CHALLENGES FOR LEARNING ABOUT DISTRIBUTIONS IN COURSES FOR FUTURE MATHEMATICS TEACHERS

Marcos N. Magalhães
University of Sao Paulo, Brazil
marcos@ime.usp.br

Knowledge about distribution is important and necessary for the process of learning other statistical concepts, in particular, formal or informal inference. The differences between empirical and theoretical distributions are sometimes hard to understand for students of all levels. At the University of Sao Paulo, Brazil, future mathematics teachers attend two basic statistics courses as part of their curriculum structure. In Brazilian middle and high schools, these are the professionals responsible for teaching the statistical concepts included in the contents of the mathematics discipline. Therefore, good knowledge of the concept of distribution will improve their teaching. In this paper, we focus on the misconceptions observed in basic Statistics courses attended by future mathematics teachers. We also present suggestions for activities that include the use of graphical and computer tools to improve the learning process.

INTRODUCTION

Basic Statistics concepts are taught in Brazilian middle and high schools by mathematics teachers that, usually, have a degree in Mathematics Education or in Physics and Science. In this paper, we discuss the knowledge about distributions, in particular, the distinction between empirical (frequency) and theoretical (model) distributions, in a first basic statistics course, attended by students in Mathematics Education career of the Institute of Mathematics and Statistics of the University of Sao Paulo, Brazil. These students have two semesters of statistics in their curricula, and they study combinatorial analysis, descriptive statistics, probability and discrete random variables. Generally speaking, the objectives of these courses aim at developing statistics literacy and statistics reasoning as indicated in delMas (2002).

The University of Sao Paulo is a public university and the admission is based on a score obtained in exams arranged for this purpose, covering high school content. There are different levels of competition regarding the careers and the education majors are far from the top ones. As indicated in Gatti et al (2009), the teacher career does not arouse any great interest from young people in Brazil. If there is no change in this scenario, Brazil will have big problems to fill the teacher positions in the near future.

In this context, despite their major being in the area, the Mathematics Education students have typically low levels of both general and mathematical knowledge, reflecting their previous school work standards. This poor background affects their learning processes, particularly in the junior year at university. Furthermore, we observed that several students have difficulties with reading comprehension of statistical problems, as well as with studying methods.

Garfield and Ben-Zvi (2008) discuss in their book several empirical and theoretical distributions issues. They reinforce the importance of understanding that the distinction underlying empirical and theoretical distributions is related to variation.

Another important reference to mention is Reading and Canada (2011). The authors provide an extensive discussion about the knowledge of distribution, reporting research studies related with teacher learning, both, before and while teaching. They mentioned that distribution is a key concept that depends on and is depended on many other statistical concepts.

Batanero, Tauber and Sanchez (2004) present results from research on students´ reasoning about Normal distribution in a university-level introductory course. The authors mention students´ difficulties that were not exclusive in the study of Normal model, among them, the difficulty to distinguish between empirical and theoretical distributions.

CONCEPTUAL DIFFICULTIES ON DISTRIBUTION

We have taught statistics basic courses for several years and the discussion in this paper reflects this experience. Comments on specific activities are related to a daytime course attended by future math teachers in 2013. At the beginning of this semester, seventy five students were

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

enrolled. However, despite much effort, the dropout rate is high and, at the end of semester, the class attendance was nearer fifty students. This has been a difficult challenge to overcome.

We encourage student participation and traditional lectures are combined with group work including exercises and debates about topics on the course. The final grade is based on exams, exercises and projects. Except for the exams, most of the work was done in groups.

Combinatorial analysis is the first topic of the course and takes around three weeks. Then we introduce data sets and, subsequently, tables and empirical distributions. We present several types of graph such as dot and bar plots, histograms and box plots. At that point in the course, the only measures discussed are median, quantile and the range of values. In the sequence we go on to discuss probability concepts, random variables and the discrete models: Uniform, Bernoulli and Binomial (we mention other models briefly). After that, we present central tendency measures as average, median (again) and mode and dispersion measures as variance and mean deviation. The semester is closed with the study of joint random variables. We follow this sequence in order to stress the graphical properties of the variables avoiding the tendency to resume the study of data in obtaining mean and variance. We have been used this approach with good results in the reasoning about distributions. However, difficulties still remain and the present research is part of the continuous effort to critically evaluate our teaching process.

We list below three misconceptions on distributions that frequently appeared and, in the next section, we discuss how some proposed activities helped in overcoming these errors.

- *Misconception 1: A random variable is completely unpredictable*
  The students associate the word random with wild or without control. In this case, the random variable would produce any value, without limit, range or pattern and the (wrong) notion of variability would indicate the variable does not have any control. It is possible to imagine variables behaving like that, however this is not necessary to be random variable. What these students must consider is the variability connected with the probability, that is, different values with different probabilities.
- *Misconception 2: In a random variable, all of its values have equal probability*
  When we introduce the idea of a random variable, it is natural to discuss simple examples with balanced coins and dice. Also, when using a random sample with replacement, all elements of the data set will have equal probability to be chosen. Some students do not move forward from these initial settings and assume all models have equal probability.
- *Misconception 3: There is no distinction between theoretical and empirical distributions*
  Some students have difficulties identifying the differences between theoretical and frequency distributions. They do not use probability to compute measures as mean and variance and they felt the need to create a hypothetical frequency table with 100 observations in total. This indicated a reasoning limitation in their understanding of theoretical distributions and it would affect the future study of continuous models. There were also students that had difficulties understanding that a frequency distribution of a variable does not, necessarily, indicate similar behavior to this variable in the population.

ACTIVITIES

Alongside the successive editions of the course, activities were created, modified or suppressed according to the results obtained. They have been used to reinforce concepts, trying to avoid misconceptions such as the ones mentioned above. In the following sequence we present some of the activities connected with knowledge on distributions and we commented how each activity aids in overcoming conceptual difficulties reflected in the above mentioned errors. They were presented in chronological order, as they were implemented in the 2013 edition of the course.

Project 1: Data Analysis

The objective of this activity was to use statistics tools in real contexts, that is, to create contexts so that students could experience working with research questions based on a real data set. We asked students to organize themselves in groups of 3 to 6, choosing a theme that they wanted to work on. The data set to be used was originally collected with students of the discipline in the years of 2012 and 2013, through questionnaires which variables were, supposedly, of the students' interests. The group wrote a report and presented the results to the class.

When we proposed this project, the content covered in the course did not included mean, mode or variance. The group was supposed to use frequency tables, graphs as histogram and box plot, and quantile. Our intention was to reinforce that good analysis can also be made using "simple tools", as graphs and tables, and not only with the usual summary of computed measures.

With this activity the students realized that the frequency table changed, depending on the sub-population considered. Therefore, it served as motivation and preparation for the upcoming presentation of the concept of random variable with different models according to the established conditions. The activity helped to avoid Misconceptions 1 and 2, respectively, a random variable to be completely unpredictable and to have all values with equal probability.

How to Build a Model?

We encouraged a class discussion asking for suggestion to a model for the number of children in the Brazilian families. There were several suggestions and a lively discussion on what would be the upper limit for the number of children in the family. After some time, someone suggested considering a specific geographic region since there are big economical differences in the country. We finished the lecture asking for an internet search in the site of IBGE (Brazilian Institute of Geography and Statistics) to establish models for a few regions.

The activity, a kind of informal inference, gave us the opportunity to discuss range, variability, and other issues necessary to set a model distribution. Besides that the internet search was important to connect the real data with the building of models. All the cited misconceptions were explicitly discussed. In particular, the choice of different probabilities for different values of the variable helps avoid the Misconception 2 (equal probabilities for all random variable values).

Fitting Theoretical Distribution to Data Set

We discussed in class an example from the textbook of the course, Magalhães and Lima (2013), about fitting models. A data set was presented with 100 observations of the number of newborn pigs born alive, from an in vitro insemination process. With a few simplifying hypotheses, we asked if the Binomial model with parameters n=10 and p=0.5 could be adequate to the data. Values 0 and 10 were not presented in the data set but they appeared in the theoretical model and the apparent contradiction provoked a discussion that it is particularly useful to avoid Misconception 3 (no distinction between theoretical and empirical distributions). A graph was used to compare observed and expected frequencies. We could also compute a measure of the distance between these frequencies (like chi-square). The activity was complemented by some other tasks, simulating models and fitting data. In general, students improved their knowledge about distribution after this activity.

Project 2: Didactic Material

The organization of the class - working in groups- was similar to Project 1. However, in Project 2, students were supposed to prepare a practical activity, and a respective support material to be used in middle or high school classes. They could choose the subject among the topics discussed in class, and the activity could be created by the group or adapted from an existing one. They prepared a report and a poster presentation for the class.

The activity had several benefits and it was not intended exclusively to learning on distribution. There was 10 groups and the topics choice indicated what students felt comfortable to work with when preparing the didactic material and, in the future, when teaching classes.

With respect to Misconceptions 1, 2 and 3, the activity had different effects depending on the topic developed. Two groups proposed a data collection and this was an additional opportunity for reflection on theoretical and frequency distributions (relate to Misconception 3). The other eight groups proposed some kind of game, which was useful to avoid Misconceptions 1 and 2, since they need to study the experiment in order to assign probabilities for the different outcomes.

RESULTS

The concept of distribution is essential in most of the statistical ideas discussed in class. In this way, the course overall results were affected by the knowledge about distribution. In this section, we present quantitative results based on students' final performance and comparisons in similar situations. A more detailed discussion, considering qualitative and quantitative results, is presented in Magalhães and Magalhães (in press).

The same question about distribution was proposed in 2011 and 2013: *Are there differences between random and empirical variables?* In 2011, the students in pairs answered the question in a quick class quiz; in 2013, individual students took an exam. The results are shown in Table 1. There was a slight improvement in the 2013 course edition.

Table 1. Comparative results on a question about distributions

Year	Number	Not satisfactory	Partially satisfactory	Satisfactory
2011	41	32%	46%	23%
2013	34	24%	36%	40%

Table 2 presents some figures related to the whole course in three years, 2011 to 2013. In those years, we are assuming that we have the same level of difficulty in the assessments and approximately the same criteria to attribute grades. Also, data available from the Admission Office of the university indicates there is not a clear difference in the students´ background in these years. To pass, students must reach a final grade of 5 or over in a scale of 0 to 10. The results indicate small changes from year to year. The dropout rate is relatively high, and the pass rate is moderate even after the exclusion of the students that dropped the course. Related to final grade means, it is worth to mention that these numbers are typical in mathematical courses at University of Sao Paulo.

Table 2. Comparative results on different editions of the Statistics course

Year	Enrolled	Dropout Rate	Pass Rate*	Final grade mean (sd) all students	Final grade mean (sd) passing students
2011	67	28%	77%	5.3 (1.4)	5.9 (0.8)
2012	66	32%	56%	4.6 (1.5)	5.6 (0.8)
2013	75	28%	65%	5.0 (1.6)	5.9 (1.0)

(*) The percentage was computed excluding the students that dropped the course.

As an additional analysis, in March 2014, we asked the students who succeeded the 2013 course to answer a test with 20 multiple choice items. The test involved concepts of the two statistics introductory courses from the previous year with 14 items predominantly related to the first course. Participation was voluntary and anonymous, and we got 22 responses. The same test was answered in 2010, and in 2011, for students which were completing the Education Math degree in these years. They formed a group of 38 students (named *Other years*), and it is estimated that 50% of this group have also attended additional elective courses in the area of statistics. At the moment they did the test, they may be better prepared in statistics than the 2013 students. Note that the attendance to both tests does not come from a random sample or a population census, so conclusions are limited to the group tested.

Figure 1 presents the box plot with the number of correct answers (score) for the two groups (2013 and Other years). The results are equal for mean and median, but the 2013 group is more homogeneous. The mean score would correspond to a final grade of 6.5 (13 over 20) which is higher than the final grade in the first statistics course (5.9) attended by the 2013 group. The relatively low dispersion in the 2013 group, when compared to other years, is not too different from the result presented in Table 2. We have standard deviation 0.9 (in 0 to 10 scale) against 1.0 in Table 2. It is important to mention that, small values of standard deviation could reflect more interaction and participation of the students in the classroom learning process. If the tendency to attend additional statistics courses remains among the 2013 students, one would expect they will have better conceptual formation when they leave the university.

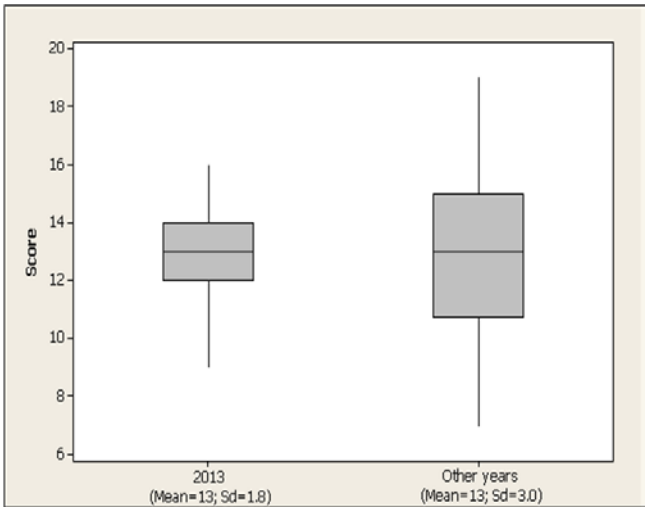

Figure 1. Number of correct answers (score) in a test with 20 items

In a complementary view of the test, we present in Figure 2, the proportion of the correct answers by items. We observe better performance of the 2013 students in 11 items. Also, we have only four items (5, 13, 14 and 18) with expressively worse results in 2013 group when compared to the other group. A more complete analysis of the differences between the two groups is underway and it will be reported elsewhere.

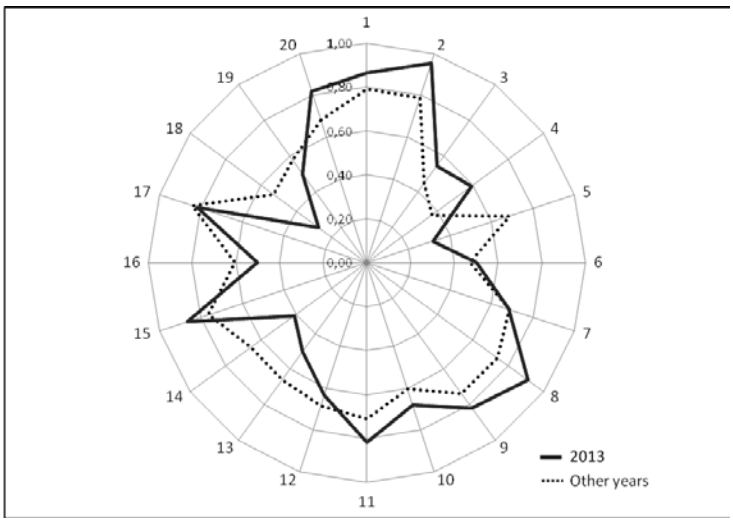

Figure 2. Proportion of correct answers by items

CONCLUSION

In the last few years we have been teaching initial statistics courses for future math teachers, knowing that the concepts related to distribution are critical to teachers preparing to teach at middle and high school levels. Mainly when we teach in large classes, the learning of the students depends on their previous background. As we see in Table 2, there are undesirable numbers to overcome. The challenge to enhance the course remains, but we feel that the activities implemented in 2013 could improve the outcomes of the course. The approach used in this course, to increase students' participation through activities, could be applied in introductory statistics courses anywhere, and particularly to future mathematics teachers.

AKNOWLEDGEMENTS

We would to thank the partial support from FAPESP (São Paulo Research Foundation), grant #2014/05160-9.

REFERENCES

Batanero, C., Tauber, L. M., & Sanchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning and thinking* (pp. 257-276). Dordrecht, The Netherlands: Kluwer.

delMas, R. C. (2002). Statistical literacy, reasoning, and learning. *Journal of Statistics Education 10*(3).

Gatti, B. A., Tartuce, G. L. B. L., Nunes, M. M. R., & Almeida, P. C. A. (2009). *Atratividade da Carreira Docente- Relatório Final*. São Paulo, Fundação Carlos Chagas (in portuguese).

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York: Springer.

Magalhães, M. N., & Pedroso de Lima, A. C. (2013). *Noções de Probabilidade e Estatística* (7th ed., 2nd reprint- reviewed). São Paulo: Edusp (in portuguese).

Magalhães, M. N., & Magalhães, M. C. C. (in press). A critical understanding and transformation of an initial statistics course. To appear in *Statistics Education Research Journal*.

Reading, C., & Canada, D. (2011). Teachers' knowledge of distribution. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching statistics in school mathematics- Challenges for teaching and teacher education. A joint ICMI/IASE study: The 18th ICMI Study* (pp. 223-234). New York: Springer.