Streamlined Statistics : An Innovative Polytechnic Approach Reviewed

Mike Camden - Wellington, New Zealand

1. Introduction

In their jobs, many people need to deal with: large sets of data; a feeling that they don't have the mathematical skills they need to deal with this data; and some software - probably a spreadsheet package or graphics package.

These people also have lots of common sense.

This paper describes an approach to statistics which aims to build numeracy and graphic skills onto people's common sense, and to give skills in using software on statistical data. It examines the subject, 3150 "Statistics" for New Zealand Science Technicians, then evaluates our experience at Wellington Polytechnic with this and similar subjects that we have presented to a wide range of students.

2. Statistics for Science Technicians (Statistics 3150)

This introductory statistics subject for the New Zealand Certificate in Science was developed by a group of statisticians and teachers, and launched in 1986. Second and third year subjects followed.

The aims of the introductory course are as follows: First, to introduce students to the most used concepts and techniques of statistics. Second, to enable students to grow in numerate judgement. Third, to achieve skills in collecting and exploring data, representing them graphically, and reporting conclusions. Its contents are set out below. Each topic is allocated a given number of class hours and computer hours. Very deliberately, the subject starts off with Exploratory Data Analysis (Topics 1, 2 and 3). The second and third subjects start off with further work on EDA.

TABLE 1
Topics for Statistics 3150

	Topics	Time (hours)		
		Class	Comp	Total
0	Statistical Literacy	0	0	0
1	Exploring Single Variables	8	6	14
2	Exploring Pairs of Variables	7	6	13
3	Exploring Series Variables	4	5	9
4	Probability Concepts	8	0	8
5	Probability Models	14	2	16
6	Sampling and Data Quality	8	2	10
7	Estimation	6	1	7
8	Statistical Quality Assurance	10	4	14
9	Significance	3	0	3
10	Least Squares Regression	2	4	6
	Total	70	30	10

We had some difficulty persuading the examining authority to accept a topic numbered zero, and with no hours allocated to it! We felt that "Statistical Literacy" should permeate the whole subject, and defined it like this:

Topic 0 : Statistical Literacy: The student should be able to ...

(i) communicate clearly:
- (a) record data and transcribe the results of calculation accurately, and to a suitable precision;
- (b) construct tables and graphs which are clear conveyors of information;
- (c) equip tables and graphs with titles and labels;
- (d) name variables correctly and state their units;
- (e) state the source of data;
- (f) where possible, check that numerical results are reasonable.

(ii) use statistical computer packages:
- (a) use a statistical package for every appropriate objective of this prescription;
- (b) understand the role of computing in data exploration and statistical inference;
- (c) approach new statistical software with confidence.

3. Principles and prejudices

The beliefs, principles, and possibly prejudices, which the group used in developing these subjects are similar to those which guide our teaching. This section states our philosophy for the learning of statistics by our client student groups.

This philosophy is designed for a client group which excludes future statisticians, but includes future and current laboratory technicians, research workers, policy developers, nurses, computer professionals, environmental health officers, electronics and engineering technicians, and various others.

Exploration versus inference: Skills in exploration (or EDA) are a much better starting point, and are much more useful to our target group, than skills in inference. This principle is based on the beliefs that:

(i) inference, especially hypothesis testing, is often done inappropriately;
(ii) if hypothesis testing is to be done at all, it should be left to the professionals;
(iii) even where formal inference is appropriate, an exploration of the data for non-standard features is necessary;
(iv) exploration often yields more information from the data than formal inference, as stated by John Tukey at the NZ Statistical Association Conference in 1977.

An alternative path to useful numeracy: The exploration of data with graphs is an alternative path to skills and confidence with numbers. These exploratory skills can be based on experience, rather than on the mathematical skills which eluded many people at school, and they are among the skills most often needed in today's workplace. We believe that this approach bypasses the blockages of "Maths Anxiety".

Exploring data with graphs involves two sorts of skills: mechanical skills in making graphs, and the much more challenging skills of reading the graphs. Many statistics texts, even some recent ones, spend large efforts on making graphs, and hardly any on interpreting them.

Humans communicated with pictures for thousands of years before they used written words or numbers. However, it took humans until 1786 (Tufte, 1982) to begin publishing statistical data graphics. Our teaching experience makes it very clear that the making and reading of graphs is a skill which must be learnt, but which feels natural once it is learnt.

A healthy diet of raw data: Students flourish if they're fed on a diet of data-sets which are raw, fresh, high fibre, locally-grown, and which contain some bugs and blemishes. The data-sets need to have bulk, which means they are multivariate. Our reasons? First, real data is full of interest for learners (and full of surprises for teachers). Second, our client group will have to struggle in their jobs with data that has missing values, clerical errors, and unexpected features. Third, our client group will need to read research literature, and will benefit from a level of cynicism about the data on which the research is based. And finally, real data-sets are always multivariate, and relationships among variables are much more interesting than the properties of a single variable.

Time-series data-sets are readily available to teachers, though they are often "cooked" rather than raw. Case data-sets are harder for teachers to come by. Hence, the NZ Statistical Association published the *Data Bundle* (Camden, 1989), a collection of 25 data-sets, with suggestions on how they can be explored.

The place of software tools: A statistical package with graphical features is just as essential to the learner as a pencil. Students should use a package from the start of the course onwards, to store and edit data-sets, to do calculations, and to draw graphs.

Software is developing very fast, but not necessarily in the ideal direction for our purposes. Perhaps non-statistical software developers may soon realise what the statistical uses of graphs really are!

We are faced with two key questions.

(i) *Does the ideal beginners' package exist?* It has to be appealing, really simple to use, really elegant in its output, and easy to print from. Ideally, it is interactive as well. I haven't found it yet but there are some strong contenders. They include Data Desk Professional (Velleman, 1989), the PC version of MINITAB (Joiner, 1985), and PC-INFOS (NZ Department of Statistics).

(ii) *How can modern spreadsheets be used in data exploration?* Spreadsheet software with graphics (such as Excel (Microsoft, 1989) and WingZ (Informix Software, 1988)) must be one of the fastest developing forms of software, and one of the most-used tools for handling data in the workplace. They merit a place in our courses.

Obsolete and emerging skills: The existence of software with statistical features renders some skills obsolete. These include the ability to calculate means, standard deviations, correlation coefficients, and regression coefficients. The whole process of classing data then doing calculations on the grouped distribution is unnecessary. Classing distorts relationships with other variables and distorts the measures which are calculated. The intricacies of histogram construction can be left out too.

The software renders some skills more important, and easier to demonstrate. These include an understanding of how the measures of centre, spread, and relationship are affected by data features like clusters and outliers. A fundamentally important skill which becomes easier is the process of finding residuals and inspecting them with graphs. Even with the help of friendly software, students find this a very challenging process to grasp.

The software suggests one new set of skills at least: the skill of designing graphics (with the help of software) which communicate clearly and represent the data accurately.

Time series: Our approach to statistics includes the exploration of time series, with plotting, smoothing, and inspection of residuals. Our reasons are as follows. Most of the graphs which we find in the public media are about time series, and a large part of the data which workers find on their desks consists of time series. Most graphics software is directed towards time series.

What to include and what to exclude: Our criteria for choosing the subject's content go as follows. First, put in the topics which are most likely to be useful to the client group, and leave out topics of doubtful usefulness. Second, put in topics which will leave clients with skills which are based on their own experience, and leave out topics which will leave clients with only a vague idea of their purposes. Third, for every technique put in, add a treatment of its limitations.

4. Our client groups

Our students are people who will not be statistical specialists but who have or will have lots of data to deal with. Some have recently left school and some are more mature in years. On the whole, they don't like mathematics too much.

The histogram (Figure 1 below) shows how a group of our school-leavers compares with the population of New Zealand school-leavers, by using the grades in National Sixth Form Certificate Mathematics. About 40% of New Zealanders take this subject, usually when they are aged about 16. The grades range from 1 (high) to 8 (low). The bars show the distribution of grades awarded in the New Zealand Sixth Form Certificate, 1988. The X's show the distribution of grades held by a group of incoming students.

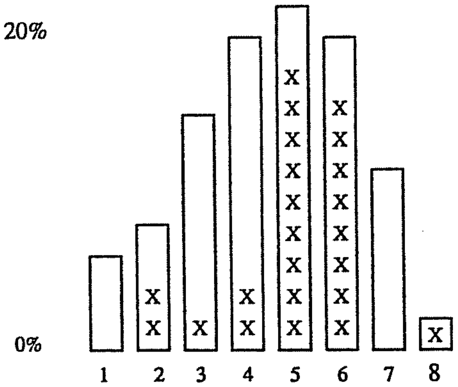

FIGURE 1
Percent of students taking Sixth Form Certificate Mathematics per grade

Our older clients have mathematical backgrounds ranging from "I was completely lost in Maths at school from fourth form onwards", to "I did Statistics for Sociology, Economics, Geography etc. and can remember that we did lots of Chi Square Tests"!

The essential statistical needs of these groups are skills in dealing with data, and skills for reading the literature in their fields.

5. An evaluation of our approach

I shall not evaluate our approach by testing hypotheses; instead I shall offer some short explorations.

5.1 *Exploratory statistics as an alternative route to numeracy*

Our student groups always contain people with very low confidence in mathematics (algebra, calculus, trigonometry). These people often handle the statistics course with confidence and success. The scatter-plot (Figure 2) shows mathematics marks against statistics marks for a small group of first year science students. The two outliers did not let their low mathematics background prevent them from succeeding with statistics.

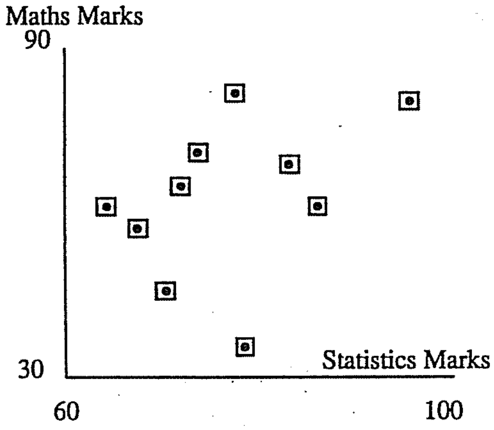

FIGURE 2
Mathematics marks against statistics marks

5.2 *Interest, usefulness and perceived new skills*

A small group of students who had completed the science statistics first course just a year before were asked to grade the interest and usefulness of each of the ten topics. They used a scale which went from 0 (not interesting or useful at all), to 4 (extremely interesting or useful). The Environmental Health interests of these students show through in their perceptions. The winning topic, 6: Sampling and Data Quality, includes practicalities like convenience and judgement sampling. The next topic, 8: Quality Assurance, is of particular interest to this group. After that came the EDA topics. Probability and Estimation are seen as interesting even when they are not useful!

In their comments, this group gave Graphics a high value, along with Sampling and Quality Assurance.

The same students applied a similar 0-4 scale to assess the extent to which the course had helped them develop new skills. The highest scores (average of 3) related to improvements in their ability to read scientific literature and graphs, followed closely (average 2.7) by improved ability to use graphics. Improvement in numeracy skills scored somewhat lower (average 2.0).

5.3 *Student perceptions of statistical software*

This is at the heart of our approach. If students feel happy to launch into exploring a data-set with the help of software tools, then we have succeeded. In fact, students feel only moderately happy about this. Here are some explanations: First, a minority of students find computers very difficult; they dislike keyboards and screens. Second, many students find it difficult to remember the software's commands, and syntax. Third, students have to learn two quite different sets of skills at once; skills in exploring data, and skills in using a particular software package. Our approach seems to require software which is extremely simple to use, and elegant in its output.

6. The ultimate streamlined statistics course

The ultimate streamlined statistics course for our client group would seem to rush onwards with so much fascination and so little turbulent difficulty that students would hardly realise that they were doing statistics. We haven't quite managed it yet. The following are the directions we are taking, or would like to take, in further streamlining our courses:

(i) A thorough coverage of EDA should remain right at the start of the course.

(ii) Topic 1 should be streamlined by the removal of stem-and-leaf plots and box-and-whisker plots. They are useful and fascinating tools, but they look like puzzles to the beginner. The dotplot is the cleanest way of displaying the features of a distribution, and the meaning of the measures of centre and spread. We would like to remove histograms too, but they are more acceptable to the public than dotplots.

(iii) Topic 2, Exploring Pairs of Variables, should cover three common situations: two measurement variables, one measurement variable and one category variable, two category variables.

(iv) Topic 3, Time Series, will be extended to include plots of raw, smooth and rough; deflating a series by an index series; and constructing demographic rates. Our clients' experience enables them to deal with series graphs much more easily than with frequency graphs or scatterplots. Spreadsheet software helps here too, and PC-INFOS adds to the enjoyment greatly.

(v) Topics 4 and 5, Probability and Probability Models, will get left out. They are of much less practical use to our clients than the other topics. In their nature, they are more akin to mathematics and mathematical modelling. I admit that omitting them does cause problems with two later concepts: sampling distributions and acceptance sampling.

(vi) Topic 6, Sampling and Data Quality, will be strengthened with work on questionnaire design and the practicalities of conducting surveys. Our students often need to do projects involving questionnaires and surveys. Taking steps to ensure data quality and dealing with data defects, are also important skills for our clients.

(vii) Topics 7 and 9, Estimation and Significance, will be directed towards skills in critically reading the relevant literature.

(viii) Topic 8, Statistical Quality Assurance, could be reduced to process control with charts, by the removal of acceptance sampling.

(ix) Topic 10, Regression, should be directed towards assessing the appropriateness of the model, and towards inspecting residuals. Software, including spreadsheet software, enables the fitting of non-linear functions, and this ties in with work some students have done in mathematics classes.

(x) A further topic, Making Quality Graphs, needs to be added. Our clients will be expected to include graphs in their reports and presentations, and software for making good (and bad) graphs is available. In a world full of graphs there are few people who clearly understand the way graphs work (Bertin, 1984) and the making of quality graphs (Tufte, 1983; Cleveland, 1985). I hope our students will be among them.

7. Sources

Some texts and software packages state a philosophy very similar to ours, explain the skills required for it, and/or provide appropriate software tools. These texts include Koopmans (1981), Chambers, Cleveland, Kleiner and Tukey (1988), Mosteller, Fienburn and Rourke (1983), the MINITAB Handbook by Ryan, Joiner, and Ryan (1985), Marsh (1988), the manual for the S package by Becker and Chambers (1984), and the Quantitative Literacy Series by Landwehr and others (1987). A number of papers at ICOTS 2 presented similar philosophies. The documentation for the Data Desk software, by Velleman (1989), contains a very clear statement of the exploratory philosophy, and a very appropriate quote from Winnie the Pooh!

References

Becker, R A and Chambers, J M (1984) S: *An Interactive Environment for Data Analysis and Graphics.* Wadsworth Duxbury, Boston.

Bertin, J (1984) *Semiology of Graphics.* University of Wisconsin Press.

Camden, M D (1989) *The Data Bundle.* Education Subcommittee of NZ Statistical Association, Wellington.

Chambers, J M, Cleveland, W, Kleiner, B and Tukey, P (1983) *Graphical Methods for Data Analysis.* Wadsworth Duxbury, Boston.

Cleveland, W (1985) *The Elements of Graphing Data.* Wadsworth.

Davidson, R and Swift, J (1986) *ICOTS II Proceedings.* ISI/University of Victoria, British Columbia, Canada.

INFOS Services (1989) *PC/INFOS* (with *dX* software: Econdata Pty, Australia, 1989). Department of Statistics, Wellington.

Landwehr, J M and Watkins, A E (1986) *Exploring Data.* Dale Seymour Publications, California.

Koopmans, L H (1981) *An Introduction to Contemporary Statistics.* Wadsworth, Duxbury, Massachusetts.

Marsh, C (1988) *An Introduction to Data Analysis for Social Scientists.* Polity Press, Cambridge, UK.

Mosteller, F, Feinburg S E, and Rourke R E K (1983) *Beginning Statistics with Data Analysis.* Addison Wesley, Massachusetts.

Research and Statistics Division (1989) *1988 NZ School Certificate and Sixth Form Certificate Statistics.* Ministry of Education, Wellington.

Ryan, T, Joiner, B and Ryan, B (1985) *MINITAB Handbook* (Edition 2). PWS.

Stirling, D (1987) *STATLAB.* NZ Statistical Association/Massey University, Palmerston North.

Tufte, E (1983) *The Visual Display of Quantitative Information.* Graphics Press, Conneticut.

Velleman, P (1989) *Data Desk Handbook* and *Statistics Guide.* Odesta Corp., Northbrook, Illinois.

Excel, Version 2.2 (1989) Microsoft Corp., Remond WA.

WingZ (1988) Informix Software Inc, Lenexa KS.