Assessing Shiny apps through student feedback: Recommendations from a qualitative study

José Antonio González[1*]
Email: jose.a.gonzalez@upc.edu
Mireia López[1]
Email: mireia.lopez.beltran@upc.edu
Erik Cobo[1]
Email: erik.cobo@upc.edu
Jordi Cortés[1]
Email: jordi.cortes-martinez@upc.edu

[1] Statistics and Operations Research Department, Barcelona-Tech, UPC, Barcelona, Spain
[*] Corresponding author: J.A. González, Statistics and Operations Research Department, UPC; C/ Jordi Girona 1, C5-221; 08034 Barcelona, Spain; +34934015867

Abstract

Teaching statistics has benefitted from Java applets, the successful technology that appeared in the late 90s and which allowed real interactivity on an internet browser. Combining dynamic functionality with the web provides an inspirational complement to the contents of many subjects in undergraduate statistics courses, especially for active learning activities. Since Java applets are becoming obsolete, we explore a different technology based on Shiny and R, the latter being a popular statistical language nowadays. Although the pedagogical value of these tools has been implicitly accepted so far, our aim is to consider the students' perspective while investigating more suitable means to accompany using apps in statistics. We conducted a qualitative study in which we tested ten of our applications and collected student opinions through questionnaires and regular meetings. Our conclusions indicate that the students view these resources positively, although they may not be very efficient without enough support to facilitate both getting started and using the tools effectively. In addition, programming in R is surely more accessible and satisfying to statistics lecturers than other languages and, consequently, implementing instructional activities can be specially tailored by the teacher.

Keywords

Applets, Statistics education, qualitative study, focus group, Shiny

1. Introduction

Research on statistical learning has been developed greatly over recent decades, and a large part of this development is related to the parallel and impressive growth of Information Technologies (IT). The main support for educational theorists comes from the constructivist paradigm [3, 8], which states that students construct their knowledge by transforming and reorganizing previous knowledge through fitting new information into their existing cognitive structures [16, 21]. Furthermore, under this paradigm, instructors facilitate learning to their pupils by means of several methods devised to boost an active student role [18]. Although the gap between the theory of constructivism and educational practice “seems to be difficult to bridge” [9], many efforts to fill this gap have been made in statistical education for “making the student a more active learner” [14] with the use of various tools, for example, simulations [10, 27]. Through a comprehensive review of studies, authors of [25] have provided robust empirical evidence which proves that simulations enhance traditional instruction.

In the nineties, a “great explosion” of Java applications (i.e., *applets*), took place on the web. At that time, they were “easily and freely accessible” [14], but nowadays the situation has changed profoundly. While Java is still being used and maintained for the web pages of thousands of powerful companies, there are calls for Java to be displaced in favor of simpler technologies (JavaScript, CSS, HTML5). Recently, some browsers have implemented changes that prevent Java from running. Java is also a resource-demanding platform and does not work well on mobile devices and tablets. Moreover, Java updates quite often, which causes their users some discomfort.

The consequences are already apparent. Many of the innumerable and superb collections of applets devised to enhance learning on statistics subjects (VESTAC [4]; RVLS [13]; WISE [2]) have been forced to either update their material or warn their visitors that they have to either use the programs locally or modify the configuration of their computers. Most collections of statistical resources (applets, among them) are practically useless, due to the increasing number of missing links to the original material. To illustrate, the STAT-ATTIC page [6] includes links to more than 600 applets, but only a few of them could be reached successfully.

In consequence, we have developed a set of small routines based on the technology we call (in short) Shiny. The interactive applications are written in the R language and can be made available to others via the web, while the source can also be shared with the community. The result is both robust and attractive with not much effort. The Shiny apps cover a wide spectrum of probability and statistics issues for undergraduate students, and we built them for educational purposes in order to allow users to interactively discover key concepts in probability and statistics. Although there is general agreement that a guided approach is needed to ensure students improve their learning [11, 14], we were unable to find enough guidelines about the best design for boosting student discovery. To determine the educational benefits, we posed two research questions:

- Do the students find the Shiny apps useful enough to become part of their course materials?
- How can the apps be improved to be more useful?

In order to answer these questions and reach our research objectives, we designed a study with a group of volunteers, who: tested a set of applications that were specifically chosen for their statistics course; answered a questionnaire for each tested app; and attended regular meetings so that we could collect their impressions.

This paper will first describe the Shiny applications, followed by a presentation of the study design and an analysis of the data collected from the focus group and questionnaires. Finally, we discuss the results and make some recommendations in the Conclusions.

2. Shiny applications

2.1. A sketch of Shiny

Shiny is a web framework designed by Rstudio [23] for developing applications using the R environment. The developer has to be familiar with the R language, but does not necessarily have to know elements such as HTML, CSS, JavaScript or internet architecture, though it may help in reaching the goal one has in mind.

After obtaining a license for academic use, we prepared a web server on a virtual machine with Linux S.O., where the shiny server would be hosted. The instructions given by the provider were quite straightforward, with only minor issues that were quickly solved by email. A user with root privileges created an account for each person in charge of developing the programs. After that, each developer created a folder under the special folder called *Shinyapps*, and wrote the application code in two files named *ui.R* and *server.R*, which were placed in the subfolder. The name of the web app is the name of the specific program folder, after the server URL and the user name. For instance: http://shiny-eio.upc.edu/josean2/twomeans/. For developing and testing tasks, it is notably more comfortable to use the Rstudio software, which allows the user to work either locally or remotely (with RStudio Server) on a desktop computer.

There are likely many sites with applications developed using Shiny, but they have not been sufficiently disseminated. Post [20] presented at eCOTS his collection of Shiny applications for undergraduate and graduate mathematical statistics courses, and Doi et al [7] extensively report on several applications from their collection; yet, both lack explanations of their teaching experience. Show Me Shiny [26] is a large and heterogeneous directory pointing to different sites; there are some interesting demonstrations, and it covers mostly data exploration. For didactical purposes, Rstudio [24] maintains a "Teaching" section in its gallery, which is fed by many contributors. As far as we know, the big picture for Shiny is that it is still not nearly as prominent as Java applets.

2.2. Principles and examples

The basic features of our applications include: interaction with the user (the student), graphical output and often summary results. Each one is an educational packet containing a brief lesson. We did not contemplate developing self-contained courses. These applications are created as a complement to face-to-face teaching. Our apps were designed to be "useful in understanding the course" instead of "the course being necessary to understand the application", although a minimum background about previous concepts in the subject is expected of the student. These applications attempt to ensure that the resources are used autonomously by the students and allow them to extract the relevant conclusions on their own. A proposed introduction and guidelines have been drafted, and they will be improved upon with student contributions.

In accordance with Rossman and Chance [22], in order to facilitate learning we have tried to maintain some features from one application to another, such as the use of introductory texts, the positioning of the input elements on the left and the distribution of content among several tabs when the application becomes complex. However, we decided that the applications would be more stimulating if we deliberately avoided using a tight standard in our designs. Most of the applications expand on some aspect that is usually presented in introductory courses on probability and statistics. We have occasionally included advanced topics to complement a fundamental topic. For instance, the “VAC” app introduces users to the characteristics of continuous random variables by emphasizing the properties of a probability density function: positiveness and area under the curve equal to 1. In any event, the curious student can approach the topic of parameter estimation by means of the same model employed in the probability section, which includes a brief explanation of the maximum likelihood method.

Some of the main topics for each application are summarized in the table provided in Appendix A. The list, which is maintained and updated at the first author’s web address (http://www-eio.upc.es/~josean/shinyweb/jag_shiny.html), is expected to grow eventually. As reflected in the table, simulation is an important resource that is employed in most apps related to statistics. By means of large volumes of data, it is possible to obtain either expected results (such as the Normal distribution for the sample mean, a common procedure in introductory courses) or unpredictable results (such as the strange distributions that appear whenever a fundamental premise is violated, e.g., in the “modlin” app).

Allowing the students to experience what happens if they do or do not comply with premises can be an efficient method for capturing their attention. One example can be found in Figure 1. The “mnas” app addresses the estimation of a proportion and allows violating the essential independence premise applied to the sample (that is, observations in the sample must not be correlated). An input control allows the user to set a parameter called “affinity”, which operates as an autocorrelation in the response. When the affinity is set to 0.5, the 95% confidence interval for the true proportion is systematically shorter than necessary: the depicted simulation reveals that only 81.9% of CI’s include the actual value. The accompanying picture illustrates the cause: the sample proportion coming from non-independent data has a greater than expected spread. Appendix B includes an extensive description of one of the applications, “anova”, which can be useful in giving the reader a more precise idea of the general operating mode.

3. Method

3.1. Study design

In order to appraise the degree to which students accept the applications and find them useful, we decided to carry out a study with a focus group in which they discussed their thoughts, ideas and opinions. We wanted to create a comfortable environment to help them share information [19]. At the beginning of the spring semester of 2016, we called for participation from among the students in the subject of “Probability and Statistics”. All the students were enroled in the Bachelor’s Degree in Informatics Engineering, a four-year degree offered at the Technical University of Catalonia (UPC). The call was presented in class and with a message in the intranet in the two groups taught by two of the authors. Each group had approximately twenty students. Five students (four men and one woman) showed interest. This proportion was consistent with the 8.17%[1] of female students in the degree in Informatics Enginering at UPC in 2016-2017. All five were college sophomores and shared similar educational backgrounds. We found five to be a suitable number because the meetings were planned around precise and defined topics [12].

From the beginning, the students were informed that in order to acknowledge their contribution to the study, they could raise their course grades with the continuous assessment activities designed for the whole group or they could choose to be evaluated by their contribution to the study through questionnaires and attendance at the meetings. This grade is a factor that increases the grade of the exam[2] by up to 5%, and it is always offered to all the students. Thus, the participants in the study were given an additional way to obtain this increase.

We planned six parts, according to the breakdown of topics in the course. Within the period corresponding to each part (about two weeks), we sent an email to the participants with a link to the respective questionnaire and the internet addresses of the applications. Specifically, the following 10 applications were employed in the endeavor. All of them were conceived while contemplating key topics in the “Probability and Statistics” syllabus:

- Part 1 (probability): “pcondic”
- Part 2 (random variables): “vardisc”, “VAC”
- Part 3 (models for random variables): “bombas”, “binopois”
- Part 4 (statistical inference): “mnas”
- Part 5 (two-group testing): “pares”, “twomeans”
- Part 6 (linear model): “a-ojo”, “modlin”

Two tools collected the data for each part: a questionnaire and a focus group meeting. We asked the students to try each application on their own and followed up by collecting their opinions. The questionnaire consisted of five open-ended questions to be answered before the meetings, with the same ones being used by the moderator during the meetings. Initially, the questions were:

1. Explain briefly what operations you have performed with the application.
2. Did you make a “discovery”? Did something already explained by the teacher become clearer after your exploration?
3. Did you try to get some particular result and not succeed? Describe the attempt.
4. Could you highlight a positive aspect of this application? And a downside?
5. Finally, have you any ideas that may be helpful in enhancing this application?

At the request of the students, we added an open question from the second questionnaire: “Here, you have space to add any other comments”. Furthermore, we eliminated questions 4

[1] https://gpaq.upc.edu/lldades/indicador.asp?index=1_1_5
[2] The grades in Spain are in the range from 0 to 10. In this case the students can increase the grade of the exam by multiplying by a factor from 1 (no variation) to 1.05 (5% increase).

and 5 from the questionnaires but not the focus group meetings in Part 3, since we realized that the answers focused on technical issues rather than those related to the subject.

Then the five students met with the authors (ML and JAG) for 30 minutes to complement the data collected with the questionnaires. The meetings were recorded with the students' permission, and they were informed that their viewpoints would be used only for academic and research purposes. JAG acted as the moderator and ML as the assistant. The meetings were held in the conference room of the department and we used a computer and projector to help the students share their views.

The answers to the questionnaires were used to prepare the meetings by foreseeing potential problems of student comprehension related to key aspects of the current topic covered by the app. The discussions were also intended to highlight the conceptual gaps in a climate of trust that made the students feel free to express their opinions. After two meetings, we modified their structure: one of the students was tasked with an active, five-minute presentation of the current application, followed by an open discussion. In the last meeting, we also recorded the computer screen in order to complement the oral information given by the students and to be able to follow their explanations of the specific data simulation that was used.

Finally, we asked them to answer a closing questionnaire, which was not related to any particular application. It included only three questions: 1) Which are the three most useful apps for learning the subject topics? 2) Which are the two least useful apps? 3) Please summarize which features an app should have in order to be a good learning tool.

3.2. Data analysis

The six meetings were held on a regular basis between February 29th and June 10th, and the students answered all the questionnaires and attended each meeting (with the exception of one student on the 29th of April). They participated actively, both in remarking on different aspects of the applications and in asking questions that arose in the discussion. Overall, the students' engagement was highly satisfactory.

After each meeting, one of the authors (JAG or ML) listened to the focus group tape and then created an abridged transcript, focusing on the research questions and transcribing only the parts that "assist in better understanding of the phenomenon of interest" [19]. Then the final version was completed by a second author (ML or JAG), who repeated that process confirming and adding observations. All the transcriptions were made in the following two-week period after the meeting. The transcription was organized in six blocks, each corresponding to a meeting, so each statement of the transcript was coded with the number of the block (1 to 6) and a numeration within the block to facilitate the subsequent localization for the analysis.

From the transcription and focusing on answering the two research questions, we selected 98 statements from among the blocks and 111 statements from among the anwers of the questionnaires. After this first organization of the data, a Constant Comparison Analysis was then undertaken [15]. We coded the different statements to identify those that express similar positions. An abductive methodology was used to define codes (codes that had emerged from the data iteratively). First, half of the set of data was read to identify the first list of codes. Next, by means of the previously identified codes and an *Others* code, the entire set of data was processed. After a first round, we coded the statements that take into account mainly the topic expressed. At this stage, we realized that it was important to gather also the attitude that the student expressed in relation to the topic. After a second round, we established two code groups: one according to the topic of the statement (we called it Topic) and the other regarding the attitude of the student with the topic (we called it Attitude). In the Topic group, we established the following codes: *First contact*, *Technical issues*, *Graphics and interface*, *Data redundance*, *Statistics related* and *Usefulness*. In the Attitudes group, we identified the following codes: *Confusion*, *Troubles/difficulties*, *Suggestions/proposals*, *Judgment/impression*, *Appreciation* and *Association/connection*. An *Others* code was added in both groups. Each statement has, at least, one code from each group.

For example, the seventh statement in B2 “It consolidates what has been taught in class but it is just difficult to see the whole relationship” was coded as *Statistics related* in the Topic group, and *Troubles/difficulties* in the Attitudes group. In this case, the student expressed some difficulties (Attitudes) with the connection that was found between the contents in the syllabus and the task performed with the app (Topic).

This process is illustrated with an example in Table 1 where we selected all the statements codified with *First contact* and *Troubles/difficulties*.

4. Results

4.1. What the students told us

After the codification process, we obtained the classification shown in Table 2. From this table, the authors created the Figure 2 to visualize the links between the topics and the attitudes analysed. The authors have determined the position of the labels after considering the neighboring relationships expressed in Table 2. An ancillary analysis, based on a Principal Components Analysis on the data matrix coded with 0 and 1, provided a factor map where the codes of both classification groups were projected on the two first components, capturing essentially the same result.

During the interviews and in the questionnaires, the students alluded also to issues related to the subject. The strongest relationship determined in the analysis was found between the *Statistics related* topic and the *Connection* attitude. The students frequently referred to the *Graphics and interface* topic when expressing their perspective, which turns out to be relevant since the apps were designed to visually estimulate the student’s curiosity. The interfaces of the apps were well received in general; they often emphasized in the questionnaires that the interfaces were “*visual*”, “*intuitive*” or “*easy to use*”. In the meetings, one student assessed the interface very positively, describing it each time as “*very clear*”.

Considering their training in Information Technologies, the students paid notable attention to technical issues like data input. They preferred the apps for which the data input was “*easier and more intuitive*” and more than one method was possible. For instance, they repeatedly asked for a keyboard input for entering numerical data, in addition to the slider. These suggestions were sensible and led to improvements in the code.

The students also appreciated that the data was shown in more than one way (for example, both tabular and graphically, or with two different types of charts). The participants revealed that they would like to have the tables with raw data as well. During the meetings, it became clear that, even though they appreciated having more than one graphic, they did not always use them all. We were able to detect difficulties when interpreting some charts. The students also experienced some confusion when using statistical concepts.

The usefulness of the apps was also well appraised. In the first meeting, one student found the apps to be a "*complement to learning*" because when "*you want to practice, you want to learn, (...) you can change the data, you can try things*". He added, "*It helps a lot, at least for me, seeing things graphically helps me to understand things and I internalize them more easily*". He concluded, "*I find the app useful*". In the last meeting, another student told us that he had already learned the theory of the linear model for the exam but that the graphics in the app *modlin* "*helped him to completely understand it*".

The students requested guidance for the apps, such as a "*tutorial with instructions*" or some "*example or problem*" that could help them in their first steps. Without any assistance, conceptual gaps often appeared in the app description or in the subsequent deliberation. From the third meeting on, we also discussed the key statistical aspects in the apps, in order to collect information about the difficulties they were having. We corroborated that guidance was needed both in their introduction to using the apps and in helping them learn the key statistical issues.

The closing questionnaire emphasized the need for guidance and for examples. They suggested "*the inclusion of examples to provide the student with context*", and they pointed out that the apps should "*show examples*" in order to be a good learning tool. The example of the bombs (usually linked to the Poisson lesson) was appreciated. Regarding guidance, they suggested "*a short video or tutorial*" in order to introduce them to the first steps. The final conclusions also remarked on the need to keep the app simple by "*avoiding too much concentrated information*" and encouraging the use of sections for better structure. The usefulness of the apps was also mentioned (although not directly asked about). One student found that most of the apps "*have helped me to reinforce class knowledge*".

4.2. Discussion

In our study, we have seen that students can adopt different postures when operating with our apps: they can feel themselves confused, but they can respond also positively, through positive feedback or acknowledgment if the experience was fruitful. Nonetheless, they generally required additional support in order to fully discern the topics covered in the apps. When asked, the students agreed that guidance would be improved with introductory, explanatory videos (they reported a good experience in another subject, in the previous semester). Along similar lines, [10] remark that successful discoverers follow a plan whereas those who are unsuccessful proceed more randomly. Thus, it seems suitable to merge information with directions in the introductory material.

Our starting hypotheses assumed that the apps would help the students on their own. Under this premise, we provided them with only minimal information about their use. However, our study suggests the need for introductory mechanisms that foster the development of complementary resources rather than providing tools intended to be autonomous and self-sufficient. In this regard, a systematic review [1] revealed a significant advantage to explicit instruction as compared to unassisted discovery, although the discovery process was enhanced more than other forms of instruction.

According to [28], errors and misconceptions may be addressed by different methods (e.g., particular activities using a simulation tool or web applet) to help build correct reasoning. [17] suggest also that online multimedia material can be improved by including the discussion of misconceptions. We think that the developed applications extend this research along the right lines if it allows violating statistical premises as a possible scenario in some topics. For instance, showing that a failing premise causes a substantial change in the actual hit rate of a 100(1–α)% confidence interval can be an appealing lesson. The app "Linear model and premises", which makes extensive use of this principle, was selected as one of the three most useful apps by all the students in the final questionnaire, supporting that the approach was appreciated by the students. The message was clearly anticipated in one of the answers from the fourth questionnaire: "*With the help of the sample proportion [app], I got a bit of a clearer idea about how random samples are not simple, since we have seen only simple ones in class. So it helped to see the difference between both types*".

4.3. Programming issues

In our opinion, R and the Shiny environment are far more suitable than other programming languages like Java for developing tools such as those presented here. Thanks to the many packages that are freely available to R users, practically any of them can be approached and integrated into an application, so long as the computation time is compatible with the short response time expected for an interactive tool. Packages from the R community may also be useful for high quality graphics, thus freeing the developer from the responsibility of their programming.

Our advice for statistics teachers who are willing to develop shiny apps is to pay attention to the weaknesses in R and program their applications carefully. Many simulations may require substantial amounts of space and computation time, and R is particularly slow if the code does not take advantage of vectorized operations. Moreover, if the app is available only on a web server and the code is run and accessed remotely, multiple simultaneous users can cause long response times. We suggest also that lecturers employ local versions with their students, especially if the app consumes substantial resources.

The current versions of our codes can be found in the repository
https://github.com/joseanglez/Shinyapps.

5. Conclusions

For those teachers who feel confident about their programming skills with R and believe that their students can benefit from exploration and discovery, we encourage them to carefully design and implement tools like the apps presented here. They are able to instill intuition first and expedite understanding later. Based on our experience in developing the apps, our study and comparing the results with published research in the field [14, 5, 22], we have developed the following recommendations:

- Conceive the app according to precise objectives. Prioritize the key topics you want the students to focus on.
- Design a clear and intuitive app. Incorporate different data entry tools and different ways to visualize the results.
- Include the student as an active agent in your app. Avoid making the student a passive observer with routine operations.

And for every teacher keen to make use of such tools, we recommend always providing orientation, especially when the task deviates from the standard procedure:

- Equip the app with instructional materials: a) introduction, for learning the basics; b) examples, to provide context; c) complements, to facilitate the connections with other topics.
- Keep the introduction short in order to facilitate getting started.

Furthermore, take into account the pupil responses:

- Assess student progress; ask questions before and after their interactions with the app.
- Make them aware of the relevance of their own discoveries.
- Listen to their opinions and use their feedback to improve the tools.

After our experience of putting the students under a microscope, we feel the need to attract the attention of many teachers concerned with development of instructional materials. We believe that the effort to check the acceptance of their applications before leaving them available to a wider community is worth. Moreover, early evaluation may capture *a priori* credibility of the tools, which is necessary to undertake successfully any confirmatory study. Otherwise, the waste of resources to create ineffectual materials for students could be substantial.

Acknowledgements

The authors wish to express their gratitude to the students participating in this endeavour. Furthermore, we are thankful to Rstudio Inc. and their employees for providing the Shiny license and quick assistance when needed.

References

1. Alfieri, L., Brooks, P. J., Aldrich, N. J., Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1-18. doi:10.1037/a0021017

2. Berger, D. E. Saw, A. T. (2008). WISE Hypothesis Testing Tutorial.http://wise.cgu.edu. Accessed 04.07.2016.
3. Bransford, J. D., Brown, A. L., Cocking, R. R. (1999). How people learn: Brain, mind, experience, and school. National Academy Press.
4. Darius, P., Ottoy, J. P., Solomin, A., Thas, O., Raeymaekers, B., Michiels, S. (2000). A collection of applets for visualizing statistical concepts. In *COMPSTAT* (pp. 253-258). Physica-Verlag HD. http://lstat.kuleuven.be/newjava/vestac. Accessed 04.07.2016.
5. delMas, R., Garfield, J., and Chance, B. (1999). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Journal of Statistics Education* [Online], 7(3). (ww2.amstat.org/publications/jse/secure/v7n3/delmas.cfm)
6. DePaolo, Concetta. "Statistic Applets For Teaching Topics In Introductory Courses". *http://sapphire.indstate.edu/~stat-attic/index.php*. N.p., 2010. Web. 7 July 2016.
7. Doi, Jimmy; Potter, Gail; Wong, Jimmy; Alcaraz, Irvin; & Chi, Peter. (2016). Web Application Teaching Tools for Statistics Using R and Shiny. *Technology Innovations in Statistics Education*, 9(1). uclastat_cts_tise_27492. Retrieved from: http://escholarship.org/uc/item/00d4q8cp.
8. Garfield, J. (1995). How students learn statistics. *Int. Stat. Rev*., 63, 25–34.
9. Gijbels, D., van de Watering, G., Dochy, F., & van Den Bossche, P. (2006). New Learning Environments and Constructivism: The Students’ Perspective. Instructional Science, 34 (3), 213-226. doi:10.1007/s11251-005-3347-z
10. De Jong, T., Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research*,*68*(2), 179-201.
11. Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 46, 75-86.
12. Krueger, R. A. (1994). Focus groups: A practical guide for applied research (2nd ed.). Thousand Oaks, CA: Sage.
13. Lane, D. M. (1999). The rice virtual lab in statistics. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 24-33. http://onlinestatbook.com/rvls.html. Accessed 04.07.2016.
14. Lane, D. M., Peres, S. C. (2006). Interactive simulations in the teaching of statistics: Promise and pitfalls. In *Proceedings of the Seventh International Conference on Teaching Statistics*.
15. Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for qualitative data analysis triangulation. *School Psychology Quarterly*, 22, 557–58
16. Mills, J. D. (2002). Using computer simulation methods to teach statistics: a review of the literature. *Journal of Statistics Education,* 10(1). https://ww2.amstat.org/publications/jse/v10n1/mills.html
17. Muller, J. Bewes, M. Sharma, P. Reimann (2007). Saying the wrong thing: improving learning with multimedia by including misconceptions. *J COMPUT ASSIST LEAR*. 24: p. 144–155.
18. Mvududu, N. (2005). Constructivism in the statistics classroom: From theory to practice. *Teaching statistics*, 27(2), 49-54.

19. Onwuegbuzie, A.J., Dickinson, W.B., Leech, N.L. and Zoran, A.G. (2009). A qualitative framework for collecting and analyzing data in focus group research. *International Journal of Qualitative Methods*, 8(3), 1-21.
20. Post, J. (2016). Interactive Math Stat Visualizations Using R Shiny. Poster presentation at eCOTS, May 11 2016. https://www.causeweb.org/cause/ecots/ecots16/posters/c/8. Accessed 12.07.2016.
21. Prince, M. & Felder, R.M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education* 95 (2): 123–38.
22. Rossman, A. J. and Chance, B. L. (2014), Using simulation-based inference for learning introductory statistics. WIREs Comput Stat, 6: 211–221. doi:10.1002/wics.1302.
23. RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URLhttp://www.rstudio.com/.
24. Rstudio (2016). Shiny User Showcase. https://www.rstudio.com/products/shiny/shiny-user-showcase/. Accessed 19.11.2016.
25. Rutten, N. Van Joolingen, W.R. & van der Venn J.T. (2012). The learning effects of computer simulations in science education. *Computers & Education,* 58, 136–153
26. Show Me Shiny. Gallery of R Web Apps. http://www.showmeshiny.com/. Accessed 19.11.2016.
27. Tishkovskaya, S., Lancaster, G. A. (2012). Statistical education in the 21st century: a review of challenges, teaching innovations and strategies for reform.*Journal of Statistics Education*, 20(2), 1-55.
28. Zieffler, A., Garfield, J., delMas, R., Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 40-58.

Appendix A. List of applications

Name	Description
pcondic	Venn Diagram, probabilities of events, conditional probability
probco	Conditional probability
VPs	Predictive values, sensitivity, specificity
XYZ	Frequency tables for three dichotomous variables; confusion and interaction; odds ratio with confidence interval
efectos	A version of “VPs” focused on the elements that make science reliable
vardisc	Discrete random variables through dice; indicators; simulation, and a step towards the law of large numbers
VAC	Continuous random variables; modelling; simulation, and reference to parameter estimation
corre	Correlation of two discrete random variables; independence of variables; conditional distributions
bombas	Poisson model; a tribute to R.D. Clarke, and his study on the London bombing during WWII
binopois	Comparison of Poisson and Binomial models
ROC	ROC curves for both Normal and discrete responses
estima	Estimator properties, based on four different estimators for the maximum of a variable; analytical and simulated solutions
distmed	Distribution of the sample mean; simulation; Central Limit Theorem
polling	Random and nonrandom samples; estimation of a proportion with confidence intervals
mnas	Nonrandom sampling; sample proportion with no independent data; loss of confidence in confidence intervals; simulation
pares	Paired samples; testing the difference of means by the mean difference; correlation
twomeans	Two independent samples; two-mean test; simulation; standard error; one-side and two-side testing
regre	Parameters of the linear model; the population line and the regression line; conditional distribution of response
a-ojo	Least squared residuals method; line estimation
modlin	Premises of the linear model; distribution of the statistics; checking the violation of the premise; simulation; type I error risk

BLUE	Ordinary Least Squares; Least Absolute Deviations; properties of both estimators: bias and efficiency; linear estimator; simulation
anova	One-way analysis of Variance; variability within and between the group; sum of squares simulation; testing the equality of means
balanzas	Overdetermined system; increasing accuracy of measures
hangman	As its name suggests, just to play (yes, with R!)

Appendix B. Example of an application

The “anova” app is presented in the following, as it comprises many of the features we think are worthwhile. It is lively and interactive, since the student must perform actions and also observe the program’s reaction. Furthermore, it relies on intense computations as well, since simulation is an important component in its operation.

The app consists of two parts, which are accessible through their respective tabs. In the “Decomposition of sum of squares” tab, the student can explore the basics of the analysis of the variance, that is: why the study of variability is important for determining whether several groups have different means. In the “Statistical properties” tab, particular attention is given to showing the relationship between the population means of three groups (either equal or different) and the distribution of the F statistic.

The first approach starts with the input of data. Instead of using real data or generating it at random, we let the student enter *pseudo-observations* by clicking with the mouse in an initially empty diagram (see Figure B1, left). We account for just three groups to be compared, and three shades of grey are employed to differentiate between them (Figure B1, right). The points are shifted horizontally to improve their visibility in case of coincidence. The sample means $\bar{y}_j$ for each group *j* are displayed below the diagram and are also visible as dashed lines. A thick, solid line represents the global (or grand) mean $\bar{Y}$.

Whenever the three groups have data and an estimate of the within group variance is possible (that is, when there is more than one observation of some group), the app shows a graphic with a representation of the decomposition of sum of squares, also known as *SS*. The student chooses between two options: the simplest form is a piechart (Figure B2, left), where six portions are drawn. Three are blue-colored, tagged with numbers 1 to 3, and represent the deviation of each group mean with respect to the global mean. The three orange-colored portions are also tagged with numbers 1 to 3, and they represent the amount of deviation of the individual observations with respect to the group mean. Thus, the cold colors are related to variability explained by the group (deterministic), and the warm colors are related to residual variability (differences by chance). In either case, the intensity of colors is related to the intensity of grey shade in the diagram, in order to simplify the identification of the group.

The second option for the chart is called *puzzle* because of its appearance (Figure B2, right). It looks more or less like a jigsaw with its pieces in place, though separated from each other, as if the jigsaw were exploding. The pieces —rectangles, actually— are also blue- and orange-colored, and the group is indicated by the intensity in color, as in the piechart. The height-width ratio of the rectangles is irrelevant. Their main attribute is the surface, which is proportional to the amount of variability the topic provides: for the *j*-th blue piece, it comes from:

$$SS_{Ej} = n_j\left(\bar{y}_j - \bar{Y}\right)^2,$$

as does the corresponding portion in the piechart. However, the puzzle chart also shows the decomposition of each orange portion in the piechart, with as many rectangles as observations in the group, and the contribution of observation *i* from the group *j* is:

$$SS_{Ri,j} = \left(y_{i,j} - \bar{y}_j\right)^2.$$

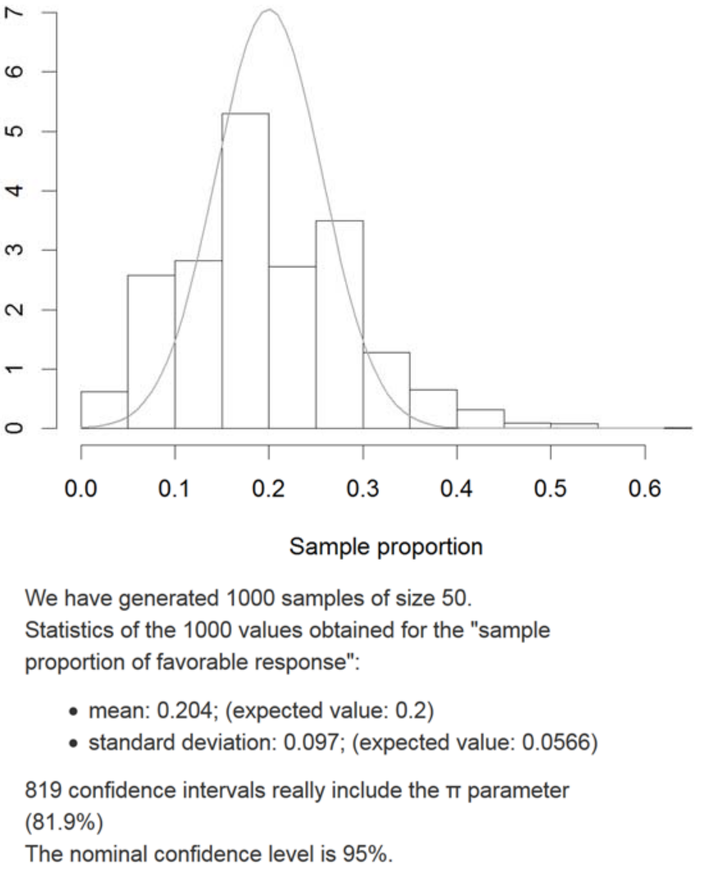
At the right side of the page, a classical ANOVA table is filled in with the information that was entered into the diagram: that is, the decomposition of variability into either explained or residual terms, so that the student has a view of the same information in both visual and numerical format.

The student can be aware of the magnitudes in action by clicking on any of the puzzle pieces. If she chooses a blue rectangle, she will see in the diagram a double arrow line going from the corresponding group mean to the global mean. And if she chooses an orange rectangle, the line is drawn from the chosen observation to its group mean. The association relating large blue pieces with large differences between groups, or large orange pieces with small differences between groups, is easy to establish. This is still clearer if the student enhances or reduces the differences among means by strategically adding more points, since the changes are automatically translated to the charts. The app does not allow erasing individual points, but a button allows restarting the experience by removing all the points.

The second tab addresses the statistical properties of the ANOVA test, so it is important to remind the student about the numerical nature of the decomposition. In the absence of differences, we deliberately skip the rationale about the two estimators (MS_E and MS_R) of the residual variance, and the student is invited to set the parameters: total sample size, means and residual standard deviation σ. As before, three groups are considered, and the application assigns at random the group for each observation, and it generates independent, Normally distributed values according to the previous parameters. With the simulated data, the MS_E, MS_R and their ratio F are computed and stored. This process is repeated 1,000 times. Finally, the empirical distributions of MS_E, MS_R and F (Ratio) are graphically represented.

Figure B3 shows an instance obtained with: sample size = 40 (randomly divided into the groups); expected means equal to 10, 10 and 12.5 units, respectively; and standard deviation = 4 units. The expected value is σ^2 for MS_E (Between MS, top left) and MS_R (Within MS, bottom left), both indicating σ^2 with a purple dashed line. One assumes that each group has the same expected value, but the student can observe that the hypothesis does not hold for the distribution of MS_R, and only MS_E is sensitive to it. Indeed, the sample mean is also represented with a green solid line; in this case, it is particularly clear that the between-group sums of squares are larger than expected. This effect also translates into the ratio (right), which shows the curve of the expected distribution of F and the Fisher-Snedecor probability distribution with 2 and 37 degrees of freedom, which categorically does not fit the empirical results.

Considerations about the power of the test are left out, though the picture may be used as a basis to start a discussion in the classroom. For instance, a lecturer who uses the Figure B3 as a basis could ask the students for an estimate of the proportion of times the test would fail to detect the actual differences between the means. From the histogram of ratios, they could determine that more than half of the tests fall under the 95% quantile of the distribution; so, from a Neyman-Pearson perspective, these could not reject the identity of means.



We have generated 1000 samples of size 50.
Statistics of the 1000 values obtained for the "sample proportion of favorable response":

- mean: 0.204; (expected value: 0.2)
- standard deviation: 0.097; (expected value: 0.0566)

819 confidence intervals really include the π parameter (81.9%)
The nominal confidence level is 95%.

Figure 1. A fragment from the “mnas” app output: the response was generated with affinity parameter = 0.5 (positive autocorrelation).

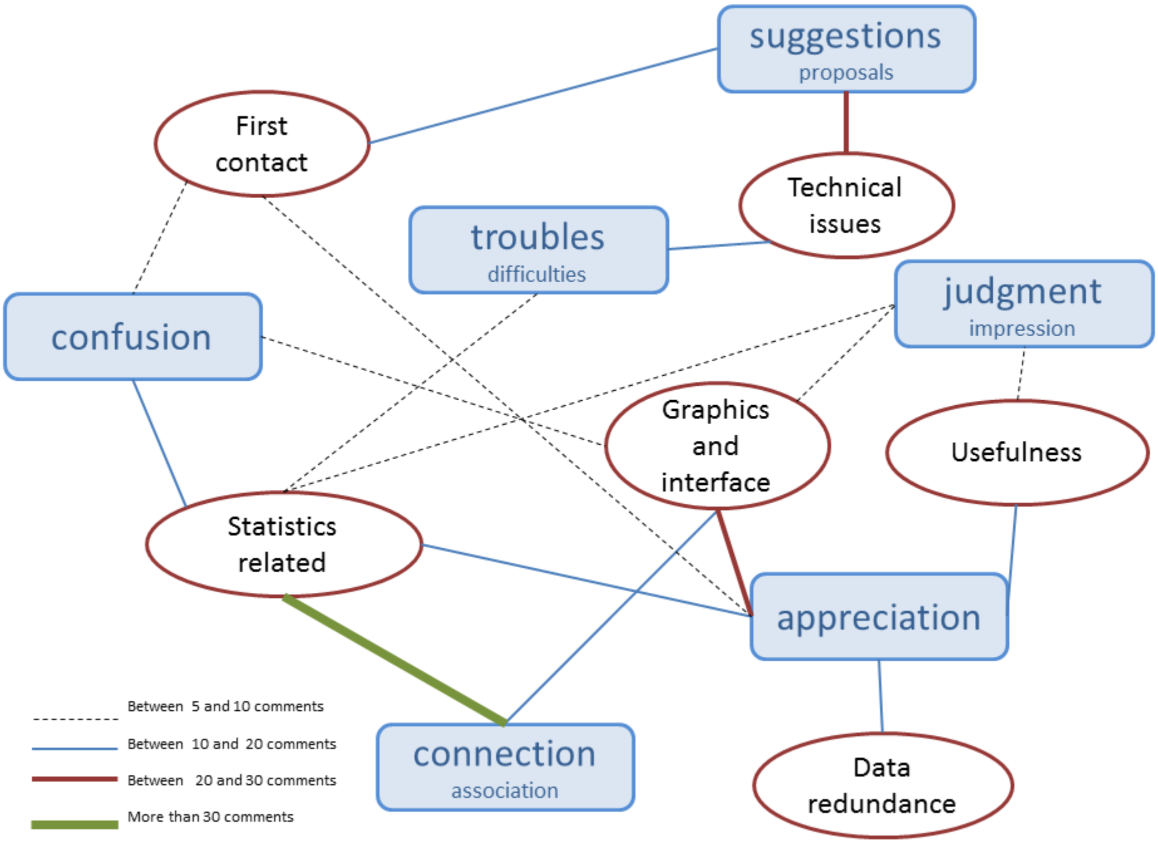
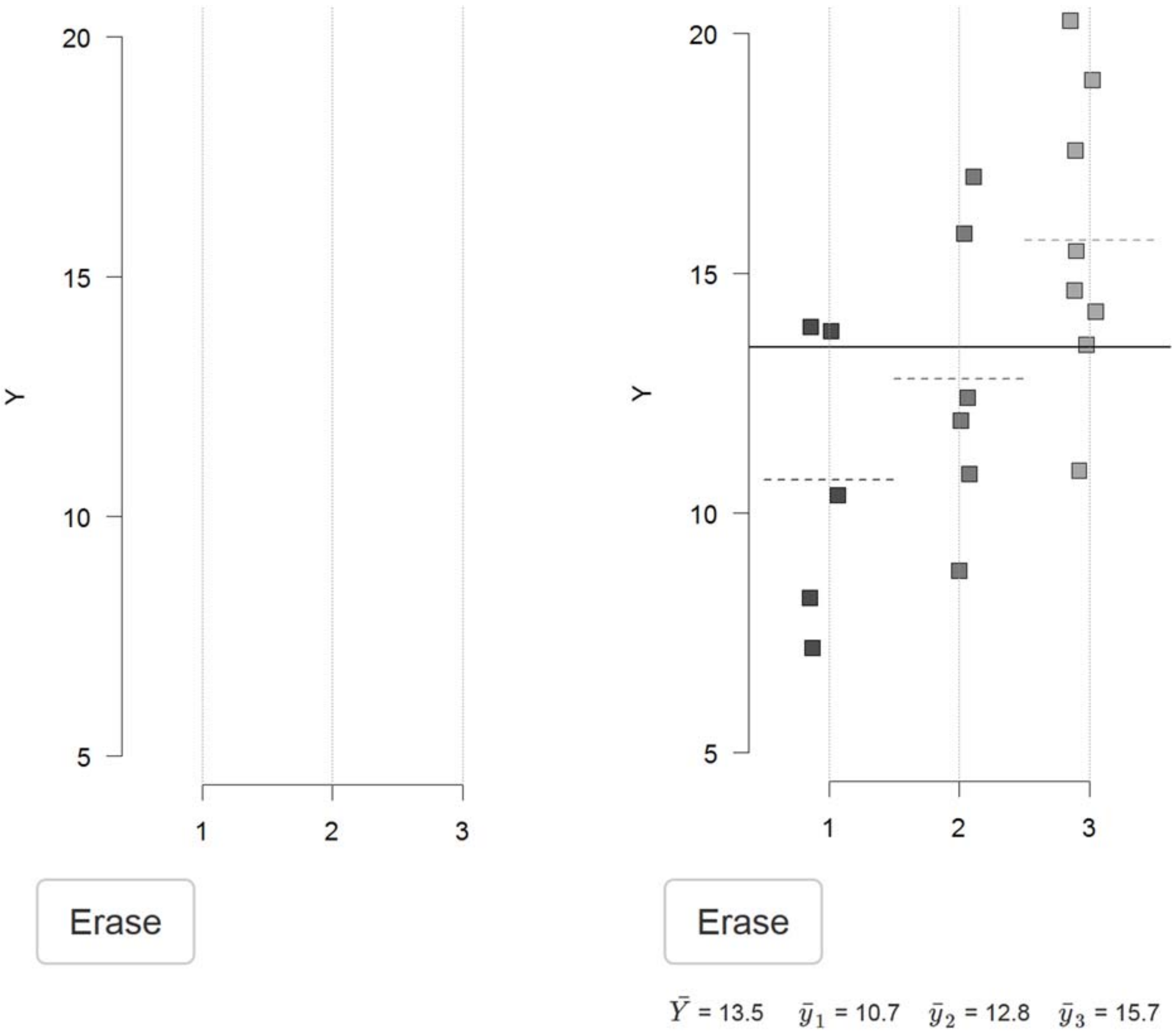
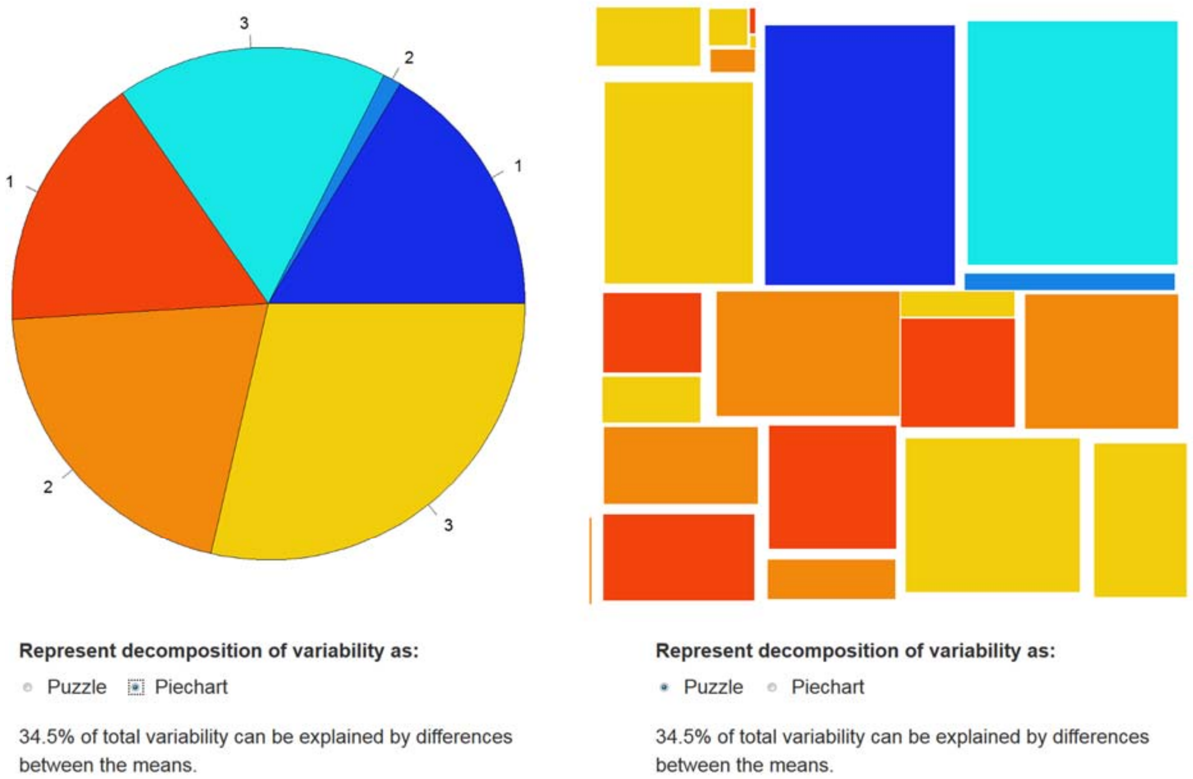


Figure 2. Inside oval labels, codes for Topic classification of the students' comments and answers in questionnaires. Inside rectangular labels, codes for Attitudes for the same comments and answers. The authors have determined the position of the labels after considering the neighboring relationships between the topics, expressed in Table 2. The labels have been connected according to the number of common comments.



$\bar{Y}$ = 13.5 $\bar{y}_1$ = 10.7 $\bar{y}_2$ = 12.8 $\bar{y}_3$ = 15.7

Figure B1. Left, the diagram before the student begins setting data in the “anova” app; right, the diagram after the student has clicked some purported observations.



Represent decomposition of variability as:

○ Puzzle ● Piechart

34.5% of total variability can be explained by differences between the means.

Represent decomposition of variability as:

● Puzzle ○ Piechart

34.5% of total variability can be explained by differences between the means.

Figure B2. Piechart (left) and puzzle chart (right) representing the decomposition of variability in the "anova" app

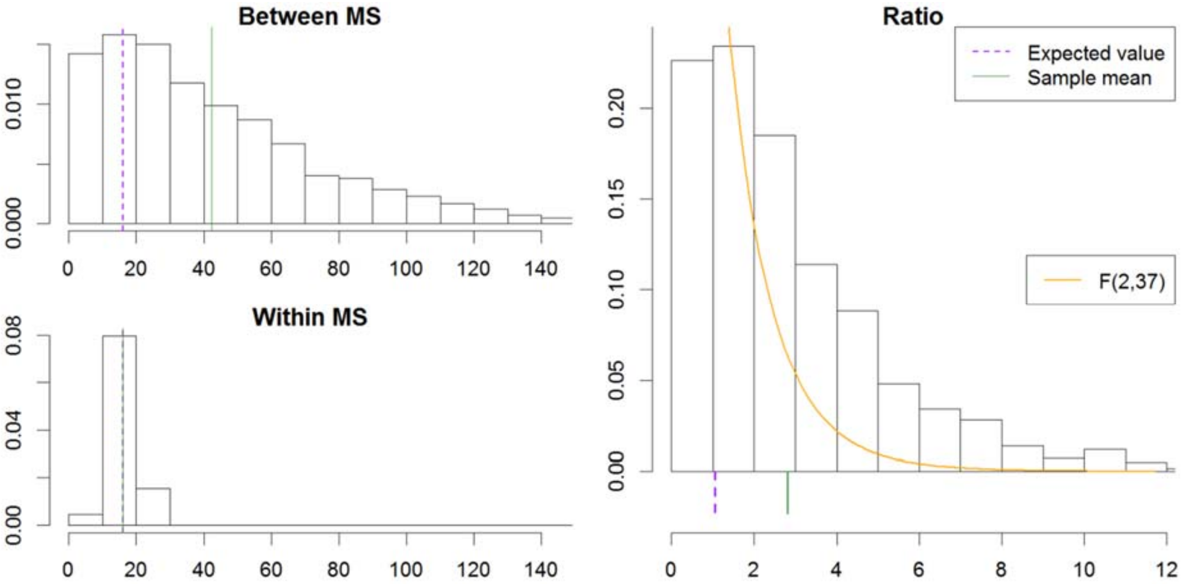


Figure B3. A snapshot from the “anova” app with the empirical distributions of MSE, MSR and F after a simulation.

Table 1 - Data analysis example

Block – Number	Statements	Codification Topic	Codification Attitudes
B2 – 5	Exercise has also been very hard for me to understand (...) but in the end I ended up interpreting that what it was asked was that everything must be positive and that the area must be 1	*First contact*	*troubles / difficulties*
B2 – 6	I don't completely understand the task	*First contact*	*troubles / difficulties*
B2 – 14	I did not understand very well what it was	*First contact*	*troubles / difficulties*
Q5b - 8	At first, it was difficult for me to understand the graphics. And the variation of the difference between averages has helped me very much to understand it	*First contact* *Graphics and interface* *Statistics related*	*troubles / difficulties;* *appreciation*

Table 2 - Classification of the statements from the students in the data analysis

		Attitudes						
		confusion	*troubles* *difficulties*	*suggestions* *proposals*	*judgment* *impression*	*appreciation*	*connection* *sssociation*	*others*
Topics	*First contact*	6	4	13	0	7	0	0
	Technical issues	3	17	20	1	5	0	2
	Graphics and interface	8	5	5	9	29	12	0
	Data redundance	2	0	5	5	10	3	0
	Statistics related	10	9	2	6	19	37	1
	Usefulness	0	0	1	7	14	5	0
	Others	1	1	0	2	4	0	0