New Technologies revolutionize the applications of statistics and its teaching

Borovcnik, Manfred
University of Klagenfurt, Dept. of Statistics
Universitätsstraße 65-67
9020 Klagenfurt, Austria
E-mail: manfred.borovcnik@uni-klu.ac.at

Summary

The impact of New Technologies may be seen from three perspectives:

Firstly, it changes the applications of statistics outside the universities completely. A lot more statistics is applied nowadays than 25 years ago – i. e. before the computers have spread to nearly every household in the western world. A lot of large-scale datasets, all multivariate, are nearly automatically generated and wait for their analysis. New methods, more methods, and more variety in used methods signify the applications of statistics. This builds up a pressure from outside universities to change our curricula completely so that our graduates will qualify for their future work.

Secondly, New Technologies open the way to teach the subject matter of statistics completely different than in the pre-computer era. Films may be distributed on DVDs and watched on one's own PC. Software like R is free of cost and easily installed to avoid tedious calculations. Spreadsheets are often integrated on the basic equipment of PCs and allow e. g. teachers to design their own animations to illustrate a difficult concept, which would have cost a lot of time and a production studio for animated cartoons in the pre-computer age.

Thirdly, the computer enables (but also necessitates) new forms of communication. While a small team of students (2 or 3) work together in front of the PC, the teacher gets more and more into the role of an advisor. Students learn from sources from the computer (the used software, a hypertext, or the Internet) and the teacher intervenes only on request, or to advise the students about their potential strategies to find their way through.

Not that we have to re-invent all of teaching statistics but we have to adapt a lot in our curricula to meet the challenges of the New Technologies. The reward will be far more complex knowledge, which our students will be able to master making them fit for their demanding jobs in the future.

The change in the applications

Only two examples are discussed to give an impression how applications of statistics have changed.

In order to classify rocks for their presumptive characteristics (e. g. abrasiveness, brittleness, and compressive resistance) the rocks are inspected in the quarry by infrared spectroscopy. The results are curves of absorption of light for different wavelengths for each rock. The problem is to define a fast procedure to separate the rocks on the conveyor by a classification procedure. The procedure has to operate online while the rocks are passing by. The solution does not use the direct data but the Fourier series of the curves and gives quite good attributions with respect to the quality of the rocks. Very often in applications, the original data have to be transformed, here to Fourier coefficients, in order to solve the task. It is not sufficient to know classification methods from statistical point of view, it is also necessary to combine it with other complex mathematical methods.

To exploit the stock of clients to the best of a company and to hold one's own stockholding at a minimum, or produce at the right time, it is necessary to study the clients from their past behavior and from a (short) investigation of them by a survey. The aim is to find a segmentation of clients that will be a basis for predicting their future behavior, which would enable oneself to prepare in time to act instead of reacting to their demands. This necessitates dealing with multivariate data with different methods of clustering and classification methods and so on. One statistical method alone will not suffice to deal with the problem adequately. Only in comparison of results of different methods will it be possible to give reliable predictions. As a special case of the previous, consider online-shops, which have sophisticated procedures how to react 'online' to requests of customers. This necessitates to collect automatically their past data and to analyze them accordingly. It also requires a conceptual hierarchy on the goods they offer and to exploit conceptual distance between goods in order to supply the client target-oriented with more information, which might lead to further purchases. – This approach is similar to the organizational memory in supporting learning, as is presented by Chaput (2007).

Clearly, if our students will leave the university with an introductory course mainly focusing on normal distribution and one-dimensional data analysis, they will not be able to do their work in their jobs. We have to teach them authentic applications also. However, in that endeavor we will come to our end (we will not be able to understand all the details of such applications on our own) and we will overburden our students so that finally they would not be able to understand anything. To teach applications it is necessary to develop new styles of teaching mathematics without all the details yet enabling the students to grasp the main structure of the methods, its important pre-assumptions for application, and the statements, which they enable, with a proper interpretation of its potential and its restrictions. Mougeot (2007) in her presentation gives a detailed discussion on the pressure of applications outside university and how to react accordingly in university courses to meet the challenge of the statistics revolution in the applications.

New Technologies open new ways to teaching statistics

Films on DVDs are easily spread. No further technical devices are needed in order to look at e. g. Moore's "Against all Odds". Such films contribute to students' motivation as they authentically show that and how statistics is applied.

Software may be used to avoid lengthy calculations. This might not seem attractive enough but if you remember the difficulties students had to calculate Binomial, or normal probabilities, or a confidence interval for a mean, it helps a lot to get the message through. Moreover, software enables also to apply more complex statistical methods, e. g. classification methods for multivariate data. There is, however, the danger of blindly applying the software without really understanding the results, their restricted interpretation and the restrictions necessary to fulfill in case of applying it. However, the time saved by automatic calculation could be used to focus teaching efforts on the conceptual issues in the background.

Software may be used to demonstrate the impact of more complex methods dynamically. This may start from as simple an example to illustrate the robustness of the median as compared to the mean only by moving one point of the data, implemented in Excel by Borovcnik (2007). And come to more complex examples as showing the recursive adjustment of a separation line for discriminating between two sets of data as shown in the presentation of Mougeot (2007), programmed in R.

Software may be used to complete an analysis of data by graphs to communicate the results better (also done in the course, which was discussed by Mougeot 2007). To give boxplots in amendment to an analysis of variance might likewise help as profile curves for the mean in the various groups might help to support an understanding of the results, which are formally backed by the p values of the F statistics telling us, whether one or the other factor is significant (or if there are significant interactions).

Software might enable us to teach authentic applications holding technicalities at a minimum level and letting our students to focus on the subject content and the questions therein. In a systems analysis of the problem at hand one has to model the situation accordingly so that the later experiment will produce valid data on the problem. In first stages of analysis, data cleaning, missing values and the like has to be dealt with (again Mougeot 2007 gives a plea for such an approach in her presentation). Only then one can revert to statistical models and analysis, which at the end have to be translated back to the original problem to answer (partially) the posed questions with all the restrictions of the interpretation in mind.

Software might enable us to illustrate statistical concepts more easily than it could be done in the mathematical framework alone. That comprises not only graphs to study the effects of the concept under investigation; it includes also the powerful tool of simulation. By this technique not only unknown probabilities may be calculated (which otherwise involves too complex mathematics; sometimes not only for students too difficult). It also helps to understand basic features of methods: For example, the influence of the sample size on statistics to estimate an unknown parameter may easily be shown by simulating the process of taking a sample repeatedly and end up with a graphical representation of the estimating statistics. If this is compared for two different sample sizes, it may easily be seen that increased size leads to higher precision, i. e. smaller variance of the statistic at hand. Likewise it could be illustrated what the method of least squares does, what the maximum likelihood approach involves etc.

Software might also let us introduce newer approaches into teaching like the resampling approach. Instead of sampling from the true population, which is represented either physically or by probability hypotheses (e. g. a normal distribution with specified mean and standard deviation); sampling could be done from a distribution, which is estimated on the basis of the first sample. Repeated sampling from the first dataset (which establishes this estimation) with replacement and calculating the value of the required statistic (e. g. the mean) leads very fast to an impression about the "reliability" of the first estimate of the unknown parameter (here the mean of the population). The variability of this reapplied estimation (of the mean) in repeating to take a re-sample is a sign how precise the estimation from the first sample could be. From a long series of re-samples the distribution of the estimating statistics is well approximated and a confidence interval might be taken as the 2.5% and 97.5% quantiles of this distribution. The method of resampling was caricatured as Munchhausen's strategy to pull oneself out of the swamps by pulling himself on his hair – but its rationale is sound. Without the help of computers it cannot be performed. This author takes great hopes to 'renew' introductory statistics by the resampling approach as it is conceptually much easier than any procedure on inferential statistics (cf. Howell 2007, Christie 2004, or Arnholt 2007).

New technologies open and necessitate new forms of communication in teaching

The presentation of Chaput (2007) deals with e-Learning and how it can be made more effective by introducing a hierarchical system of coherence upon the notions, which have to be learned. If knowledge, as in e-Learning is done, is presented on the computer in a hypertext then the student has to go his/her own way through. The information could be presented as usual in a linear sequencing, which is quite ineffective for self-study. The hierarchy of notions, however, can be re-analyzed in this context to yield a "neural network" between them, characterized by conceptual distance. This embodiment could help learners to "browse" one's way through until finding the concepts required for solving a specific problem. Chaput (2007) presents details of how her group has designed an organizational memory to help the learning path to become more efficient. E-Learning is the high-end of computerized teaching and necessitates the most sophisticated infrastructure to supply learning to become effective. The interesting part is that the same organizational memory can also enhance learning in a traditional course as it yields another structure of relations between the concepts.

The computer may change the ways and the kind of communication completely even without the option of e-Learning. Imagine groups of two or three students working together in front of their PCs, the teacher organizing their work becomes more and more an advisor, what methods could be taken, how to find the required information, giving advice, which analysis to perform, giving some caveats about restrictions of results and their interpretation etc. The students would get their information about the concepts from a hypertext, or from an e-Learning text, and work independently. Intermittently the teacher could assemble their approaches and introduce a classroom discussion by showing them where their different approaches would lead to. To follow-up the students' work, the teacher could use screenshots supported by audiotapes of their comments etc.

Conclusion

Applications have enforced the intensified implementation of calculator devices, more recently this means mainly software used on computers. These devices have taken their own development and revolutionized not only the applications but also the theory of statistics. With the advent of EDA, a new and innovative way of interactive data analysis was borne, which was no more inextricably bound to probability. With the resampling approach, the fundamentals of probability are not questioned. However, the inferential part may be written completely from anew. The impact of applications will thus completely alter the image and the vision of statistics as it will radically decrease the key role of parametric models (among which the normal distribution is the most important) by replacing it through suitable re-sampling of the first data set. Such an approach may easily be extended to multivariate data.

For teaching statistics it is interesting to note that the New Technologies have enabled to take the role of applications serious. Prior to the spread of computers, applications could be integrated only in an artificial way, mainly in the cookbook variant. By software, tedious calculations are outsourced, which saves time to focus on conceptual issues. That brings the modeling aspects of applications to the fore: How to model the situation adequately, which restrictions are inherent to the various models, how to get valid data to answer the questions, how to interpret the used methods and the results in a feasible way? The design of a study already bears in it if the study could be a success or will definitely be a complete flop.

Here the circle closes between the impact from applications and from New Technologies and new approaches towards teaching. Admittedly, matters are not always clear and easy. Any software necessitates time to master it. Concerning didactical software for teaching to illustrate key features of concepts there are many a wish open – despite a good tradition and practice already going back to Kissane (1981); maybe a solution lies in flexible development of such demonstration pieces by the teacher and the students with tools like R (Arnholt 2007), Fathom (Maxara and Biehler 2007), or EXCEL (Borovcnik 2007). Maybe didactical analyses should focus on the design; i. e. what a good demonstration software should include for specific notions; and didactical research should concentrate on a guided implementation and monitoring of the success of such software pieces to support learning. For teaching authentic applications, the New Technologies have proven to be indispensable tools; they have revolutionized the teaching of statistics.

REFERENCES

Arnholt, A. T.: 2007 Resampling with R. *Teaching Statistics* 29 (1), 21–26.

Borovcnik, M.: 2007, Das Sammelbildproblem – Rosinen und Semmeln und Verwandtes: Eine rekursive Lösung mit Irrfahrten. *Stochastik in der Schule* 27 (2), 19-24.

Borovcnik, M.: 2007, On Outliers, Statistical Risks, and Statistical Inference. *WG 'Stochastical Thinking',* CERME 5, Larnaka.

Borovcnik, M.: 2007, The influence of software support on stochastics courses. *Contributed Paper Session*, 56. Session of the ISI Lisboa.

Chaput, B.: 2007, An Organizational Memory as Support of Learning in Applied Mathematics. *Invited Paper Session 38: How modern technologies have changed the curriculum in introductory courses*, ISI Lisboa.

Christie, D. 2004: Resampling with EXCEL. *Teaching Statistics* 26 (1), 9-14. German translation in *Stochastik in der Schule* 24(2004)3, 22-27.

Howell, D. C.: 2007, Resampling Statistics: Randomization and the Bootstrap. *Statistical Home Page*. http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.

Kissane, B.: 1981, Activities in Inferential Statistics, in A. P. Shulte and J. R. Smart, *Teaching Statistics and Probability*, NCTM, Reston, Virginia, 182-193.

Maxara, C. and Biehler, R.: 2007, Constructing simulations with a computer tool – students' competencies and difficulties. *WG 'Stochastical Thinking',* CERME 5, Larnaka.

Mougeot, M.: 2007, Impact of New Technologies for Teaching Statistics. *Invited Paper Session 38: How modern technologies have changed the curriculum in introductory courses*, 56. Session of the ISI Lisboa.