Analysis of student sampling strategies in Chocs and Blocks activity

Jiayi Xu

Supervisors: Dr. Anthony Morphett & Dr. Paul Fijn

Introduction

This project investigates sampling distributions of several common strategies that students may apply in the Chocs and Blocks activity, and attempts to model data sets of samples collected from large lecture classes.

Chocs and Blocks Activity

Chocs and Blocks is a classroom activity designed for introductory statistics students, which helps them to build understandings of sampling, sampling bias, variability, estimation and sampling distributions. (The activity is described at https://mslc.pages.gitlab.unimelb.edu.au/chocs-and-blocks/)

Students are presented with a tray of 100 chocolate pieces (Chocs and Blocks website provides a virtual representation of chocolate pieces), which are irregularly shaped and vary in size. The task for students is to estimate the average weight of chocolates on the tray.

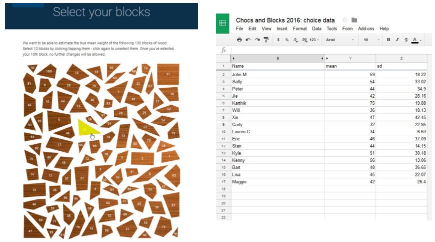


Figure 1: Chocs and Blocks selection process
After selecting 10 blocks, the website automatically calculates and records the mean weight of the blocks they've chosen.

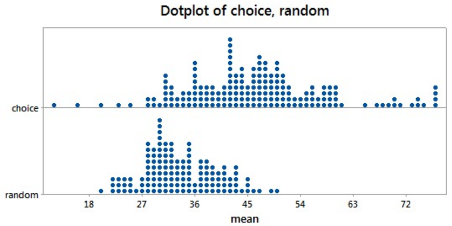


Figure 2: Example dotplots of means from student choice samples (above) & means from random samples (below)

Methodology

- First, common possible sampling strategies were determined. For each strategy, use R to generate 10,000 samples.
- Second, apply Support Vector Machine (SVM) for data training and testing, complete possible classification.
- Third, implement Gradient Descent Algorithm to find the optimal distribution of remaining sampling strategies.

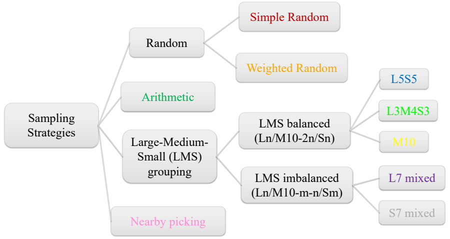


Figure 3: Possible sampling strategies

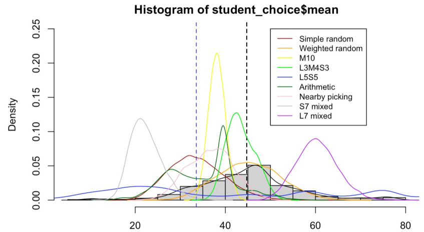


Figure 4: Density plot for different sampling strategies

Utilization of CV (computer vision)

OpenCV was employed to boost the efficiency of generating samples and reading.

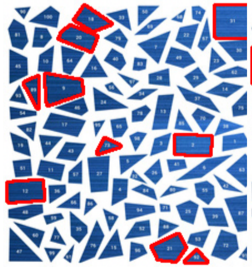


Figure 5: Visualization of Nearby picking (left)

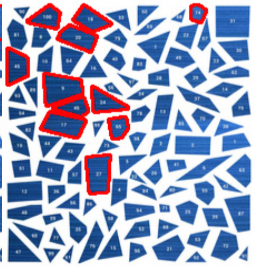


Figure 6: Visualization of L7 mixed (right)

Results

- **Part 1: Classify Arithmetic & Nearby-picking strategy by SVM**

Under strategy class: LMS balanced, LMS imbalanced, arithmetic, nearby binormal, weighted random, simple random

Use 90% data for training, 10% data for testing:

```
['LMS_balanced' 'LMS_imbalanced' 'arithmetic' 'nearby_binormal'
 'simple_random' 'weighted_random']

('LMS_balanced=>LMS_balanced', 683)
('LMS_balanced=>LMS_imbalanced', 2)
('LMS_balanced=>simple_random', 94)
('LMS_balanced=>weighted_random', 221)

('LMS_imbalanced=>LMS_balanced', 110)
('LMS_imbalanced=>LMS_imbalanced', 514)
('LMS_imbalanced=>simple_random', 174)
('LMS_imbalanced=>weighted_random', 202)

('arithmetic=>LMS_balanced', 9)
('arithmetic=>LMS_imbalanced', 1)
('arithmetic=>arithmetic', 951)
('arithmetic=>simple_random', 27)
('arithmetic=>weighted_random', 12)
```

```
('nearby_binormal=>nearby_binormal', 779)

('simple_random=>LMS_balanced', 66)
('simple_random=>LMS_imbalanced', 38)
('simple_random=>arithmetic', 1)
('simple_random=>simple_random', 728)
('simple_random=>weighted_random', 167)

('weighted_random=>LMS_balanced', 103)
('weighted_random=>LMS_imbalanced', 18)
('weighted_random=>arithmetic', 1)
('weighted_random=>simple_random', 155)
('weighted_random=>weighted_random', 723)

The mean accuracy on the given test data and labels is: 0.7575705139297456
```

Figure 7: SVM testing data classification

Mean accuracy of classifying the testing data: 75.76%.
Accuracy of classifying "Arithmetic samples": 95.1%.
Accuracy of classifying "Nearby-picking samples": 100%.

⇒ Therefore, among 1319 student choice data, there are 76 "Arithmetic" data, 129 "Nearby-picking" data.

- **Part 2: Find optimal strategy proportions by Gradient Descent Method**

Under strategy class: L5S5, L3M4S3, M10, S7 mixed, L7 mixed, weighted random, simple random

Use Gradient Descent Algorithm to fit mean and standard deviation distribution:

- Introduce hypothesis function $h_\theta(x) = \sum \theta_i x_i$
- Introduce cost function $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}[\,h_\theta(x^{(i)}) - y^{(i)}]^2$
- Start with random initial value θ
- For each iteration, approach minimal cost function

$$\theta_j' = \theta_j - \alpha \cdot [\frac{\partial}{\partial \theta_j} J(\theta)]\,, \text{ where } \frac{\partial}{\partial \theta_j} J(\theta) = [\,h_\theta(x) - y] \cdot x_j$$

$$\theta_j' = \theta_j - \alpha \cdot \sum_{i=1}^{m}[\,h_\theta(x^{(i)}) - y^{(i)}] \cdot x_j$$

[0.3463589316974409, 0.02504707088096908, 0.04002305882792069, 0.04926002229316281, 0.07345410939698645, 0.21024157857143586, 0.25561522833208417]
['L5S5', 'L3M4S3', 'M10', 'S7_mixed', 'L7_mixed', 'weighted_random', 'simple_random']

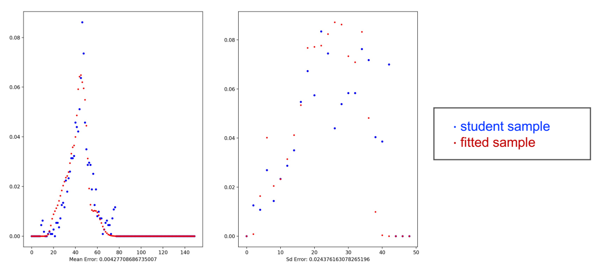


Figure 8: Fitted plots for student mean(left) and standard deviation (right)

⇒ Therefore, among the remaining student choice data, there are 34.6% L5S5 data, 2.5% L3M4S3 data, 4.0% M10 data, 7.3% L7 mixed data, 4.9% S7 mixed data, 21% weighted random data, 26% simple random data .

Conclusions

The modelling distribution obtained from the 10 simulated sampling strategies is consistent with the real density distribution of the collected student choice data, which both show an overestimation of the theoretical mean of the tray of chocolates.

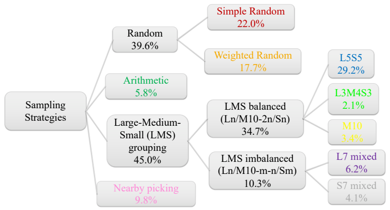


Figure 9: Estimated proportions of simulated sampling strategies

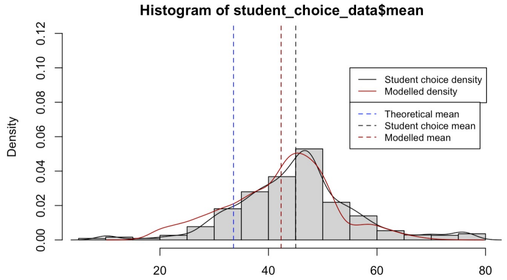


Figure 10: Modelled density distribution

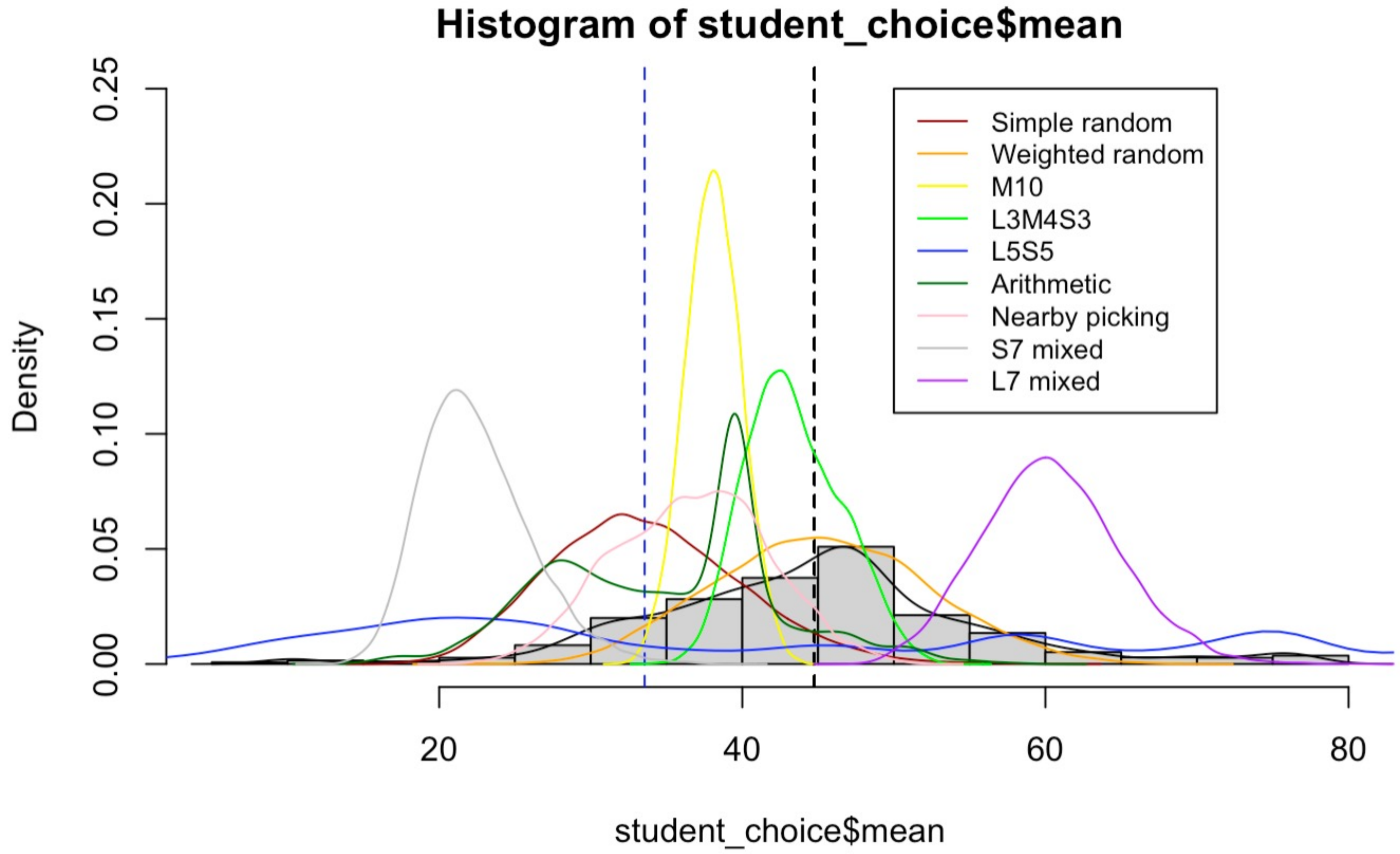
The reason of the overestimation might be resulted from the visual error that students are more likely to choose a larger-sized block than smaller-sized or median-sized one.

For further investigation, more emphasis need to be laid on the way to distinguish simple random sampling from other sampling strategies.

Acknowledgement

The Vacation Scholar Program has been an invaluable experience throughout my university life, which has allowed me to explore my interest in statistics and statistical research.

I would like to express my heartfelt gratitude to Dr. Anthony Morphett, Dr. Paul Fijn and Yanting Mu for their support and patient guidance throughout this project.


Histogram of student_choice$mean
Density
0.00
0.05
0.10
0.15
0.20
0.25
20
40
60
80
student_choice$mean
Simple random
Weighted random
M10
L3M4S3
L5S5
Arithmetic
Nearby picking
S7 mixed
L7 mixed