Yoshinari INABA & Tetsushi KAWASAKI, Gifu (Japan)

On Points of Attention about Teaching to Make Good Use of Knowledge of Data Analysis – From a survey conducted after a data analysis class in Japan

1. Introduction

In basic statistics courses, the overarching goal is the students' acquisition of statistical literacy, however, we often feel some sort of insufficiency in terms of their actual analysis of data. For example, when analyzing observed data, it is sometimes the case that students figure out a correlation coefficient despite the fact that there are clearly outlier values.

At present, basic descriptive statistics is learned through a method of data analysis that has been in use since 2012 in the high school textbook "Mathematics I" in Japan. However, we do not see examples that include outliers in the data handled in almost all textbooks. We think this is due to the difficulties in formulating questions because of the ambiguity in the treatment of issues such as "At what point do you determine data should be considered as being an outlier?"

In statistical education, it is necessary to learn how to analyze data, and the mere acquisition of knowledge such as the statistical calculation method is insufficient to solve real problems. For example, guidance that is conscious of real-life scenarios that require problem solving is required as follows.

①What kind of data should be collected to support an assertion

②What kind of arrangement should be performed with the actual data, what kind of charts and statistics should be obtained

③What judgment is to be made from the obtained chart and statistics and what kind of argument is to be linked

The authors handled 90-minute lectures on descriptive statistics and basic inference statistics 15 times in semi-annual statistics courses for vocational nursing school students. The number of participants in the course was 40.

The purpose of this research is to find out what kind of things are insufficient when students who have learned basic statistics try to analyze actual data by using learned statistical knowledge. And we would like to point out some points to keep in mind in teaching statistics education based on that. For that purpose, we assumed several scenarios of data analysis and asked them about their processing as six questions, from Q1 to Q6.

As a result, there were problems in several instances, and we found that guidance that is conscious of the actual scene of problem solving based on the actual data is required.

2. Status of questions and answers

We mainly explain questions Q1, Q3, Q4 and Q6. The first question concerned the “treatment of outliers.” The data of Q1 consisted of numbers when two groups of 10 people each ate at the conveyor belt sushi shop and one of two group includes an outlier. The average value is easily affected by outliers, but the median value is hardly affected at all. So for data that includes outliers, it is not appropriate to compare groups by mean value. As a countermeasure method, we assumed the answer “judge by excluding outliers” and “judge by median value.”

According to the answer status of Q1, “judge by excluding outliers” and “judge by median value” were 30% of the total. The most common answer was “judge by taking dispersion or standard deviation.” Although these are not inappropriate, it is obvious that the dispersion is large because of the existence of outliers.

Q3 was the question which asking whether students were aware of the two items (factors) that are the basis of their argument and that they can do the appropriate processing. “Answer the following questions when you want to investigate whether the play-time of mobile games has a negative impact on academic scores. (1) What kind of data should be gathered? (2) What quantities are to be calculated and what kind of diagrams should be used?”

According to answer status, it seemed that students could easily notice the correlation between academic performance and game playtime. And reasonable answers such as scatter diagrams and correlation diagrams accounted for 3/4 of them.

Q4 was a question to see whether students can see the correlation between multiple data correctly. Calculating the correlation coefficient accurately is troublesome in manual calculation, but we thought that correlation could be found by ingenuity such as creating a simple scatter diagram. We indicated five types of data, “Height, 100m run time, Number of double jumps, Grip strength of right hand, Math test scores” and made students judge the relationship between these data. “By looking at these data, which items are likely to be related? Also, there seems to be relevance from the data, but is there anything suspicious as to whether it is relevant?”

According to answer status, about 20% of the students considered the correlation. Only one student used the correlation matrix. There was no one using the scatter chart. Many of the answers were comparisons of the people in higher rank with ones in lower rank, respectively, or the comparison of ranking. In addition, many answers not mentioning the reason were seen. Furthermore, when we look at the correlation coefficient between “test score of mathematics” and “number of double jumps”, we see a strong positive correlation (Correl Coef = 0.84), but it is difficult to consider the relevance of these two things. There was no student who mentioned that.

The final question Q6 was about statistical judgment and the sentence is as follows. “We conducted a certain test using two inspection instruments A and B. In the measurement results using A's inspection equipment, many results were found out of the abnormal values. Since there is a possibility that the inspection equipment of A did not work normally, we would like to make a statistical judgment. Write down what kind of methods you can think of.” There is a statement “I want to make a statistical judgment” in the sentence, however the answer of re-measurement was 1/3. There were only a quarter of students who answered that they used a statistical test.

3 Conclusion and remarks

We will describe some of the things we have noticed from the above survey results. First of all, with respect to descriptive statistics, it is almost impossible to handle “outliers” in typical textbook exercises in regular classes. Therefore it is natural that students are not accustomed to processing “outliers”. However, unless we try to increase the students’ practical experience of dealing with actual data in front of them, they cannot master this kind of realistic handling. The result of Q1 shows such facts.

Next, how should we evaluate the correct answer status of Q3, which is the question concerning what kind of data should be obtained about the relationship between the two things? In the text of the question, there is a phrase “we want to investigate whether the playtime of the mobile game has a bad influence on academic scores.” Nonetheless, contrary to expectations, many students did not know what kind of data to obtain. Again, it can be said that students are unfamiliar with the process of extracting necessary data from scenes of actual problems because it is customary for data to be given in advance in routine exercises.

Also, when looking at the relationship between multiple data, we cannot tell how actual data is distributed based only on the correlation coefficient. When students think about correlation, it is recommended that they should first see a scatterplot, but they tend to rely on numerical calculations simply as a correlation coefficient. There is also a tendency to decide simply that there is a strong relationship because the correlation coefficient is high. The last question Q6 asked for statistical judgment, but the number of students who answered “statistical test” was as small as about 1/4, and many students answered that it should be re-measured. Although re-measurement is certainly a practical method, we think that the student overlooked the phrase “we want to make a statistical judgment”, or it was difficult for the students to imagine scenes that actually use a “test.” In routine problem exercises, it is usual to set the scene of the examination in advance for the exercise of “test”. Students can calculate the test statistics according to the given conditions. It is difficult to notice that “statistical test” can be used in actual problem solving scenes by such exercises alone.

From these several results, we can state some of the points to keep in mind in teaching statistical education. One of them concerns at least “the way data is handled in statistical education” in Japan. Numerous problem exercises aiming at skills to calculate statistics are handled, but it is extremely rare that outlier values are included in the object data. The students are simply unfamiliar with “how to handle data including outliers.” We should handle data that is more suitable for actual problem solving.

Another one of them relates to “asking the relationship of multiple data.” Although there are opportunities to learn scatter plots and correlation coefficients, data is given from the beginning, and in many cases it is a simple task of calculating the correlation coefficient. Students do not handle simple approximate straight lines or approximate curves. In addition, there are few opportunities to see what kind of relationship exists between three or more data. We should also deal with problems that we do not know if there is a relationship between given data.

Another concern relates to “the usefulness of inferential statistics.” There are limited opportunities to learn about inferential statistics in Japan compared to descriptive statistics. In high schools, statistical tests are not dealt with in textbooks, there are no study reference books, and easy to understand textbooks are few. Therefore, it is necessary to make efforts in instruction so that students understand that the technique of statistical inference can be useful in various scenes of daily life.

Finally, as concerns “scenes in which statistics can be utilized”, we think that it is necessary to provide practical scenarios in which learners feel the usefulness of statistical methods such as problem solving in their daily life as well as in routine problem exercises.

References

Sakai, J. & Inaba, Y. (2018) A High School Case Study of Statistical Education Practice after Completion of a Unit on “Data Analysis” – From the viewpoint of fostering the ability to utilize data (in Japanese with English abstract), Proc. Inst of Stat Math, 2018, to appear