear role in checking the predictions of individual components against experimental and observational data. Checks that model predictions over several years for different regions of the earth's surface agree with observation are encouraging, but not clinching evidence. By the time that clinching evidence of the accuracy of model predictions is available, the damage will be irreversible. Hence the importance of close critical scrutiny of the separate components of the models, of the way that those components are linked and of sensitivity analyses that check how predictions would change if there were changes to those model assumptions that are open to challenge.

Scientists from many different disciplinary backgrounds have critically scrutinised the computer models. There has been extensive refinement of the details. Qualitative model predictions have withstood these criticisms remarkably well. The most persistent criticism has come from those with a political axe to grind, usually in defence of inaction! Such critics have the option, and the challenge, to build and offer for scientific scrutiny models that give predictions that are more to their taste.

8.11 Science as a Human Activity

> I know that most men, including those most at ease with problems of the greatest complexity, can seldom accept even the simplest and most obvious truth if it be such as would oblige them to admit the falsity of conclusions which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives.
> [Tolstoy, quoted in Gleick, 1988.]

> [Scientific theories] . . . are constructed specifically to be blown apart if proved wrong, and if so destined, the sooner the better. "Make your mistakes quickly" is a rule in the practice of science. I grant that scientists often fall in love with their own constructions. I know, I have. They may spend a lifetime vainly trying to shore them up. A few squander their prestige and academic political capital in the effort. In that case – as the economist Paul Samuelson once quipped – funeral by funeral, theory advances.
> [Wilson, E.O., 1998, p.56]

Humans are not inherently rational creatures. Much of what passes for reasoned argument is rationalisation – the use of reason to defend positions that we hold for other reasons. An attitude of mind that judiciously balances openness to new ideas with rigorous critical scrutiny does not come easily to our human nature. Prejudice readily takes precedence over the demands of rationality. Scientists are not inherently different from other humans who are prey to idiosyncratic belief systems and spurious claims. Gilovich (1991) is one of many books devoted to the discussion of our irrational foibles.

Fallible Scientists

Scientists are not immune from the tendency to rationalise. Thus craniology – the measurement of the brain capacity, often with the aim of relating brain capacity to

racial differences – became a popular subject of study in the nineteenth century. Not surprisingly, much of this work collected and used data in ways that reflected the racial and sexual prejudices of the scientists who undertook it. Gould (1996), in a highly readable book, discusses this and other similar examples. Fortunately the processes of scientific criticism and re-evaluation do in the course of time tend to expose and correct such abuse. (Gould's book has itself attracted accusations of bias from academic critics.)

Still today, rationalisation and prejudice compromise science. New prejudices and new rationalisations have arisen to replace those that we hoped to have conquered. Such rationalisations find it especially easy to establish and retain a foothold in those areas where there is a dearth of external checks on the exercise of imaginative reconstruction. Dogma easily masquerades as science.

Researchers may become more concerned about maintaining their funding or their position within the profession than about truth. Science easily degenerates, in some times and some corners, into pseudo-science. There is self-deception, there is an often exaggerated deference to authority, there is deliberate manipulation, and there is a yielding to self-interest. There is a challenge to devise ways of funding and directing scientific research that reduce opportunity for manipulation, for deviousness, and for prejudice and dogma that masquerade as science.

Different scientists have different qualities. Some may be receptive to new ideas, but not good at criticism. Others may be good at criticism, but not receptive to new ideas. They may apply high standards of criticism in their own area, but make idiosyncratic judgments when the scientific demands change. They may be hypercritical, not understanding the different nature of the evidence that the new and unfamiliar area demands. Or, failing to note the different opportunities for self-deception that this new area offers, they may be unduly credulous. There are few who can examine claims in medicine or social science or physics with more or less equal critical incisiveness.

Dominant authorities

As in all communities, there are some whose pronouncements carry especial weight, or whose positions give them special authority. They may be editors of major journals, or have a large influence in the decisions of funding agencies. There are practical reasons for listening to the voices of such dominant figures. Their judgments can be effective in weeding out ideas that are not worth pursuing. At the same time they may weed too ruthlessly, their own speculative notions may acquire the force of dogma, and they may resist anything that they find too novel. This may be a particular danger if there are just one or two dominant figures — individuals who occupy the sort of position that Harold Jeffreys occupied in geophysics in the 1950s. It is healthier if the dominant figures do not altogether agree among themselves.

Jealously and backbiting also flourish. Other scientists may be seen, not as partners in a common endeavour, but as threats to one's own enterprise who must be cut down by any means available. Political concerns may influence scientific judgements. Even if such attitudes are not overt, they may lurk below the surface. Perhaps we should be surprised that the demands for scientific rationality do so often prevail over these human influences. Only an overarching insistence on rigorous criticism can keep science from becoming prey to irrationality. There will never be total success. There is however plenty of scope for improvement on the way that science is now conducted.

The Logic of Science and the Sociology of Scientific Communities

Above I noted conditions that, according to Kuhn, must be satisfied before a new theory can replace an existing theory. There must be serious cracks in the existing theory, and a new theory must be available.

However Kuhn goes further. He argues that science is driven by powerful social forces, akin to those that drive other human activities. An objective examination of the history of science shows much that confirms Kuhn's claim. A weakness in Kuhn's account is that he does not maintain a clear distinction between the logic of scientific discovery and the sociology of scientific communities[26]. Science has an inherent logic that does often, in the course of time, prevail against the sociological forces that drive one or another scientific community. At least in the physical and biological sciences, it is unusual for reactionary attitudes to hold back progress for more than a decade or two. Individuals who show unusual insight may be denied their PhDs. Their ideas, if they withstand critical scrutiny, do however finally prevail. This is a remarkable feature of scientific discovery. A science that was wholly the product of social forces would be ineffective.

The sociology of scientific communities often works against really good science. I will criticise unhelpful practices, in data collection, in data analysis and in the reporting of results, that are undesirable outgrowths of the sociology of particular scientific communities. My complaint is that they are contrary to the inherent logic of science. Some common failings are:

- uncritical reliance on expert opinion
- exaggerated expectations of what can be learned from observational data
- failure to marry subject area insights with results from statistical analysis
- deficiencies in data-based overview
- unwillingness to bring in other skills when these are clearly required
- deference to pressures from commercial interests.

Reductionist Scientists?

Scientists who wish to publish extensively and advance in their chosen research area will do well to limit their attention to a narrow range of problems that seem likely to yield easily to their skills. This narrowness of focus, which can be beneficial in making initial progress in a closely defined area of research, does not give the breadth of view needed to tackle "big issue" questions. Determining the structure of an organic chemical compound found in the river water, or using radio-isotopes to trace its progress through the river system, does not of itself give the breadth of view needed to tackle such "big picture" problems as dry land salinity.

Wilson (1998, p.40) has apt comments:

> The vast majority of scientists have never been more than journeymen prospectors. That is even more the case today. . . . They acquire the training needed to travel to the frontier and make discoveries of their own, and as fast as possible, because life at the growing edge is expensive and chancy. The most productive scientists, installed in million-dollar laboratories, have no time to think about the big picture and see little profit in it.

The skills of a "journeymen prospector" may serve well those who expect to join multi-million dollar research laboratories. A narrow training focus seems clearly inappropriate for anyone whose work is likely to demand skills different from those of their Ph.D. or other research degree, or who is likely at some time to work on "big picture" issues.

[26] For a recent wide-ranging critique of Kuhn's views, see Fuller (2000).

Commercial Pressures

Money speaks volumes. Commercial pressures may be a potent influence. Wilkinson (1998) offers a series of case studies that highlight some of the issues. Edmeades (2000) is an interesting study of the aftermath to a celebrated defamation claim that occupied the New Zealand High Court for 135 days. What were the rights and duties of fertiliser scientists who wished to make the results of their research available to the farming community that they had a responsibility to serve?

The Uses of Controversy

Controversy can be helpful in drawing attention to areas of weakness in the science. It offers an interesting window both into the sociology of scientists and into the logic of scientific discovery. It is an advantage when the different parties to the controversy come from different disciplines, and accordingly offer different perspectives. Novice researchers sometimes find themselves caught, uncomfortably, between the different sides of a controversy. From time to time the views of a PhD examiner will, in spite of care in the choice of supervisors and examiners, be seriously at odds with the ideas and insights that shaped a smaller or larger part of the thesis. It is with these points in mind that I will now comment on controversies that have surrounded the study of human abilities and human nature.

8.12 The Study of Human Nature and Abilities

> Know then thyself, presume not God to scan,
> The proper study of mankind is man.
> [Alexander Pope (1688-1744): An Essay on Man.]

The scientific study of human nature and abilities is a sensitive area, for all sorts of reasons. Are humans able to pursue such studies objectively, with the detachment that science demands? Supposed scientific objectivity readily becomes a vehicle for particular prejudices.

The Heritability of IQ

Studies of the genetic basis of IQ have had a long and tangled history. A key and greatly overplayed concept has been the *heritability* coefficient, the proportion of variation (measured using the statistical variance) that is due to genetic variation. The heritability coefficient has been widely used in animal and plant breeding studies, where the outcome variable of interest has been weight or milk production. A high heritability suggests a potential to get further improvements from breeding. Comparison between heritability coefficients from different trials makes sense only if environmental variation is comparable. This may be reasonable if, as in many animal and plant breeding studies, conditions are similar across different trials.

Studies of twins, both identical and non-identical and including separated pairs, have been the main source of evidence for the heritability of IQ in human populations. As one might expect, the two members of a separated pair are often reared in very similar circumstances, more similar than for two randomly chosen members of the population. Thus the studies tell us nothing about heritability in a section of the population where the range of social disadvantage is large. Lewontin (1979) has argued, rightly in my view, that

> . . . there is no way in human populations to break the correlation between genetic similarity and environmental similarity, except by randomised adoptions.

One would need to randomly assign adoptees to the whole range of social circumstances to which it was intended to generalise results. Such an experiment is surely out of the question.
There is a further issue. Twins share a common maternal environment. Daniels et al. (1997), in a meta-analysis of more than 200 studies, estimate that the shared maternal environment of twins accounts for 20% of the total variance. The ignoring of this component in earlier analyses of data from twin-adoption IQ studies led to a substantial over-estimate of heritability. Assigning to the wrong source a component that turns out to be 20% of the total is perhaps excusable in the initial tentative investigations. Long before one has the 212 sets of results that Daniels et al. analysed, this surely has acquired the status of a fundamental biological error! This analysis still leaves large questions unanswered. What is the relevance of these studies, if any, to a wider population where the range of environmental effects may be far larger than those typically experienced by the separated twins?
IQ tests capture a small part of the rich texture of human abilities. Mental and other abilities continue to change and develop through into old age. *Mind Sculpture* (Robertson 1999) is the title of a book that discusses evidence on how our brains develop and change as a result of demands placed on them. The emphasis should perhaps move from the study of mental testing to the study of *mind sculpture*.

Sociobiology

In his 1975 book *Sociobiology: The New Synthesis*, Wilson defined sociobiology[27] as "the systematic study of the biological basis of all social behaviour". Wilson hoped to find a genetic basis for behaviour. Sustained controversy followed its publication. While most of the book was devoted to the study of animal and especially insect societies, the final chapter speculated on genetic influences on human behaviour. Why all the fuss? The discussion that now follows draws at several points on the account in Segerstråle (2000).
Any initial foray into an area that is as complex as genetic effects on animal behaviour must over-simplify. But what if the simplifications that seem required are precisely those that readily feed into racial, sexual, national and other such forms of prejudice? Wilson was aware of the risks of the area into which he had ventured, and took care to protect his words from such misuse. His critics were not satisfied, either with his science or with the care that he had exercised. Criticisms were of several different types:

- Wilson was charged with specific scientific errors.
- Notwithstanding the generally liberal tenor of Wilson's views, it was argued that they lent support to those opposed to steps that would ameliorate the position of socially and economically disadvantaged groups.
- Criticism of Wilson's book became a convenient starting point for promoting wider scientific and political agendas. In some instances statements were taken out of context, charging Wilson with views that were at variance with specific statements in the surrounding text.

There is a succinct statement of the criticisms in Rose et al. (1984). Segerstråle attempts to disentangle the various strands of this controversy. It is worth noting that a wide spectrum of political views is found both among those who emphasise genetic

[27] Note also the more recent term *evolutionary psychology*, used to describe an area of study that has a large overlap with sociobiology.

influences on human behaviour and abilities, and among those who emphasise environmental effects.

The first tentative steps in a new area of study may use overly simplistic models, which will be refined as understanding advances. Problems arise when there are perceived implications for the way that we regard or treat fellow humans. There is a long history of misusing claimed scientific results that is the theme of Gould's *The Mismeasure of Man*[28]. Where such implications are perceived, it behoves scientists to tread with extreme care, to acknowledge obvious limitations in their models, and to acknowledge the tentative character of their results. This may conflict with the motivation that researchers feel to persuade themselves and others of the importance and significance of their work.

One outcome of the controversies in sociobiology has been a closer scrutiny of the scientific methodology than has been common in other areas of biology that rely extensively on observational data. This scrutiny needs to go further. Such statistical methodologies as regression are too often used uncritically, without regard to traps such as were discussed in section 5.2. Even if the models are correct, estimates of key parameters may be wrong.

References and Further Reading

Box, Joan Fisher 1978. Fisher — the Life of a Scientist. Wiley, New York.

Clark, W.R. 1995. At War Within. The Double-Edged Sword of Immunity. Oxford University Press, Oxford, UK.

Clarke, D. 1968. Analytical Archaeology. Cambridge.

Coram, Jane (ed.) 1998. National classification of catchments for land and river salinity control. Rural Industries Research and Development Corporation (Australia), no. 98/78.

Crowley, T.J. 2000. Causes of climate change over the past 1000 years. Science 289: 270-277.

Daniels, M., Devlin, B., and Roeder, K. 1997. Of genes and IQ. Chapter 3 of Devlin, B., Fienberg, S.E. and Roeder, K., eds., Intelligence, Genes and Success. Springer, New York.

Diamond, J. M. 1997. Guns, Germs, and Steel: the Fates of Human Societies. Random House, London.

Edmeades, D.C. 2000. Science Friction. The Maxicrop Case and the Aftermath. Fertiliser Information Services Ltd., P.O. Box 9147, Hamilton, N.Z.

Fuller, S. 2000. Thomas Kuhn: A Philosophical History for Our Times. University of Chicago Press.

Gilovich, T. 1991. How we know what isn't so. The Free Press, New York.

Gleick, J. 1987. Chaos: making a new science. Viking, New York.

Gould, S. J., revised and expanded edition, 1996. The Mismeasure of Man. Penguin Books.

Greenhalgh, T. 1997. How to Read a Paper: the basics of evidence-based medicine. BMJ Publishing Group, London.

Hallam, A., 2nd. edn 1989. Great Geological Controversies. Oxford University Press.

Harré, R. 1967. The principles of scientific thinking. In Harré, R., ed.: The Sciences. Their Origin and Methods, pp. 142-174. Blackie and Son Ltd., Glasgow.

Jeffreys, H. 1926. The Earth, its Origin, History and Physical Constitution. Cambridge University Press.

Kuhn, T., 2nd edn, 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lewontin, R.C. 1979. Sociobiology as an adaptionist program. Behavioural Science 24: 5-14.

Little, M.A. and Haas, J.D., eds. 1989. Human Population Biology. A Transdisciplinary Science. Oxford University Press.

Maindonald, J. H. 1986. Apple transport in wooden bins. New Zealand Journal of Technology 2: 171-176.

[28] Gould's account has itself attracted strong criticism from a number of academic reviewers.

Robertson, I. H. 1999. Mind Sculpture. Bantam, London.
Sagan, C. 1997. The Demon-Haunted World. Science as a Candle in the Dark. Headline Book Publishing, London.
Schneider, S.H. 1996. Laboratory Earth. The Planetary Gamble We Can't Afford to Lose. Weidenfeld and Nicholson, London.
Segerstråle, U. 2000. Defenders of the Truth. The Battle for Science in the Sociology Debate and Beyond. Oxford University Press, Oxford.
Silverman, W. A. 1985. Human Experimentation. A Guided Step into the Unknown. Oxford University Press, Oxford.
Taubes, G. 1998. The (political) science of salt. Science 281: 898-907 (14 August).
Wilkinson, T. 1998. Science Under Siege: The Politicians' War on Nature and Truth. Johnson Press, Boulder CO.
Wilson, E.O. 1975. Sociobiology: The New Synthesis. Harvard University Press, Cambridge MA.
Wilson, E.O. 1998. Consilience. The Unity of Knowledge. Abacus, London.

9. Critical Review

> To give a basis for independence of judgement it is, I believe, of far more importance than is generally supposed that the worker should allot a considerable fraction of his working time to making himself acquainted with the published literature. . . . The student's reading may have been well directed, but it has covered almost certainly only a small fraction of the published researches bearing on his problems. The junior worker should receive encouragement, and his duties should allow him to read, with adequate care, far beyond the limited series of papers which his chief may indicate to him as necessary for the work of his department. The object should be to familiarise the reader with the stages whereby current opinions have been developed, and to train him, by scrutinising the results of past experimentation, to exercise his own judgement on the value of the experimental evidence available on different disputable points.
> [Fisher, R.A., in Bennett 1989.]

Critical review of previous research is the appropriate starting point for a new study. The aim is, as far as possible, to start from what is already known. New research should build on and learn from what others have or have not done. Look also for other ways of getting a research consensus, such as talking directly to `experts'.

The principles of critical review have wide application. They are pretty much those of evidence-based medicine. One can apply them to the use of medical advice, and one can apply them to research.

Statistical insights are often crucial in assessing the literature. Not all studies are of equal quality. It is necessary to decide, as objectively as possible, which studies are relevant and of high quality. This requires careful and critical scrutiny of each individual study. Watch for confounding, i.e. more than one explanation is available for an observed effect. Ask about possibilities for bias? Ask whether the study had sufficient precision to detect the effects that are of interest. Influences on precision include measurement instruments, experimental or sampling design, and sample size. Inadequate description of methodology may be a warning sign of methodological inadequacies. Do not automatically give authors the benefit of the doubt.

Some researchers must contend with a large number of papers bearing on their chosen topic. A first step is to determine whether someone else has already done a thorough competent critical overview. If one or more overview studies are available, how careful and reliable are they? Are studies that show a clear effect more likely to be represented? What are the possibilities for bias? Is it possible that an effect that has shown up in a data-based overview is the result of a similar bias that has affected all studies?

Before starting one's own research, there should be a good sense of what other workers have achieved. It is well to be sure than any new piece of research has a good chance of providing new, relevant, information. In some instances the research supervisor may provide a research question that he/she is sure is unworked ground. At the other extreme, it may be impossible to decide on a sensible research question until one has canvassed the state of existing knowledge, and examined openings for new research.

The examination of existing data may be a desirable preliminary to the gathering of new data. A first step will be to examine the highly summarised data that appear in published papers. If access is then needed to original data, this may not be easy to

get. In rare cases, the data may already be available from an internet site. Some researchers are meticulous in keeping their data on file, while some are not. Some make their data freely available to other workers, while others may not.

9.1 A Springboard for New Research

Canvassing the state of existing knowledge may involve reading and digesting a small number of relevant papers. Or it may require getting a grip on a huge literature. If the amount of literature is large, then one needs to look for ways of getting quickly to the nub of the matter. Even dealing with the highly summarised data that appear in the published literature may be a major task.

All studies are not of equal quality. One must decide, as objectively as possible, which studies are relevant and of high quality? One needs to strike a balance between undue scepticism and taking at face value everything that appears in the published literature. Watch for vagueness in the description, and for claims that are made without giving the rationale. Inadequate description of methodology is often a warning sign of methodological inadequacies. Do not automatically give authors the benefit of the doubt.

If an experiment, do the authors describe their experimental design? Do they describe the manner in which the analysis reflects the experimental design? Do they describe their sampling design? Do they describe the steps that they took to minimise non-response? Do they describe their analysis in enough detail that anyone with the appropriate competence could repeat the analysis? Does their analysis reflect the sampling design?

If the authors of a report on a clinical trial are vague about how they handled the randomisation, or how they handled dropouts, it may be that the protocol was inadequate in these respects. Carefulness in giving complete information, on study design, execution, sample sizes, relevant effect sizes and relevant standard errors, may be matched by carefulness in other aspects of the study. Vague descriptions of the experimental design and field layout in agricultural field trials may likewise be an indication that design issues have not been thought through and hence that the design may have been inadequate.

Careful authors will give graphs that demonstrate that models are a reasonable description of the data. They will check, to the extent that current technology allows this, whether covariates really do have a linear effect. They will give an assessment of effect sizes and standard errors. They will be careful to say which other factors or variables are held constant for purposes of this assessment.

I have found a surprising number of instances where authors have fitted straight lines to data that are clearly non-linear. They may present a graph from which the reader can draw his/her own conclusions. Or they may give data that the reader can use to draw his/her own graph.

The use of correlations in place of regression lines, and of R^2 when one would really like to know the accuracy of prediction, are bad signs. Extensive quoting of p-values is sometimes a recourse when authors cannot think what else to do. What are the statements that these p-values support? Are these points of consequence? Consider whether there should have been some global test of significance, rather than many individual p-values.

Be sceptical of meta-analyses that do not examine trial quality, and/or that lump together results from different types of trials. Have the authors been meticulous in their search for all relevant papers? Have they searched for unpublished studies?

The skills that are needed for critical evaluation of published research papers are not much different from the skills needed for critical evaluation of what you read in the newspapers. Try practicing those skills when you read the daily paper or watch reports on television! Appendix I has a checklist for use when reading published papers.

9.2 Is Salt Bad for Health?

In chapter 3 we made extensive reference to an overview of the evidence that appeared in Science (Taubes 1998). It is unlikely that a novice researcher would directly tackle the question: "Is salt consumption of around 10gm/day bad for blood pressure?" One might however tackle a more limited question, aimed at teasing out some related issue that existing studies have not settled.

Points that emerge are:

1. Different experts have held very different views.
2. These different views are, in part at least, a result of reliance on different types of study.
3. The statistical analyses that are used in at least two of the overview studies to which Taubes refers have been challenged as seriously flawed.

The Taubes article highlights, in a severe form, the problems that can be involved in coming to terms with the existing literature and with existing expert opinion. Few beginning researchers will find themselves faced with such a plethora of deeply entrenched opinions, all able to claim the support of their own preferred choice of research evidence. On the other hand, few beginning researchers will have such a rich resource of existing literature and review papers on which to draw. One cannot have it all ways!

9.3 The Importance of Overview

Too often, statistical analysis fails to place the analysis of data in context. Where multiple sets of data are available that bear on the same question, they are analysed separately. If results are to be generalised, it follows that they must be valid for multiple sets of data. Ehrenberg (1990) makes this point forcefully. It is thus important to design data collection so that we can demonstrate repeatability, a point made in Lindsey and Ehrenberg (1993). Hubbard and Armstrong (1994) found that replication had been unusual in the published marketing literature. In the few instances where there was an attempt to replicate, over half the results contradicted the original study. Chatfield (1995) makes the comment: "If a result is not worth replicating, it is not worth knowing".

These are the sorts of reasons why multiple studies, and the use of data overview to form an overall assessment of their evidence, have become highly important in clinical medicine. Is it beneficial to inject albumin into patients who come into critical care, in order to stabilise them? Among the many studies that bear on this question, some seem to support injection with albumin, and some to argue against it. The weight of the evidence is that albumin increases the risk of death. (See Cochrane Injuries Group Albumin Reviewers 1998.)

Researchers who contend with a large number of papers bearing on their chosen topic must somehow get an adequate overview. Overview may be informal, largely supported by qualitative judgements. Or it may follow approaches that have been developed by specialists in the art of overview, and may be supported by quantitative analysis. In either case the file drawer problem is a concern; how complete a sample does the published literature provide of the evidence? Typically, studies that show an effect are over-represented among those that find their way to publication. Also, with high apparent precision available from the meta-analysis of a large number of trials,

any systematic bias that affects a large number of the trials becomes important. Any reviewer needs to pay attention to possibilities for bias.
You may find that one or more overview studies are already available in the literature. You then have the task of assessing the quality of this work. What have the authors of any overview studies done to attend to the difficult issues noted in the previous paragraph?

The Demand for Data-Based Overview (Systematic Overview)

Data-based overview places the individual studies under critical scrutiny, and places them in context. Here is an example from field crop studies. In a recent review of yield-density studies on green asparagus, Bussell et al. (1998) found large differences within the same locality. Based on commercial experience, it is likely that fertilizer and soil effects, and variety, were the main factors explaining yield differences between trials. Information on relevant factors was so incomplete that it was impossible to draw from the trials themselves any certain inference on factors affecting yield. Two only of the 15 trials gave any information on climate, irrigation and terrain. Four trials gave no information on soil type. The trials give benchmarks against which growers in a local region can compare their own yields. This aside, none of the recent trials have added anything of consequence to what commercial growers already knew – use a modern variety on a sandy or light silt loam soil, plant at the highest density that is practical, and use a fertilizer that is at least as effective as farmyard manure!
Above, I noted the problems that the 'file drawer' problem creates for data-based overview. Results from a proportion of research studies do not find their way through to publication; they remain in the file drawer. Unless a register is kept of all studies, as happens in some jurisdictions, it may be difficult or impossible to identify relevant studies. For those studies that are identified, it may be difficult or impossible to get access to raw data. In such areas as clinical medicine, an insistence on some form of international registration of trials at the time of commencement seems desirable. This would allow ready identification of all trials relevant to a particular overview study.

Data-based Overview – An Example

Human albumin solution has been used in the treatment of critically ill patients for over 50 years. There have been three recognised indications for its use — emergency treatment of reduction in blood volume as a result of shock, acute management of burns, and clinical situations associated with loss of protein from the blood. There are medical physiological reasons why administration of albumin might assist survival. But does it really help?

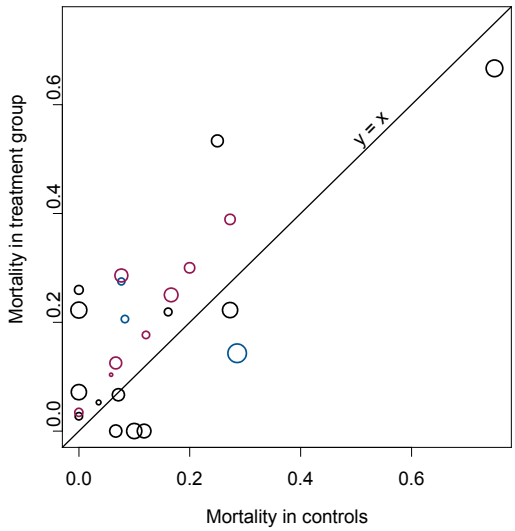


Fig. 6: Summary of results from 24 randomised controlled trials that compared an albumin treatment group with a control group. The diameter of the circle is proportional to the standard error for that trial. If there is no treatment effect, points will scatter about the line y=x.

Here we discuss results from a data-based overview (Cochrane Injuries Group Albumin Reviewers 1998). Fig. 6 presents results from the 24 trials in which there was at least one death, in either or both of the treatment or control group. These represent 1204 patients in all. The authors were thorough in their searching for information on randomised controlled trials. They searched various trials registers as well as international journals. They identified 30 trials (1419 patients) that met their criteria, and for which mortality data were available. There were two further trials (44 patients) where the mortality data were not available. All compared an albumin treatment with a control that did not involve albumin.

Fig. 6 suggests that, contrary to previous expectation, albumin may actually be dangerous to patients. Most trials, and almost all of the larger and hence more accurate trials, have points that lie above the line $y = x$, i.e. mortality was higher in the treatment group than in the control group.

A meta-analysis indicates that giving human albumin to patients in critical illness increases the risk of death, by around 1 death for every 17 critically ill patients[29] who receive albumin. The results of this study go against what had been received medical wisdom. They build a picture that was not available from any individual trial. Theoretical justifications for the use of albumin, based on its presumed ability to restore blood volume, have yielded to hard data.

The authors checked several possibilities for bias, to the extent that the data allowed it. Most of these were small trials; all except two had less than 40 patients. Small trials are not always conducted to the same standards as larger trials. Strict adherence to a pre-determined protocol is a necessity for a large trial, where in a smaller trial there may be less stringent planning and procedures. If this were a consideration here, one would expect the effect size to change with the size of the trial. It does not. Even so, the small size of most of the trials is a reason for interpreting results with caution.

[29] The 95% confidence interval was 9 – 32. This is a Number Needed to Harm (NNH) form of presentation of the results, which makes better intuitive sense than relative risk. The estimated relative risk from using albumin rather than an alternative was 1.68 (95% C.I. 1.26 - 2.23).

A further issue is that, in some of the trials, allocation concealment was inadequate or unclear. Is it possible that some clinicians did not follow the protocol strictly, giving albumin to more seriously ill patients? In order to check this, the reviewers did an analysis that excluded trials where the protocol may not have been strict. Exclusion of such trials made almost no difference to the estimates of relative risk.
There is now evidence that albumin has a variety of effects, some perhaps unhelpful. Interestingly, cohort studies that have measured the levels of albumin in the blood of seriously ill patients have shown that the risk of death reduces with increasing levels of albumin. Fig. 6 suggests that it is dangerous to add to the albumin that is already present.

Systematic Overview in Medicine

In clinical medicine, *systematic review* is a name for data-based overview. It has been a strong emphasis in Clinical Epidemiology and related areas of medicine. Its approaches to the summarization of evidence are useful models for other areas. Systematic Overview is a key methodology for the conduct of studies such as are fostered by the Cochrane Collaboration (Sackett and Oxman 1994), and for Evidence-based Medicine (Sackett at al. 1997; ScHARR 1998; Moynihan 1998, pp. 213-241). Smith (1996) asks how an 'evidence-based' human society would conduct its business. Cochrane type evidence bases are required in many other areas than medicine.
Lessons from experience with medical databases are highly relevant to efforts now under way to collect other types of data, often from disparate sources, in large databases. Draper et al. (1990) describe areas where data-based overview is important. An interesting application is to the estimation of physical constants. Data-based overview seems especially important when the literature is extensive, uneven in quality and different biases may be associated with different types of study.
The advice and insights of evidence-based medicine are in the first instance directed towards medical clinicians. The publisher's blurb for the journal *Evidence-Based Medicine*, directed to clinicians, argues:

> With 2 million new papers published each year how can you be sure you read all the papers essential for your daily practice, and how can you be sure of the scientific soundness of what you do read?

Researchers have the same interest as clinicians in getting a sense of the conclusions that ought to be drawn from studies to date, as a starting point for their own research. Systematic overview identifies secure knowledge and highlights gaps in research-based knowledge.
A particular widespread gap in clinical medicine is in evidence that would assist in tailoring treatments to the special requirements of individual patients. Some papers may have no information on a key covariate, e.g. baseline blood plasma zinc levels in a zinc supplementation trial. Too many papers focus on single end-points where the interest should be in the response profile, i.e. in the pattern of response over time.
There may be several overview studies from which to choose. Just as some papers are so flawed that they merit scant attention, so also for overview studies. Advice and training is needed that will help discriminate the good from the bad. Sackett et al. (1997) and Greenhalgh (1997) emphasize this point, and give advice on the critique of overview studies. See also Chalmers and Altman (1995). If no up-to-date and clearly authoritative study is available, the researcher's first step must be to attempt his or her own overview.

The demands of data-based overview studies that meet Cochrane Collaboration standards are severe. It may be easier, though less useful, to do a new study than to undertake a fully adequate overview of existing studies. The technical demands are such that Cochrane Collaboration studies have so far covered only a small proportion of health care. The conduct of overview studies requires special skills that are different from or additional to those of subject area experts. There is evidence that subject area experts do a poorer job than non-experts with experience and skills in the conduct of overview studies (Oxman and Güyatt 1983.)
The perspectives of evidence-based medicine, and the importance of Cochrane Collaboration type studies, seem not to be widely recognized outside of medicine. Pressures for change may come from three sources:

1. Researchers in e.g. psychology or education who work on the borderline of clinical medicine may get direct exposure to the ideas and insights of evidence-based medicine.
2. Funding bodies may demand evidence that researchers are following an 'evidence-based' approach.
3. The logic of this general approach to marshalling research evidence is compelling.

Kuhn (1970) and others have argued that research traditions change only when the pressures for change are overwhelming. The inherent logic of the approaches of evidence-based medicine and of the Cochrane Collaboration studies will not, on its own, be enough to bring about widespread adoption of these ideas and insights. Experts whose authority relies on the use of more traditional informal means for assessing the weight of evidence may feel their authority threatened.

The File Drawer Problem (Publication Bias)

Studies with human and animal subjects now require, in most countries, approval from an Ethics Committee. It is then possible to follow up all studies that have received approval, to see how many are published. One such investigation (Easterbrook et al. 1991) found that of 285 studies submitted, 52% had been published. Clinical trials were more likely to be published than were observational and laboratory based studies. Studies with statistically significant results were more likely to be published, as were studies with large sample size. Publication bias increases the likelihood of detecting treatment effects when there are none.

The Bias to Noise Ratio

The combining of data from a number of trials will reduce the effect of noise on the mean. If however there are consistent biases, it will make no difference to the biases. Thus inter-population studies of the link between salt consumption and blood pressure are susceptible to biases that arise because differences in salt consumption are likely to be linked with other dietary differences. Combining data from different studies may reduce the noise, but leaves any consistent bias unchanged. Not only the signal to noise ratio, but also the bias to noise ratio increases.
In an analysis of data from salt reduction trials Law et al. (1991) include data both from randomised controlled trials and from cross-over trials. In a cross-over trial, patients receive, first one diet, and then the other. There may be more than one cross-over. There are no details on how these cross-over studies were conducted, though more detailed information is available by reading the original papers. What is interesting is that the cross-over designs (where subjects receive first one diet and then the other) show a much larger effect for the difference between the low and high salt diets than does the randomised design. Swales (1991) argues that there is a bias

in this use of data from cross-over designs, so that putting all the different results together into a meta-analysis only highlights the bias. In the light of results reported in Sacks et al.(2001), it now seems that the it was the cross-over designs that had the precision needed to detect effects from salt intake. Randomised controlled trials may, unless there is control for other dietary factors, be too inaccurate to detect the effects that were under investigation.

The Neglect of Data Overview

There are many reasons for the past relative neglect of data overview issues. One is that these studies have traps for the unwary, of the type discussed above. No amount of pooling of information that is biased in one direction can remove the bias. There are severe problems in deciding how to weight the separate sources of evidence. How does one deal with issues of trial quality? Should trials of a type that are thought to yield poor quality evidence be ignored?

Note that these technical difficulties are difficulties for any use of the data. It is a helpful side effect of systematic overview that it brings them to light. Historical reasons, rather than such technical difficulties, are probably the main reason for the neglect of data overview. Some form of data overview, formal or informal, is inevitable when research results are brought together and their implications for practical decision-making assessed.

An adequate statistical theory, for use in data-based overview, was slow to develop. For a long time there was more than adequate challenge to theoretical skills from developing a theory that would handle data from an individual field site or from an individual clinical trial. Scientists have often been protective of their experiments and their data, which they may believe should stand on their own independently of the work of other scientists. The tradition of analysing separately data from each field experiment or each trial became firmly established. It remains firmly entrenched in horticulture, and in other research areas also. Experimenters who have worked on different sites may each claim the other is ‘wrong’, where it is unclear whether the difference is a geographical effect, or perhaps due to differences in experimental procedure.

Data-Based Overview – Examples and Further Comment

1. There are numerous instances where the relative weighting of different sources of evidence and the pooling of evidence are key issues. The Taubes study, quoted earlier, was an example. There are a number of medical examples that are modern re-runs of the discovery that blood-letting, so far from making you better, is (for the great majority of conditions) actually dangerous.
2. Many of the agricultural fertilizer trials that were conducted in New Zealand over several decades prior to the 1980s were for a long time not analysed. Not until the 1980s did a series of papers appear in the *New Zealand Journal of Agricultural Research* that provided the first careful overall quantitative evaluation of evident major effects. They highlighted areas that had been over-researched, and identified remaining gaps. There was an inevitable and implicit criticism of individual trials. Nowadays, a reasonable expectation is that such data will feed into a fertilizer database, with data analyses regularly updated to take account of data from new trials.
3. McGuinness (1997) provides evidence of several different competing schools of thought, each convinced it is right, on the teaching of reading. This may be an area where theory has grown like a weed, too little constrained by data from

experiments that follow strict protocols such as are now demanded for medical clinical trials. The book is a careful overview of the current evidence, though perhaps overstating the case for her own approach. She rightly criticises the quality of much reading research, to the extent that there has been no direct comparison with competing approaches or that claims have been based on loaded comparisons that have not used appropriate controls. McGuinness's account has some of the elements of the thorough data-based overview that is required.

McGuinness uses research evidence to identify a range of sub-tasks that must all be mastered if children are to learn to read. There is an inexorable logic to the approach that she defends, which includes tests for identifying failure in any sub-task. A key insight is that children should be able to identify the 43 or 44 sounds of spoken English before learning letters or letter combinations that represent these sounds. The attempt to work in the other direction, from letter combinations to sounds, introduces too many complications. There are too many letter combinations to master quickly. The theory that she develops seems compelling, because it seems relatively complete and is backed up at key points by research evidence. She presents limited research evidence that shows that her methods work.

While her arguments are persuasive, they do not quite clinch her case. Much of the research that will show the efficacy of her methods has still to be undertaken. The jury seems to me still out in respect of her more extreme claims.

4. Meredith Wilson, a Ph.D. student in Archaeology and Anthropology at the Australian National University, is using published information to undertake an overview of rock art in the Pacific. An inventory of rock art sites found in this region was initially compiled by Specht (1979) and later added to by Ballard (1992). Wilson is drawing on this inventory to specifically conduct a comparison of rock art motifs. How should one group the sites and districts that are represented? What insight does the art, and the groups into which art items fall, shed into historical cultural connections in the region?

Her study has the potential to make, from relatively disconnected items of published information at a site level, a coherent account. As well as forming the building blocks of that account, those individual published site reports will surely have a more regional relevance and meaning within the framework of her account. Moreover, better understanding of the chains of connection between the art at the various sites must lead to better understanding of the individual motifs.

Just as with other types of overview, there have been reporting inadequacies that create difficulties for the study. Future studies of individual sites will do well to note those criticisms.

9.4 The Historical Sciences

There are broad principles that apply across different areas of research. There are also issues that are specific to particular areas of research. The historical sciences – including history, archaeology, evolutionary biology, and geology – raise issues that rarely arise in physics and chemistry. Obviously there is a use of the technical methods from chemistry and physics. Archaeologists may need to do chemical analyses of soil samples, and to measure the amount of radiocarbon in fragments of wood. The questions that are of interest go well beyond chemistry and physics. How does one use the results of these tests to make inferences about events that took place ten or twenty thousand years ago?

Data will enter in different ways into different forms of research synthesis. Historians and archaeologists can learn from the methods of physical scientists, biologists and experimental educationalists, but they will not be able to take them over as they stand. While the study of patterns of human history can learn from the methods of the physical sciences, the research approaches must be different.

In his book *Guns, Germs and Steel* Diamond (1997) seeks to explain striking differences between the long-term histories of peoples on different continents and islands in the past 13,000 years. The book is in a sense a sequence of data-based overview studies that are welded into a splendid continuous narrative. The data that he quotes are broad brush – numbers of plant and animal species domesticated in different geographical locations, differences in land area and population size, differences between continents in the diffusion rates of crops and artefacts that seem a result of their different geography, one-sidedness in the transfer of diseases between Europe and the Americas, and a variety of archaeological and phylogenetic data. He limits attention to data that seem to have a clear and relatively unequivocal story to tell. He is not afraid to criticise his sources.

As is inevitable in a book that is intended for a wide audience, the casual reader must largely take Diamond's facts and figures on faith, accepting that they are adequately accurate for his purpose. Specialist readers will wish to refer to his sources, described in a chapter by chapter list of references. There is a brief commentary that makes clear the relevance of each of the books and articles that he cites.

Diamond concludes that environmental and resource differences explain the striking differences between the long-term histories of different peoples, and not innate differences in the peoples themselves. Why? There have been, historically, numerous experiments in which individuals from one environment have migrated to another environment – European farmers to Greenland or the Great Plains, farmers stemming ultimately from China to the Chatham Islands, the rain forests of Borneo, or the volcanic soils of Java or Hawaii. Depending on the environments and what they had brought with them, the ancestral peoples either ended up extinct or returned to living as hunter-gatherers, or went on to build complex states. It was no trivial matter, and in some cases impossible, to transplant existing farming practices to the new environment. It is similarly interesting to note how Aboriginal Australian hunter-gatherers became hunter-gatherers with an unusually simple technology once in Tasmania, in South Australia became canal builders running a productive fishery, and ended up extinct when transplanted to truly appalling conditions on Flinders Island.

Diamond identifies four groups of factors that help explain inter-continental differences affecting the different technological patterns of development of different human societies. They are:

1. Wild plant and animal species available for domestication
2. Factors affecting rates of diffusion and migration within continents.
   [Most rapid in Eurasia.]
3. Factors affecting movement between continents.
   [Easiest from Eurasia to sub-Saharan Africa.]
4. Differences in area or total population size.
   [A large area or population means more innovations and potential inventors, more competing societies, and more pressure to adopt and retain innovations.]

Diamond presents evidence on the way that each of these types of factors has affected the different histories of the different peoples living on the different continents.

Thus, in respect of the first point above, he presents a table that compares the distribution of large-seeded grass species, that once domesticated might have provided food crops:

		Number of Species
West Africa, Europe, North Africa		33
Mediterranean zone	32	
England	1	
East Asia		6
Sub-Saharan Africa		4
Americas		11
North America	4	
Mesoamerica	5	
South America	2	
Northern Australia		2

Diamond defines a mammalian candidate for domestication as a species of terrestrial, herbivorous or omnivorous, wild mammal weighing on the average over 100 pounds. This is the basis for another list:

	Continent			
	Eurasia	Sub-Saharan Africa	The Americas	Australia
Candidates	72	51	24	1
Domesticated species	13	0	1	0
% of candidates domesticated	18%	0%	4%	0%

Diamond discusses the characteristics of species that were suitable for domestication, and argues that there were good reasons why none of the African species were domesticated.

In regard to point 2, he argues that diffusion will happen most readily along lines of similar latitude, i.e. between regions with similar climate and able to grow similar types of plant species. So one expects that diffusion of domesticated plants and animals, and of human populations, will be more rapid in Eurasia than along the predominant longitudinal axes of the Americas and sub-Saharan Africa. He discusses such archaeological data as are available on rates of diffusion. He handles points 3 and 4 in much the same way, making general points and backing these points up with whatever archaeological and evidence is available.

The Future of Human History as a Science

Particularly relevant to my discussion is Diamond's last chapter, on "The Future of Human History as a Science". Diamond proposes a research programme that would gather quantitative information intended to test his major claims, and that would provide more accurate quantitative estimates of e.g. the different diffusion rates of crops, artefacts, etc., in the different continents. Diamond's research synthesis sets the scene for an ongoing research programme. This leads into a wide-ranging discussion of 'historical science'. There is an overlap of interest with the historical content of astronomy, climatology, earth science and evolutionary biology. A view that sees history as a series of 'natural experiments' can be illuminating and insightful. We have referred to the experiment, frequently repeated in the past 13,000 years, involved in taking people from one continent and culture and placing them, with quite different environmental resources, on another continent. It is important to look at those movements where migrants have not been able to take with them any substantial new plant or animal or other material resources from the country from which they have come.

Imaginative reconstruction and synthesis readily gets out of hand. Hence the importance of using all available data-based reality checks, and all sources of evidence, that we can summon. Hence also the importance of using sources critically, recognising their limitations. One should not base elaborate reconstructions on individual pieces of shaky evidence. Diamond critically evaluates evidence from archaeological artefacts, plants, animals, linguistics, and genetics. He brings together multiple sources of evidence for his major claims, to create a coherent account. His research synthesis has wide-ranging implications that further research can check. He looks for a confluence of evidence.

Why do I consider that Diamond is broadly right, but reject the elaborate imaginative historical reconstructions of Immanuel Velikovsky, which Sagan (1979) dissects? Velikovsky makes individual items of shaky evidence the basis for elaborate reconstructions. One does not find a confluence of different sources of evidence.

References and Further Reading

Andersen, Bjorn 1990. Methodological Errors in Medical Research: an incomplete catalogue. Blackwell Scientific Publications, Oxford.

Appel, L. J. et al. 1997. A Clinical Trial of the Effects of Dietary Patterns on Blood Pressure. The New England Journal of Medicine 336: 1117-1124.

Ballard, C. 1992. Painted rock art sites in Western Melanesia: locational evidence for an 'Austronesian' tradition. In J. McDonald and I.P. Haskovec (eds.), State of the Art. Regional rock art studies in Australia and Melanesia. Occasional AURA Publication No. 6. Australian Rock Art Research Association. Melbourne: pp 94-106.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639.

Bennett, J.H. (ed.) 1989. Statistical Inference and Analysis. Selected correspondence of R. A. Fisher, letter to J. R. Baker, pp. 343-346. Oxford.

Bussell, W. T., Maindonald, J. H. and Morton, J. R. 1997. What is a correct plant density for transplanted green asparagus? New Zealand Journal of Crop & Horticultural Science 25: 359-368.

Chalmers, I. and Altman, D. G., eds. 1995. Systematic Reviews. BMJ Publishing Group, London.

Chatfield, C. 1995. Uncertainty, data mining and inference (with discussion). Journal of the Royal Statistical Society A, 158: 419-466.

Cochrane Injuries Group Albumin Reviewers 1998. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. British Medical Journal 317: 235-240.

Diamond, J. M. 1997. Guns, germs, and steel : the fates of human societies. Random House, London.

Draper, D; Gaver, D P; Goel, P K; Greenhouse, J B; Hedges L V; Morris, C N; Tucker, J R; Waternaux, C M 1992. Combining Information. Statistical Issues and Opportunities for Research. National Academy Press, Washington D.C.

Easterbrook, P. J., Berlin, J. A., Gopalan, R. and Matthews, D. R. 1991. Publication bias in clinical research. Lancet 337: 867-872.

Ehrenberg, A. S. C. 1990. A hope for the future of statistics: MSOD. The American Statistician 44: 195-196.

Greenhalgh, T. 1997. How to read a paper. The basics of evidence based medicine. BMJ Publishing Group, London.

Hubbard, R. and Armstrong, J.S. 1994. Replications and extensions in marketing: rarely published but quite contrary. International Journal of Research in Marketing 11: 233-248.

Law, M. R., Frost, C. D., and Wald, N. J. 1991. By how much does dietary sodium lower blood pressure? III – Analysis of data from trials of salt reduction. British Medical Journal 302: 819-824.

Lindsey, R. M. & Ehrenberg, A. S. C. 1993. The design of replicated studies. The American Statistician 47: 217-228.

McGuinness, D. 1997. Why our Children Can't Read. The Free Press, New York.

Moynihan, R. 1998. Too Much Medicine. Australian Broadcasting Corporation.

Oxman, A. D. and Güyatt, G. H. 1983. The science of reviewing research. Annals of the New York Academy of Sciences 703: 125-131.

Sacks, F.M., Svetkey, L.P., Vollmer, W.M., Appel, L.J., Bray, G.A., Harsha, D., Obarzenek, E., Conlin, P.R., Miller, E.R., Simons-Morton, D.G., Karanja, N., and Lin, P.-H. 2001. Effects of blood pressure on reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. New England Journal of Medicine 344: 3-10.

Sackett, D. L. and Oxman, A. D., eds. 1994. The Cochrane Collaboration Handbook. Cochrane Collaboration, Oxford.

Sackett, D. L., Richardson, W. S., Rosenberg, W. M. C. and Haynes, R. B. 1997. Evidence-Based Medicine. Churchill Livingstone, New York.

Sagan, C. 1979. Broca's Brain. Random House, New York.

ScHARR (School of Health and Related Research, University of Sheffield). 1998. Netting the Evidence. A ScHARR Introduction to Evidence Based Practice on the Internet. Available at http://www.shef.ac.uk/~scharr/ir/netting.html

Smith, A. F. M. 1996. Mad cows and ecstasy: chance and choice in an evidence-based society. Journal of the Royal Statistical Society A 159: 367-383.

Specht, J. 1979. Rock art in the western Pacific. In S.M. Mead (ed.), Exploring the Visual Art of Oceania. Australia, Melanesia, Micronesia, and Polynesia. University of Hawai'i Press. Honolulu: pp 58-82.

Swales, J. D. 1991. Dietary salt and blood pressure: the role of meta-analysis. Journal of Hypertension 9, supplement 6: S42-S46. See also the discussion: S47-S49.

Sacks, F.M., Svetkey, L.P., Vollmer, W.M., Appel, L.J., Bray, G.A., Harsha, D., Obarzenek, E., Conlin, P.R., Miller, E.R., Simons-Morton, D.G., Karanja, N., and Lin, P.-H. 2001. Effects of blood pressure on reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. New England Journal of Medicine 344: 3-10.

Taubes, G. 1998. The (political) science of salt. Science 281: 898-907 (14 August).

10 Styles of Data Analysis

> *A cautious investigator will use existing data, or data collected from a pilot study, to determine the form of analysis that is appropriate for the main body of data. In all studies, it is necessary to check for obvious data errors or inconsistencies. In addition there should be checks that the data support the intended form of analysis.*
> *Exploratory data analysis is both a methodology for examination of data, and a collection of techniques for data exploration. Exploratory data analysis may be especially important for data from a project with which you have had little previous contact.*
> *The data may contain information on more than just the research question that was under investigation. The information may for example suggest fruitful new lines of research. Even if this information does not directly relate to questions that were in view at the beginning of the study, it is undesirable to lose it.*
> *The validity of formal data analysis depends crucially on the choice of a model that accurately describes the data. The model should reflect relevant theoretical understanding, the design of data collection, and what has been learned from exploratory data analysis. How does one decide which model assumptions matter, and need careful checking? Where normality is assumed, how close to normality is adequate?*

One is presented with a new set of data. There may be very limited clues from examining what other researchers have done with similar data. Or there may be obvious inadequacies in published analyses. What is the best way to begin? What forms of data exploration will draw attention to obvious errors or quirks in the data, or to obvious clues that the data contain. What are the checks that will make it plausible that the data really will support the intended analyses. What mix of exploratory analysis and formal analysis is appropriate? Should the analysis be decided in advance, as part of the planning process? To what extent is it legitimate to allow the data to influence what analysis will be performed?
This chapter will cover general issues relating to the approach to data analysis. They are issues that you should keep in mind when you plan your study, well before you do the analysis or bring the data for analysis. Questions that the data analyst should ask include:

1. What is the research question?
2. What has been measured, and how does it bear on the research question?
3. What was the design of data collection? Was a randomised design used?
4. What is the structure of the data? What are the explanatory variables? Which, if any, variables were under experimental control? What are the outcome variables?
5. What are the sources of variability?
6. What is the population to which it is hoped to generalise results? Have the data been collected in a manner that allows this generalisation? What are the implications for data analysis?
7. Is there prior information, data-based or theoretical, about likely effects or about the form of the likely response?

It will, finally, be necessary to give a clear and lucid report of results. Keep this in mind from the beginning. Start practicing the report, verbally and in writing, as soon as possible.

Windfalls

Often data contain information different from what they were designed to provide. A careful data analyst will watch out for such information. It may be more interesting than the information that the data were intended to collect, perhaps suggesting new lines of research. It does not follow that the design of data collection is unimportant to the usefulness of the data. On the contrary, such windfalls are likely only if data have been collected according to a strict design. It matters less than one might think that the design would have been optimal for getting this different and interesting information.

The Different Demands of Different Studies

In an experiment or study where the outcome is to an extent predictable, it should be possible to plan the analysis in advance. This reduces room for preferring the analysis that gives a result that conforms to expectations or preferences! Even so, preliminary graphical checks of the data should precede formal analysis.
Where there is no previous experience of working with data that are entirely comparable with the data from the current investigation, it is essential to take a careful preliminary look at the data. One should examine

1. whether there are outliers;
2. whether, aside from outliers, data for each treatment group are roughly normal;
3. whether a transformation of the data might be helpful, e.g. in order to make them more symmetrically and normally distributed;
4. whether there are clusters in the data;
5. whether any of the pairwise scatterplots show evidence of clustering or of unexpected patterns;
6. whether variances are homogeneous, i.e. similar for different groups;
7. whether there are unanticipated time trends associated, e.g. with order of collection.

Standard forms of analysis assume that observations are independent. Consider whether this is a plausible assumption for your data. Diagnostic plots are in general not much help in checking for independence.

10.1 Exploratory Data Analysis

Exploratory Data Analysis (**EDA**) is a name for data display techniques that are intended to let the data speak for themselves, prior to or as part of a formal analysis. EDA looks for what may be apparent from a direct, careful and (as far as possible) assumption-free examination of the data. An effective EDA display presents data in a way that will make effective use of the human brain's abilities as a pattern recognition device.
In all studies, checks against obvious errors or quirks in the data are essential. Also, researchers will not want to miss obvious clues that the data contain. So all data analyses should have some elements of exploratory data analysis. Extensive use of exploratory data analysis may be essential where the research breaks radically new

ground, or where the data are different in character from that of earlier workers. Alternatively, there may be obvious inadequacies in published analyses that make the analyses carried out by earlier researchers poor models to follow. The burden of rectifying the deficiency then falls to later researchers. One result of a need to put a relatively heavy reliance on EDA is that research results are less definitive than one might have hoped.

Exploratory Data Analysis – More Detailed Comments

There are helpful discussions of exploratory data analysis in

JMP Start Statistics: chapters 5 (85-114) and 13 (299-318)
SPSS for Windows Base System Users' Guide: chapters 9 (pp.181-199), 16 (291-301) and 27 (553-567).

EDA has at least three roles:

1. EDA examines data, leaving open the possibility that this examination will suggest how data should be analysed or interpreted. This use of EDA fits well with the view of science as inductive reasoning.
2. An exploratory data analysis may however go further, challenging the theoretical understanding that guided the initial collection of the data. EDA then acquires a more revolutionary role. It has become the catalyst, in the language of Kuhn (1970), for a paradigm shift.
3. EDA allows the data to influence and criticise an intended analysis. EDA gives checks on assumptions, needed so that subsequent formal analysis can proceed with confidence. EDA formalizes and extends the use that competent statisticians have always made of graphs to check their data.

Diagnostic statistics and graphs carry careful data scrutiny over into an examination both of the model used and of output from the analysis. They allow an 'after the event' form of EDA.

There is a risk that data analysts will see things that are just a result of looking hard. That is why the tools of conventional statistical analysis remain important. (Under torture, the data have confessed.) Inferences from an analysis that has been chosen to suit the data should be scrutinised with more than ordinary care. Where possible, significance levels and/or standard errors should be adjusted to take account of the extent to which the analysis has been chosen from some wider class of analyses. If doubt remains the only recourse is to try the analysis on new data, preferably from a new experiment.

An effective data analyst will use both EDA and conventional data analysis. Careful practical statisticians have always made extensive use of graphical displays. The main change brought by EDA is that ideas on how data should be explored prior to formal analysis have become more systematised.

Even if data have not been collected in a way that makes them suitable for formal statistical analysis, exploratory data analysis may give useful clues. The analyses will however be no more than suggestive. Results must be checked in more carefully designed studies.

The Merging of EDA into Mainstream Statistical Analysis

Areas where there have been big advances in recent years, cutting across both EDA and more conventional styles of analysis, include the analysis of counts, and data smoothing. Methods that check for and accommodate non-linearity are becoming pervasive. Another class of methods use the data for extensive sampling experiments that provide the information needed for testing hypotheses or estimating confidence

intervals. Many of the new methods are "computer intensive"; they would be unthinkable without modern high speed computers.
A smooth curve that estimates the pattern of the relationship may be a useful aid to interpreting a scatterplot. The methodology does not insist on a particular mathematical form of curve. Where a mathematical curve is available that seems satisfactory, it may be useful to compare it with the curve given by the data smoothing routine.
Among several relatively new types of graphical display that have been associated with EDA, perhaps the most widely used is the boxplot. Note also the stem and leaf display, which we consider first.

10.2 EDA Displays

The Stem-and-Leaf Display

The stem and leaf display is a finer grained alternative to a histogram, for use in displaying a single column of numbers. It provides a good place to start, in order to discuss modern approaches to displaying data. It provides a ready way to get medians and quartiles. I demonstrate a simple form of stem and leaf diagram, for the data that appear in the boxplot below. After sorting from smallest to largest, the numbers are:

-8.4, -6.0, 0.7, 1.0, 1.8, 2.0, 2.9, 3.0, 3.5, 3.6, 5.1, 6.1, 7.0, 7.5, 9.3

Here is the stem and leaf diagram:

```
N = 15   Median = 3
Quartiles = 1, 6.1

Decimal point is 1 place to the right of the colon

 -0 : 86
 -0 :
  0 : 11223334
  0 : 56779
```

[The numbers have been rounded to one decimal place. Then the first 8 is a -8, the 6 is a -6, and so on.]

Boxplots

How does one summarise, in a useful and readily assimilable way, the information that is presented in a histogram? The boxplot strikes a good balance between the coarse summary and fine detail. It picks on particular features of the distribution of data, and shows those. Fig. 7 gives information needed to interpret a standard form of boxplot.

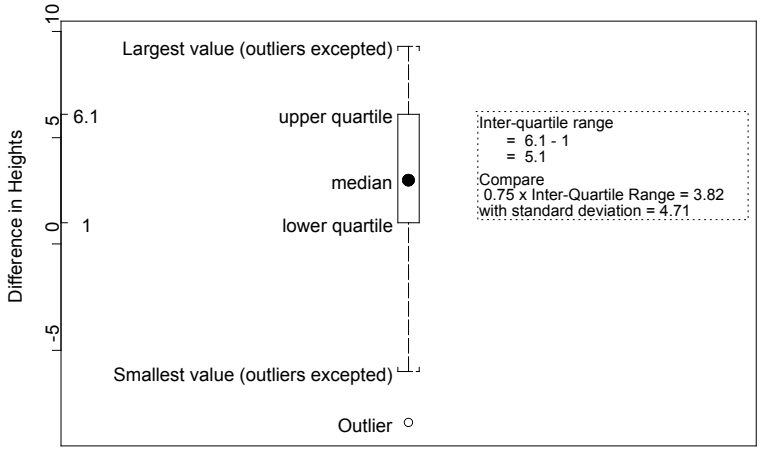


Fig. 7: The Interpretation of Boxplots. The boxplot shows the distribution of differences of heights of cross-fertilised and self-fertilised plants. (n =15). Data are from Charles Darwin.

The Scatterplot – The Challenger Disaster

The scatterplot, of which Fig. 8 is an example, is an indispensable exploratory tool. We will see frequent examples of its use. An extension of the scatterplot is the scatterplot matrix, which shows plots of every variable against every other variable in a scatterplot layout.

On January 28 1986 two large rubber O-rings on the space shuttle Challenger leaked, leading to an explosion and the death of the seven astronauts. The rings (on the booster rockets) had lost their resiliency because of the low temperature – around 0ºC (32ºF). The engineers, from Morton Thiokol Inc, had recommended delaying the flight. Their argument was not persuasive, and the launch proceeded.

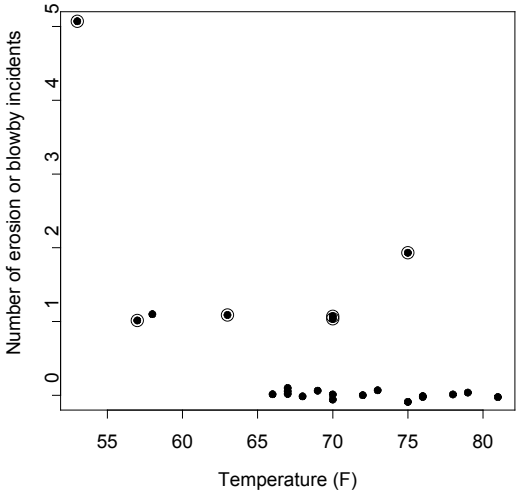


Fig. 8 : Total O-ring incidents versus temperature, in shuttle flights prior to the Challenger disaster of January 28 1986. Points have been jittered slightly, to separate noverlapping points. Circles points are for data that were presented in one of the 13 pre-launch charts.

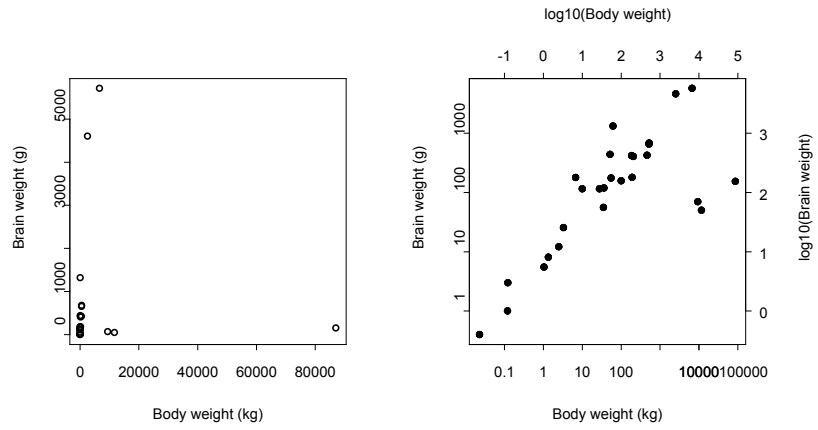
Fig. 8 shows the data from earlier shuttle flights. What is the correct way to interpret the data? Would you have advised proceeding with the flight?

Note: There were two main types of damage – *erosion* and *blowby* damage. Blowby damage occurred when hot gas leaked through and burned, causing blackening of the

O-ring. This was not supposed to happen. The gap occupied by the rings expanded as pressure built up in the rocket. The rubber needed to expand fast enough to close the gap. It seems obvious that temperature will have a large effect on the resilience of the rubber. There should have been experiments to find, for various temperatures, how fast the rubber expands. Feynman, as a member of the committee investigating the disaster, did a simple experiment during the hearing that checked out the response of the rubber at 0ºC. He put the rubber in a clamp, then left it in cold water for a while. On undoing the clamp it did not spring back. (Feynman 1988.) As a demonstration this was highly effective. However there was no standard with which to compare the outcome of Feynman's demonstration. How hard did Feynman clamp the rubber? What happens at intervening temperatures? This would not rate well as a scientific experiment, but might serve well enough as part of the preliminary investigation. Tufte (1997) has a fairly complete discussion of the Challenger disaster, emphasising the importance of clear and accurate presentation of information. See also Wainer (1997).

10.3 What is the Appropriate Scale?

Figs. 9a and 9b plot brain weight (gm) against body weight (kg), for a number of different animals.



(a) Untransformed scale (b) Logarithmic scale, both axes

Fig. 9: Brain weight versus body weight, for different animals.

Fig. 9(a) is almost useless. We need scales that spread the data out more evenly. We can do this by choosing a logarithmic scale. Multiplication by the same factor (e.g. for the tick marks in Fig. 9(b), by a factor of 10) always gives the same distance along the scale. If we marked points 1, 5, 25, 125, … along an axis, they would also lie an equal distance apart.

Quantities for which a logarithmic scale is appropriate change in a multiplicative manner. If cells in a growing organism divide and produce new cells at a constant rate, then one will get multiplicative growth. Random changes in the growth rate, following perhaps a normal distribution, will produce adult organisms such that the logarithm of the size (height or weight, or another measurement) is normally distributed.

The logarithmic transformation is so commonly needed that I have felt it necessary to introduce it at this point. Biologists, economists and others need to be comfortable working with it. As I have indicated, there are many circumstances in which it makes

good sense to move to working on a logarithmic scale, i.e. to use a logarithmic transformation. A transformation is often necessary when working with percentages.

10.4 Data Mining and Exploratory Data Analysis

Data are a valuable resource. As such, perhaps one can *mine* the resource for its nuggets of gold. In part the interest in *data mining* has been driven by individuals and organizations who find themselves with large data holdings, that they feel ought to be sources of valuable information. They may have little idea what to do with them. Hardware and software computer vendors, looking for new market niches, have fanned the interest. "Data mining" is a term that has been used to sell an idea – that large data bases may hold information additional to what was in mind when they were collected.

There is no firm distinction between data mining and statistics. Any adequate attempt at data mining will use statistical insights. Much commercial data mining activity uses relatively conventional statistical methods. A difference is that data miners may be working with quite huge data sets. Hence Friedman's (1998) definition of data mining as the "computer automated exploratory data analysis of (usually) large complex data sets." A data set with values of twenty or thirty variables for each of several hundred thousand records is, in the context of commercial data mining, small. Some data mining approaches are fairly specific to individual research areas, such as astrophysics at one extreme or business data processing at the other.

A simple example of 'exploratory' data mining is the use of medical practice variations as a starting point for questions about operating or prescribing practices. McPherson (1990) quotes standardised rates for hysterectomy that were six times as high in the United States as in Norway. Such a huge difference calls for investigation and comment. In a classical statistical sense, the data miner is looking for outliers. Detection of fraud, in large clinical trials, or in business records, provides another example. What sorts of unusual patterns might make closer scrutiny desirable?

The exploratory form of data mining applies a search process to a data set, often a very large data set, and looks for interesting associations. While the data may initially have been collected to answer some primary question or questions, the expectation is that there will be other interesting and potentially useful information in the data. Most experienced statisticians have at some time encountered unexpected and interesting results when, as a prelude to the main analysis, they have set out to do a careful exploratory analysis. Is it possible to set up automatic processes that may bring such results to attention? Jorgensen and Gentleman (1998) cite examples of data sets where there is bound to be unmined interesting information – fisheries data collected by Australian and New Zealand agencies over a number of years, secondary information in databases on clinical trials, and databases of routinely collected business information.

Much of the focus of data mining research has been on ways to find views of data that highlight 'interesting' or unusual features – a search for what statisticians would call 'outliers'. Friedman (1997) lists a number of approaches. *Exploratory Data Analysis* is in the spirit of Friedman's description of data mining. Research on data visualization is in this same tradition.

In spite of the huge size of the data sets, there are the usual problems of knowing what information will generalise to future data and what will not. This is true even if the data sets that are mined are in some sense complete. We will say more about this in the chapter on data structure.

10.5 Formal Analysis

Planning the analysis is one aspect of planning the research study. The formal analysis should be planned in advance, but with the possibility of limited changes as a result of the exploratory analysis. The best situation is where you are able to find data similar to what your study will provide, and can practice the intended analysis on that. The information from this practice analysis may be invaluable for designing your own study.

This is one of several reasons why research data ought to be archived in such a way that later researchers can access it. Published results rarely provide the particular information that is needed for designing later studies.

What sorts of departure from the plan are acceptable, and which are not? If the exploratory analysis makes it clear that a transformation is needed to achieve normality, then you should use the transformation. There is too much risk that the untransformed data will yield misleading results.

Selecting the Variables

On the other hand, substantial variable selection, especially using stepwise or best subset regression, may introduce huge biases. You should specify in advance which covariates you will use. Any investigation of other covariates may be undertaken as an exploratory investigation, a preliminary to your next study. If for example you have 100 observations, and 40 explanatory variables, variable selection is a hopeless task. You may be able to use all explanatory variables to get an equation that you can use for prediction, but you will not know which explanatory variables make the prediction work.

One way to demonstrate the problem is to set up a regression where all 40 variables are random noise. Selection of the best three explanatory variables, then evaluating the result with no allowance for selection effects, will most often give an equation in which two out of the three variables appear significant at $p<0.05$. The use of variable selection procedures without allowance for selection effects is tailor made to generate spurious associations. The p-values that regression programs provide assume that there has been no selection of the variables.

10.6 Inference – Asking the Data Specific Questions

The questions we ask of data may be simple: "Does increasing the amount of an additive in milk make it seem sweeter? If so, by how much does its sweetness increase?"

Or they may be questions about relationships: "What is the relationship between the stretch of a rubber band that we hold over the end of a ruler and the distance the rubber band moves when it is released? How accurately can we predict the distance?"

Here again are the taste experiment data that first appeared in section 4.3:

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
4 units	72	74	70	72	46	60	50	42	38	61	37	39	25	44	42	46	56
1 unit	58	69	60	60	54	57	61	37	38	43	34	14	17	54	32	22	36
Diff.	14	5	10	12	-8	3	-11	5	0	18	3	25	8	-10	10	24	20

We may wish to ask how much the assessment of sweetness changed when we went from one unit of additive to four units. The mean difference is 7.4, with an SD of 10.9 . The SE of this difference is thus $10.9/\sqrt{17} = 2.74$. There are now several ways to report this:

1. The mean change is 7.4 [SE 2.74]. (One should report this anyway.)

2. The t-statistic is $t = 7.4/2.74 = 2.66$, on 16 degrees of freedom. In other words, the difference is 2.66 times the standard error.
3. A 95% confidence interval for the change is $(7.4 - 2.12 \times 2.66, 7.4 + 2.12 \times 2.66)$,
   i.e. (1.7, 13.1).
   [The multiplier, equal to 2.12, is the 5% two-sided critical value for a t-statistic on 16 (= 17 -1) degrees of freedom.]
4. We reject the null hypothesis that the true mean difference is 0 (p =0.02).
   [The two-sided p-value for t = 2.66 on 16 d. f. is 0.02]

Item 1 is simple. Often, and especially if the difference is more than four or five times the SE, it is all we need. Item 2 gives us a rough way to compare the change with its standard error. If *t* is more than about 2, we can begin to worry whether the 95% confidence interval in item 3 contains 0, or (equivalently) whether the p-value in item 4 is less than 0.05 .

Many researchers find significance tests are hard to understand. Misunderstandings are common in the literature, even among mature researchers. A p-value does not allow the researcher to say anything about the probability that either the null hypothesis or its alternative is true. We will pick this point up below. So why use them? Perhaps the best that can be said is that hypothesis tests often provide a convenient and quick answer to questions about whether effects seem to stand out above background noise. But if all that emerges from an investigation are a few p-values, one has to wonder what has been achieved.

Because of these problems, there are strong moves away from hypothesis testing (item 4) and towards confidence intervals (item 3). Formal hypothesis testing (significance testing), which at one time had become almost a ritual among researchers in psychology, is now generating a huge controversy, reflected in the contributions to Harlow et al. (ed. 1997). Krantz (1999), which is a review of the Harlow et al. book, is an interesting guide to the controversy. See also Gigerenzer (1998).

Examining Relationships

The examination of pattern and relationship is much closer to the central themes of science, and hence to what ought to be central themes of statistics, than is hypothesis testing. So consider an experiment where a rubber band is pulled over the end of a ruler that is held at an angle of 20° to the ground, and let go. How does the distance travelled by the rubber band change with the amount of stretch? Fig. 10 shows data from such an experiment.

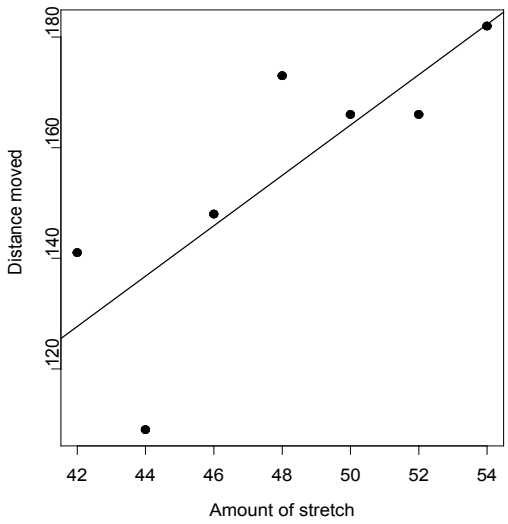


Fig. 10: Distance moved by rubber band, versus amount of stretch.

This is much more informative than doing repeated trials, some with a stretch of 44mm and some with a stretch of 54mm. With the data we have, it would be daft to do a significance test to compare, for example, stretches of less than 47mm with stretches of more than 47mm. Our data allow us to do a lot better than that. We can study the relationship and ask whether a line is an adequate description. We can ask how we could change the experiment so that we can get a more accurate line. [For greater precision, we can increase the range of values of "Amount of stretch".]

10.7 The Limits of Confidence Intervals and Hypothesis Testing

Formal statistical hypothesis testing should not be used when there is strong prior information. Criminal investigations and diagnostic tests offer two examples. In contexts where there is strong prior information, the reporting of a p-value that ignores this information can be highly misleading.

Suppose a diagnostic test for HIV infection has a specificity of 0.1%. This implies that, for every 1000 people tested, there will on average be one false positive. (This is a very high specificity. Even though a test may be able to do as well as this with highly skilled operators, such a high specificity may be beyond less skilled and careful operators.) Now suppose an adult male who does not belong to any of the recognised risk categories has the test, with a positive result.

Using the hypothesis testing framework, take the null hypothesis H_0 to be the hypothesis that the individual does not have HIV. Given this null hypothesis, the probability of a positive result is 0.001 . So the null hypothesis is rejected.

But wait! There is strong prior information. The incidence of HIV in adult Australian males (15-49 years) may be 1 in 10,000. Assume that 10,001 people are tested. One will, on average, have HIV and test positive. (We assume 100% sensitivity, i.e. everyone who has HIV will test positive.) Among the remaining 10,000 who do not have HIV, we expect that 10 will test positive. So if we incorporate the prior information, the odds that the person has HIV are 1:10, i.e. less than 10%. Pinker (1997, p.348) considers examples of this type, in the context of a discussion of human abilities with probabilistic reasoning:

Not Infected	Infected

10,000 × 0.001 = 10 (false) positives	1 true positive
Table 1: Expected Numbers of Positives, in a Population of 10,001 that includes one true positive.	

In serious criminal cases the police may examine 10,000 or more potential perpetrators. Suppose there is a form of incriminating evidence that occurs for one person in 1000. One will net 10 suspects. Suppose one of these is later charged. The probability of such incriminating evidence, assuming that the defendant is innocent, is indeed 0.001, which is the relevant probability for a test of hypothesis.
However assuming that the police screening of 10,000 potential perpetrators is guaranteed to net the perpetrator, the police screening will net around ten innocent people along with the one perpetrator. This evidence leads to odds of 1:10, i.e. less than 10%, that the defendant is guilty. On its own, it should be discounted. If the police screening is not guaranteed to net the perpetrator, the odds of guilt are lower still.

Not the Perpetrator	The Perpetrator
10,000 × 0.001 = 10 (false) positives	1 true positive
Table 2: An example from forensic screening that illustrates how the probability of matching incriminating evidence must be weighed against the extent of any searching that has led to the identification of a match.	

References and Further Reading

Andersen, T. F., and Mooney, G. 1990. The Challenges of Medical Practice Variations. Macmillan Press, London.

Cleveland, W. S. 1993. Visualizing Data. Hobart Press, Summit, New Jersey.

Cleveland, W. S., 2ed. 1994. The Elements of Graphing Data. Hobart Press, Summit, New Jersey.

Feynman, R. P. 1988. What do you care what other people think? W. W. Norton & Co., New York.

Friedman, J. H. 1997. Data Mining and Statistics. What's the Connection? Proc. of the 29th Symposium on the Interface: Computing Science and Statistics, May 1997, Houston, Texas.

Friedman, J. H. 1998. Statistics 315B: Statistical Aspects of Data Mining (Winter 1998). Available from http://www.stanford.edu/~jhf/Stat315B.html

Gigerenzer, G. 1998. We need statistical thinking, not statistical rituals. Behavioural and Brain Sciences 21: 199-200.

Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds.) 1997. What If There Were No Significance Tests? Lawrence Erlbaum Associates, Mahweh, N. J.

Intersalt Cooperative Research Group. 1988. Intersalt: an international study of electrolyte excretion and blood pressure: results for 24 hour urinary sodium and potassium excretion. British Medical Journal 297: 319-328.

Jorgensen, M. and Gentleman, R. 1998. Data mining. Chance 11: 34-39 & 42.

Krantz, D. H. 1999. The null hypothesis testing controversy in psychology. Journal of the American Statistical Association 44: 1372-1381.

Kuhn, T., 2nd edn, 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Maindonald J H 1992. Statistical design, analysis and presentation issues. New Zealand Journal of Agricultural Research 35: 121-141.

Pinker, S. 1997. How the Mind Works. Norton, New York.

Tufte, E. R. 1997. Visual Explanations. Graphics Press, Cheshire, Connecticut, U.S.A.

Wainer, H. 1997. Visual Revelations : Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot. Copernicus Books.

11. Statistical Models

> Models are to be used, not believed.
> [Henri Theil, Principles of Econometrics, 1971. Wiley.]

> *Models underlie all analyses. Unless there are model assumptions, an analysis is impossible. As one makes stronger model assumptions, the chances of getting clear results improve. There is a price for stronger assumptions – if the assumptions are wrong then results may be wrong.*

Some assumptions are fairly harmless. We say that the method used is *robust* against those assumptions. Other assumptions matter a lot. How do we know which is which? Much of the art of applied statistics comes from knowing which assumptions are important, and need careful checking. There are no hard and fast rules.

Ideas of what model is appropriate stay somewhat in the background in initial efforts at exploratory data analysis. The choice of model is of crucial importance for the main analysis. I will comment below on considerations that, together with what has been learned in the exploratory data analysis, should influence the choice of model.

Research planning should include a provisional assessment of the model that the data are expected to follow. Any sample size calculation will work from this provisional assessment. But unless the initial assessment was based on acquaintance with closely comparable data, it may be necessary to revise the assessment following exploratory data analysis on the new data. It is often necessary to move between exploratory data analysis and formal analysis.

11.1 Rough and Smooth

Many models have the form[30]

Observed value = Model Prediction + Statistical Error

In electrical engineering terminology, the model prediction is the "signal", while the statistical error is "noise". The model prediction that we get from statistical analysis is an estimate, which we might call the "smooth". The differences between observed values and model predictions are the residuals, which we might call the "rough"

In Fig. 11, the points predicted by the straight line (the `smooth') are shown as circles, while the residuals (the `rough') are shown as dotted vertical lines. The residuals for points that lie above the line are positive, while residuals for points that lie below the line are negative.

[30] Mathematically, one may write

$$Y = \mu + \varepsilon$$

(Often μ is a function of explanatory variables.)

The model prediction (μ) is the `smooth'. The statistical error ($\varepsilon$) is the `rough'. Using the mathematical idea of *expected value,* it is usual to define $\mu = E(Y)$, where the $E(Y)$ denotes `expected value of Y'. The expected value generalises the idea of mean.

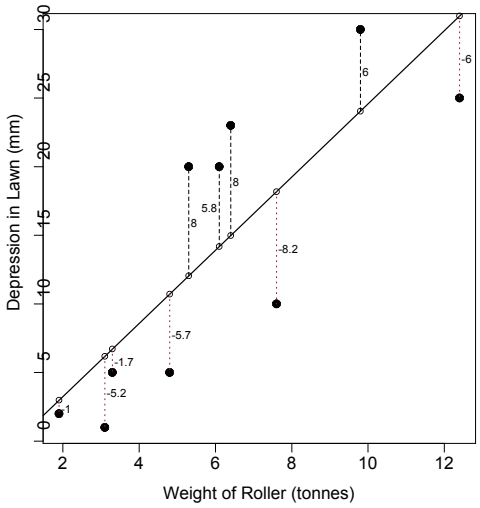


Fig. 11: Lawn Depression, for Various Weights of Roller, with fitted straight line. Positive residuals are shown with dashes, while negative residuals are shown dotted.

An alternative is to fit a smooth curve. This may help indicate whether a straight line really is appropriate. In Fig. 12 the smooth has been obtained using a general "smoothing" algorithm.

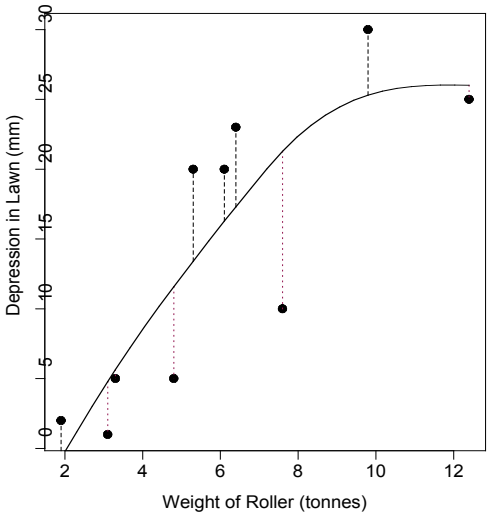


Fig. 12: Lawn Depression, for Various Weights of Roller. Also shown is a Fitted Curve that used the S-PLUS Loess Smoothing Routine.

Note that there is just one point that seems to be causing the line, and the fitted curve, to level out. So even if this effect had approached statistical significance, it would not be convincing.

11.2 Why Models Matter

Routine "black box" analysis of data, with no regard to data structure, can lead to silly results. In the two examples that I give below, the structure comes from factors other than those examined in the initial faulty analysis.

An Example from Bowling Averages in Cricket

In the 1st innings bowler A takes 4 wickets, for 40 runs. Bowler B takes 5 wickets for 70 runs.

The 2nd innings is much less happy for the bowlers. Bowler A takes 6 wickets for 240 runs. Bowler B takes 1 wicket for 50 runs.
Here is how the averages total up:

Bowler A			Bowler B		
First Innings	Second Innings	Overall	First Innings	Second Innings	Overall
40 runs	240 runs	280 runs	70 runs	50 runs	120 runs
4 wickets	6 wickets	10 wickets	5 wickets	1 wicket	6 wickets

One can look at the table this way

Bowler A			Bowler B		
First Innings	Second Innings	Overall	First Innings	Second Innings	Overall
10 r/w	40 r/w	28 r/w	14 r/w	50 r/w	20 r/w
4 wickets	6 wickets	10 wickets	5 wickets	1 wicket	6 wickets

Notice that bowler A did better than bowler B in both innings, but ended with a poorer average overall. This is because bowler A did more of the bowling when the going was tough.
From an experimental design point of view, this is a block design, with innings as a 'block'. The issue is how we should combine results from the two innings. The usual method, which adds up wickets and runs, weights runs/wicket according to the number of wickets! It thus favours bowlers who, when there is a feast of runs, are not used!
From a modelling point of view, adding up wickets and runs ignores the effect of the pitch. The model is incomplete, and hence the answer it gives is misleading. It is essential to piece the separate pieces of evidence together in ways that truly reflect the data structure. An effective model will encapsulate all those features of the data structure that are important for the inferences that are to be drawn.
A reasonable model is

Runs/wicket = Mean for innings + Effect Due to Bowler
= Overall Mean + Effect Due to Innings + Effect Due to Bowler

The 2 by 2 table of runs per wicket information is:

	Innings 1	Innings 2	average	(Total runs / total wickets)
A	10	40	25	(28)
B	14	50	32	(20)
average	12	45	28.5	

We might summarise results thus:

Overall mean	28.5	= 0.5 × (12 + 45)
Effect due to Innings 1	-16.5	= 12 - 28.5
Innings 2	+16.5	= 45 - 28.5
Effect due to Bowler A	-3.5	= 25 - 28.5
Bowler B	3.5	= 32 - 28.5

Why Models Matter – Adding up Contingency Tables

The following contrived example shows admission patterns in two separate faculties

Engineering			Sociology			Total		
	Male	Female		Male	Female		Male	Female
Admit	30	10	Admit	5	10	Admit	35	20
Deny	30	10	Deny	15	30	Deny	45	40

Because both the admission rates and the gender balance of applicants differ between the two faculties, simple addition of the numbers gives a misleading result.
This is an example of Simpson's paradox. The method that is needed is the Maentel-Haenzel method, which combines the odds ratios for the admission rates. (The odds of admission in each faculty are the same for the two sexes. Hence the odds ratios, with gender as the classifying factor, are 1:1 in both faculties.)
The data do highlight different male:female application rates, and different overall admission rates, in the two faculties.

Choosing the Model

Factors that should influence the choice of model include:
1. Scientific understanding, e.g. a well-tested theory that predicts what to expect.
2. Previous experience with similar or broadly similar data.
3. Indications from the exploratory data analysis.
4. Diagnostic information from your first tentative model fits.

There may be a well-attested theory that fits or approximately fits the experimental facts. In that case the theoretical relationship may be a useful starting point, even if some modification is necessary so that it accurately describes real data.
Other researchers may be able to give useful pointers. Question other researchers closely about reasons for their choice of model. Do not automatically assume that they are using correct models. For example there has been widespread incorrect use of probit models in disinfestation research, especially for time-mortality data.
It is particularly important to check that the model is capable of reproducing the major features of the data. It should also be consistent with theoretical knowledge of the qualitative behaviour of the system that has generated the data. If there is a clash between the behaviour predicted by the theory and clear features of the data, then the data must win, unless the data are flawed. All models are approximate and tentative. They should be modified if the data demand it. New data will often call for some rethinking of the model.
This may in part be a sample size issue. Data from six seals does not allow one to say much about the relative growth rates of different organs. Data from 30 seals is likely to present a markedly different picture, just because of the more adequate sample size.
In the next section I discuss common model assumptions.

11.3 Model Assumptions

All models make assumptions. It is a nice point of judgement to decide when departures from assumptions are serious. We discuss common assumptions.

Normality of the Error Distribution or Distributions

Approximate normality is enough. This assumption is, typically, less crucial for large samples than for small samples. In large samples, a process of averaging may move the distribution of statistics that we wish to examine close to normality. There is a key theorem, known as the "Central Limit Theorem", that describes what happens to the distribution of the mean as the sample size increases.

The problem is sometimes with a small number of outliers that seem to come from a distribution different from that for the main body of data. Of particular concern are observations that, as well as being outliers, are highly influential in the model fit. When fitting a line, points at the two extremes have the greatest influence. It is the same effect that one gets on a see-saw; sitting as far as possible away from the pivot gives the maximum leverage.

Independence

Failure of independence assumptions is a common source of wrong statistical inferences. For example, the assumption of independence is violated when there is clustering in the data. Within a cluster, responses are correlated. Whenever possible, statisticians like to avoid clustering and other forms of non-independence by gathering data in such a way that the independence assumption is guaranteed. This is why randomisation is so important in designed experiments, and random samples are so important in designed sample surveys.

Tests for independence may be of little use unless one has some idea of how the assumption may have failed, and the sample is reasonably large! So failure of the independence assumption is not only serious, it can be hard to identify.

Homogeneity of Variance

We'll say more about this as the discussion proceeds. Is the variation about predicted model values the same for all predicted values? Plots of residuals against fitted values may give useful clues on whether the variance really is homogeneous. For example, residuals may tend to fan out as fitted values increase, giving a "funnel" effect. This would indicate that the variance increases as fitted values increase.

Nonparametric Statistics

Nonparametric statistics such as the Wilcoxon tests or rank tests are not the answer to every problem of failure of assumptions. They still make assumptions, and one still has to worry whether the assumptions are realistic. In addition nonparametric statistics are generally much better suited to hypothesis testing than to estimation and model-building. They may give little insight.

The price for making weaker assumptions may be a reduced sensitivity to effects of interest. Often, reasonably strong parametric or other assumptions are necessary in order to find the effects that are present. For example, tree-based regression may be unsatisfactory in small or medium sized data sets where there is a parametric regression type structure.

11.4 Model Validation Issues

Model validation involves checking model assumptions, to the extent that this is possible. It also involves checking on the extent to which model estimates are affected by a small number of influential observations. The problem with highly influential observations is that, because they are so influential, any lack of fit is unlikely to show up when the usual diagnostics are used.

In very small data sets, e.g. with less than 10 or 15 data points, checks will reveal only gross departures from assumptions. Unless previous experience with similar data provides confidence that assumptions will be satisfied, inferences from such small data sets may be hazardous.

There are even more serious problems when the number of explanatory variables or factors is large relative to the number of observations. Broadly, there should be at least ten times as many observations as explanatory variables. (For any qualitative factor, subtract one from the number of levels, and count this as the number of variables contributed by that factor.) If there is no other basis for pruning down the number of variables and factors, it may be necessary to make an informed guess.

11.5 Broad Principles of Model Construction

Model structure must reflect data structure. Broadly, it must reflect all relevant fixed and random sources of variation. Elementary statistics courses typically emphasise fixed effects, assuming that there is a single random source of variation. In practice, multiple random sources of variation are the rule rather than the exception.

Consider as an example a study on the weight and other physical features of custard apples that treats different regions (North Queensland, Central Queensland, Southern Queensland, New South Wales) as fixed effects. Sources of variability may be differences between orchards, differences between trees within any one orchard, and differences between fruit within any one tree. Thus the model may be

y = region effect + orchard effect + tree effect + fruit effect

For purposes of comparing regions, comparisons must stand up relative to variation between orchards. So we want multiple orchards to be selected at random within each region.

In order to simplify the analysis we can use the mean for each orchard as the data. If there were 2 randomly chosen orchards in each of the 4 regions there would be 8 data items. The use of a convenience sample of orchards would invalidate the analysis. In addition issues of normality, which are not amenable to testing when the sample is so small, become important in such very small samples.

References and Further Reading

Chatfield, C. 1988. Problem Solving. A Statistician's Guide. Chapman & Hall, London.

12. Types of Data Structure

> The statistician is no longer an alchemist expected to produce gold from any worthless material offered him. He is more like a chemist capable of assaying exactly how much of value it contains, and capable also of extracting this amount, and no more.
> [Fisher, R.A., quoted in Bibby 1983: Quotes, Damned Quotes, and . . . Demast Books, Halifax, U.K.]

Data structure is in part determined by the way the data are collected, by the design. The design should be chosen to give results that are as accurate as possible. The statistical Design of Experiments, and Sampling Design, provide principles that should guide data collection. The analysis that one performs, if it is to be valid, must then reflect the data structure. It is a concern with observational data that the data structure may not be totally clear.
Data structure has both a fixed and a random component. The structure of the random component, although frequently an issue for the analysis of real data, typically gets little or no attention in introductory statistics courses.

It is scarcely possible to over-emphasise the importance of data structure issues. Yet elementary textbooks commonly ignore them. The ignoring of these issues is a common reason for faulty analyses. The central ideas can be understood without mathematics. The reverse is also true. It is possible to understand the mathematics without understanding the practical implications! Understanding the ideas and mastering the mathematics, while not totally separate activities, are not the same. Of course, those who want to get into the theory will find it essential to master the mathematics. For now, we can circumvent it.
We need to speak of fixed and random effects. Consider clinical trials. Experimental treatments are fixed effects. Any overall effect from age, in an experiment with humans or animals, is a fixed effect. Variation between subjects is a random effect; we want to generalise from the subjects in the trial to the wider population. In a multi-centre trial, there will be random effects associated with centres. We will want to generalise results to all centres that might have been chosen to participate in the trial.
The simplest case is where all effects are fixed, and errors can be neglected. We have a deterministic model. The result can be predicted so accurately that there is no need for statistics. Claims for such accuracy must of course be open to testing and checking. Some experimenters can be grossly over-optimistic about the accuracy of their equipment.
We will begin with an example that illustrates a simple but important aspect of data structure.

12.1 Example

Ten apples are taken from a box. A randomisation procedure assigns five to one tester, and the other five to another tester. Each tester makes two firmness tests on each of their five fruit. Firmness is measured by the pressure needed to push the flat end of a piece of rod through the surface of the fruit. Here are the results, in N/m^2:

	Fruit 1	2	3	4	5
Tester 1	6.8, 7.3	7.2, 7.3	7.4, 7.3	6.8,7.6	7.2, 6.5

	Fruit 6	7	8	9	10
Tester 2	7.7, 7.7	7.4, 7	7.2, 7.6	6.7, 6.7	7.2, 6.8

For comparing the testers do we have five results from each tester, or ten? The answer is that we have five experimental units for each tester. One way to do a t-test is to take means for each fruit. We then have five values (means) for one treatment, which we can compare with the five values for the other treatment.
What happens if we ignore the data structure, and compare ten values for one tester with ten values for the other tester? This pretends that we have ten experimental units for each tester. We will get results that suggest that the treatment means are more accurate than is really the case. We get a pretend standard error that is not the correct standard error of the mean. We may (though it does not happen for these data) underestimate the standard error of the treatment difference.

12.2 Fixed Effects, and a Simple Form of Error Structure

In the lawn roller example, the roller weight was a fixed effect. Earlier we discussed an experiment where Francis Bacon applied five different treatments to wheat seeds: water mixed with cow dung, urine, and three different wines. These five treatments were fixed effects
We do not need to spend much time discussing such experiments. Anyone who has done any kind of course on statistics probably feels that they understand the idea of a fixed effect. Problems arise when a fixed effects analysis, assuming independent and identically distributed errors, is used for data that have a more complex structure. What are "independent and identically distributed errors"? Recall again the lawn roller experiment. There is an error term that has a different value for each different roller. We assume that there is no correlation between the error terms for different rollers, and no tendency for the size of the error term to change with changing roller weight. If for example one started with the smallest roller on one side of the lawn, and moved systematically across the lawn as the roller became heavier, the independence assumption is unlikely to hold. The compressive strength of the soil may change systematically as one moves from one side to the other.

12.3 Two or More Nested Random Components

In the example that used a pressure tester on apples there were two levels of random variation – between measurements on the one fruit, and between different fruit. In that example we were able to calculate a mean for each fruit. This meant that we could forget about variation between different measurements on the one fruit. Matters get more complicated when the number of measurements is different for the different experimental units. It is then necessary to find a computer package that can handle such calculations. Fortunately, most reputable packages do now have such routines.
In the apple example, one level of variation (between measurements on the one apple) was nested within the other (between apples). Situations where the two levels are not nested are more complicated to handle.

Several Random Components

We might for example have variation between orchards, variation between trees within an orchard, variation between apples on the one tree, and variation between measurements on the one apple. A complication is that some of the variation on the tree, for example between the side most exposed to the sun and the side that is most shaded, may be systematic.

Several levels of random variation are common in survey data. For example, we may have multiple hospitals, multiple specialists within a hospital, and multiple patients for each specialist. Often an important issue is: To what population does one want to generalise? If the aim is to generalise to other hospitals, then hospitals must be treated as random. The hospital is the *primary sampling unit*.
Suppose that we calculate means for several primary sampling units. Then variation between those means is affected by components of variation at the primary sampling unit level and at all lower levels. In any analysis, greatest importance attaches to modelling variation at the primary sampling level. Failure to model variation at lower levels of the hierarchy of variation has effects that are much less serious. Variation at lower levels is important for the way that it affects variation at the primary sampling unit level.

12.4 Time Series Data

An example may be midday temperatures on successive days, over several years. There will be seasonal effects that will be more or less constant from year to year. There is likely to be a strong correlation between temperature for one day and temperature for the next.
Bivariate time series require special treatment. It is in general wrong to use ordinary least squares regression to regress one variable on the other, ignoring the time series structure. Fig. 13 plots all-Australia rainfall and (in the lower panel) the Southern Oscillation Index (SOI), for 1910-1992. (Data are from Nicholls et al. (1996). These data are discussed further in Chapter 14 Section 4. The Southern Oscillation Index is the difference in sea barometric pressure at sea level between Tahiti and Darwin.)

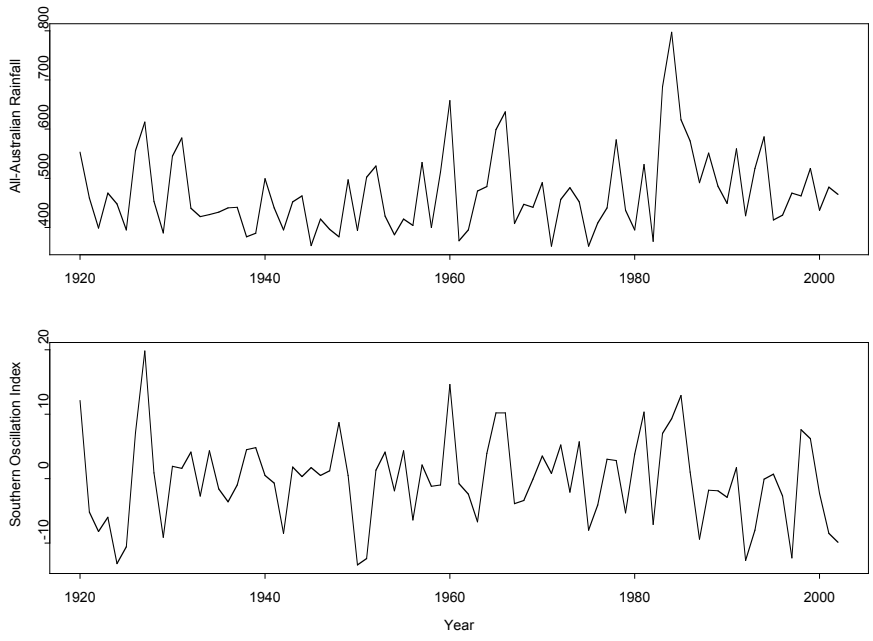


Fig. 13: All-Australian rainfall versus year, and Southern Oscillation Index versus year, for 1910-1992.

As can be verified by the appropriate analysis, figures for later years are correlated with results for earlier years. This complicates investigation of the relationship between All-Australian rainfall and the Southern Oscillation Index. There are further plots, and further discussion of these data, in section 14.4.

12.5 Repeated Measures Data

Fig. 14 is an example[31]. The points that are joined are the weights for one rat. Each rat has its own profile. Clearly there are overall trends for each treatment group. Weights at successive time points are likely to be correlated, with the correlations decreasing as points move further apart. An analysis that assumes independent rat weights at different times may be misleading.

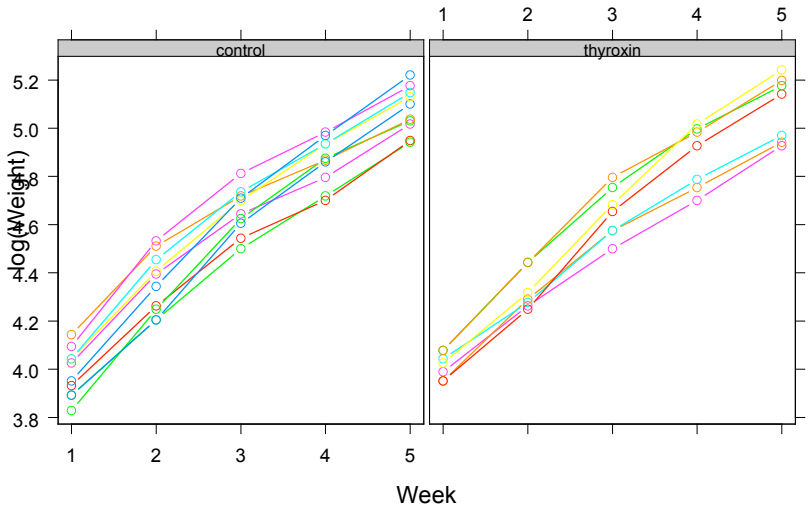


Fig. 14: Rat weight versus time, for alternative diet regimes.

Repeated measures data have a time-dependent component to their error structure, just as for time series data. A difference is that in repeated measures there may be many short series, while in most time series problems there is just one series[32]. In technical language, there may be many realisations in repeated measures, whereas in time series there is typically just one realisation.

12.6 Data Mining and Data Structure

Earlier, in chapter 10, we presented data mining as a form of exploratory data analysis. In practice, data mining may be a mix of exploratory data analysis and analyses that are concerned with predictive accuracy. A difference between data mining and traditional statistics is that data miners may have quite huge data sets. A data set with values of twenty or thirty variables for each of several hundred thousand records is, in the context of commercial data mining, small. Data structure is just as important with these large data sets as with smaller data sets. One has to ask:

- What are the major sources of variability?
- Relative to the population to which results are to be generalised, how well are those sources of variability represented? [How big a sample do we really have?]
- What is the likely potential for bias?

Because of their limited sampling of major sources of variability many physically large datasets are, from a statistical point of view, small. Data from a huge number of patient records from just ten hospitals may have a sample size n =10, for purposes of

[31] Data appear in Box, G.E.P. 1950. Problems in the analysis of growth and wear curves. Biometrics 6: 362-387.

[32] The series may however be multivariate, i.e. there may be measurements on several variables at each time point.

generalising to all hospitals. Worse, if this is a convenience sample, e.g. selected because these were the hospitals that made least fuss about making their records available, this may be an extremely biased sample. The sample population consists of those hospitals of which these ten might plausibly be a random sample, while the target population is all hospitals. Results may not generalise to other hospitals.

Methods for Analysing Large Data Sets

Standard statistical methods cannot, often, be used directly with very large data sets. There will almost inevitably be substantial structure in the data. Methods that ignore this structure, and that ignore the relationship of the sample population to the target population, will give optimistic estimates of the predictive accuracy of estimates. The wider population of hospitals may not give results similar to those from the selected ten.

If we know the data structure (which is not always the case), then in principle we could model it. The sheer size of the calculations that are required may make this problematic or impossible. One answer is to work with summary data. I will pursue the hospital example. In a first pass of the analysis, we might calculate relevant means or medians for each of the ten hospitals. These means or medians would then become the data for further analysis. Our massive data set has suddenly become rather small!

A more common approach may be to use estimation or classification methods that ignore major aspects of the data structure. Because major aspects of the data structure have been ignored, one cannot use classical methods to estimate predictive accuracy, even when they are available. So the approach is to develop the predictive model on a subset of the data, which is called the *training* set. The remainder of the data, the *test* set, is kept in reserve to use to get an estimate of the predictive accuracy of the model.

How should the test set be chosen? Consider those hospitals again. For generalising to the population from which the hospitals have been taken, we need data from a different set of ten hospitals. If we intend to apply our results to those same ten hospitals, then the test set might be a randomly chosen subset from the data from all ten hospitals, with the remaining data making up the training set.

The Targeting of Data Collection

It is typical of data mining exercises that they search for information different from what the data were collected to provide. This will often reduce the chances of finding useful information. Where data mining or other forms of exploratory data analysis do provide evidence of valuable ancillary information, it is likely that better targeting of the data collection process would yield even better information. Data mining exercises are far more likely to yield useful information if data collection has regard to the information that data miners may hope to get from it.

Questions: 1. An insurance company is developing a model for predicting fraud. It has a large database on insurance claims, going back fifteen years, with known fraud cases identified. What sorts of structure would you expect to find in the data? How would this affect your choice of training set and test set?

2. A petroleum exploration group has a large data set on sites that contained oil-bearing geological structures, and on superficially similar sites that turned out to contain no oil. What structure would you expect to find in the data? How might you choose your training and test sets?

Testing Predictions on New Data

This is the crucial test, however the estimates and associated accuracy assessments have been made. Internal assessments of accuracy should be regarded as provisional, pending this decisive test. For further discussion, see Chatfield (1995).

12.7 Outliers

Outliers are data values that seem anomalous. Often they stand out, in a graph, as quite separate from the rest of the data. Or a check for outliers in the course of an analysis may identify certain data values as outliers.
The first step should be to check whether a mistake has been made in measurement, or in recording or transcribing data. If there has not been a mistake, and nothing in the way that the data were collected explains the anomaly, the value must stand. However it is important to ensure that a few aberrant values do not unduly distort the analysis. It is often reasonable to separate outliers from the rest of the data, and examine them separately.

The Thinning of the Ozone Layer over Antarctica

This is an interesting example of the message that outliers may have for us. It warns us that noise in the data, from a variety of sources, may obscure their message.
Ozone (O_3) is an extremely reactive gas that is present in small quantities in the stratosphere, the region between about 10 km and 50 km above the earth. Ozone molecules are formed from the action of ultraviolet radiation from the sun on molecules of oxygen. They are in turn destroyed as they absorb ultraviolet radiation at slightly longer wavelengths. This filtering is fortunate, because longer wavelength ultraviolet radiation that reaches the earth's surface breaks down DNA and other complex molecules that occur in living organisms. The severe depletion of the ozone layer first came to the attention of the scientific world in 1985. It is largely a result of the escape into the earth's atmosphere of CFCs (complex fluorocarbons) used in refrigerators and aerosol spray cans.
Since the launch of the Total Ozone Mapping Spectrometer aboard the Nimbus-7 polar orbiting satellite in 1978, NASA has provided daily high-resolution maps of global ozone levels. NASA scientists noticed and began examining unusually low ozone values from the October 1983 data in July 1984. Up until 1983 no reliable measurement of ozone had been flagged that was lower than 180 Dobson Units (DU[33]). For this reason values less than 180 DU were flagged as outliers. The October 1983 data showed a sudden increase in the number of flags for values under 180 DU. Since this could have been the result of an instrument error, they checked their result against data from the Amundsen-Scott ground station at the South Pole. Unfortunately this station was, because of an error, reporting values of around 300 DU when the satellite instrument was reporting values under 180 DU. However they could find nothing wrong with the satellite data and finally, in late 1984, decided to submit it for reporting at a meeting that was held in Prague in August 1985.
In the meantime, Japanese and British scientists had been studying data from their ground stations. The Japanese published preliminary results from their Syowa station in a little read journal in December 1984. The NASA scientists do not seem to have had access to these data. An article reporting the anomalously low results from the British Halley Bay ground station appeared in Nature in May 1985, and quickly attracted wide attention.
Folklore has grown up around these events, suggesting that the NASA scientists had programmed their computer to ignore the low readings. In fact, it seems that they

[33] Amounts of ozone are measured in Dobson Units (DUs). One imagines that all the ozone above an area on the earth's surface brought down to the surface and spread evenly across the area, at standard temperature and pressure. Then 1 DU is a 0.01 mm thickness. The layer is about 260 DU at the tropics. Historically, it has increased in thickness as one moves to polar regions.

took the steps that one would expect from careful scientists when confronted with anomalous results. Their difficulty was that they were confronted with conflicting data. It took some months to sort out the anomaly. Komhyr et al. (1986) describe the mistake made with the data from the Amundsen-Scott ground station.

References and Further Reading

Data Structure Issues

Chatfield, C. 1995. Uncertainty, data mining and inference (with discussion). Journal of the Royal Statistical Society A, 158: 419-466.

Hubbard, R. and Armstrong, J.S. 1994. Replications and extensions in marketing: rarely published but quite contrary. International Journal of Research in Marketing 11: 233-248.

Data Mining

Friedman, J. H. 1997. Data Mining and Statistics. What's the Connection? Proc. of the 29th Symposium on the Interface: Computing Science and Statistics, May 1997, Houston, Texas.

The Thinning of the Ozone Layer

Chubachi, S. 1984. Preliminary result of ozone observations at Syowa station from February 1982 to January 1983 (Abstract). Memoirs of the National Institute of Polar Research, Special Issue No. 34, Proceedings of the Sixth Symposium on Polar Meteorology and Glaciology, National Institute of Polar Research, Tokyo, December 1984.

Farman, J. C., Gardiner, B. G., and Shanklin, J. D. 1985. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x fluctuation. Nature 315, May 1985, 207-210.

Komhyr, W. D. Grass, R. D. and Leonard, R. K. 1986. Total ozone decrease at South Pole, Antacrtica, 1964-1985. Geophysical Letters 13, November 1986 supplement, 1248-1251.

Nicholls, N., Lavery, B., Frederiksen, C. and Drosdowsky, W. 1996. Recent apparent changes in relationships between the El Ni™o – southern oscillation and Australian rainfall and temperature. Geophysical Research Letters 23: 3357-3360.

Pukelsheim, F. 1990. Robustness of statistical gossip and the Antarctic ozone hole. The Institute of Mathematical Statistics Bulletin 19: 540-542.

13. Presenting and Reporting Results

It is easy to lie with statistics. It is hard to tell the truth without statistics.
[Andrejs Dunkels]
The setting out of conclusions in a way that is vivid, simple, accurate and integrated with subject matter considerations is a very important part of statistical analysis.
[D. R. Cox 1981.]

Keep in mind from the beginning the required style and content for the eventual report, paper or thesis. This will help plan and structure your project. It is a good idea to include a provisional list of chapter or section headings in the research plan. This outline can be filled out and modified as the project proceeds.
Much of the focus of this chapter is on the presentation of statistical results. Efficient and cost-effective collection of quality data, and analysis that gets from the data all the information that is reasonably available, are central to research. The endpoint is the presentation of clear and coherent results. How does one present the message so that it accurately reflects the data, so that it is clear, and so that it will be heard and used?
Appendix II has a checklist for the authors of reports. Appendix III has a checklist of statistical presentation issues for the use of authors and referees. These supplement the comments in this chapter.

13.1 Keep the End Result in Clear Focus!

Right from the beginning of your study it is helpful to keep in view the required style, framework and content for the final report or thesis. Include a provisional list of chapter or section headings in the research plan. This skeleton framework will then be filled out and modified as the project proceeds. While changes may be needed, it is much more satisfactory and productive to modify a well-considered framework than to start careful planning only when something goes wrong!
Think carefully about the details of the information that will be presented, perhaps preparing provisional templates for the entry of data summaries as they become available. Along with these data summaries, there should be annotations that explain in detail the sources of information, details of instruments used, and other background information.
Depending on the specific research project, reporting serves several types of user. Busy professionals will wish to get quickly to the nub of what is presented, and may pay little attention to the details. Research peers or supervisors will look for a presentation that assists critical review. If a commercial organization has commissioned the report, their interest will be in knowing the key conclusions or recommendations.
Always, demonstrate that conclusions are soundly based. This may require a modest level of technical detail. In a report for a commercial client it is often best to consign technical detail to an appendix. Research theses may include substantial appendices.
Try to put yourself in the shoes of a reader of your report or thesis. Does it start with a summary that presents the major insights and conclusions? Does it present a clear coherent story? Does your report read well? Is the supporting evidence in place? Does the text focus on the major issues?
The next two sections are largely adapted from Maindonald (1992). The advice is set out in a pithy tutorial style. It is intended as a basis for consideration and debate.

13.2 General Presentation Issues

Here I set out broad principles. For published papers, it is necessary to follow the style that is laid out in a set of Instructions to Authors. Individual university departments may have their own preferred style for theses.

Summary Details

In a report, start with a half or one-page summary that sets out the main conclusions in a clear concise form. The abstract at the beginning of a published paper serves the same purpose. A research thesis may have an extended summary.

Scientific Background

Describe the scientific background and the rationale both for the study design and for the analysis.

Critical Comment

Acknowledge sources of information. Try to demonstrate that you have taken reasonable steps to find all relevant sources of existing information, and that you have evaluated it fairly and critically.

Major Patterns

Ensure that your presentation highlights the major patterns or effects evident from the data. Begin with a brief lucid summary that gives the main conclusions.

Models for your own Presentation

For a paper, critically examine papers that others have published in that journal. For a thesis, critically examine a well-regarded earlier thesis. If there is a scholarly book that canvasses themes similar to yours, examine how it is structured. It may serve as a starting point for developing a layout for your own work.

13.3 Statistical Presentation Issues

Substantial or Scientifically Important Effects

Focus first on the effects that are substantial and/or have special biological interest. Give the magnitudes involved in the text as well as in the tables, perhaps with a note on statistical significance given in parenthesis. Be sure to give standard errors, if available. These may be supplemented with tests of significance, if this seems necessary.

If less important though perhaps statistically significant effects are discussed at all, leave them till last. A reference to tables may be adequate.

Avoid unnecessary complication

It is not necessary to take the recipients of your report through the whole tortuous chain of reasoning that you have followed yourself. With hindsight, the argument can be simplified and streamlined. The graphs that you used to explore data may need substantial modification, if they are appropriate at all, when you come to present the data. Output from computer packages is rarely suitable for direct use – you will need to modify and adapt it.

Scientific Interpretation

Interpret all statistical results, as far as possible, in subject matter terms. Use the statistic that translates easily into subject matter terms in preference to a statistic that does not easily translate. Translate regression coefficients into rate of change terms whenever this seems helpful. Instead of reporting the relative risk of two medical treatment regimes, it is often more meaningful to report the number needed to treat (NNT) to avoid one death.

Translate all transformed values back into meaningful units for presentation. On graphs you may wish to plot using transformed units, with the axes labelled using the original units.

Economic Implications

It is often helpful to give an assessment of economic implications. But be realistic about uncertainties and limitations. Present calculations of economic return in such a way that it is straightforward to work out how results would be different under different economic conditions.

Scientific models

Analyses that use models that are motivated by scientific understanding are in general more insightful than analyses that use ad hoc and/or empirical models. Use any scientific understanding that is available to help direct the study design and the analysis. At the same time, be sensitive to questions that the data may raise for current scientific perceptions. Allow the data to speak for themselves.

Description of the design

Describe the study (experiment, sample survey, . . .) accurately and fairly. Be careful to identify experimental or sampling units and the units on which measurements were made. Where experimental data are reported describe the blocking structure, the exact form of randomisation, and other details of the experimental design. Explain the reasons for your choice of design. In field experiments either provide a drawing of the field layout, or else describe it in sufficient detail that the reader can sketch a diagram.

Describe realistically and accurately the population to which results apply.

Measures of Precision

Include SEs or SEDs (or their equivalent) and sample sizes wherever relevant. Where there are multiple error strata, be sure to quote the SE that is relevant to the comparison that is made. If results do not have the replication that would allow determination of the relevant SE, note this.

Note sources of variability that have been excluded in determining standard errors.

If the data allow it, present one SE rather than different SEs for different groups.

Curve fitting

When estimating a particular point on a fitted curve (eg. time to 99% mortality, or a maximum), it is crucial that the curve fits well in the neighbourhood of that point. If necessary, the fitting procedure should omit points that are at one (or both) extreme(s) from the point that is of interest.

Consider the use of a smoother as an alternative to the use of a curve that follows a specific mathematical form.

Measures of Relationship

The standard Pearson product-moment correlation is a measure of straight-line association. Use it only if you can justify restricting attention to linear association. Scatterplots will highlight gross departures from linearity. In addition there are statistical methods for testing linearity against specific curvilinear forms of response. Correlation and regression calculations should ordinarily be supported by relevant plots.

Reserve multiple or adjusted R^2 for comparisons across similar experimental or sampling designs. Use adjusted R^2 in preference to multiple R^2.

Note that a high correlation or multiple R^2 does not automatically imply that the relationship is adequate. The size of R^2 must be judged against the scatter in the data. If there is little scatter, it will require a correspondingly high R^2 to justify the claim that the fitted curve adequately captures the data.

Unless experience with earlier comparable results has shown what magnitude of R^2 to expect, do not rely on R^2 as a measure of model adequacy. Instead use a graphical check, perhaps backed up with a formal test for absence of systematic departure from the assumed form of response.

Significance Tests

Use p-values, if appropriate, to back up what you see as the major points that you have to make. Otherwise be abstemious in the use of p-values. Be sensitive to alternative ways of presenting the data that may reveal its major patterns.

Highlight the Trends

Where effects are quantitative use a trend curve or response surface analysis in preference to individual tests of significance. Multiple range tests are not appropriate for structured data.

Overall Analyses

Where work is widely extended in space and time, present an overall analysis that captures the major results. This extends to results that have been obtained by different workers, but carrying out closely related studies. Such analyses will identify how, after allowing for systematic effects due eg. to geography and soil type, local results stand up against site to site variation. In the absence of such an overview the effort that has gone into the individual trials may be largely wasted.

Consider the relevance of results to those who may use them. Farmers and horticulturists are interested in effects that apply to their farm or orchard. They can be confident in using results that have appeared consistently over different locations and years. Doctors are interested in results that apply to their patients.

Studies that have not yielded statistically significant results must be included in overview analyses.

Graphical Presentation

Put major conclusions into graphical form. Make captions comprehensive and informative. Use appropriate graphical presentations to reduce reliance on tables and on verbal description.

The best statistical software links statistical analysis closely with graphical presentation. Effective presentation of data and of statistical results will similarly link the results of the analysis with graphical presentation.

Design graphs to make their point tersely and clearly, with a minimum waste of ink. Avoid distracting irrelevancies. Label as necessary to identify important features. Use scatterplots in preference to e.g. bar graphs whenever the horizontal axis represents a quantitative effect. Keep the information to ink ratio in mind.

Use graphs from which information can be read directly and easily in preference to those that rely on visual impression and perspective. Thus in scientific papers contour plots are much preferable to surface plots or two-dimensional bar-graphs.

Draw graphs so that reduction and reproduction will not interfere with visual clarity.

Explain clearly how error bars should be interpreted — ± SE limits, ± 95% confidence interval, ± SD limits, or .... You must explain what source of error is represented. It is pointless to present information on a source of error that is of little or no interest.

References and Further Reading

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639.

[The checklist that appeared as part of this statement can be found at: http://www.ama-assn.org/public/journals/jama/jlist.htm]

Cleveland, W. S. 1993. Visualizing Data. Hobart Press, Summit, New Jersey.

Cox, D. R. 1981. Theory and general principle in statistics: the address of the President (with Proceedings). Journal of the Royal Statistical Society, A 144: 289-297.

Finney, D.J. 1988-1989. Was this in your statistics textbook? Experimental Agriculture 24:153-161; 24:343-353; 24:421-432; 25:11-25; 25:165-175; 25:291-311.

Gardner, M. J.; Altman, D. G.; Jones, D. R.; Machin, D. 1983. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? British Medical Journal 286: 1485-1488.

Maindonald, J.H. 1992. Statistical Design, analysis and presentation issues, New Zealand Journal of Agricultural Research 35: 121 - 141, 1992.

Murray, A.W.A. 1988. Recommendations of the editorial board on use of statistics in papers submitted to JSFA --- guidelines to authors as formulated by A W A Murray. Journal of the Science of Food and Agriculture 42, no. 1, following p. 94. Reprinted in vol. 61, no. 1, 1993.

Perry, J. N. 1986 Multiple-comparison procedures: a dissenting view. Journal of Economic Entomology 79: 1149-1155.

Wainer, H. 1997. Visual Revelations. Springer-Verlag, New York

14. Critical Review – Examples

> The popular impression that disproof represents a negative view of science arises from a common, but erroneous, view of history. . . . In this view, any science begins in the nothingness of ignorance and moves towards truth by gathering more and more information, constructing theories as facts accumulate. In such a world, debunking would be primarily negative, for it would only shuck some rotten apples from the barrel of accumulating knowledge. But the barrel of theory is always full; sciences work with elaborated contexts for explaining facts from the very outset. . . . Science advances primarily by replacement, not addition. If the barrel of theory is always full, then the rotten apples must be discarded before better ones can be added.
>
> [Gould, S. J. 1981. The Mismeasure of Man, pp. 321-322. Penguin, London.]

Typically, a paper will present summary information from an analysis of the data. You must assess whether the data really do address the research question, whether the analysis is correct, and whether the results support the interpretations that are placed on the results. Thus large questions of statistical design and interpretation may arise in reviewing of the literature, before you start on your own research.

Examples will illustrate issues that may arise in examining the results of other workers. Some demonstrate serious faults. Others are included because they illustrate interesting and important points.

14.1 Inadequate or Faulty use of Data

Straight Line or Curve?

Clutton-Brock et al. (1999) were interested in how the percentage of time that adult meerkats spent on guard varied with the size of the group.

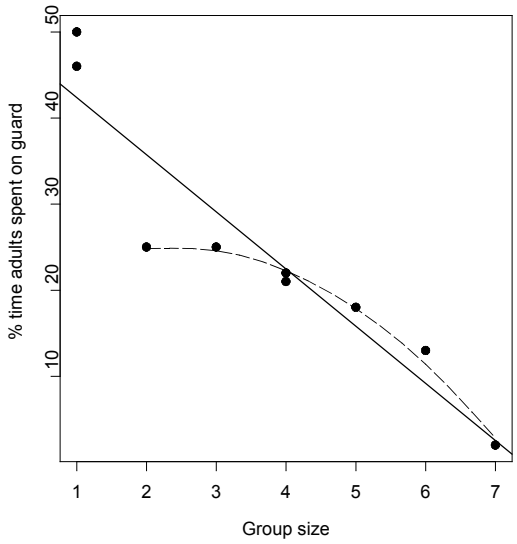


Fig. 15: Mean % of foraging time spent by adult meerkats on guard in groups of different sizes. The fitted straight line is clearly inappropriate.

Fig. 15 reproduces a straight line that, in that paper, was fitted to the data. I have fitted the dashed curve, for group sizes of two or more. The dashed curve seems an

accurate summary, for groups of more than one. It is entirely plausible that animals that are isolated from other adults will spend a much larger proportion of their time on guard. It is hard to understand why these authors have insisted on fitting a line, and surprising that it escaped Science's editorial processes.

Roots of Kiwifruit Vines

Reid and Petrie (1991) compared the root system of a waterlogged vine with the root system of a vine that was not waterlogged. The authors had a large number of samples from each vine. However, because each set of results is from one vine only, there is no way to know how much of the difference was due to differences between the vines and how much to the waterlogging. This is an example of what is sometimes called pseudo-replication. Elsewhere (Maindonald 1992) I suggest that it would be better called Clayton's replication, i.e. replication that is not really replication.

The 0.3m x 1.6m acrylic window of an observation chamber or *rhizotron* gave a view of the vertical section of the part of the root system that bordered on to the window. Over time, some roots will die, and some new roots will appear. Tracings of the root system, taken at intervals of between 1 and 15 days, were used to identify the dates at which roots were first and last observed. Counting ceased when 80 roots had been observed at each time point. This made it possible to estimate, at each time, an approximate age for each root then visible. Here are results:

Average root age (days) on	`Control' Vine	`Waterlogged' Vine
9 Jan	36.7 ± 4.5	38.2 ± 4.0
30 Jan	45.6 ± 4.7	42.7 ± 4.5
9 May	100.4 ± 9.4	72.9 ± 7.63

Table 1: Comparison between roots on waterlogged vine and roots on control vine. Results are given as mean ± SEM, in each instance based on $n = 80$ roots

The paper is vague on how the roots that were observed at each time were made up to a total of 80. How were roots selected? It appears that the authors used a "convenience sample", taking whatever roots were at hand until the total reached 80. In fact the root systems of eight vines were visible from the rhizotron. Focusing all the measurement effort on just two vines therefore appears a poor use of resources. Less accurate information on all vines would surely have been far more useful. The authors are aware that variation between vines may be huge. In another paper they document this large variation. At some sites vines took two months to show adverse symptoms, while at others vines showed stress within a few days of flooding and died within a month. So how useful is detailed information that is based on the response of just one vine?

Custard Apples

Here is part of a table from the 1996 biennial review of the Horticulture Postharvest Group in the Queensland Department of Primary Industry:

	CQ	NQ	SEQ	NSW
Weight	423 a	451 a	425 a	451 a
Days to eating soft	5.8 a	7.2 b	5.1 ac	4.5 c
Woodiness (0-5)	3.1 a	1.6 bc	2.1 b	1.5 c
Pointiness (1-3)	1.4 ab	1.5 a	1.2 bc	1.0 c

Table 2: Data are for mature "African Pride" custard apples, harvested from two farms in each of the production areas Central Queensland (CQ), North Queensland (NQ), Southeast Queensland (SEQ), and Northern New South Wales (NSW).

The letters (a, b, c) are presumably designed to allow an assessment of statistical significance at the 5% level. Means that are identified with the same letter cannot be distinguished. It is however unclear whether this is a comparison-wise significance level or an experiment-wise significance level, i.e. whether the least significant difference has been adjusted to the number of comparisons that are to be made. (Often the letters are called `Duncan' letters. Duncan was the originator of various versions of a widely used multiple comparison test.)
There are two levels of variation that need to concern us here – between fruit within any orchard, and between orchards. If results are to apply to all orchards in a production area, then we need a random sample of orchards. If the same size of sample is taken from each orchard, then the correct analysis can be based on orchard means. There is a question whether this is how the analysis was done, or whether results from individual fruit were treated as replicate values.
Let us however assume that the analysis was done correctly, assuming random sampling of orchards. There are then two values per production area. Were the orchards really chosen randomly from a list, or were two orchards that were close at hand used as a convenience sample?
Finally, how close to normal is the distribution of orchard means? With such a small sample of orchards, the normality assumption becomes important?
The bottom line is that a sceptic can find a number of reasons for not taking these data too seriously. At this point there can be no firm judgement on how custard apples differ between regions.

Allometric Relationships

In animal growth studies, there is an interest in the rate of growth of one organ relative to another. For example, heart growth may be related to increase in body weight. As the animal must be sacrificed to get the weights of organs, it is impossible to get data on growth profiles for individual animals. For seals and dolphins and some other protected marine species, the main source of information is animals that have died, often snared in trawl nets, as an unintended consequence of commercial fishing. For each animal, the data provide information at just one point in time, when they died. At best, if conditions have not changed too much over the lifetimes of the animals in the sample, the data may indicate the average of the population growth profiles. If for example sample ages range from 1 to 10 years, it is pertinent to ask how food availability may have changed over the past ten years, and whether this may have had differential effects on the different ages of animal in the sample.
Allometric growth implies that, for any two measurements x and y

$$y = a\,x^b$$

where x may for example be body weight and y heart weight. It may alternatively be written

$$\log(y) = a + b \log(x)$$

i.e. $y' = a + b\,x'$, where $y' = \log(y)$ and $x' = \log(x)$. If $b = 1$, then the two organs (e.g. heart and body weight) grow at the same rate.

Thus we have an equation that can be fitted by linear regression methods. As we will consider below, it is doubtful whether it is the regression relationship that we really want; this allows us to predict values of y' given a value for x'. I will return to this point later.

Gihr and Pilleri (1969, p. 43) present such data for particular species from the genera Phocaena (porpoises), Stenella (a seal genus) and Delphinus (dolphins). Fig. 16 shows the plot of heart weight against body weight for 17 individuals from the species *Phocaena phocaena*.

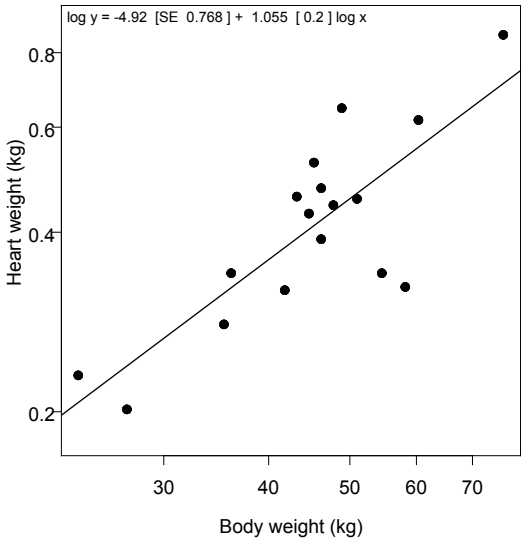


Fig. 16: Heart weight versus body weight, for 17 members of the species Phocaena phocaena.

Although the estimate of the exponent b (= 1.055) differs from 1.0 by less than a quarter of the standard error, Gihr and Pilleri state that heart weight is increasing more rapidly than the body weight. They make comparable "increasing more rapidly than" and "increasing more slowly than" statements about other organs, for this and for other species.

Now in fact these authors present extensive calculations that include (Table 10 on pp. 40-41) details of the standard errors of regression coefficients. There are coefficients for seven organs, in most instances for each of three species. They present p-values for a test of significance that the coefficients are zero. Even though the focus of the discussion in the text is on the comparison with 1, they seem uninterested in whether this difference is statistically significant or even whether it is more than SE[b]. Having demonstrated that a coefficient is statistically different from zero, they then treat it as, to all intents and purposes, exact.

Of the twenty coefficients that they present, only two differ from 1 by more than statistical error (p=0.05). Both of these are much less than 1, and are for small samples (n=11). One of these has b=0.042, and is anyway not significantly different from zero. In passing, note that where there is large scatter about the line, estimates

of *b* will be biased down. In small samples there will sometimes be large variability about the regression line just as a result of chance.
Gihr and Pilleri have not distinguished what is likely to be an accident of this particular sample from features that are less readily explained as accidents of sampling variation. They have thus failed in what is surely a major responsibility of the writer of any paper, to draw to the reader's attention those scientifically interesting features that may be independent of the accidents of sampling.
We return to the issue of how the equation is to be estimated. There are in fact two regression relationships, one for predicting Y given X, and one for predicting X given Y. These coefficients may be dramatically different. Here, using related data, is an example. For liver weight in *Delphinus delphis*, the regression slope for log(liver weight) on log(body weight) is 0.043. If on the other hand one regresses log(body weight) on log(liver weight) and turns this around to express log(liver weight) in terms of log(body weight), the coefficient is 17. Neither coefficient is meaningful, as Fig. 17 makes clear.

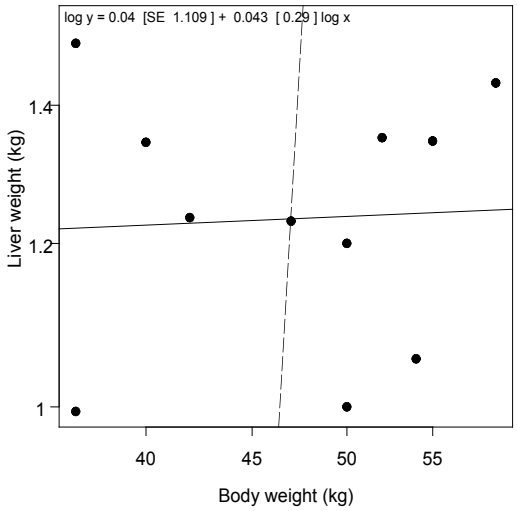


Fig. 17: Plot of liver weight versus body weight, both on log scales, for 11 members of the species Delphinus delphinus. The solid line is the regression of log(liver weight) on log(body weight), while the dashed line is for log(body weight) on log(liver weight). Neither line is meaningful.

By comparison with Fig 16, observe that there is a much more restricted range of body weights. The range of body weights is not wide enough to allow detection of the relationship (if any) between liver weight and body weight.
A plausible point of view for the present application is that there is an underlying functional relationship. The analysis assumes that observed values of log(organ weight) and log(body weight) differ from the values for this underlying functional relationship by independent random amounts. The line for the underlying functional relationship will lie between the regression line for Y on X and the line for X on Y.

An Insect Disinfestation Experiment

A formal target for research aimed at developing treatments that will remove insect pests from export fruit (disinfestation) is to design a treatment that will allow the survival of at most one insect in 30,000. This is equivalent to a mortality of 99.9968%. There are broadly similar standards in many different countries.

Experiments that are of modest size, e.g. 300 insects at each of seven or eight treatment times or doses, are used to predict a treatment that will achieve this very high mortality. This prediction usually involves a large extrapolation from the experimental data. The estimate is then tested out in a large-scale trial, perhaps with 100,000 insects. If more than one insect out of 100,000 survives in this large-scale trial, this is taken to indicate a lack of assurance that no more than 3.3 in 100,000 would on average survive. The large-scale trial must be repeated, using a higher treatment time or dose.

Jessup and Baheer (1990) give fruitfly mortalities after various times in low temperature storage. Around 200 larvae were used at each time point. They quote an estimate of 29.4 days for the time to 99.9968% mortality. The authors then undertook a large-scale experiment in which they tested out a storage time of 12 days. Clearly they did not believe the analysis that they presented. One must assume that they examined the graph directly, then making a guess that 12 days would be adequate.

It appears that these authors fitted a straight line to the plot of the probit of `treatment induced' mortality against log(time) that is shown in Fig. 18.

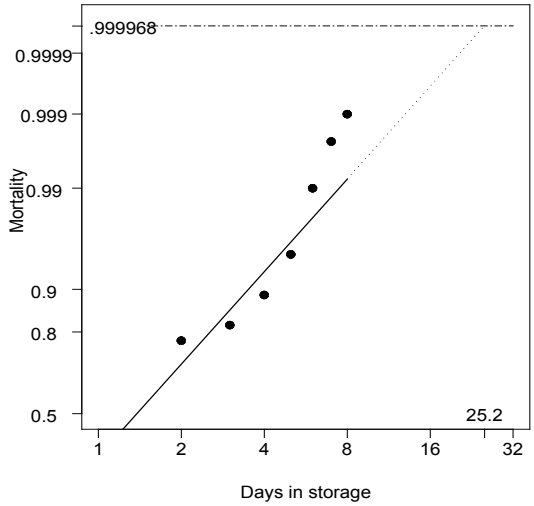


Fig. 18: Probit model fitted to first instar data of Jessup & Baheer (1990).

The plot makes it clear the response is far from linear. Extrapolation exaggerates the effect of the bad fit. My estimate, from extrapolating the straight-line, of the time to 99.9968% mortality is 25.2 days. This assumes equal numbers of insects at each time point. The discrepancy with Jessup and Baheer's estimate of 29.4 days most likely arises because the authors used estimates, not presented in the paper, of total numbers for each time point.

The authors give results for three further insect stages. The LT99.9968 is the time required to kill 99.9968% of the relevant insect stage. The table below gives the LT99.9968 estimates for eggs and the three fruit fly stages:

Stage	LT99.9968	95% CI
Egg	17.3 days	17.7-18days
1st instar	29.4 days	28-30.9
2nd instar	14.4 days	13.8-15.1
3rd instar	24.4 days	23.2-25.7

Table 3: LT99.9968 estimates for eggs and the three fruit fly stages of Queensland fruit fly (*Dacus tryoni*).

The plots for the egg, 2nd instar and 3rd instar are in fact not too far from linear, so that for those instars the main reason for complaint is that there is a huge extrapolation. Thus we have one LT99.9968 that is based on a line that does not fit, and three LT99.9968s derived from huge extrapolations of lines that fit moderately well. On the basis of the longer estimated time to 99.9968% mortality for the 1st instar, the authors claim that the first instar is more resistant to low temperature storage than the second and third instars[34].

Such inappropriate fits seem a commonplace in the disinfestation literature. Few authors present evidence that their data are consistent with the assumed form of response. It has been common to assume, without further investigation, a straight line relationship between the probit of mortality and either log(x) or x, where x is either dose or time. Disinfestation experiments often use huge numbers of insects, and a plot will readily show serious departures from the assumed form of response.

In part, these practices may reflect the history of dose-mortality studies. Conventional animal dose-mortality experiments typically used relatively few animals, in the tens or twenties. Systematic departures from the assumed form of response did not stand out against the scatter in the data.

Confused and Erroneous Claims

Articles in the *Journal of Economic Entomology* which demonstrate statistical misunderstanding are disturbingly common. Thomas and Mangan (1997) is astonishing in this respect. First note that the tolerance distribution plots the expected mortality against dose or exposure. The population mortality curve that corresponds to the points in Fig. 18 plots mortality against number of days in storage.

Thomas and Mangin claim that sample size affects the tolerance distribution. As the tolerance distribution is a property of the population of insects, how could it? One incorrect formula is correctly derived from another incorrect formula. They suggest, wrongly, that a lognormal tolerance distribution makes it appropriate to assume a logit or complementary log-log distribution. They misrepresent results that are presented in authors from whom they quote. They claim, wrongly, that a model that gives a narrower prediction interval should be preferred to one that gives a wider prediction interval. (If the model is wrong the prediction interval will also be wrong, and may well be too narrow.) They claim to use the Maentel-Haenzel test for testing goodness of fit to an assumed tolerance distribution, an entirely inappropriate use. This is an incomplete list.

[34] The argument is then that effort can be focused on developing a treatment that is effective with the "most resistant stage", that any treatment that is effective for the "most resistant stage" will be equally or more effective with later stages.

This paper is unusual in the extent of its errors and confused statements. If a paper seems to make nonsensical claims, one should not rule out the possibility that they are indeed nonsensical. The literature places a huge burden of discrimination on the reader.

A Claimed Link Between Fluoridation and Cancer

Yiamouyiannis and Burk (1977)[35] compared the pattern of cancer death rates for the 10 largest fluoridated cities in the United States with the ten largest cities not fluoridated in 1969, but with a cancer death rate of 155 per 100,000. Fig. 19 shows the comparison.

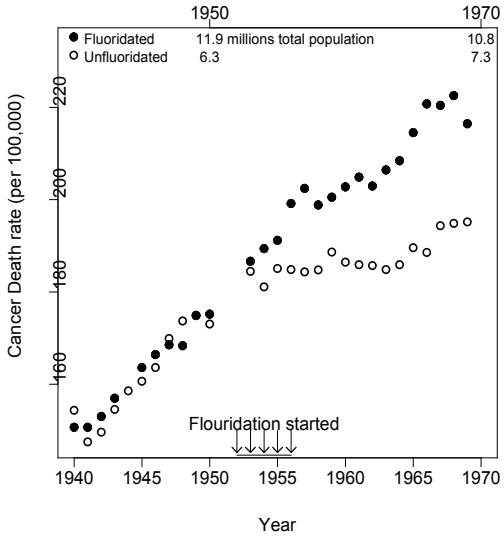


Fig. 19: Cancer rates in the ten largest fluoridated US cities, compared with cancer death rates in ten large unfluoridated cities. The ten largest unfluoridated cities were chosen that had comparable cancer death rates (>155 per 100,000) in 1953.

In the fluoridated cities, fluoridation began in one of the years 1952-1956. Yiamouyiannis and Burk suggest that this led to the subsequent divergence between the cancer rates in the two groups of cities. This divergence occurred over 1955-1970.

Because fluoridated cities were not randomly assigned to one or other regime, the divergence in crude death rates of itself proves nothing. Other changes were taking place at the same time. There were large changes in the population age structure in both sets of cities, as might be expected given the changes in total population in the two sets of cities.

Subsequent discussion focused on comparing 1970 with 1950. Yiamouyiannis and Burk acknowledge that the age structure of the two sets of cities changed in different ways over 1950 - 1970. They then made the comparison separately for the age groups 0-24, 25-44, 45-64 and 65+. There was no difference in either of the under 44 age groups, but large differences in the 45-64 and 65+ groups.

[35] The journal (`Fluoride') in which this paper appears is not a recognised scientific journal. It provides a vehicle for articles which support an anti-fluoridation point of view. However to the extent that the arguments are soundly based on reliable data, they must be taken seriously. There are, as we will see, serious flaws in their argument. Teasing out what is wrong with their arguments requires a fair amount of subtlety. It is not hard to find articles in mainstream scientific journals where there are comparable faults.

They are aware that these differences may be due to differences in the sex, age and racial breakdown within these groups. They then proceed as follows

1. They check the age distribution within these 45-64 and 65+ categories. They argue that because the percentages never differ by more than about 3%, the age distribution within these groups is "virtually identical" for fluoridated and unfluoridated, both in 1950 and 1970.
2. They note also that there is a greater percentage of non-whites in the fluoridated than in the unfluoridated group. They then regress age-corrected (i.e. corrected for the broad categories) mortality rates against % non-white population in each city. None of the correlations were significant.
3. They did the same regression, but now limiting attention to the 45-64 age group. Again the correlations were not significant.

The "virtually identical" claim of point 1 will not do. Examination of Fig. 20 will show why.

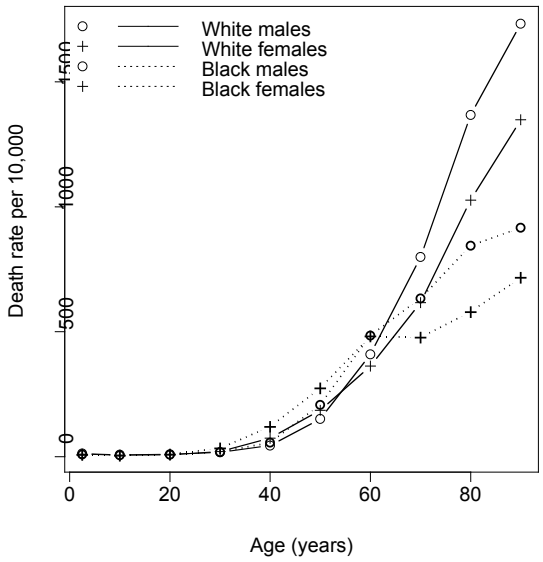


Fig. 20: Age specific death rates from malignant neoplasms (cancer). Note the different patterns of mortality for white males, white females, non-white males, and non-white females.

Observe how mortality increases with increasing age. A slight shift towards more old people, and especially towards more older males, may lead to a large increase in cancer rates. By comparison, changes in the age structure for the under 30s will have almost no effect.

There are two problems with use of correlations as in points 2 and 3 above:

1. We are dealing with time series, not with independent data values. So the "usual" tests for significance are not valid.
2. Even if these tests of significance were valid, they would not be relevant. There is a difference in the population structure of the two sets of cities (fluoridated and unfluoridated). We want to know whether the differences in population structure can explain the different cancer mortality rates, not whether the difference is in some sense statistically significant. While this may appear a subtle point, it is an important one, that may trip up statistical neophytes.

The way to find out whether the differences in population structure can explain the different cancer mortality rates is to calculate the cancer mortality rates that would

apply to some specific mix of ages, sexes and races. Oldham and Newell (1977) did exactly that, and found that the difference all but disappeared.

Apparent Biological Activity at Impossibly Low Dilutions

In June 1988 the journal *Nature* published a report, by Davenas et al., which appeared to demonstrate the biological activity of solutions that initially contained anti-IgE antibodies, but at dilutions so extreme that most of the diluted solutions should contain no antibodies. A later issue published a report, by Maddox et al. (1988), that gave the report of a team that had visited the laboratory where these experiments were performed.

By the usual standard of laboratories for pharmacology and allergy, Davenas et al. may have carried out their experiments with care. They were however looking for effects where very minor and ordinarily unimportant contamination will invalidate results. Possibilities for contamination that were identified by the investigators included: possibly misplaced test-tube stoppers, slight contamination of unintended wells during the pipetting process, and general laboratory contamination that may have arisen because experiments were carried out on an open bench. Wells were not counted in duplicate.

The investigating team also carried out their own double blind experiment, following their own very strict protocol. Results turned out negative, i.e. no effect was found. In investigations at or near the limits of scientific detectability, biases (here sources of contamination) that are ordinarily unimportant may vitiate results.

Other Examples

Other examples of poor or erroneous statistical analyses are discussed in Andersen (1990), Chanter (1981), Gardner et al. (1983), Gates (1991), Maindonald and Cox (1984), Padak (1989) and Thomas (1978).

14.2 Probing the Reasons for Differences in Results – An Example

Penetrometer and Operator Effects in Measuring Fruit Firmness

It often happens that some researchers find a claimed effect. Others do not. Who should one believe? Where the evidence is experimental, an important issue is whether some experiments were inherently more precise than others. Precision is affected both by the measurement instruments and by the statistical design.

Here is an example, from research (Harker et al. 1996) that compared different instruments for measuring fruit firmness. In addition there was interest in possible operator effects. There were four measurements on each kiwifruit. On apples more than four measurements should be possible, but none of the papers I have seen uses more than four.

The following types of experimental design are possible:

1) Use just one device and one operator per fruit, so that fruit to fruit variation affects comparisons both between operators and between devices.
   a) Devices are compared, but not operators (Abbott et al. 1976)
   b) Both devices and operators are compared (Blanpied et al., 1978).
2) For each fruit, compare multiple devices. Use one operator per fruit. (Bongers 1992; Lehman-Salada 1996).
3) For each fruit, compare multiple operators. Use one device per fruit. (Lehman-Salada 1996, in a further experiment).

4) For each fruit, compare multiple device-operator combinations (Harker et al. 1996).

Harker et al. (1996) document the extent to which, in their study
1) Comparisons made on the same fruit are typically much more accurate than comparisons that use different fruit for different devices etc.
2) The gain is much larger for fruit at harvest (when they are firm) than for fruit after storage, when they are relatively much softer.

So another issue is whether the different studies used harvest fruit or storage fruit. Harker et al. used both storage and harvest fruit, as did Lehman-Salada. Blanpied et al. seem to have used storage fruit. Others used storage fruit.
These various pieces of information give us a context in which to interpret the results. Where comparisons are made within fruit, one expects relatively good precision. Where comparisons use the between fruit level of variation, we expect relatively poor precision. Thus, it is not surprising that Bongers et al. found no differences between operators.

14.3 Instructive Examples

Fickle Mice

A standard approach in mice studies investigating a suspected behavioural role for a gene has been to knock the gene out, then investigate the effect on the behaviour. There have however been frequent cases where later researchers have been unable to reproduce published claims for a gene/behaviour association, or have even found an effect that goes in the other direction. This was the impetus for a study, reported in Crabbe et al. (1999) where researchers in three laboratories tested six mouse behaviours simultaneously, using exactly the same inbred strain and one null mutant strain. Stringent precautions were taken to ensure identical test apparatus, testing protocols and animal husbandry.
For some behaviours the authors report consistent effects across all three laboratories. For others there was no consistency. In one of the strains tested a receptor for the neurotransmitter molecule seratonin was knocked out. In one location there was more maze activity than for controls with intact receptors, in another there was less activity, while in the third location the loss of the receptor appeared to make no difference. A weakness of the study is that the experimental procedure was not repeated at the three separate laboratories. A fully adequate replication at a laboratory would require, as well as repeating other aspects of the experimental setup, the use of separate and different operators.

An Interesting Case of Confounding

Cohen (1996) describes a trial where researchers were looking for a difference between two analgesic drugs. An alert reviewer spotted that some of the treatment groups contained more women than men, and proposed a re-analysis to determine whether this accounted for the results. In fact, almost the whole effect could be ascribed to sex differences in the response.
Other related work by the same researchers can be used to illustrate the effect. Here is the allocation of subjects:

	Pantazocine + Placebo	Pantazocine + Baclofen
Females	7	15

Males	9	3

Treatment A was Pantazocine plus placebo, while treatment B was Pantazocine plus Baclofen. The difference in the two treatments is between Placebo and Baclofen. Suppose we do an analysis that ignores the sex effect. Any difference we find may be a difference between Baclofen and Placebo, or it may be due to the greater preponderance of females in the Baclofeb treatment group. Fig. 21 shows the separate results, as a function of time after administration.

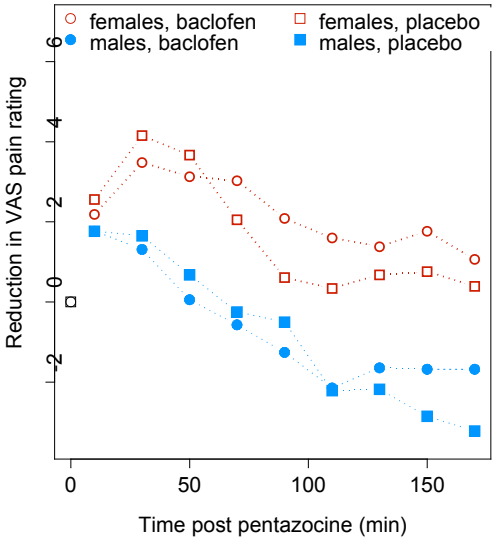


Fig. 21: The effect of pentazocine on post-operative pain, with (circles) and without (squares) preoperatively administered baclofen.

Fig. 22 shows how, if we ignore the gender effect and combine the two sets of results, the Baclofen result is weighted towards the result for females.

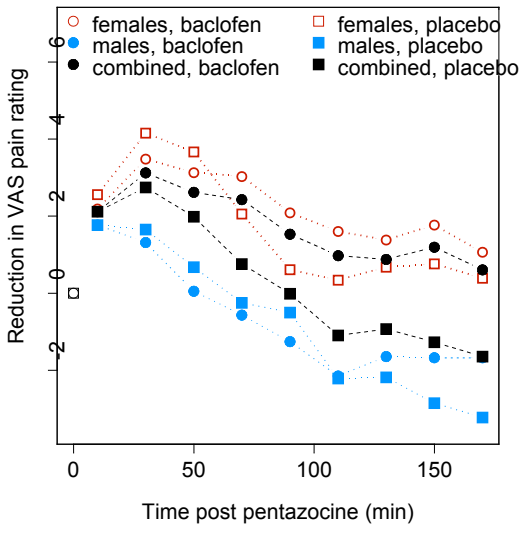


Fig. 22: The effect of pentazocine on post-operative pain, with (circles) and without (squares) preoperatively administered baclofen. Also shown (solid black circles & squares) is the combined result, obtained by averaging over all patients.

*14.4 Bivariate Time Series

Southern Oscillation and Australian Rainfall

The plots in Fig. 23 are reproduced from data in Nicholls et al. (1996)[36].

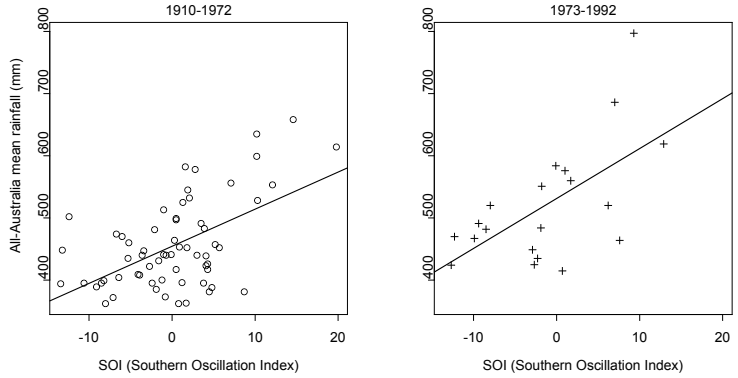


Fig. 23 : All-Australia mean rainfall versus Southern Oscillation Index. Scatterplots, with fitted lines, are shown separately for 1910-1972 & 1973-1992.

The authors were looking for a relationship between the Southern Oscillation Index (SOI) and climate information. They present a time series of the All-Australia spatially averaged rainfall, and for the SOI values. (In addition they give all-Australia maximum temperatures.) Fig. 23 shows separate plots of rainfall versus SOI, for 1910-1972 and 1973-1992. The two plots follow a relatively similar pattern. There are several objections to their analysis.

1. There seems no justification, independent of the data, for fitting a different response line for 1972-1992.
2. The graphs suggest that the relationship is curvilinear rather than linear.
3. Their analysis did not allow for the sequential correlation structure in the data.

I have done an analysis that attempts to account for the time-dependent structure in the data. This suggests that a clearer relationship between rainfall and SOI than the least squares analysis indicates.

14.5 Multiple Papers, and the Task of Overview

This is an important issue for, among others, medical researchers. For example, suppose you want to study the effect of zinc supplementation on the growth of children. How does one start? A careful critical overview, if you can find it, is likely to be the best place to start. You should look for indications that the paper has weighed the reliability and usefulness of the evidence that is presented in the different papers, and has provided a summary that gives greatest weight to the most complete and reliable sources. Overview papers along the lines: "I've read all these different papers, and here is my assessment" ask you to put a lot of faith in the judgement of the reviewer. You may be getting little more than another opinion to set alongside the opinions that are expressed in the various papers. Chalmers and Altman (1995) is a good summary of the uses and hazards of systematic reviews. Sackett et al. (1977) press very strongly the merits of well-conducted systematic reviews.

Chapter 13 of Linus Pauling's 1986 book *How to Live Longer and Feel Better* reviewed evidence on the claimed benefits of Vitamin C in preventing colds. For all

[36] Dr Nicholls kindly supplied me with a copy of the data.

his scientific prowess, Pauling did a poor job of systematic review. He says nothing about the design of his survey of Vitamin C trials. It is unclear how he chose the particular trials that he included, whether methodological quality was assessed without knowledge of the outcome. It is not clear how he decided whether the result of any trial was positive, negative or indecisive. A more recent study, that of Kleijnen, ter Riet and Knipschild (1992), has addressed these deficiencies. They did a thorough literature search and systematic review. Their conclusion was that Vitamin C, even in gram quantities per day, will not prevent a cold. Once one has a cold, a large dose of Vitamin C may however slightly reduce its duration and severity.
In their review of the quality of medical overview papers, Oxman and Güyatt (1983) found that content expertise was inversely related to the quality of the review. Moreover there was poor agreement among content experts on the methodological quality of reviews. The implication may be that the conduct of systematic reviews requires skills that must be learned. This is true as much for content experts as for those who have no special content expertise. Sackett (1983) gives a more extreme view.

Smoking and Health

Physicians first noticed an apparent increase in death rates from lung cancer in the 1920s. Initially there seemed a possibility that it might be an artefact of improvements in diagnosis. By the 1940s there was general agreement that the increase was real, and investigators began to focus on a search for possible causes. Smoking was one of several theories. Pearl (1938), in a report that was widely criticised, compared death rates of smokers with those of non-smokers, finding that heavy smokers had higher death rates than non-smokers.
The first reports on two studies that had a major impact appeared in 1950, one in the UK and one in the USA. The UK study, by Doll and Hill, was a hospital-based case-control study. Cases were persons admitted to hospital after diagnosis with lung cancer. Controls were patients admitted for other reasons. The investigators then classified both cases and controls according to whether or not they smoked. Here is what they found:

	Cancer diagnosis (case)	Other diagnosis (control)
Smoker	1,350	1,296
Non-smoker	7	61

Notice that non-smokers are rare in both groups, much rarer than in the population at large, where at least 40% (more women then men) would have been non-smokers. The key point is that lung cancer is nearly 5 times more common among smokers (1350/(1350+1296) = 51%), than among non-smokers (7/(7+61)=10.3%). If hospitalised patients without lung cancer were a random sample from the general population, it would follow[37] that the relative rates are much greater than 5.
Patients are not hospitalised at random. Nevertheless the results had troubling implications, whatever their relevance to the wider population. There was a

[37] In the population cancer is a (relatively) rare disease, both among smokers and non-smokers. One can then argue that the smokers/non-smokers odds ratio, i.e. 1350/7/(1296/61) ≈ 9, should be the same as in the general population. Odds ratios for rare events approximately equal the relative rates. It then follows that smokers in the general population were nine times as likely to get lung cancer as non-smokers.

formidable list of critics, including the statisticians Berkson (in a paper published in 1955) and Fisher (in a book published in 1959). Fisher (himself a smoker) proposed that there was a genotype that led both to a predisposition to smoke and to a predisposition to lung cancer.

Other results were soon obtained that pointed in the same direction. Doctors, impressed by the evidence on the health effects of smoking, gave up smoking in large numbers. Between 1951 and 1965, about half of the doctors who used to smoke gave up. Rates of lung cancer among UK doctors dropped from 60 to 37 per 100,000 between a 1953-57 study and a 1962-65 study. In the general population the rates increased slightly, from 113 per 100,000 to 120 per 100,000.

To refute Fisher's genetic hypothesis, Finnish epidemiologists (Kaprio and Koskenvuo 1989) studied smoking-discordant monozygotic twins, i.e. twins where one smoked and the other did not. Limiting attention to the 22 instances where one twin had died, the smoking twin died first in 17 cases out of 22. Looking now at cause of death, here is what they found:

	Smokers	Non-smokers
Coronary Heart Disease	9	0
Lung cancer	2	0
Other causes	6	5

These results make it hard to maintain the genetic hypothesis.

The strength of the case against smoking comes from its coherence. Many different sources of evidence, including laboratory studies that have directly demonstrated that tar in tobacco is a carcinogen, all point in the same direction. They are coherent in the following ways (Freedman 1999; Evans 1993, pp. 186ff):

1. There is a dose-response relationship, with the risk of disease greatest for heavy smokers.
2. The risk increases with the duration, i.e. with the number of years that a person smokes.
3. Ex-smokers have a risk that moves closer to the risk for non-smokers as the time from quitting increases.

Assessing the Performance of Surgeons

Heart surgery is undoubtedly effective. However a good outcome depends on high levels of skill, both from the surgeon and from supporting staff. It is very unusual for data on the success rates of such operations to become public. That is however exactly what happened in New York state in 1991. A newspaper used the Freedom of Information Law to gain access to the 1989 data, broken down by surgeons. There were surprising outcomes.

Chassin et al. (1996) give summary data for coronary artery bypass grafting (CABG). The data showed that the low-volume heart surgeons, those doing less than one operation a week, had higher than expected mortality rate. Here are two sets of figures:

Risk-Adjusted Mortality Rates for CABG: 1990-1992

27 low volume surgeons (1990 until contract terminated)	11.9% [of ~18 thousand patients]
All New York State	2.9% (4.2% in 1989; 2.5% in 1992)

One hospital had unusually high death rates for one specific groups of patients, those requiring CABG on an emergency basis. Here are the figures:

Risk-Adjusted Mortality Rates for Emergency CABG Patients

St Peter's Hospital, Albany	27%
All New York State	7%

Inevitably, there was an investigation into the source of the problem. It turned out that steps taken to stabilise patients before surgery were inadequate. From 7 deaths from 42 patients in 1992 the hospital went to no deaths from 54 patients in 1993. There are several points:

1. One can only make such comparisons if there are very large numbers. This study was effective because it could compare one hospital with hospitals as a whole, and an individual surgeon or groups of surgeons with surgeons as a whole. It is terribly important to bring all the data together, and to study it in context.
2. If one pools data from low-volume surgeons, there are enough data to make useful comments. Only where individual low-volume surgeons had an exceptionally high mortality rates could one argue that an individual surgeon was not performing well. Some of the low volume surgeons might actually have been quite good.
3. The true long-term figure, for an individual surgeon who performed 50 operations over the 3-year period and had a 6% mortality rate, might be anywhere between 1.4% and 22%. We have very little idea, with such scant data, as to how good that surgeon really is.
   [I have given a 99% confidence interval.]
4. It was essential to make adjustments to allow for the higher number of high risk patients operated on by some surgeons and in some hospitals. Use of the figures without such adjustment would have been an abuse of statistics.

These sorts of comparative figures are open to serious abuse. If two surgeons have each performed 200 operations, one with 5 deaths and the other with 10 deaths, it would be wrong to try to make anything of the difference. Some reporters focused on just these sorts of differences when the data were first reported. The New York State Department of Health started a program to educate reporters on how to interpret the figures, leading to huge improvements in reporting standards.

Example – Meta-analyses of Trials Studying the Link Between Salt and Blood Pressure

Law et al. (1991) is a paper in three parts. The first paper examines observational data from 24 communities – twelve from undeveloped countries and twelve from developed countries. They developed equations that predict, for a subject of a given age who makes some given change in salt consumption, the change in blood pressure. They then use these results to predict expected changes, by age, in 14 sets of within population data. They `correct' the within population data for random error in the sodium (salt) measurements. The correction is needed because the random error is uncomfortably close to the total population variation. With this correction, they claim that predictions from the between population study match the within population results well enough.

Finally, they use the between population results to predict changes from low to high salt diet in 78 sets of clinical trial data. Again, they claim fair agreement.

I have a number of comments:

1. Between population studies are, as I have argued repeatedly, susceptible to serious biases. While it may be reasonable to examine whether they can be made to agree with within population studies, the use of results from such studies to interpret data from clinical trials is surely the use of a highly suspect instrument to check out a much more trustworthy instrument.
2. It is not clear how the 14 within population studies were chosen.
3. About 60% of the numbers in the within population studies were derived from one large and rigorous Scottish study. On its own, the Scottish study showed no effect when confounding factors were taken into account. Law et al. add another 13 studies, almost certainly less rigorous, and claim to find an effect. They did not adjust for confounding factors. So their result may reflect the effects of confounding factors that the Scottish study adjusted for.
4. There is evidence (Easterbrook et al. 1991) that small studies are less likely than large studies to find their way through to publication. Thus the inclusion of all these smaller studies (one with as few as 58 subjects) has the potential to increase publication bias.
5. Law et al. combine data from different types of trials, from crossover trials as well as from randomised controlled trials. Even when the order of treatments in a crossover trial is randomised, interactions between the successive treatments (here the high and low sodium diets), or differential interactions between the washout period and the immediately following treatment, have the potential to give spurious effects. This may explain why, when the difference between high and low sodium is calculated separately for the two types of trial, the crossover trials gave a difference that was about twice that for the randomised controlled trials.
6. Law et al. do not seem to have looked in detail at quality issues. It is unclear whether the crossover trials were all conducted to similar standards of care, and indeed unclear whether the order of the diets was randomised. There may well be more than two types of trial.
7. Law et al.'s clinical trial data combined 21 trials for subjects with normal blood pressure with 57 trials for subjects with high blood pressure. It assumes that blood pressure increases linearly with salt intake. Their interpretations of the data rely strongly on this linearity assumption, which may well be wrong.

Clearly Law et al.'s analysis raises many different issues. Their approach to the analysis of this data, and their lumping together of disparate studies, surely introduces more confusion than light.

14.6 Measuring Instrument and Study Type Issues

Diet surely is an influence in the onset of some diseases. Attempts to nail the connection down precisely seem fraught with difficulty, at least for studying the effect on incidence of relatively rare diseases. Thus dietary fat is thought to be important. A standard approach has been to get people to write down details of what they eat. Unfortunately, such records are liable to be notoriously inaccurate. There are methods that seem more accurate, but they are too expensive for use in the large studies needed to investigate any link with such relatively rare diseases as breast cancer.

Diet and Breast Cancer – Nutrient Fat Measurement

For some time, it has been suspected that nutrient fat intake promotes breast cancer. Women who have a higher fat intake may be more likely to develop breast cancer. Here is a summary of the evidence, for and against:

- Direct manipulation of diets shows this effect in animal studies
- When breast cancer rates are compared between different countries, there is a clear correlation between nutrient fat intake and breast cancer rates
- Breast cancer rates, which are low (though rising) among Japanese women in Japan, increase over a period of time to U.S. rates when they move to the U.S. This is thought to be a result of their increased fat intake.
- Meta-analysis of case-control studies shows a nutrient fat intake effect on breast cancer.
- (Against) There have in addition been prospective cohort studies that maintain records both of dietary behaviour and of breast cancer history. None of these studies has ever found a statistically significant effect of nutrient fat intake on breast cancer.

None of the evidence for a link between nutrient fat and breast cancer is conclusive. The animal results do not necessarily carry across to humans. Countries differ in many ways, not just in nutrient fat intake. Case-control studies are susceptible to a variety of forms of bias. Nevertheless the confluence of these different sources of evidence does create a prima facie case for a link.

The problem with the prospective studies is that fat intake is hard to measure. With minor exceptions, every large prospective study has used a food frequency questionnaire (FFQ). These have large and acknowledged errors. Are the errors so large, or of such a nature, that it is altogether to be expected that the prospective studies will fail to show an effect?

There are two keenly contested points of view:

1. The confluence of the different sources of evidence does create a prima facie case for a positive association between fat intake and breast cancer, at least to the extent that the issue warrants a search for more conclusive evidence. The evidence from prospective studies should be discounted because their measuring instrument (FFQ) is unreliable.
2. The evidence from prospective studies is compelling. It is consistent over different studies. While the FFQ has problems, they are not large enough to explain the consistency of the failure to find an association.

The search for more conclusive evidence has led to the huge expensive randomised clinical trial, extending over ten years, that is being conducted as part of a National Institutes of Health WHI (Women's Health Initiative) study in the U.S.. Women in the `treated' group will be counselled to undertake a healthy diet, involving as one component a large reduction (from 35% to 20% or less) in the proportion of calories coming from fat. We will have to wait ten years or more for the answers.

Of more immediate interest is work that examines sources of error in the use of FFQ information to estimate nutrient fat intake. Everyone acknowledges that there are large errors in the estimates of fat intake. The effect of these errors is to flatten the regression line, and make it more difficult to find statistical significance. The effect is not strong enough, if one assumes that errors for any individual average out over time, to explain the consistently negative results from prospective trials.

Suppose however that there are individual-specific biases. Some people consistently over-report their fat intake, while others consistently under-estimate it. The model is

FFQ = fat intake + person-specific bias + measurement error

We now have a model where there are two sources of error. One is individual-specific and will be assumed to vary randomly between individuals, while the other will be a source of occasion to occasion variation within individuals.

None of the available data allow estimation of a person-specific bias. The effect of the person-specific bias is to further flatten the regression slope. Kipnis et al. (1999) have investigated the magnitude of the effect, for a range of possible values of the person-specific bias. It is at least possible that this explains the negative results from prospective studies.

Perhaps the most important point that emerges is that it is impossible to know too much about one's measuring instrument. Apparently harmless assumptions can have large consequences. Investigations are now under way that will collect data that should allow estimation of the distribution of the person-specific bias.

There are better alternatives to the FFQ. They are however so expensive that it is not feasible to use them in the very large numbers of participants that are required for prospective studies. For example, breast cancer rates in Australian women aged 50-60 are of the order of a few per thousand. Such studies require, at a minimum, some hundreds of thousands of participants.

Investigations into the characteristics of the FFQ can get useful results from relatively small numbers of subjects, of the order of a hundred or two. In such studies it is possible to use more expensive alternatives to the FFQ as benchmarks against which to compare FFQ estimates.

References and Further Reading

Advice and Criticism Directed at Specific Application Areas

Andersen, Bjorn 1990. Methodological errors in medical research : an incomplete catalogue. Blackwell Scientific.

Chanter, D. O. 1981. The use and misuse of regression methods in crop modelling. In: Mathematics & Plant Physiology, ed. D. A. Rose & D. A. Charles-Edwards. Academic Press.

Chassin, M.R.; Hannan, E.L.; DeBuono, B.A. 1996. Benefits and hazards of reporting medical outcomes publicly. New England Journal of Medicine 334: 394-398.

Gardner, M. J.; Altman, D. G.; Jones, D. R.; Machin, D. 1983. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? British Medical Journal 286: 1485-1488.

Gates, C.E. 1991. A user's guide to misanalyzing planned experiments. Hortscience 26: 1262 - 1265.

Johnson, T. 1998. Clinical trials in psychiatry: background and statistical perspective. Statistical Methods in Medical Research 7: 209-234.

Maindonald, J. H. and Cox, N. R. 1984. Use of statistical evidence in some recent issues of DSIR agricultural journals. New Zealand Journal of Agricultural Research 27: 597-610.

Padak, P. M. 1989. Inconsistencies in the use of statistics in horticultural research. Hortscience 24: 415.

Thomas, D. H. 1978. The awful truth about statistics in archaeology. American Antiquity 43: 231-244.

Towers, N. R., Gravett, I., Smith, J. F., Smeaton, D. C., Knight, T. W. 1983. Guidelines for production response trials. In Grace, N. D. (ed.) *The Mineral Requirement of Grazing Ruminants*. New Zealand Society of Animal Production, Hamilton.

Application Area Papers

Blanpied, G. D., Bramlage, W. J., Dewey, D. H., LaBelle, R. L., Massey, L. M. Jr., Mattus, G. E., Stiles, W. C. and Watada, A. E. 1978. A standardised method for collecting apple pressure test data. New York's Food and Life Sciences Bulletin 74: 1-8.

Bongers, A. J. 1992. Comparison of three penetrometers used to evaluate apple firmness. Washington State Tree Fruit Postharvest Journal 3: 7-9.

Cohen, P. 1996. Pain discriminates between the sexes. New Scientist, 2 November, p. 16.

Davenas, E. et al. 1988. Human basophil degranulation triggered by very dilute antiserum against IgE. Nature 333: 816-818. See also 333: 787 and 334: 285-286.

Gihr and Pilleri 1969. Anatomy and biometry of Stenella and Delphinus. In Pilleri, G., ed.: Investigations on Cetacea. Berne, Switzerland.

Gordon, N. C., Gear, R. W., Heller, P.H., Paul, S., Miaskowski, C. and Levine, J. D. 1995. Enhancement of Morphine Analgesia by the GABAB against Baclofen. Neuroscience 69: 345-349.

Harker, F. R., Maindonald, J. H. & Jackson, P. J. 1996. Penetrometer measurements of apple and kiwifruit texture: operator and instrument differences. Journal of the American Society for Horticultural Science 121: 927-936.

Jessup, A. J. and Baheer, A. 1990. Low-temperature storage as a quarantine treatment for kiwifruit infested with Dacus tryoni (Diptera: Tephritidae). Journal of Economic Entomology 83: 2317-2319.

Law, M. R., Frost, C. D., and Wald, N. J. 1991. By how much does dietary sodium lower blood pressure? I – Analysis of observational data among populations; II – Analysis of observational data between populations; III – Analysis of data from trials of salt reduction. British Medical Journal 302: 811-824.

Maddox, J., Randi, J. and Stewart, W. W. 1988. "High-dilution" experiments a delusion. Nature 334: 287-290. [See also Benveniste's response in the same volume on p. 291.]

Nicholls, N., Lavery, B., Frederiksen, C. and Drosdowsky, W. 1996. Recent apparent changes in relationships between the El Ni™o – southern oscillation and Australian rainfall and temperature. Geophysical Research Letters 23: 3357-3360.

Oldham, P. D. and Newell, D. J. 1977. Fluoridation of water supplies and cancer – a possible association. Applied Statistics 16: 125-135.

Reid, J. B. and Petrie 1991. Effects of soil aeration on root demography in kiwifruit. New Zealand Journal of Crop and Horticultural Science 19: 423-431.

Thomas, D. B. and Mangan, R. L. 1997. Modelling thermal death in the Mexican fruit fly (Diptera: Tephritidae). Journal of Economic Entomology 90: 527-534.

Yiamouyiannis, J. and Burk, D. 1977. Fluoridation and cancer. Age-dependence of cancer mortality related to artificial fluoridation. Fluoride 10: 102-125.

Dietary Measurement Models

Kipnis, V., Carroll, R.J., Freedman, L.S. and Li Li 1999. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. American Journal of Epidemiology 150: 642-651.

Fickle Mice

Crabbe, J.C., Wahlsten, D., and Dudek, B.C. 1999. Genetics of mouse behaviour: interactions with laboratory environment. Science 284: 1670-1672.

Enserink, M. 1999. Fickle mice highlight test problems. Science 284: 1599-1600.

Studies with Historical Interest (including the Smoking Controversy)

Evans, A. S. 1993. Causation and Disease: A Chronological Journey. Plenum, New York.

Freedman, D. 1999. From association to causation: some remarks on the history of statistics. Statistical Science 14: 243-258.

Kaprio, J. and Koskenvuo, M. 1989. Twins, smoking and mortality: a twelve-year prospective study of smoking-discordant twin pairs. Social Science and Medicine 29: 1083-1089.

Pearl, R. 1938. Tobacco smoking and longevity. Science 87: 216.

Systematic Reviews

Cochran Injuries Group Albumin Reviewers 1998. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. British Medical Journal 317: 235-240.

Kleijnen J., ter Riet, G. and Knipschild, P. 1992. Vitamin C and the common cold; review of a megadose of literature. In: Kleijnen J. Food supplements and their efficacy (dissertation). University of Limburg, Maastricht.

Sacks, F.M., Svetkey, L.P., Vollmer, W.M., Appel, L.J., Bray, G.A., Harsha, D., Obarzenek, E., Conlin, P.R., Miller, E.R., Simons-Morton, D.G., Karanja, N., and Lin, P.-H. 2001. Effects of blood pressure on reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. New England Journal of Medicine 344: 3-10.

Taubes, G. 1998. The (political) science of salt. Science 281: 898-907 (14 August).

The Science of Systematic Review

Chalmers, I. and Altman, D. G., eds. 1995. Systematic Reviews. BMJ, London.

Easterbrook, P. J., Berlin, J. A., Gopalan, R. and Matthews, D. R. 1991. Publication bias in clinical research. Lancet 337: 867-872.

Greenhalgh, Trisha 1997. How to read a paper: the basics of evidence-based medicine. BMJ, London.

Oxman, A. D. and Güyatt, G. H. 1983. The science of reviewing research. Annals of the New York Academy of Sciences 703: 125-131.

Sackett, D. L. 1983. Proposals for the Health Sciences. I. Compulsory retirement for experts. Journal of Chronic Diseases 36: 545-547.

Sackett, D. L., Richardson, W. S., Rosenberg, W. M. C. and Haynes, R. B. 1997. Evidence-Based Medicine. Churchill Livingstone, New York.

15. The Research Process

> The map of the material world, including human mental activity, can be thought a sprinkling of charted terrain separated by blank expanses that are of unknown extent yet accessible to coherent interdisciplinary research.
> [E.O. Wilson, 1998, p. 299.]

> *The scientific community is not good at looking critically at its own processes. Few journals conduct regular critical evaluations of all papers that have appeared over a period of a year or two. Even fewer publish the results of such review. Publication or archiving of data should ordinarily be mandatory, allowing other workers to impose their own checks on the analysis and/or on the data. This would ensure access to the data for use in planning or in overview studies. Few journals have clear and adequate standards for the reporting of data summaries, so that the completeness of information in published papers varies widely, even within the one journal. There are particular inadequacies when research demands skills that are outside of the primary areas of expertise that the journal represents. In all these respects, it is reasonable to expect eventual change and improvement. Many different forces are driving change, not all of them benign.*

Research is paid for by public organizations, by business and industry, and rarely by individuals. All these demand varying degrees of control over the research they fund. Funding bodies will continue to look for ways, not always well-conceived, to get better value from the research dollar. We should have no doubt that there will be change, particularly in publicly funded research, and not all for the better. The research community has been slow to initiate, from within its own ranks, changes that would genuinely improve the research process. I have hinted at some desirable changes earlier in these notes.
An examination of changes of the past two decades gives clues on how research demands may change in the next two decades. We should expect future changes of a similar or larger magnitude. Some changes will arise from the attempt to fix problems with our current approaches. Some will be driven by technological advance. Some will be demand driven.

Publication Pressures

One direction of change in the past two decades ought to be reversed. Pressures to publish fill the literature with increasing quantities of trifling verbiage. Disturbingly often, this material is buttressed by shoddy statistical analysis[38]. The rewards too often go to those who pay as much attention to publishing the inconsequential as to results of substance. This places a huge burden of discrimination on the reader. There are particular problems when authors stray into areas that are outside of their specialist discipline or disciplines.

[38] See Andersen (1990) for a wide-ranging account of problems in the medical literature.

Openings for Improvement

Notwithstanding these various pressures, the standards of the best research are improving all the time. Some of the most interesting innovations that affect the use of scientific data have been in clinical epidemiology and clinical medicine.

A good starting point is to examine changes in those areas (such as clinical epidemiology) where in the past twenty years there have been large advances in approaches to the design of data collection and data analysis, then consider the implications of these changes for other research. These trends can then be extrapolated a limited distance into the future. This is the motivation for the following list. We might expect:

1. Requirements to place data and supporting documentation in the public domain, inviting scrutiny and facilitating incorporation, where this is pertinent, into overview studies.
2. Identification of skill gaps that compromise research that otherwise demonstrates high levels of technical skill.
3. Development in other areas of mandatory reporting standards comparable to those for randomized controlled trials that are set out in the CONSORT statement (Begg et al. 1996).
4. In medicine, health and education, the development of mechanisms that will replace multiple small trials by large carefully co-ordinated multi-centre trials.
5. The establishment of international registers for major studies in specific areas, making it easier for anyone who is conducting an overview study to identify relevant studies.
6. Except where they break radically new ground or where experimental approaches are impossible for ethical or other practical reasons, there may in clinical medicine and related areas be an increased reluctance to fund non-experimental studies.
7. Insistence, where appropriate, that researchers use qualitative and quantitative approaches to complement each other.

There is urgent need for items 1-3. Items 1 and 2 could be implemented without making any substantial change to refereeing processes. There are existing models for both these steps. Moves to collect data into data bases that operate as commercial entities (Transborder 1998) may to a greater or lesser extent work in the other direction to item 1, restricting access to data. Clinical medicine has made limited progress with items 4-7.

Attention to potential skill gaps in ancillary disciplines is more than ever important as we respond to the seductions of new technology — molecular biology, informatics, machine learning, data mining, and so on. The use of a new technology, or of an old technology under a new name, should not be a new opening to dispense with the complementary disciplines needed for an effective study.

Insights from Evidence-Based Medicine

Evidence-Based Medicine aims to base clinical practice, as far as possible, on the best research evidence. Its methods and insights have far-reaching implications for research, both in medicine and in other areas. They are a good point of departure for discussing how research approaches and research training ought to change. They provide a perspective from which to assess the contribution that new computing and other technologies may be able to make to research. A key issue for Evidence-Based Medicine is to distinguish the data and associated interpretation that merit attention from what is valueless or of substantially inferior worth.

Three key developments have been:

1. the Cochrane Collaboration (Sackett et al. 1994), an international network that facilitates the conduct of systematic overview studies that provide expert assessments of the available evidence on particular medical issues;
2. Initiatives (Sackett et al. 1997) that aim to identify established research results, make them accessible to clinicians, and get clinicians using them. Guidance on ways of getting quickly to the most useful and relevant evidence has been an important thrust;
3. the Consort statement (Begg et al., 1996) that sets standards for the reporting of clinical trials.

Clinicians do not have time to wade through and assess a mountain of papers, in order to decide on the best treatment for each and every condition. Nor do they have the skills needed to decide between rival claims. They require a reliable and up to date research consensus, expertly done. The Cochrane Collaboration, and other similar initiatives, provides such a consensus.

Other areas of science might with advantage take on board the approaches and insights of the Cochrane Collaboration studies, and of evidence-based medicine. The researcher's need to use an assessment of existing knowledge as a point of departure for new research is not greatly different from the clinician's demand for a research assessment on which he/she can base clinical decisions. While the researcher may finally need to wade through some part of the mountain of paper, it is a huge help to have guidance on what is there.

Unless authors explicitly report all relevant details of their procedures, it is difficult or impossible to judge the quality of the work, to assess its specific contribution to the total body of evidence, and to assess the relevance of the research for clinical practice. This is the motivation for the attention that the Consort statement (point 3 above) gives to reporting standards. Again the point applies with equal force to other areas of science.

Cost-Benefit Analysis

Funding agencies increasingly demand cost-benefit or other economic analyses. In principle this is useful. A serious deficiency is that these analyses may focus on the costs of a narrow group of stakeholders, ignoring spill-over environmental or social effects that do not incur an immediate financial cost. There are typically major benefits and losses that are not costed! Assumptions may be simplistic. It is important to keep assessments of medical benefit separate from economic assessments, allowing other investigators to vary the cost structure.

The Relevance of Information Technologies

There should be demands to gather better data, to gather more comprehensive data, to organize data better, and to make better use of the data we then have. While computing has from its beginnings carried with it the promise to address these concerns, computing technologies are not the most appropriate place to start in trying to address such demands. Changes should be driven by the demands of the research process itself.

Areas where information technologies seem relevant are:

1. There have been huge advances in the methodology for data analysis, taking advantage of advances in computing hardware and software. Statistical packages differ greatly in the extent to which they have taken up these methodological advances.

2. There has been a large emphasis on methods for *Exploratory Data Analysis*. There is an overlap with the methods and approaches of Data Mining.
3. There have been substantial advances, again taking advantage of the increased power of computing systems, in the methodology for analysing data from overview studies.
4. Database technology, already a powerful tool for storing and accessing data, has extended to the networking of physically separate databases. There is a new emphasis on the key role of data aggregation for the scientific enterprise. Statisticians, while rightly emphasizing the hazards of making inferences from data that are from disparate sources, have been slow to take this new emphasis on board.
5. New data analysis challenges arise from the sheer size of some databases.
6. The computer science perspective seems appropriate for attempts at automating the task of making data based inferences. To date, these attempts have had very limited success.

There has been a large growth in methodologies that require a pooling of the skills of computer scientists, statisticians and subject area specialists. It is important that the demands of the research process should drive the use of computer technology, that the research process should not be driven by technology. There is further discussion of these issues in Maindonald (revised 2000).

Access to Data

A serious impediment to scrutiny or further use of published results is that the data are rarely readily accessible. It may turn out, when a request is made, that the data have been lost or misplaced. Requests for access to data may not be well received, in some instances because authors do not want to risk exposing their own analyses to scrutiny. One author that I contacted agreed to make the data available on the condition that they were not used to reach a conclusion different from that in the published paper! Unless there are strong ethical or privacy reasons, all data that are the basis of published results ought to be archived and placed in the public domain. Placing data of suitable interest and quality in the public domain should of itself count as a publication. This is needed for reasons that, in total, are surely compelling:

1. It allows various forms of post-publication scrutiny of results.
2. The skills needed for undertaking the research that generated the data are different from those needed for design of data collection and for data analysis.
3. Separation of the analysis task from the main research task that generated the data would often result in more effective eventual use and interpretation of the data.
4. If data are of sufficient value, the work that led to their collection should be rewarded, independently of any attempt at analysis and interpretation.
5. The data are then available to other scientists who may want to use them as a basis for planning their own studies.
6. The data are available for inclusion in overview studies.
7. Fraud is then harder to hide. In investigations into fraud (e.g. Hagmann 2000) a first step may be to ask for the original data.

Published analyses should often be treated as preliminary assessments, pending more definitive analysis.

Inter-disciplinary Research

Often large advances take place at the boundaries between disciplines. Physicists who had moved into biology did most of the early work on bacteriophages. Crick, the co-discoverer of the structure of DNA, was trained as a physicist. Crick and Watson relied heavily on the X-ray data of the chemists, Rosalind Franklin and Maurice Wilkins (Drlica 1994.)

The funding of work that spreads across current disciplines might therefore seem a priority. Present funding systems have difficulty with inter-disciplinary research. While the problem is widely recognised, little seems to be done by way of remedy. The following comment is from a U.S. report on the mathematical sciences (Senior Assessment Panel 1998). It notes that interactions between mathematicians and other groups are often "obscured by the inward focus of mathematics and science departments". It goes on to argue:

> **The structure of universities mitigates against interdisciplinary research.** While the above finding criticizes mathematical scientists for not collaborating more actively with other scientists and engineers, part of the fault lies with the organization and culture of universities, here and abroad, which restrains collaboration across scientific boundaries. The academic award system does not encourage collaboration; in fact, individuals who straddle fields reduce their chances of tenure. …

Evaluation of the Research Process

A good principle is that, until they have been subjected to peer review, scientific claims should be treated with extreme caution. Forms of commercial secrecy that interfere with this scrutiny can readily become cloaks for incompetence. The peer review process does not however guarantee quality. The quality of the review process varies greatly from one area to another and from one journal to another. Even after peer review, scientific claims must be closely scrutinised. The process requires much better evaluation than is currently common. Review of the papers that have appeared in one or other journal over the course of a year or two seems unusual. Why? Maintenance of quality, over all the skill areas that are relevant to papers that appear, is surely desirable if journals are to serve their presumed primary function as repositories of the results of research. There should be checks on rejections as well as on acceptances.

Journals have acquired another function that has partly displaced their primary function. They have become a vehicle by which researchers can demonstrate their academic worth. Review and maintenance of quality, over all skill areas relevant to papers that appear, is likewise necessary if they are to perform this function credibly.

References and Further Reading

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639.

Chalmers, I. and Altman, D. G., eds. 1995. Systematic Reviews. BMJ Publishing Group, London.

Drlica, K.A. 1994. Double-Edged Sword. The Promises and Risks of the Genetic Revolution. Addison-Wesley, Reading MA.

Hagmann, M. 2000. Panel finds scores of suspect papers in German fraud probe. Science 288: 2106-2107.

Maindonald, J. H. revised 2000. New Approaches to Using Scientific Data – Statistics, Data Mining & Related Technologies in Research & Research Training, v + 39pp. Graduate School Occasional Publications GS98/2.

Sackett, D. L. and Oxman, A. D., eds. 1994. The Cochrane Collaboration Handbook. Cochrane Collaboration, Oxford.

Sackett, D. L., Richardson, W. S., Rosenburg, W. and Haynes, R. B. 1997. Evidence-based medicine. How to Practice and Teach EBM. Churchill Livingstone, New York. See also the web site http://cebm.jr2.ox.ac.uk/

Senior Assessment Panel 1998. Report on the senior assessment panel of the international assessment of the U. S. mathematical sciences. Available from http://www.nsf.gov/pubs/1998/nsf9895/start.htm

Transborder 1997. Bits of Power. Issues in Global Access to Scientific Data/Committee on Issues in the Transborder Flow of Scientific Data, U.S. National Committee for CODATA, Commission on Physical Sciences, Mathematics, and Applications, National Research Council. National Academy Press, Washington D. C. Available from http://www.nap.edu/readingroom/books/BitsOfPower/index.html/

Appendix I: Checklist for Use with Published Papers

Aims and Purpose

1. Do the authors explain their scientific reasons for undertaking the study?
2. Is there a clear statement of what they aimed to achieve?
3. Did the authors review current knowledge, before embarking on their study?

Data Collection

4. How were the data obtained? Some of the possibilities are Sample, Experiment, Informed opinion, Guess.
   [How many tens of thousands of people did the papers say marched across Sydney Harbour Bridge? Did someone count them all? Was the number a stab in the dark?]
5. Do the data make sense; are they free of apparent serious anomalies?
   [Some numbers may be impossible? Or, e.g., a height/weight ratio may be impossible.]
6. Do any of the claims go beyond what the data could support?
7. Do the data answer the research question?
8. Are the measurements/questions clear? Or is there ambiguity?
   [e.g. using data from a limited local study to support claims that relate to another geographical location.]
9. Are the data valid for the intended use?
10. In a study of human subjects, who had contact with the participants and how?
11. Who/what was studied and what was the selection process?
12. Are the data sampled from the population to which the researchers wish to generalise?
    [A sample of Sydney-siders is not a good basis for generalising to what Canberra residents think.]
13. Was the study capable of detecting effects of a magnitude that were of interest?
    [Influences on precision include measurement instruments, experimental or sampling design, and sample size.]
14. What biases may have been present in the data?
    [Consider, measuring instrument bias, observer bias, selection bias, etc.]
15. Where groups are compared are there extraneous differences?
    [e.g., in clinical trials, differences that have nothing to do with the treatment.]

Data Analysis

16. Is the arithmetic correct?
17. Does the analysis take account of data structure (fixed effects, random effects, clustering, etc.)
18. Is the description of the method of analysis clear and complete, with a reference given if the methodology is at all non-standard?
19. Has account been taken of clear grouping (e.g. males/females, different species, etc.) in the data? If results were combined across groups, is justification given?
20. Is statistical significance distinguished from practical significance?
21. Do the authors present graphs or tables that allow the reader to assess agreement with the assumed model?

Interpretation and Presentation

22. Do the authors give a clear statement of what they claim to have achieved?
23. Do the data support the claims that are made?
24. Do the authors distinguish substantial effects from effects that, even if perhaps statistically significant, are insubstantial?
    [Large studies may detect effects that are of little practical consequence.]
25. Do authors seem to rely uncritically on the claims of other authors?
26. Are the interpretations plausible? Do the data support them? Do the data rule out other interpretations?

Appendices III and IV have more detailed checklists, with greater attention to technical statistical issues, for use in evaluating the statistical presentation in published papers. See also the checklists in Greenhalgh (1997). References appear at the end of Appendix III.

Appendix II – A Checklist for Authors

This is primarily directed to the writing of reports and theses. Most of it is also relevant to the writing of scientific papers. Note however that each journal has its own style, which papers published in that journal need to follow.
Here is the checklist:

1. *Did you begin with a brief intelligible summary that gives the main conclusions?*
2. Have you given a clear description of the research question?
3. *Have you given clear information on the technical background that explains why the project was needed, gives technical information that will help understand your report, and places your report in context?*
4. Have you given a clear description of the design of data collection, and of special difficulties that arose in implementing the design?
5. Have you given a brief clear explanation of your methods of analysis?
6. Are your statistical analyses appropriate? Are they correct? Are they reasonably complete?
7. *Do you highlight the main points that emerge from your analyses? Do detailed technical information and the details of computer output, where these seem necessary, appear in an appendix?*
8. *Is your discussion of results clear, critical and incisive? Do you focus on the key issues?*
9. *Have you used clear and appropriate forms of graphical and tabular presentation? Is all the material that you include pertinent?*
10. Have you included references that will assist readers who want more information on technical background and methods of analysis?
11. Have you used a consistent style (e.g. the Harvard style) for all references?
12. Have you addressed potential challenges to the interpretation of results, including challenges that may arise from inadequacies in the design of data collection?
13. *Is the layout and general presentation attractive? Consider page margins, headings, line and other spacing, type fonts, graphs, division into sections and paragraphs.*

Points that will quickly attract the casual reader's attention appear in italics. In a report for a commercial client, these will often be the main focus of attention. They may become important to a commercial client (and to the report writer) when claims made in the report are challenged, when the report goes to other consultants for review, or when other specialists make use of information in the report.
Other points relate more directly to statistical or other professional concerns. They are intrinsic to doing a thoroughly professional job. In a research thesis these are likely to be the major focus of attention.

Appendix III – Checklist for Presentation of Statistical Results.

This checklist may be useful both to authors and to referees

1. Is the objective (purpose) of the study sufficiently described?
2. Is an appropriate study design used, having this objective in view?
3. Is the study design adequately described? If an experiment, is it clear
   - **i** how the experiment was laid out?
   - **ii** what were the experimental units, and what measurements were made or samples taken within experimental units?
   - **iii** how treatments were assigned to experimental units?
   - **iv** what sources of variability were represented – different error strata etc?

4. Is all information given that is relevant to analysing or assessing results?
   - i Is the standard error of mean or of difference or of other statistics given when appropriate?
   - **ii** Are standard errors or other measures of variability based on the appropriate source of variation?
   - **iii** Where standard errors are not available or not appropriate, are there other indications of precision?
   - **iv** Are results presented to an appropriate numerical accuracy? (Thus means should be given to around 10% of the SEM.)
5. Were there sufficient replicates to give the precision that was desirable?
6. Were trend or response surface methods used when the data required this?
7. Do the statistical analyses connect closely to points that are of scientific interest?
8. Are the statistical methods used appropriate?
9. Are there statements describing or referencing all statistical tests or estimation methods?
10. Does it seem that the validity of the statistical methods – e.g. homogeneity of variance or the form of response curves – has been adequately checked?
11. From your examination of (i) text, (ii) tables and (iii) figures determine
    - **i** Is there an adequate overview of the data?
    - **ii** Is the focus on effects that are substantial and of major interest?
    - **iii** Is the presentation of statistical material clear?
12. Is an appropriate/correct conclusion drawn from the statistical analysis?
13. Are results translated, as far as possible, into subject matter terms?
14. Do graphs convey information tersely and clearly, avoiding irrelevant and/or distracting features?
    - **i** Are graphs adequately labelled?
    - **ii** If there are multiple standard error bars, are they all necessary? (But take care that when there clearly are standard errors that are very different, this is reflected by the use of the requisite number of error bars.)
15. Is assistance with the design and/or statistical analysis and/or interpretation acknowledged by
    - **i** authorship?
    - **ii** acknowledged help?
16. From the statistical viewpoint is the paper of acceptable standard to be published?
17. Comment on any points not covered by the above questions.

[Adapted from the checklist on page 1486 of Gardner, Altman, Jones and Machin (1983).]

References and Further Reading (Appendices I, II and III)

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639.
[The checklist that appeared as part of this statement can be found at: http://www.ama-assn.org/public/journals/jama/jlist.htm]

Cleveland, W. S. 1993. Visualizing Data. Hobart Press, Summit, New Jersey.

Gardner, M. J.; Altman, D. G.; Jones, D. R.; Machin, D. 1983. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? British Medical Journal 286: 1485-1488.

Greenhalgh, Trisha 1997. How to read a paper: the basics of evidence-based medicine. BMJ, London.

Maindonald, J.H. 1992. Statistical Design, analysis and presentation issues, New Zealand Journal of Agricultural Research 35: 121 - 141, 1992.

Murray, A.W.A. 1988. Recommendations of the editorial board on use of statistics in papers submitted to JSFA --- guidelines to authors as formulated by A W A Murray. Journal of the Science of Food and Agriculture 42, no. 1, following p. 94. Reprinted in vol. 61, no. 1, 1993.

Index