USING DATA, STUDENT EXPERIENCES AND COLLABORATION IN DEVELOPING PROBABILISTIC REASONING AT THE INTRODUCTORY TERTIARY LEVEL

H.L. MacGillivray
Queensland University of Technology, Australia
h.macgillivray@qut.edu.au

In the focus over the past decade on data-driven, realistic approaches to building statistical literacy and data analysis curriculum, the explicit development of probability reasoning beyond coins and dice has received less attention. There are two aspects of probability at the introductory tertiary level: for use in introductory data analysis; and as foundation for further study in statistical modelling and applications, and increasingly in areas in information technology, engineering, finance, health and others. This paper advocates a minimalist objective-oriented approach in the former, and a constructivist, collaborative and data-linked approach in the latter. The latter is the main focus here, with strategies to help students unpack, analyse and extend what they have brought with them to tertiary study, enabling them to consciously develop coherent probabilistic understanding and linking with real investigations and processes.

INTRODUCTION

Research and curriculum development in statistical education over the past decade have focussed on accessing statistical understanding, on statistical literacy and statistical reasoning with emphasis on data (for example, Cobb, 1999; delMas, 2002; Chance, 2002). A considerable portion of this work has been motivated by, and oriented to, introductory tertiary data analysis courses, particularly for non-quantitatively-inclined students (Garfield *et al*., 2002), and using technology (Ben-Zvi, 2000). Calls for statistical educators to consider carefully and in depth the objectives, goals, contexts and content of introductory tertiary statistics courses (Hogg, 1991; Vere-Jones, 1995; Moore, 1997) are echoed in the general higher education literature on objectives (James *et al.*, 2002) and on assessment (Angelo, 1999).

Probabilistic understanding has been part of recent research on: children's development of concepts of statistical reasoning (Watson and Callingham, 2003); how statisticians think and to acquire such thinking skills (Wild and Pfannkuch, 1999); and the strength of intuitions students bring to statistics courses (Konold, 1995) including those linked with public perceptions (Watson, 2004). Although such research has not ignored the development of probabilistic reasoning, the focus has tended to be on its role in the development of inferential reasoning from data.

Sowey (1998) comments that it is difficult to persuade publishers to break new ground, but the continuation in most introductory statistics texts of traditional chapters on probability and random variables placed between exploratory data analysis and introductory inference, and the lack of literature on introductory probability modelling comparable to introductory data analysis, indicate both challenges and opportunities for statistical researchers and educators. Questions from school teachers such as "How can I teach anything in statistics without first doing events and Venn diagrams?" (MacGillivray, 2005) also demonstrate these challenges.

However, despite such questions, the work on data-driven development of statistical thinking has influenced school curriculum across many countries. Although there are some problems in coherent progression from early development into middle and senior school (MacGillivray, 2005), early and prolonged exposure to the collection and exploration of data and the language of chance contribute to a valuable basis that students bring to tertiary study. Such a basis can be used in a constructivist way in developing both statistical and probabilistic reasoning, with *both* linking with data, modelling and real contexts, but with coherent and logical flow that builds strong consistent foundations for all statistics.

This paper gives just one example to illustrate using only necessary probability in introductory inferential data analysis. The paper then focuses on a development of probabilistic and distributional concepts and modelling that helps students analyse their existing skills, connect with what they see around them, link with data, learn to consolidate *and transfer* their mathematical skills into statistics, and work in groups for a range of learning, including a project that collects and analyses data in real stochastic situations. The teaching scenario of this paper involves a separation of inferential data analysis and probability modelling at the introductory tertiary level, as explained in the next section, but whether the requisite scenario is separate or overlapped units, the principles are the same – clear and logical pathways developing probabilistic and statistical thinking, linking with real processes, real data and prior learning.

THE TEACHING CONTEXTS

The extent to which the statistics education "reform movement" has been influencing tertiary introductory data analysis courses for mathematics/statistics majors appears to depend on a country's educational system, and in particular on the extent of separation of "service" courses from those for mathematics/statistics majors. In countries such as the UK, Australia and NZ, the data-driven, holistic approach has been influencing courses for mathematics/statistics majors as much as, and occasionally more than, "service" courses, with the speed and extent of influence depending on local conditions, personnel and management. However, in contrast to statistical data analysis, the challenge of probabilistic development at the introductory tertiary level appears to have received little investigation or airing in discussions.

The practical teaching scenario of the work reported here is a situation in which all mathematics and statistics majors, including education students with mathematics as a teaching area, take as core compulsory courses both the general introductory data analysis course that services all the sciences, and a separate introductory course in probability and distributional modelling. This probability course can be taken before, after, or at the same time as the data analysis course. It can also be an elective in science and other degrees, and some information technology cohorts and approximately half the engineering cohorts take a subset of it.

The data analysis courses in the teaching contexts here include a whole semester group project involving the planning, carrying out, analysing and reporting of a real, many-variabled data investigation in free-choice contexts (MacGillivray, 1998a). The statistical techniques in the course range up to the *use* (via statistical software) of ANOVA and multiple regression. Engineering students take a similar course with project (MacGillivray, 2002), but including some distributions and reliability. The probability course includes a project that collects and analyses data in real stochastic situations. Communication with and about statistics and probability is integrated throughout both courses.

MINIMALIST APPROACH TO PROBABILITY IN INTRODUCTORY DATA ANALYSIS

Moore (1997) argues that "our teaching must avoid the professionals' fallacy of imagining that our first courses are a step in the training of statisticians." In Australia (at least) this is not necessarily the case, as indicated above, but the spirit of Moore's comment is similar to those of James *et al.* (2002) and Angelo (1999) that the objectives of *each* course should be clear to the cohort, including full acknowledgement of cohort diversity, with the course structure, learning experiences and assessment combining in a coherent purposeful program that addresses the objectives of that course. In this spirit, statistical educators should take care that the probability in an introductory data analysis course is oriented to the course objectives and does not take students into unnecessary side paths. One example is given here.

The chi-square tests of a given set of proportions and of independence are ideal vehicles for introducing the concepts of hypothesis testing and p-values, requiring only the basic concepts of probability commonly developed in school syllabi, and separated from the complications of dealing with continuous variables and their parameters, and of the sampling behaviour of the mean and variance. Data from one or two categorical variables are the simplest for students at the introductory level, occur across all disciplines, and this test fills a natural logical place in a structure after handling and exploring different types of data (MacGillivray, 2004). The value of this placement is universal, but was first realised for engineering students (MacGillivray, 1998b), then for Master of Business Administration (MBA) (MacGillivray, 2003). Although different cohorts in their quantitative backgrounds, they shared the characteristic of time-pressured courses, and wanting, as quickly as possible, practical methods to use and interpret in real, non-trivial data, with their focus on understanding being on the procedures "making sense."

All that is needed from probability for comfort with this test are the concepts of chance, of number expected in a category from its probability, and basic concepts of data variation. For the test of independence, estimation of probabilities by relative frequencies and multiplication of probabilities in the case of independence are required. The former is an everyday experience, and at the introductory level, multiplying probabilities tends to be regarded by students as the norm rather than something special. The form of the test statistic "makes sense," as does its "largeness" providing evidence against the null assumption. Interpreting the "smallness" of the chance of obtaining its value or greater is *the* significant conceptual step that requires significant discussion (one that requires constant re-visiting as the course progresses in its building of the statistical toolbox), but using tables or a calculator/computer menu to obtain this chance is a technological rather than conceptual step for today's students. Utts and Heckard (2002) also use this strategy but restricted to twoway tables initially, with the more general situation later.

Feedback and performance from a wide range of both quantitative and non-quantitative cohorts have consistently demonstrated how much easier it is for students to grasp the concept of testing and p-values in these tests than in tests for parameters of continuous variables. A crucial pedagogic point is that using only what is necessary and familiar from prior probability concepts, and avoiding unnecessary digressions into formal notions of probability, maximises students' freedom to assimilate a big concept like "chance of getting our data or more extreme under the assumption of the null situation" that is fundamental to understanding hypothesis testing.

INTRODUCTORY PROBABILITY AND DISTRIBUTIONAL MODELLING

The topics of this course are:

- basic properties and rules of probability; independence and system reliability
- conditional probability, including applications of law of total probability and Bayes
- introductory Markov chains
- general concepts of random variables, distributions, mean, median, variance, quartiles
- special distributional models – their circumstances
- the mathematics and applications of some special discrete and continuous distributions
- introductory queueing processes – modelling and collecting data from them
- goodness-of-fit for discrete (chi-square), and continuous (Kolmogorov) distributions
- introductory bivariate and correlation properties; conditioning arguments.

The course commences with a probability reasoning questionnaire which aims to seed thought and discussion, with some questions taken from Konold (1995), Garfield (2003) and Watson and Callingham (2003), and others based on probability misconceptions seen at the school/tertiary interface. This questionnaire is used later in a class discussion and reflection forum. As the course progresses, each of the above topics has preliminary experiences or exercises or discussion points. These enable students to perceive, analyse and extend the knowledge or skills they bring to the topic, and are much-valued. As one student so aptly put it at the class forum, "Using what we already knew to learn other stuff was really good and helped us learn other stuff." These preliminaries and other strategies in the course have been gradually built in an ongoing research cycle of development, trialling and evaluation, and given as a complete combination in 2004 and 2005. Examples of some of these preliminaries are given below.

Early linking with data in learning probability and distributional modelling is valuable for students to connect with prior experiences and to notice everyday processes with new eyes. Such linking using collaborative learning in this course takes a number of forms, outlined briefly below.

- *The preliminaries for the basic properties and rules of probabilities*

These are in three parts: group activities in quick data collection to estimate probabilities; a group activity attempt to mathematically model the result of throwing a dart non-randomly at a dart board; and a series of incorrect probability statements requiring identification of the mistake.

The probability estimation activities focus on contrasts, such as (a) estimating the probability that a student doing the course lives on the southside and the probability that a student living on the southside takes the course, and (b) data on number of children in a family with at least one child and number of children in an arbitrary "family." The attempt to model throwing a dart non-randomly at a dart board involves realising what "at random" means in a continuum and that mathematical and probability modelling meet in complex real processes. Students can also be surprised in seeing the diversity amongst themselves of individuals' ways of tackling it.

Each of the incorrect probability statements relates to a probability rule (*before* any rules are stated). Except for one, they tend to give few problems and are designed to demonstrate to the students how much they already know. However, a statement such as, "The probability that a family has a TV is 0.89, the probability that a family has a dishwasher is 0.61, and the probability they have both is 0.41" confronts the students with a number of possible prior beliefs they did not realize they held, the most common of which being the automatic multiplication of probabilities.

- *The preliminaries for independence and conditional probability*

The definition of independence is given, and the preliminaries for independence are the series and parallel building blocks of systems/networks reliability, and a simple 3-state, 3-step example of what is essentially a branching process. But it is the preliminaries for conditional probability that surprise the students by the knowledge and skills they already have, handling percentages and generalizing these to probabilities. An example of a % question is:

> Suppose that 5% of people who receive a flu immunisation will still contract flu, and that 35% of those without immunisation will contract flu.
>
> (a) If 20% of people are immunised,
>
> (i) What % of the population will contract flu?
>
> (ii) What % of those who contract flu were immunised?
>
> (b) What % of people should be immunised if the authorities want no more than 20% of people to contract flu?

Analysing and generalizing such questions opens up problems ranging across challenges and contexts, all involving applications of conditional probability. Students are encouraged to suggest and use diagrams that include representations of the sizes of the conditional probabilities.

- *The preliminaries for Markov chains*

The emphasis on Markov chains in the introductory course is on the modelling aspects – the setting up of the matrix. The preliminary class experience is to consider a simple model for levels of discount in car insurance, with movement in level from one year to the next depending on the customer's starting level and the number of claims in that year. For example, a customer

> moves to the next level of discount, or remains on the top level, after a year with no claims;
> moves down a level of discount after a year with exactly one claim;
> moves back to the no discount level after a year with two or more claims.

This also implicitly introduces a probability model for a discrete random variable (number of claims per year) which is picked up in the next topic. An added advantage in this exercise is that it raises the question that the insurance company would be interested in the proportion of customers on the different discount levels, leading naturally to the concept of probabilities in stable conditions.

- *The preliminaries for special distributional models*

These preliminaries are a little different, aiming to provide separation in students' minds between the three aspects of special distributions: assuming a distributional model; the mathematics of the distributions; and using the distributions. The preliminaries consist of information giving names and circumstances for four or five special discrete models, and three or four special continuous models. The students then do computer-based and paper-based examples in which they suggest appropriate distributions (by name) for a variety of circumstances, such as "the number of phone calls to find a plumber who can attend to a job in under a week." No data are involved – the emphasis is on the description of the situation. The small computer module was written and shared with the author more than a decade ago - when it was discovered we approached this topic in the same way - by the Statistics Department, University of Glasgow, where it is planned to develop this module further.

- *Linking with data*

As indicated above, linking with data in estimating probabilities, particularly conditional probabilities, plays a key role in such a course. The other main data activity is a group assignment in which each group collects data on two possible Poisson processes of their choice. For one process they collect data on the number of events in intervals of time (or space), and in the other they collect the time (or distance) between events. They are required to use a combination of appropriate goodness-of-fit tests (the chi-square goodness-of-fit and the Kolmogorov tests are covered in this course) and graphs to investigate and report on the appropriateness of assuming a Poisson model for each process. Hence sequential collection of the data is required to investigate graphically for trends or relationships. Students are encouraged to consider everyday processes around them, and to brainstorm and do small pilot trials to decide, based on practicalities, whether to observe a process via its discrete or its continuous representation. Just a few examples of the students' creative and varied choices are:

> Time between catching waves (surfing); number of people using a beach shower per minute; number of lecturer's "ums" per minute; time between scoring shots in the Australian Rules grand final; the number of times per page the word "Harry" appears in the first 100 pages of a Harry Potter book; the distribution of leaves on tiles; the distance between coffee shops

- *Collaboration*

A number of types of collaboration are key features of learning experiences throughout this type of course. Apart from the group assignment and the class activities, particularly the preliminaries, students are encouraged to work together in tutorials, computer laboratory sessions, and outside class time. More subtle, but just as important in developing comfort and confidence, are class discussions on the roles of the skills and concepts in other disciplines, and on the range of individual difficulties they had met in themselves or others and how these could be helped. The various forms of collaboration contribute significantly to students' feelings of belonging, demonstrated by such examples as students' taking photos during the last class session of 2005.

EVALUATION AND FURTHER RESEARCH QUESTIONS

The course provides an environment in which students are comfortable and forthcoming in ongoing informal feedback. As it was found that using a standard survey for formal feedback resulted in very few student comments, class forums were held in both 2004 and 2005 with a note taker. Discussion reflected on the initial questionnaire, and flowed with seeding assistance in the form of questions such as "Which are the most difficult topics?" The notes were collated and summarized. In 2005, a subsequent survey was given to all students based on opinions expressed at the forum. Students welcomed the opportunity to provide further comments. The preliminaries and the practical group assignment are very highly valued. Generally the problem interpretation aspects of applications of conditional probability and using special distributions are thought most challenging, but also most interesting by many, along with setting up Markov chains.

Remembering that both the data analysis and the probability courses are compulsory for these students, their feedback that they would prefer to do the data analysis course first raises interesting research questions. Noting that the probability course is one of five compulsory first year mathematics courses for mathematics/statistics majors, an interesting measure of the value of the generic skills of analysis, synthesis and problem-tackling developed by the course emerged recently during separate research, when student achievement in this course was found to be best predictor of performance in a second level linear algebra course (MacGillivray and Turner, 2005).

CONCLUSION

The principles used over the past decade in helping statistical data analysis to be more real, relevant and accessible to students, can, and should, be used with introductory probability and distributional modelling, even if the potential student cohorts are much smaller. Connecting and analyzing prior experience, linking with data and real processes, facilitating collaborative work, and ensuring coherent, objective-oriented curriculum all contribute to enabling student ownership of probabilistic understanding and skills that have previously tended to be misplaced, under-represented, or placed in the "too hard basket" at the introductory tertiary level.

REFERENCES

Angelo, T. (1999). Doing assessment as if learning matters most. *Bulletin of the American Association for Higher Education*, http://aahebulletin.com/public/archive/angelomay99.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning*, 2(1-2): 127-155.

Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3), www.amstat.org/publications/jse/v10n3/chance.html.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, 1, 5-43.

delMas, R. (2002). Statistical literacy reasoning and thinking. *Journal of Statistics Education*, 10(3), www.amstat.org/publications/jse/v10n3/delmas.html.

Garfield, J., Hogg, R. V., Schau, C. and Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education*, 10(2), www.amstat.org/publications/jse/v10n2/garfield.html.

Hogg, R. V. (1991). Statistical education: Improvements are badly needed. *The American Statistician*, 45, 342-343.

James, R., McInnis, C., and Devlin, M. (2002). Assessing learning in Australian universities. Melbourne: The University of Melbourne Centre for the Study of Higher Education.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, 3(1), www.amstat.org/publications/jse/v3n1/konold.html.

MacGillivray, H. L. (1998a). Developing and synthesizing statistical skills for real situations through student projects. In L. Pereria-Mendoza, L. S. Kea, T. W. Kee, and W-K. Wong (Eds.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics*, Singapore, (pp. 1149-1155). Voorburg: The Netherlands: International Statistical Institute.

MacGillivray, H. L. (1998b). Statistically empowering engineering students, *Proc. Engineering Maths and Applications Conference*, (pp. 335-338). Institution of Engineers, Australia.

MacGillivray, H. L. (2003). Making statistics significant in a short course for graduates with widely-varying non-statistical backgrounds. *Journal Applied Mathematics and Decision Sciences*, 7(2), 105-113.

MacGillivray, H. L. (2004). *Data Analysis: Introductory Methods in Context* (1st edition). Pearson Education Australia.

MacGillivray, H. L. (2005) Coherent and purposeful development in statistics across the educational spectrum. In G. Burrill and M. Camden (Eds.), *Proceedings of IASE 2004 Roundtable on Curricular Development in Statistics Education*, Lund, Sweden. Voorburg, The Netherlands: International Statistical Institute.

MacGillivray, H. L. and Turner, I. (2005) Components of learning and assessment in linear algebra. *Proceedings of the 5th Delta Conference on Undergraduate Mathematics and Statistics Teaching and Learning*, (to appear).

Moore, D. (1997). New pedagogy and new content: The case of statistics (with discussion). *International Statistical Review*, 65(2), 123-137.

Sowey, E. (1998). Statistics teaching and the textbook – an uneasy alliance. In L. Pereria-Mendoza, L. S. Kea, T. W. Kee, and W-K. Wong (Eds.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics*, Singapore, (pp. 151-158). Voorburg: The Netherlands: International Statistical Institute.

Utts, J. and Heckard, R (2002). *Mind on Statistics* (1st edn). Belmont, CA: Duxbury.

Vere-Jones, D. (1995). The coming of age of statistical education. *International Statistical Review*, 63, 3-23.

Watson, J. and Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3-46.

Watson, J. M. (2004). Quantitative literacy in the media: An arena for problem solving. *Australian Mathematics Teacher*, 60(1), 34-40.

Wild, C. J. and Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, 67(3), 223-265.