IDENTIFYING MISCONCEPTIONS ABOUT CONFIDENCE INTERVALS

Pawel Kalinowski
School of Psychological Science, La Trobe University, Australia.
p.kalinowski@latrobe.edu.au

Although confidence intervals (CIs) have many benefits over null hypothesis significance testing (NHST) they can still be misinterpreted. Identifying CI misconceptions is a first step in designing teaching tools that can be used to prevent or reduce them. I surveyed graduate level students and found they hold several misconceptions about CIs. Many believe there is a uniform likelihood distribution across a CI, with a high proportion of these showing a cliff effect (a sudden major drop in likelihood at each limit of a CI). Many students also misunderstand the relationship between the width of a CI and the confidence level. In this paper I present a taxonomy of CI misconceptions identified by empirical studies, and explore faulty conceptual models that may be the source of the misconceptions. I also propose an educational tool that could be used to confront CI misconceptions, particularly misconceptions about CI distributions.

BACKGROUND

Many proponents of statistical reform in social and behavioural sciences recommend CIs as an alternative to reporting *p* values (Harlow, 1997). CIs have also been recommended by The American Psychological Association (APA) *Publication Manual*, and by various APA and other peer reviewed journals. A reason to recommend CIs over other statistical techniques is that CIs offer far more information than *p* values, while remaining in a familiar Frequentist framework. Despite the advantages of CIs, statistical educators should stay vigilant to the possibility of CI misconceptions. By identifying student misconceptions I can gain some insight into the best ways to teach CIs to prevent or address them.

Fidler (2005), categorized CI misconception using the terms “definitional” or “relational” (p. 210). Definitional misconceptions refer to misconceptions of what a CI measures, its inferential nature, or what it estimates. These definitional misconceptions can be stated directly by the participant or inferred through experimental results such as surveys or judgment tasks. Relational misconceptions refer to expectations of relationships between confidence level, width and sample size. Fidler surveyed 180 undergraduate students and found that students held several definitional and relational misconceptions (Table 1).

Table.1 Percentages of students who held definitional (D) and relational (R) misconceptions of CIs (Fidler, 2005)

Type	*%*	*Misconception*
D	38	CIs are a range of plausible values for the sample mean.
D	8	CIs are range of individual scores.
D	11	CIs are a range of individual scores within one standard deviation.
R	20	CI width increases as N increases.
R	29	Change in N has little effect on CI width.
R	73	90% CI is wider than a 95% CI for the same data.

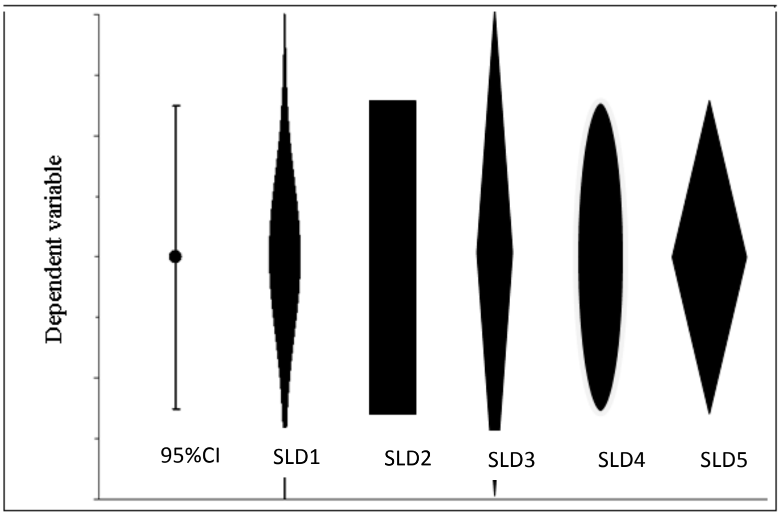
N = 180

This current study investigates these relational misconceptions further. I do this by exploring relative likelihood distributions. In the long run a 95% CI has a 95% chance of capturing the population mean μ, but the relative likelihood of each point across the interval falling on μ is not equal (Cumming, 2007). Instead, the relative likelihood of points across the CI (given sigma known) is distributed normally, with points closer to the sample mean (centre of the CI) being relatively more likely to fall on the population parameter than those further away (Figure 1a.). If sigma is not known, the relative likelihood follows a *t* distribution. I refer to these distributions as relative likelihood distributions.

I assume that students have at least an implicit belief about the relative likelihood distribution of a CI; some may be able to express such beliefs, others may not. I refer to student

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

beliefs about the relative likelihood distribution of a CI as their subjective likelihood distribution (SLD). Figure 1 provides examples of some of the possible SLDs that could be held by students.



SLD1 is the normative standard indicating the calculated relative likelihood that each point of the CI will fall on μ

Figure 1. A 95% CI and five representations of possible SLDs for a 95% CI

By identifying student CI SLDs I hope to find some basis for the types of misconceptions identified by Fidler (2005). For example if a student held SLD2, they would treat a CI as a dichotomous decision making tool, may not understand accurately the relationship between sample size and precision, and would believe that a 50%CI would be roughly half the width of a 95%CI. I developed several procedures aimed at identifying SLDs.

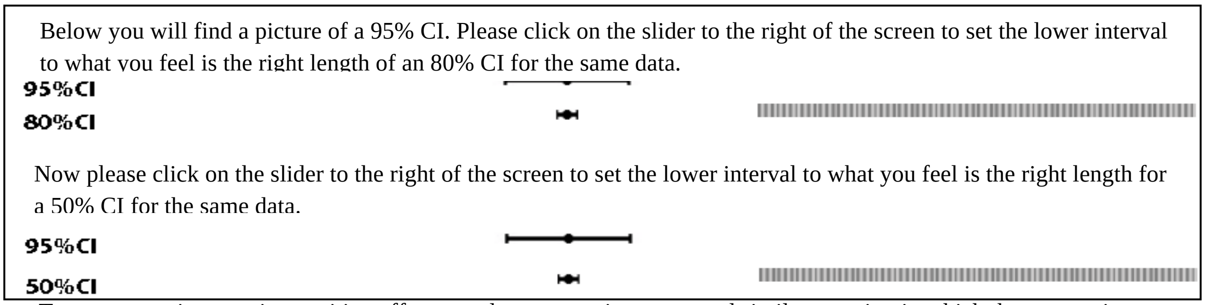
METHOD

Ninety-four honours (fourth year undergraduate) and postgraduate students from Psychology, Ecology, Medicine and other science disciplines replied to a web survey distributed by email and social network web pages The survey consisted of four separate tasks aimed at tapping students' CI SLDs. In the current paper, I report results of two of these tasks.

Survey Tasks

A *width choice task* (Figure 2) presented students with 95%CI and asked them to adjust accompanying lines to create a 80%CI and an 50%CI for the same data. To balance starting position effects students were then given a 50%CI and asked to adjust accompanying lines to create 80% and 95%CIs.

The *SLD forced choice task* (Figure 3) presented students with 6 potential likelihood distributions for a 95%CI. Students selected from these options the distribution which most closely matched their own SLD.



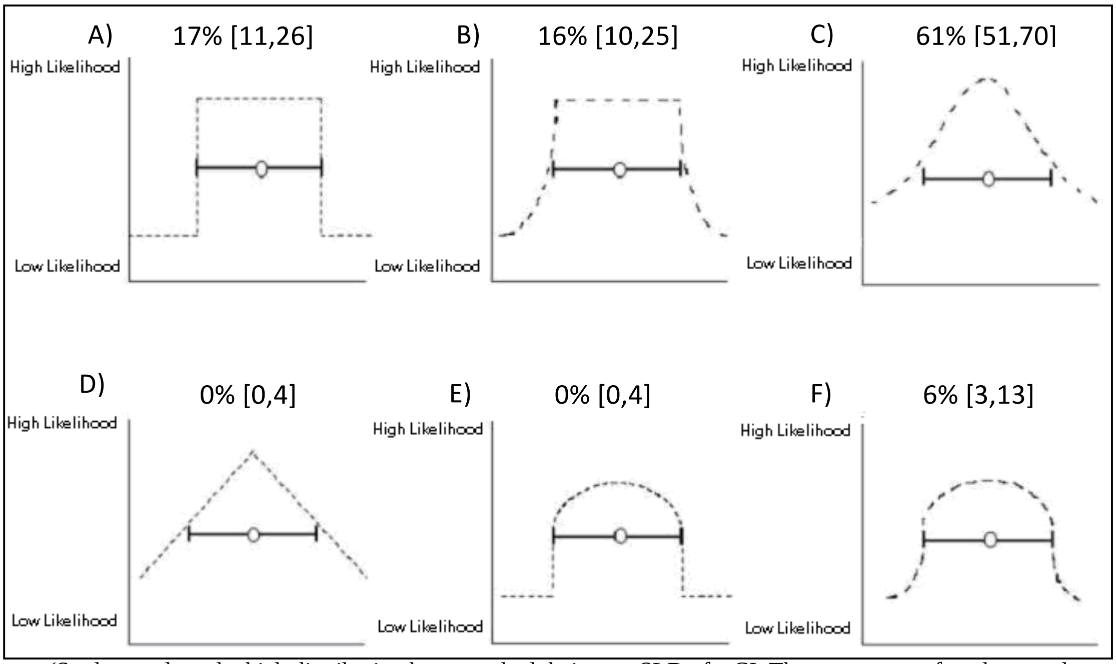
To protect against starting position effects, students were given a second similar question in which they were given a 50%CI and asked to adjust the width of 80% and 95%CIs.

Figure 2. The *width choice* task: Students were given a 95% and used the grey slider to adjust the accompanying CI to create to create 80% and 50% intervals for the same data

RESULTS

Width choice task: Almost one quarter of students (23%, 22 of 94, 95%CI [16, 33]), incorrectly believed that CI width would *decrease* as the confidence level *increased*, that is, they created 95%CIs that were shorter than 80%CIs, and 80%CIs that were shorter than 50%. The remaining (77%, 72 of 94, [67, 84]) of students adjusted accompanying intervals in the correct direction. On average their estimated width of a 50%CI, 80%CI and 95%CI corresponded to 60%CI, 82%CI, and 90%CI respectively. In summary, they overestimated the width of the 50% (and to a lesser extent the 80%CI), but underestimated the width of the 95%CI. This suggests that their understanding of the relational concept of CI width does not match that of a normal distribution. In 14% [8, 22] of students this was particularly pronounced: their responses indicated a belief that a 50% CI is approximately half the length of a 95%CI. In fact, a 50%CI is approximately one third of the width of a 95%CI, because the normal likelihood distribution flattens out towards the ends of a 95%CI.

SLD forced choice task: One third of students (33%, 31 of 94, [24, 43]) falsely believe that a 95%CI has an underlying uniform distribution. (Of course, it has an underlying normal distribution). The percentage of students that chose each of the likelihood distributions as the best representation for their SLD is shown in Figure 3. Selection of the shape A) and Shape B) distributions are considered evidence of a uniform SLD in this study.



(Students selected which distribution best matched their own SLD of a CI. The percentage of students and (the 95%CI for that percentage) selecting each option is shown directly above the distribution.)

Figure 3. Six likelihood distributions proposed to underlie a 95%CI

Students holding a uniform SLD may be more likely to interpret CIs in a dichotomous way, that is, to interpret the interval only in terms of whether it contains a certain parameter or not, thus ignoring important information about precision and other values of importance within the CI. These distributions also suggest a cliff effect—a sudden drop in confidence beyond the end of the interval. The bottom right distribution also suggests belief in a cliff effect, bringing the total of cliff effect responses to 39% (37 of 94, [30-49])

DISCUSSION

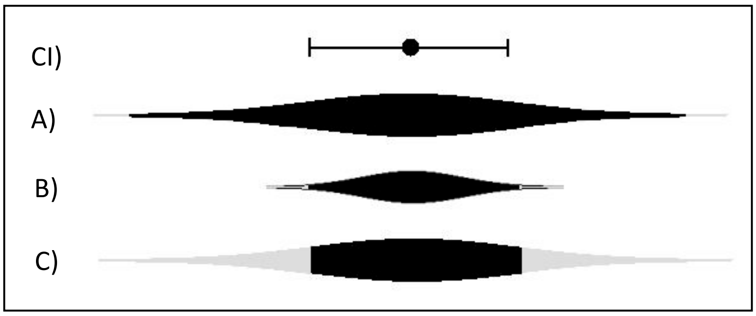
My results identify three misconceptions about CIs. Firstly, I found substantial evidence of a relational misconception identified by Fidler (2005), that is, the false belief that as the confidence level increases, the width of a CI decreases. Fidler found 75% of undergraduate (first year) students held this misconception. At postgraduate level this figure is still a worrying 23%.

Secondly, a small but still substantial number (14%) of students believe a 95%CI is approximately double the width of a 50%CI. This is disturbing given the education level of these students and the amount of statistics training they have received by this stage.

Finally, one third of students (33%) believe that the underlying distribution of CI is uniform. This suggests they will be more prone to interpreting intervals in a dichotomous fashion, and thus not using intervals to their best advantage. One of the main proposed advantages of CIs is that they help alleviate dichotomous thinking. The current results suggest that, for at least that proportion of students who hold uniform SLDs, this is unlikely. The misconceptions identified in this survey have important implications for teaching. By presenting students with visual cues relating to the likelihood distribution of a CI, perhaps CI misconceptions can be reduced or even eliminated.

TEACHING IMPLICATIONS

These results suggest that students would benefit from extra information about likelihood distributions when learning about CIs. Giving students visual cues such as the bulging CI (figure 1, SLD1) could reduce confusion about CIs. Further, an interactive bulging CI may help students understand the relationship between CI width, sample size and confidence level. For example, lowering the confidence level (say from 95% to 50%) decreases CI width. Increasing sample size (at a given confidence level) also reduces CI width. But the bulge produced in each case is different (figure 4). Visual cues could be particularly useful in demonstrating the relationship between confidence level, likelihood density and width. I suspect such pictures would reduce CI misconceptions such as 'a 95%CI is approximately double the width of a 50% CI'. The next step in my research program is to evaluate whether bulging CIs improve students' conceptual understanding of CIs and reduce relevant misconceptions.



Bulging CIs such as A, B and C could provide students with more information about the relationship between confidence level, sample size and width. For example, by adding more data to A the resulting CI would look like B. The interval is narrower but the shape is the same because the confidence level is constant. If sample size is held constant and the confidence level was reduced, A would look like C. The interval is narrower but the shape is different.

Figure 4. CIs are usually presented using the 'CI' figure

ACKNOWLEDGEMENTS

I thank Jerry Lai, Fiona Fidler and Geoff Cumming for valuable comments.

REFERENCES

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (in review). Confidence Intervals permit, but don't guarantee, better inference than statistical significance testing.

Cumming, G. (2007). Inference by eye: Pictures of Confidence Intervals and thinking about levels of confidence. *Teaching Statistics, 29*, 89-93.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* (Unpublished PhD thesis, University of Melbourne). Online: www.botany.unimelb.edu.au/envisci/docs/fidler/fidlerphd_aug06.pdf.

Harlow, L. L. (1997). Significance Testing in Introduction and Overview. In L. L. Harlow, S. A. Muliak & J. H. Steiger (Eds.), *What If There Were No Significance Tests?* (pp. 1-17). Mahwah, NJ, USA: Lawrence Erlbaum.