Journal of Mathematical Modelling and Application
2014, Vol. 1, No. 9, 14-21
ISSN: 2178-2423

An Example of Statistical Modeling for Count Data Analysis in Secondary Education

Yoshinari Inaba
Ritsumeikan Uji Senior High School, Japan
y-inaba@ujc.ritsumei.ac.jp

Tetsushi Kawasaki
Gifu University, Japan
tetsushi@gifu-u.ac.jp

Abstract
In this paper we report on introductory lessons to the Poisson distribution and on statistical modeling in regards to probability distributions in secondary school in Japan. "Statistical modeling" is the use of data analysis and modeled data to describe, explain and predict real-world phenomena. In statistical modeling, we often see some examples which treat probability distributions as a separate data model. When students analyze some given data, we think that it is important they understand the statistics models to be used more deeply. We also think that the ability to determine "what kind of model applies to what kind of data" is also important from a viewpoint of "statistical literacy." We therefore prepared some lessons about the Poisson distribution for high school students with modeling in mind. After that we instructed some task-based lessons as an example of count data analysis. In statistical analysis of count data, there are various cases. Some will fit the normal distribution, and some will fit the binomial distribution. However it is well known that the case with small integral-value observed data tend to fit the Poisson distribution. By the way, we cannot usually find the Poisson distribution in mathematics textbooks in secondary schools in Japan, but we can find it in many overseas textbooks. By knowing some properties of the Poisson distribution, students can get a typical model in the analysis of various observational data, especially count data. As a result, students were actually verifying the example with which the observed data fit the Poisson distribution. In the viewpoint of statistical modeling, we think that students get the technique of analyzing an actual phenomenon through a model by recognizing some typical statistics models.

Keywords: statistic education, mathematical modeling, Poisson distribution

1. Introduction

In secondary education in Japan, we have been unable to find learning material related to statistics in the government authorized mathematics textbooks except for some select texts from more than ten to twenty years ago. We believe the educational concept called "Yutori" education, which means "relaxed education" brought about this situation. As a result, almost all students complete their high school education without learning statistics in their mathematics classes. We have to say that this fact shows a remarkable lack of balance from a comprehensive viewpoint in regards to the learning of such mathematical concepts as quantity, space and shape, change and relationship and uncertainty as presented in the OECD-PISA.

As a result, it caused the fall in the ranking of students' global mathematics academic ability, which came to be called the "PISA shock." A backlash from this shock was also felt later in Japan. At present, the curriculum of secondary schools has been changed into one that has some contents that

involve mathematical activity and task-based study. Moreover, statistics education itself has become a subject that is considered as required study at high school. On the other hand, the age of the current generation of math teachers is also advancing as a new problem. Under these circumstances, the development of teaching-materials utilizing modeling is highly desired as such an educational policy will gives students the mathematical literacy and statistical literacy they need.

Now, the well-known phrase “Mathematical modeling” is used in the field of mathematics for equations, functions, graphs, statistical tables, and so on, to describe and explain real-world phenomena or to investigate important questions about the observed world. Furthermore, we may say that it is this activity which predicts phenomena or solves problems in the real world through mathematical models. Statistical modeling is a process to make a statistical model from the analyzed and characterized data extracted out of the real world and to assist in the re-analyzing of the real world. In this process, we can expect to apply the so-called “PPDAC cycle” (Problem -> Plan -> Data -> Analysis -> Conclusion) which is one of the cyclic learning activities. On the other hand, we have many ways with which to materialize modeling. We can often see examples that use probability distributions.

Data analysis takes the appearance of data characterized as a part of a collective model. It includes characterization by the typical statistics value in descriptive statistics, or by graphical things such as a regression line, or by a quantitative value such as a correlation coefficient. The data that is given or is observed and accumulated is a numerical or qualitative collection with some properties of the groups. Analyzing the data involves applying a probability distribution to them, which is well known as a mathematical model (statistics model), and it is an attitude which attempts to analyze the features of the data. It is also an attempt to guess and analyze the distinctive features of the source group of data. It is important that students can know or guess what kind of data group fits to what kind of statistics model from such a viewpoint in respect to the of acquisition of statistical literacy. We therefore tried to give a description about a Poisson distribution and tried to make a task-based study about it as an example of an analysis of count-data. Statistics education in Japan is still at a developing stage. We hope that students develop the ability to process and analyze data.

2. Poisson distribution model

2-1 Poisson distribution with count data analysis

In a statistical analysis of count data, there are various methods and cases, and some will fit the normal distribution, and some will fit the binomial distribution. However, it is well known that data consisting of small integer values will tend to fit a Poisson distribution. Here, we are going to introduce a statistics lesson in which we will use a Poisson distribution. Usually, we cannot find a Poisson distribution in the math textbooks of secondary schools in Japan, but we can find it in many overseas texts.

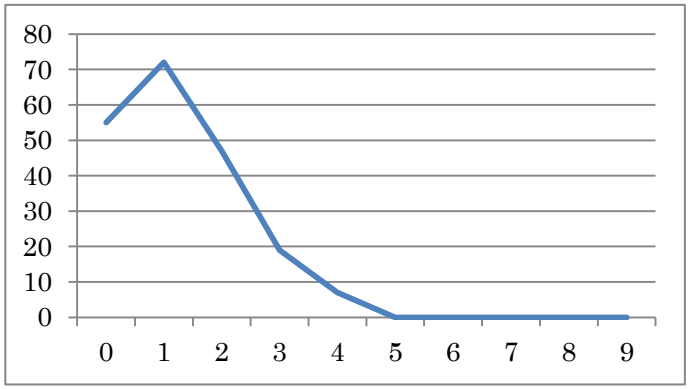
Usually, a calculator with a formula manipulation function, called CAS, is frequently used in overseas statistics courses. We cannot help feeling a difference in the classroom environment of our country which depends for this on non-electric calculation in many cases. We think it is clear that overseas educators feel a familiarity with the Poisson distribution and consider it as suitable teaching material for secondary education level students.

2-2. Introduction to the Poisson distribution through some applicable examples

When introducing an actual Poisson distribution to a classroom, although various methods can be considered, we are not going to introduce it in such a way that it presents a probability function or could said to be a limit of a binomial distribution. Instead, we will introduce it by focusing on its usefulness through some familiar examples of count-data which fit it well. Therefore, at first, we will

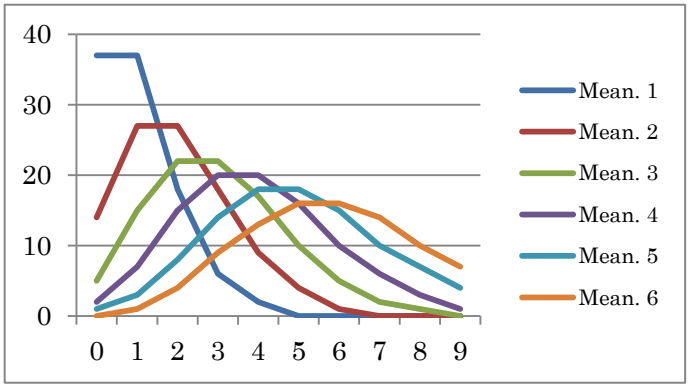
begin with the statistics from games of soccer which high school students can understand whether they are soccer fans or not. If we guess intuitively, it is very rare that a soccer score is in double figures. Probably, in most of the cases, it has a non-negative integral value smaller than 5. Then, we searched all the game results for the year 2011 of the J. League on the Internet and got the scores.

We showed students the table and made them count the number of each score, figure the graph by handwriting, and calculate the mean value of them. As a result, a graph with a bias towards the left (close to 0) was obtained. In fact, students had already experienced the following training in a classroom in advance. Using dice with many faces, such as dodecahedron dice and icosahedron dice, students tossed them many times and got graphs near to the binomial distributions, B (12, 1/12) and B (10, 1/10) respectively. In this case, the mean value is 1 and it results in a graph considerably skewed to the right (long tail on the right) as a binomial distribution. The following graph was obtained from the actual observed values which nearly fit to B (10, 1/10), where each value should be divided by 200 on Graph.1.



Graph.1 The number of times the number ten or twenty appeared on the icosahedron dice when tossed 200 times (students' experiment value)

Next, we started with an explanation of the Poisson distribution. However, it is difficult to explain the graph from the formula of a probability function because most Japanese high school students do not know a natural logarithm. In such a case, it is common to compare the graph which is composed by a spreadsheet software program and the graph of the Poisson distribution. Still, this way was too simplistic, and was not suitable, so we considered another way. First, we prepared some graphs of a Poisson distribution which have an integer mean value by using a spreadsheet on a computer (Graph.2).



Graph.2 Poisson distributions with mean (parameter) is 1 to 6.

We handed out such a graph to students and explained that these are curves by the name of “Poisson distribution.” Next, by using the graph, we made students guess the distribution curve with a mean value of 1.4 which was previously calculated. Of course, this is a something of a crude method. However, since they can trace between the two curves with a suitable estimate, it presented no great difficulty to them. In other words, it can be called a work which finds an approximated curve by interpolation (Fig.1). Furthermore, the curve can be compared with the handwritten graph about the soccer scores the students made. The two will be unexpectedly close. Of course, a re-comparison is carried out by the spreadsheet to verify that later. The reaction of students in such a scene is frank, and they are unabashedly impressed. Students commented, “I was surprised that the two curves are so mutually close,” and “Does the baseball score give us the same result as the soccer score does?” and “For what can I use this result ?” etc.

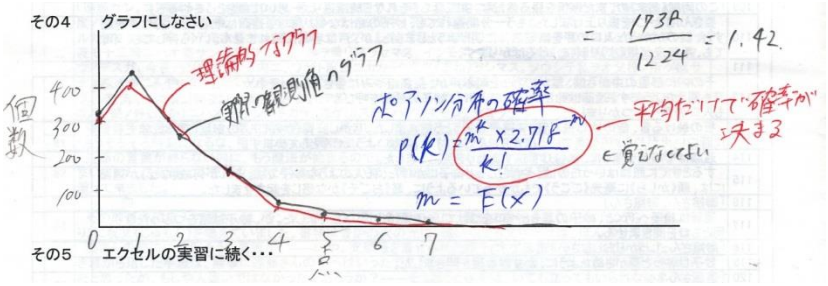
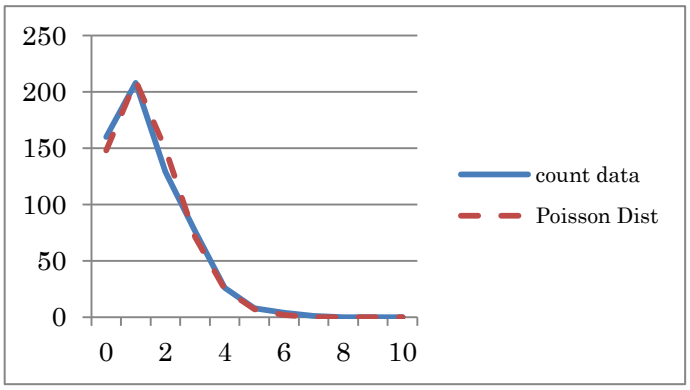


Fig.1 From a students’ notebook

Now, one of the greatest features of a Poisson distribution is that a probability function is a one-parameter function of only mean value. That is, a graph will be obtained only if the mean value of the observed data is given. Although we have used a method of introduction at the point of the count data analysis, we think that a Poisson distribution should be dealt with more extensively in secondary education in Japan than it is now.

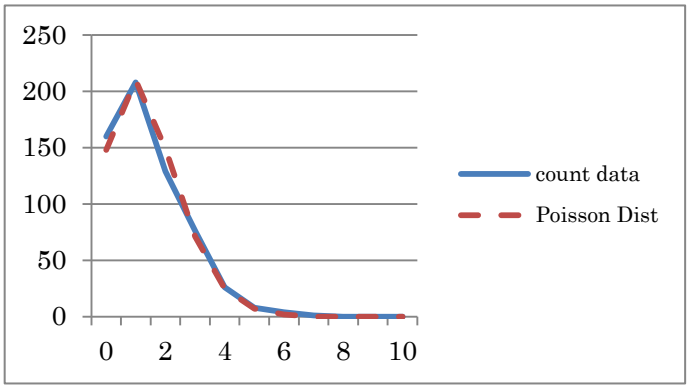


Graph.3 Graph of score record of 2011 J. League, and the Poisson distribution of mean value 1.42

2-3. Other examples in which we can easily obtain data about the Poisson distribution

We handed out a data sheet in which the number of times earthquakes occurred per day from May 1st, 2012 to October 31st 2012 in Fukushima Prefecture where there had been a series of aftershocks after the Tohoku earthquake was recorded. These data also fit well to a Poisson distribution. As another example, we showed the number of specific characters which appear in the text of a short novel, the number of the traffic accidents per day in Kanagawa Prefecture, and the number of runs in Japanese professional baseball games. We also mention it is important that all of that data can be found on the web. In such illustrations, students felt that data with a low frequency

and small mean value were suited to a Poisson distribution.



Graph.4 The graph of earthquakes which were observed in Fukushima Prefecture and the Poisson distribution with mean value 1.75

When we apply a statistical distribution to an actual phenomenon, it is necessary to get to know the character of statistical distribution and to know what kind of phenomenon can be explained. This is indispensable when we are conscious of modeling on data analysis.

2-4. Example of task-based class-work supposing a Poisson distribution

We regarded a Poisson distribution as one of the models of rare count data, and actually assigned students a data observation as a home task as follows.

Contents of Assignment

Subject name : Modeling of count data by a Poisson distribution

Aim: A Poisson distribution is treated as one of the tools in statistical modeling. It can be considered as an example which allows secondary school students to experience the example of modeling. Students can consider events which do not happen very frequently, and students can characterize it by a Poisson distribution.

Contents of a subject :

(1) Decide an observation theme.

(2) Consider and plan the method of data observation.

(3) Observe them.

(4) Calculate the mean value and distribution of observed data, and figure graph.

(5) Compare the Poisson distribution with graph of observed data.

Deadline : After a week

Evaluation: Give standard evaluation, if the above-mentioned directions are observed. What is especially rich in originality gets added points. Those with a defect in observational data, and handed in late after the deadline had points taken off..

Supplementary directions :

・The number of data to observe is 50 or more and 75 or less.

・The web is available for hint.

・When the survey is difficult, use of the data on the web is possible.

・Results contrary to one's anticipation are also acceptable.

From a point of view of statistical modeling, each process (1),(2),(3) and (4) imply "Problem," "Plan," "Data," "Analysis / Conclusion" respectively.

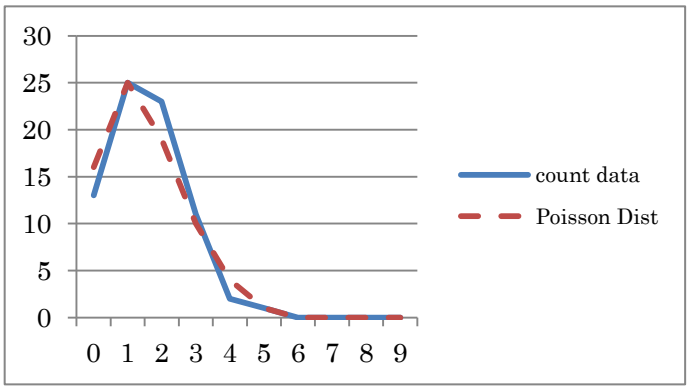
2-5. Result of task-based class-work

Almost all of the students submitted their own data within one week. They gathered or observed the data and regarded it as being close to the image of the Poisson distribution. Some of the themes of the data which the students actually observed are as follows:

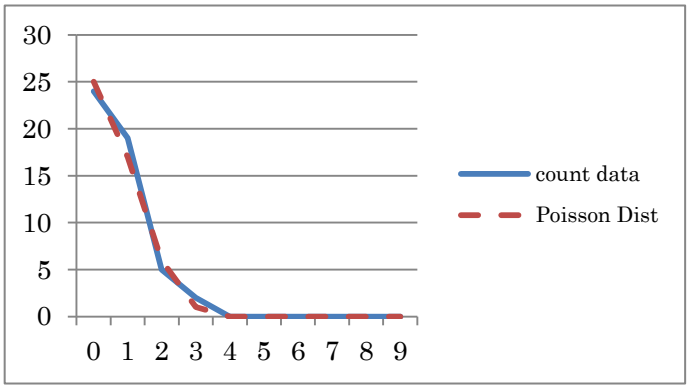
Theme lists which students set up

・The number of the Chinese characters which have 15 or more stroke counts in a page of the Kanji Test book.
・The number of hairs left on a comb after each combing.
・The number of times that a beginner tennis player put the tennis ball into the service box continuously.
・The number of the home runs which Yomiuri Giants (Japanese pro baseball team) recorded in a game from a season opener to 75th games in 2012.
・The number of missed shots per every 10 shots during tennis practice.
・The number of the single portraits on a page of the school's album.
・The number of times in which a person wearing glasses passed the ticket gate machine of the station in Kyoto per 5 seconds.
・The number of stray cats seen per hour.
・The number of the vegetables seen per dish.
・The number of the typhoons that caused a death which was observed in 1961-2011 at Japan Meteorological Agency.
・The number of sneezes per hour.
・The number of mails which received at Hotmail per day.
・The number of capital letters in a German sentence.
・The number of red pens which are picked out of a pen-case filled with many colored pens.
・The number of times the letter "R" appeared in the title of a foreign book.
・The score of the losers in a high-school baseball autumn Nara tournament and Kinki tournament.
・The number of cars which passed through an intersection near their house every 10 seconds.

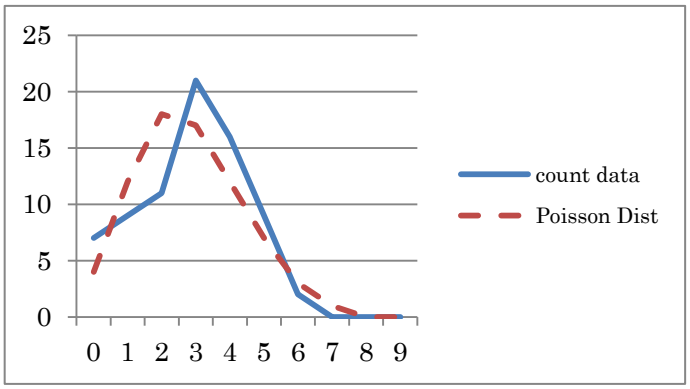
These show that students collected data with good sense. We think that it shows that students can do imagining with the Poisson distribution model. Students said it was difficult to determine the theme and that it was difficult to actually observe data although the theme was predetermined. They also said it was uncertain that data fit the Poisson distribution, etc in the next lesson. After that, students verified the result visually by using a PC. We show some observed data and Poisson distribution they obtained below.



Graph.5 The number of sneezes per hour.



Graph.6 The number of times that a beginner tennis player put the tennis ball into the service box continuously.



Graph.7 The number of capital letters in a German sentence.

2-6. Evaluation by chi squared test

After several hours of task-based class-work, we got an opportunity to give a brief explanation about the chi squared test. Then, we took some themes as an example of a statistical test. Here, the observed data and theoretical values of theme "The number of sneezes per hour" are as follows.

Frequency	0	1	2	3	4	5	6	
Observed val.	13	25	23	11	2	1	0	
Theoretical val.	15.8	24.6	19.2	10.0	3.9	1.2	0.3	
								χ^2 value
	0.50	0.01	0.75	0.10	1.07			2.43

χ^2 value: chi squared value becomes 2.43 (at frequency 4 – 6, values are integrated). On the other hand, the 5% point of upper probability in chi squared distribution with 3 degrees of freedom is 7.81. Therefore, we cannot reject the null hypothesis "The data do not fit a Poisson distribution."

Also for example, the observed data and theoretical values of theme "The number of the capital letters in a German sentence." are as follows.

Frequency	0	1	2	3	4	5	6	7
Observed val.	7	9	11	21	16	9	2	0
Theoretical val.	4.3	12.2	17.5	16.8	12.0	6.9	3.3	1.3

							χ^2 value
	1.75	0.85	2.43	1.08	1.33	0.02	7.46

In this case, χ^2 value becomes 7.46 (at frequency 5 – 7, values are integrated). On the other hand, the 5% point of upper probability in chi squared distribution with 4 degrees of freedom is 9.49. Therefore, we cannot reject the null hypothesis as well.

3. Conclusions and Remarks

As we mentioned above, a Poisson distribution was introduced to students. We think that we were able to make them feel the merit of fitting of a Poisson distribution to actual rare data. And after that we implemented a modeling which involves theme setting and data collection. As students have actually found the examples of count data, we can also often see them in our life apart from the textbook. Therefore, we think that it is important to show a Poisson distribution in addition to a normal distribution and a binomial distribution. Furthermore, from the viewpoint of modeling, we think that it can make students have an intuitive image of a Poisson distribution from rare data. On the other hand, since the high school students do not have appropriate standards to evaluate, it is regrettable that they cannot help evaluating intuitively on a graph. But, from the results: chi squared test above, the students have very good sense to the model. In the stage of studying the foundation of statistics, we think that an intuitive understanding of this level is enough and that if students need more rigorous theory they should reinforce it in the next academic stage. In statistical modeling, students can select a suitable statistics model in which they can feel an affinity in their image by recognizing some statistics model beforehand. From these kinds of reasons, it is especially important that Poisson distribution should be treated in secondary education. And in order that students may get statistical literacy, we think that statistical modeling should be utilized more.

4. Bibliography

[1] Kawasaki,T. Inaba,Y. Kihira,T. Maesako,T.(2013). A Practical Study of Statistical Modeling in Secondary Education [in Japanese with English abstract], *Annals of Educational Studies Osaka University, 18*, (pp.3-19), Osaka, Japan: Osaka University.

[2] Kubo,T. (2012). *Data Kaiseki no Tameno Tokei Modeling Nyumon (An Introduction to Statistic Modeling for Data analysis)* [in Japanese], Tokyo, Japan: Iwanami Shoten, Publishers.

[3] Watanabe,M. (2007). A new framework of statistics education - What the new curriculum guidelines imply in Mathematics [in Japanese with English abstract]-, *Japan journal of mathematics education and related fields, 48(3.4),* (pp.39-51), Mathematics Education Society of Japan.