STATISTICS COMPUTING LABORATORY

Alan J. Branford

ABSTRACT
The utility of Statistical Science to a vast range of disciplines, from biological and agricultural sciences, through medical science, education research, and social sciences, is evident by the plethora in a typical university of service subjects variously labelled 'Data Analysis', 'Statistics', 'Quantitative Methods', or 'Research Methods'. Although there have been considerable advances in computing and graphics technologies, many statistics service subjects have merely grafted these innovations into a traditional curriculum. *Statistics Computing Laboratory* has been created ab initio, mindful of the requirements of client disciplines, of the background of the students, and of the opportunities provided by modern computing technologies. This paper discusses the curriculum design and pedagogy, and evaluates the *Statistics Computing Laboratory* paradigm.

KEYWORDS
Statistics service teaching, data analysis, statistics, quantitative methods, research methods.

1 Introduction

Statistical Science is quintessentially an interdisciplinary field of study. Although its theoretical foundation is in mathematics, its practice transcends disciplinary boundaries, from biological and agricultural sciences, through medical science, education research, and social sciences – we shall term these the "user disciplines". Indeed, the pioneers of modern statistics number a geneticist (Fisher) and a geophysicist (Jeffreys); long before the University of Cambridge had a Chair of Statistics, or even a formal departmental home, there was a Statistical Laboratory of practitioners, founded by Wishart, an agricultural scientist.

A university typically will have a Department of *Mathematical* Statistics, or its academic statisticians will be subsumed by a Department of Mathematics. Meanwhile, service subjects variously labelled 'Data Analysis', 'Statistics', 'Quantitative Methods', or 'Research Methods' will abound in the user disciplines, taught by academics in that user discipline. In many cases, the academic statisticians have no desire to be involved with service teaching, preferring to concentrate on the pure aspects of their profession. Meanwhile, the user disciplines often jealously guard their home-grown statistics subjects and resent interference by what they would view as purists divorced from the reality of their needs.

Three principal justifications for user-discipline-based statistics subjects can be postulated.

1. There are techniques important in the practice of the user discipline that are rarely required in other areas;
2. there is a body of received wisdom concerning problems that occur in the user discipline;
3. students of the user discipline will best relate to the statistics if they study as a cohort immersed in their home discipline.

As an example of the first, economics students would have a need for the study of time series, whereas this area of statistical modelling and analysis would be rare in the biological sciences. For this reason, a generic statistics service subject, or one provided by another user discipline, would be deficient in time series content. Hence the impetus for an economics-based subject in econometrics.

The second can be illustrated by considering, say, students of environmental science. It may be known amongst environmental scientists that the variable *hydrocarbon pollution potential index* is approximately normally distributed after a logarithmic transformation. This is an extract from a body of received wisdom. How could a generic statistics service subject be guaranteed to transfer that accumulated body of knowledge?

However, the vast number of service subjects, based in far-flung user disciplines in the one university, raises certain issues.

1. Are resources being used efficiently?
2. Are the curricula and the pedagogical techniques employed appropriate?

Statistics Computing Laboratory is a recently developed and evolving paradigm for the service teaching of statistics within a university. Its teaching strategies and implementation, discussed in Section 4, provide a generic framework whilst incorporating user discipline aspects. This structure allows for efficient delivery and facilitates the incorporation of modern curriculum and pedagogical developments in a controlled manner across the client groups and yet retains user discipline input to address the three principal justifications for user-discipline-based statistics subjects postulated above (see Section 6).

Pedagogical issues in the teaching of a statistics service subject are discussed in Section 2, which includes a typical example of pedagogical deficiencies in many traditional statistics service subjects.

Section 3 develops from first principles the educational aims and expected learning outcomes for a statistics service subject. From this base, the paradigm was developed – its structure is described in Section 4. Assessment methods are discussed in Section 5. An evaluation of this teaching method is given in Section 7.

2 Pedagogy in a Statistics Service Subject

It is important to realise that students in a statistics service subject typically lack quantitative skills and confidence. Certainly the subject should have the effect of raising these qualities in the student, but care must be taken to address the aims and objectives of the subject and not distress or distract students unnecessarily with intricate computations or mathematical formulae.

Consider as a typical example the analysis of variance for the one-way layout. The experimental design of the one-way layout arises when samples are collected separately from each of r different populations and the value of a continuous variable is recorded on each subject sampled. The purpose of the analysis is to draw conclusions about the population averages on the basis of the samples. Under appropriate modelling assumptions, this form of data may be analysed using the technique of analysis of variance.

Analysis of variance apportions the total variation in the data (quantified by the "total sum of squares") to two sources: variation due to systematic differences between the populations (quantified by the "between groups sum of squares") and natural variation from subject to subject (quantified by the "within groups sum of squares"). From this, the significance of the sample differences between the groups may be assessed.

Typically, a student will be presented with the formulae for the various sums of squares, for example, the between groups sum of squares

$$SS_{between} = \sum_{i=1}^{r} n_i (y_{i\cdot} - y_{\cdot\cdot})^2 \ ,$$

where n_i is the number of observations in group i, $y_{i.}$ is the sample average for group i, and $y_{..}$ is the average of all the data.

A student with a mathematical background would presumably not be fazed by the mathematical symbolism, and would have sufficient quantitative confidence to implement the formula in (smaller) applications. Furthermore, it would be expected that this student would appreciate from the formula that this is indeed a sensible measurement of the variation between groups (and may even see its connection with projective geometry!).

A typical student of the statistics service subject, however, would be overwhelmed by the mathematical formula and require considerable coaching just to successfully implement it. The subtler appreciation of the formula as a measurement of the variation between groups would be lost completely. Given that any practitioner would use a computer package in any practical situation, the student in short would be wholly distracted by something totally irrelevant.

Surely, what is important is simply the fact that the total variation in the data is apportioned to two sources: variation due to systematic differences between the populations and natural variation from subject to subject; from this, the significance of the sample differences between the groups may be assessed. A presentation of the formulae is not required to convey this point, nor to implement this technique. Rather, suitable computing and graphics technologies will cement this principle and allow the student to confidently implement it in practice.

Thus, the pedagogical technique required by a statistics service subject is a clear exposition of statistical principles and concepts, confirmed and explored using modern computing and graphics technologies, thereby also giving rise to practical implementation skills. Mathematical formulae for their own sake have no role in this. These pedagogical principles underpin the development of the *Statistics Computing Laboratory* paradigm.

The following example illustrates the importance of exploiting the power of modern computing and graphics technologies in statistical analysis.

> *Distribution of toxic substances in rivers–*
> Source: Jaffe, P.R., Parker, F.L., and Wilson, D.J. (1982), "Distribution of toxic substances in rivers", *Journal of the Environmental Engineering Division*, **108**, 639-649.
>
> Jaffe, Parker and Wilson (1982) investigated the concentration of aldrin, a hydrophobic organic substance, in the Wolf River in Tennessee. Measurements were taken downstream of an abandoned dump site that had previously been used by the pesticide industry to dispose of its waste products.
>
> It was expected that this hydrophobic substance might have a nonhomogeneous vertical distribution in the river because of differences in density between the

> compound and water, and because of the adsorption of the compound on sediments, which could lead to higher concentrations on the bottom. It is important to check this hypothesis because the standard procedure of sampling at six-tenths of the depth could miss the bulk of the pollutant if the distribution were not uniform.
>
> Grab samples were taken with a La Motte-Vandorn water sampler of 1 litre capacity at various depths of the river. Ten surface, ten mid depth and ten bottom samples were collected, all within a relatively short period. The aldrin concentrations (in nanograms per litre) in these 30 samples are given.

An analysis of variance is required here. Hand calculations will reasonably easily produce the analysis of variance table, but post hoc multiple comparisons would be impractical. But more importantly in this example, the use of graphics tools will allow the student to demonstrate that the assumptions required for the one-way analysis of variance are not met by these data as they are presented, but that they are met for the inverse concentrations. This observation fundamentally alters the model, and relies heavily on graphics capabilities through inspection of appropriate residual diagnostic plots.

3 Aims, Learning Outcomes and Syllabus

The aim of teaching is to facilitate learning, not simply to present a range of subjects or to conduct a series of classes. Subject planning therefore starts with a statement of the educational aims of the subject and the learning outcomes desired in terms of the development of students' knowledge, understanding, skills and attitudes.[1]

By examining from first principles the reasons for having a statistics service subject we may formulate an Educational Aim as follows.

> **Educational Aim**
> This subject aims to introduce students to standard statistical analyses which are routinely encountered in the life and social sciences. This aim will be achieved by involving the students directly in canonical problems in the computer laboratory.

This process of development then further yields the following Expected Learning Outcomes.

[1] Adapted from the Flinders University's Education and Research Policy.

Expected Learning Outcomes
On successful completion of this subject, a student is expected to:-

- have acquired experience in identifying the various techniques required;
- have the ability to assess the validity of underlying assumptions;
- have the knowledge to implement analyses in a commonly used computing package; and,
- be able to present informed and critical summaries.

A syllabus for a first subject in statistics for user disciplines which has no prerequisites or assumed knowledge is illustrated in a modular fashion in Figure 1 in the form of a dependency chart. The teaching strategies employed (see Section 4) and the assessment methods used (see Section 5) suggest that this subject is best suited for the second year of a three-year undergraduate degree.

Figure 1: Dependency Chart for the Syllabus

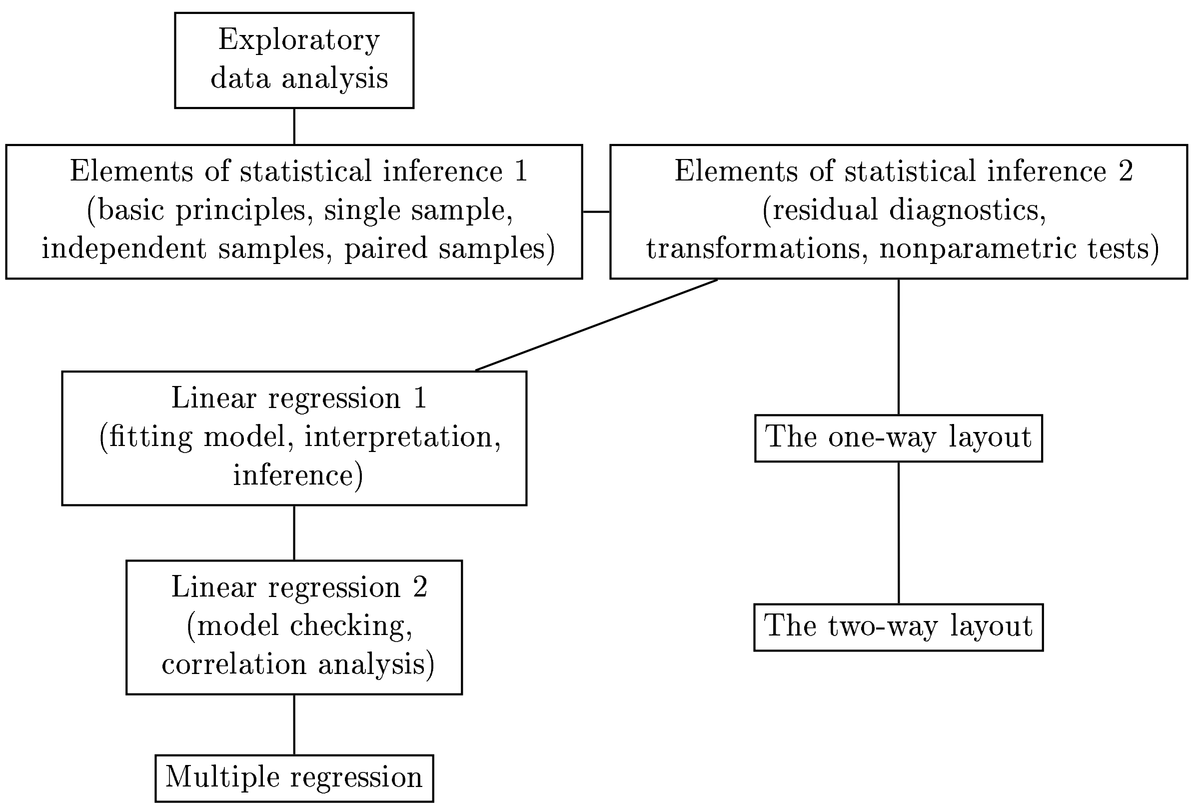
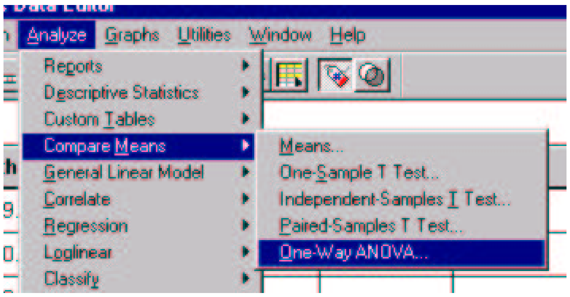
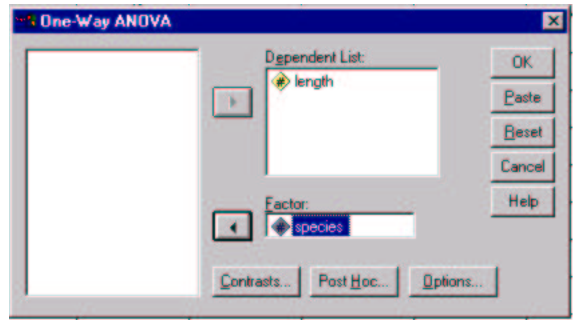


Figure 2: Example of pictorial assistance



Selecting the One-Way ANOVA item



Completed dialog box

4 Teaching Strategies and Implementation

Statistics Computing Laboratory is presented as a number of modules (usually 8), plus an introductory lecture in week one of teaching, according to the dependency chart in Figure 1. For each module there is a one-hour lecture to present concepts and to overview and signpost the practical implementation of the statistical analyses. For each module, there is also a two-hour computer laboratory practical in which the student executes practical tasks and achieves various checkpoints related to the module. In our implementation of this paradigm, we use SPSS 9.0 for Windows as the vehicle of instruction.

In any given semester, there will be students from a number of user disciplines. Students formally will be enrolled in a subject associated with their home discipline, for example, "Statistics for Medical Science", "Statistics for Forensic Science", and so on. However, the students all attend the same lecture; they execute the same practical sessions, although in groups according to their home discipline. The examples in the presentation of the module are deliberately drawn from across the spectrum of statistical applications, to illustrate the utility of statistics to a vast range of scientific disciplines. The user-discipline input is via the assessment instruments (see Section 5).

The text book is an internal publication– a workbook containing a detailed discussion of the statistical principles and concepts, and a detailed guide to the implementation. Mathematical formulae for their own sake are eschewed. This workbook not only has a key role in the delivery of the subject, but also serves as a reference book for the students' later career.

Figure 2, which signposts the execution of an analysis of variance of the one-way layout of 'length' (of eggs) grouped according to 'species' (of bird), illustrates the nature of pictorial assistance which accompanies the discourse.

Thus the teaching strategies and implementation satisfy the pedagogical requirements discussed in Section 2.

5 Assessment

Assessment is by the achievement of computer-laboratory-based checkpoints, by practical assignments (one per module), and in some variants by a project, which is consistent with the aim of providing a practical problem-solving course in routine statistical analysis in the life and social sciences. Each expected learning outcome is addressed in this assessment strategy.

5.1 Assignments

Whereas the examples in the presentation of the module are deliberately drawn from across the spectrum of statistical applications, to illustrate the utility of statistics to a vast range of scientific disciplines, the assignments are drawn from the student's home discipline. A particular user-discipline cohort will be presented with a situation drawn from their discipline and they will be provided with the data. They are required to analyse the data and to synthesise a consultant-style report, complete with appropriate tables and figures, which they submit for assessment.

In our implementation, each assignment is designed to require 5 hours of work on the part of the student. As discussed in Section 4, each module has 3 hours of class contact. Allowing a further 2 hours per module of general study and review, our implementation has a student workload of 10 hours per module.

The following two examples are both assignments from the multiple regression module, one for commerce students, the other for environmental science students.

5.1.1 Example from commerce

> *Predicting Appliance Sales–*
> Source: *Business Statistics*, US Department of Commerce
>
> The data gives factory shipments (domestic) of dishwashers (thousands) in the United States from 1960 to 1985. These and other data are published currently in the Department of Commerce's Survey of Current Business, and are summarised from time to time in their publication, *Business Statistics*. Also included are durable goods expenditures and private residential investment in the United States, both in billions of 1972 US dollars. You are to explore the relationship between dishwasher shipments and durable goods expenditures and private residential investment.

The utility of basing the subject on computing and graphics technologies is that students can work with actual data, and deal easily with issues of model development otherwise inaccessible to them. In this example, the student will discover from the residual diagnostic plots, having first fitted a model of multiple regression of dishwasher shipments on durable goods expenditures and private residential investment, that the average dishwasher shipments appear to depend *quadratically* on durable goods expenditures. They may then repeat the analysis, this time with three predictors: the two original predictors and the square of durable goods expenditures.

5.1.2 Example from environmental science

> *Evaporation from the soil–*
> Source: Freund, R.J. (1979), "Multicollinearity etc., some "new" examples", *American Statistical Association Proceedings of Statistical Computing Section*, 111-112.
>
> Observations represent 46 consecutive days from June 6 through July 21. The variables measured were the average soil temperature (integrated area under daily soil temperature curve), the average (integrated) air temperature, the average (integrated) relative humidity, and the total wind (miles per day), as well as the daily amount of evaporation from the soil.

In this assignment, the environmental science students were required to explore the model of multiple regression of daily amount of evaporation from the soil on the other variables, so as to derive an appropriate prediction model. The examination of appropriate exploratory plots and residual diagnostic plots, and the fitting of submodels as well as the full multiple regression model are all required– again the key use of computing and graphics technologies.

5.2 Project

A project allows the combination in the one endeavour of many of the statistical skills acquired, and also the collection of data in the user-discipline context. Some user disciplines may instead wish to have separate project subjects as follow-ons to this one.

6 Addressing the Requirements of User-Disciplines

The assignments and, if used, the project provide a vehicle for transferring the user-discipline-specific body of received wisdom. Students will study in part as a cohort from their discipline, as the practical classes are formed according to the user disciplines. The

Table 1: Enrolled Students by Major

Major	Number	Major	Number
Aquaculture	20	Education	5
Biology	7	Information Technology	4
Chemistry	3	Mathematics	2
Earth Sciences	2	Science & Business	4
Economics	2	Science (miscellaneous)	5

students will not only see the relevance to their home discipline (through the assignments and project) but also to the wider context (through the classes).

Addressing the situation of any techniques important to the practice of the user discipline that are rarely required in other areas is more problematic. These may be able to be addressed in the project. Alternatively, it may be desirable to develop these as further modules within this paradigm, for the use of some but not all client disciplines. It may be necessary for these to be taught within the home discipline as a follow-on subject. In any event, *Statistics Computing Laboratory* provides an appropriate first subject in statistical analysis for user disciplines.

7 Evaluation

A generic, elective version of the subject has been offered and developed since 1998, as have two small user-discipline variants. A large full-scale implementation will commence in 2001. In 2000, 54 students enrolled in the generic, elective version. These students were drawn from majors as tabulated in Table 1. A questionnaire was administered through the Advisory Centre for University Education (ACUE) of the University of Adelaide, with questions selected from the ACUE's Student Evaluation of Teaching (SET) Question Bank (version F4.2). There were 39 responses received, although not every respondent addressed every question. (The ACUE procedure does not track respondents by major.) Histograms of the results are displayed in Figure 3. For each question, the response is on a 7-point likert scale, with 1 = 'Strongly disagree', 4 = 'Undecided', and 7 = 'Strongly agree'.

Questions (a)–(d) address the aims of the subject and demonstrate that the students appreciated why the subject was being offered and confirm that it achieved its aims. Question (e) reveals that the students found the workbook useful. Students' perception of their learning (question (f)), motivation (question (g)), and abilities and skills (question (h)) were investigated and found to be very positive. Finally, questions (i) and (j) show that the students were satisfied by the assessment process.

Figure 3: Student evaluation of teaching

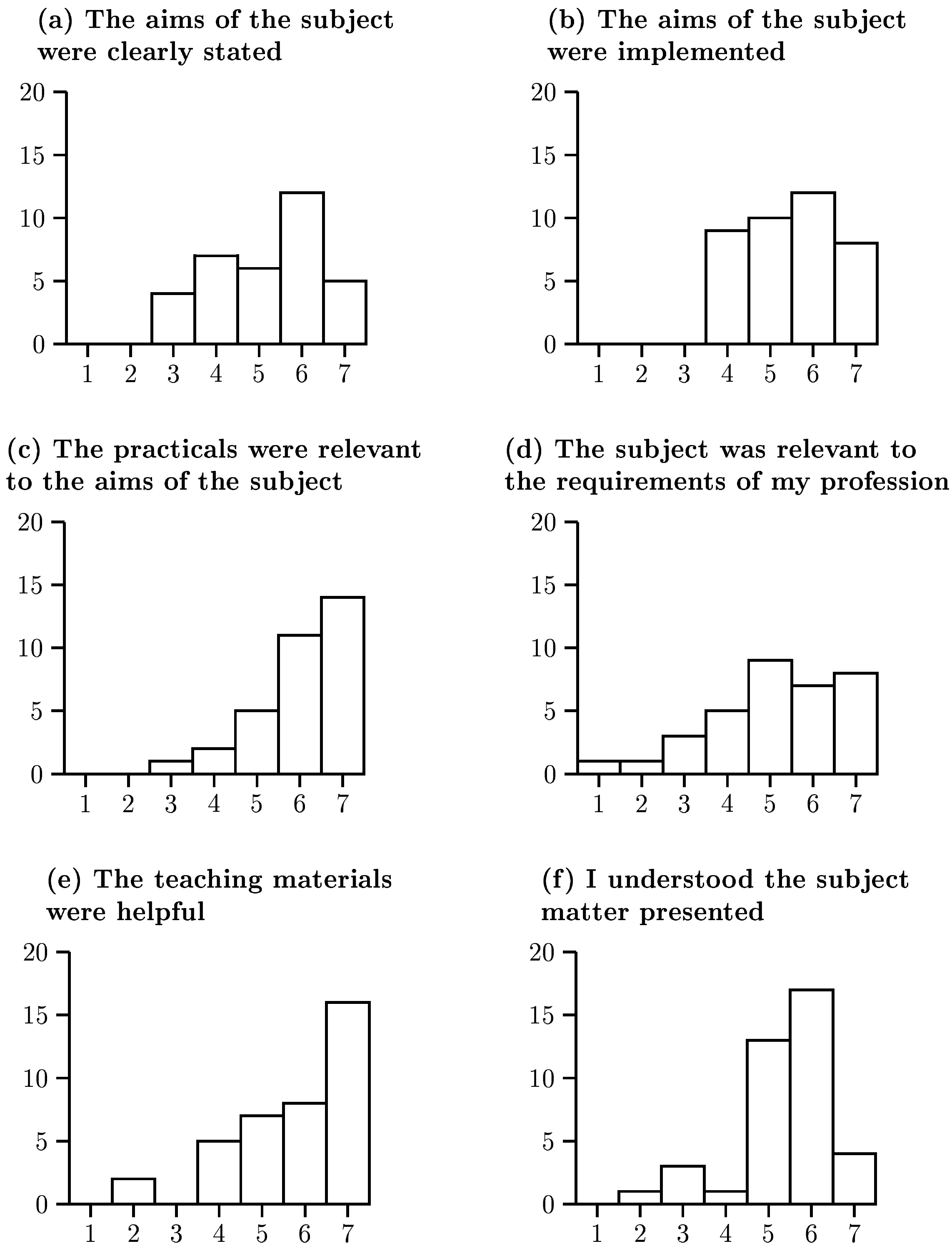
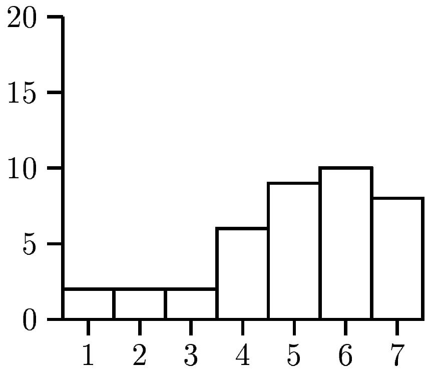
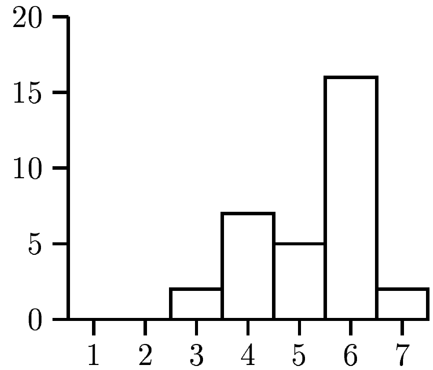


Figure 3 (continued): Student evaluation of teaching

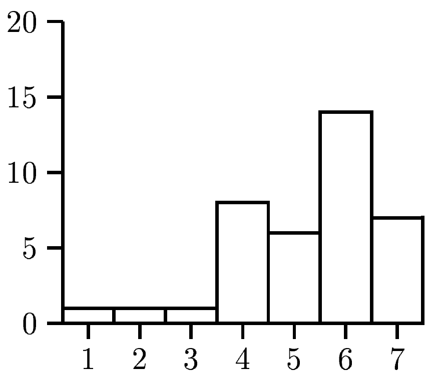
(g) I have a positive attitude to this subject



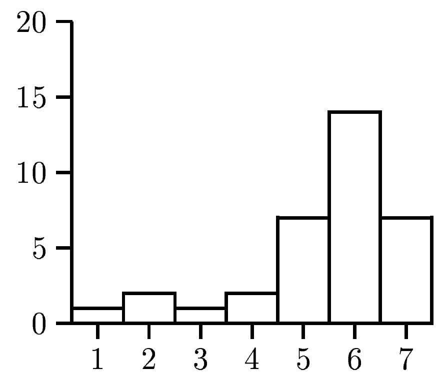
(h) My ability to work independently has been increased



(i) Overall, the assessment of the subject was fair



(j) The assigned work could be completed within the allocated time



Alan J. Branford
Flinders University
GPO Box 2100
Adelaide SA 5001
Australia

Email: alan.branford@flinders.edu.au