THE ROLE OF COMPUTER BASED TECHNOLOGY IN DEVELOPING UNDERSTANDING OF THE CONCEPT OF SAMPLING DISTRIBUTION ®

Kay Lipson
Swinburne University of Technology
Australia

Traditionally, the concept of sampling distribution has been seen as fundamental to an understanding of introductory statistical inference. As a result many computer packages have been developed which offer activities intended to support the development of this concept. However, we need to recognise that the concept of sampling distribution is complex and multi-faceted, with many different mathematical and symbolic representations possible. Computer simulations of the sampling distribution tend to address only the empirical representation of this concept, and leave the linking of representations to the user. And it is the development of these links which is critical to the development of understanding in statistical inference. This paper reports some results of a study analysing the role of the computer-based technology in the development of understanding of sampling distribution.

INTRODUCTION

Two of the most important underlying abstractions in inferential statistics are the concepts of *population* and *sample*. The behaviour of a population is often described by a mathematical model, known as a *probability distribution*. This model is then used to make predictions concerning the nature of a sample to be selected. The properties of a sample are described by *statistics*. It is the information contained in these statistics, together with knowledge of the behavior of the probability distribution used to model the population that underlies the process of *statistical inference*. That is, that inferences are made about the properties of a population, usually unknown, based on the information contained in the sample, which is known.

Statistical inference requires students to recognise that the sample with which they are working is just one of a potentially infinite set of samples which may be drawn from that population. The student then needs to appreciate that, in order to make an inference, the distribution of all such samples must be known, or modelled. The distribution of a sample statistic is called a *sampling distribution*, and this is a key concept in the study of statistical inference. Many statistics educators (for example Rubin, Bruce, & Tenney, 1990; Shaughnessy 1992; and Tversky & Kahneman, 1971) have suggested that the sampling distribution is a core idea in the understanding of statistical inference, something that many teachers of the subject have intuitively recognised. One need only look at the proliferation of computer activities dedicated to the Central Limit Theorem to confirm this. Yet, despite its critical role in understanding inference, experience and research have shown that the idea is generally poorly understood (for example, Cox & Mouw, 1992).

Why has the sampling distribution proved so difficult for students? Because the concept of the sampling distribution is multi-faceted and complex, being associated with both the selected sample and the dynamic process of sampling, there are many images that can be, and arguably should be, associated with a schema for sampling distribution. Thus it follows that there are a range of mathematical, symbolic and computer generated representations that can be associated with a concept. Those of particular relevance here, the alternative mathematical representations of sampling distribution, are discussed in the next section.

REPRESENTATIONS OF THE SAMPLING DISTRIBUTION

A particularly important idea in the development of an understanding of statistical inference is the recognition that when samples are drawn from a population they will vary, and that this variation will conform to a predictable pattern. This idea has been traditionally introduced in statistics courses using a deductive approach based on probability theory (for example Johnson & Bhattacharyya, 1992). Such explanations are usually expressed in a highly

symbolic form which tends to make the argument inaccessible to all but the mathematically able, now a very small minority of the students taking introductory courses in inferential statistics. But perhaps more importantly, the sampling distribution described by a probability density curve is a theoretical development that is difficult to relate to the actual physical process of drawing a sample from a population.

Educators have come to recognise that there are deficiencies with a random variable based explanation and it is now often replaced with the long run relative frequency argument. The sampling distribution is viewed as the result of taking repeated samples of a fixed size from a population where the value of the sample statistic is calculated for each sample and then the distribution of the sample statistic values formed. It is here that many have seen potential for computer based sampling, with the dynamic visualisations which modern technology is able to offer. In this approach, the probability density function which forms the theoretical sampling distribution is seen as the limit of the empirical sampling distribution.

Thus, sampling distributions may be defined through either of two alternative statistical arguments, one from a theoretical, random variable perspective and the other from an empirical perspective. Whichever approach is used, and despite the closer relationship of the relative frequency approach to the sampling process, the end product of the discussion on the sampling distribution is generally seen as a probability density curve. This theoretical distribution can usually be described mathematically in terms of a well-known distribution, such as the normal distribution. Once the idea of a theoretical sampling distribution has been established, students are generally not again reminded of the link between the sample statistic and the empirical form of the sampling distribution which arose from the sampling process. Determination of the P-value becomes an abstract exercise in the calculation of the probability in the tail of a probability density curve. And any cognitive link which was established between the process of hypothesis testing and the sampling process is possibly unlikely to remain over time, as it is no longer made explicit.

The empirical approach to the development of the sampling distribution sees the sampling distribution as arising from repetitive sampling, has the advantage of being more readily related to the actual physical process of sampling than the theoretical approach, which derives the sampling distribution as a probability density curve. Analysis of the concepts associated with statistical inference show that it is this view of a sampling distribution, as a way of describing the sampling variability of a sample statistic, that is needed to interpret the results of the inference. From this perspective the observed value of a sample statistic is readily viewed as one of many possible values, some of which were more likely to be observed in a sample drawn from a particular population than others. Similarly, a confidence interval is seen as a variable interval estimate of the population parameter, which is quite likely to contain that parameter.

Thus, the empirical view of sampling distribution is an essential component of a schema for sampling distribution which facilitates understanding of statistical inference. It is important then for both the empirical and theoretical mathematical representations of sampling distribution to be part of a student's schema. And, more than this, it is desirable that the schema associated with sampling distribution contain links between these two mathematical representations, so that whichever part of the schema is activated, the alternative mathematical representation is available.

From the point of view of understanding the process of statistical inference, it is useful for the sampling distribution to be seen by students as the distribution of a sample statistic, based on the observation of many, many samples, which can be modelled by a particular theoretical probability distribution under certain assumptions (rather than just as a probability density curve).

THE STUDY

This paper reports in part the results of a study carried out to investigate the development of student understanding in statistical inference. The participants in the study were twenty- three part-time mature age students with little or no mathematical background. These subjects were exposed to an instructional treatment that was designed to encourage the development of the concept of sampling distribution and to facilitate the formation of links between the sampling distribution and hypothesis testing in the students' schemas. In particular, the computer sampling packages *Sampling Laboratory* (Rubin, 1990) and *Models'n'Data* (Stirling, 1991) were key

components of the teaching and learning strategy employed. Both packages provide dynamic linked representations of the sampling process and the formation of the sampling distribution.

EVALUATING UNDERSTANDING OF THE SAMPLING DISTRIBUTION

Prior to the study concept mapping exercises were carried out by the researcher and a colleague in order to deconstruct the structural knowledge implicit in a study of statistical inference, and to identify important links between key concepts in the schema constructed by experts. Previous research has shown that as students learn, the schema they create becomes closer in structure to those of their instructors, and thus that the students' knowledge structure can be evaluated by comparing the students' maps with the expert maps (Jonassen, Beissner & Yacci, 1993). Several concept maps were constructed by both of the experts for various stages of the course, and, by a process of negotiation, the maps arrived at met with unanimous approval. These concept maps were then termed the *expert* maps, in that they exhibited all the key features required at that particular stage in terms of concepts and connections between concepts included. From these expert maps, certain propositions could be identified, which summarised both the knowledge domain and the connections between aspects of knowledge, which identify a connected schema.

To investigate the students' knowledge development they constructed several concept maps using sets of terms that were provided by the researcher. Six maps were completed over a six-week period, as follows:

Map 1 Concerned with the sampling distribution of the sample proportion.
Map 2 Concerned with the sampling distribution of the sample mean.
Map 3 Concerned with the sampling distribution.
Map 4 Concerned with the hypothesis testing.
Map 5 Concerned with the estimation.
Map 6 Concerned with the statistical inference.

A list of terms used as a starting point for each of these maps was given to the students. The purpose of the concept mapping exercises was to document the students' schemas at particular points in time. This would enable the researcher to identify in the maps the propositions formed by relating the terms given and which indicate understanding of particular statistical concepts by a student. Over a period of time, any changes in the nature of this understanding could be documented by analysis of the sequence of maps. Those maps concerned with the sampling distribution were constructed by students after they had participated in intensive computer simulation sessions, where sampling distributions where generated dynamically and their key feature discussed. Analysis of the students' maps allowed general trends to be identified concerning relationships between concepts that were generally recognised or generally not recognised, and any overall group features. The analysis shows that the students' conceptual structures are highly individual, showing varying degrees of cognitive reconciliation, and also points to some misconceptions which may be held by individual students.

THE RESULTS OF THE STUDY

The purpose the concept maps was to give the researcher some insights into the student schemas for the sampling distribution. By comparing the sequence of maps for an individual student, the development of links between concepts could be determined (as well as the loss of links in some cases). By comparing the sequences of maps between students, the researcher could not only document the variety of schema constructed, but also identify any similarities between students which could lead to the determination of a classification of stages of understanding. Whilst qualitative analysis of the maps was their main purpose, the percentage of students who included particular key propositions (those developed by the content experts) in each of the concept maps in given in Table 1, for information. Note that the number of students who completed each map varies, particularly for Map 2. The students were asked to complete their map in class each week, except for week 2 when they were asked to complete it at home and submit it in class the next week. The poor rate of return of this map ensured that this strategy was not used again.

Table 1
Summary of Propositions

Key Propositions	Map 1 n=21	Map 2 n=9	Map 3 n=23	Map 4 n=21	Map 5 n=22	Map 6 n=23
Populations give rise to samples.	95	78	87	71	77	91
Population distributions are described by parameters.	90	89	91	62	9	26
Parameters are constant.	19	44	91	10	9	0
Sample distributions are described by statistics.	71	67	83	71	73	74
Statistics are variable.	33	44	61	48	0	13
Sample statistics form a distribution known as a sampling distribution.	67	100	43	43	27	48
The sampling distribution of the sample statistic can be modelled by a known probability distribution.	43	33	9	5		
The sampling distribution of the sample statistic is characterised by shape, centre, spread.	52	56	48			
The spread of the sampling distribution is related to the sample size.	10	67	35			
The sampling distribution is centred at the population parameter.	5	22	9			
Hypothesis testing is about populations.				43		39
Hypotheses are about parameters				43		13
The test statistic is formed from the sample statistic				57		
The exact sampling distribution depends on the null hypothesis				5		
The test statistic and its sampling distribution together help to determine the P-value.				19		30
The P-value depends on the alternate hypothesis				0		
A decision is based on comparing the P-value to the significance level				48		70
The decision is concerned with the null hypothesis				24		4
Estimation is concerned with population parameters					41	22
Point estimates for a parameter is the sample statistic					64	39
Knowledge of the sampling distribution enables us to calculate an interval estimate.					14	13
This interval estimate is called a confidence interval.					68	26
Statistical inference is concerned with both hypothesis testing and estimation						78
Both aspect of inference are concerned with knowing more about a population parameters						30
Consideration of the confidence interval is an equivalent act, leading to the same conclusion.						26

Whilst generalisations from samples of size 23 or so are questionable at best, it can be seen from the table that there were some key propositions concerning samples and populations which were included by most students. For example, consider Map 3, the map for the general sampling distribution, which required students to synthesise the features of a sampling distribution. Most students understood that there was a relationship between populations and samples (87%), that parameter is the term used for a measure which describes population distributions (91%), and statistic is the term used for a measure which describes sample distributions (83%). As well, 61% of students went further and correctly identified sample statistics as variable, a fundamental concept when building a schema for sampling distribution.

Qualitative analysis of the maps (Lipson, 2000) showed that or some students knowing that sample statistics were variables led to propositions concerning distributions and their features. On the basis of this analysis student were able to be divided into three groups. These were:

- Those who correctly linked the sampling distribution as the distribution of the sample statistic (43%). These students indicated in their maps that the sample statistic was determined from the sample, and that the variability of the sample statistic could be described by the sampling distribution.
- Those who incorrectly designated the distribution of the sample (the sample distribution) as the sampling distribution (22%). For these students the sampling distribution described the distribution of the sample, and sample statistics were calculated to summarise features of the sampling distribution.
- Those who did not clearly indicate whether the sampling distribution was describing the distribution of the sample or the sample statistic (35%). This points to possible confusion on their part over the basis of the sampling distribution.

This confusion between sample distribution and sampling distribution, which occurred early in the development of the concept of sampling distribution for many students, proved to be startlingly common. Whilst earlier research has identified student misconceptions with sampling this confusion as a source of the problem has not previously been specifically stated, and thus provides an insight into student thinking which was not anticipated by the researcher.

Since the students prepared these maps over a six-week period, they evidence changes in the students' conceptual structures over time. Whilst the percentage of students including some propositions remained reasonably constant, others showed large increases or decreases over time. For example, the percentage of students recognising that the sampling distribution is characterised by shape centre and spread was fairly constant at around 50% on maps 1, 2 and 3. However, the percentage of students knowing that the spread of the sampling distribution is related to the sample size was only 10% on map 1, rose to a high of 67% on map 2 and then dropped back to 35% on map 3. This observation supports the contention of Vinner (1983) that the student concept image is dynamic, and does not necessarily retain all the desired features over time unless attention is paid to these features. For example, a necessary concept for the interpretation of a P-value as obtained from a standard hypothesis test such as a *t*-test is that the sampling distribution can be modelled by a known theoretical probability distribution. The students in this study participated in an instructional sequence which was designed to establish and reinforce this concept, and thus from an educator's viewpoint it can be said that attention was continued to be paid to important features of the student schema. However, it can be clearly seen that whilst 43% of students included the appropriate proposition in their maps after the computer based session in which this relationship was recognised, only 9% included this key concept in the maps constructed later in the semester.

During this instructional strategy emphasis was given to the nature of the sampling distribution. From Table 1 it may be seen from the figures for Map 3 that, whilst 48% of the students noted that the sampling distribution was characterised by shape, centre and spread, only 35% related the spread of the sampling distribution to sample size and a very small 9% noted that the sampling distribution was centred at the value of the population parameter. Identification of the centre and spread of the sampling distribution, and the relationship of these to both the population and the sample size are fundamental for both the application of the theoretical distribution model and the interpretation of the results of inference. However, these results suggest that for many students these ideas were not understood.

The theoretical analysis of the content domain carried out prior to this study suggested that formation of conceptual links between the empirical sampling distribution and the determination of P-values and confidence intervals was necessary to facilitate conceptual understanding in statistical inference. The analysis of the Map 6 in Table 1 shows that only 30% of the students explicitly linked the sampling distribution to the determination of the P-value, whilst a very small 13% of the students explicitly linked the sampling distribution to the determination of the confidence interval. This lack of connections suggests due to inadequate or incomplete schemas, many students may later exhibit a lack of conceptual understanding in these areas.

CONCLUSION

From this study it seems possible that the instructional treatment, including the extensive use of the sampling software, was helpful in elucidating some important concepts of sampling distribution in each of the specific contexts in which it was applied for some students. However, the nature of these software packages is such that they are always distribution specific, and as such they have no specific role in illustrating the concepts and links, which together form a schema for the generalised sampling distribution. That is, the recognition of each of the contexts as a particular example of a general concept is an act of integrative reconciliation cannot be assumed to follow from experience with computer based technology.

Is it possible to design computer software which, when appropriately integrated in a teaching/learning strategy, could encourage the formation of the generalised sampling distribution concept as well as facilitate the formation of links to hypothesis testing and estimation? This would seem to be the next challenge for the educational software developer. What appears to be desirable is computer software which supports the dynamic visual linking of the empirical and theoretical representations of the sampling distribution, and enables students to take both the empirical and theoretical paths to investigating inference. By explicitly recognising, investigating and comparing the alternative representations in a variety of contexts, students may develop not only and understanding of the steps involved in carrying out statistical inference, but also of the limitations of this process in practice.

REFERENCES

Cox, C., & Mouw, J.T. (1992). Disruption of the representativeness heuristic: Can we be perturbed into using correct probabilistic reasoning. *Educational Studies in Mathematics*, *23*, 163-178.

Johnson, R., & Bhattacharyya, G. (1992). *Statistics Principles and Methods* (2nd edn). New York: John Wiley and Sons.

Jonassen, D.H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for Representing, Conveying and Acquiring Structural Knowledge*. Hillsdale, NJ: Lawrence Erlbaum.

Lipson, K.L. (2000). The role of the sampling distribution in developing understanding of statistical inference. Ph.D. Dissertation, Swinburne University of Technology, Melbourne.

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the 3rd International Conference on Teaching Statistics* (pp. 314-319). Voorburg, The Netherlands: International Statistics Institute.

Rubin, A. (1990). Sampling Laboratory, computer program. Unpublished.

Shaughnessy, J.M. (1992). Research in probability and statistics: Reflections and directions. In D.A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp.465-494). New York: MacMillan.

Stirling, D. (1991). Models'n'Data (Version 1), computer program. Santa Barbara: Intellimation.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105-110.

Vinner, S. (1983). Concept definition, concept image and the notion of function. *International Journal of Mathematics Education in Science and Technology*, *14*(3), 293-305.