STUDENTS' UNDERSTANDING OF CONFIDENCE INTERVALS

Robyn Reaburn
University of Tasmania, Australia
Robyn.Reaburn@utas.edu.au

The aim of this study was to gain knowledge of students' beliefs and difficulties in understanding confidence intervals and to use this knowledge to develop improved teaching programs. This study took place over four consecutive teaching semesters of a one-semester tertiary statistics unit. The study was cyclical, in that the results of each semester were used to inform the instructional design for the following semester. Over the semesters the following instructional techniques were introduced: simulation with and without a computer, encouraging students to write about their work, and the use of alternative representations. As the interventions progressed, a higher proportion of students successfully defined and used confidence intervals to estimate the value of the population mean. This study also identified sources of confusion for students that can be a basis for further research.

This paper describes a study that examined students' problems in understanding confidence intervals for the mean in a first-year tertiary statistics unit and the results of an intervention that aimed to improve this understanding. Confidence intervals are used to estimate the values of population parameters. They give a range of plausible populations from which a random sample might produce the observed sample statistic.. One of the problems with statistical inference is that not only is each sample not exactly representative of the population, but no two samples are exactly alike either. Therefore, any estimate made from a sample is made with uncertainty. We do know, however, the relationship between a sample and the parent population. In particular, the Central Limit Theorem (CLT) states that if the sample size is large enough the mean of this sample belongs to a normal distribution that is centred on the mean of the initial population. In addition, this distribution has a standard deviation (known as the *standard error* of the mean) that is equal to the standard deviation of the initial population divided by the square root of the sample size. This knowledge allows us to make estimates of the population mean with pre-determined levels of uncertainty.

To understand the process of confidence intervals, students need to realise that the distribution of the sample means has the characteristics of any other normal distribution, including that approximately 95% of the possible sample means will lie within two standard errors of the population. Therefore it is likely that the mean of a random sample will be within two standard errors of the population mean. If the sample mean is within two standard errors of the population mean, adding and subtracting two standard errors to and from this sample mean will produce a range of values that will include the value of the population mean.

There is a body of literature that suggests that not only should confidence intervals be used to estimate the value of population parameters but that confidence intervals should replace the use of *P*-values in the Null Hypothesis Test (NHT). *P*-values can vary widely from sample to sample (Cumming, 2010) and can vary even with the same data, depending on the method of analysis chosen by the researcher (Hubbard & Lindsay, 2008). In addition, *P*-values do not reflect the effect size (Hubbard & Lindsay, 2008; Wagenmakers, 2007). Another concern is about the validity of the way *P*-values are calculated. Assuming the null hypothesis is true, *P*-values are the probability of the observed data and the probability of more extreme data, yet these more extreme data are not actually observed (Hubbard & Lindsay, 2008). If confidence intervals should replace NHT in the future, then it becomes even more important that students should understand how they are derived.

Confidence intervals are very easy to calculate, as they just require the addition and subtraction of a number (determined by the level of *confidence* required) to the sample mean multiplied by the standard error. However, understanding what the results of these calculations represent has been shown to be problematical. A 95% confidence interval for the mean, for example, tells the educated reader that the process used to give this interval will include the value of the population mean 95% of the time it is used. Therefore it is not absolutely certain that the value of the population mean is within this interval. In a study of undergraduate students, delMas,

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Garfield, Ooms, and Chance (2007) found that about one third of students believed that a confidence interval indicated the percentage of population values within the range of the interval, and the majority indicated that the level of confidence denoted the percentage of all sample means that lie between the confidence limits. In a study of researchers, Cumming (2006) found that there was a common misconception that the level of confidence denoted the percentage of sample means that would fall within the original confidence interval if replicate samples were taken.

METHODOLOGY

Participants

This study was part of a doctoral project (Reaburn, 2011) and was carried out over four consecutive teaching semesters of a one-semester introductory statistics unit at a tertiary institution. The subjects of the study for each of the four semesters were volunteers from this unit. Because the students were volunteers, there was considerable variation in the proportions of students who volunteered data over the period of the study. In the first semester there were 12 volunteers out of a possible 20, in the second semester there were 23 out of a possible 26, for the third semester there were 6 out of a possible 27, and for the fourth semester there 12 were out of a possible 26.

Methodology

The study was in the form of action research (Mills, 2007) where the researcher was the lecturer of the unit. The study was cyclical, in that the results of each semester were used to inform the instructional design for the following semester. The first semester of the study, the pre-intervention semester, was used to gain knowledge of students' understanding of confidence intervals before the teaching program was altered.

Over the four semesters of the study the following strategies were added sequentially. In the first cycle of the intervention (the second semester of the study) simulation was used to demonstrate the CLT. This was designed to encourage students to think that the distribution of sample means would have the same distribution as the original population, and then to confront them with conflicting evidence. In the second cycle of the intervention, students were given data for a population of 100 people and then took samples of size 10, calculated the sample means, and placed their answers on a number line. This was to give students the idea that not all sample means would have the same value, that these sample means would probably be 'close' to the value of the population mean, and that they could be used to give an estimate of the value of the population mean. This was followed by the simulation to demonstrate the CLT. In each case, the CLT was not introduced in a formal lecture until after the simulation. In the third cycle of the intervention, in addition to the simulations described, the students were asked to draw diagrams to illustrate how confidence intervals were derived. In addition, they were asked to write down the principles of the derivation of confidence intervals and also what they thought they were for. It was intended that by discussing their answers with each other and with the lecturer, students would become aware of the gaps in their knowledge and try to fill in these gaps (Morgan, 2001; Pugalee, 2001).

Assessing Students' Understanding

To assess students' understanding of confidence intervals the students were required to answer two questions in a test held in the final week of each semester, with the same wording used each time. The questions were:

> The 95% confidence interval for the expected number of visits by Tasmanians to a doctor during 1998 is 7 to 11.
>
> a. In completely non-technical words, explain what this statement means.
> b. What does the 95% refer to?

Their answers were coded according to the level of understanding shown in their answers. This coding is shown in Table 1.

Table 1. Codes given for the answers to the confidence interval questions.

	Answer	Code
Part a	Mean number/expected number of visits was between seven and eleven	2
	Mean number/expected number of visits was between seven and eleven 95% of the time/On average, Tasmanians visited a doctor between 7 and 11 times	1
	Seven to eleven Tasmanians visited a doctor/no answer	0
Part b	The process used would include the value of the population mean 95% of the time it was used	2
	95% of population is within two standard deviations of the mean/95% of sample means will be within two standard errors of the population mean – no further explanation	1
	95% of the sample means will be in the stated range/95% of the population means are in stated range/95% of population visited a doctor between 7 and 11 times/95% of population is within two standard deviations of the mean/no answer	0

RESULTS

Figure 1 shows the distribution of the codes given to the student answers over the four semesters of the study. For Part (a) the biggest difference between the third cycle of the intervention and the previous semesters was that no student received a score of "0" in the third cycle. For part (b) a higher proportion of students in the third cycle of the intervention received a score of "2". The differences among the scores of the semesters for part (a) were not significant (Kruskal-Wallis test $P = .397$). For Part (b) the differences among the scores of the semester were significant, with the third cycle of the intervention having the highest mean rank (Kruskal-Wallis test ($P = .009$).

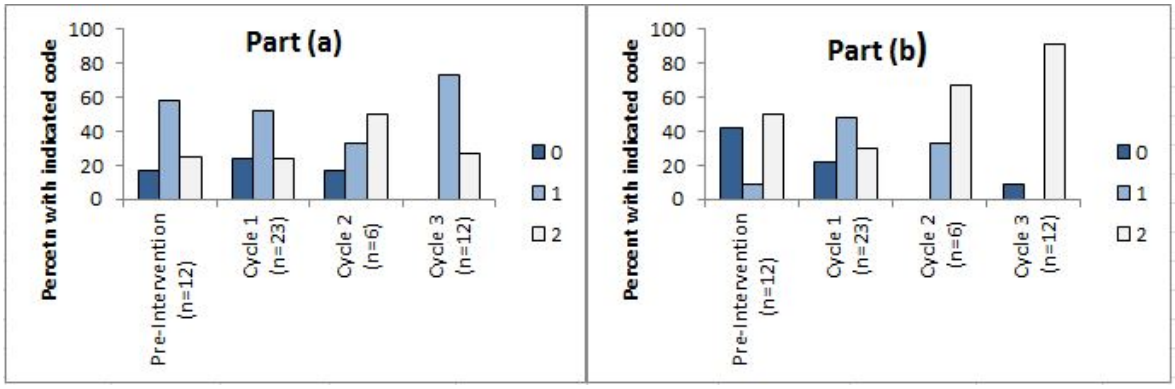

Figure 1. Percentage of each code given to the students' answers to parts (a) and (b) of the confidence interval questions in the test.

DISCUSSION

Over the period of the study there was a general improvement in the understanding of confidence intervals, but it was evident that problems in understanding still remained. One of the most common of the incorrect statements was that 95% of the population visited a doctor between seven and eleven times. Other common misconceptions, whilst in themselves were correct statements, did not show a full understanding of what confidence intervals are. Such statements included that 95% of the population is within two standard deviations of the population mean and that 95% of the sample means are within two standard errors of the population mean. It was also apparent that some students believed that 95% of the sample means are included in the confidence interval or that 95% of the population will be in the stated interval (also found by delMas, Garfield, Ooms & Chance, 2007). Other students indicated that the mean is within the confidence interval 95% of the time. The most disturbing interpretation was that between seven and eleven people

visited the doctor during the year. Several of the answers also showed that there was confusion between the terms *standard deviation* and *standard error*.

Overall, the proportion of correct answers to these questions rose from one quarter of the students in the pre-intervention semester to three quarters at the end of the third intervention. It is the belief of this researcher that no one strategy was of significant benefit in helping students understand the principles behind confidence intervals. Instead, it was a combination of strategies that led to an improvement in students' understanding. The simulations, both by computer and by hand, demonstrated that no two sample means will be identical, and the principle of the Central Limit Theorem. The addition of the diagrams was intended to give a visual connection to the theory, and the writing about their understanding was to help students realise where they had gaps in their knowledge and motivated them to search for the understanding they needed.

Concurrent with this study was an investigation into students' understanding of *P*-values (Reaburn, in press). Overall the understanding of *P*-values showed greater improvement than for confidence intervals. The same techniques of computer simulation, alternative representations and encouraging students to write about their work were used but with one important addition. This was that the students were given an example that was easily understood to use as a basis for their reasoning. This example asked the students to consider how likely it would be to see people dressed in winter clothes if the day was actually hot (Shaughnessy & Chance, 2005). The students could readily see that this observation would be unlikely on hot day and used this reasoning in their later work on *P*-values. Finding such an example to serve as a template for further reasoning for confidence intervals may be fruitful.

This study has identified sources of confusion and misunderstanding of confidence intervals that may form the basis of further work in this area.

REFERENCES

Cumming, G. (2006). Understanding replication: Confidence intervals, *p*-values, and what's likely to happen next time. In B. Phillips (Ed.), *Developing a statistically literate society* (Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

Cumming, G. (2010). Understanding, teaching and using *P* values. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society.* (Proceedings of the 8th International Conference on Teaching Statistics, Ljubljana, Slovenia). Voorburg, The Netherlands: International Statistics Institute.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 25-58.

Hubbard, R., & Lindsay, R. (2008). Why *P* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*(1), 69-88.

Mills, G. (2007). *Action research: A guide for the teacher researcher*. Upper Saddle River NJ: Merrill Prentice Hall.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Morgan, C. (2001). The place of pupil writing in learning, teaching and assessing mathematics. In P. Gates (Ed.), *Issues in Mathematics Teaching* (pp. 232-244). New York: Routledge Falmer.

Pugalee, D. (2001). Writing, mathematics, and metacognition: Looking for connections through students' work in mathematical problem solving. *School Science and Mathematics, 101*(5), 236-245.

Reaburn, R. (2011). *Students' understanding of statistical inference: Implications for teaching* (Doctoral thesis). Tasmania: University of Tasmania.

Reaburn, R. (in press). Students underestanding of *P*-values. *Statistics Education Research Journal*.

Shaughnessy, J., & Chance, B. (2005). *Statistical questions from the classroom.* Reston, VA: NCTM.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of *P* values. *Psychonomic Bulletin and Review, 14*(5), 779-804.