Analysis of student sampling strategies in an activity with chocolate

Vacation Scholar: Christopher Lee
Supervisors: Dr Anthony Morphett & Dr Paul Fijn

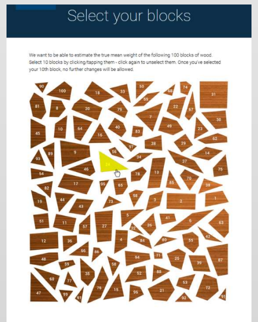
Introduction

This project analyses the most common sampling strategies that students adopt in the Chocs and Blocks activity. By deriving sampling distributions and modelling data sets from large lecture class, the theoretical model obtained from the data sets can be compared with data obtained from simulations. The intention of this project is to aid the learning and teaching of key statistical concepts to university and high school students.

Chocs and Blocks Activity

The Chocs and Blocks sampling activity is designed to introduce students to statistical concepts including sampling, sampling bias and sampling distributions.

Having been presented with a tray of 100 chocolate pieces, students are asked to estimate the average weight of a piece of chocolate on the tray. Following a series of questions from students regarding the chocolate pieces, they are directed to the Chocs and Blocks website. After selecting 10 blocks on the screen, the website calculates the mean weight of the chosen blocks.


Select your blocks

We want to be able to estimate the true mean weight of the following 100 blocks of wood. Select 10 blocks by clicking/tapping them - click again to unselect them. Once you've selected your 10th block, no further changes will be allowed.

Figure 1. The Chocs and Blocks website allows students to select a sample from a population of 100 chocolate blocks. After selecting 10 blocks, the sample mean is calculated and automatically recorded on a spreadsheet.

Following many repetitions of this activity, it has been found that students consistently overestimate the mean by approximately 10 units. The actual mean of the chocolate blocks is 33.5.

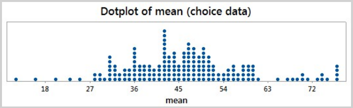


Figure 2. An example dotplot of sample means, where most of the data lies to the right of the true mean.

Methodology

The various sampling strategies adopted by students were first collated based on known strategies discussed in lectures.

For each strategy, 100,000 samples were simulated on R. The resulting sampling distributions were graphed.

With the data of 1258 individuals who had previously participated in the Chocs and Blocks activity, each sample was categorised according to its strategy. Using this, the likelihood that a student adopted a certain strategy was determined and used to create a combined sampling distribution of all strategies.

Results

Part 1: Determine most frequently used sampling strategies

Using data sets from large lecture classes along with information from post-activity discussion, the following categories were found to be the main strategies adopted by students:

- Systematic – samples selected based on location or fixed intervals of block numbers
- Stratified – samples selected by estimating proportion of block sizes in population and sampling according to estimated proportions
- Random clicking – samples selected by clicking randomly on Chocs and Blocks website
- Random – samples selected by entering block numbers into a random number generator

Further specifying these strategies, the following tree diagram was formulated.

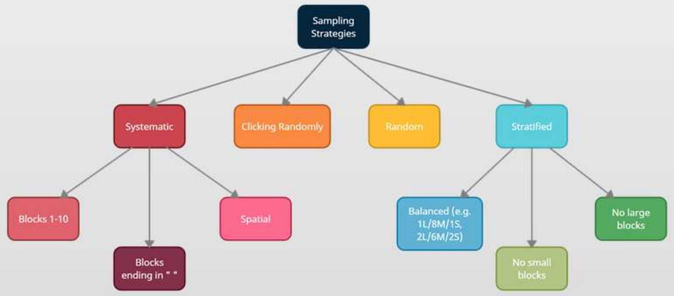

Sampling Strategies
Systematic
Clicking Randomly
Random
Stratified
Blocks 1-10
Blocks ending in " "
Spatial
Balanced (e.g. 1L/8M/1S, 2L/6M/2S)
No small blocks
No large blocks

Figure 3. Common sampling strategies adopted by students

Part 2: Determine sampling distribution of frequently used sampling strategies

After simulating 100,000 samples of each strategy, the sampling distributions were collated onto the same graph.

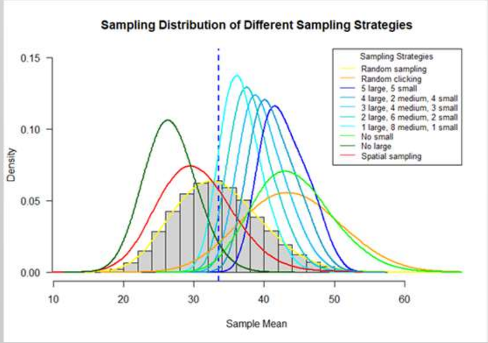


Figure 4. Comparison of sampling distributions of different sampling strategies

Some strategies did not involve randomness, with a certain sample mean achieved each time the strategy was used (systematic strategies based on block numbers).

Part 3: Determine combined sampling distribution of all sampling strategies

The 1258 samples available from students participating in the Chocs and Blocks activity were categorised according to the most likely strategy utilised. Using these data, the likelihood of a student choosing a particular sampling strategy was estimated.

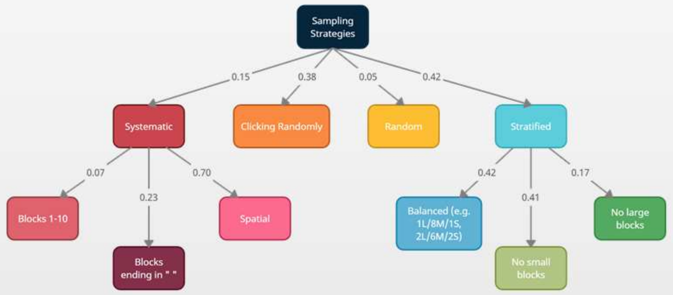

Sampling Strategies
0.15
0.38
0.05
0.42
Systematic
Clicking Randomly
Random
Stratified
0.07
0.23
0.70
0.42
0.41
0.17
Blocks 1-10
Blocks ending in " "
Spatial
Balanced (e.g. 1L/8M/1S, 2L/6M/2S)
No small blocks
No large blocks

Figure 5. The likelihood that students chose a particular sampling strategy was estimated using the 1258 samples of student data available

Implementing the sampling tree above into R, 100,000 samples were generated and the sampling distributions of the sample mean, number of large blocks selected and the minimum spanning tree were graphed.

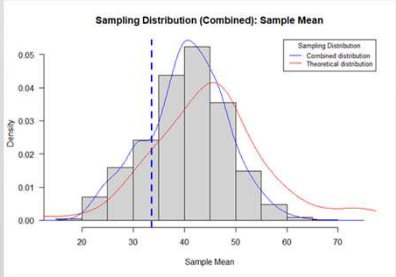


Figure 6. The sampling distribution of the sample mean incorporating all sampling strategies and their likelihoods against the theoretical distribution

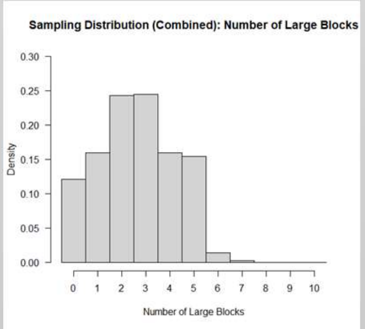


Figure 7. The sampling distribution of the number of large blocks chosen incorporating all sampling strategies and their likelihoods

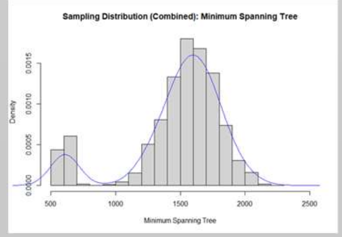


Figure 8. The sampling distribution of the minimum spanning tree of blocks chosen incorporating all sampling strategies and their likelihoods

Conclusion

The results obtained from the simulations and consequent sampling distributions are consistent with the observation that students consistently overestimate the true mean of the tray of chocolates.

As evident in Figure 4, the stratified sampling methods resulted in an overestimation of the mean. A sample obtained through randomly clicking chocolate blocks had a similar outcome. Furthermore, through categorizing the 1258 samples available, these sampling methods were found to be the most popular and frequently used (as depicted in Figure 5).

Incorporating all sampling strategies and likelihoods, the combined sampling distribution of the sample mean is depicted in Figure 6. Once again, the sample mean is most frequently found to be approximately 10 units greater than the true mean (33.5).

However, in comparison to the dotplot in Figure 1, the combined sampling distribution does not have multiple peaks. Furthermore, the combined distribution peaks at a lower sample mean compared to the theoretical distribution. This is likely a result of inaccuracies of the probability estimates of a student choosing a particular sampling strategy, having been determined using a limited sample of 1258. To further refine these estimates, Bayesian inference can be used.

For future lectures and for educational purposes, the sampling strategies that most accurately estimate the mean appear to be spatial sampling and balanced stratified sampling weighted heavily towards medium-sized chocolate blocks.

Acknowledgements

The Vacation Scholar Program has been an invaluable experience which has allowed me to explore my interest in statistics and statistical research. I have been able to greatly develop my technical skills and ability to code in R.

I would like to thank Dr Anthony Morphett and Dr Paul Fijn for their guidance, patience and unwavering support throughout the project.

References

Gunn, S. and Morphett, A., n.d. *Chocs and Blocks Activity - notes for teachers*. [online] Melbapplets.ms.unimelb.edu.au. Available at: <https://melbapplets.ms.unimelb.edu.au/activities/chocsandblocks/chocs-blocks-teacher-notes.html>.

https://melbapplets.ms.unimelb.edu.au/activities/chocsandblocks/index.html