A Case Study in Multiple Correspondence Analysis: Clicker Question Responses to Cluster Different Types of Students

University of California, Los Angeles

Justin Yee

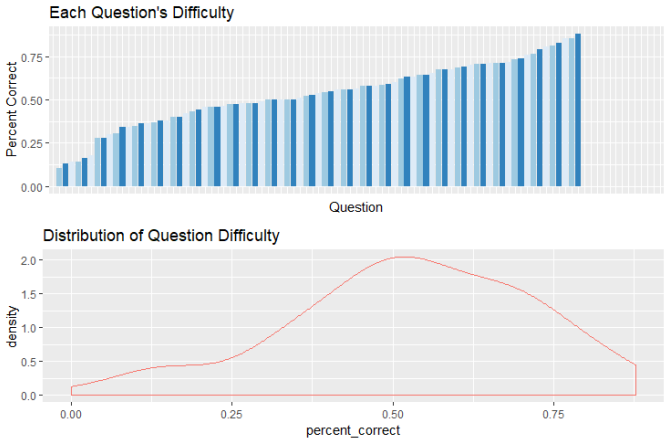
June 10, 2019

Introduction:

For the purposes of this study, the main focus of analysis was to draw insights of the different types of both questions and students from the dataset of clicker question responses registered from an Undergraduate Introductory Statistics course at UCLA. The nature of the dataset led to the main method of statistical analysis – Multiple Correspondence Analysis. Dealing with categorical responses (Multiple Choice Answers), and no labels or semantic meaning behind any of the answer choices given, besides the key to the correct response, clustering methods and data analysis that did not rely on unavailable information were necessary. As such, the analysis of this study distinguishes between only two types of student responses: 1. Student responses that are either identical (Student 1 answering Question 1 with 'A' and Student 2 answering Question 1 with 'A'), or 2. Student responses that are not identical, with each type of non-identical response being given the same 'weight' in terms of dissimilarity or distance metric.

Exploratory Analysis:

To first understand the dataset, exploratory analysis of summary statistics was conducted. To start, I first looked at the summary statistics of the Questions variable, to assess the distribution of Question difficulty, as measured by total percent correct by each question.

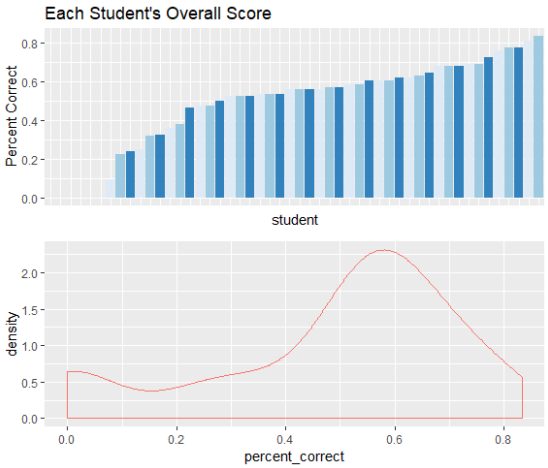


As seen by the bar chart and density plot above, the distribution of Question difficulty is slightly skewed left, as most questions had above 50% correct. The summary table below confirms the distribution.

Table 4:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
percent_correct	84	0.525	0.191	0.000	0.415	0.674	0.878

Next, we take a look at the exploratory analysis of the student distribution.



Looking at the bar chart and density plot, we can see that the distribution of students' overall percent correct score is similar to that of the question difficulty distribution (skewed left). The below summary table confirms this similar distribution, although with slightly lower averages than that of the question difficulty distribution.

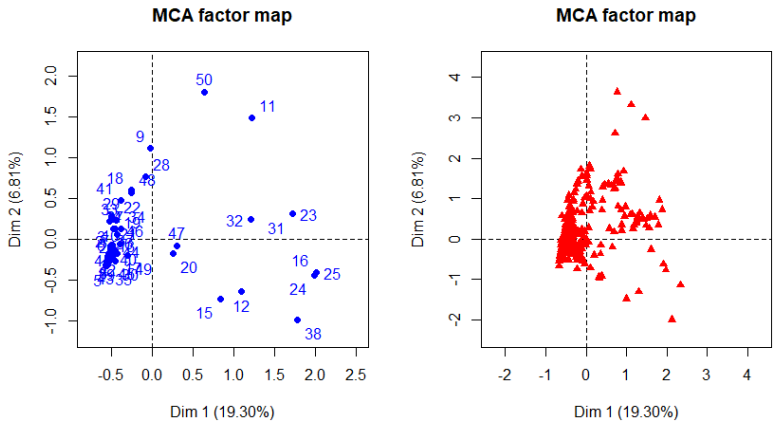
Table 5:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
percent_correct	50	0.485	0.237	0.000	0.363	0.628	0.833

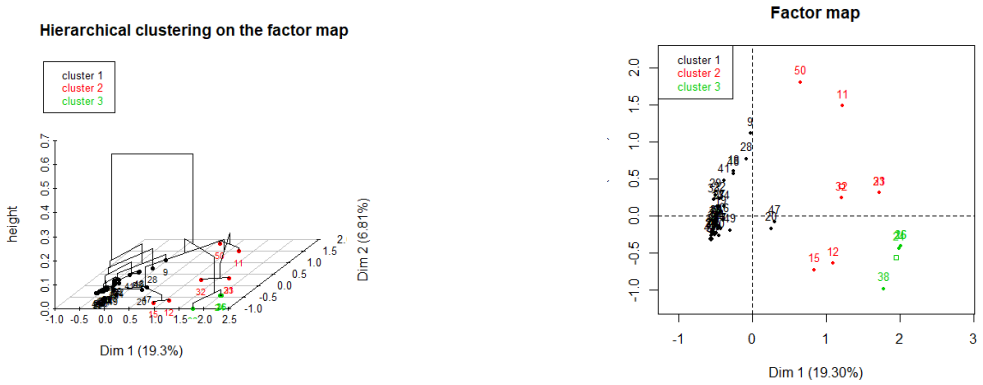
From this exploratory analysis, we can see that while the most common answer by far is the correct answer, there is still much room for improvement in the student answers, and perhaps there are trends and patterns present within students who usually get answers correct versus students who have a lower total percent score. We will explore these patterns further in our analysis and statistical methodologies.

Multiple Correspondence Analysis:

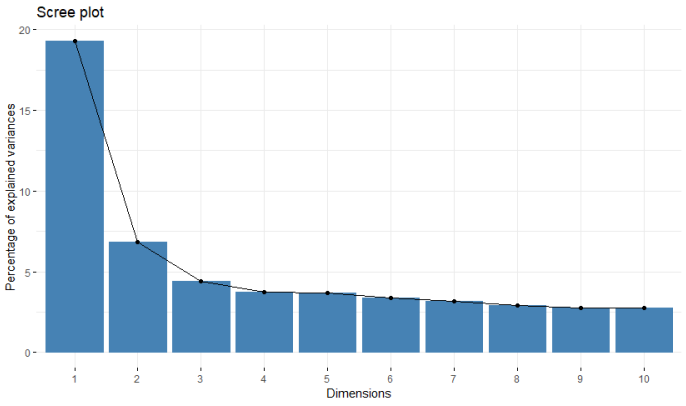
In hopes of finding distinctive and informative clusters about the different types of students through an unsupervised learning approach, I employed the statistical method of Multiple Correspondence Analysis. Similar in nature to the method of Principle Components Analysis, Multiple Correspondence Analysis is a dimensionality reduction technique that is specifically designed to handle categorical variables as opposed to numeric, continuous variables. Below are the results from the Multiple Correspondence Analysis depicted in the form of a separated biplot. The left MCA factor map in blue displays the individuals (Students), while the right MCA factor map in red displays the variables (Questions).



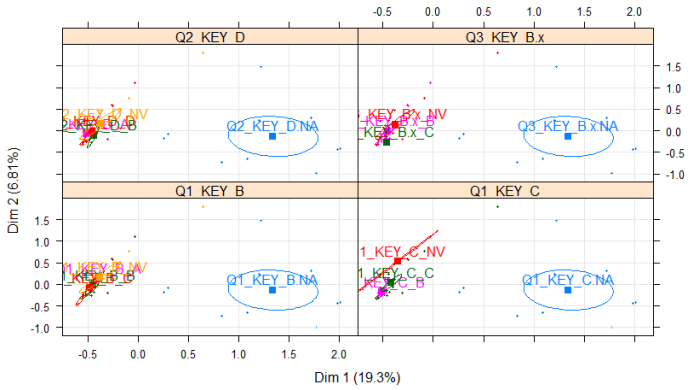
Following this projection of the data into two Principal Components (dimensions), hierarchical clustering was performed to obtain three distinct clusters. We will later dive into the results of these clusters and their interpretations within the context of our dataset.



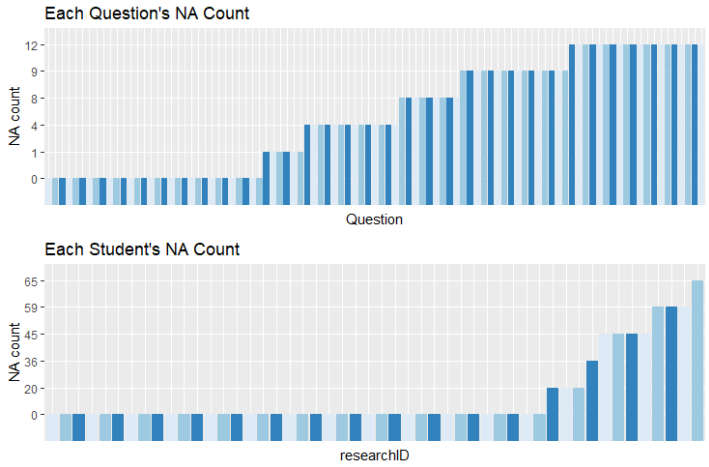
Scree Plot results from the Multiple Correspondence Analysis indicate that the first two dimensions, and the first dimension in particular, capture the variance of the data fairly well (26.11%), considering that there are 97 Questions (dimensions) to start with the original dataset.



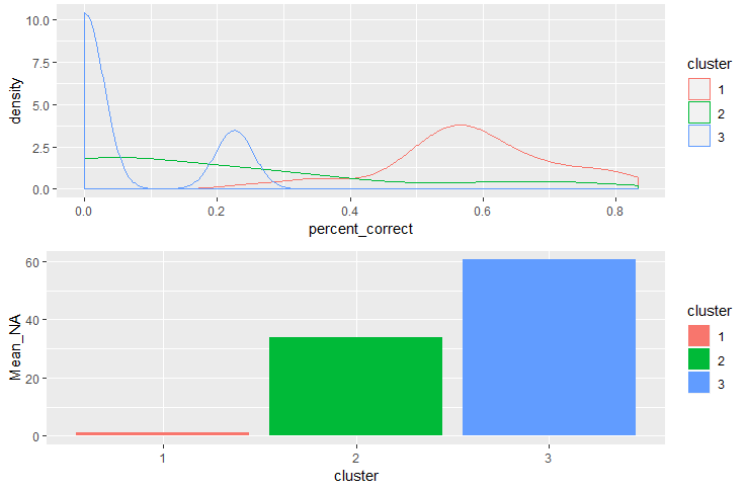
95% confidence ellipses of the first four Questions reveal that the most distinguishing answer type is the “NA” answer choice, which denotes that the student was not accounted for in the clicker question registration. The “NA” answer choice also differs from the “NV” answer choice, as “NV” indicates that the student had registered their clicker, but did not respond to the question, while the “NA” response indicates a student who had not yet registered their clicker.



Looking at the counts of “NA” responses by Question and by Student (researchID), we get the following results. The distributions for both Questions and Students are left skewed, telling us that most questions and students had registered clicker responders. These results may give us insight into the meaning behing the hierarchical cluster results derived from the Multiple Correspondence Analysis. The analysis of the “NA” response values and the hierarchical clusters may reveal a connection between students who register their clickers later in the quarter and their performance on the questions they answer following their late registration.



When looking at the mean count of "NA" student responses within the context of the clusters produced from the hierarchical clustering algorithm, we see that the "NA" responses have a clear and substantial effect on the creation of the clusters of different types of students. Additionally, when analyzing the density plots of clusters in the context of the students' Percent Correct score, we see that cluster 1 has higher percent correct students compared to cluster 2 and cluster 3. Therefore, we can deduce that there are definite "types" of students who register their clickers later in the class, thus negatively impacting their performance on clicker questions for the remainder of the quarter.



To gain further insight into the interpretation of the hierarchical clusters, I derived my own version of a "correlation" matrix for the categorical response variables (Multiple Choice Answers). To do this, I matched up each question and answer while preserving the correct ordering of the question and answers in order to directly compare two students to each other. This "proportion of exactness" ranged from the values 0-1, mimicking the correlation coefficient of numerical variables. While the answer responses themselves have no meaning, outside of the key (correct answer), NV, or NA values, they serve to show if certain students are answering in the same exact way, thus denoting a measurement of similarity. This will output a very large matrix (50 by 50 matrix), since we must compare all students against each other. Since this output is impossible to look at, we will look at the highest "proportion of exactness" coefficient for each row.

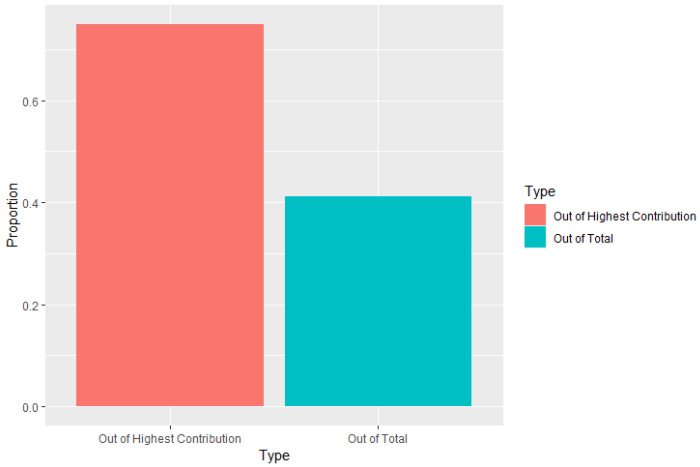
I have displayed a table of the results below, denoting cluster 2 pairs in red and cluster 3 pairs in green. The results of the clusters match the highest pairs exactly. Therefore, we can say that the clustered groups also have high "correlation" with each other in terms of answering the same exact ways.

Table 6:

row	column	proportion
25	16	1
31	23	1
24	16	0.990
4	43	0.804
5	43	0.804
43	4	0.804
32	23	0.763
38	16	0.763
11	23	0.742
44	5	0.742
45	5	0.742
1	5	0.711
2	43	0.711
36	45	0.711
40	4	0.711
10	40	0.691
14	27	0.691
17	4	0.691
21	36	0.691
27	14	0.691
35	17	0.680
39	17	0.680
8	2	0.670
26	4	0.660
50	11	0.660
3	5	0.649
7	45	0.649
12	15	0.649
15	12	0.649
37	36	0.649
46	3	0.649
22	41	0.639
41	22	0.639
42	5	0.619
29	33	0.608
33	29	0.608
49	36	0.608
6	4	0.598
13	14	0.588
34	2	0.588
19	26	0.577
30	8	0.557
18	8	0.536
20	43	0.526
48	41	0.526
9	50	0.515
28	18	0.464
47	20	0.464

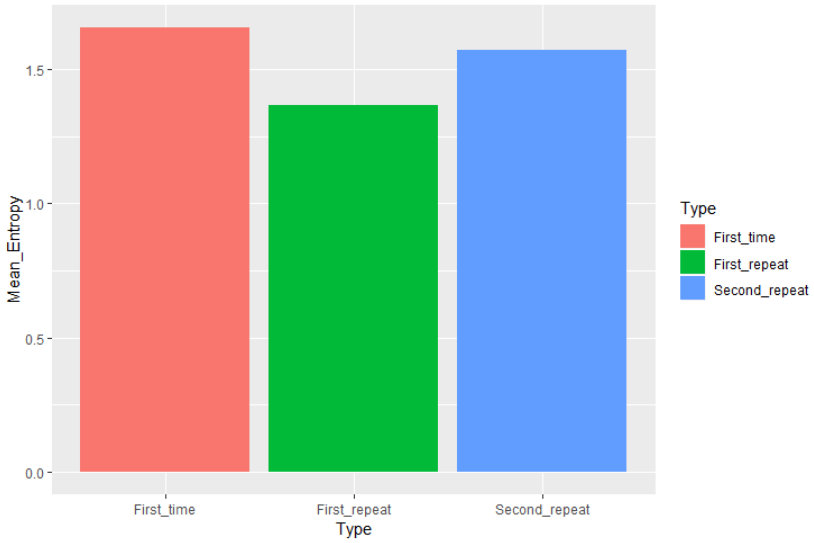
Now that we have gained some insight into the different types of students, we will turn our attention to the different types of questions. An advantage of this dataset is the labeled question types, in that there are certain questions that are repeated. Questions that the professor deemed more difficult based on the total percentage correct of the students were often repeated. I sought to determine the effect of Repeated Questions on the creation of the Principal Components from the Multiple Correspondence Analysis.

Of the 16 highest contribution variables(Questions) to the first Principal Component, each question having the same amount of contribution, 75% of those questions fell in the category of a Repeated Question. In contrast, Repeated Questions only made up 41% of the total Questions in the dataset. These results indicate that the variation present in the types of Students can be greatly attributed to these Repeated Questions.



Another lense to look at the informational value of Repeated Questions is the amount of Shannon's entropy for each question. Shannon's entropy increases as the level of uncertainty for a response increases. In other words, Shannon's entropy is at a maximum when the response variables are distributed uniformly, and at a minimum when only one response variable is observed and the other response variables have a frequency of zero. According to information and psychometric theory, a question is considered to be more valuable for a researcher if the question carries a high value of entropy.

My intution led me to believe that Repeated Questions might show a significant difference in entropy compared to Non-Repeated Questions, however they did not. However, if we break down the Repeated Questions into three different categories – First Time, First Repeat, and Second Repeat, we can see some noticeable changes in mean entropy levels between these categories. Looking at the bar chart below, the entropies are fairly similar, however the First Time is the highest, followed by the Second Repeat, which is followed by the First Repeat. These entropy levels could signal that the First Time is the most difficult type of question, and there is a significant amount of improvement under the First Repeat. However, the Second Repeat entropy level is greater since only very difficult questions were repeated a second time (2 questions in total).



Conclusions and Further Studies:

Through statistical analysis of the clicker question dataset, we were able to draw insights from the Multiple Correspondence Analysis, Hierarchical Clustering following the Multiple Correspondence Analysis Principal Components, and Shannon's Entropy Analysis.

From the Cluster Analysis, we were able to conclude that the "NA" response values play a significant role in the formation of distinct clusters and in the performance of Students in regards to their Percent Correct Score. From this, we can conclude that it is particularly important for Students to register clickers early on in the class and building a strong foundation of statistical knowledge in order to perform well and understand the material going forward.

By creating a "proportion of exactness" and comparing the highest "correlated" pairs of students to the results of the hierarchical clustering, we were able to gain even an additional insight into the interpretation of the clusters from the Multiple Correspondence Analysis. We found that these clusters of students tend to answer questions in the same exact way for many questions, thus giving us a way to group students based on their mode of thinking, if we were to give semantic labels to the question answers retroactively.

Lastly, we looked at the types of Questions in the dataset, Repeated and Non-Repeated, to gain information concerning the performance of Students on different types of Questions. In general, from the measurements of Shannon's Entropy, we saw that a greater level of consensus was reached when a question was repeated, thus showing the benefits of repeating a question for the students to discuss amongst themselves and try again.

In future studies, we can take a deeper dive into specific questions and specific pairs of students to gain semantic meaning behind the questions and answers by categorizing them beforehand. This study was meant to gain general insight into the different types of students and questions, but more specific conclusions can be reached in future studies.

References:

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01