TEACHING STATISTICAL ANALYSIS FOR UNIVERSITIES: A PRINCIPLED APPROACH

Su-Chen Wang and Wing-Keung Wong
National University of Singapore, Singapore

The question of HOW data analysis can best be taught calls up two prior questions, and these are WHO will be taught and WHY. This paper mentions some of the principles on which decisions concerning the content and the methodology of any proposed data analysis teaching programme for a school can rationally be made. The principles centre on the learner and on his real world goals.

INTRODUCTION

Today's real world is rapidly becoming an increasingly number-crunching world, and as educationists, we may therefore have to re-think our understanding of what it is for the average schooled person to be numerate in such a world. Not to do so may be to leave our charges unprepared to understand the language of data-analysis that goes on in the real world, as well as to expose them to the possibility of the skullduggery of massaged data machination.

WHY SHOULD WE TEACH DATA ANALYSIS?

The first question that needs to be settled is whether there is a real need for us to design a specific data analysis package for a specific body of learners. A possible answer is that we hope that our custom designed course will serve as a bridge, not necessarily a massive structure, for our target learners to enable them to access their real world.

WHO ARE OUR LEARNERS?

In designing a data analysis package, we need to determine who the learners are in terms of factors such as their age, their perception of whether they need to know how to handle data, their background knowledge, their anxieties (Ellis et al 1993), learning inclinations and the like. In our example, our target students are undergraduates majoring in subjects other than Statistics. These students need to be able to use data statistical procedures in their own fields of study and to interpret the results.

WHAT SHOULD WE TEACH?

The contents of the basic data analysis package will in this instance be governed by at least two factors. First, our learner needs to have a firm understanding of the principles governing data analysis on the basis of which he can make sensible use of the vast array of data analytical tools, which he may otherwise apply indiscriminately (Blejec, 1992). Second, the student may need to know in greater detail specific areas of analytical knowledge that he can use for his major. Here, he needs more in-depth knowledge, which he can apply to concrete problems.

Suppose that we have a Psychology and Social Work student, who wants to investigate the relationship between mothers' educational attainment and their attitudes towards the role of women. We will include ordinal categorical data analysis in our basic data analysis package for him because our student will need to be able to use it to analyse his data (see the example below).

HOW SHOULD WE TEACH?

We suggest that a problem-oriented approach (using problems that the student encounters in his main course work) will be a useful way of introducing him to a basic data analysis course. The example below illustrates how an item can be taught, say, to our Psychology and Social Work student.

EXAMPLE

This is an example of applying ordinal categorical data analysis to investigate the relationship between the mothers' educational attainment and their attitudes towards the role of women in society. It also explores the attitudes towards the role of women as one of the predictors of the mothers' decision to work.

A structured questionnaire was used to collect the data required in the study. A total of 450 mothers were randomly selected and interviewed to gather the data for the study. The fieldwork was conducted by a professional research firm. Out of the 450 respondents, there were 25 who did not provide complete data. As such, their responses were excluded in our data analysis.

The statement in the questionnaire that we analysed in this paper is that "A woman should give up her career to look after her young children". 2 response choices of "Agree" and "Disagree" were provided.

Table 1 shows the profile of the respondents in terms of their highest educational attainment and their attitude. It is noted that about 7.76% of the respondents received less than PSLE level of education, 37.18% received up to PSLE level of education, 41.18% received up to GCE "O" level while 13.88% received at least GCE "A" level and above (which includes tertiary education).

With regard to working status, we can see from Table 2 that 34.35% of the respondents were working full-time (that is 40 hours or more of work per week), 8.94% were working part-time (less than 40 hours of work per week) and 56.71% were not working at all.

Table 1: Sample Composition by Educational Attainment

Educational Attainment		Attitude		Total
		Agree	Disagree	
LESS THAN PSLE	Frequency Percentage Row Percentage	19 4.47 57.58	14 3.29 42.42	33 7.76
PSLE	Frequency Percentage Row Percentage	87 20.47 55.06	71 16.71 44.91	158 37.18
GCE 'O' LEVEL	Frequency Percentage Row Percentage	97 22.82 55.43	78 18.35 44.57	175 41.18
GCE 'A' LEVEL AND ABOVE	Frequency Percentage Row Percentage	22 5.18 37.29	37 8.71 62.71	59 13.88
Total	Frequency Percentage	225 52.94	200 47.06	425 100.00

We apply the goodness-of-fit tests to Tables 1 and 2. The chi-square values obtained are 6.807 and 7.542 respectively while the corresponding p-values are 0.078 and 0.023. Based on these values, we do not reject at 0.10 level of significance but reject at 0.05 level of significance the hypothesis that the mothers' attitudes that women should give up their career for the childcare role is independent of the mothers' educational attainment. The values also lead us to conclude that the mothers' attitudes that women should give up their careers for childcare varies according to the working status of the mothers at 0.05 level of significance.

Table 2: Sample Composition by Working Status

Working Status		Attitude		Total
		Agree	Disagree	
FULL TIME	Frequency Percentage Row Percentage	64 15.06 43.84	82 19.29 56.16	146 34.35
PART TIME	Frequency Percentage Row Percentage	23 5.41 60.53	15 3.53 39.47	38 8.94
NOT WORKING	Frequency Percentage Row Percentage	138 32.47 57.26	103 24.24 42.74	241 56.71
Total	Frequency Percentage	225 52.94	200 47.06	425 100.00

We denote A for attitudes towards the role of women, E for educational attainment and W for working status. To study the joint effect of A, E and W, we apply the loglinear models to fit the data, and find that the best loglinear model is the one that has linear-by-linear A-W and E-W associations.

The model is:

$$\log(m_{ijk}) = \mu + \alpha_i^A + \alpha_j^E + \alpha_k^W + \beta_1(A-\overline{A})(W-\overline{W}) + \beta_2(E-\overline{E})(W-\overline{W}) + e_{ijk}$$

where A=1 for agree, 2 for disagree; E=1 for less than PSLE, 2 for PSLE, 3 for GCE 'O' level, and 4 for GCE 'A' level and above; W=1 for full time, 2 for part-time and 3 for not working; and m is the frequency of the observations. By applying this model, we obtain the results in Table 3.

The Goodness-of-Fit test statistics for this model is 16.46279 with 15 degrees of freedom and the corresponding p-value is 0.352. The A-E association is not significant at the 0.10 level and hence we exclude it in the model. Both the E-W association (t value = 4.3098) and the A-W association (t value = 2.4754) are significant at the 0.05 and hence they are included in the model.

Table 3 : Estimates for Parameters

Parameter	Coefficient	Std. Error
α_i^A	-.23066	.12668
α_j^E	-1.913360 .33268	.28054 .10412

	1.05656	.12286
α_k^W	-.84733 -1.04238	.25804 .11402
β_1	-.26098	.10543
β_1	-.27540	.06739

A COMMON PROBLEM ENCOUNTERED IN TEACHING CATEGORICAL DATA ANALYSIS

Often we have students whose Mathematics background is poor and who are not particularly interested in Statistics, except for when they need to use it. Such students may misuse statistics. For example, many students will apply regression modelling technique to the example given above. However, the observations are not from a continuous distribution and the length of interval between two values is not meaningful. So, regression modelling is not appropriate for this particular problem.

THE TEACHING APPROACH

We should probably have to teach them how to use a computer package, such as SAS, MINITAB or SPSS, use graphic simulations (Gordon and Gordon, 1992), and give them general guidelines to help them select the most suitable statistical procedures for the problems they are likely to encounter. We propose teaching them basic concepts such as Goodness-of-Fit, breach of appropriacy conditions, level of significance, etc. For the categorical data analysis, we need to teach them how to interpret odds ratios and the estimates of the parameters.

Concerning the question of HOW we should teach these students, we propose that we skip teaching them difficult mathematical proof as far as possible. Instead, we could concentrate on teaching the practical applications of statistical tools, as well as teach them to consult the statistician for advice on matters of appropriacy and interpretation.

REFERENCES

Blejec, A. (1992). Teaching of Statistics --- Why, What and How?. In L. Pereira-Mendoza, (Ed), Introducing Data Analysis in the Schools: Who Should Teach It and How? *Proceedings of the International Statistical Institute Round Table Conference, International Association for Statistical Education.* Lennoxville, Quebec, Canada.. The International Statistical Institute with the financial assistance of UNESCO.

Ellis, H. C., Varner, L. J. and Becker A. S. (1993). Cognition and Emotion: Theories, Implications, and Educational Applications. In L. A.Penner, G. M. Batsche, H.d M. Knoff and D. L. Nelson, *The Challenge in Mathematics and Science Education, Psychology's Response.* American Psychological Association: Washington, DC.

Gordon, F. S. and Gordon S. P. (1992). Sampling + Simulation = Statistical Understanding: Computer Graphics Simulations of Sampling Distributions. In F. S. Gordon, and S. P. Gordon (Eds.), *Statistics for the Twenty-First Century*. MAA Notes, Number 26, The Mathematical Association of America.