Utilization of large-scale disaster prevention data in data science education and research

○Shizue IZUMI, Michinori HATAYAMA

ABSTRACT: we propose a design for an elementary statistics course incorporating a massive open online course (MOOC), real-data-oriented project-based learning (PBL), and artificial intelligence (AI). In April 2017, Shiga University established the first department of data science in Japan. We intend to teach data science by enabling students to accumulate experience in solving real-world problems using various data and creating value from the data. The course contents include development of critical thinking, communication, and teamwork. Whereas students utilize MOOC to acquire knowledge outside the classroom, they acquire skills in communication and teamwork inside the classroom through PBL practice by examining large-scale disaster prevention data. We hope that our proposed model will help provide the students with a balanced introduction to statistical concepts, methods, and theory.

● Course design

Figure 1 shows the first grade's curriculum tree of data science program in Shiga University. Here, the term first grade refers to the first year of college. As part of courses on data analysis, freshmen learn real data handling and the basics of statistical thinking. In the course titled "Basics of Data Analysis," which is part of the first semester, freshmen learn about different types of data (discrete and continuous) and how to view a statistical graph. In addition, they learn cross-tabulation and descriptive statistics such as mean and correlation. Further, they briefly study linear regression and time-series data. In the course titled "Fundamentals of Statistical Inference," which is part of the second semester, freshmen learn the concepts of population and sample, which are key to the development of statistical thinking (takemura et al., 2018).
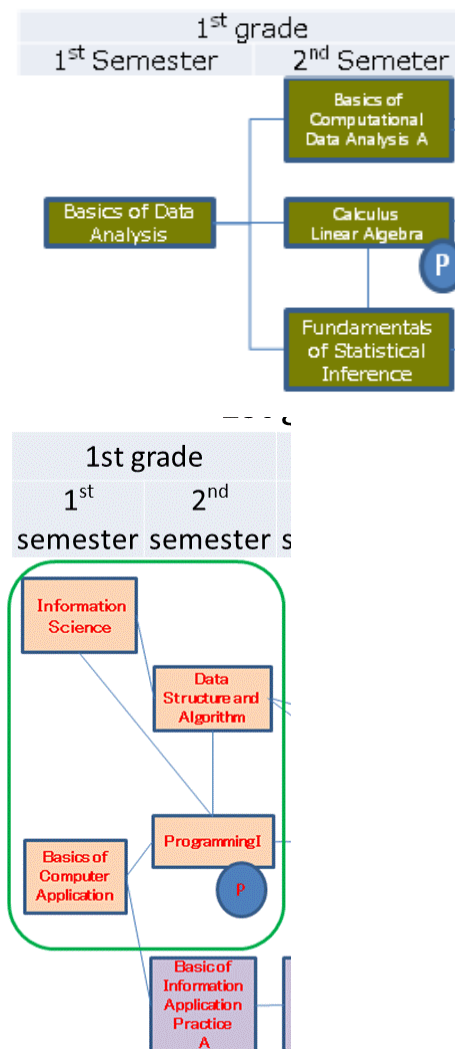

Figure 1. Courses on data analysis (top figure) and information (bottom figure) in the first grade's curriculum tree of data science program in Shiga University (Shiga University, 2017)

Further, to understand statistical estimation and hypothesis testing, students learn about the random variable and its distribution. In the second semester, they attend the calculus and linear algebra courses to build a sound mathematical knowledge foundation, which is important in understanding statistical theories.

These courses use an extensive range of active learning techniques. In both "Basics of Data Analysis" and "Fundamentals of Statistical Inference," a massive open online course (MOOC) is used outside the classroom. On the other hand, during the second semester, the course titled "Fundamentals of Statistical Inference" comprise face-to-face classroom activities, including a guest lecture, problem exercises, and PBL practices. One of the themes of the guest lecture in 2019 was data management in insurance mathematics. As part of PBL practices, five students form a group and accumulate experience in solving a problem using the Problem-Plan-Data-Analysis-Conclusion (PPDAC) cycle for large-scale disaster prevention data collected from city hall web pages in Japan. Our PPDAC cycle includes data cleansing in the Data phase and (written, oral, and visual) communication of results in the Analysis phase. In hybrid (blended)-style active learning, students express various statistical concepts using their own words (Izumi, et al., 2016). This type of courses enables more flexible learning than traditional lectures since the former involves lesser limitation of lecture time than the latter.

Further, the course titled "Basics of Computational Data Analysis A" teaches how to analyze data using the tools and functions available in Microsoft Excel. The course contents include computation of descriptive statistics, such as average, statistical estimation, hypothesis tests, linear regression analysis, and analysis of variance. In the first year of college, Microsoft Excel is chosen as the primary tool of data analysis for students.

The quality assurance of statistical education is performed by the Ministry of Education, Culture, Sports, Science and Technology and an external advisory board involving our department. In their fourth semester, all students take a grade 2 examination to qualify for the Japan Statistical Society Certificate.

● PBL practices

Students accumulate their experiences in real-data-oriented PBL practices from the first year of college onward and learn various aspects of problem-solving. They tackle exercises using the real data received from collaborative institutes and companies and public big data such as e-Stat (https://www.e-stat.go.jp/) and Regional Economy Society Analyzing System (https://resas.go.jp/#/13/13101). By learning these practices for four years, students will acquire not only expertise in data analysis but also extensive knowledge in various fields.

In the course titled "Fundamentals of Statistical Inference," we set up a theme of PBL practices and created educational materials in cooperation with the Disaster Prevention Research Institute, Kyoto University. As part of this course, a mini project is allotted to a group of five students to analyze data on climate-, earthquake-, and geography-related regional disaster prevention. The results of PBL practices are summarized in a report and a statistical graph poster. Such activities aim to enhance students' knowledge of statistics and emphasize the report and graph techniques of statistical expression. Following a poster tour, a group presentation is evaluated by teachers and the students of other groups, whereas a report is evaluated by a teacher alone. Rubric SPART is created based on three viewpoints: statistical literacy, reasoning, and thinking (Fukazawa, et al., 2018). Based on the rubric SPART for peer review, we create and use a checklist for each phase of the PPDAC cycle. The results of the evaluation are instantly fed

back to students through an electric portfolio to encourage students to reflect on what they have learned.

Other courses on PBL practices use consumption and purchasing, annual health examination, social networking service, public statistic, and regional mobile data provided by collaborators. Topics chosen by students include a trend of purchasing chocolates, the relation between blood pressure and structure, Twitter data analysis for sport events, and a comparison of sex- and age distribution of visitors between cities.

● Comparison with other universities

Our educational model is consistent with the models prevalent in other countries. Our model is based on the Guidelines for Assessment and Instruction in Statistics Education college reports 2016 (ASA, 2016a). Our data science program includes a series of courses in data analysis and information science. When building our curriculum, we acknowledged the Curriculum Guidelines for Undergraduate Programs in Statistical Science (ASA, 2014) and Curriculum Guidelines for Undergraduate Programs in Data Science (ASA, 2016b). We find some similarities between the undergraduate programs in statistics and data science program offered by Yale University (2017) and Ohio State University (2018). According to Coursera MOOC catalogue, the courses "Exploratory Data Analysis" and "Statistical Inference" offered by Johns Hopkins University (2018) seem to have similar contents as the courses "Statistics I: Basics of Data Analysis" and "Statistics II: Methods for Statistical Inference" used in our model.

Since our data science program has just completed its third year and we have no information regarding its performance during the coming years, it will be interesting to thoroughly examine the features in our model that do and do not work well after the first four years.

● CONCLUSION

Our model is structured around an integrated combination of lectures in Japanese MOOC, problem exercises, and small-group PBL practices. An interactive teaching method requires several hours of preparation in advance; the use of LMS and MOOC reduces this burden. In addition, it is important to maintain the motivation of students to learn outside the classroom and increase their interests in PBL practices. In this respect, choosing a familiar topic among students may be helpful. Furthermore, it is important to appropriately evaluate the results of implementation and instantly provide feedback to the students. By reviewing the merits and demerits of applying this model, we can improve our proposed model further. In the future, we will attempt to develop virtual teaching materials and textbooks, as well.

From a preliminary review of our model based on first-year experiences, we tentatively conclude that our model is a promising one from the perspective of future data science education. While students are satisfied with traditional lectures inside the classroom, they find it challenging to learn mathematical and statistical materials using MOOC. PBL practices motivate them to think about the possibilities of data science applications. Students may also recognize a reason why they want to become data scientists.

We hope that our proposed model will provide students with a more balanced introduction to statistical concepts, methods, and theory. Our model may encourage both science and liberal arts major students to consider statistics as a potential career path or, at the very least, a topic of high interest. Further, our model may guide teachers to employ student-centered pedagogy in their future classes.

● ACKNOWLEDGEMENTS

REFERENCES

American Statistical Association. (2014). Curriculum Guidelines for Undergraduate Programs in Statistical Science. www.amstat.org/asa/education/Curriculum-Guidelines -for-Undergraduate-Programs-in-Statistical-Science.as px

American Statistical Association. (2016a). Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Reports 2016. www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_ Full.pdf

American Statistical Association. (2016b). Curriculum Guidelines for Undergraduate Programs in Data Science. www.amstat.org/asa/files/pdfs/EDU-DataScienceGuid elines.pdf

Statistical Inquiry Process and Assessment. to appear in Proceedings of the Institute of Statistical Mathematics, 66. (in Japanese)

Izumi, S., Sakurai, N., & Fukazawa, H. (2016). Interactive class design and its assessment in undergraduate statistical education. Statistics Education and Research: Institute of Statistical and Mathematics, 362, 5-10. (in Japanese)

Ohio State University. (2018). Undergraduate data analytics major. data-analytics.osu.edu/

Shiga University. (2017). Website of Faculty of Data Science. https://www.ds.shiga-u.ac.jp/

Takemura, A., Izumi, S., Saito, K., Himeno, T., Matsui, H., & Date, H. (2018). Shiga-University Model of Data Science Education. to appear in Proceedings of the Institute of Statistical Mathematics, 66. (in Japanese)

Yale University (2017). Undergraduate Statistics and Data Science Programs of Study 2017–2018. catalog.yale.edu/ycps/subjects-of-instruction/statistics/