Title: Suggestions for Teaching Exploratory Data Analysis in Engineering Education within the European Credit Transfer System

Authors: Mónica Blanco, Marta Ginovart (Department of Applied Mathematics III - Technical University of Catalonia, SPAIN)

Sub-theme: Mathematical activity in the 21st-century classroom.

Abstract: Exploratory data analysis methods form an integral part of many engineering programs, as a component of a course in basic statistics. In this article a number of useful approaches are suggested to aid the instructor of the exploratory data analysis methods in engineering education within the framework of the European Credit Transfer System. These include ideas for oral presentations, projects, writing component, student-generated data, and cooperative learning. We believe these techniques help students develop an appreciation for the field of exploratory data analysis and the broad range of its applications in practice.

Key Words: Cooperative learning; jigsaw model; exploratory data analysis, statistical project; student-generated data.

1. Introduction

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions and develop suitable models. The EDA is not a mere set of techniques, but an attitude about how a data analysis should be carried out. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics *per se*. The seminal work in EDA is *Exploratory Data Analysis* by Tukey (1977). Over the years it has benefited from other noteworthy publications such as *Data Analysis and Regression* by Mosteller and Tukey (1977), *Interactive Data Analysis* by Hoaglin (1977), and *The ABC's of EDA* by Velleman and Hoaglin (1981). It has gained a large following as "the" way to analyze a data set. Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, uncovering the structural secrets of the data, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides unparalleled power to carry this out. Instruction in EDA has thus been recommended as an important part of an undergraduate instruction in connection with the statistical applications to the engineering field.

In this article we outline some suggestions for teaching EDA methods in the framework of the European Credit Transfer System (ECTS). This is one of the topics of a compulsory undergraduate course in basic statistics. We are developing this course mainly at the School of Agricultural Engineering of Barcelona (Technical University of Catalonia, Spain) but could be applied in any engineering educational context. We believe that the very nature of the subject calls for special consideration in the teaching of the subject, especially with regard to the new European higher education context. The ideas given in this paper gather our experiences in teaching the courses in statistics offered by our school and other engineering schools we have been involved with. These consist of an undergraduate course primarily encompassing basic statistics. Many of these proposals will, of course, need to be tailored to the level of the course being taught, depending on the scope of the institutions where the course is taught.

So far little has been published about the teaching of EDA in the new framework of ECTS. ECTS makes study programmes easy to read and compare, when it comes to the recognition of study periods undertaken abroad by mobile students through the transfer of credits. It can be used for all types of programmes, either distance and in-person courses or blended learning courses, covering in turn self-study and work experience. ECTS is a student-centred system based on the student workload required to achieve the objectives of a programme of study. These objectives should preferably be specified in terms of learning outcomes to be acquired. Learning outcomes are sets of competences, expressing what the student will know, understand or be able to do after completion of a process of learning. Competences represent a dynamic combination of attributes, abilities and attitudes, which should correspond to specified learning outcomes. Student workload in ECTS consists of the time required to complete all planned learning activities such as attending lectures, seminars, independent and private study, preparation of projects and examinations. In most cases one credit ECTS stands for around 25 to 30 working hours.

Since the university policy is heading towards the integration into the European higher education area, we set to plan the teaching of EDA methods guided by the ECTS philosophy. Hence, aside from improving the teaching and learning process, we were committed to accomplish the following specific aims:

1. To increase the students' interest and to convince them of the importance and usefulness of EDA methods. The accomplishment of this aim may help to keep up the attendance and, therefore, to instill constant work.

2. To plan and run activities for the students both in and outside the classroom, to ensure the successful completion of the work required.

3. To adjust workload and/or educational activities through an effective use of students' ratings.

4. To analyze and assess the learning outcomes achieved.

This paper opens with an outline of class meetings or lectures, and other educational activities, designed with the ECTS philosophy in mind. The paper proceeds with a sketch of one specific activity, namely, the project work, which derives into a discussion on cooperative learning, the writing component and oral presentation in subsequent sections. Assessing the results is the next step, leading to the final conclusions.

2. Class Meetings

As a part of the general process we outlined in Blanco et al. (2006), teaching and learning EDA is accomplished through a combination of lectures, problems classes, computer practicals and students' private study. We try to help students develop their independent work as well as their team work capacity. Right at the beginning of the course our students are given a handout with the specified learning outcomes, classified in the following three outcomes according to Bloom's taxonomy (in Bloom, 1956):

- Knowledge, defined as the remembering of appropriate, previously learned information.

- Comprehension, as grasping the meaning of informational materials.

- Application, meaning the use of previously learned information in new and concrete situations to solve problems that have single or best answers.

For instance, the learning outcomes regarding graphs have been articulated as follows:

> After attending the course the student will:
> - List and characterize a variety of graphs (dot plot, pie chart, bar chart, stem-and-leaf diagram, histogram, scatter plot). [Knowledge]
> - Identify the most convenient graph to display a certain data set. [Comprehension]
> - Plot a graph of a data set, considering all possible options according to the nature of the data. [Application]

One goal of the course is to convince students of the importance and usefulness of EDA methods. This is best accomplished by showing, rather than telling. An effective way to help students see the value of EDA methods is to demonstrate their strong points through examples, preferably real ones as opposed to contrived ones. For instance, the very first computer practical class the students are required to create a small data set from their own student record, to start exploring it graphically and numerically. Handouts and problem sheets can also fulfill this need very effectively, so that the students are provided with 'living' data. Additional sources of real data sets are the *Journal of Statistics Education Data Archive* [http://www.amstat.org/publications/jse/jse_data_archive.html], the Data Bank section of *Teaching Statistics* [http://www.rsscse.org.uk/ts/] and *The Data and Story Library* [http://lib.stat.cmu.edu/DASL/]. Ideal data sets for demonstrations of EDA methods are those which are clearly from non-Gaussian or asymmetric distributions, those for which the population median is more relevant than the population mean, and data sets with outliers (Micceri, 1989).

Presenting examples of assorted EDA methods provides the instructor with an excellent opportunity to engage the students in discussion of the assumptions made by the procedures and whether they are met or not. Discussion on which procedure, graphic or summary data is appropriate for the problem at hand is suggested here for pedagogical purposes. Reading selected news from newspapers turns out to be very profitable as a means to discuss real examples.

We would even suggest that advanced students could also be encouraged to read journal articles related to another subjects and prepare statistical reports on them.

To increase the interest and the level of achievement, it is our preference to apply cooperative learning, through activities such as statistical jigsaws or project works. The use of projects and/or cooperative learning activities in any statistics course has been advocated by Jones (1991), Dietz (1993), Garfield (1993) and Fillebrown (1994), among others. At every step not only the dynamic interplay, but also the independent learning must be guaranteed, through self-reflection and evaluation of their own work processes.

As an initial effort to set up cooperative learning groups we tried the Jigsaw model (originally presented by Aronson and colleagues, 1978). In the Jigsaw model the student becomes a member of both a learning group and an expert team. Each puzzle piece must be worked out to form a complete picture. We prepared three different "expert sheets" with the following topics: 1) descriptive statistics of central tendency, 2) descriptive statistics of dispersion, and 3) descriptive statistics of location. These materials gathered the puzzle pieces for every member of a learning group. After having studied their corresponding sheet, the members joined expert teams to discuss about their particular piece of the learning puzzle in the classroom. Upon completion of the expert teams' work, the members returned to their original learning groups and shared the results. Oral presentations and a question-and-answer session allowed the groups to share their findings. To assess whether the learning group's goal was achieved the students had to solve independently a particular exercise regarding the topics of the puzzle. The final discussion about their results - first, with the rest of the group, and then, with the instructor - illustrated both peer-assessment and self-assessment, respectively.

3. Projects

In this section we discuss some issues related to the use of projects in a course in statistics and, in particular, in EDA. A project in statistical education is a learning experience which aims to provide students with the opportunity to synthesize knowledge from a complete statistical analysis of a problem, and critically and creatively apply it to real life situations. The main feature is that they can decide independently how and in which order to solve the tasks necessary to successfully cope with the project (Kubinova et al., 1997). This process, which enables students to acquire skills like collaboration, communication and independent learning, prepares students for lifelong learning and the challenges ahead. As a cooperative activity, project works are to be done in groups. This methodology aids the student gain competence in working both independently and in team, managing time effectively and using computer resources appropriately. Some of the aspects included, such as the discussion of statistical methods used, critique of assumptions, analysis of data, and conclusions, can be worked during the course as homework assignments, for instance.

Our students must carry out a project from real data derived from a survey conducted among the students at the beginning of every semester since 2000. The data basis gathers a collection of attractive real data such as gender distribution of students, the amount of time students spend surfing the World Wide Web, the amount of time students spend watching TV, students' mobile phone expense, students' job situation, favourite sports or grades of specific subjects. Students are clearly involved in the process, all the more so because dealing with their own data allows them to combine their personal interests with statistics, thus increasing student enthusiasm in the course.

When planning to use projects in a statistics course, one likely question is: should students analyze their data on their own? Or had the teacher better provide them with some guidelines throughout the process? Unlike previous years, this year we are going for guidance. At the beginning of the course we suggested our students what to do at every step, including an estimation of the time required for every activity. In addition, a handout collected the assessment criteria, which encompassed aspects from demonstration of knowledge, conclusions and structure, to effective collaborative work and creativity. Discussion with the teacher about the work done at every step was also favoured insofar as it ensured immediate feedback for the completion of the project.

4. The Writing Component and Oral Presentation

A fundamental part of the project work is the writing component and oral presentation. The incorporation of a writing component in statistics courses has been encouraged in recent years by Radke-Sharpe (1991) and Garfield (1994). Writing helps students to think about the assumptions behind statistical or graphical procedures, to formulate these assumptions verbally, and to critically examine the suitability of a particular procedure based on its assumptions. The inclusion of some writing facilitates student comparison of procedures. Putting such comparisons into words and verbally

justifying the use of a particular procedure are tools that will serve students well in their future scientific or academic writing.

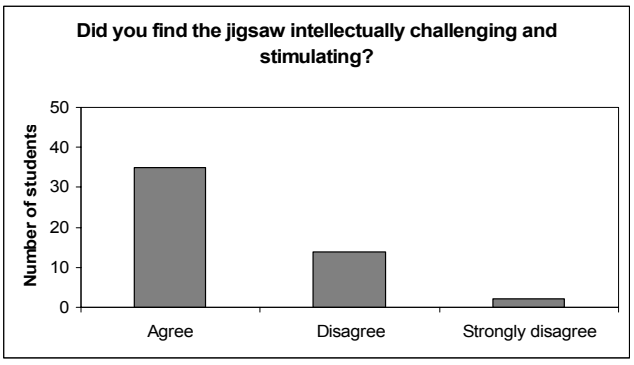
There are many ways to include writing as a part of a course in statistics. Students might be asked to compare and contrast the use of different EDA procedures. Writing assignments can also be used to ensure that students understand the various meanings of the terminology. In a data analysis setting, one could ask the students to select the most appropriate procedure and justify their choices. Another possibility is to ask, individually, for a short, purely verbal description of a particular aspect of the project. For example:

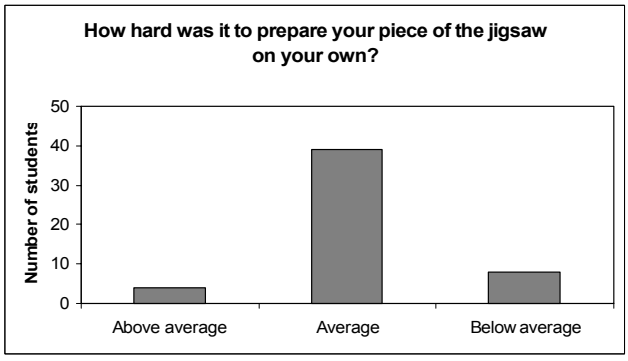
1. Why did you select the items you have treated in the project?
2. How did you set up the analysis of the data? Which difficulties did you meet?
3. Describe briefly some the methodological tools you employed in the project.
4. Comment on some of the most interesting outcomes in your opinion.

Encouraging students to put concepts such as these into words will strengthen their understanding of those concepts. One of the criteria for assessing the project is precisely mechanics of writing and speaking. It is worth emphasizing that the above-mentioned Jigsaw activity offers students the chance to orally develop a statistical topic.

5. Results and Conclusions

To increase the students' interest and to ensure the successful achievement of the learning outcomes, we moved towards a novel approach in our teaching activity and, accordingly, tried out several educational activities, both in and outside the classroom. All through the teaching and learning process we used student feedback for formative evaluation. The improvement of one's own teaching relies largely upon the knowledge of how a class goes and where changes may be needed or attempted. By the end of the semester our students usually rate the importance of items regarding learning, satisfaction, course characteristics, assignments and workload. This year they were also invited to comment on the development of some specific activity during the course. Though not the only source of feedback, student ratings provide an excellent guide for designing the teaching process. At the end of the Jigsaw session the students were asked to rate the activity:



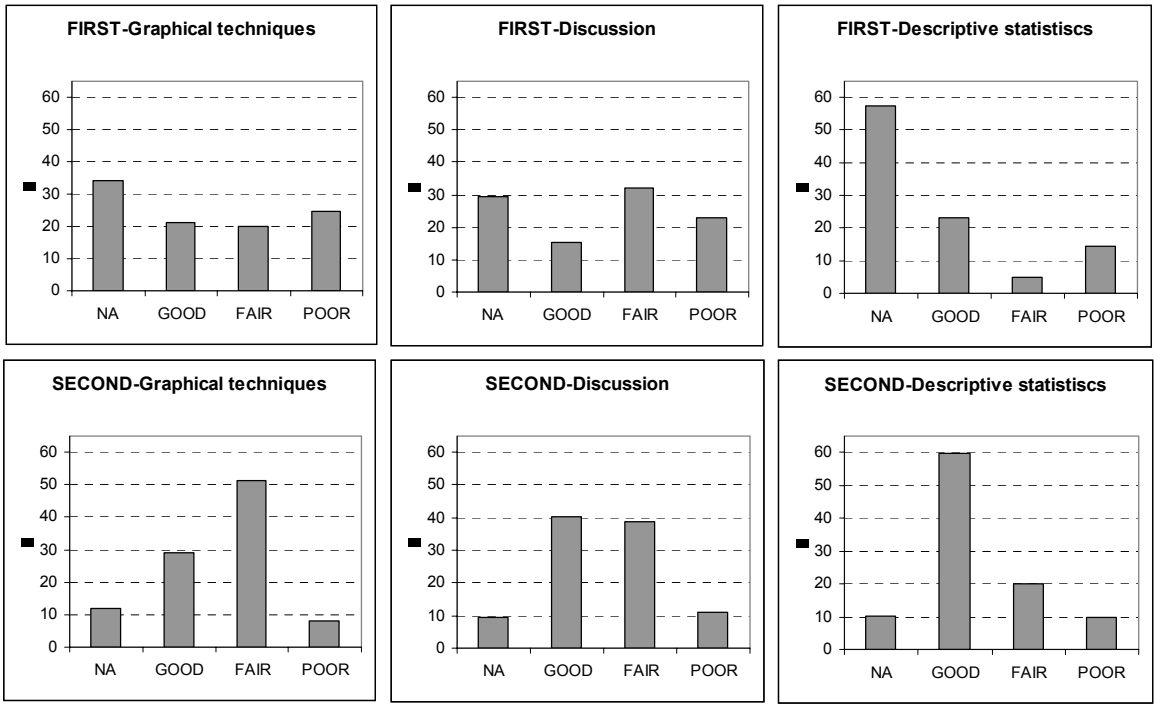


When questioned whether they would like to run another similar activity, 86% of the students answered affirmatively. We single out the following additional comments concerning the Jigsaw session:

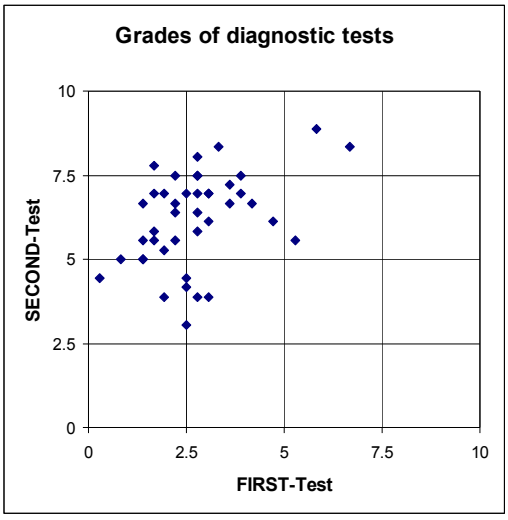
"We are required to understand the material first in order to explain it to the others later on."
"It helps to settle down knowledge."
"It facilitates a quick understanding."
"Team work is fun."
"Learning is rendered active, instead of passive."

We used again students' ratings as a source of feedback to adjust the student workload and/or educational activities. Students were asked about the "real time" they invested in the reading of the corresponding sheets for the puzzle and for the fulfillment of the project work. From the students' answers the course planning will be updated next semester.

Since our fourth aim was to assess our students' level of achievement, we set out to analyze any variation in the achievement of every individual student. At the beginning of the semester a diagnostic test provided us with information about the students' statistical background. The diagnostic test consisted of a set of questions grouped into the following items: graphical techniques, computation of descriptive statistics and discussion of results. In order to quantify the students' improvement, a similar test was run in the middle of the semester, without warning the students in advance. This way we attempted to assess what the students had effectively learnt without making an extra effort oriented to the purpose. In both tests every question was marked with either NA (non-answered), poor, fair or good. From the graphics below it is clear that in the second test the number of non-answered questions and questions marked as poor has essentially dropped, whereas the number of questions with good or fair marks has increased considerably.



On the other hand, we were also interested in tracing individual improvement. To that purpose, we translated the former marks into numerical, which added up to a comprehensive numerical mark for every test. Then, from the numerical marks got by those students who took both tests the following scatter plot was obtained:



Since all the points lay above the grid's diagonal, the scatter plot reveals a tendency towards improvement in the second test with respect to the first test. Hence we can state that our students' learning achievement shows a satisfactory progress.

Furthermore, these tests along with the jigsaw activity contributed to control the index of absenteeism. Of the sixty students who took the first test, around fifty students took part in the two subsequent activities, namely, the jigsaw activity and the second test. Compared with previous semesters, this is quite a negligible ratio of absenteeism.

In short, showing real examples, having students do some writing and oral presentations, assigning individual and group works, and exposing students to computing and graphical methods for the procedures taught in class, certainly made our students change their attitude towards the course, improve their level of achievement and keep up the attendance.

Acknowledgements

The authors are very grateful to their students for their valuable contribution to the material presented here.

References

Blanco, M.; Ginovart, M.; Estela, M. R.; Jarauta, E. (2006), "Teaching and Learning Mathematics and Statistics at an Agricultural Engineering School." *Proceedings of the CIEAEM 58 Congress. Changes in Society: A Challenge for Mathematics Education*, University of West Bohemia, Plzen, 152-157.

Bloom, B. S. (ed.) (1956), *Taxonomy of Education Objectives: Handbook I: Cognitive Domain*, David McKay Company, New York [Major Categories in the Taxonomy of Educational Objectives]

Dietz, E. J. (1993), "A Cooperative Learning Activity on Methods of Selecting a Sample." *The American Statistician*, 47, 104-108.

ECTS Users' Guide (2005), [Online]. [http://ec.europa.eu/education/programmes/socrates/ects/doc/guide_en.pdf]

Fillebrown, S. (1994), "Using Projects in an Elementary Statistics Course for Non Science Majors." *Journal of Statistics Education* [Online], 2(2). [Journal of Statistics Education, V2N2: Fillebrown]

Garfield, J. (1993), "Teaching Statistics Using Small-Group Cooperative Learning." *Journal of Statistics Education* [Online], 1(1). [Journal of Statistics Education, V1N1: Garfield]

Garfield, J. (1994), "Beyond Testing and Grading: Using Assessment to Improve Student Learning." *Journal of Statistics Education* [Online], 2(1). [Journal of Statistics Education, V2N1: Garfield]

Jones, L. (1991), "Using Cooperative Learning to Teach Statistics." *Research Report* No. 91-2, The L. L. Thurstone Psychometric Library, University of North Carolina.

Kubinova, M., Novotna, J., Littler, G. H. (1999), "Projects and mathematical puzzles-a tool for development of mathematical thinkin." *European Research in Mathematics Education Proceedings of the First Conference of the European Society for Research in Mathematics Education*, Forschungsinstitut für Mathematikdidaktik, Osnabrück, [Online], II, 53-63. [Cerme 1 - Proceedings]

Lawall, M. L. (1998), *Students rating teaching, How student feedback can inform your teaching*. University teaching services, The University of Manitoba [Online]. [http://www-ice.upc.edu/pro_accio/seeq/millora.pdf]

Micceri, T. (1989), "The Unicorn, The Normal Curve, and Other Improbable Creatures." *Psychological Bulletin*, 105, 156-166.

Moon, J. "Linking levels, learning outcomes and assessment criteria." Exeter University, [Online]. [http://www.bologna-bergen2005.no/EN/Bol_sem/Seminars/040701-02Edinburgh/040701-02Linking_Levels_plus_ass_crit-Moon.pdf]

NIST/SEMATECH e-Handbook of Statistical Methods [Online]. [http://www.itl.nist.gov/div898/handbook/eda/eda.htm]

Radke-Sharpe, N. (1991), "Writing As a Component of Statistics Education." *The American Statistician*, 45, 292-293.

Southwest Consortium for the Improvement of Mathematics and Science Teaching (1994), "Models That Promote Cooperative Learning."*Classroom Compass*, [Online], 1 (2). [http://www.sedl.org/pubs/classroom-compass/cc_v1n2.pdf]