STATISTICS & DATA SCIENCE EDUCATOR

The North Island has the bigger population, but does the South Island have the bigger motor vehicle engines? An introduction to bootstrapping

Mark Hooper
Otago Boys' High School, New Zealand
Email: mark.hooper@obhs.school.nz

Overview of lesson

The lessons introduce the concept of resampling through bootstrapping, leading to the formation of a confidence interval of the difference between two means or two medians. The students experience sampling variation and bootstrapping visually through the aid of technology and manually through a physical exercise involving data cards of a sample from a very large population. A confidence interval of the difference between two population parameters is formed and interpreted. Lastly the experience using technology is formalised.

Learning objectives

Within the PPDAC investigative cycle framework students should be able to:

- pose a comparative question
- recognise sampling variability
- form a confidence interval using the bootstrapping method
- make a sample-to-population inference
- understand the principles behind bootstrapping

Suggested age range

For Year 13 students (17 to 18 years old)

Time required

4 to 5 lessons (240 to 300 minutes)

Keywords

bootstrap, resampling, confidence intervals, sample, inference, population

Introduction

Increasingly in recent years there has been an encouragement to use technology to better demonstrate the ideas behind many statistical processes instead of being hidden under the mathematics. Two such ideas are sampling variation and resampling through bootstrapping. The starting point is for students to understand the nature of the variables to be studied and how to ask good comparative statistical investigative questions. Contact with local businesses in our city has thrown to the fore how many "real life" statisticians are using datasets that are millions of rows deep with dozens of measures as column headings.

Big Data for many of our students when they leave high school will be part of their next learning journey. My guiding principles are to educate my students to ask meaningful questions and to guide their learning with a data set that is meaningful and of relevance to them within the scope of the available tools at high school and the curriculum. Of course the choice of dataset, level of teacher guidance and student led engagement have to be balanced with the needs of the curriculum. The hope is that these lessons can provide a framework for teaching and learning some key concepts for determining estimates and confidence intervals using methods such as resampling, in a way that combines a good variety of modern technology and pedagogy.

Lesson outline

1. Asking good comparative statistical questions

Guided by the PPDAC cycle for a statistical investigation, the first stage was to build the confidence and skill level of students to ask good questions. The following Figure 1 was shown to the class. The data was compiled from the UK and the New Zealand Census at School websites a few years ago. Unfortunately the UK version of this excellent resource is now defunct, but the New Zealand version is an excellent repository for statistics teachers of students of all ages (see new.censusatschool.org.nz).

Look at the following start of a dataset, collected from a couple of sources based in the UK and New Zealand
- copy down the column headings

gender	age	waketime	Base from 7am	Nationality
boy	16	7:15	15	UK
boy	15	6:45	-15	UK
boy	15	5:30	-90	UK
boy	15	6:45:00	-15	NZ
boy	15	6:00:00	-60	NZ
boy	15	7:15	15	UK
boy	16	7:45	45	UK
boy	15	6:45:00	-15	NZ
boy	15	7:40	40	UK
boy	17	7:00	0	UK
boy	16	7:05	5	UK

What could be the point of analysing such a data set?
What could be a good question?

Figure 1: Data set used to stimulate students to formulate questions.

Before we started to formulate questions, an emphasis was placed on why such variables were measured in the first place. What is the point or purpose of such an investigation that comes from a question being asked?

> The students gave good, imaginative and somewhat humorous answers. An idea of which country was laziest seemed to be a common theme! Already in the lessons the depth of thinking was impressive. Students asked questions that revealed they were thinking beyond what was just presented to them. *How was the data collected? Are all the students at school? Are girls included in the sample?*

Eventually for the teacher guided part of the lesson, the following Figure 2 was shown and some time spent on discussing the preciseness required over structuring a good statistical investigative question at this level.

Investigative question – the forensic structure of a good question

What is the difference between the median of wake up times (in minutes from 7 a.m.) of New Zealand Year 11s and the median wake up times of UK Year 11s?

Difference direction of comparison (leave open ended)
Median the population parameter (key feature) being estimated
wake up times the variable being measured (with units too) (numerical)
New Zealand Year 11s & UK Year 11s categorical or grouping variable

Figure 2: Discussion exemplar for considering the elements of a good question.

The text in the bottom half of Figure 2 describing each highlighted part was initially blanked out. The students were required to offer suggestions. Key ideas such as numerical variables, categorical variables and sample to population inference were introduced and/or reinforced during this stage.

> The students conjectured that the question posed in Figure 2 could allow us to conclude which of the two countries had the "laziest" Year 11s. Some students, seemingly influenced by images from programmes such as Coronation Street, offered suggestions that the nearest secondary school due to greater population density would be closer than typically in New Zealand thus permitting UK Year 11 students to get up later each school day.

Once students had grasped how to ask good questions, they practised writing their own with the activity shown in Figure 3.

Below shows the variables collected across three different data sets. For each data set, write a well-structured question in the format we have just learnt. Justify why your question is worthwhile for study and analysis. In other words, what's the POINT?

Accident data from a sample of reported motor crashes in New Zealand in 2017 involving drivers that had some alcohol in their system											
Gender	Age	Risk group (High or Low)	licence type (L, R or F)	Crash severity (serious or moderate)	blood alcohol level (mg)	Distance driven (until crashing) (km)	Distance from home (km)	Vehicle age (years)	Insurance payout ($NZ)	medical expenses ($NZ)	Time off work (days)

A selection of the variables as part of the 2018 senior academic data																		
Course effort - Average - Year	Course progress - Average - Year	Periods - Absent (Year)	House	Year Level	Type – domestic or international	Boarder Status	Ethnicity - Level 1 Group	Credits Earned Percentage - Previous Year	NZQA - Credits Last Year - Attempted	NZQA - Credits Last Year - Total	NZQA - Credits Last Year - Not Achieved	NZQA - Credits Last Year - Achieved	NZQA - Credits Last Year - Merit	NZQA - Credits Last Year - Excellence	NZQA Results - L1 NCEA Endorsement	NZQA Results - L2 NCEA Endorsement	NZQA Results - L3 NCEA Endorsement	Total naughty entries on KAMAR

Variables collected as part of the New Zealand Household Savings Survey										
Gender	Employment	Qualification	Ethnicity	Partnered	Age	AgeP	Income from all sources	WagesSalary (from main job)	Debt	Networth

Figure 3: Data set variables given to students to stimulate posing.

> Whilst this seemed like a natural next step after the UK and NZ wake-up time activity, more value probably would have been added by allowing more time for the students to gauge their understanding of what the data sets were all about. Despite my assumption that the contextual understanding would magically appear just through a brief chat and seeing the column headings, a deeper discussion with more time and encouraging some research links would have resulted in better questions. Whilst students demonstrated a structural understanding of how to ask a question, which was the objective of this part of the lesson, it seemed the heart and soul had been shoved to one side. There was an abstractness about this activity following a very concrete and meaningful episode with the wake-up time data.

2. Introduction to resampling and bootstrapping interactive visualisations

The start of the next lesson began by getting the students to recall the constituent parts of what made a good question. The core nature of why we needed to use a resampling process for what we wanted to achieve was demonstrated through the website stat.auckland.ac.nz/~wild/VITonline and selecting the sampling variation link. The wake-up time data set, followed by the data sets that were used to get the students to ask their own questions were then used to demonstrate sampling variation. Sampling variation was emphasised because we were dealing with a sample and we needed to have an awareness of what could happen if we took another sample, and another, and another and so on. VITonline is an excellent tool for demonstrating visually some of the key statistical processes and concepts we cover in high school statistics learning in New Zealand.

The screenshot in Figure 4 demonstrates the sampling distribution for the mean of the blood alcohol level from the accident data, assuming the data we have is the data for the whole "population." We sampled 1000 lots of samples of size 20. The screenshot in Figure 5 shows the impact of increasing the sample size to 80. By the time New Zealand students have reached Year 13, they have been exposed to sampling variation and in various stages discussed and utilised it in order to be ready at Year 13 to estimate a population parameter from a sample (see, for example, Arnold, Pfannkuch, Wild, Regan & Budgett, 2011; Pfannkuch, Regan, Wild & Horton, 2010; Pfannkuch, Wild & Parsonage, 2012; Wild, Pfannkuch, Regan & Horton, 2011). At our school technology and the wrapping of a formal structure, which is possible using online tools such as confidence interval coverage around the interval construction are not introduced until Year 13.

It was fascinating to hear a trend of the word "accurate" being mentioned by the students. The sample of 80 is more accurate. This seems to be a bit of a battle in recent years where the word accurate is being used instead of "more precise".

At this stage the students have an awareness of what happens when we resample and also why collecting a larger sample of data may be useful. Very informally the students were asked what values they could say the "true" mean blood alcohol level might be using the idea of sampling variation and the output shown in Figures 4 and 5.

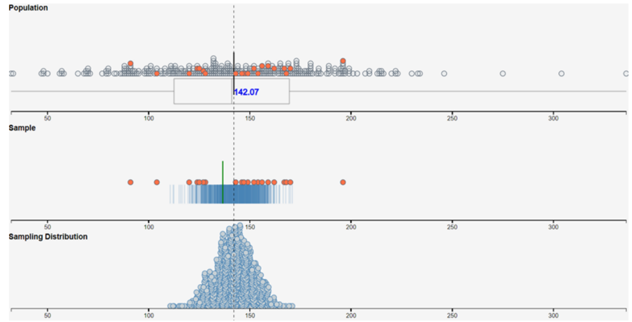


Figure 4: Screenshot of sampling distribution generated with VITonline tool of mean blood alcohol level from motor crash accident data (n=20).

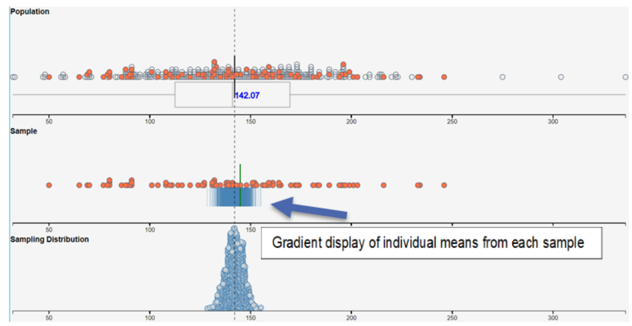


Figure 5: Screenshot of sampling distribution generated with VITonline tool of mean blood alcohol level from motor crash accident data (n=80).

Tentatively and reasonably quickly a student offered the range of values using the base of the sampling distribution. I missed an opportunity at this stage once the students used the base of the sampling distribution to link to what was happening in the gradient display (see Figure 5). This may have allowed some deeper thinking and sharing of ideas to be elicited to do with the very extreme sampled means.

3. Understanding the data set and posing an investigative comparative question

The data set planned for use for the overall objective (a bootstrap confidence interval to estimate the difference in means of a numerical variable split by a binary categorical variable) was introduced. Through some rather random searching for an interesting data set that would likely engage my students, I came across the New Zealand Transport Agency (NZTA) open data portal. One data set was the New Zealand motor vehicle register (opendata-nzta.opendata.arcgis.com/datasets/motor-vehicle-register-1).

A global discussion with the students was held about the nature of the variables (or attributes as the NZTA call them) measured in this data set (Figure 6).

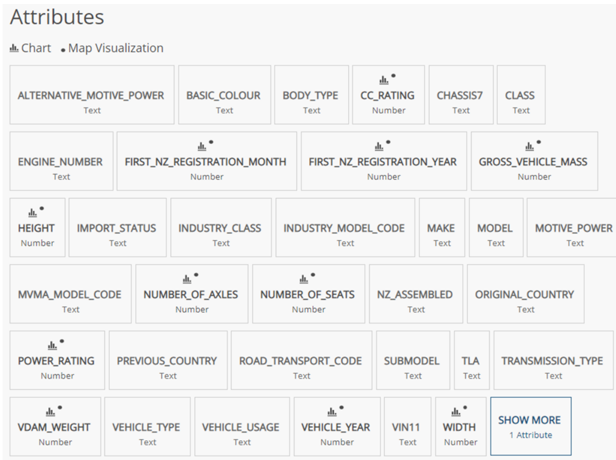


Figure 6: Variables associated with the NZ motor vehicle registrations data set.

A good 15-minute discussion, largely student led was then held. The students were fascinated how such a data set even existed (as was I!). Part of the discussion revolved around whether the data held allowed an individual's privately-owned vehicle to be identified. The consensus on this was that it was but that it would be difficult.

On reflection I missed a key point here about ethics and legal structures with government agencies and companies holding data. Whilst I hold very little knowledge about this myself it would have been a very good discussion. Additionally a discussion about metadata could have been held and its importance (via this link opendata-nzta.opendata.arcgis.com/pages/mvr-data-field-descriptions)

I knew the question we wanted to investigate and subtly led the students towards this. During the whole session, the fact we had a population of 5.2 million vehicles in our hands and analysis was impossible due to its sheer size was emphasised to the students, which was partly true due to my own lack of coding skills at the time! It was also discussed, through some research I conducted, that New Zealand is unique amongst all countries in the world in having such a high rate of vehicles per capita. Students did seem genuinely interested that we were dealing with a data set that was all New Zealand based and so large.

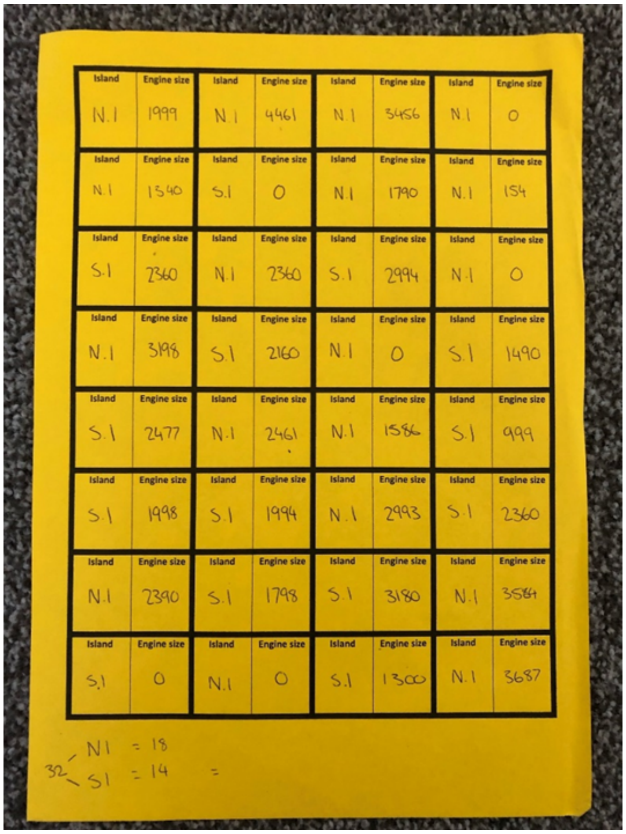
Thus the reason for needing a sample from the 5.2 million was established. It was mentioned at this point that there existed a statistical technique that we could use that enabled us to use a sample of just 30 vehicles and hence I asked each student to select a vehicle at random from the NZTA database. This would allow us to answer our question. The question was a comparison of engine size (CC_RATING) between the North Island and South Island (TLA). That is, *What is the difference between the mean engine size of North Island motor vehicles and the mean engine size of South Island motor vehicles?* This was considered a worthwhile question as the hillier and rougher roads of the South Island could mean vehicle engine sizes needed to be bigger.

Students selected one vehicle at random using the random number generator on their calculators, the query going through RStudio where the entire database was imported. This was a good moment to show the students how Excel could not cope with such a large dataset as the number of rows it holds by default is just over a million. Note if you do not use RStudio then I would suggest using the first one million rows of Excel if a dataset is bigger than this number. A random number generator could be used to pick the random sample, for example, if the random number 234,564 was displayed, the data value on this row number would be selected as part of the sample. For the purposes of demonstrating sampling variability and confidence interval construction, this number of rows is more than adequate. Initially in the planning phase of the lesson this was my intention. Had this turned out to be the case I was more than confident that the students would accept that a million rows of data could be considered big enough to be a population.

Each student filled in a pre-made data card set as shown in Figure 7. This was the end of this lesson, which was a good point. We had asked a meaningful question, understood how and why we had to take a sample and had the sample ready to process in the next lesson.

4. Learning the resampling procedure manually

Copies of the data card set in Figure 7 were made and each data card set and a pair of scissors was given to a group of students. Over the course of the next two lessons the students worked in groups to manually resample from the 32 data cards, following the instructions shown in Figure 8. The name Reuben is mentioned as he was the first student to answer what was required to be able to answer our question – a process to establish, with greater confidence, about what was happening in the whole population in relation to the estimate of our parameter. I explained the process of resampling with replacement



Island	Engine size	Island	Engine size	Island	Engine size	Island	Engine size
N.I	1999	N.I	4461	N.I	3456	N.I	0
N.I	1540	S.I	0	N.I	1790	N.I	154
S.I	2360	N.I	2360	S.I	2994	N.I	0
N.I	3198	S.I	2160	N.I	0	S.I	1490
S.I	2477	N.I	2461	N.I	1586	S.I	999
S.I	1998	S.I	1994	N.I	2993	S.I	2360
N.I	2390	S.I	1798	S.I	3180	N.I	3584
S.I	0	N.I	0	S.I	1300	N.I	3687

Figure 7: Data card set created by each student randomly selecting one vehicle (32 students in class).

Tasks to do Monday's, Tuesday's and Wednesday's learning:

- Working in groups of 3, 4 or 5 (so no pairs or sixes please).
- Before anything, share and practice describing a sample of data from the NZTA database, this time looking at the weight of vehicles and fuel type. As you're working in groups, let's go for 9 sentences.
- Once you've accomplished that ... now onto dealing with our wee sample of 32 engine sizes from 5,216,303 NZ motor vehicles.
- Agree on the wording of the question, a hypothesis and why it's worth studying for the sample of data collected on the yellow A4 sheet. Find one piece of research, quote at least one line (maybe from the abstract) that backs up the point of doing this study.
- Follow Reuben's advice and let's get some more data by doing the following – this is the guts of the concept you are learning (IMPORTANT STUFF)
  - Cut out the 32 'domino' cards
  - Separate the NI and SI into different piles
  - "Randomly" resample from each NI and SI sample another 18 for the NI and another 14 for the SI – sample with replacement. This means once you've selected one engine size, put it back in the pile for the chance to be resampled again.
  - STOP once you've resampled 18 NI and 14 SI
  - Calculate the mean from the new 18 NI sample the mean for the new SI sample
  - Calculate the difference in means and record this value, being VERY careful you know whether the difference in means is larger for the NI or SI
  - Do this 4 times each in your group and on your posters plot the difference between the two means on a suitable number line.
  - What do you think you've plotted allows you to conclude?
  - What are the good and bad aspects about the way we followed Reuben's advice (the way we got more data)?

Figure 8: Instructions given to students on how to manually resample.

and covered, using student input, why resampling without replacement would be pointless. Before the students started this activity a relatively lengthy discussion took place about the goal of statisticians, to some extent, is to answer questions about the whole population as reliably as possible. In an ideal world the whole population would be available but time, money, and other constraints play their part in making this unfeasible. However, it was established, initially by Reuben, that if a larger sample could be obtained with relatively little time and expense, then this would be a good thing.

Each group produced their own poster with the plots of the differences in the two means on a suitable scale. The original intention was to collate these results into all the differences in means calculated. Logistics and the need to move the learning on prompted the decision not to do this. The groups were working close enough to each other to see the variation in the results. Concentrating on smaller groups allowed me to pinpoint questions that were more relevant to each group. For example, if a particularly large difference in the means was plotted, it meant further questioning of the group took place to ensure they understood the bootstrapping resampling process more acutely.

It is important to note that I had prepared a different comparative sample of data from the NZTA database that compared weights of vehicles by fuel type. This is mentioned towards the start of Figure 8 and was designed to allow students to explore sample data and write down relevant, succinct points on what it showed. On reflection, whilst a useful and necessary activity to allow full coverage of the PPDAC cycle, the slightly different context did confuse some students. *Are we doing engine size or weights of vehicles sir?* was a question asked more than once.

Once the students had experienced this process a few times as per the instructions, the class was brought back together and through the VITOnline package the dataset was bootstrapped 1000 times to produce the

Figure 9 output. Initially one bootstrap resample was demonstrated to show the students an electronic version of what they had done manually – the point being that although re-sampling with replacement was a good way to picture the population it needn't be done manually. Computers can do the process very quickly and repeatedly.

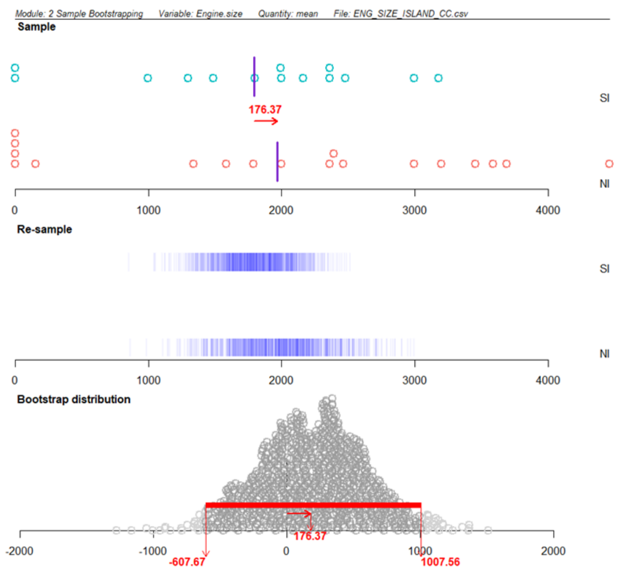


Figure 9: Screenshot of bootstrap confidence interval generated by VITonline tool for the difference between the means of engine size (cc) for North Island and South Island motor vehicles.

> One of the aims during the transition from the manual to the computer bootstrap resampling was to emphasise the amazing power of being able to simulate what was happening in the population with a resample of the original sample (but with replacement). I don't think in hindsight my ambition of wanting students to class this moment as one of their academic highlights at high school was reached!

5. Drawing a conclusion

During a whole class discussion, we built towards the conclusion of how the output in Figure 9 could be interpreted in relation to our question. The positive and negative values, including zero, in the confidence interval were pointed out and emphasis was placed on what they meant in terms of the context. The repeated emphasis of what each grey dot on the bootstrap distribution represented was needed to ensure that they did not think it was the mean of a sample but a value that represented the difference in mean engine sizes between a sample of 18 vehicles from the North Island and 14 from the South Island. To reinforce this idea the grey dot near the 2000 mark was discussed and how it had occurred. Most students understood it probably meant the largest and larger South Island values were resampled a few times in one resample of 14 and the lowest North Island engine sizes were repeatedly resampled in one resample of 18.

To complete the PPDAC cycle of this series of lessons, the students wrote an inference by interpreting the bootstrap confidence interval and drew a conclusion. For example:

"We can be fairly sure that the mean North Island engine size is somewhere between 1008cc higher and 608cc lower than the mean South Island engine size."

OR

"It could be that the mean North Island engine size is bigger than the South Island and if so, we can be fairly sure the difference could be as much as 1008cc or it could be that the mean South Island engine size is bigger that the North Island and, in this case, we can be fairly sure that the difference could be as much as 608cc."

AND

"I am unable to say that either island tends to have mean engine sizes which are bigger than the other. It is not possible to say the South Island has larger engine sizes, on average, than the North Island."

Powerful evidence and statistical processes, and slightly disappointed students!

> Given the size of the NZTA database and the question being asked, after the inference and conclusion I should have encouraged the students to think more deeply about possible reasons and also to re-start the PPDAC process with further questions. This development of new questions based on the experience of the PPDAC cycle is an important element of my teaching and it was a missed opportunity.

Adaptations

In a class with several international students, on reflection, I should have been more prepared to break down the technical language requirements of the series of lessons. Due to the mixture of learning styles engaged through the lessons I am confident that all students at least experienced and saw the bootstrapping process. The more able students, again on reflection, could have been encouraged to find out more about the development of the bootstrapping process (see Rutgers University article in Teacher notes) and also to explore a little of the statistical coding elements in RStudio to do with bootstrapping.

Within the iNZight VIT package (stat.auckland.ac.nz/~wild/iNZight/) there is a module dedicated to Confidence Interval Coverage. To deepen the students' understanding of the purpose, concepts and potential pitfalls of confidence intervals, I should have directed students in an activity to use this module. I cannot remember a part of the series of lessons where there was an opportunity from our engine size manual bootstrapping exercise to point out the fundamental idea of the confidence interval enclosing the true population parameter, for example, 95% of the time.

Teacher notes

I have found this article has helped improve my technical and historical understanding of the bootstrap process - stat.rutgers.edu/home/mxie/rcpapers/bootstrap.pdf.

At our school we are developing statistical coding modules from Year 11 upwards. This is at the very basic level and based on advice mainly from previous students who have studied Stage 100 papers or above at university and fed back that high school students learning statistical programming language would be an incredibly useful skill. RStudio allows some very basic skill development of the statistical programming language R, and has the advantage of a "ready to use" graphics suite incorporated into its user interface. Before the students were asked to undertake these series of lessons they spent a week with some basic coding activities, typically producing scatter graphs, boxplots, dot plots and summary statistics for datasets with at most a few hundred rows and a mere 5 variables at most. I noticed at the end of this experience several students (around a third of the class and mainly more able) had gone beyond the coding examples they were encouraged to complete and were, for example, adding extra aesthetics to their graphical displays.

When first learning the bootstrapping process in 2010 to 2012, I contacted the USA based statistician, Professor Brad Efron, Stanford University, who introduced it to the world. He was gracious enough to reply within a day.

My email to Brad:

> Dear Professor Efron
>
> Sorry to trouble you. I work as a high school mathematics and statistics teacher in the South Island of New Zealand. Our senior statistics curriculum is currently undergoing some very exciting changes, largely based around using technology for estimating statistical parameters.
>
> Teachers have been undergoing some pretty intensive professional development. However, one question keeps popping up with regards to bootstrapping. If the original sample to be resampled from is an extremely poor rogue one then no bootstrapped samples will ever 'cover' the true parameter.
>
> Our high school students are bound to ask and I have been unable to find resources which answer this.
>
> Is there anything you know of that could help?
>
> Thank you
> Mark Hooper

His reply to me:

> Dear Mark,
>
> No statistical method can protect one from rogue samples. The usual confidence interval for a mean xbar+-1.96*sigmahat will be far off too in such a case. All that bootstrap confidence intervals are supposed to do is cover the true value most of the time, say 95% of the time in the situation above. In other words, they Are supposed to miss 5% of the time! The bootstrap just makes this more obvious instead of hiding it under the mathematics.
>
> New Zealand seems to be far ahead of Palo Alto in terms of the high school curriculum.
>
> Hope the new curriculum goes well,
> Brad Efron

References

Arnold, P., Pfannkuch, M., Wild, C., Regan, M., & Budgett, S. (2011) Enhancing students' inferential reasoning: From hands-on to "movies". *Journal of Statistics Education*, 19(2), 1–32. http://www.amstat.org/publications/jse/v19n2/pfannkuch.pdf

Pfannkuch, M., Regan, M., Wild, C.J., & Horton, N. (2010). Telling data stories: essential dialogues for comparative reasoning. *Journal of Statistics Education*, 18(1). http://www.amstat.org/publications/jse/v18n1/pfannkuch.pdf

Pfannkuch, M., Wild, C.J., & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM – The International Journal of Mathematics Education*, 44(7), 899–911. doi 10.1007/s11858-012-0446-6

Wild, C.J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 247-295.

Materials required

- It is helpful to be able to access and use the online Visual Inference Tool VIT Online (stat.auckland.ac.nz/~wild/VITonline/) for the demonstration of some of the key concepts.
- Technology that is capable of installing and running the statistical software package "R" (r-project.org) that allows RStudio to operate (rstudio.com). The lesson does not require the use of RStudio. The majority of datasets that are used can be directly imported into other packages such as iNZight (stat.auckland.ac.nz/~wild/iNZight/).
- Card, scissors

Copyright information