DEVELOPING A MASSIVE OPEN ONLINE COURSE OF INTRODUCTORY STATISTICS IN JAPAN

Fumitake Sakaori
Chuo University, Tokyo, Japan
sakaori@math.chuo-u.ac.jp

We developed an introductory statistics course on a Japanese massive open online courses (MOOC), and held the course several times. This course provides basic knowledge about necessity of statistics, descriptive statistics, and usage of official statistics for wide range of people, mainly for business person. Also, we provided a (toll) one-day flipped classroom for applicants. In this paper, we will compare it with the ordinary university's course in assessment and percentage of correct answers, compare the time series variation of participants, and discuss additional flipped classroom and problems of manage a MOOC course.

INTRODUCTION

Open education such as MOOC (Massive Open Online Course) is educational system with texts and video lectures available to the public free of charge via the web, and without any academic admission requirements. It appeared in the United States with a conflict situation of aiming to promote lifelong education and advertise university education, and increase of school expenses due to subsidy and staff reduction. Also, the MOOCs can be seen as an important tool to achieve a Sustainable Development Goal (the goal 4) of the 2030 agenda.

The MOOC is an integration system of massive open courses, e.g., the Coursera and edX, in which some Japanese universities participates. There are plenty of the MOOC courses of statistics and data science, which cover a lot of topics of basic statistical theory such as descriptive statistics, inferential statistics, more advanced and applied statistics, and statistics tools such as R and Python.

Although these MOOCs are huge and cover wide range of statistics, one of the big problem for Japanese people, same as most of Asian country, is the language barrier of English. Then the Japan Massive Open Online Courses Promotion Council (JMOOC) was established in 2013, and started to provide open online courses in April 2014. Until now JMOOC has several platforms such as “gacco”, “OpenLearning, Japan”, “OUJ MOOC”, and “Fisdom”.

Our MOOC course opened in the gacco, the biggest platform of JMOOC, which is based on the Open edX, an open source MOOC system. There are over 400 thousand members and were about 200 classes in gacco until now.

The name of our course is “Statistics I – the basics of data analysis” and opened 5 times until July 2018 (actually, the last course will start in April 2018, after the deadline of this paper). The author is one of five instructors of this course. The course description will be given in the following section. Upon the needs of statistics, the number of participants in this course is very large: about 15,000 at the first time, and 7,500 at the fourth. Now we have several courses about statistics in gacco, e.g., “Statistics II – methods of inferential statistics”, “Statistics III – multivariate data analysis”, “Introduction to the data science for business person”, “Introduction to the data science for high school students”, and “Open data of official statistics for everyone”.

There exist few studies about MOOC statistics courses, but many about online classes. Mills and Raju (2011) gave detailed review of online statistics classes. Everson et al. (2008) and Gazioglu (2012) emphasized the importance of group discussion via BBS and forum instead of active learnings in the classroom. Also, they stressed that the instructors should communicate with participants via chat or e-mail. Several papers compared the education effect of online classes to ordinary university classes, and most of them reported no significant difference between them. For graduate school students, Hahs-Vaughn et al. (2017) compares hybrid classes and online classes for teaching statistical literacy, and these delivery formats have no significant differences. Some paper, however, reported higher rate of dropouts in online courses.

Anyhow the online courses give different perspectives with ordinary university course, and need various devises to obtain sufficient education effect. Further, it should be noted that the

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

background of participant of MOOCs are diverse compared with ordinary class. Also, the MOOC courses are so huge, and one-to-one contact to the participants is practically impossible.

In this paper, we will give the course description and discuss about education effect of the course.

COURSE DESCRIPTION

The main content of our MOOC course is basics of descriptive statistics. By the importance and the purpose, we included slightly advanced topics such as multiple contingency tables, Lorenz curve, and regression analysis. Also, we added some topics of official statistics. Five instructors are in charge weekly. The followings are weekly theme and content.

- Week 1: Motivation and an overview. Understand the importance of decision making with data, and study the development of statistics. Also, study the process of data analysis (PDCA or PPDAC cycle) and types of data.
- Week 2: Description of qualitative data and contingency tables. Study point to draw graphs such as bar, pie and band graphs. Also, study how to make and read contingency table, the difference between association and causal relationships, and multiple contingency tables.
- Week 3: Description of quantitative data. Know how to grasp the tendency of data with graphs such as histograms and box-and-whisker plots, and numbers such as central values and dispersion. Also, study deviation values and Lorenz curves.
- Week 4: Correlation and descriptive time series analysis. Study how to express the relationship between two variables via scatterplots and correlation, and how to use regression models. Also, understand index, change rate and growth rate to grasp time series data.
- Week 5: Official statistics, mechanism of statistical survey, and summary of the course. Study how to use the portal website of the government statistics and Ministry of Internal Affairs and Communications. Also, mention the importance of statistical inference for further study.

The main contents are video lectures, assessments, and final exam. Free online text (handouts of the slides) are available. Also, the participant can discuss on the BBS. Additionally, we had a one-day toll flipped classroom for applicants. The detail is given as follows.

Video lectures and assessments

Each week consist of about 8 video lectures. Both the instructors and the slides are shown in the video. The video lectures are subtitled and vary playback speed. Each video is about 10 minutes long, and the participant can watch in pockets of time, e.g., in the commuter train, via their smart phones.

There are 10 assessments (1 point each) from first to fourth week, and 30 problems of final exam (2 points each). The passing score is a total of 60 point or more out of 100. These are all five-option questions about statistical literacy, which refers to understand fundamental statistical concepts and terms, and statistical reasoning, which refers to interpret based on sets of data or summaries and graphs, and combine several statistical ideas. The assessment can be answered up to third times, but the final exam can be only once. This style of the five-option questions follows the Japan Statistical Society Certificate.

In this course, we have never used the function that the participants score their reports each other yet, which is installed on the Open edX system.

Discussion

In gacco system, there exist online discussion system of thread-type BBS. It can be used for discussing within the participants, and questions to the instructors or management side. As mentioned above, this kind of group discussion on the BBS or forum is quite important in online courses.

Compared with ordinary university classes, we felt the broad background and active enthusiasm of the participants: the number of final threads at the first time was 229, and the number of comments went up to 913. We should have checked the discussion many times every day during the course period, but still it was hard to check all the writes. In addition, there were many questions that we did not know how to answer in various ways, such as being vague

questions or being lost in judging how far to tell. Anyway, confirmation and correspondence of the discussion is overloaded beyond imagination.

Documents and supplementary book

In this course, the documents can be downloaded via the web. The documents are the handouts of all slides in the video. Also, the data used in the lecture can be downloaded and the participants can examine the calculation and make graphs as in the lecture.

Additionally, we provided toll (not complementary) supplementary book, "study notes". It provides summary of each page with fill-in-the-blanks. The participants can study actively and effectively by filling the blanks. Further, we provided simulated assessments and final exams, their explanation, how to use R and Excel to draw graphs and calculate summary statistics, and a glossary of statistical terms.

Flipped classroom

We had a (toll) flipped classroom for applicants. In this one-day classroom, lecture and active learning session are provided. The number of applicants is about 180 at the first time, where the seventy percent of applicants is male and their age is from 20's to 70's. Most of them are businessman.

The lecture is about supplement of online course and some topics that had many questions on the BBS. Also, some participants give the presentation about their data analysis for optional submissions. As the active learning, group discussion and data analysis practice using R were examined.

It is quite interesting that the participants' responses were different to that of students: they tried to interpret the statistical results based on their experience of real society. They were almost zealous, but the difference in their level of knowledge of statistics and R made the group discussion and data analysis practice difficult.

ANALYSIS OF EDUCATION EFFECT

In this section, we describe tendency of participants and education effect of this MOOC course by several aspects.

The number of participants and Response rates

The number of participants are 14,961, 10,888, 6,946, and 7,457 for the first (in 2014), the second (in 2015), the third (in 2016) and the fourth (in 2017). It declines slightly, but still many people are participate compared with other courses.

Figure 1 is the average response rate of assessments and final exam. The completion rates were not obtained, but the response rates of the last week is approximately 10% or 15%, whilst Jordan (2015) reported the current average completion rate in MOOC was 15%, almost the same level. Average response rates slightly decreasing every week, no special difference is found between four times.

The percentage of correct answers

The average percentage of correct answers in the assessments and the final exam is about 85%. There is no big difference between four times. Most of the percentage are over 90%, but under 50% in some difficult problem. Figure 2 shows a histogram of the percentage of correct answers at the fourth time.

The problems with a low rate of correct answers are as follows:

- A statistical literacy problem of choosing one table with the rows correspond to individuals and the columns correspond to variables out of 5 tables. The other tables are contingency table for example. Many participants chose the contingency table by mistake. They might misunderstand the meaning of individuals by variables, or be hard to understand what the problem means.
- A statistical reasoning problem of reading the tendency of distribution from a histogram and summary statistics. This problem can be solved if they could understand the relationship

between the shape of distribution and the central values (mean and median). However, many participants could not do so.

- A statistical literacy problem of interpreting the result of simple linear regression. Many participants did not understand the true meaning of the slope and coefficient of determination (the square of multiple correlation coefficient).
- A statistical literacy problem of asking the knowledge of the system of census and statistical survey in Japan. Their knowledge is simply insufficient.
- A statistical reasoning problem of explaining what is the population and the sample of a national survey giving the information of the detail of the survey and mechanism of sampling (two-stage stratified random sampling). They could not understand the population of this survey.

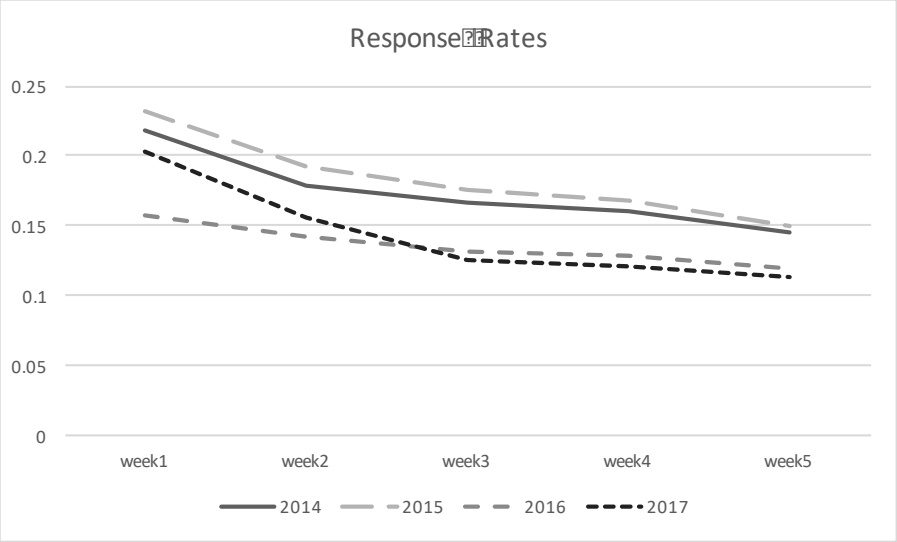

Figure 1. The average response rate

Their correct answer rate was very high although there were many complicated problems of reading graphs and statistics. This tendency seems to be somewhat different from ordinary lecture at university from the experience of the author; students are relatively good to memorize terms and concepts, whereas reading complex information tends to be weak.

The reason for this may be as follows. The MOOC participants is more enthusiasm and voluntary than in ordinal classes. Also, there may be many students who are not, but there is a possibility that such students did not answer the assessments. That is, there might be a selection bias by looking at the correct answer rate of only respondents. Therefore, it may be problematic to measure the discriminating power of the problem only from these correct answer rates. Also, the correct answer rate of the last questions in the final exam of 30 questions tends to be very low, and there is a possibility that the concentration ability is not continued so far. Of course, we should consider that they could answer up to three times for weekly assessment. In practice, the average number of answer in some complex problem is over 2. That is, they missed once in average.

CONCLUSION

In this paper, we developed the introductory course on statistics on the Japanese MOOC and discuss its characteristics through several times of implementation. We should consider the point that the background of the participants is different from the ordinary university class. Also, we found that it is difficult to measure the discriminating power and the educational effect due to the selection bias of response which is caused by the high dropout rate. In particular, it is also necessary to consider that multiple responses are allowed in the weekly assessment.

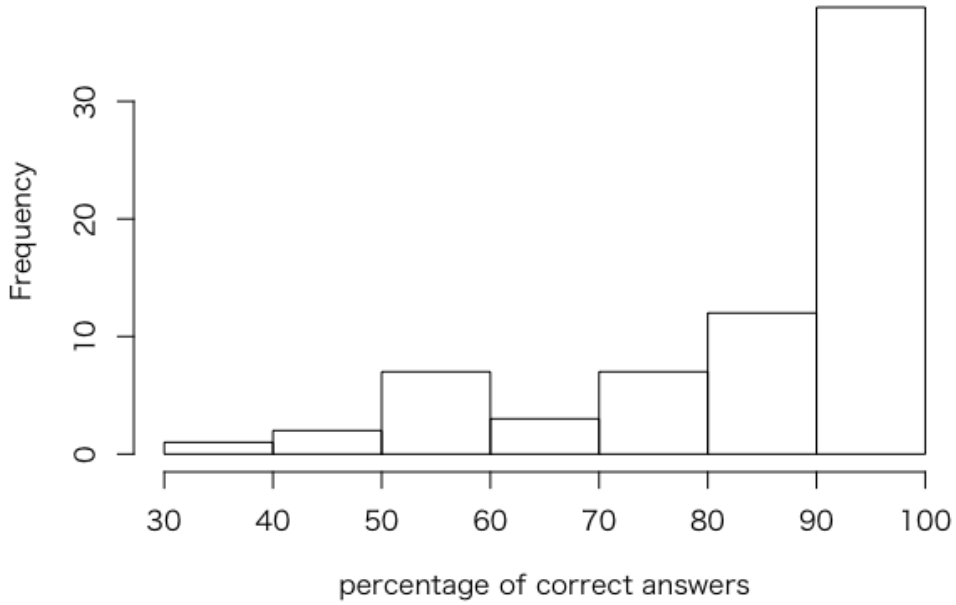

Figure 2. The percentage of correct answers

REFERENCES

Everson, M., Zieffler, A. & Garfield, J. (2008). Implementing new reform guidelines in teaching introductory college statistics courses. *Teaching Statistics*, *30*(3), 66-70.

Gazioglu, S. (2012). Online course design: a statistics course example. *Teaching Statistics*, *35*(2), 98-102.

Hahs-Vaughn, D.L., Acquaye, H., Griffith, M.D., Jo, H., Matthews, K. & Acharya P. (2017). Statistical literacy as a function of online versus hybrid course delivery format for an introductory graduate statistics course. *Journal of Statistics Education*, *25*(3), 112-121.

Jordan, K. (2015). MOOC completion rates: the data. Retrieved February 27, 2018 from http://www.katyjordan.com/MOOCproject.html

Mills, J.D. & Raju D. (2011). Teaching statistics online: A decade's review of the literature about what works. *Journal of Statistics Education*, *19*(2).