INTRODUCING THE NORMAL DISTRIBUTION IN A DATA ANALYSIS COURSE: SPECIFIC MEANING CONTRIBUTED BY THE USE OF COMPUTERS ®

Liliana Tauber
Universidad Nacional del Litoral
Argentina
Victoria Sánchez
Universidad de Sevilla
España

In this work, we describe the elements of meaning related to normal distribution, which appear in a data analysis course based on the use of computers. The course was directed to students in their first year of university studies. We study the elements implemented in a teaching unit for the normal distribution in which computers were introduced as a didactic tool. We pay special attention to the specific meaning conveyed by the use of computers as well as to the meaning attributed by the students throughout the teaching sequence.

INTRODUCTION

One main problem in an introductory statistics course at university level is making the transition from data analysis to inference (Moore, 1997). The scarce time available and the students' poor previous knowledge make a complete study of probability difficult. It is necessary, however, to provide the students with an intuitive knowledge about probability distributions, since the difficulties in understanding these distributions might lead to errors in the application of inferential procedures, such as interval estimation or hypothesis testing. In this paper, we will focus on the normal distribution.

The relevance of the normal distribution in statistics has been highlighted by different authors, who remark on its role as a link between descriptive and inferential statistics. Cohen and Chechile (1997) also suggested that the normal distribution is an important topic, even in an introductory course and that statistical software can be used as a tool to make data analysis more accessible and to facilitate the teaching of this distribution.

THEORETICAL FRAMEWORK

A basic epistemological assumption in our theoretical perspective (Godino, 1999) is that mathematical objects (for example, the normal distribution) emerge from problem solving activity mediated by the semiotic tools that are available in specific institutional contexts. The meaning of mathematical objects is then conceived as the system of practices related to specific problem fields and five different types of elements are distinguished:

1. *Extensive elements*: The situations and problems from which the object emerges.
2. *Ostensive elements*: The semiotic tools available to represent or to operate with the problems and objects involved.
3. *Actuative elements*: Procedures and strategies to solve the problems.
4. *Intensive elements*: Characteristic properties and relations to other object such as definitions and theorems.
5. *Validative elements*: Arguments that serve to justify or to validate the solutions.

Based on this semiotic-cognitive perspective (Godino, 1999) and before implementing the teaching sequence, we analyzed some textbooks directed to Human and Social Sciences students, to identify the elements of meaning that commonly appear in the study of normal distributions (Batanero, Tauber, & Meyer, 1999; Tauber, Batanero, & Sánchez, 2000). This was considered the *institutional meaning of reference*.

From the results of this analysis, we selected the elements of meaning that were better suited to our didactical approach. We also added those elements that were specific from introducing the computer in the teaching (more detail will be given later). Once these elements were determined, we designed activities to be solved with the computer and other traditional activities, and organized a teaching sequence with them.

A content analysis of this material was made, with detail of what elements of meaning were included in each session, to obtain the *predicted institutional meaning* and finally the course was carried out. In the *sessions developed in the traditional classroom,* problems aimed towards the students' discovering of some properties of normal distributions were proposed. In addition, we organized and oriented the solutions debate. Interpretative activities and applications to real problems were emphasized. These activities were solved with paper and pencil, calculators and drawing tools.

In the *computer lab sessions,* we provided the students with data files, where they could fit the normal distribution to some of the variables and where this was not possible with other variables. For each session, the students produced a word-processed report, where they included those data analysis results that they considered necessary to answer the questions. Wide opportunities to exercise their argumentative capacity were given.

The development of both types of sessions was based on a written material that was given to the students, where the normal distribution and the software "Statgraphics" were introduced. From the content analysis of recorded observation of the sessions, we determined the *observed local institutional meaning*. We described the meaning elements and the relations among them that were put into play in each session. This allowed us to compare with the *predicted institutional meaning*, and to analyze what problems arose. To evaluate the students' *personal meaning* we analyzed their working documents, from which the diverse elements of meaning they had used were categorized. Introducing the computer led to the introduction of new elements of meaning as regards to what is included in the majority of textbooks. According to our aims, we detail these specific elements in the next section.

DEVELOPMENT OF THE STUDY

Analyzing the meaning of normal distributions acquired by the students in a specific teaching experiment was the general aim in our research. In particular, we were focused on the specific meaning induced by the computer. Below we describe the institutional context where this study was developed, as well as the data collection process.

Institutional Context

The teaching experiment was developed within an optional data analysis course, 90 hours long, of which about 12 hours were devoted to the topic. 60 students followed this course, which was carried out at the University of Granada (Spain) in the academic year 1999-2000. Most of the students came from Education, Business, Psychology and Engineering majors. Their statistical knowledge was very varied, though prior to the course, they had never been introduced to the use of software or to the practical aspects of statistics. Before studying the normal distribution, the students worked with the software along the introduction of previous themes. Half the sessions were carried out in a traditional classroom and the remaining in a computer lab.

Data Collection and Analysis

Data were taken from the researcher's observations and from the documents produced by the students for each session. The development of all the sessions was registered. The observation followed a previously designed protocol to take into account the specific points of the written material given to the students, the activities and the development that were planned for the given session, in the a priori analysis of the teaching sequence. In addition, we collected the students' written productions in relation to the tasks proposed.

In the analysis of the observation records, we selected those paragraphs in which meaning elements or its relationships were clearly shown, as well as the most outstanding interactions between the students and the lecturer. This served to identify the basic elements of meaning in the implemented teaching sequence and to interpret the students' written production, where correct and incorrect application of the meaning elements was identified, with discrimination of the elements specific of the normal distribution and those related to other concepts. Frequency tables of these elements were produced and examples were given to clarify the different categories. These frequencies served to make a qualitative analysis of the student's application of meaning elements. In the next section, we describe the different types of meaning considered in our research, which served to characterize the teaching experiment.

PREDICTED INSTITUTIONAL MEANING

Following the categorization of elements in our theoretical frame, we have considered the following elements of meaning in our didactical sequence:

Extensive elements: Problem fields and contexts. We considered the following problem fields that lead to the normal distribution:

P1. Fitting a curve to frequency histograms or polygons for an empirical data distribution, as an approximated theoretical model in fields such as Psychology, Biometry, or Theory of errors.

P2. Approximation of discrete distributions in variables with many different values; for example in the binomial distribution for a high value of the n parameter.

These problems were proposed in diverse contexts in agreement with the students' interest. Working with data files served to create open-ended problems and to introduce a multivariate and exploratory philosophy in the analysis of data. In this way, we were able to work with rich tasks, where diverse elements of meaning were integrated, and to establish more complex relations among these elements.

Ostensive elements: Representations. The following types of representations of the abstract objects were used, with a double symbolic and instrumental function:

Graphical and numerical representations in traditional support. Frequency histograms, polygons, density trace, box-plot and stem and leaf display, representation of tail areas in the normal curve, representation of the central intervals in a normal curve, frequency tabulation and statistical summaries of central tendency, dispersion and shape.

Verbal and symbolic representations of normal distributions in traditional support. Words such as normal, statistics, parameter, density function, symbolic representations, etc.

Representations that are specific to the computer. The use of software offered a great variety of representations, both numerical and graphical, that can simultaneously appear on the computer screen. In addition to the usual representations previously described, it was possible to represent several density functions or the frequency histogram superimposed on the density trace in the same reference framework. Numerical analyses based on the study of the critical values and tails areas for any normal distribution were made, whereas in textbooks they are only offered for the standard normal distribution. This variety of representations facilitates the data analysis, although it involves a higher semiotic complexity.

Actuative elements: Specific techniques to solve problems. The computer served to introduce the following types of strategies and procedures to solve the above problems:

A1. *Descriptive study*. The computer was used to produce various numerical and graphical analyses with the purpose of determining the goodness of fit between the frequency histogram or polygon and the density curve, and to decide if the given variable could be well approximated by a normal distribution.

A2. *Standardizing*. This was only used to compare distributions with different means and standard deviations.

A3. *Computing probabilities and critical values*. The computer served to introduce two computation procedures:

1. Given one or two variable values, computing the probability that the variable falls in the interval determined by that or those points (tail areas).
2. Given a probability, finding the limits of the interval including this probability (critical values computation).

A4. *Visual comparison*. This is an important element contributed by the computer. The frequency histogram or polygon shape can be visually compared to the density curve and, in this way the empirical distribution the goodness of fit to the normal one can be graphically assessed.

Other practices were: *changing the numbers of intervals in the histogram, changing the parameters, and computing the limits of central intervals that include a given percentage of cases.*

Intensive elements: Definition, properties and relation to other concepts. The computer was used to introduce the normal distribution by means of simulation, using a data set referring to the intelligence coefficient for a group of students. The normal distribution was presented as a model that approaches the relative frequency polygon when the sample size is increased and the width of intervals is diminished. This served to study the following properties:

- *Symmetry and kurtosis*: relative position of mean, median and mode, interpretation of asymmetry and kurtosis coefficients, area below and on the mean, central intervals probability.
- *Properties related to the normal distribution parameters*: relation between the standard deviation and the curve inflexion points, geometric meaning of parameters, variation of the curve of density with variation of parameters.
- *Statistical properties*: total probability under the curve, property of central intervals.
- *Relation with other concepts*: statistical and random variable, empirical and theoretical distribution, measures of central position, dispersion, symmetry, etc.

Validative elements: Types of proofs and arguments. In presenting the subject we avoided excessive formalization, using some types of validation that are not computer specific, such as:

- *Verifying some cases*. Particular cases were used to verify some properties.
- *Generalization*: With formal or informal arguments the students reached general conclusions that extended particular initial cases.
- *Analysis*: Discovering the peculiarity or initial features in a situation that can later lead to generalization or synthesis.
- *Synthesis*: When a conclusion is made from all the properties or conclusions drawn in the analysis phase or where several elements of meaning are included.
- *Validation by use of graphical representation and simulation*. The computer served to produce validations based on graphical representations of the empirical distributions and on simulations of fitting normal distributions.

PERSONAL AND INSTITUTIONAL OBSERVED MEANING

Focusing on *the computer lab session*, the analysis of observations showed some difficulties in the use of computers, such as errors in using the secondary menus, which lead to take the default options, which are not always appropriate to solve the task. In these sessions, the progression in learning and relating the diverse elements of meaning was not clear. This was possibly due to the fact that new elements of meaning were introduced in each session, and many students did not have the time necessary to achieve a meaningful understanding of these elements, and the relations among them. Different analysis situations and very different variables were considered in each computer lab session, which required the application of very different relations between the elements of meaning for each task. All this implies the integration of the elements, which entails a great semiotic complexity and this causes the tasks to be more significant but also more complex.

These results were considered in the study of the students' *personal meaning* that they constructed in the didactic sequence. Below we will first focus in the form in which the elements of meaning incorporated by the use of the computer appear in these *personal meanings*. On the other hand, we will describe the difficulties observed in relating these elements of meaning. We will use a qualitative approach, due to the exploratory character of this study.

Extensive Elements

Introducing the computer allowed us to use some data sets variables to study the fitting of empirical distributions to the model. We worked with a counterexample, taking a discreet distribution with few values that was not well approximated by a normal distribution.

Ostensive Elements

The main difficulties were confusing theoretical and empirical density trace, using absolute frequency histogram instead of relative frequencies, using the histogram instead of the polygon, incorrect use of frequency tabulation.

Actuative Elements

The use of the computer produced some difficulties related to the confusion between the theoretical density trace and the empirical data, leading to many errors in visual comparisons. As a result, some students made an incorrect descriptive study of the data because they were based exclusively on the density trace shape, without analyzing other elements and without considering the source of data. Incorrect computation of critical values or incorrectly represented density curves resulted from not modifying the default options. Many problems arose in computing percentages in central intervals, which requires the integration of diverse actuative elements (manual actions and computer actions).

However, we emphasize that most students managed to make correct and suitable actions in each activity. This suggests acceptable learning, considering the diversity of menus in the software. This does not just represent a higher semiotic complexity as compared to traditional class, but also requires learning new actions related to new elements and goals.

Intensive Elements

The most difficult elements were interpreting asymmetry and kurtosis coefficients, quartiles and percentiles. The central intervals property led to diverse computing errors (actuative element) and to errors in applying and interpreting this property in the real situations of data analysis.

The interpretation of probabilities as tail areas was difficult for students, especially when they needed to operate with probabilities in different intervals. Some students who were unable to relate the ostensive element they were representing to the situation and consequently, provided erroneous answers.

Incorrect understanding of the normal distribution as a model led to a failure to differentiate empirical and theoretical distributions or statistics and parameters. These elements are difficult to apply, since they require relating other types of elements and understanding the network of semiotic relationships between them. Consequently, the data analysis tasks presented in the computer lab session were far more complex than the traditional tasks.

Validative Elements

Graphical representation was a main validative element applied by the students. Although there were many activities in which a previous analysis of the situation could be made, in most of the cases the students only visualized diverse graphs to draw final conclusions from them. Few students arrived at a generalization from some properties of the normal distribution.

In general, the students only analyzed a single property when a conclusion was requested. For example, they observed that the asymmetry coefficient was close to zero, from which they concluded that the distribution was normal, without analyzing the graphs in which it was clear that the distribution was discreet and was bimodal. Some students displayed difficulties to produce a synthesis and they only gave partial validations. Although they uses some representations in a meaningful way, they did not manage to integrate them, in agreement with what was found by Ben-Zvi and Friedlander (1997).

Relating Meaning Elements

The tasks to be solved with the computer raised a philosophy different from the classic tasks. There an integration of the elements was needed to reach a coherent conclusion and to take a decision about whether: “the empirical distribution fits or does not fit well a normal distribution”.

Most of the students who previously had managed to correctly apply many of the elements described in isolation, failed in these new tasks. This was easily perceived when the students analyzed the property of central intervals. In that case, they firstly need to recognize the software options (ostensive elements), secondly they need to make calculations with the computer and by hand (actuative elements). Once they obtain the results they need to interpret frequencies, intervals and percentages (intensive elements). Finally, they have to integrate all these elements making a synthesis (validative element) to justify the final conclusion. In general, the students failed when applying this property because they made mistakes in the application of some of these elements. For this reason and because this process was too complicated, they decided to analyze other properties or characteristics where using such a great diversity of elements is not needed.

CONCLUSIONS

The results obtained show the complexity of the concept *"normal distribution",* the understanding of which requires understanding of the different meaning elements and relating them to the problem fields where the normal distribution is applied. Our study has shown that the use of computers is not trivial. On the one hand, the computer presents specific difficulties that should initially be considered, since they can interfere in the introduction of the specific concepts. On the other hand, the use of the computer incorporates new meaning elements or new approaches of them that in general increases the semiotic complexity of each property or concept. Consequently, we should be conscious of this and consider these particularities when planning the teaching.

In addition, the computer facilitates the simultaneous working with many elements. This fact is positive in the sense that time of representation is reduced, and can be spent in the interpretation of each element. However, we should realize that providing simultaneously a great amount of information could increase the difficulty to obtain a coherent integration of all the implicit meaning elements. All of this warns us to be cautious with the possibilities and problems that the use of the computer in teaching Statistics involves. We believe that the complexity of statistics education, increased by the introduction of computers should be studied in future Statistical Education research.

REFERENCES

Batanero, C., Tauber, L., & Meyer, R. (1999). From data analysis to inference: A research project on the teaching of normal distributions. *Proceedings of the 52nd Session of the ISI.* Helsinki.

Ben-Zvi, D., & Friedlander, A. (1997). Statistical thinking in a technological environment. In J.B. Garfield, and G. Burrill (Eds.). *Proceedings of the IASE 1996 Round Table Conference* (pp. 11 – 23). Voorburg: ISI and IASE

Cohen, S., & Chechile, R. (1997). Overview of ConStats and the ConStats assessment. In J.B. Garfield and G. Burrill, (Eds.). *Proceedings of the IASE 1996 Round Table Conference*. (pp. 109 – 118). Voorburg: ISI and IASE.

Godino, J.D. (1999). Implicaciones metodológicas de un enfoque semiótico-antropológico para la investigación en didáctica de la matemática. In T. Ortega (Ed.), *Actas del III Simposio de la SEIEM* (pp. 196-212). Valladolid.

Moore, D.S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review, 65*(2), 123-137.

Tauber, L., Batanero, C., & Sánchez, V. (2000). Comprensión de la distribución normal en estudiantes universitarios. In C. Loureiro, F. Oliveira, and L. Brunheira (Eds.), *Ensino e Aprendizagem da Estatística.* (pp. 117-130). Lisboa: SPE, APM y Universidad de Lisboa.