Some Undergraduates' Understanding of the Meaning of a Correlation Coefficient

John Truran

University of Adelaide

Abstract

The paper examines the responses of more than 300 non-naïve students of first-year Business Data Analysis to a question asking what information is provided by a value of r *in a certain context. It shows that few students appreciate that they can find information about both association and variance from this parameter. Many responses were either incomplete or aberrant, and a summary of the most significant responses is provided.*

Introduction

The National Statement on Mathematics for Australian Schools, (AEC, 1991) with its accompanying document *Curriculum Profile for Australian Schools* (AEC, 1994) have made stochastics, erroneously referred to as 'Chance and Data' (Truran, 1994), a significant part of the school mathematics curriculum in both primary and secondary schools.

Statistical inference is seen as an important part of this course. 'Interpreting Data' constitutes one aspect of the Chance & Data strand from Level 2, although only reading and describing skills are required from Levels 2 to 5. However, in Level 6 students who might be expected to be aged about 12 or 13 are expected to '[interpret] collected and published data from tables, diagrams, graphs, plots and summary statistics and [report] on data collection processes and results' (AEC, 1994, p. 13). This interpretation includes '[reporting] on what their own displays and summary statistics show about similarities and differences between two data sets' (AEC, 1994, p. 109. In Level 7 students are required to '[interpret] scatter plots, considering whether there is positive, negative or no association ...' (AEC, 1994, p. 125).

It would be reasonable to assert that in general throughout Australia appropriate pedagogies for teaching stochastics are usually poorly developed in primary schools and often poorly developed in secondary schools. In particular, such pedagogies as do exist tend to emphasise the deterministic aspects of the topic, rather than the non-deterministic, interpretative aspects. So children are provided with lots of experience in calculating relevant parameters, but much less experience in making reasonable interpretations of the values they have produced.

Students' understanding of statistical ideas has only recently attracted the interest of research workers and much more work remains to be done (Shaughnessy, 1992). However, the topic is starting to attract attention among research workers, often those concerned with teaching statistics to mathematically under-prepared undergraduates.

For example Lipson (1994) has found little correlation between a tertiary student's ability to carry out a standard hypothesis test and his or her capacity to explain what is being done. There is some debate, summarised in Pfannkuch & Brown (1994) about whether a probabilistic approach to understanding statistics is a help or a hindrance.

The work reported here is a contribution to building up a wide appreciation of how students interpret statistical parameters. It is related to work done by Estapa & Batanero (1994) which assessed secondary children's interpretation of scatter diagrams and which isolated a number of incorrect strategies adopted by students. This investigation focuses on the meanings which non-naïve undergraduates seem to have for the correlation coefficient

$$r = \frac{\text{cov}\,(x\,,y\,)}{s_x\, s_y}$$

when presented with an open-ended question in a written examination at the end of an introductory one-semester course in data analysis.

Source of Data

This paper analyses the responses of 304 examinees to part (b) of question 3 in the examination in Business Data Analysis 1 (9101) held at the University of Adelaide in November 1994. There were 26 students (8·6%) who made no attempt at question 3 (b). The three hour examination came after instruction lasting for one semester which was a traditional balance of lectures, class exercises, tutorials and workshops. The course was co-ordinated and lectured by an experienced, senior member of the Department of Economics.

Analysis of student numbers makes it clear that at least 257 were in their first year at a tertiary institution. Their mathematical background is not known, but anecdotal evidence suggests that it is very varied. Some will have studied a formal academic mathematics course in secondary school, some applied or business mathematics. Their levels of achievement will have varied substantially. Those who studied formal courses in South Australia will have had little experience of stochastics. Formal courses from other states may well have provided them with counter-productive experiences (Truran, 1992, pp. 122 - 125).

For these reasons it is reasonable to assume that in most cases whatever these students understand about correlation coefficients has come either from the course itself or from their own less formalised experience which may well not have included discussion of correlation coefficients. The course will have provided their primary and most authoritative source.

Questions 3 was as follows:

The table below gives descriptive statistics for a sample of sales of food and other items in a supermarket.

	food	*other*
Mean	28.64091	9.682727
Standard Error	7.813259	3.444853
Median	14.99	8.12
Mode	#N/A	#N/A
Standard Deviation	25.91365	11.42529
Sample Variance	671.5171	130.5371
Kurtosis	-1.13184	7.294715
Skewness	0.900091	2.527169
Range	62.5	40.69
Minimum	6.49	1
Maximum	68.99	41.69
Sum	315.05	106.51
Count	11	11
Confidence Level(95.000%)	15.31368	6.751778

The correlation coefficient between the two variables has been calculated as: r = 0.636

The manager is interested in the size and variability of the transactions and also in what relationship there may be between the amount customers spend on food and on what they spend on other items.

(a) What is the average amount spent on food? Give a 95% confidence interval for the true mean.

(b) Explain what the correlation coefficient tells the manager about the relationship between the two types of expenditure.

(c) Calculate a linear regression which describes the amount spent on 'other' items as a function of the amount spent on food.

(d) If someone comes in and spends $30.00 on food, how much on average will they be likely to spend on other items?

The question came sufficiently early in the paper to assume that all students who felt able to answer the question were not hindered by time constraints from presenting an answer. An extensive formula sheet was provided during the examination which included three formulae for r, two for the covariance of x and y, and also the relation $b = r \frac{s_y}{s_x}$. Seven marks were awarded for the question out of 87 for the whole examination. Calculators could be used.

One mark was awarded for part (b) of this question, which is all we are concerned with in this paper. Half of this was awarded for stating that $r = 0.636$ indicated a positive relationship between the amounts spent on the two types of commodities. The other half was awarded for calculating r^2 and stating that about 40% of the variance in expenditure on food is explained by the variance in expenditure on other items. Awareness of the symmetry of this relation was rarely stated and not required for full marks.

This question may be seen as a 'pure' question dressed up in 'economic' clothing. The specific context in which it might be economically sound for a manager to allocate time to interpret the correlation coefficient is not specified. Students responded in terms of the supermarket environment, but their answers were almost always 'pure' in form rather than 'applied'. However, the open-ended nature of the question has meant that the examinees' responses provide a valid indication of what they knew and thought was relevant to any interpretation of the data. Most wrote at length; it is reasonable to assume that they said all that they believed was relevant.

Method of Analysis

Marking of the examination scripts suggest the structure for a spreadsheet which recorded:

Personal Details

ID number (only the first two digits were retained as a measure of the first year of enrolment at the University of Adelaide.)

Total score in examination

Gender

Information about r

Awareness that r indicates a general association

Awareness that the sign of r is significant

Belief about the strength of the association between the variables

Information about r^2

Calculation of r^2

Belief that r^2 explains a fraction of

variance

movement

variables

Whether the student saw correlation as measure of causation

Any other responses

This classification proved robust and easy to apply without an unwieldy number of cases needing to be listed in the final section. It has enabled a fairly straightforward listing of the types of responses made (especially those which indicate misconceptions) together with some indication of the relative frequency of these responses among the chosen population. What I have judged to be the most important of these responses are now discussed.

Belief that Correlation Measures Cause rather than Association

The eradication of this widely held myth must be one of the prime aims of every teacher of introductory statistics. Even so, 33 students (11·9% of respondents) gave interpretations which used words like 'affect'. While very few of these answers provided blatantly wrong statements like 'the amount spent on food is a direct cause of the amount spent on other items', the complex sentences which students tended to compose made it easy for expressions of causation to slip in unrecognised.

Interpretation of *r* as a Measure of Association

There were 168 students who stated that r was a measure of general association between the variables. Only 150 of these (89·3%) observed that the association was positive. Four explicitly stated that a positive value of r meant that a rise in x implied a fall in y . Two explicitly stated that r was between 0 and 1. So less than half the class saw it as important that the variables were positively correlated. What a professional statistician would see as obviously relevant is not seen as such by many non-naïve students.

The course had not provided any ways of assessing the significance of r . However 78 students provided some verbal measure of how strong they believed this association was. I summarised these on a four-point scale where the level of association was described by the terms Significant, Strong, Fairly Strong, Moderate and Small. Two students used the word 'significant', probably in its vernacular, rather than its technical meaning. Four used the word "strong', 29 the words 'fairly strong', 21 'moderate' and 16 'small'. Only two students gave any indication of have any rule-based algorithm. This was that if r were greater than 0·5 then for one student the association was strong and for the other student the association existed. An unrecorded small number of students seemed to believed that if r^2 were greater than 0·5 the association would be significant, but they did not make this rule explicit. No student mentioned the size of the samples as relevant for assessing the significance of r .

Even allowing for the imprecision of language and the difficulties in classifying some of the responses it is clear that there is no agreement among the students about what is meant by a strong or significant association. The wide diversity of intuitive interpretations placed on r strongly suggests that this issue needs to be addressed formally within the course.

Aberrant Interpretations of *r*

None of these occurred in large numbers, but their presence gives some important clues about students' intuitive interpretations. The most common (eight cases) was to see r as equivalent to the slope of the regression line. Although r is symmetrical about x and y the slope was almost always seen as referring to x and y in their conventional orientations. An unusual variation was to argue if food consumption increased by 10% then consumption on other items would increase on average by 6.36%. But one student believed that a change in other items would cause a 63.6% change in food.

Others interpretations were less predictable. One student argued that 63.6% of customers bought both food and other items from the store while 36.4% of customers bought only food. Another saw r as a measure of the number of times the two expenditures are related. Another, more precisely, argued that 64% of the time customers bought food they would buy other goods. One student believed that r could tell the manager how much was spent on each product. One believed that r gave the minimum sum of squared deviations from the regression line while another saw a high correlation as indicating a large deviation from the regression line.

These answers were fairly deterministic, but some clearly saw r as a probabilistic measure. One argued (though not in these words) that r was the conditional probability that people would buy other goods, given that they bought food. Another argued that r was the conditional probability that there would be a change in expenditure on other goods, given that there was a change in expenditure on food while yet another saw it the other way round.

Understanding of the Relevance of r^2

One of the reasons why the sign of r was not mentioned may have been because the course had also emphasised the interpretation of r^2, which is of course always positive and always between 0 and 1.

There were 68 examinees (24·6% of respondents) who realised that r^2 was a relevant parameter, and of these 62 actually calculated its value. However, there was great confusion about what its relevance was. There were 38 of the 68 (55·9%) who realised that it was a measure of what fraction of the variance of one variable could be explained by the variance of the other. But some of students may have confused variance and variable and wrote answers which were not entirely consistent but which would be accepted in an examination context. Some students used the term 'variation' where 'variance' was meant. My data do not isolate either of these cases, but if mathematics is 'meaning what you say and saying what you mean', then the distinctions are important and may be an unrecognised source of confusion for some.

There were 11 (16·2% of those who considered r^2) who believed that r^2 was a measure of how much the *change* in one variable could be explained by the change in the other and of these 3 also believed that it was a measure of how much variance could be explained. This provides further evidence that the technical meaning of variance may not be well understood. There were 17 (25% of those who considered r^2) who believed that r^2 was a measure of how much the amount of one variable could be explained by the amount of the other and 1 who believed that it was a measure of both the variance and the variable. So nearly half of the students who realised that r^2 was important had serious misconceptions about the nature of its relevance.

Furthermore eight respondents believed that it was r, not r^2 which indicated what fraction of some aspect of one variable could be explained by the same aspect of the other variable. Of these, five connected r with the variances of the variables, and three with the amounts of the variables.

Aberrant Interpretations of r^2

A small number of students saw r^2 as the slope of the regression line; presumably they had not consulted their formula sheet. Another saw it as the percentage spent on food compared with other items.

Some saw r^2 as being a probabilistic measure. One argued that there was a 40% probability that other items would be bought, another that there was a 40% chance that expenditure on x or y would affect expenditure on the other.

Conclusion

Little is known about school children's intuitive understandings of correlation. It is not possible to conclude that the understandings expressed by non-naïve tertiary students after instruction will be the same as those expressed by naïve primary and secondary students. But the tertiary students have shown misconceptions and restricted understandings which are sufficiently frequent to argue that the findings of this paper constitute a fruitful basis for investigations of understandings held by younger children.

In particular, mathematical symbolism is information rich. Only 15 students scored full marks on this part—i.e., only 15 saw that *r* yield information about both association and variance. The data presented here suggest that there is a *prima facie* case for explicitly teaching students that parameters are capable of providing a variety of types of information, and that they should not assume that parameters are information-poor.

Acknowledgements

I should like to express my appreciation to Mrs Margaret Meyler, Deputy Head of the Department of Economics, University of Adelaide, for allowing me to use her question and answer as a basis for this paper, and for allowing me access to the students' answers. I also thank Dr Brian Sherman, Department of Education, University of Adelaide, for some helpful comments on the nature of the investigation being conducted here.

References

Australian Education Council (1991). A National Statement on Mathematics for Australian Schools. Carlton, Victoria: Curriculum Corporation.

Australian Education Council (1994). Mathematics—A Curriculum Profile for Australian Schools. Carlton, Victoria: Curriculum Corporation.

Estapa, A. & Batanero, C. (1994). Judgements of Association in Scatterplots: An Empirical Study of Students' Strategies and Preconceptions. In Joan Garfield (ed.) (1994) Research Papers from the Fourth International Conference on Teaching Statistics (ICOTS 4). Minneapolis, Minnesota: International Study Group for Research on Learning Probability and Statistics

Lipson, Kay (1994). Assessing Understanding in Statistics. In Joan Garfield (ed.) (1994). Research Papers from the Fourth International Conference on Teaching Statistics (ICOTS 4). Minneapolis, Minnesota: International Study Group for Research on Learning Probability and Statistics.

Pfannkuch, Maxine & Brown, Constance M. (1994). Building on and Challenging Students' Intuitions about Probability: Can We Improve Undergraduate Learning? in Joan Garfield (ed.) (1994). Research Papers from the Fourth International Conference on Teaching Statistics (ICOTS 4). Minneapolis, Minnesota: International Study Group for Research on Learning Probability and Statistics.

Shaughnessy, J. Michael (1992). Research in Probability and Statistics: Reflections and Directions. in D.A. Grouws (ed.) (1992). Handbook of Research on Mathematics Teaching and Learning (pp. 465- 494). New York: Macmillan.

Truran, John M. (1992). The Development of Children's Understanding of Probability and the Application of Research to Classroom Practice. M.Ed. Thesis: University of Adelaide.

Truran, John M. (1994). Chance & Data Or Probability & Statistics - Are We Oversimplifying? Möbius, 21, 46 - 48.