International Statistical Review (1999), **67**, 3, 223–265, Printed in Mexico

Statistical Thinking in Empirical Enquiry

C.J. Wild and M. Pfannkuch

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

Summary

This paper discusses the thought processes involved in statistical problem solving in the broad sense from problem formulation to conclusions. It draws on the literature and in-depth interviews with statistics students and practising statisticians aimed at uncovering their statistical reasoning processes. From these interviews, a four-dimensional framework has been identified for statistical thinking in empirical enquiry. It includes an investigative cycle, an interrogative cycle, types of thinking and dispositions. We have begun to characterise these processes through models that can be used as a basis for thinking tools or frameworks for the enhancement of problem-solving. Tools of this form would complement the mathematical models used in analysis and address areas of the process of statistical investigation that the mathematical models do not, particularly areas requiring the synthesis of problem-contextual and statistical understanding. The central element of published definitions of statistical thinking is "variation". We further discuss the role of variation in the statistical conception of real-world problems, including the search for causes.

Key words: Causation; Empirical investigation; Statistical thinking framework; Statisticians' experiences; Students' experiences; Thinking tools; Variation.

1 Introduction

> *"We all depend on models to interpret our everyday experiences. We interpret what we see in terms of mental models constructed on past experience and education. They are constructs that we use to understand the pattern of our experiences." David Bartholomew (1995).*

"All models are wrong, but some are useful" George Box

This paper abounds with models. We hope that some are useful!

This paper had its genesis in a clash of cultures. Chris Wild is a statistician. Like many other statisticians, he has made impassioned pleas for a wider view of statistics in which students learn "to think statistically" (Wild, 1994). Maxine Pfannkuch is a mathematics educator whose primary research interests are now in statistics education. Conception occurred when Maxine asked "What *is* statistical thinking?" It is not a question a statistician would ask. Statistical thinking is the touchstone at the core of the statistician's art. But, after a few vague generalities, Chris was reduced to stuttering.

The desire to imbue students with "statistical thinking" has led to the recent upsurge of interest in incorporating real investigations into statistics education. However, rather than being a precisely understood idea or set of ideas, the term "statistical thinking" is more like a mantra that evokes things understood at a vague, intuitive level, but largely unexamined. Statistical thinking is the statistical incarnation of "common sense". "We know it when we see it", or perhaps more truthfully, its *absence* is often glaringly obvious. And, for most of us, it has been much more a product of experience, war stories and intuition than it is of any formal instruction that we have been through.

There is a paucity of literature on statistical thinking. Moore (1997) presented the following list of the elements of statistical thinking, as approved by the Board of the American Statistical Association (ASA) in response to recommendations from the Joint Curriculum Committee of the ASA and the Mathematical Association of America: *the need for data; the importance of data production; the omnipresence of variability; the measuring and modelling of variability*. However, this is only a subset of what the statisticians we have talked to understand by "statistical thinking" or "thinking statistically". In the quality (or more properly, process and organisational improvement) area, much has been written, but addressing a specific audience. Snee (1990, p. 118) defined statistical thinking as "*thought processes, which recognise that variation is all around us and present in everything we do, all work is a series of interconnected processes, and identifying, characterising, quantifying, controlling, and reducing variation provide opportunities for improvement*". (See also Britz *et al.*, 1997; Mallows, 1998; and Dransfield et al. 1999).

The usual panacea for "teaching" students to think statistically is, with apologies to Marie-Antoinette, "let them do projects". Although this enables students to experience more of the breadth of statistical activity, experience is not enough. The cornerstone of teaching in any area is the development of a theoretical structure with which to make sense of experience, to learn from it and transfer insights to others. An extensive framework of statistical models has been developed to deal with technical aspects of the design and analysis that are applicable once the problem and variables have been defined and the basic study design has been decided. An enormous amount of statistical thinking must be done, however, before we ever reach this stage and in mapping between information in data and context knowledge throughout the whole statistical process. We have little in the way of scaffolding to support such thinking (see Mallows, 1998). Experience in the quality arena and research in education have shown that the thinking and problem solving performance of most people can be improved by suitable structured frameworks (Pea, 1987, p. 91; Resnick, 1989, p. 57).

The authors have begun trying to identify important elements from the rich complexity of statistical thinking. In addition to the literature and our own experience, our discussion draws upon intensive interviews with students of statistics and practising professional statisticians. One set of eleven students, referred to as "students" were individually given a variety of statistically based tasks ranging from textbook-type tasks to critiquing newspaper articles in two one hour sessions. They were interviewed while they solved the problems or reacted to the information. Another set of five students, referred to as "project students" were leaders of groups of students doing real projects in organisations which involved taking a vaguely indicated problem through the statistical enquiry cycle (see Fig. 1(a)) to a solution that could be used by the client. Each was interviewed for one hour about their project. The six professional statisticians were interviewed for ninety minutes about "statistical thinking" and projects they had been involved in. The "project students" and statisticians interviews were structured around the statistical enquiry cycle and were in the form of a conversation which reflected on their approach and thinking during the process of an investigation. This paper is not a report on this particular research (that is being done elsewhere, e.g. Pfannkuch, 1996, 1997), but an attempt to synthesise a more comprehensive picture from these interviews and the literature.

We are not concerned with finding some neat encapsulation of "statistical thinking". Our concerns are deeper than this. We are investigating the complex thought processes involved in solving real-world problems using statistics with a view to improving such problem solving. We are thus interested in developing a framework for thinking patterns involved in problem solving, strategies for problem solving, and the integration of statistical elements within the problem solving. We do not address the thinking involved in developing new statistical methodology and theory. We recognise that much statistical thinking can beneficially take place in day-to-day activities, particularly in the interpretation of information in media and other reports. In interpreting reports, we recognise the applicability of parts of our statistical knowledge about the production, behaviour and analysis of data to the type of information we are receiving and are thus able to critically appraise aspects of that information. The

type of thinking required is very similar, if not identical, to fragments of the thinking performed by someone involved in an enquiry. We see the enquiry cycle as providing a coherent structure that links the fragments and, thus, as an ideal place to start. In subsequent work, we have been specialising and further developing the ideas given here for interpreting statisticial information in reports.

This discussion is organised into a statistical thinking framework for empirical enquiry in Section 2. Section 3 explores "variation". It looks at statistical approaches to real-world problems from the starting point of omnipresent variation. Section 4 takes lessons learned in Section 2 and gives a fragment of a thinking tool for improving investigative skills. Section 5 contains a discussion.

2 A Framework for Statistical Thinking in Empirical Enquiry

Applied statistics is part of the information gathering and learning process which, in an ideal world, is undertaken to inform decisions and actions. With industry, medicine and many other sectors of society increasingly relying on data for decision making, statistics should be an integral part of the emerging information era. Statistical investigation is used to expand the body of "context" knowledge. Thus, the ultimate goal of statistical investigation is *learning* in the *context* sphere. Learning is much more than collecting information, it involves synthesising the new ideas and information with existing ideas and information into an improved understanding.

From the interviews we have built up the four dimensional framework shown in Fig. 1 which seeks to organise some of the elements of statistical thinking during data-based enquiry. The thinker operates in all four dimensions at once. For example the thinker could be categorised as currently being in the planning stage of the Investigative Cycle (Dimension 1), dealing with some aspect of variation in Dimension 2 (Types of Thinking) by criticising a tentative plan in Dimension 3 (Interrogative Cycle) driven by scepticism in Dimension 4 (Dispositions). Who is doing this thinking? Anyone involved in enquiry, either individually or as a member of a team. It is not peculiar to statisticians, although the quality of the thinking can be improved by gaining more statistical knowledge.

2.1 Dimension One: The Investigative Cycle

The first dimension in Fig. 1(a) concerns the way one acts and what one thinks about during the course of a statistical investigation. We have adapted the PPDAC model (Problem, Plan, Data, Analysis, Conclusions) of MacKay & Oldford (1994). The elements of this model should be self-explanatory to statisticians. The statisticians we interviewed were particularly interested in giving prominence to the early stages of PPDAC, namely, to grasping the dynamics of a system, problem formulation, and planning and measurement issues (see Pfannkuch & Wild, 1998).

A PPDAC cycle is concerned with abstracting and solving a statistical problem grounded in a larger "real" problem. Most problems are embedded in a desire to change a "system" to improve something. Even ostensibly curiosity-driven research is usually justified by the idea that the accrued understanding will have long term practical benefits. A knowledge-based solution to the real problem requires better understanding of how a system works and perhaps also how it will react to changes to input streams, settings or environment. Certain learning goals must be met to arrive at the desired level of understanding. A PPDAC investigative cycle is set off to achieve each learning goal. Knowledge gained and needs identified within these cycles may initiate further investigative cycles. The conclusions from the investigations feed into an expanded context-knowledge base which can then inform any actions.

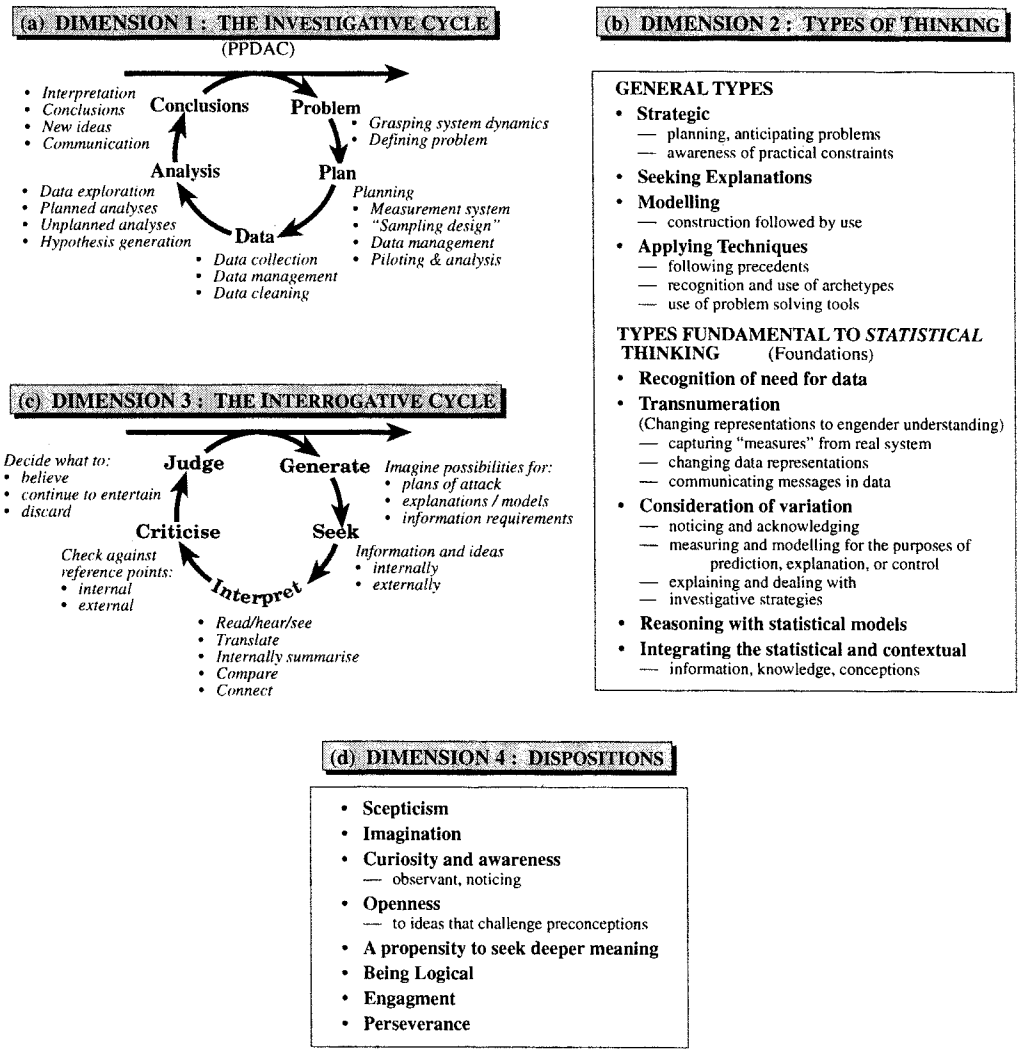

Figure 1. *A 4-dimensional framework for statistical thinking in empirical enquiry*

2.2 Dimension Two: Types of Thinking

A number of types of thinking emerged from the statisticians' interviews and were subsequently refined and modified when we applied them to the student and project-student interviews. The resulting categories are shown in Fig. 1(b). Some types of thinking were common to all problem solving. We will relate these general types of thinking to the statistical context in Section 2.2.2. First, however, we concentrate on types of thinking that are inherently statistical.

2.2.1 Types fundamental to statistical thinking

The types of thinking categorised under this heading in Fig. 1(b) are, we believe the foundations on which *statistical* thinking rests.

Recognition of the need for data: The recognition of the inadequacies of personal experiences and anecdotal evidence leading to a desire to base decisions on deliberately collected data is a statistical impulse.

Transnumeration: The most fundamental idea in a statistical approach to learning is that of forming and changing data representations of aspects of a system to arrive at a better understanding of that system. We have coined the word *transnumeration* to refer to this idea. We define it as "numeracy transformations made to facilitate understanding". Transnumeration occurs when we find ways of obtaining data (through measurement or classification) that capture meaningful elements of the real system. It pervades all statistical data analysis, occurring every time we change our way of looking at the data in the hope that this will convey new meaning to us. We may look through many graphical representations to find several really informative ones. We may re-express the data via transformations and reclassifications looking for new insights. We might try a variety of statistical models. And at the end of the process, transnumeration happens yet again when we discover data representations that help convey our new understandings about the real system to others. Transnumeration is a *dynamic* process of changing representations to engender understanding. Mallows (1998, Section 2) would appear to be advancing a similar idea.

Variation: Thinking which is statistical, in the modern sense anyway, is concerned with learning and decision making under uncertainty. Much of that uncertainty stems from omnipresent variation. The ASA resolution, and Moore and Snee's discussions of statistical thinking all emphasise the importance of variation. The last element of the list following "variation", namely "*for the purposes of explanation, prediction, or control*" is in the original statement of Snee (1990), albeit with a process-improvement spin, but has been dropped from the ASA statement. It is a critical omission. We do not measure and model variation in a vacuum. The purpose influences the way in which it is done. Our concerns with variation also extend beyond "measuring and modelling" to investigative strategies such as randomisation and blocking. In Section 3, we consider the variation theme in much greater detail.

A distinctive set of models: All thinking uses models. The main contribution of the discipline of statistics to thinking has been its own distinctive set of models, or frameworks, for thinking about certain aspects of investigation in a generic way. In particular, methods for study design and analysis have been developed that flow from mathematical models which include random components (see Mallows, 1998). Recently, however, there is a growing desire (enlisting a phrase from David Moore) to nudge "statistics a little further back towards its roots in scientific inference". Large parts of the investigative process, such as problem analysis and measurement, have been largely abandoned by statisticians and statistics educators to the realm of the particular, perhaps to be developed separately

within other disciplines. However, there are more valuable generic lessons that can be uncovered about these parts of the investigative process using other modelling tools. There is a need to expand the *reach* of our statistical models.

Context knowledge, statistical knowledge and synthesis: The raw materials on which statistical thinking works are statistical knowledge, context knowledge and the information in data. The thinking itself is the synthesis of these elements to produce implications, insights and conjectures. One cannot indulge in statistical thinking without some context knowledge. The arid, context-free landscape on which so many examples used in statistics teaching are built ensures that large numbers of students never even see, let alone engage in, statistical thinking. One has to bring to bear all relevant knowledge, regardless of source, on the task in hand, and then to make connections between existing context-knowledge and the results of analyses to arrive at meaning. Ideally, all of this knowledge would be resident in the same brain, but this is often not possible. Major investigations are team efforts which bring together people of differing expertise. Fig. 2 emphasises the synthesis of ideas and information from the context area and from statistics.

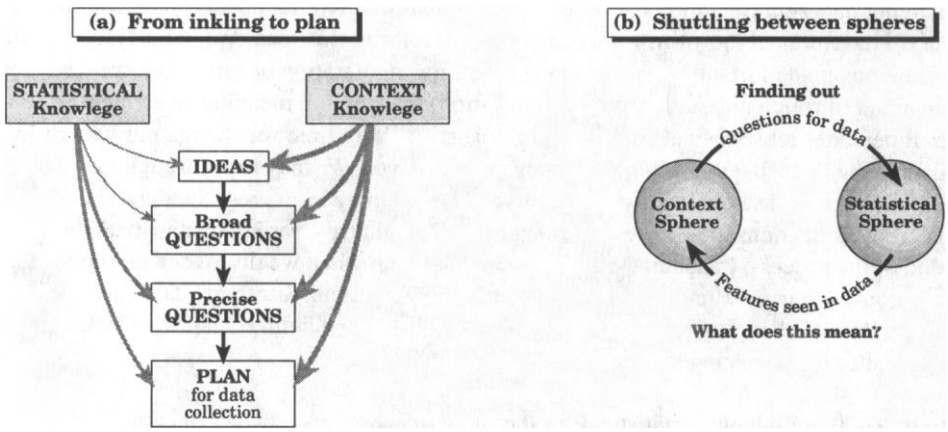

Figure 2. *Interplay between context and statistics*

Fig. 2(a) traces the (usual) evolution of an idea from earliest inkling through to the formulation of a statistical question precise enough to be answered by the collection of data, and then on to a plan of action. The earliest stages are driven almost entirely by context knowledge. Statistical knowledge contributes more as the thinking crystallises. Fig. 2(b) illustrates the continual shuttling backwards and forwards between thinking in the context sphere and the statistical sphere. This goes on all the time throughout PPDAC. For example, at the analysis stage questions are suggested by context knowledge that require consulting the data—which temporarily pushes us into the statistical sphere—whereupon features seen in the data propel us back to the context sphere to answer the questions, "Why is this happening?" and "What does this mean?"

2.2.2 General types of thinking applied in a statistical context

Strategic thinking

By strategic thinking, we mean thinking aimed at deciding upon what we will do (next or further into the future) and how we will do it. This includes such things as: planning how to attack a task; breaking tasks down into subtasks; setting deadlines for subtasks; division of labour; and anticipating problems and planning to avoid them. An important part of strategic thinking is having an awareness of the constraints one is working under and taking them into account in planning.

Real statistics is less about the pursuit of the "correct" answer in some idealistic sense than about doing the best one can within constraints. Many factors limit the quality and effectiveness of the thinking. Some of these factors are internal to the thinker. *Lack of knowledge* obviously constrains thinking. Unfortunately, what we "know" is not only our greatest asset but also our biggest curse because the foundations of what we "know" are often not soundly based. Our *preconceptions* can lead us astray in many ways, for example, by blinding us to possibilities because what we "know" determines where we look, and by desensitising us to important information. The challenging of something we "know" and take for granted can remove an obstacle and lead to new insight. This often occurs when people with different backgrounds discuss the same problem. Consulting statisticians see it at work in their clients when a quite innocent question surprises the client, loosens a previously held preconception, and leads to the client seeing the problem in a new way. We tend to solve problems by following "precedents". In applied research, this happens all the time and often the statistical methods of the *precedents are inadequate*. As far as *dispositions* (Dimension 3) are concerned, someone who is not curious, imaginative, sceptical and engaged will be less effective than someone who is. There is also an *ability* factor operating. Faced with the same set of information, some people will be better at making useful connections and grasping the essential features than others. And *inadequate communication skills* limit the ability to extract vital information and ideas from clients and others.

Other constraints are due to the environment the thinker is operating in. These include the general *time, money and materials* constraints, the imperfection of all human communication which results in misunderstandings and gaps in transmission of essential knowledge, and *limitations of the data* available. Very often, the problem we would like to solve is simply not soluble on the basis of the information we can get. For example, it may be impossible to capture with feasible measurement processes the characteristics we would like to capture. It may be impossible to sample the desired population or even a good approximation to that population, and so on.

This paragraph relates to particular constraints faced by statistical consultants, but students and other researchers are subject to some closely related constraints. The consultant works on problems owned by someone else. In other words, the statistician is in the position of having to satisfy "clients". This brings additional constraints which run deeper than time-and-materials constraints. Major decisions are made by, or must be cleared with, the client. The problem territory tends to be mapped out and even ring-fenced by the client. The client is often the chief source of context information so the statistician is not only constrained by the quality of communication and the extent of the client's knowledge, but will also tend to take on board the client's preconceptions. As the client is the final arbiter, the statistician is constrained by what the client can understand and accept. This can be strongly influenced by a number of what might be described as *psychological factors*. Statisticians have to gradually build up the client's trust in their judgement and abilities. An important consideration in "building trust" is not taking clients too far from territory in which they feel secure. An important element in client security is, in the words of colleague Chris Triggs, "what has been done in the field before". We call this the *first-in-the-field effect*. Early work in a field tends to take on an authority of its own whether or not it is warranted. It can influence every decision in the investigative process, right through to presentation. A related psychology of measurement effect

concerns the *sanctity of the measured variable*. To many clients, the way in which a variable has been measured takes on a meaningfulness and inviolability that a statistician might disregard, given the arbitrary elements in the initial choice of the variable. (This is not universal. Some client groups such as engineers are very sophisticated in this area.) The use of transformations in analysis is an area in which these issues come into sharp focus. Pfannkuch & Wild (1998) give a much more detailed and wide-ranging discussion, derived from the statisticians' interviews, of the realities of working with clients.

Modelling

Constructing models and using them to understand and predict the behaviour of aspects of the world that concern us seems to be a completely general way of thinking. All models are oversimplifications of reality in which information is necessarily discarded. We hope that we have caught the essential features of a situation and the loss of information does not invalidate our conclusions. Fig. 3 illustrates the way in which we learn about the context reality as a statistical investigation proceeds. As our initial quotation from David Bartholomew makes clear, "understanding" builds up in mental models of the context reality. These models are informed by information from the context reality, e.g. incorporating "expert knowledge". In an ideal world, we would be continually checking the adequacy of the mapping between model and reality by "interrogating" the context reality. Some of the information we seek and get from the context reality is statistical data. We build statistical models to gain insights from this information ("interpret") which feed back into the mental model. "Statistical models" here is more general than something like logistic regression. It refers to all of our statistical conceptions of the problem that influence how we collect data about the system and analyse it. Fig. 3 also incorporates the role of statistical knowledge and experience. Most obviously, it is a major determinant of the statistical conceptions we form in order to obtain and analyse data. Additionally, depending on the problem and the education and experience of the thinker, statistical elements can also be part of the way we think about the world and thus be integral parts of our mental models of the context reality.

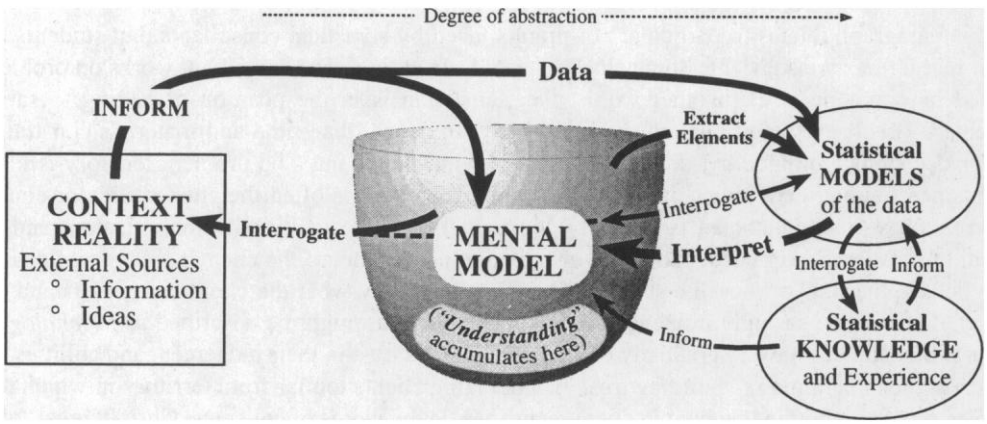

Figure 3. *Learning via statistics*

Applying techniques

A basic problem solving technique in the mathematical sciences is to find a way of mapping a new problem onto a problem that has already been solved so that the previously devised solution can be applied or adapted. The whole discipline of statistics is itself a manifestation of this strategy. Statistical theory makes the mapping process efficient by creating problem archetypes and linking them to methods of solution. To use statistics, we first recognise elements of our context that can be usefully mapped onto a model (a process of abstraction from the particular to the generic), operate within that model, and then we map the results back to context (from the generic to the particular). (Additionally, applied statisticians are always borrowing problem-solving ideas from previous experience with other problems and other data sets.)

Implementation of the problem-archetype strategy, and indeed the practical application of any technique, algorithm or concept, involves the three steps shown in Fig. 4. Instruction tends to focus on step 2, mechanical application. However, steps 1 (recognition) and 3 (interpretation in context) are: first, vital to step 2 having any utility, and second, inordinately more difficult. This is particularly true for the recognition step. (The project students needed to make constant external checks with their supervisor about whether they were on the right track.) One can deal with the mechanics of procedures by simply talking about them, establishing them with a few exercises and then moving on. Synthesis, insight, critical thinking and interpretation happen in the realm of the particular and require exposure to large numbers of disparate situations (cf. Wild, 1994).

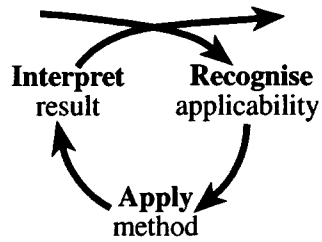

Figure 4. *Using any technique*

It is characteristic of statistics that we apply relatively sophisticated statistical models to the analysis of data and experimental design. Of all our statisticians, however, only the one operating in quality improvement seemed to use tools (e.g. cause-and-effect diagrams and process-mapping techniques) to analyse the nature of the problem itself. For the others, it seemed to be a process of imaginative construction of a mental model of the system, without discernible organisation. (The type of thinking "***seeking explanations***" has not been discussed in this section, but will be an important theme in Section 3.)

2.3 Dimension Three: The Interrogative Cycle

The Interrogative Cycle illustrated in Fig. 1(c) is a generic thinking process in constant use in statistical problem solving. From a detailed analysis of the project-students' and students' transcripts, it appears that the thinker is always in one of the interrogative states while problem solving. The cycle applies at macro levels, but also at very detailed levels of thinking because the interrogative cycle is recursive. Subcycles are initiated within major cycles, e.g. the "checking" step of any cycle can initiate a full interrogative subcycle. The ordered depiction on a wheel is an idealisation of

what perhaps should happen. In reality steps are often missed. We discuss the Interrogative Cycle as we observed it, being applied to statistical enquiry and statistical critique. The "thinker" is anyone involved in these activities. We now explore the components in more detail.

Generate: By this we mean imagining and brainstorming to generate possibilities, as an individual or in a group. We might be applying this to a search for possible causes, explanations and mechanisms, to the ways parts of a system might interrelate and to other building blocks of mental and statistical models. We might be applying it to the types of information we need to seek out to fill an information gap or to check out an idea, or to plan an approach to a problem or subproblem. The generation of possibilities may be from the context, the data or statistical knowledge and apply to the present problem, or may be registered for future investigation (hypothesis generation).

Seek: Generation tends to be followed by a seeking or recalling of information. This may be internal or external. For internal seeking, we observe people thinking "I know something about this" and digging in their memories for the relevant knowledge. External seeking consists of obtaining information and ideas from sources outside the individual or team. Working statisticians talk to other people about their problems—clients, colleagues, context-matter experts, people "working in the system". Seeking includes reading relevant literature. At the macro level it includes the collecting of statistical data, while at a more detailed level it includes querying the data in hand.

Interpret: By this we mean taking and processing the results of our seeking.

$$\text{Read/see/hear} \longrightarrow \text{Translate} \longrightarrow \text{Internally summarise} \longrightarrow \text{Compare} \longrightarrow \text{Connect}$$

This process applies to all forms of information including graphs, summaries and other products of statistical analysis. "Connect", the endpoint of "interpret" refers the interconnecting of the new ideas and information with our existing mental models and enlarging our mental models to encompass these interrelationships. Some of the problems observed in student thinking involved making one connection and then rushing to "judge" rather than trying to make multiple connections or going through the criticism phase.

Criticise: The criticism phase applied to incoming information and ideas involves checking for internal consistency and against reference points. We ask, "Is this right?" "Does this make sense?" "Does this accord with what else I or others know?" We check against *internal* reference points—arguing with ourselves, weighing up against our context knowledge, against our statistical knowledge, against the constraints we are working under, and we anticipate problems that are consequences of particular choices. We may also check against *external* reference points such as: other people (i.e. talk to clients, colleagues, experts, "workers in the system"); available literature and other data sources (e.g. historical data).

We can similarly try to take a mental step back and monitor our own thinking. Educational theorists talk about metacognition, of recognising and regulating one's normal modes of thought (see Shaughnessy, 1992). Reference points to check against here include the following: (1) *The purpose of the thinking*: for example, "Does this address the question the client wants answered?", or some sort of agreed objectives. (2) *Belief systems*: "Am I being unduly guided by unwarranted preconceptions—my own, my client's, or my community's?" Pfannkuch & Wild (1998) have some good cautionary tales from the experiences of our statisticians. (3) *Emotional responses*: One of our project students was worried about how the company's treatment of her seemed to be influencing the way she was approaching the problem and viewing the data.

Judge: This is the decision endpoint of criticism. What we keep, what we discard or ignore, what

we continue to tentatively entertain, what we now believe. We apply ***judgement*** to such things as: the reliability of information; the usefulness of ideas; the practicality of plans; the "rightness" of encapsulation; conformance with both context-matter and statistical understanding; the relative plausibility of competing explanations; the most likely of a set of possible scenarios; the need for more research; and the many other decisions involved in building and reasoning from models.

The result of engaging in the interrogative process is a ***distilling and encapsulating*** of both ideas and information. Internal interrogative cycles help us extract essence from inputs, discarding distractions and detail along the way (Fig. 5).

Figure 5. *Distillation and encapsulation*

2.4 Dimension Four: Dispositions

In this subsection, we discuss personal qualities categorised in Fig. 1(d) which affect, or even initiate, entry into a thinking mode. The nature of these dispositions emerged from the statisticians' interviews and we could subsequently recognise them at work in the students. We think these elements are generic, but again we discuss them as we observed them—in the context of statistical problem solving.

Curiosity and Awareness: Discoveries are triggered by someone noticing something and reacting to internal questions like "Why?", or "How did that happen?", or "Is this something that happens more generally?", or "How can I exploit this?" Being observant (aware) and curious are the well-springs of the question generation process that all innovative learning results from. Wild (1994) formed the slogan "Questions are more important than answers" to emphasise this point. Statistician Peter Mullins stressed the importance of "*noticing variation and wondering why*" for generating ideas for improving processes and service provision. We hazard that this very basic element of statistical thinking is actually at the root of most scientific research. "Noticing and asking why" is also critical for successful data exploration and analysis.

This brings us to ***engagement***. When the authors become intensely interested in a problem or area, a heightened sensitivity and awareness develops towards information on the peripheries of our experience that might be related to the problem. We suggest that this experience is fairly general.

People are most observant in those areas that they find most interesting. Engagement intensifies each of the "dispositional" elements curiosity, awareness, imagination and perseverance. How do we become engaged? Spontaneous interest is innate. Background knowledge helps—it is hard to be interested in something one knows nothing about. Being paid to do a job helps, as does the problem being important to people we care about. This may be our main difficulty in getting statistics students to think. They simply do not find the problems they are asked to think about interesting enough to be really engaged by them. We observed the effects on performance of engagement with some tasks and not others in the statistics students.

Imagination: It is hard to overemphasise the importance of imagination to statistical thinking. This is somewhat ironic given popular stereotypes of statisticians. The formation of mental models that grasp the essential dynamics of a problem is a deeply imaginative process, as is viewing a situation from different perspectives, and generating possible explanations or confounding explanations for phenomena and features of data.

Scepticism: By scepticism, we mean a tendency to be constantly on the lookout for logical and factual flaws when receiving new ideas and information. It is a quality all our statisticians both possess and value. Some writers refer to this as "adopting a critical attitude". Gal *et al.* (1995) and Pfannkuch (1996) discussed critical thinking in the interpretation of statistically based reports and media articles. Scepticism here was basically targeted towards, "Are the conclusions reached justified?" There may be worries about the motivation, predispositions and objectiveness of the writer which would effect the level of trust in anything that had been done. Experienced statisticians are likely to evoke automatically technical "worry questions" concerning the appropriateness of the measurements taken, the appropriateness of the study design, the quality of the data, the suitability of the method of analysis, and whether the conclusions are really supported by the data. Postulated explanations create worries about whether this really is the only plausible explanation.

Another aspect involves *a sense of number* and scepticism. A precursor step towards "Is this information/conclusion *justified*?" is "Is this information/conclusion even *credible*?" One of our statisticians told the simple story of reported attendance rates at a free outdoor concert in Auckland. If the figures were correct, that would mean that one in every three Aucklanders, one in nine New Zealanders, would have needed to have attended and that was, frankly, incredible. The information is discounted at this first hurdle. However it should be noted that one is much less inclined to be sceptical when conclusions fit one's own preconceptions. A conscious effort may be required to counter this.

Being logical: The ability to detect when one idea follows from another and when it does not, and to construct a logical argument is clearly important to all thinking. Synthesis of new information with existing knowledge is largely a matter of seeing *implications*. Logical reasoning is the only sure way to arrive at valid conclusions. To be useful, scepticism must be supported by an ability to reason from assumptions or information to implications that can be checked against data.

A propensity to seek deeper meaning means not simply taking things at face value and being prepared to dig a little deeper. Of the other "dispositions", ***openness*** helps us to register and consider new ideas and information that conflict with our own assumptions and ***perseverance*** is self evident.

Can "dispositions" be taught?

Schoenfeld (1983) analysed the mathematical problem solving experience within individuals in terms of a "manager" and an "implementer" working in tandem. The manager continually asks questions of a strategic and tactical nature deciding at branch points such things as which perspective to adopt and which direction to take or abandon. We have described the characteristics above as "dispositions". They tend to initiate manager functions. We first thought of the dispositions as innate

characteristics of the thinker but had to modify this with the idea of "engagement". A person's "dispositions" are problem dependent—they change according to the degree to which the person is engaged by the problem. One of our statisticians was adamant that some people are sceptical, others are credulous, and there is little one can do about it. The authors are less pessimistic. It seems to us that credulousness in a particular area is a result of ignorance. As you gain experience and see ways in which certain types of information can be unsoundly based and turn out to be false, you become more sceptical. Moreover, all we want in operational terms from scepticism is a prompting to raise certain types of "worry" question [cf. Gal *et al.*'s (1995)] concerning the reliability of information, which can be taught (see Section 4).

3 Variation, Randomness and Statistical Models

3.1 Variation as the Starting Point

The centrepiece of the quality and ASA definitions of statistical thinking is "variation" or "variability". Any serious discussion of statistical thinking must examine the role of "variation". The "variation" terminology and message seem to have arisen in one small area of statistical application, namely that of quality, and their penetration into other areas would appear to be slight. If "variation" (as a major source of uncerainty) is indeed to be the standard about which the statistical troops are to rally, we need to arrive at a common conception of statistics in terms of "variation". This section attempts such a conception. Moreover, we are striving for a view of statistics "from the outside".

The first three "variation" messages are: variation is omnipresent; variation can have serious practical consequences; and statistics give us a means of understanding a variation-beset world. Subsequent messages concern how statistics goes about doing that.

Omnipresence: Variation is an observable reality. It is present everywhere and in everything. Variability affects all aspects of life and everything we observe. No two manufactured items are identical, no two organisms are identical or react in identical ways. In fact, individual organisms are actually systems in constant flux. The aforementioned refers only to real variation inherent in the system. Fig. 6 depicts how, when we collect data from a system, this real variation is supplemented by variation added in various ways by the data collection process.

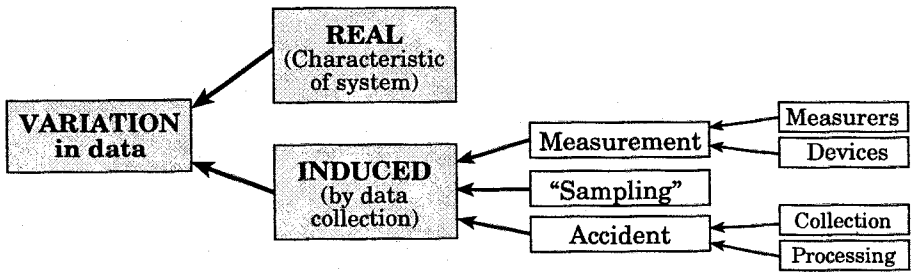

Figure 6. *Sources of variation in data*

Practical impact: Having established that variation is everywhere, we have then to demonstrate the important practical impacts of this variation on peoples' lives and the way they do business. It is variation that makes the results of actions unpredictable, that makes questions of cause and effect difficult to resolve, that makes it hard to uncover mechanisms. Variation is the reason why people

have had to develop sophisticated statistical methods to filter out any messages in data from the surrounding noise.

3.2 *Predict, Explain and Control*

Fig. 7 categorises rational responses to variation in a system in the world of action. This is idealistic. The way people actually *do* react to variation can be quite another story! (See Joiner, 1994, Chapter 10).

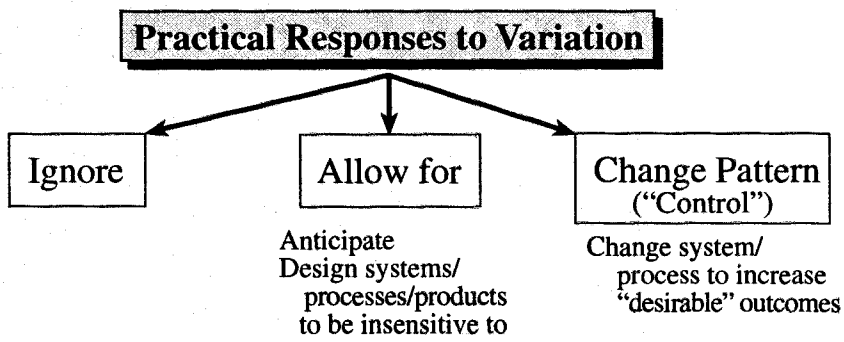

Figure 7.

First, we can pretend that the variation does not exist, e.g. behave as though every object or organism is the same or differs in some deterministically known way. In some circumstances this works admirably. If it did not we would have to write off all of applied mathematics and every field it fertilises. Second, we can investigate the existing pattern of variation and come up with ways of working around it as in our system of clothing and shoe sizes. Variation is allowed for at the design stage in quality management approaches to manufacturing where one wishes to design a product that is "rugged" or "robust" to the variability of uses to which it will be put and conditions to which it will be subjected. Third, we can try to change the pattern of variation to something more desirable, e.g. to increase average crop yield or reduce a death rate. We do this by isolating manipulable causes, or by applying external treatments. The former approach is often used in quality improvement or in public health, the latter is frequently used in agriculture or in medical research aimed at the treatment of individual patients.

Statisticians model variation for the purposes of prediction, explanation, or control. *Control* is changing the pattern of variation to something more desirable. *Prediction* is the crucial informational input to "Allow for" in Fig. 7. *Explanation*, gaining some level of understanding of why different units respond differently, improves our ability to make good predictions and it is necessary for control. Causal and mechanistic explanation is the goal of basic (as opposed to applied) science. As soon as we ask "Why?", we are looking for causes. While on the one hand variation may obscure, it is the uncontrolled variation in a system that typically enables us to uncover causes. We do this by looking for patterns in the variation. Fig. 8 picks up this idea in a way that relates back to the goals in Fig. 7.

Statisticians look for sources of variability by looking for patterns and relationships between variables ("regularities"). If none are found, the best one can do is estimate the extent of variability

and work around it. Regularities may or may not correspond to causes. In terms of solving practical problems, causes that cannot be manipulated are operationally equivalent to any other observed regularity, although they will give us more confidence in our predictions. The presence of regularities enables us to come up with predictions and measures of variability that are more locally relevant, e.g. more relevant to an individual patient. Manipulable causes open the option of control.

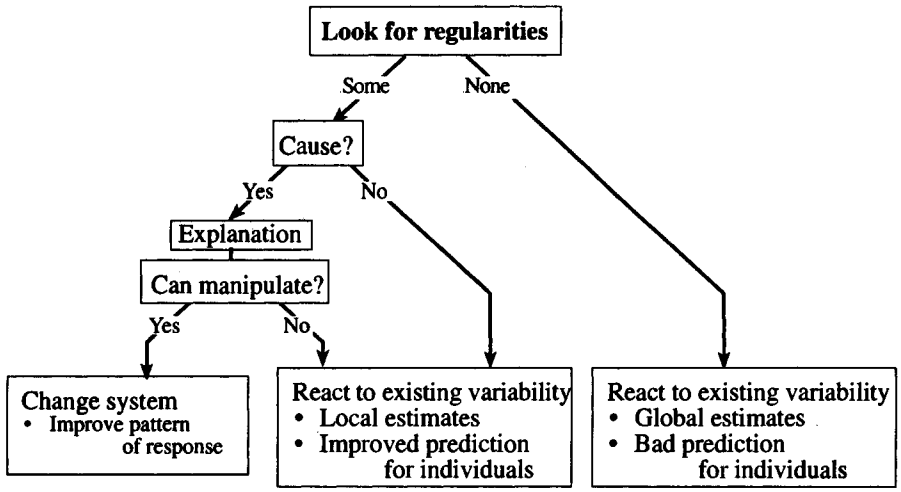

Figure 8.

3.3 The Quest for Causes

In our research, two groups of first-year students were given media clippings and similar items containing statistical information, and then interviewed individually about their responses. Initially, our approach to the student transcripts was that of teachers marking term papers, of looking for mistakes and gaps, for what the students had "done wrong". One item was based on Tversky & Gilovich (1989). The streaks that sports fans see in sports data, and then proffer all sorts of causal explanations for (e.g. falters under pressure), can often be explained entirely in terms of a random, e.g. binomial, model (see also Moore, 1990; Falk & Konold, 1992; Biehler, 1994 presents other perspectives). The item run concerned a basketball player with a 70% success rate of free throws succeeding with only 2 out of 5 throws. Under a binomial model, this is not a particularly unusual event. We think of this as "the statistics teachers' point". Our students proffered all sorts of causal explanations. As statistics teachers, we thought they had missed the point. Mark that one wrong! For the next group, we loaded the item entirely in favour of the statistics teacher's point: "*The team manager attributed her performance to normal variation, that she scored 70% in the long run and that 70% was only an average so that you had to expect some low scores now and again.*" Even then we saw the tip of the deterministic-causal-thinking iceberg. One student said, "*the manager's comments are OK if that is the way he wants to look at the score and not on 'we want to win'* " and then he gave possible causes.

This comment eventually overturned our attitude. The student was right. There is a real problem underlying this item. Coaches and managers of sports teams are seeking to learn from their observations so that they can work on improving player skills and strategy, and better deploy their pool of available players. A random model is of no help at all in this regard. The statistics teacher's concerns

do not replace the need to search for causes and predictors of good and bad performance. It is that search that is of primary importance.

The real problem underlying the statistical problem very often involves searching for and isolating causes of some response. We have run a variety of stories including medical stories and prison suicides. Whenever students have contextual knowledge about a situation, and their experience has given them some notion of the nature of the real problem, they will come up with a range of possible causal explanations with little or no prompting. This appears to be a well developed impulse that has not been explicitly taught. It is a first and direct step towards solving the primary problem. The real purpose of many of the new statistical ideas we teach is simply to moderate the search for causes by preventing a premature jumping to conclusions—"Hey, not so fast ...." This role is secondary and subtle. It is probably not surprising then, that even after some statistical instruction, the randomness ideas are much weaker in students than the impulse to postulate causes. Probabilistic thinking is not so much an alternative to deterministic thinking, as some statistics educators (Shaughnessy, 1992) and statisticians (Hoerl *et al.*, 1997) have suggested, but something to be grafted on top of the natural thinking modes that directly address the primary problem. As an interesting aside, if an explanation or cause has already been suggested to students for a particular set of data or if the data has been presented stratified in some particular way, it can take a great deal of prompting for the student to go beyond the explanation given, to think that there may be other explanations and start coming up with ideas. This latter is a quite different incarnation of "Hey, not so fast, ..."

What does statistics education have to say about causation? By far the loudest message is, "correlation is not causation". This is the statistician as Cassandra, the harbinger of doom saying "this way lies disaster". True, we usually go on to make the important point that the randomised experiment is the most convincing way of establishing that a mooted relationship is causal. But, as stressed by Holland (1986), Cox (1992) and few others outside of quality and epidemiology, this greatly undersells the true importance of the search for causes. Solving most practical problems involves finding and calibrating change agents. Statistics education should really be telling students something every scientist knows, "The quest for causes is the most important game in town". It should be saying, "Here is how statistics helps you in that quest. Here are some general strategies and some pitfalls to beware of along the way .....". It should not just be preventing people from jumping to false conclusions but also be guiding them towards valid, useable conclusions—replacing Cassandra by a favourite literary detective.

Thinking about causes

It is ironic that the uncontrolled variability in a system provides us with the best opportunities to do the detective work that uncovers causes. By checking for relationships between upstream variables and downstream responses, we can identify possible causal factors. Observation precedes experimentation. All *ideas* for possible causal relationships originate in observation, whether anecdotal or from formal studies. And as we continually stress, randomised experiments provide the most convincing way of confirming or refuting the causal nature of an observed relationship.

Conducting any sort of study to detect causes and estimate their effects proceeds from ideas about profitable places to look, ideas which draw almost exclusively on context-matter knowledge and intuition. Ideas about possible causes and other factors that might be important predictors of the behaviour of the response are translated into a set of variables to measure (transnumeration) and data is collected to facilitate investigation of relationships between measured variables and the responses of interest. The primary tools of analysis in the search for causes are models of the regression type, i.e. models for exploring how Y-behaviour changes with changes in X-behaviour. (The humble scatter plot falls into this class.)

Cox (1992) distinguishes between: response variables (those whose behaviour we want to find

causal explanations for); intermediate response variables (which measure intermediate effects that happen along the way from initial state to response state) and explanatory variables (those we want to use to explain or predict the behaviour of the response). Explanatory variables are further categorised into possibly causal variables, intrinsic properties of entities under study, and non-specific (e.g., different countries). Intrinsic variables are those whose values cannot be manipulated. Intrinsic variables are often included to improve our ability to detect relationships, improve the precision of estimation of effects and to explore how a cause may act differently for different types of entity (interactions). Detection of strong relationships between non-specific variables and a response lead to a search for new explanatory variables, variables associated with the non-specific variable which could conceivably explain the response. For example, when disease rates differ greatly between countries, we start looking among factors that differ between the countries as possible causes. We note that the above distinctions between variables are much more than distinctions for analysis. As a set they constitute a general thinking tool which adds clarity to the way in which the context-matter problem is conceived. They are an example of the way in which statistical knowledge or training can feed into a core mental model of the context reality that is understandable by statisticians and non-statisticians alike, the "inform" arrow linking statistical knowledge and mental model in Fig. 3.

Some consequences of complexity

Most real systems are enormously complex with variation in innumerable components, each of which could contribute to the response of interest. We are incapable of handling such complexity and need strategies to "sift" the large numbers of possibilities for a much smaller number of promising leads. The quality control distinction between *special cause* and *common cause* variation can be seen in this light. It gives a means of distinguishing situations (special-cause variation) in which the seemingly instinctive human reaction of looking around for something unusual occurring in the system just prior to the problem event is likely to be a profitable strategy for locating the cause, from situations (common-cause variation) in which this strategy is unlikely to be profitable and may even be harmful.

The main statistical "sifting"-strategy is to restrict attention to variables which have strong associations with the response of interest. We have no hope of identifying a cause and characterising its effects if it acts in complexly different ways for different individuals or at different times. The only causes that we can hope to find are those that act in a reasonably uniform or regular way. Moreover, we will only detect the existence of a cause if we think of some way of looking at the situation that will reveal that regularity (transnumeration). There must be sufficient "natural variability" in a cause-variable in the system for the effect of this variability on the response to be seen. Causal variables that we miss using the above strategy are unlikely to be good agents for making substantial changes unless settings are used that lie far beyond the range of variability seen in that variable in the data.

From association to causation

It is at this point that Cassandra makes her entrance. And the world really does need her warnings. It is clear that people do jump far too quickly to causal conclusions. But "correlation is not causation" is simply a "Hey, not so fast" warning and we need to supply ways of moving on from there. The search process has not given us a set of causes. It has only given us a set of promising contenders for causal status. Our main worry at this point stems from the fact that we have not considered the universe of relevant variables, but just that subset that happened to come to mind. We are worried that other unconsidered factors, those sinister lurking variables of textbook fame, may be producing the relationships we are seeing. Thus, we challenge the causal assumption, whether our own or

somebody else's. We rack our brains for other possible explanations and for strategies for testing these explanations. This behaviour has to be learned. It comes naturally to very few students. The goal of the scientist is to reach a position at which there are no other plausible explanations at the current level of understanding. To do this, we need strategies which use observation, experimentation and analysis to discount all other plausible alternatives.

Where experimentation is not possible and one must make decisions based upon using observational studies, there is a range of ideas about what strengthens the impression that a causal contender is in fact a cause. The criteria of A.B. Hill (1965, see also Gail, 1996) are a good starting point. In epidemiology and quality, the finding of causes with a view to improving systems is not a philosophical problem but a pressing practical imperative. Substantial literatures have grown up in these fields. Cox (1992, 1993) and Holland (1986) also view questions of causation with practical applications clearly in mind. In view of the fundamental importance of the search for causes, there is a real need to synthesise this material into accounts which are more accessible for practising investigators and for teachers.

Levels of "causal proof"

Decisions to take action tend to be made on the basis of a "best guess" in the light of the available information. They seldom wait for incontrovertible evidence of causality. The results can be spectacularly good. Take cot death in New Zealand. Research showed strong relationships between cot-death rates and certain behaviours, e.g. the way the baby was put down to sleep. There was no incontrovertible proof that the behaviours caused cot death but the idea was sufficiently plausible to mount publicity campaigns and the advice given to new mothers by doctors. Cot death rates halved. There is a level of assurance at which decision makers are prepared to take what they consider to be a small chance and take action. There are many factors affecting this level of assurance. The degree of causal proof it takes will probably depend on many factors including the difficulty of making (and reversing) changes to the system, the consequences of making a wrong call, and the number of people who must be convinced before action is taken. We are all brought up on the smoking-cancer debate as the primary example of the difficulties in establishing causality. In that debate, there were (and are) entrenched and powerful vested interests with a high political profile. Not surprisingly, the level of proof required in such circumstances is extremely high. An industrial production manager would have made the call long before, with the greater attendant risk of getting it wrong.

3.4 Modelling Variation

A number of statisticians have told us that the biggest contribution of statistics is the isolation and modelling of "signal" in the presence of "noise". The base problem with statistical data, is how to make some sort of sense of information that is, if one considers the details, of mind-boggling complexity. The main statistical approach to solving this problem begins by trying to find patterns in that data. Context knowledge may give us some ideas about where to look and what to expect. Statistical methodology gives us tools to use in the search. Common experience tells us that studies conducted under very similar conditions always give results which are different in detail, if not in broad thrust—patterns seen in data from one study are never repeated identically in another. The base problem, then, is to come up with strategies for separating phenomena which are "likely" to persist more generally from those that are purely local, to sift the enduring from the ephemeral. Patterns which persist provide the basis for forecasting, control and insight. Statisticians have evolved particular sets of strategies for "solving" this problem—strategies based, in the main, on probabilistic modelling. We often say that an important function of probability models and statistical inference is to counteract a human tendency to "see" patterns where none exist. As statistician (and also zoologist)

Brian McArdle put it so vividly in a personal interview, "*The human being is hard-wired to see a pattern even if it isn't there. It's a survivor trait. It lets us see the tiger in the reeds. And the downside of that is that our children see tigers in the shadows on the wall.*" It is not entirely true that no patterns appear in purely random phenomena. These patterns are real to the brain in the sense that we can recognise features that would help us reproduce them. However, such patterns are (i) ephemeral, and (ii) tell us nothing useful about the problem under study. In other words, they are meaningless. Part of our reasoning from random models is to say that we will not classify any data-behaviour as "enduring" if it closely resembles something that would happen reasonably frequently under a purely random model.

The distinction between "explained" and "unexplained" variation is important here. We generally try to find meaning in explained variation, the patterns which we have not discounted as ephemeral, the "signal". Unexplained variation, or "noise", is what is left over once we have "removed" all the patterns. It is thus, by definition, variation in which we can find no patterns. We *model* unexplained variation as being generated by a (structureless) random process. We have no idea whether this variation really is random; this is not something that bothers us. If random sampling really has occurred, there is an element of randomness in the noise. However, measurement error and components of the variation in the original process typically contribute to the unexplained variation and there is no way of knowing whether these behave randomly or not. In fact, randomness is just a set of ideas, an abstract model, a human invention which we use to model variation in which we can see no pattern. The very physical models we use to illustrate randomness are, with sufficient knowledge, actually deterministic (see Stewart, 1989, Chapter 14). It is all part of an attempt to deal with complexity that is otherwise overwhelming, and it tends to be a model-element of last resort. The level at which we impose randomness in a model is the level at which we give up on the ability to ask certain types of question, questions related to meaning and causation.

Language such as "real or random" or referring to the possibility that "the observed difference is due to chance" actively obscure the distinction between the underlying problem and a statistical approach to its solution. In talking about a project he did on mangroves one student said "*My teacher explained it [t-test] to me that the results I got were due to chance. I still don't think that statement makes any sense. I can understand what chance is when you are rolling a dice. I don't really understand what chance is when you relate it to biological data. Everything you could possibly measure is going to be due to some environmental impact.*"

Some writers in quality have taken to saying, "*all variation is caused*"; e.g. Joiner & Gaudard (1990), Pyzdek (1990). The latter repudiates the "outdated belief that chance causes should be left to chance". These claims seem to be predominantly motivated by concerns about human psychology. Tomorrow, with new information, insight or technology, we may be able to find patterns in what today looks random, to trace causes from those patterns, and to improve the system (Pyzdek gives examples where this has occurred). The propensity to do so may well be lost if the idea is internalised that this variation is "just random". In commenting on the difficulties people have with coming to grips with statistics, Shaughnessy (1992) wrote "*the real world for many people is a world of deterministic causes . . . there is no such thing as variability for them because they do not believe in random events or chance.*" We do not need to ask them to. Variability is a demonstrable reality. Randomness need not relate to any belief system about the true underlying nature of reality. It is simply a response to complexity that otherwise overwhelms us. The unexplained variation may well be the result of "a multiplicity of causes", to use the phrase of Falk & Konold (1992). Few would dispute that much unexplained variability is of this type. But, the statistical response is that if we can see no structure there, we will model it as having been generated randomly.

From these models, we make inferences. We assume that the data has been randomly generated according to the model and use probability as the link between population/process and data. This is the very heart of the statistics we teach. Our models, including their random components, stand or

fall on the practical usefulness of the answers they produce. There are some clear success stories, e.g. the insurance industry. To use models with random components, we have to be able to: first, recognise that such models provide a useful framework for considering the problem; second, build and fit an appropriate model; and third, deduce implications from that model. The third step involves some understanding of how random models behave. There is an emerging literature on the difficulties in gaining that understanding; see, for example, Pfannkuch & Brown (1996), Garfield & Ahlgren (1988), Konold (1994). Our inferential paradigms are also subtle and difficult to grasp, but we will not discuss that here (see Mallows 1998, Section 7; and Cox, 1997, Section 5).

One of the stories we have been showing students, and our reaction to it, niggled at us for a long time. It was a news story about an apparent jump in prison suicides, the sort that leads to accusatory finger pointing and the pushing of different causal explanations by different sectional interests. We automatically did a quick check against Poisson variation. The figure was within reasonable limits. We sensed a tendency, as a consequence of this calculation, not just to disregard the hype, but to disregard the problem. However, prison suicides are an important problem and people should be looking for causes. It took a long time to realise that what the lack of significance really tells us is to adopt the common-cause-variation strategy of in-depth study rather than the (popular) special-cause-variation strategy of looking among recent changes for a cause.

Relating the "variation" words

We conclude this section by putting some summarising organisation into the "variation" terminology. *Special-cause* versus *common-cause* variation is a distinction which is useful when looking for causes, whereas *explained* versus *unexplained* variation is a distinction which is useful when exploring data and building a model for them. An understanding of variation in data could be built on these suppositions: (1) variation is an observable reality; (2) some variation can be explained; (3) other variation cannot be explained on current knowledge; (4) *random* variation is the way in which statisticians model unexplained variation; (5) this unexplained variation may in part or in whole be produced by the process of observation through *random sampling*; (6) randomness is a convenient human construct which is used to deal with variation in which patterns cannot be detected.

4 Thinking Tools

Gal *et al.* (1995) used the term "*worry questions*" when discussing the critical appraisal of reports—questions to invoke worries about the way information had been obtained and how inferences had been drawn from it. *Trigger questions* (e.g. "Why?" and "How?") are their creative cousins. They tend to initiate new thinking in certain directions. We will use the term "trigger question" for both roles. Such questions can be very effective. Many times in our interviews when no thinking was taking place, some small prompt opened flood gates of thought.

Experienced statisticians working in a collaborative or consulting environment learn to generate trigger questions which elicit pertinent context information, ideas and explanatory inferences from clients. The success of the dialogue between statistician and client may depend upon the quality of the trigger questions. No one taught our statistical consultants to ask the questions they do. Our consultants' statistics education had relied on the process: stimulus + experience + disposition $\longrightarrow$ pertinent trigger questions $\longrightarrow$ gaining critical ideas and knowledge about the context. This completely unstructured approach puts an enormous premium on experience. If statistical thinking is something that we teach rather than something simply to be absorbed by osmosis, then we have to give it structure. Structure can stimulate thinking, prevent crucial areas from being overlooked, and provide something to fall back on when you hit a brick wall.

The idea from the quality armory that we have found most powerful is the simple idea of intercon-

nected processes, with associated process diagrams, as a framework for analysing problems. It gives a structured way of breaking down activities, or mental processes, into components. It emphasises principal steps and their interrelationships and hides detail which can be uncovered subsequently when analysed as (sub)processes. Joiner's 7-Step Process (see Joiner 1994) is a thinking tool for working through for a quality improvement project. It leads the thinker through the steps of a project in time order and, within each step, gives lists of questions to prompt the crucial types of thinking that should occur there. The above approach is not new. Polya (1945) used it in *How to Solve It*, the most famous of all works on mathematical problem solving. In the quality arena, we see many serious attempts to capture essential elements of expert experience through creating thinking models/tools which can be used to approach specific types of problem. Underlying all of the above are two simple principles, which we have combined to form *systemise what you can, stimulate what you cannot.*

Schoenfeld (1987) distinguishes between a *description* which characterises a procedure and a *prescription* which characterises a procedure in sufficient detail to serve as a guide for implementing the strategy. PPDAC is a high-level description of a systematic approach to investigation. It identifies major elements. It can be used as the foundation for something that is much more of a prescription. This is a huge undertaking so what is presented here is merely indicative. The principles involved in our model fragments (Figs. 9 and 10) are:

- Systemise what you can, stimulate what you cannot.
- Use trigger questions to do the stimulation.
- Work from overviews and zoom in for the next level of detail.
- Keep the number of steps in any view of the process small to emphasise the most important relationships.

At any level, one drills down for more detail by clicking on a node in the process diagram (e.g. in an internet-type application). The area we have applied this to (in Figs. 9 and 10), is drilling down into the "Plan" node of PPDAC and then further down again into the "Measurement" node of the model of "Plan". We stopped at this level of detail and used sets of trigger questions about measurement (derived from the interviews) which are very general. Context-matter disciplines have built up enormous amounts of expertise about how to measure the things that are of great importance for research in their discipline. We have simply pointed to that with our questions. Models targeted at a particular application area could build in much more of that local expertise. Still, a general model such as ours could be useful for someone doing applied research in a less developed area, and for building in statistics students a more holistic feel for statistical investigation and the broad issues that need to be addressed. It prepackages some of the "strategic thinking" of breaking major tasks down into subtasks.

An attractive model element that we have not incorporated here, though it might be useful to do so, are lists of the tools that are helpful at particular nodes of a process. For examples, see Hoerl & Snee (1995). Process analysis tools provide us with a means of building up new bodies of "statistical theory" addressing critically important areas of the statistical process that statistics teachers are currently rather silent about. The results will be oversimplifications and sometimes gross oversimplifications, but then so are all our mathematical models. The theories should give students two opportunities to learn about and make sense of the statistical endeavour. First, the theory provides a scaffolding to use in forming a picture of some very complex processes. Second, once such a picture has been established, a more sophisticated understanding can be gained by considering ways in which the models are inadequate.

We conclude this subsection with a story related to us by some project students in a quality improvement course that sheds light on the complementary roles of theory and experience. The students first learnt some theory about quality improvement (including the role of statistical tools) via lectures and readings and found it all rather abstract and meaningless. On their first practical

project they floundered. The theory did not seem to help. But from those experiences the theory started to make sense. And by the second project it had started to work for them—its value had become obvious.

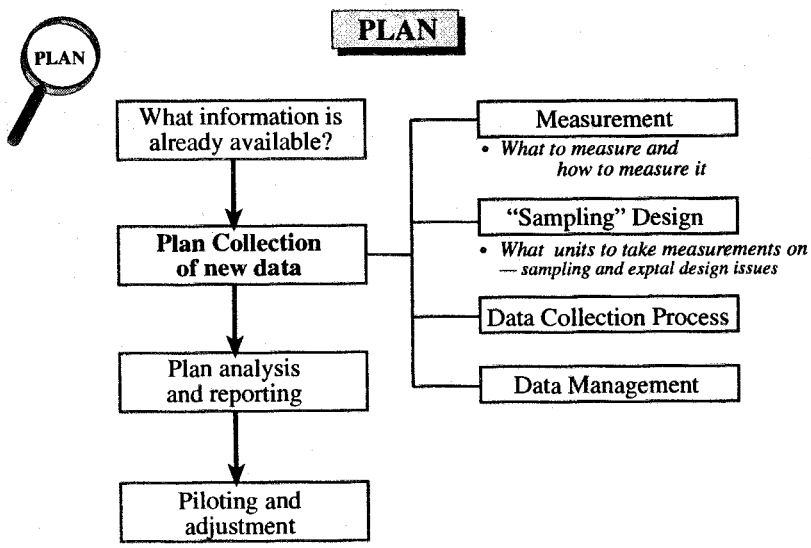

Figure 9. *Drilling down into the "Plan" node of PPDAC*

5 Discussion

The ultimate aim of statistical investigation is learning in the context domain of a real problem. Learning consists of the construction of mental models of the system under study. Statistics is itself a collection of abstract models ("models" is used in a very broad sense) which permit an efficient implementation of the use of archetypes as a method of problem solution. One abstracts pertinent elements of the problem context that map onto a relevant archetypical problem type, uses what has been worked out about solving such problems, and maps the answers back to context domain. There is a continual shuttling between the two domains and it is in this shuttling or interplay, that statistical thinking takes place—where the statistical rubber meets the real-world road. When it works, we gain real traction. Our abstraction processes bring clarity to thinking and efficiency to problem solving. However, when we use archetypes to solve problems, an enormous amount rides on the ability to do the mappings. And this is where the wheels so often fall off. Statistics education spends little time on developing the mappings. We must take more cognisance of the fact that the getting from the first stirrings of a practical problem to something like $y = \beta^T x + \varepsilon$, the point at which the theory of analysis typically kicks in, does not involve driving blithely across some small crack in the road, but rather it involves the perilous crossing of a yawning chasm down which countless investigations and analyses plummet to be lost without trace.

For successful problem solving, statistical thinking is not a separable entity. There is only holistic thinking that can and should be informed by statistical elements. The more relevant knowledge one

MEASUREMENT

(Including Classification)

Identify Key Characteristics

What features of the system are you interested in?
For each feature:
- What ideas about this feature are you trying to capture?
- Can you substitute something more specific?
- How well does this capture the idea?
- Is the idea you wish to capture by measurement inherently multidimensional in important ways or should a single measurement do?

Do these "key characteristics" adequately capture the essence of the real system?

Decide how to measure them

Experience in the field
Is there a generally accepted way of measuring this charactersitic in the field?
Is this currently accepted in the field as the best way to measure this entity?
If there is an accepted method and you want to use something else:
- What is your justification?

Have others tried to measure this? How did they do it?
- Are there known problems with their measure?

A fall-back
Can I draw on the experience of others in measuring similar charactersitics?

Anticipate problems

Validity & reliability
To what extent does this measurement really capture the characteristic I want to measure?
What are the practical implications of the extent to which it fails?
Will repeat measurements on the same units give very similar results?
Will different people making such measurements obtain very similar results?
Will different measuring instruments give very similar results?
If not, what impact will this have on the usefulness of any conclusions?

Analysis
Will I be able to analyse data containing measurements like this?
Will this measure make the analysis unnecessarily difficult?
Will another choice confer greater statistical efficiency?

"Audience" reaction
Will others be able to understand this measure?
Will the audience for the results accept that this is a sensible way to measure this?
Will I be able to understand the results of an analysis based on these measures?
Will I be able to communicate the results of an analysis based on these measures?

Practical implement-ation
Can I implement these measures in practice on the scale needed for the study?
Is the equipment/personnel required for this measure available? affordable?
Is the measure unacceptably or unnecessarily difficult? expensive? invasive?
Are there cheaper/easier/less invasive alternatives that will serve almost as well?

People: Do these measures take account of the psychological, cultural and perceptual differences of the people to be measured?

Can I do better?

Figure 10. *Drilling down further into the "Measurement" node of "Plan"*

has and the better one can connect it, the better one can do. In many research environments, statistical thinking is like breathing—everyone does it all the time, seldom being aware that it is happening. Statistics, the discipline, should be teaching people to "breathe" more effectively. However, we are dealing with complex and sophisticated thinking processes. We cannot expect, and indeed should be highly suspicious of, what W. Edwards Deming called "*instant pudding solutions*".

In Section 2, we identified several dimensions of the statistical thinking used in empirical enquiry: (1) the *investigative cycle*; (2) *types* of thinking; (3) the *interrogative cycle*; and (4) *dispositions* (see Fig. 1). We further discussed factors constraining the effectiveness of the thinking. Much of what we have identified relates to general problem solving skills being applied in a statistical context. One might think that for such skills a general thinking skills course such as those developed by de Bono is all that is needed. According to Resnick (1987), however, there is no empirical evidence that even these general skills are transferred to specific subject areas. She believes (p. 35) that thinking processes should be embedded into the discipline itself because, "*it provides a natural knowledge base and environment in which to practice and develop higher order* (thinking) *skills as . . . one must reason about something . . .* (and) *. . . each discipline has characteristic ways of reasoning . . .*." To carry out this embedding, we need more research into how these broad thinking skills are specifically used in statistics.

Omnipresent variation was presented as providing an important raison d'être for statistical thinking. In Section 3, we took the observable reality of variation in the concrete world as a starting point and endeavoured to cut through many of the confusions surrounding such abstract notions as "random variation" and their application to practical problem solving.

In Section 4, we discussed some techniques people have used to improve thinking. It seems to us that the rest of statistics can only benefit by following the lead of our colleagues in process and organisational improvement and develop tools that help us to think about, and to think through, parts of the statistical process that we are presently rather silent about. We can develop other forms of statistical model, other forms of statistical theory to deal with these areas. We stress that thinking tools are not a substitute for experience with investigation and data. Probably their most important purpose is to help us understand our experience and extend the range of situations to which we can apply it. But they may also re-initiate thinking that has become stalled.

As for the usefulness of the models presented, we have subsequently used this framework to develop judgement criteria to help students interpret statistically based information such as in media reports. We can see multiple uses even for a very simple model like the interrogative cycle (Fig. 1(c)). It could be used: to monitor thinking during problem solving; to help students become aware of their own thinking; as a tool for evaluating student thinking; and as a reference point against which to check learning opportunities provided to students. Do they, at least collectively, provide good opportunities for the students to experience all of these modes?—It turns out that many of the tasks we gave students did not! Nor did they measure up in terms of types of thinking.

Can thinking tools work? The people in process and organisational improvement and Polya and his successors in mathematics believe so. Are they panaceas? There is nothing certain or cut-and-dried in applied statistics. The real world is a messy, complicated place. We are reminded of David Moore's distinction between mathematics and statistics, "Mathematical Theorems are true; statistical methods are sometimes useful when used with skill." We cannot expect more from our new tools than from our traditional ones. Statistics is not some hiking tent that can be erected in an afternoon. It is an enormous edifice. Most of the work in our writing and teaching, however, has gone into constructing its upper levels. But, with advancing technology inexorably shifting the balance of human statistical effort from processing and computation to thinking, we need to do some emergency work on the foundations to ensure that the whole structure stands steadily on the ground.

References

Bartholomew, D. (1995). What is Statistics? *J.R. Statist Soc. A*, **158**, Part 1, 1–20.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning and the search for causes: Do we need a probabilistic revolution after we have taught data analysis? In *Research Papers from The Fourth International Conference On Teaching Statistics, Marrakech 1994*, Ed. J. Garfield. MN: University of Minnesota.

Britz, G., Emerling, D., Hare, L., Hoerl, R. & Shade, J. (1997). How to Teach Others to Apply Statistical Thinking. *Quality Progress*, June 1997, 67–79.

Cox, D.R. (1992). Causality: Some Statistical Aspects. *J. R. Statist. Soc. A*, **155**, Part 2, 291–301.

Cox, D.R. (1993). Causality and Graphical Models. In *Bulletin of the International Statistical Institute, Proceedings of the 49th Session* Vol. 1, pp. 365–389. Voorburg: International Statistical Institute.

Cox, D.R. (1997). The current position of statistics: A personal view (with discussion). *International Statistical Review*, **65**, 261–290.

Dransfield, S.B., Fisher, N.I. & Vogel, N.J. (1999). Using statistics and statistical thinking to improve organisational performance. *International Statistical Review*, **67**, 99–150.

Falk, R. & Konold, C. (1992). The Psychology of Learning Probability.In *Statistics for the Twenty-First Century*, Eds. F. & S. Gordon, pp. 151–164. *MAA Notes*, Number 29. Washington, DC: The Mathematical Association of America.

Gail, M. (1996). Statistics in Action. *Journal of the American Statistal Association*, **91**, 1–13.

Gal, I., Ahlgren, C., Burrill, G., Landwehr, J., Rich, W. & Begg, A. (1995). Working group: Assessment of Interpretive Skills. *Writing Group Draft Summaries Conference on Assessment Issues in Statistics Education*, pp. 23–25. Philadelphia: University of Pennsylvania.

Garfield, J. & Ahlgren, A. (1988). Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education*, **19**, 44–63.

Hill, A.B. (1965). The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine*, **58**, 295–300.

Hoerl, R., Hahn, G. & Doganaksoy, N. (1997). Discussion: Let's Stop Squandering Our Most Strategic Weapon. *International Statistical Review*, **65**, 147–153.

Hoerl, R. & Snee, R.D. (1995). Redesigning the introductory statistics course. *CQPI Technical Report No. 130*, Center for Quality and Productivity Improvement, University of Wisconsin-Madison.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945–970.

Joiner, B. & Gaudard, M. (1990). Variation, Management, and W. Edwards Deming. *Quality Progress*, December, pp. 29–37.

Joiner, B. (1994). *Fourth Generation Management*. New York: McGraw-Hill Inc.

Konold, C. (1994). Understanding Probability and Statistics through Resampling. In *Proceedings of the First Scientific Meeting of the International Association for Statistical Education*, Eds. L. Brunelli & G. Cicchitelli, pp. 199–211. Perugia: University of Perugia.

MacKay, R.J. & Oldford, W. (1994). *Stat 231 Course Notes Fall 1994*. Waterloo: University of Waterloo.

Mallows, C. (1998). The zeroth problem (1997 Fisher Memorial Lecture). *The American Statistician*, **52**, 1–9.

Moore, D. (1990). Uncertainty. In *On the shoulders of giants: new approaches to numeracy*, Ed. L. Steen, pp. 95–137. Washington, DC: National Academy Press.

Moore, D. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review*, **65**, 123–165.

Pea, R. (1987). Cognitive Technologies for Mathematics Education. In *Cognitive Science and Mathematics Education*, Ed. A. Schoenfeld, pp. 89–122. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Pfannkuch, M. & Brown, C. (1996). Building on and Challenging Students Intuitions about Probability: Can We Improve Undergraduate Learning? *Journal of Statistics Education*, www.amstat.org/publications/jse/v4nl/pfannkuch.html.

Pfannkuch, M. (1996). Statistical Interpretation of Media Reports. In *New Zealand Statistical Association Research in the Learning of Statistics Conference Proceedings*, Eds. J. Neyland and M. Clark, pp. 67–76. Wellington: Victoria University.

Pfannkuch, M. (1997). Statistical Thinking: One Statistician's Perspective. In *People in Mathematics Education. Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia*, Eds. F. Biddulph and K. Carr, pp. 406–413. Hamilton: MERGA Inc.

Pfannkuch, M. & Wild, C. (1998). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. (unpublished manuscript).

Polya, G. (1945). *How to Solve It: A New Aspect of Mathematical Method*, Princeton: Princeton University Press.

Pyzdek, T. (1990). There's no such thing as a common cause. *ASQC Quality Congress Transactions–San Francisco*, pp. 102–108.

Resnick L. (1987). *Education and Learning to Think*. Washington DC: National Academy Press.

Resnick, L. (1989). Treating Mathematics as an Ill-structured discipline. In *The Teaching and Assessing of Mathematical Problem Solving, Volume 3*, Ed. R. Charles and E. Silver, pp. 32–60. Reston, VA: Lawrence Erlbaum Associates NCTM.

Schoenfeld, A. (1983). Episodes and Executive Decisions in Mathematical Problem-Solving In *Acquisition of Mathematics Concepts and Processes*, Ed. R. Lesh and M. Landau, pp. 345–395. New York: Academic Press.

Schoenfeld, A. (1987). Cognitive Science and Mathematics Education: An Overview. In *Cognitive Science and Mathematics Education*, Ed. A. Schoenfeld, pp. 1–31. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Shaughnessy, M. (1992). Research in Probability and Statistics: Reflections and Directions. In *Handbook of Research on Mathematics Teaching and Learning*, Ed. D. Grouws, pp. 465–494. New York: MacMillan.

Snee, R. (1990). Statistical Thinking and its Contribution to Quality. *The American Statistician*, **44**, 116–121.

Stewart, I. (1989). *Does God Play Dice?* London: Penguin Books.

Tversky, A. & Gilovich, T. (1989). The "hot hand": Statistical reality or cognitive illusion? *Chance*, **2**, 31–34.
Wild, C.J. (1994). On Embracing the "Wider View" of Statistics. *The American Statistician*, **48**, 163–171.

Résumé

Le présent article concerne les processus mentaux impliqués dans la pensée statistique prise dans un sens large, depuis la formulation de problèmes jusqu'á leur solution. Il tire ses sources de la littérature sur le sujet ainsi que d'entrevues auprès d'étudiants et de praticiens en statistique, conçues pour identifier leurs processus de raisonnement statistique. De ces entrevues, nous avons identifié un cadre conceptuel quadridimensionel applicable à la pensée statistique dans le domaine de la recherche empirique. Ce cadre est composé d'un cycle d'investigation, d'un cycle d'interrogation, de types de pensée et de dispositions. Nous avons amorcé la caractérisation de ces processus par des modèles pouvant servir de base à la création d'outils ou cadres intellectuels aidant la résolution de problèmes. Des outils de ce types pourraient complémenter les modèles mathématiques déjà utilisés en analyse en plus de couvrir certains aspects de la recherche statistique que les modèles mathématiques ne peuvent pas satisfaire, particulièrement les aspects associés à la synthèse des types contextuel et statistique de compréhension. L'élément central apparaissant dans les définitions de la pensée statistique ayant fait l'objet de publication est celui de la "variation". Nous discutons aussi le rôle de la variation dans l'approche statistique de problèmes pratiques, incluant la recherche de causes.

Discussion

T.M.F. Smith

University of Southampton, UK

The topic of this thoughtful paper is the modes of thinking that underlie efficient empirical enquiries. The authors, henceforth WP, title the paper statistical thinking but, in fact, the topic is much wider and embraces the thinking and learning processes that form the basis of all scientific investigations. If throughout the paper you replace the word statistical by scientific, or more generally systematic, then very little would have to be changed. The models represented by Fig. 1–10 remain valid and provide an excellent framework for discussing scientific thinking. The only term that is missing is creativity, the mode of thinking that leads to the greatest advances in science. But creativity is impossible to teach and WP are concerned with processes that can, and should, be taught, and that form the basis of all scientific enquiry from the most humdrum to the most creative.

A key word that is repeated throughout the paper is context. All thinking, including statistical thinking, takes place within a context which needs to be understood by all who are conducting the enquiry. WP make a strong case for multi-disciplinary teams who challenge assumptions, bring different backgrounds to enquiries, and generate the synthesis of ideas that lead to understanding and future progress. WP argue that statisticians should frequently be members of these teams, but they include examples where conventional statistical methods, as currently taught, do not add much value to the enquiry. To justify team membership, statisticians must be seen to add value that cannot be added by scientists from other disciplines. In the examples from both students and consultants this is not always the case. In discussing context we learn that "Of all our statisticians, however, only the one operating in quality improvement seemed to use tools (e.g. cause and effect diagrams and process mapping techniques) to analyse the nature of the problem itself." How many statisticians have been taught to use these tools? Later on, in the discussion of causation, we find the statistician as Cassandra, the harbinger of doom, warning that correlation does not imply causation and that results may be meaningless if they could have arisen by chance. Statistical thinking can be very negative at times and can stifle imaginative enquiry. Is that really our main contribution? Sometimes yes, since there is a danger in jumping too quickly to conclusions in areas such as medicine. But this danger may not exist in other areas and statisticians must be aware of the context in which they are operating

and modify their approach accordingly. Trying to apply statistical methods uniformly without regard to context gives statistics a bad name. But how do you teach context?

Where does statistical thinking add value? To me the greatest contributions of statistics to scientific enquiry have been at the planning stage. Our systematic methods for designing experiments and surveys, based on randomisation to avoid biases, have been of major importance. The second great contribution is mathematical, it is the application of probability theory to intrinsically random processes, such as quantum theory, genetics and natural selection. Applied probability modelling, and the related thinking, have transformed large areas of science and the very way that all of us perceive the world. The deterministic models of much of physics are no longer the only alternative to religious mysticism. Although statisticians have helped in the development of applied probability models the main work has been by mathematicians working either within a science or with scientists. Many of the methods of current interest in statistics bear the names of scientists and engineers, Gibbs sampling, Kalman filters etc., not of statisticians. The exception again is medicine, where Cox models have transformed population epidemiology.

WP quote many statisticians who claim that variation or variability is the centrepiece of statistics, and that thinking about variability is the main message of statistics. This is a grandiose claim since the whole of science is concerned with variation. Where there is no variation there is no scientific problem. The problem is the nature of that variation. The physical sciences have made dramatic advances by modelling deterministic variation, and if statistics is to "give us a means of understanding a variation-beset world", as WP would like, then the nature of statistical variation needs more careful specification.

Statistical variation is random variation and it is the nature of the randomness in different situations, and the uncertainty that this generates, that needs exploring. Where there is intrinsic randomness and a given applied probability models is generally accepted, then likelihoods have real meaning and Fisherian or Bayesian methods can be justified. The randomness of measurement processes superimposes itself on all investigations, statistical or deterministic. Statisticians have a contribution to make in designing and analysing measurement experiments, but most scientists ignore this contribution and concentrate either on designing new instruments that give more precise measurements or on increasing sample sizes. Randomisation in the design of a study induces a probability distribution which may be used in the interpretation of results. The alternative, which is to employ models that capture the properties of the data and the design, is harder to justify. In what sense is there a random process generating the data beyond that of the randomisation? The greatest problems arise in observational studies of non-random samples where there is no intrinsic random process that generates the data. As WP say "most real systems are enormously complex with variation in innumerable components". The scientific strategy is to search for regularities, but these regularities may be specific to the context in which the data were generated. For example, in economics the effect in the growth period of a cycle may be different from the effect in a downturn. In opinion polling, one of the few areas in statistics where the results can be validated, the accuracy of the polls appears to be context specific. There are many possible sources of error, of which the non-sampling errors are the most important, not the choice of quota or random sampling. In some elections the combined effects may be small while in others they all move in the same direction giving wildly inaccurate results, as in the UK in 1992. What is required is studies that measure the interactions between the study variables and the context variables, followed by methods for forecasting the values of the context variables. If the interactions are small then regularities in the study variables are likely to persist, if they are not small then the regularities are local and of more limited use. Studies of this type are rarely performed due to their cost and complexity and statisticians frequently hide the difficulties by fitting a context specific model and then adding an error term which purports to measure unexplained heterogeneity. Since this error term varies over individuals they then assert that it can be treated as if it were random. This added randomness, which features in so much of statistical modelling,

is hard to justify. A common argument is to say that if the individual were drawn at random then the error would be random, but this is random sampling error not random process error. It was this type of argument that caused problems for WP's students and they were right to be troubled. If as statisticians we claim to have "correct" methods of model specification then in many areas we will be ignored and our valuable contributions to design and measurement will be lost. In Smith (1997) I argue that we must recognise the limitations of statistical thinking and not make grandiose claims for the universality of our discipline based on the fact that variation is universal.

The real challenge of this paper is to teachers. How can the ideas of scientific thinking be incorporated into the teaching of statistics? If problems are context specific then would it be better to teach statistical thinking to applied scientists, who should be familiar with at least one context, than to mathematicians who work in a context free zone? Certainly I find teaching engineers a more rewarding experience than teaching mathematicians because they are problem driven. Perhaps mathematicians should be forced to study an applied science before they embark on a statistics course.

Additional reference

Smith, T.M.F. (1997). Social surveys and social science. *The Canadian Journal of Statistics*, **25**,1, 23–44.

Discussion: What Shall We Teach Beginners?

David S. Moore

Department of Statistics, Purdue University, West Lafayette, IN, USA

Wild & Pfannkuch provide much food for thought in their extensive and valuable discussion of statistical thinking (Section 2) and of variation and randomness (Section 3). To keep my discussion within bounds, I shall raise questions related to just one issue: *What in the way of statistical thinking ought we to attempt to teach beginners in a first statistics course*?

Statistical Thinking for Beginners

I fully agree that "development of a theoretical structure with which to make sense of reality" is an important goal of instruction, and that "the thinking and problem solving performance of most people can be improved by suitable structured frameworks". I most emphatically agree that *mathematical* structure is inadequate (even when it exists and can be grasped by students) as a basis for understanding and doing statistics in practice. Thus the working teacher should ask what kinds of structures she can offer students as frameworks for statistical thinking. Wild & Pfannkuch offer help.

Not, however, help specifically intended for teachers. Their discussion of "the thought processes involved in statistical problem solving" is quite complex. This surely reflects the actual complexity of these processes, but the resulting sets of cycles and epicycles bear a daunting resemblance to Ptolemaic astronomy. Hence my most important overall comment: *Beginning students need a more selective introduction to statistical thinking.* Beginners often lack intellectual maturity; they often lack the context knowledge needed for full statistical problem-solving; we know that beginners do not find elementary statistics an easy subject, so that we must be hesitant to add to their cognitive

load. I find that industrial statisticians, whose teaching deals with mature and motivated company employees, often have quite unrealistic notions of the extent and sophistication of content that typical undergraduates can assimilate.

A fair response to Wild & Pfannkuch's mild criticism of the ASA/MAA committee's brief discussion of statistical thinking is therefore that the committee's statement appears in the explicit context of a first course in statistics and so is quite appropriately "only a subset" of what statisticians understand about statistical thinking. I believe that, in fact, the ASA/MAA committee has a good grasp of the elements that are most important in teaching beginners. Their longer discussion (Cobb, 1992) is very much worth reading.

Let me offer some personal opinions on the kinds of statistical thinking that any first course ought to provoke in students. First, important as they are, issues of measurement and of problem formulation require substantial grounding in the details of a specific context. Beginners can be expected to deal with only elementary levels of these aspects of statistical thinking. That is, I think current practice shows good judgment in generally assuming that variables are satisfactorily measured and that the problem posed is indeed one that clinicians or ecologists or civil engineers find worth investigating.

Second, we can start by mapping more detailed structures for the "Data, Analysis, Conclusions" portion of the investigative cycle, that is, for conceptual content currently central to elementary instruction. Here is an example of such a structure:

> When you first examine a set of data, (1) begin by graphing the data and interpreting what you see; (2) look for overall patterns and for striking deviations from those patterns, and seek explanations in the problem context; (3) based on examination of the data, choose appropriate numerical descriptions of specific aspects; (4) if the overall pattern is sufficiently regular, seek a compact mathematical model for that pattern.

This "suitable structured framework" for thinking supports yet more detailed frameworks in more specific settings. Wild & Pfannkuch rightly emphasize that "subcycles are initiated within major cycles". Students learn in each setting what standard graphs may be helpful, what typical overall patterns to look for, what numerical measures and mathematical models may be appropriate. For example, faced with data on a single quantitative variable, we can expect students to choose wisely among at least histograms and stemplots, to look for the shape, center, and spread of the displayed distribution, to weigh the five-number summary against $\bar{x}$ and s as a description of center and spread, and to consider standard density curves as possible compact models. Structures such as these are specific enough to guide beginners, yet general enough to be genuinely helpful. They are not, of course, simply recipes on the order of the algebraic recipes that filled our first courses before the blossoming of technology. In particular, data always have a context, and students must learn by (rather simple) examples and experience to, as Wild and Pfannkuch nicely put it, pursue a synthesis of context knowledge and statistical knowledge.

I hope it is clear that this discussion does not indicate disagreement with Wild & Pfannkuch's principles. I simply want to illustrate the work that teachers must do to make explicit the aspects of statistical thinking that we will emphasize in helping beginners learn. I also want to emphasize the need to be selective by reminding readers how much explicit content lies behind the structures.

Finally, though, I agree with Wild & Pfannkuch's implicit judgment that our current instruction is too narrow. We have done well to place much more emphasis than in the past on the design of data production and on exploratory analysis of data, and to situate formal inference more solidly in a setting shaped by design and exploration. Yet, as technology continues to automate the details, we must continue to ask what broader intellectual skills our students should carry away from a modern introduction to the science of data. I make some preliminary comments in Moore (1998), whose title "Statistics among the liberal arts" suggests the status I think statistical thinking deserves.

Randomness and Variation for Beginners

Wild & Pfannkuch offer an excellent and thoughtful discussion of variation, randomness, and causation. These are all issues that we must address in teaching beginners, and the discussion here should be helpful to any teacher. I want in particular to endorse their discussion of "Modelling Variation" with its emphasis that the "random" portion of statistical models is our way of describing unexplained individual variation and that "We have no idea whether this variation really is random." We would be wise, I think, to continue to reduce the place of formal probability in teaching statistical practice to beginners. I find that "unexplained individual variation" is clearer to students (because more concrete) than "random variation". Elementary inference procedures often assume that this variation is roughly described by a normal distribution. Normal distributions are not primarily "probability distributions" in this setting, but simply compact descriptions of the overall pattern of some sets of data.

What, no Bayes?

No doubt others will make the point that, for all the thought and empirical investigation behind it, the framework offered by Wild and Pfannkuch is itself "only a subset" of statistical thinking. Omission of the Bayesian paradigm for organizing statistical problems is striking, for example. That omission will bring no complaints from me, as I think (Moore, 1997) that there are compelling reasons to avoid Bayesian inference when we teach beginners. Given the broader aims of this article and the prominence of Bayes methods in current research literature, however, it would be helpful if Wild & Pfannkuch commented on where these ideas fit. Did the statisticians they interviewed show no traces of formal Bayesian thinking?

Conclusion

I am sure that readers will agree that Wild & Pfannkuch have made a stimulating contribution to our continuing reexamination of the nature of statistical thinking. I hope that we will continue to reexamine our teaching in the light of this and other discussions.

Additional references

Cobb, G. (1992). Teaching statistics. In *Heeding the Call for Change: Suggestions for Curricular Action*, Ed. L.A. Steen, pp. 3–43. Washington, D.C.: Mathematical Association of America.

Moore, D.S. (1997). Bayes for beginners? Some reasons to hesitate. *American Statistician*, **51**, 254–261.

Moore, D.S. (1998). Statistics among the liberal arts. *J. Amer. Statist. Assoc.*, **93**, 1253–1259.

Discussion: Statistical Thinking in Practice

N.E. Breslow

Department of Biostatistics, Box 357232, University of Washington, Seattle, WA, 98195-7732, USA

When queried by students seeking a general recipe on how to apply statistical methods to any particular area of application, I usually punt. This, I explain, is part of the "art" of statistics whose practice is perfected through immersion in the subject matter area, through careful study of statistical applications in that area and, if one is lucky, through apprenticeship under a master practitioner. Wild

& Pfannkuch's (hereafter W&P) stimulating contribution mandates a re-evaluation of this position. Their arguments suggest that the apprenticeship period may be shortened through a structured approach to statistical reasoning that is applicable across subject matter disciplines. While it is not clear that their proposals will fully satisfy student desires for a formal set of rules, broader discussion of these issues among statisticians, their students and their clients is certainly long overdue. The comments that follow focus on three issues: (i) consideration of variation; (ii) a plea for more rigorous statistical thinking in clinical and epidemiological research ; and (iii) the need to couple the structured reasoning approach with concrete examples of its application.

I agree completely with the prevailing view that consideration of variation is central to statistical thinking. Once the study goals are defined and response variables are identified, the next step is to consider sources of variation in those variables. This goes well beyond the identification of potential explanatory variables for inclusion in a regression equation. It is important to recognize that there may be multiple sources and multiple levels of unexplained (random) variation that could or should be taken into account. Some of this is alluded to under the subsection titled "Validity & reliability" in W&P's Fig. 10. At its most basic, consideration of variation entails the realization that the number of primary sampling units, rather than the total number of observations, is often the critical determinant of statistical precision. The fundamental statistical concepts taught in courses in experimental design and components of variance analysis are essential to developing student awareness of the multiplicity of sources of (random) variation, an awareness that leads to good statistical thinking. It is important that these concepts and courses continue to receive appropriate emphasis in the modern, computationally oriented statistics curriculum.

As a medical statistician, I am appalled by the large number of irreproducible results published in the medical literature. There is a general, and likely correct, perception that this problem is associated more with statistical, as opposed to laboratory, research. This undoubtedly contributes to the low esteem in which the statistical profession is held by some and a general lack of confidence in statistical investigations. Laboratory scientists often take pains to ensure that their results are reproducible, at least in their own laboratories, before submitting them for publication. Epidemiologists and clinical investigators are less likely and less able to impose such constraints. This is partly because they work with human subjects and partly due to the observational nature of much of their data. I am convinced, however, that results of clinical and epidemiological investigations could become more reproducible if only the investigators would apply more rigorous statistical thinking and adhere more closely to well established principles of the scientific method. While I agree with W&P that the investigative cycle is an iterative process, I believe that it works best when it is hypothesis driven. Thus I would put "hypothesis generation" or perhaps "hypothesis specification" at the beginning of the cycle, before collection of the data, rather than afterwards. Protocols for randomized clinical trials generally do state hypotheses explicitly. The fact that many of these still yield irreproducible results has more to do with their multiplicity, with small sample sizes, with subgroup analyses and with consideration of endpoints that were not explicitly defined at the outset. The epidemiology literature is replete with irreproducible results stemming from the failure to clearly distinguish between analyses that were specified in the protocol and that test the *a priori* hypotheses whose specification was needed to secure funding, and those that were performed *post-hoc* as part of a serendipitous process of data exploration.

Statisticians have an important role to play as referees, both of their colleagues' work before publication and as anonymous reviewers afterwards. Their greater involvement in editorial decisions should help to reduce the level of "noise" in the medical literature. In an effort to improve scientific rigor, many clinical journals now make a formal statistical review part of the editorial process. This involves more than the application of standard criteria for statistical reporting (e.g., Bailar & Mosteller, 1988), important as these may be. A healthy skepticism and the imagination needed for alternative explanations, dispositions mentioned by W&P in Fig. 7, are desirable qualities in

a referee. I look forward to W&P's promised future work on the interpretation of information in statistical reports, which is an excellent arena in which to hone one's skill at statistical thinking.

W&P's Fig. 1–10 outline a rather abstract, albeit progressively less so, framework for the planning of statistical investigations and the conduct of statistical reasoning. While they provide a useful and stimulating basis for discussion among professional statisticians, I am less sure regarding their potential to "reach" the apparent intended audience: students of statistical science with no prior experience with applications. I am sure W&P would agree that for teaching purposes it is essential to supplement the abstract presentation of guidelines and concepts with examples of statistical thinking in practice. This need not necessarily imply student participation in projects or case studies that require data analysis. There is a place also for study and discussion of examples from the literature, which allows a wider array of issues to be covered in the limited time available. In this spirit, I conclude my comments with two examples of statistical thinking that represent my earliest and latest forays into the medical literature. The first example was used for several years by my faculty colleagues to illustrate concepts of statistical variation in introductory classes for health sciences students. Both examples are statistical critiques and re-evaluations of articles published in the medical literature.

One of the first facts I learned upon becoming statistician for the Children's Cancer Group was that the presence of a large fraction of lymphocytes in the bone marrow cell population often presaged relapse in patients with acute lymphocytic leukemia. Thus it was with some surprise that my colleagues and I read an article that called into question the then standard criteria for evaluation of response to therapy (Skeel *et al.*, 1968). This placed children with an elevated bone marrow lymphocyte count (BMLC) in the good response category. The authors reported that patients whose BMLC remained below 20% throughout remission tended to have shorter remissions, as determined by the fraction of immature blast cells in the marrow, compared with patients whose BMLC exceeded this threshold at least once. Turning to similar data from an ongoing clinical trial, in which bone marrow exams were performed at six week intervals during remission, I sought to demonstrate that this result most likely was a statistical "artifact" resulting from measurement error and the tendency of the maximum of a random sequence to increase with its length (Breslow & Zandstra, 1970). Sure enough, when we separated our patients into three groups depending on the maximum BMLC during remission, those with the highest maxima had the longest remissions and those with the lowest maxima the shortest. When patients were divided into three groups based on the average BMLC during remission, the results were reversed. Nowadays sophisticated methods of longitudinal data analysis could be used to predict future relapses from the level or change in BMLC during remission.

More recently, I was intrigued when a "nuclear morphometric score" was reported to predict relapses in children with the childhood kidney cancer known as Wilms tumor (Partin *et al.*, 1994). Nuclear morphometry is an image processing technique designed in part to overcome the subjectivity of the pathologist in the grading of cancer cells. In this particular example, slides of tumor tissue from a case-control sample of 42 patients who had relapsed (cases) and 40 who remained disease-free after treatment (controls) were processed by the imaging device. It identified 150 nuclei on each slide and calculated 16 numerical shape descriptors, including the diameter, roundness factor, degree of ellipticity, etc., for each nucleus. Then the commercial software summarized each of the 16 frequency distributions using 17 statistical measures including mean, variance, skewness, kurtosis, range, etc., for a total of $16 \times 17 = 272$ measurements per child. The software next selected the two measurements that showed the greatest difference between cases and controls, which turned out to be the skewness of the nuclear roundness factor and the minimum value of ellipticity as measured by the "feret diameter" method. After calculating the linear discriminant using these two measurements plus age, a known prognostic variable, the resulting "morphometric score" was demonstrated (on the same data) to be associated with outcome at a high level of statistical significance. Needless to say, when I re-evaluated this morphometric score using a prospectively acquired sample of 218 patients of whom 21 relapsed,

the coefficient for age remained about the same (and significant) whereas the coefficients for the two morphometric variables were close to zero (Breslow *et al.*, 1999). The discussion emphasized the importance of using appropriate statistical methods, such as bootstrapping, for obtaining unbiased estimates of sensitivity and specificity based on linear discriminant analyses. More on the positive side, I suggested that Bayesian variable selection techniques might have a role to play in the next generation of software for nuclear morphometry, whose future as a diagnostic tool at this point is somewhat uncertain.

Additional references

Bailar, J.C. & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals. *Annals of Internal Medicine*, **108**, 266–273.

Breslow, N. & Zandstra, R. (1970). A note on the relationship between bone marrow lymphocytosis and remission duration in acute leukemia. *Blood*, **36**, 246–249.

Breslow, N.E., Partin, A.W. *et al.* (1999). Nuclear morphometry and prognosis in favorable histology Wilms tumor: A prospective re-evaluation. *Journal of Clinical Oncology*, **17**, 2123–2126.

Partin, A.W., Yoo, J.K. *et al.* (1994). Prediction of disease-free survival after therapy in Wilms tumor using nuclear morphometric techniques. *Journal of Pediatric Surgery*, **29**, 456–460.

Skeel, R.T., Henderson, E.S. & Bennett, J.M. (1968). The significance of bone marrow lymphocytosis of acute leukemia patients in remission. *Blood*, **32**, 767–773.

Discussion: Development and Use of Statistical Thinking: A New Era

Ronald D. Snee

Sigma Breakthrough Technologies, Inc., 10 Creek Crossing, Newark, Delaware 19711, USA
E-mail: RDSnee@AoL.Com

Statistical research, practice, and education are entering a new era, one that focuses on the development and use of statistical thinking. The advancement of computer technology and globalization of the market place are greatly affecting the role of statistics and statisticians. The ubiquitous nature of personal computers and statistical software have thrust statisticians into strategic "thinking" roles as well as operational roles of solving problems with statistical tools. Market globalization has forced companies to improve to stay competitive. Many of those who implement the teachings of W. Edwards Deming (1986) and use the Six Sigma approach (Hoerl, 1998) are working in strategic roles of developing and implementing statistical approaches to process and organizational improvement. In order for an endeavor to be effective it needs activity at three different levels: strategic, managerial, and operational (Snee, 1990). Statistical thinking provides the discipline of statistics with this strategic component, which has been largely ignored until recently. It is also important to note that strategy is the principal work of top managers of organizations.

Joel Barker (1985) points out that the need for a new approach, a new paradigm, is apparent when the paradigm in use cannot solve the existing problems. I believe that we are in such a situation today. The rapid advancement of computer technology and global competition has created the need for a new way of using statistics. This new paradigm will define a new era. Also, few are satisfied with the way statistics is taught, particularly at the introductory level and to nonstatisticians. Wild & Pfannkuch have recognized this new era and need for a new view of statistics and have developed models for "statistical thinking in empirical inquiry". Their approach is a welcome contribution. I

applaud their use of models. My comments focus on needed model refinements, the elements of statistical thinking and some implications of the authors' research.

Developing Useful Models

George Box and other leaders of the statistical profession have admonished us that models must be simple and parsimonious if they are to stimulate insight and be useful. With this thought in mind, I believe that the authors' model shown in Fig. 1 must be considerably simplified if it is to have broad effectiveness. The current model has 4 main components and 27 sub-components. Four to seven key elements would be a good target. The authors and others who work in this area would do well to also identify a small set of tools to integrate with the elements, or steps, in the model. Roger Hoerl and I (see Hoerl & Snee, 1995) and others have learned that teaching tools and models separately only creates confusion. A good example of this integration is the Six Sigma process improvement methodology (Hoerl, 1998). Here, a 4-step approach (measure, analyze, improve, control) and 8–9 tools are integrated to produce a methodology that is both effective and easy to understand and use. I believe that such an effort would be a good direction for further development of the authors' model for the use of statistical thinking in empirical inquiry.

It is important to recognize that George Box's "Iterative Nature of Experimentation" and "Inductive —Deductive Learning" models are also useful for describing the interplay between context and statistics as shown in the authors' Fig. 2 (Box, Hunter & Hunter, 1978). Unfortunately, focusing on the identification of theories to be checked by experimentation, and data collection to drive process improvement is greatly underutilized in the teaching and practice of statistics.

Elements of Statistical Thinking

The authors' build statistical thinking into their model in "Dimension 2: Types of Thinking". They note the importance of "process" and "need for data". Some discussion of these elements of statistical thinking is needed. The key elements of statistical thinking are process, variation, and data (Figure 1). These elements are connected in the following way. First, all activity, work or otherwise, is a process. A process is defined as any activity that converts inputs into outputs. The problems of empirical inquiry are associated with one or more processes. Hence, the process or processes involved provide the "context" for the statistical work.

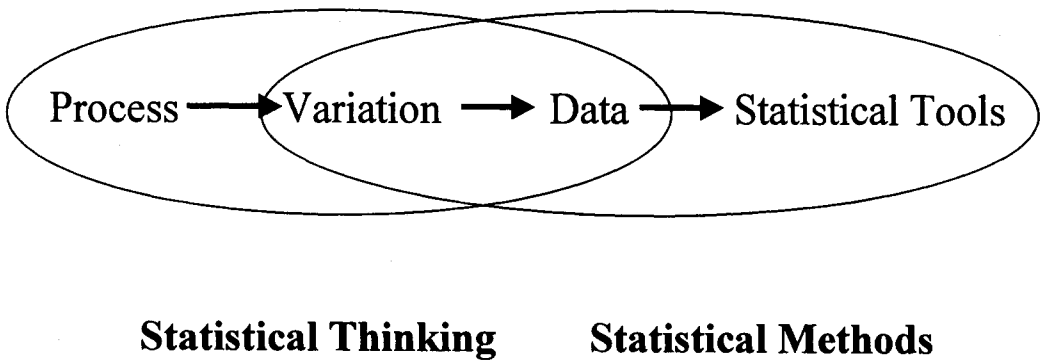

Figure 1. *Relationship between statistical thinking and statistical methods*

Process improvement and problem solving get complicated quickly because all processes vary. Our need to deal with the variation leads us to make measurements as a way of characterizing the process being studied; and creating a basis (numerical) for comparison. The result of the measurement process is data. Hence, it is the need for measurement (a basis for comparison) that produces data. Now the data vary for two reasons as shown in the authors' Fig. 6; the process varies and there is variation produced by the measurement and data collection system. We use statistical tools to analyze the process data, which vary because of the process and measurement system. The elements of statistical methods are thus variation, data and statistical tools (Fig. 1). It is important to note that the Six Sigma approach utilizes both statistical thinking and statistical methods (i.e. process, variation, data and statistical tools).

The broad utility of the "process concept" is clear when we think of the following examples of process: The U.S. Budget process, the peace process, the educational process, etc. Other examples of processes are the political, learning, cell growth, data collection, surgery, and mail delivery processes. The list goes on and on. The process concept extends beyond manufacturing and business processes to almost all, if not all, areas of science and engineering. The broad utility and provision of context for empirical inquiry provides a useful fundamental for statistical thinking. Statistical thinking as defined by the American Society for Quality (1996) has three fundamentals: All work occurs in a system of interconnected processes, variation exists in all processes, and understanding and reducing variation are key to success. Understanding variation enables us to change the average level of the process as well as reduce process variation.

The importance of variation to statistical thinking becomes clearer when we recognize that variation is present in processes whether or not we have data on the process. This observation leads us to the conclusion that we can use statistical thinking without data. For example, we know that, in general, decreasing the variation of process inputs decreases the variation of process outputs, that students have different backgrounds and different learning methods, that many products are used differently by different customers creating different customer segments, etc. Deming (1986) emphasized that reducing the number of suppliers for a process decreases the variation in process output. Hence, many companies have instituted efforts to significantly reduce the number of suppliers.

One can more deeply understand variation when one thinks about working as a statistician and not being able to collect data. Can statisticians make a contribution in this situation? I believe that the answer is yes! Helping managers understand the importance of variation in their processes and organization is of major importance as emphasized by Deming and others. While data are critical to the effective use of statistical thinking, and should be used whenever possible, data are not absolutely essential to the use of statistical thinking.

So Where Do We Go From Here?

What are the next steps in the development of "statistical thinking for empirical inquiry?" I noted earlier directions for further model development: model simplification and the integration of tools with the model. There are others. First we must recognize that understanding variation is a core competency of statisticians (Snee, 1999). This defines the uniqueness and competitive advantage of statisticians and others who focus on the use of statistical thinking and methods. What data are relevant and how to collect good data are important considerations and might also be considered core competencies of statisticians. Understanding variation is, however, a much more important core competency whose successful application is enhanced by understanding what data are relevant and how to construct proper data collection methods. If there was no variation, there would be no need for statistics and statisticians. This leads us to the conclusion that we must focus statistical research, education, and practice on understanding variation and using the resulting knowledge to improve the performance of processes.

A focus on statistical thinking in statistical education will require changes in both the "content" and "delivery" of courses. The use of projects noted by the authors is a delivery issue. Projects are effective because they enable students to "learn by doing" and to see and experience the utility of statistical thinking and methods in their field of interest. Projects also help us teach students with widely varying backgrounds and interests in the same class with the project providing the context and connection of the statistical methods to their field. Projects have been used most widely in introductory courses but are equally applicable in applications and methods courses at all levels and provide useful context for courses on statistical theory. The research work of many leading statisticians started with the need to solve real problems e.g. R.A. Fisher, G.E.P. Box, J.W. Tukey, W.G. Cochran, *et al.*

The focus on utilizing statistical thinking in statistical education and practice will identify new areas for research. This is already happening with the rapidly growing use of the Six Sigma process improvement approach which properly focuses on improving the processes that the organization uses to do its work. Process measurement and methods for deciding on whether to focus on methods for process control, process improvement or process reengineering have already been identified as important research opportunities.

I see the development of statistical thinking as the next step in the evolution of the field of statistics. In the 1950's and 60's and earlier, the focus was on tools and methods to solve problems. In the 1960's and 70's the focus turned to the mathematical aspects of statistics. In the 1980's and 90's the emphasis moved to data analysis. In Fig. 1 we see that we are now moving further upstream from a focus on statistical tools and data to focusing on variation and the process that produced the data and its associated variation. When you focus simultaneously on all four elements: process, variation, data and statistical tools, you have a richer, more robust and effective statistical approach.

Additional references

ASQ Statistics Division (1996). *Statistical Thinking*, Special publication available from American Society for Quality, Quality Information Center, Milwaukee, WI (1-800-248-1946). This publication contains several references and examples on statistical thinking.

American Society for Quality (1996). *Glossary and Tables for Statistical Quality Control.* Milwaukee, WI: American Society for Quality.

Barker, J.A. (1985). *Discovering the Future: The Business of Paradigms.* St. Paul, MN: I.L.I. Press.

Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters.* New York, NY:, Wiley Interscience, John Wiley & Sons.

Deming, W.E. (1986). *Out of the Crisis.* Cambridge, MA: MIT Center for Advanced Engineering Study.

Hoerl, R.W. (1998). Six Sigma and the Future of the Quality Profession, *Quality Progress*, June, pp. 35–42.

Hoerl, R.W. & Snee, R.D. (1995). *Redesigning the Introductory Statistics Course.* University of Wisconsin Center for Quality and Productivity Improvement, Report No. 130, Madison WI.

Snee, R.D. (1990). Statistical Thinking and Its Contribution to Total Quality. *The American Statistician*, **44**, 116–121.

Snee, R.D. (1999). Statisticians Must Develop Data-Based Management and Improvement Systems as well as Create Measurement Systems, Comment on "Using Statistics and Statistical Thinking to Improve Organizational Performance" by S.B. Dransfield, N.I. Fisher, and N.J. Vogel, *International Statistical Review*, **67**. 99–150.

Discussion: Learning to Think Statistically and to Cope With Variation

Rolf Biehler

Universtität Bielefeld, Institut für Didaktik der Mathematik (IDM), Postfach 100131, 33501 Bielefeld, Germany. E-mail: rolf.biehler@uni-bielefeld.de

It is a widely shared concern that students should learn to think statistically. However, it is still unclear what we mean exactly by "statistical thinking" and, secondly, it is unclear how we can organize adequate learning processes supporting its development. Wild & Pfannkuch rightly criticize that the advice "let the students do projects" is not enough. Learning statistical thinking just by "osmosis" cannot be the answer. The authors are convinced that the "... cornerstone of teaching in any area is the development of a theoretical structure with which to make sense of experience, to learn from it and transfer insights to others." (p. 224).

This is the background motivation for the authors' fresh look at the question "What is statistical thinking?" They develop a general model of statistical thinking, based on existing literature, but also on in-depth interviews they did with students and statisticians as well as on experience from working as consultants in statistics.

My comments will be related to three aspects: the framework the authors propose, its relevance for organizing teaching and learning processes, and the importance of the notion of variation for rethinking statistics education.

The Framework for Statistical Thinking

The paper differs from others that narratively characterize essential elements of statistical thinking by developing a more formal four-dimensional framework for statistical thinking which is diagrammatically represented (Fig. 1), and further elaborated into a hypertext-like structure (Fig. 8–10). This form of representation is a valuable result of their intention to develop thinking tools that can be used in education and consultancy, and a merely narrative text would not be sufficient for this purpose.

The PPDAC cycle (dimension 1, problem $\longrightarrow$ plan $\longrightarrow$ data $\longrightarrow$ analysis $\longrightarrow$ conclusion) is similarly used by other authors for structuring the *process* of statistical inquiry into stages on a macro level. Wild & Pfannkuch add to this the process dimension 3 (interrogative cycle), which is more a psychological micro level description of problem solving stages that occur within every stage of the macro level process, and whose characteristics are not specific for statistics. The dispositional dimension 4 that more or less controls the quality of the interrogative cycle is related to dimension 3.

Adding these dimensions permits the authors to distinguish different quality levels of thinking. Different metacognitive qualities can now be distinguished, such as the ability to critically questions one's own inferences instead of "jumping to conclusions", different information-gathering strategies (searching in one's own memory or seeking in external resources) and different abilities for imagination and generation of ideas. The dispositional dimension 4 identifies general scientific attitudes and belief systems that are very important for higher quality statistical thinking, such as skepticism, openness, engagement, or a propensity to seek deeper meaning. These aspects can be well used for diagnosing and observing the quality of students' thinking and widen the statistics' educators awareness beyond the specificities of statistics.

A very interesting idea is to consider the PPDAC diagram as a kind of hypertext or underlying tree structure, where we can zoom into all nodes to see more detail. The authors demonstrate this zooming process convincingly by zooming into the *PPDAC plan/plan collection of data/measurement* node. *Measurement* is now on depth level 3. Fig. 10 represents 3 further levels. One path could be *measurement/anticipate problems/validity & reliability*. At the bottom of this path, we find 6 worry questions, which everybody should habitually ask while working on statistical problems.

This seems to indicate a further research program of the authors, namely to elaborate the PPDAC cycle in all respects up to the deeper levels. In Fig. 6 we find about 30 questions. If we assume a similar structure behind all the other nodes, we have to multiply this by 5 (level1) and 4 (level 2) and 4 (level 3), that would result in 2400 (!) questions. The resulting complexity could certainly only be managed by implementing these levels in a real computer-based hypertext.

This, obviously, constitutes a problem. From a theoretical point of view, one can ask whether it would be helpful to describe statistical thinking and statistical expertise more and more detailed along these lines, or whether there are alternatives. For further elaboration, I think it would be good advice to look into what Artificial Intelligence and Cognitive Science might offer about analyzing the thinking of experts. Expert thinking is sometimes characterized by reducing complexity, by pattern recognition strategies, by relying on prototypical situations to which actual situations are compared. That is of course different from working through checklists of many questions, although such checklists could still be considered a valuable starting point for novices in the field.

Since the beginning of the eighties we have seen an intensive discussion on "statistical expert systems". Even if the conclusion of this debate may be that people cannot be replaced by expert systems, the discussions contains a lot of thoughts about "what is statistical thinking and statistical knowledge" and how can we represent it. The authors have not yet exploited this debate for their purposes. Moreover, the debate on developing computer-based support systems for users of statistics instead of replacing people by machines are also relevant for further elaborating Wild & Pfannkuch's model and objectives. It may well be a result of further thinking and experience that effective statistical thinking in practice can, at least in some domains, be only performed with more advanced computer support than just data analysis systems. Mere mental thinking tools may not be sufficient.

The Model of Statistical Thinking and Learning to Think Statistically

Wild & Pfannkuch refer to various developments in mathematics education (Polya, 1945; Schoenfeld, 1987) who suggest the use of thinking tools such as lists of worry and trigger questions in order to improve thinking processes. Polya's book is a prototypical example, where he suggests that starting to ask these questions would be a first step in developing a new problem solving "mental habit". Garfield & Gal (1999) refer to similar approaches with regard to developing students' critical appraisal of statistical reports. Wild & Pfannkuch refer to similar developments in quality improvement.

The advantage of these systems of questions is their relative simplicity. Wild & Pfannkuch's system has a much higher degree of complexity and it is difficult to imagine that this could be implemented as a "mental" tool without any substantial reductions. Nevertheless, their comprehensive model may provide a sounder basis for pedagogically motivated reductions or transformations, and we may think of developing learning and teaching programs including computer-supported aids that implicitly rely on Wild & Pfannkuch's model without aiming at conveying the complexity in a more direct way to students. In this sense, the range of applications of their model may be much wider than the authors sketch themselves in their paper, namely mainly referring to the use as thinking tools.

Another aspect has to be added. Recent learning theories in mathematics education regard learning as a process of enculturation, as learning to participate in a certain cognitive and cultural practice, where the teacher has the role of a mentor and mediator. This is especially true with regard to

statistical thinking, which maybe better thought of as a "statistical culture"—including value and belief systems, habits of asking and writing, general scientific dispositions, and specificities of statistics such as the appreciation of data etc. The kind of collaborative and communicative processes that are stimulated in the classroom would be very important for statistical enculturation. The extent to which the teachers' behavior and knowledge represents an authentic model of this culture would be important. Wild & Pfannkuch's analysis could be re-interpreted as a valuable analysis of a culture. To a certain extent, disseminating this culture by "osmosis" cannot be completely replaced by communicating thinking tools. This does not mean that we just hope that something will develop spontaneously in the right way. On the contrary, planning statistical education as enculturation would be ambitious. My point is that focussing on "thinking tools" as tools for the individual thinker may be too limited. Moreover, we have to re-introduce differences between different cultures we intend to develop at a certain point. We have to take into account the differences between consumers and producers of statistics, between expert statisticians and those who should be educated so that they know when to consult an expert. Furthermore, the domain specificity of applications of statistics such as quality control or epidemology cannot be completely ignored.

Variation, Causation and Probability

I have not yet commented on dimension 2 (types of thinking). Wild & Pfannkuch concentrate on elaborating the notion of "variation" from this perspective and devote their whole chapter 3 to this topic. This chapter can be well read and used independently of the other chapters and has a value of its own. The central topic is the relation between probability, statistical thinking and the search for causes with regard to the practice of statistics and with regard to learning probability and statistics.

I am really glad that Wild & Pfannkuch have contributed further facets to a problem that has haunted me for a long time (Biehler, 1994 and Biehler, 1982) and that has not yet got the attention in statistics education it deserves. Wild & Pfannkuch make a plea to share with Pyzdek (1990) the basic belief that "all variation is caused". We use probability models to model "unexplained" variation. We do this as "a response to complexity that otherwise overwhelms us. The unexplained variation may well be the result of 'a multiplicity of causes' ..." (p. 20). This approach is important in order to avoid an attitude of stopping the questioning if something can be described by a probability model. In essence, this is Laplace's view of probability in an essentially deterministic world. Although the authors may not share Laplace's ontological claim, they plead for pragmatically using this view as a preferable strategy for dealing with real world problems. This perspective is also closer to students' habits in everyday thinking of looking for causes, and probability should build on top of this practice not as something completely distinct from deterministic thinking as some stochastics educators seem to suggest.

Although I agree with this basic line of argument, I think that one important aspect is missing that I tried to put forward in Biehler (1994). From the above perspective, probability is merely regarded as a last resort, something that we have to use because we cannot deal with complexity in any better way. But don't we get anything from adopting a probabilistic perspective? Well, first of all we get something from the transition from focussing on an individual event to looking at a system of events, which can be characterized by a (probability) distribution. This transition is intimately connected to a probabilistic perspective. We can analyze causes of why an individual accident occurred, but, in addition, we may wish to collect data on many accidents under various conditions. This transition has to be interpreted as a transition that can reveal new types of knowledge, new causes, explanations and types of factors that cannot be detected at the individual level. These new factors may explain features of a group as a whole that reflect boundary conditions. Nevertheless, the perspective of the individual case should not be eliminated but maintained as a genuine complementary view of a situation.

In the history of statistics and probability, the notion of *constant* and *variable* causes was coined. The metaphor of constant and variable causes is well materialized in Galton's Board (quincunx): The whole physical device represents the "constant causes", and the "variable causes" are responsible for the variation of the individual paths of the balls. One can become aware of constant causes when changing them; for example by changing the angle (to the left/right; forwards/backwards) of the Galton board with regard to the ground. Usually, the effect of the change cannot be observed in individual trials, but manifests itself in different long run distributions. The complementarity of individual and system level is often suppressed when probability educators draw students' attention only to the systems level of the Galton board, saying that individual paths cannot be further analyzed. The authors would rightly criticize this attitude, but we have to emphasize the positive gain of the system level as well. The item Wild & Pfannkuch used with their students, namely modeling the variation of success of basketball players by a binomial distribution, is another interesting case at the other extreme. As long as students cannot be convinced that the system perspective is of any value from any important perspective they will tend to reject this perspective as a possible but not as a convincing one in this concrete case, as one of Wild & Pfannkuch's students did. We may have to look for other prototypical examples that better show the complementarity of analysis levels. How does a reduction of the maximum speed limit on a road affect the distribution of car velocity on that road? This problem may be better suited for our purpose.

Wild & Pfannkuch seem to think in this direction by introducing the difference between common cause and special cause variation, a distinction taken from quality control, whose general importance would have deserved more space. My belief is that a further elaboration of this conception together with a more positive or constructive view of probability would result in further progress. I am looking forward to future papers of the authors on this topic.

Additional references

Biehler, R. (1982). Explorative Datenanalyse—Eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie. [Exploratory Data Analysis—A Study from the Perspective of an Empirical Philosophy of Science] IDM Materialien und Studien 24. Bielefeld: Universität Bielefeld, Institut für Didaktik der Mathematik.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes—Do we need a probabilistic revolution after we have taught data analysis? In *Research Papers from ICOTS 4*, Marrakech, Morocco, 1994, Ed. J. Garfield. Minneapolis: University of Minnesota. [Copy is available from the author.]

Garfield, J.B. & Gal, I. (1999). Assessment and statistics education: current challenges and directions. *International Statistical Review*, **67**(1), 1–12.

Polya, G. (1945). How To Solve It: A New Aspect of Mathematical Method. Princeton: Princeton University Press.

Pyzdek, T. (1990). There's no such thing as a common cause. In *ASQC Quality Congress Transactions*, pp. 102–108. San Francisco.

Schoenfeld, A.H. (1987). Cognitive Science and Mathematics Education: An Overview. In *Cognitive Science and Mathematics Education*, Ed. A.H. Schoenfeld, pp. 1–31. Hillsdale: Lawrence Erlbaum.

Wild, C.J. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, **67**, 223–265.

Response

C.J. Wild and M. Pfannkuch

We thank the discussants for their interesting and insightful comments. There is much food for thought and many valuable suggestions for further work. We agree with almost everything they say. Any disagreements we do have are with very subtle levels of emphasis rather than with the broad thrust of any of their arguments. We will limit our comments to just a few of the issues raised.

Snee and Moore gently (and colourfully in Moore's case) take us to task for the complexity of some of our models, particularly Fig. 1. Our models were developed in an attempt to analyse what happens in successful statistical problem solving for an audience of statisticians and statistics educators. They were not developed to be transmitted unfiltered to statistical novices (although there are elements that can be). We have always recognised that prioritization and simplification will be required for beginners. What we wanted to arrive at first was, in Biehler's words, "a sounder basis for pedagogically motivated reductions".

For beginners, the only part of Fig. 1 that we *explicitly* teach is PPDAC. A simple model for the analysis phase such as Moore's may suffice. Although beginning statisticians cannot become experts in measurement (or anything else), they can and should be given experiences that show them just how hard it can be to capture an important idea in measurement and experiences of situations where a measurement quite patently fails to do so. An important part of preparing our students to interpret and critically appraise media reports is the idea that if the investigators are measuring "the wrong thing" then the argument falls at the first hurdle. Over time, we would make students aware of some other elements of Fig. 1, but not all at once. A better use for Fig. 1 in teaching is as a yardstick against which teachers can measure the set of experiences they provide to students.

As Breslow surmises, we advocate abstract guidelines being used to complement examples of statistical thinking in practice rather than in place of them. We hope our framework can help teachers in their selection of case-studies, and help them to identify the types of thinking that are being used in the case studies so they can better draw students' attention to these aspects. To use case studies or project work effectively, we have to be able to abstract and convey to students the broader lessons to be learned from them. We think this is part of what Biehler is getting at in his discussion of "enculturation".

While we agree that the basic tenets of Snee's process-improvement paradigm apply very generally, some of the language and emphasis does not seem to us to speak to the every-day concerns of those working in the natural, experimental and social sciences. The need for "statistical thinking" there is every bit as pressing as in the organisational areas. The most powerful and easily transferable concept from the quality area is the concept of process as a means of analysing situations and problems. We take Snee's point about the desirability of simplicity and the incorporation of relevant tools in models actually used in teaching. Regarding Moore's "What, no Bayes?", we saw the Bayesian/frequentist/Fisherian debates as relating to a more detailed level of the Analysis phase of PPDAC than the levels explored in this paper. With respect to "creative thinking" (Smith), we emphasised the roles of imagination and flashes of insight which we see as the wellsprings of creativity. We would like to draw the reader's attention to the discussion in Smith (1997) about the differing, but complementary, roles of case studies and surveys in the development of social theories.

Smith is quite right in saying that the whole of science is concerned with variation. We hazard that an essential difference is that whereas the main focus of the sciences is on answering particular

questions arising from variation, the main focus of statistics is on the process by which this is done, or at least on specialised parts of that process. In reality, there are no clear boundaries between "the science" and "the statistics" and it all works best as a seamless whole. Should scientists be educated to be better statisticians? Very much so. But this in no way obviates the need for statisticians to be educated to be better scientists.

Smith worries about the tenuousness of employing random-variation models in situations beyond those in which randomness is induced by the study design. Unfortunately, all analyses of observational data aimed at trying to infer something about the nature and behaviour of underlying processes lie beyond these secure borders. We statisticians cannot claim our models are "correct" for such uses and for many reasons including those listed by Smith, investigators often come unstuck using observational data. But such analyses do address some of the most important problems humanity faces, randomness models do allow us to conceptualise unexplained variation, and they have proven to be "sometimes useful" (harking back to Box). The issues here are closely related to W.E. Deming's distinction between "enumerative" and "analytic" studies elaborated on by Hahn & Meeker (1993). A qualitative conclusion reached is that for analytic studies, the measures of uncertainty produced by statistical methods are best thought of as lower bounds, but that even these lower bounds tend to be surprisingly large. In Section 3, we have ventured into very deep waters, partly stimulated by Biehler (1994). The issues at stake relating causal and probabilistic thinking are much more fundamental to applying statistics than our traditional debates on the "foundations of statistical inference". We certainly do not have all the answers and are pleased with the discussion generated.

Like Smith, we have often cringed at claims by some statisticians for the universality of statistics and some sort of seer-like status for statisticians. Statistics as it has so often been conceived of and taught in the past has been narrowly specialised and we are only too aware of the limited nature of our own capabilities. But disciplines are not static. Many forces act to change the interests of their researchers and the content of their teaching. For statistics the biggest of those forces has been computer technology. If statistics is not to become some dusty, forgotten annex of computer science, statistics education must endeavour to *evolve in the direction* of universality. We recall that phrase from earlier writings of Moore about nudging statistics a little further back towards its roots in scientific inference. This involves, in Snee's words, "moving further upstream". But most of us have a lot to learn along the way!

The fundamental issue that we must continually confront as technology advances is the distinction between that which can profitably be automated and that which is essentially or necessarily human. It is on the latter, "the 'art' of statistics" (to quote Breslow) that we should concentrate when teaching all but very small numbers of specialists. As far as applications are concerned, knowing the technical details of what is happening inside a procedure is of benefit only insofar as it helps us to use that procedure more effectively. It *is* just like driving a car. Forming creative connections between ideas and pieces of information would seem to be fundamentally human contributions. Skills in enquiry and skills in arriving at insights from data (and these skills are intimately related) are fundamental, universal, human needs. The more successfully statistics can address these needs, the greater its value and the more secure its future.

In modern academia, the continued health of a discipline depends on student demand. Many of us have been cushioned by the demand for service teaching which will remain only so long as the skills it teaches are perceived to be valuable by those who direct students to us. A professional statistician requires graduate-level education. In the environment we are operating in—beset by competition from disciplines with stronger student appeal—a prospering graduate programme must be supported by and fed by a strong undergraduate programme. To attract students, such a programme must confer valuable and marketable skills. The skills that society needs that are closest to what we have traditionally done are skills in investigation and gaining insights from data. This is one case in which both altruism and the survival instinct point to the same conclusions.

Earlier, we tried to enunciate a distinction between the focus of statistics and that of the sciences. Philosophers of science are also interested in idealisations of scientific processes but they do not have our pragmatic tradition. So not only is the direction we are urging consistent with our history, there is an opening for us. This is statistics as a "liberal art" (Moore, 1998)—a liberal art which transmits scientific thinking and statistical analysis far beyond the boundaries of the traditional sciences and even into daily life. We agree with Moore there has been pleasing progress in this direction over the last 20 or 30 years, and we can think of no one who has had a greater impact on beginning statistics in this regard than Moore himself. We hope that our paper has generated sufficient light to guide a few more tentative steps. And while we have no doubt that an art is best learned "through apprenticeship under a master practitioner" (Breslow), financial pressures dictate most of us are unavoidably in the business of inexpensive mass education for our undergraduate teaching. To teach the art of statistics at all effectively in such an environment, we must begin to learn how to demystify that art to the greatest extent possible.

Additional reference

Hahn, G.J. & Meeker, W.Q. (1993). Assumptions for statistical inference. *The American Statistician*, **47**, 1–11.

[*Received November 1998, accepted May 1999*]