Confidence intervals: What matters most?

1. Understanding **why we need confidence intervals**
    - *(In sampling contexts) To correct for the likely extent of error in our estimates due to sampling variation.* ***"All estimates are wrong"*** *is my attention getter. It is a slight overstatement but even if we were lucky enough to get a correct estimate we'd never know it had happened.*
2. **Wanting to have them for every estimate** we get or are given
    - *(In sampling contexts) Estimates of any sort are subject to sampling errors which can be big especially from small samples. We need to know how big these errors could potentially be.*
3. Knowing **what they do <u>not</u> correct for**
    - *They do not correct for biases in the observational process (e.g. measurement biases and other nonsampling errors). These biases have to be minimised at the study design stage. "Garbage-in, garbage-out" is an unfortunate fact of life when analysing data.*
4. Knowing **what properties they have**
    - *Good coverage rates, averaged over the repeated taking of samples, of the corresponding true or population value.*
        - *Formally, a method for calculating a 95% CI should cover the true value for 95% of samples taken. If so it is a fairly safe bet that that if we follow this procedure we will capture the truth.*
5. Knowing that **the bigger the sample** taken **the narrower the confidence interval** obtained, (or the more *precise* is our estimation of the quantity of interest).
    - *A sample 4 times larger gives a CI approximately half as wide*
6. Being able to **interpret a** confidence interval **in the context** of a particular problem and **communicate this well** both orally and in writing

These absolutely fundamental understandings should be displayed in context in student work to meet AS 3.10 and are important also in AS 3.12. They are entirely generic. They remain the same whether an interval being used has been obtained from a formula or from the bootstrap and regardless of the quantity being estimated. It is particularly important to attend to the language being used in teaching and assessment for point 6 above.

Calculating Confidence Intervals

In essentially all real analyses of real data the actual confidence intervals are automatically produced by software, *diminishing calculation as a data analysis skill*. But understanding what it all means is crucial.

- Note that reports written by others often give summary statistics and being able to turn these into useful CIs is a valuable *statistical literacy skill (cf. AS 3.12)*.

Methods for obtaining an actual confidence interval come via two main routes:

- **mathematical theory** obtained under idealised assumptions (e.g. normal distributions)
- **computer intensive methods** of which the most generally useful is the bootstrap

The first is the historical route, necessitated by a lack of computational power. ***The two big pedagogical advantages of the bootstrap route are*****:**

(i) we can get there with many fewer and easier concepts giving a much more transparent relationship between the method of generating an interval and the problem we are trying to solve (errors due to sampling variation)

(ii) once this one idea has been established we can put useful confidence intervals around estimates of vastly more quantities; we are doing the same basic thing every time (see later).

Why have you seen CIs for means and proportions but never for measures of spread, or for medians, or for correlations?

It is not as though there is any lesser need for estimates of these other quantities to be accompanied by confidence intervals. Indeed they tend to be subject to even bigger sampling errors than means. But in the theory-based approaches we need a different recipe for each new quantity so it takes a long, long time to build up the superstructure and which makes it very difficult to see the wood for the trees. You can easily see, however, (e.g. using the iNZight**VIT** modules http://www.stat.auckland.ac.nz/~wild/VIT) how the patterns of variation generated by sampling from a population are approximated by the patterns of variation generated by bootstrap resampling for all these other quantities as well (provided the samples are not too small), and how for each quantity the generation of a bootstrap confidence interval is obtained in exactly the same way. This gives quick, understandable access to CIs for quantities including medians, proportions, quartiles, measures of spread such as interquartile ranges, differences in means, medians and proportions, ratios of spreads, regression slopes, correlations and many, many more besides.

The current situation and understanding "$\pm z\,\sigma/\sqrt{n}$"

In terms of what has been happening in schools, many students coming to university seem to have no understanding of where the "$\pm z\,\sigma/\sqrt{n}$" in the traditional confidence interval for the mean comes from and how it relates to the extent of sampling variation, or even really why we do this at all. Assessment of this area has been an external and simply a plug-numbers-into-the-formula exercise with the harder questions involving some algebraic rearrangement of the formula. Neither advances student's ability to think statistically about data at all. From 2013, AS3.10 is an internal, is investigation based and will require demonstration of statistical understanding. "Getting the numbers" will just be one small component.

BACKGROUND DETAILS

Methods obtained from mathematical theory

Theoretically derived mathematical methods are deduced from idealised assumptions that will never be satisfied exactly so deriving a method is just a first step. We then have to check the extent to which the method will work under departures from its assumptions. This is the idea of robustness and we investigate robustness using computer simulations. That is how we know that CIs for means from normal distribution theory are fairly robust. It is also how we know that normal theory CIs for standard deviations are so sensitive to even small departures from normality that they are not fit for practical use.

We can also relax the strict normality assumption and show theoretically that the traditional CIs for means work "asymptotically", as a consequence of the central limit theorem. "Asymptotically" means 'as the sample size tends to infinity'. (The n=30 rule of high school statistics has no theoretical basis.) The speed at which asymptotic behaviour starts to work depends on the parent distribution being sampled from. This behaviour has been investigated, once again, by using simulation. The results of these simulations are how we know that asymptotic behaviour for means works fairly fast for distributions not "too far" from normal. It takes large samples to get reliable coverage rates for a proportion – very large as p gets closer to 0 or 1.

Methods obtained from the Bootstrap

Although confidence intervals from the bootstrap can be generated by simulation they too are justified by very high-powered asymptotic theory (some of it involving advanced versions of the central limit theorem). For practical usefulness their behaviour has similarly been investigated using simulation. Interestingly, for means and proportions, the bootstrap generates the same large sample confidence intervals[1] as those derived from the central limit theorem because the standard deviation of the bootstrap distribution is the same as the theoretical standard error of the mean[2]. Thus these intervals experience the same sorts of small sample deficiencies as the normal-based methods we are accustomed to. To address the small sample problem, theoretical statisticians have come up with some very clever ways of improving the bootstrap's small sample behaviour. Statisticians are generally a pragmatic bunch. They will use anything that works regardless of its genesis. We often report traditionally obtained intervals in situations where these methods work because of their familiarity. When the theoretical development gets too hard (something which is happening more and more frequently with very complicated forms of analysis, or when methods from theory are too sensitive to departures from assumptions) one of the first tools we reach for in our own work is the bootstrap because it is so widely applicable.

Looking to the future

In the face of hard choices necessitated by very little teaching time, the new curriculum is setting us up for the future at some cost in terms of bridges to traditional approaches. We must remember that the last curriculum was sealed in at Year 13 level for almost 20 years. There have been huge advances in statistics over the last 20 years and the pace of change and expansion in the subject is accelerating. One of our needs going forward is to be able to open up more of the statistical world more quickly and development via the bootstrap facilitates this.

We finish by quoting George Cobb (2007), one of the USA's most respected thinkers in statistics education, "Much of what we currently teach to beginning students of statistics – a curriculum shaped by its once-necessary but now-anachronistic reliance on the normal as an approximate sampling distribution – is technically much more demanding, and substantively much more peripheral, than the simpler and more fundamental ideas that now, thanks to computers, we could and should be teaching. Before computers, there was no alternative. Now, there is no excuse."

[Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1* (1), 1-15. Online at: http://escholarship.org/uc/item/6hb3k0nz]

Google's Tim Hesterberg (2006) says that bootstrapping and randomisation "increasingly pervade statistical practice. They offer ease of use: the same basic procedures can be used in a wide variety of applications, without requiring difficult analytical derivations. This frees statisticians to use a wider range of methods, not just those for which easy formulas for confidence intervals or hypothesis tests are available."

[Hesterberg, T. (2006). Bootstrapping students' understanding of statistical concepts. In G. Burrill (Ed.), Thinking and Reasoning With Data and Chance. Sixty-eighth National Council of Teachers of Mathematics Yearbook (pp. 391–416). Reston, VA: NCTM.]

Chris Wild (U. Auckland), Jennifer Brown (U. Canterbury), Michelle Dalrymple (Cashmere HS), *8 Nov.2012*
http://www.stat.auckland.ac.nz/~wild, http://www.math.canterbury.ac.nz/~j.brown

[1] If you use a large number of bootstrap resamples and the ± 2 standard error version of the bootstrap CI. The version of the bootstrap being advocated for schools is the easiest to grasp – the so-called percentile bootstrap which also has the practical advantage of adapting automatically where there is skewness in the sampling distribution to produce sensibly asymmetric confidence intervals.

[2] to within an asymptotically negligible factor of $\sqrt{(n-1)/n}$.