# DEVELOPING AND SYNTHESIZING STATISTICAL SKILLS FOR REAL SITUATIONS THROUGH STUDENT PROJECTS

Helen MacGillivray, Queensland University of Technology, Australia

*Much is needed from tertiary introductory statistics subjects, whether mainstream or service. Students need to acquire not only essential statistical concepts and techniques, but also sufficient understanding to form a basis for future development and to enable them to synthesize their knowledge and judgement in tackling data aspects of real situations. To help such learning, the author has developed strategies including own-choice group projects in introductory statistics subjects, particularly in engineering classes of up to 400 students. The benefits of such projects for students and lecturers include the students' sense of ownership and coming face to face with key aspects of statistical problem-tackling and thinking, and the highlighting of classic learning problems in statistics. Challenges include the demands of individual practical problems, assessment and student guidance.*

## IDENTIFYING THE DEMANDS ON THE FIRST TERTIARY STATISTICS SUBJECT

For many reasons, not least being parity and maintenance of options, a mainstream first statistics subject should reasonably parallel service subjects, particularly those in areas like the sciences, engineering and business. This paper considers introductory statistical data analysis subjects of between 40-55 total class hours including laboratories, for students with at least some quantitative, but not necessarily mathematical, orientation. It does not consider subjects with a small component of statistics buried within another area.

No matter what the statistical inclinations of the lecturer, there are key concepts, principles and needs common to all statistical data analysis. Students need to be able to identify types of data, and hence variables, and choose and interpret appropriate methods from their toolkit, or recognise that they need to go beyond their current knowledge. They need to understand that analysis of data is quantifying variation in a justifiable way, often seeking patterns while allowing for variation. Interpretation may be open to debate, but the methods for quantification must be clearly justified and repeatable. They need to see that distributions are at the heart of all statistics and data analysis, and recognise the practical difference between random variables and non-random variables, and between parameters and the random variables we call statistics. This is essentially understanding that all statistical data analysis, irrespective of theoretical leanings, is model-referenced, involving model building, model fitting, model assessing, model diagnosing, model

interpretation, and model building. To be able to apply statistical thinking in real situations, a student needs to be able to "do" all these aspects, even if the tools he or she has are only basic.

It is important to identify what is needed for students' areas and their background, but considering real datasets across disciplines indicates core needs. The basic kit should contain graphical and descriptive tools for discrete (categorical, ordinal and count) and continuous data. The binomial and normal situations should be familiar, and at least one other, preferably two, distributional models considered. For example, it may be important to meet the circumstances of the Poisson and exponential models. There should be tools for estimating parameters, putting errors on estimates and finding interval estimates. There should be tools for comparing two or more groups for the models considered, for example, two or more proportions and means (or medians), and variances, using statistical software if necessary. Tools for assessing models or assumptions about models are needed, and at least some tools for investigating relationships or non-relationships. Perhaps some introductory tools for investigating data over time might be included, and perhaps also touched on for incomplete data. Note that probability enters the above as a servant, but that variables and distributions are pivotal in every way.

The tools given in an introductory subject should be clearly and logically explained and linked, to facilitate good use and interpretation. Explanations do not have to be mathematical to be logical, but their underpinning commonality and sequencing need to be very clear to enable both immediate use and future development - the last now often called "life-long learning". This tool provision aspect of the subject is best served by as many and varied "small" examples for the exposition and logical linking. These examples should involve real contexts and as far as possible real data, but it is the variety of contexts that students need during the process of being introduced to, and getting to know, the toolkit. Occasionally, with different types of student groups, I have trialled gradually building on one or two contexts; the response has each time been a resounding no from the students. They seem never to have enough variety of examples during the "what is this?" stage.

Facilities permitting, inclusion of use of a statistical computing package is essential for many reasons: to enable inclusion of valuable practical tools without algebraic or time implications; to maximise focus on understanding of the use of tools; to facilitate understanding of the concepts and underpinning; and to acquire an introductory

familiarity with a statistical package as a basis for all. Just as in pre-package days, when observers over-noticed the algebra amongst the statistics, now observers focus too much on the package amongst the statistics. The package should be carefully chosen to support and complement the use and understanding of the toolkit, but it is the servant not the master.

So after looking at the students' backgrounds, current and future needs to identify the toolkit possible within time and student constraint; choosing a package; carefully designing a logical, coherent and well-sequenced structure; and building the structure with as many relevant and real examples as possible, what else is needed? Will the students be equipped - in a basic way - to tackle the statistical aspects of real situations? To do this, they need to be able to set up a problem from first thoughts and follow it through, choosing appropriately from their toolkit, interpreting results, and recognising their limitations.

## SYNTHESIZING KNOWLEDGE FOR MEANINGFUL USE; A PROJECT COMPONENT

There would be little disagreement that statistics is not only a doing subject, but that there is a significant step from choosing and using basic tools in real but small, clean examples to making good use of statistics in real scenarios that can be messy, and often vague. Problem based learning is a popular term currently, and those involved in service teaching in some areas will be familiar with the just-in-time philosophy. This latter appears to be little more than an excuse for depriving students of the opportunity to get to know a toolkit before having to use it, based on belief in the existence of some motivational magic that will sweep all before it. Problem based learning on the other hand seems to be a broad term that covers a range of strategies, from inclusion of whole problems in some way to total problem immersion, with single large problems dictating the structure of the subject. The inclusion of whole problems is important in the development of skills for real problems in their entirety, but this must not substitute for meeting the toolkit in an ordered way, and practising use of individual tools. The understanding of this in the STEPS resource contributes to its value for supporting and consolidating understanding, while using whole problems for both motivation and synthesis development.

The concept of own-choice group projects was introduced by the author in 1993 for first year science and mathematics majors, mainly to provide hands-on experience of

the practical aspects of formulation and data collection in real problems, and to start the development of communication skills as early as possible. Trialling the concept in 1994 in an engineering statistics subject with up to 400 students, demonstrated its potential as a vehicle for developing and synthesizing statistical understanding and skills in parallel with the carefully planned structure building up the toolkit. Because the students chose a context or problem of interest to them, they wanted to immediately try the tools they were meeting. Even if their context choice is of only marginal interest initially, most students cannot help becoming involved to at least a certain extent. The exceptions are usually extreme cases.

Hence since 1994, own-choice group projects have been used with considerable success in first year statistical data analysis for science and mathematics students, and for the engineering statistics subject. The project has more weight in the engineering subject because it is given at the second or third year level, and because it is particularly valuable in combatting the deterministic tendencies of engineering students and their general suspicion of statistics. Support resources have been developed and trialled over the past few years with input from students. The principle of own-choice contexts has also been introduced in a smaller project into statistical modelling subjects, with emphasis on observing queueing and other processes that could be Poisson, to assess the assumptions. This has also been very successful in bringing together for students in a meaningful way, theory plus observation plus model assessment.

## SOME PRACTICAL ASPECTS AND THE EFFECTS OF OWN CHOICE

Two key aspects of the success of the scheme seem to be: as indicated above, running the project as a parallel activity to a well-structured coherent course; and the own-choice principle. This principle was intended to give first hand experience of the formulation and data collection stages, and to give a sense of ownership, but the benefits are even greater, mainly due to the absence of any hidden agenda that cannot help but accompany set projects, and to the amazing variety of ideas that could not possibly be dreamt up by one or two people, no matter how experienced. The natural downside to this is the need for guidance and suggestions during choice of context, but this contributes to the whole learning experience. By no later than mid-semester, students are required to submit, in writing, the group names, a brief description of their context or problem, identifying questions of interest, and data collection proposals. Feedback on this

submission should be quick, and concentrate on: is the proposed project feasible for a student project; have the variables been clearly identified; can we measure what we want to; is the data collection process appropriate and practical; should anything else also be considered; and is a trial necessary. Group sizes of three or four are recommended, although two can also work well. Although part-time students often wish to be by themselves, experience suggests trying to guide or suggest practical topics suitable for at least pairs of part-time students.

Initial ideas range from very ambitious to amazingly vague, corresponding to the range from wanting to find out everything about a topic of interest, to vague ideas such as "we want to do something with golf". The formulation can constitute one-third and sometimes more of the project, and the learning involved in the group process of formulation plus interaction with their lecturer and/or tutor, is part of the heart of the project, being both substantial and influential. Observing lecturers and tutors wrestle with students' ideas and with identification of problem and procedures, makes a lasting impression. For example, trying to assess the effect of red light cameras on driver behaviour was very difficult to formulate. It is interesting how many students want to conduct surveys of people, on everything from movie or fast food habits, to transport, to school students' ambitions to gym use to superstitions. Surveys seem to students a straightforward option until they try writing a questionnaire and then dealing with the data. The experience demonstrates well the extent of thought required. I suggested alternatives to one group who had drafted a questionnaire aiming to ask people about their graffiti habits, and another concerned about the morality of M(A) programs.

Using existing datasets may appear to reduce the work but often the opposite occurs, particularly as clearly identifying variables and the method(s) of gathering data may be difficult. If at least one participant was involved in the data gathering, for example in their work, this can be very beneficial, especially in identifying the questions of interest and which aspects of a large dataset to consider, although very large datasets tend to be unsuitable for student projects, requiring too much at the formulation stage. Examples include complete records of computer usage over three campuses, and complete records of boat registrations at every locality in Queensland over many years. However a most interesting range of topics emerge from students' workplaces or their increased consciousness about questions in everyday matters. For what settings does the bag sealer work best? What do we mean by "work best"? Do trains or buses tend to run late due to

chance or outside effects? Which file compressor works best, on which type of file? Do people tend to respect time limits on pickup zones? How does the effect of alcohol depend on individuals? Do graduates' salaries depend on elective choices and grade point average? Just a few examples serve to demonstrate the increased awareness that such questions involve explaining and/or allowing for variation, and that their new tools can actually help in considering such questions. Appreciation of the value of trials or pilots also increases.

## SOME EFFECTS ON TEACHING AND LEARNING THAT HIGHLIGHT KEY TEACHING POINTS

The downsides to the own-choice principle are of course the other side of the upsides: the many different topics and the challenge of formulation of ideas and practicalities, with the associated teaching need to guide without doing. The keys to this are experience in listening to and observing students, sizing up a situation, and asking questions that guide. It is no coincidence that these are key aspects of good statistical or mathematical teaching, and indeed, good consulting. It can also be disheartening but valuable for students to learn that clearcut answers are not the norm, that non-results can be just as important, and that the "mere" formulation of the problem can be a significant part of problem-solving, particularly in tackling variation.

There are a number of classic points that will be found in well-constructed and well-taught courses, but which involvement in own-choice projects highlight. Variables need to be clearly defined with their possible values identified, and classified as categorical, discrete, continuous, with possible distributional models given if appropriate. Continuous variables should not be arbitrarily categorised. The "prettiest" pictures are often the least informative. Data should be eyeballed but in appropriate ways not by every conceivable graph. Similarly, analysis is neither just comments on pictures, nor using every possible inference tool that can be forced on the data. Estimates of parameters are more useful with standard errors or intervals. Normality should be assessed *after* fitting a model, not just *before* fitting a regression or other model, with a comment that it doesn't look normal. And, of course, the results of analysis should always be discussed and summarised in context.

The feedback from four years of projects has justified course content and orientation, but has influenced the order and emphasis within the structure, so as to

increase the highlighting of the above points, and orient the structure more around types of variables, introducing estimation and testing concepts as early as possible. For example, introducing the core ideas of testing through testing goodness-of-fit for categorical and then discrete variables, emphasizes the simplicity of the concept, namely, how consistent are the data with the hypothesized model - how likely were we to obtain such data if the model is true.

## ASSESSMENT AND ITS IMPLICATIONS

Assessment of own-choice projects must not only be criteria-based but also a balance of contributions to the criteria. A project that required extensive thought in the formulation and planning, such as monitoring compliance with time limits in a passenger pickup zone, would receive more weight on the formulation stage than others. Conversely, a project using workplace data that demonstrated very good understanding and use of analysis, would have more weight on the analysis and discussion stage. The criteria should be clear to students, as well as this balanced approach. An essential aspect of the report should be that detail be sufficient for a reader to repeat the procedures and analysis, and that practical problems (including amusing or embarrassing ones) and suggestions for improvement be included. Such aspects are often neglected in real world reports to the frustration of readers. As with guidance, assessment of such projects is demanding, requiring both concentration and situation assessing abilities. Good projects use basic tools wisely rather than sophisticated tools badly. Such use can only happen in the presence of understanding of the tools and the principles underlying the tools. For the students who acquire this, it can be easier in the context of an own-choice project, to help them up a step to tools and ideas not usually possible in an introductory course. Thus, despite the challenges for both students and staff, such projects help not only the development of understanding of the introductory material, but also the building of the basis for future development.