Survival Analysis by Students

Wim van den Camp and André Heck

AMSTEL Institute, University of Amsterdam

Kruislaan 404, 1098 SM Amsterdam, The Netherlands

wvdcamp@science.uva.nl, heck@science.uva.nl

Dutch textbooks about statistics and data analysis at secondary school level are filled up with small, non-realistic examples. In this way students get the wrong impression that data handling is in fact nothing more than applying some standard recipe. They are not confronted with questions such as "Do the data make sense?", "What statistical model seems applicable and for what reason?", "Do you need more data for making sound conclusions?", and "How do you deal with large data sets?". This lack of focus on how to manipulate data meaningfully, using ICT tools, becomes a serious handicap for students when they explore data in a research project.

Statistical reasoning and thinking are crucial when data are incomplete. The handling of so-called censored observations is one of the main issues in survival analysis. We investigated whether this extracurricular topic provides a nice opportunity for students to apply their current knowledge of statistics to a real world problem and to get acquainted with a modern, much used statistical application. We created learning materials in which students must up with their own models for handling clinical data in a hospital study. The students make use of a spreadsheet program to compute survival probabilities so that they can avoid laborious tabular work and can concentrate on model construction and model documentation. In this paper we shall present the learning materials that have been tested in practice, and we shall discuss our classroom experiences from our research perspective.

Introduction

Probability and statistics was introduced in the Dutch mathematics curriculum in the late seventies. It was then seen as a good illustration of how mathematics is applied in practice and it was also considered as a good preparation for students in economic and social sciences. Nowadays, probability and statistics is a topic taught in all of the four fixed subject combinations from which a student must choose one at senior level. Initially students carried out chance experiments and ran simulations, but gradually these didactical instruments disappeared from the textbooks and exploring data by students in the classroom vanished [1]. Situations presented in current textbooks, strongly suggest that the way of handling data, drawing conclusions and predicting, are quite rigid as well as widely accepted. Textbook authors mostly avoid situations where students have to draw their own conclusions, based on reasonable assumptions of how to handle the data. Besides, most contexts do not suggest a strong connection to the real world. In short, the development of algorithmic-procedural skills is overemphasized in probability and statistics education and little attention is paid to modelling and interpretive skills. Since the last Dutch curriculum reform in 1999, the required use of ICT in the form of a spreadsheet program or graphic calculator, the increasing interest of educational researchers in mathematical reasoning and in mathematical thinking, and the fact that Dutch students must build up an examination portfolio of small

investigation tasks and a research or design project may revitalize doing simulations and statistical computations. More attention may be given to the process of mathematical modelling and to solving more realistic problems. With Internet, students can collect data from public databases in the context of their practical investigation work.

In this paper we shall discuss a practical investigation task in which students compute survival probabilities for patients treated after a cardiac arrest. The data set comes from a clinical trial of patients. It consists of censored observations, which means that some of the data are not available during the whole period of the trial. Interpreting such data and drawing conclusions involves assumptions on censoring. We let the students investigate various points of view and invite them to come up with alternatives. They use a spreadsheet program as a calculating tool so that they can focus on statistical reasoning, i.e., on making sense of statistical information and on interpreting results, instead of on filling out tables. We hope that statistical thinking, which involves understanding of why and how statistical investigations are conducted and the "big ideas" that underlie statistical investigations, becomes an attainable learning goal in this way.

The students' investigation task has been developed in our educational research programme that examines the possible contribution of ICT and real life contexts to the realization of challenging mathematical investigation tasks for students. This work fits in the reform view that probability and statistics education should be organized around critical inquiry rather than mastery of algorithmic-procedural skills and that activity-based instruction is an effective way to let students get personal experiences [2,3]. We hope to present a stimulating example of an investigation task that on the one hand is understandable for students who are not confident about their mathematical competence, and that on the other hand is challenging for these students and gives them an impression of how statistics is applied nowadays.

The Learning Materials

The learning materials have been designed for students who are in their penultimate year of pre-university education (age 16-17 yr.) and have knowledge in statistics, but no experience with survival analysis. Our main objectives are to let students

- work meaningfully with real data coming from a clinical trial on treatment after a cardiac arrest;
- experience that statistical work is more than applying some cookbook recipe;
- look upon a statistical problem from various points of view and comment on methods;
- develop their own statistical models, explain their choices, and draw conclusions;
- learn and practise skills with respect to data manipulation in the spreadsheet program Excel.

The learning materials [4] consist of an adaptation of a text from a continuing education course for mathematics teachers [5] to a lesson text at student level and supplementary materials. Although it

looks similar to an ordinary mathematics textbook chapter, the students' activities fulfil the characteristics of open-ended tasks that stimulate students to think critically and require from them explanation of the chosen strategy(ies) at each step. There is scarcely a closed question with a single correct answer. The problem activity is based on a real world situation in order to capture students' interest and to make it relevant and convincing to the students. The students' text consists of the following four parts:

1. *The problem description.* Students are briefly introduced to survival analysis and censored observations in the context of a clinical trial of patients treated after a myocardial infarction. Censoring means that because of external causes it is not possible to get the desired information during the whole time of research from all the patients. For instance, while studying survival time after cardiac arrest, some patients might die from another cause or stop their participation to the clinical trial. In such instances it is not clear what to do with this patient information. The students work during most of the investigation task with the data shown below:

Starting point is a clinical trial of 146 patients after cardiac arrest. These 146 patients were kept track of during the next 10 years, **as far as this was possible.** The table below lists the data known about the patients:

A	B	C	D	E	F
year since entry into the study	alive at beginning of the year	number dying during the year	cumulative number of deaths	number of censored persons during the year	cumulative number of persons censored
1	146	27	27	3	3
2	116	18	45	10	13
3	88	21	66	10	23
4	57	9	75	3	26
5	45	1	76	3	29
6	41	2	78	11	40
7	28	3	81	5	45
8	20	1	82	8	53
9	11	2	84	1	54
10	8	2	86	6	60

Column E: the number of patients whose situation was no longer kept track of from that year on, i.e., the number of censored individuals.

Table 1 (from the students' text): Data from a clinical trial on myocardial infarction.

2. *Examples from 'familiar' statistics.* We use the classical gambler's ruin example of throwing a coin repeatedly until 'heads' comes up to let students recall the systematic method of computing probabilities of events through a tree structure of chances and to let them practise the statistical terminology and notation. The students apply survival analysis techniques to the gambler's ruin example and to some invented data of a clinical trial in which censoring of observations does not yet play a role. Students who feel not confident about their spreadsheet skills are invited to practise first with a supplementary tutorial note on working with Excel.

3. *Handling survival data with censoring; simple ways to estimate survival probabilities.* The data of the clinical trial with censored observations prevent the use of standard methods of statistical

summarization and inference already known to the students. The concrete problem in this part of the learning material is to estimate the five year survival probability of a patient on the basis of the given survival data. The students are confronted with three rather naive points of view:

(P) "Eliminate all individuals who are censored and use the remaining 'complete' data."

(Q) "Eliminate all persons who are censored in the first 5 years and use the remaining data."

(R) "Consider all persons who are censored in the first 5 years as survivors and count them as such, i.e., treat them as if they did not die from heart disease".

The estimated probability of surviving the first 5 years after treatment of the heart failure would range from overly pessimistic to overly optimistic using the above data set and the above points of view (11.6%, 35%, and 48%, respectively). Since these survival probabilities are so different, the importance of refining the points of view becomes quite obvious to the students. They are explicitly asked to reflect on the given points of view, to try to formulate better alternatives, and to calculate the survival probabilities of all methods under consideration. Questions run like "Which of the two points of view is most credible and on what grounds?" and "Using the tabular data, decide which point of view is taken here. How do you get to your conclusions?".

4. *The life-table or actuarial method.* This section gives a more refined technique for computing survival times used by professionals. Depending on how one looks at the censored data, one will still obtain different survival probabilities, but the differences are smaller than before. The two points of view on censoring in the actuarial method presented in the learning material are:

(A) "Anyone censored in a year is immediately censored at the beginning of that year."

(B) "Anyone censored in a year is censored at the end of that year."

Once more, the students are explicitly asked to reflect on these points of view, to try to formulate an alternative, and to substantiate their own point of view.

The Classroom Experiment

The experiment took place in a class of 18 students with the following learner profile: They had chosen the fixed subject combination "Culture and Society", which prepares them for a university study in the humanities. They were not confident about their mathematical competencies. While teachers of disciplines with a social and cultural context were very pleased with the engagement of the students in their lessons, the mathematics teacher had to deal with a passive attitude. We know this because the first author of this paper is the mathematics teacher of this class. The practical assignment was not part of the students' examination portfolio, but it was graded as a regular test in the semester. The students had to report about their work on the basis of their Excel sheets and they had to hand in a diskette with their computer results. Although the learning outcomes were assessed with respect to statistical literacy, reasoning, and thinking, the main criteria for grading the

students' work were: "Did they give a clear account of what they had done in their analysis and why?" and "Did they express a clear statement of their points of view with respect to handling censored observations and of the conclusions drawn?". These criteria were communicated at the beginning of the classroom experiment. The students worked in pairs for three weeks; the estimated workload was 8 hours. Work took place mainly on the computers in the multimedia centre during the regular mathematics lessons (of 45 minutes) with the authors as assistants. The students could also work independently at this location during school hours or at home. We chose this cooperative, student-centred, non-intimidating setting hoping that this would enhance students' statistical reasoning and thinking.

In our research experiment we were interested in the following issues:

- What do students think of the subject, the learning material, and the use of real data?
- Do the learning material and the chosen instructional setting enable students to

    (i) understand the "big ideas" in survival analysis;

    (ii) work meaningfully with statistical data, including formulating their own ideas;

    (iii) acquire the skills and practice to use a spreadsheet program effectively.

The tools used to get answers to these questions were classroom observations, audio and video recordings (also of computer work), a questionnaire, and the reports of the students.

Findings

From the questionnaire and conversations with students we learned that they very much appreciated the use of a spreadsheet program. They felt that they had acquired good skills in working with Excel during the tasks. This seemed valuable to them because they thought it useful for work in other disciplines as well. We were also quite pleased with the role that the computer environment had in this experiment, but for other reasons than the students: we found that it provoked the students to work in an active way and to talk extensively and thoughtfully about the statistical and computational problems involved within their team, with the teacher, and with other groups of students. Significantly, their discussion was almost exclusively on the tasks, a phenomenon the teacher had hardly seen before during the regular paper-and-pencil work in the classroom.

One example may give an impression of the classroom situation and the students' attitude. In one exercise, the students had plotted survival chances of patients against the number of years after treatment. This was not a straight line and they were asked to change the data in such way that a linear graph would appear. Much to our surprise we noticed that many students were unable to perform this seemingly simple task. First of all, they did not understand the question and secondly they did not consider a trial and error approach as a valid mathematical method. Probably, the concept of linearity and the connection between a straight line and a constant increase diagram had

been discussed too quickly for these students in the past and had never been done in an active way. In pencil-and-paper work, our students would have been confirmed in their belief that they are not good in mathematics, and they would probably have stopped trying to solve the problem. In the lab setting however, suggesting that some students experiment with the numbers in a column was simply enough to get them produce a desired graph (and to let them inform fellow students about such an approach).

But of course we were not satisfied with this outcome-centred approach only; we wanted an explanation and this was still not an easy task for most of the students. For this they had to relate the numbers in one column of the data sheet with those in another column and to formulate an explicit dependency. With some effort most of them succeeded, mainly because the spreadsheet program allowed them to work with formulas in an active way, which is less intimidating than the usual way of working with formulas on paper. The most important difference between spreadsheet and pencil-and-paper work is that the variables in a spreadsheet program do not have a name. Instead, the role of variables is played by cells. The cell itself corresponds to the concept of variable whereas the content of the cell corresponds to the current value of the variable. Combining variables into new values, i.e., creating formulas is done by pointing to the cells containing the values needed. This technique is called the 'gestural description of mathematical formulas' [6]. In other words, the spreadsheet program emphasizes the process character of a formula, i.e., the use of a mathematical formula as a way to describe a process of computing a result. In pencil-and-paper work however, the students' success in dealing with formulas depends to a large extent on how far they have succeeded in learning to look at formulas as mathematical objects. It is known that this shift in focus from process to object character of formulas and functions is difficult for many students [7]. Certainly this holds for our students whose algebraic thinking is weak and full of alternative conceptions. The computer results revealed that they had sometimes found good solutions to the exercises, but that they had not been able to express their answers in terms of formulas.

The active students' attitude had substantial impact on the quality of their results. We might have overlooked this if we had only read their reports because the students' understanding of the subject was better than their written reports suggested. Listening to the conversations amongst the students and discussing the answers with them brought this to the surface. We think that the question whether the learning materials and instructional setting enabled students to understand the 'big ideas' in survival analysis can be answered by "yes". The main grounds for this conclusion are that some students were able to come up with such alternative points of view on censoring which they could only have achieved because they really had a good idea of what they were doing, felt confident enough to come up with original ideas, and knew what survival analysis is all about.

Let us have a look at the solutions students gave for the final problem in which they were asked to compare the actuarial method described in section 3.2 with methods used before in section 3.1 and to reflect upon two points of view of censoring in the actuarial method. These points of view were (A) "censoring at the end of the year", with formula $s(t) = d(t)/n(t)$, and (B) "censoring right at the beginning of the year", with formula $s(t) = d(t)/(n(t) - w(t))$. Here $n(t)$ denotes the number of patients at the beginning of year t still participating in the study, $d(t)$ is the number of patients dying that year, $w(t)$ is the number of persons censored during that year, and $s(t)$ is the probability to die during that year of the study.

Two examples of students' good reasoning, but weak documentation from a mathematical point of view are: "We prefer section 3.2, we think it is better to see survival chances per year, and you also use more data that you would censor in 3.1." and "You could say that the values of section 3.2 are better than those in 3.1. The reason is that the points of view in 3.2 are more clear and logical in the first place. Secondly, the points of view (A) and (B) are derived from viewpoint (Q). In these points of view, the data censored in the year are subtracted and not, like in point of view (P), that none of the censored data are taken into account."

The students also had to formulate a better or intermediate point of view and to express this both in everyday words and in a formula. One team was unable to give a formula related with their point of view. Because of unavoidable interaction between teams, the solution chosen by half of the teams was represented by the formula $s(t) = d(t)/(n(t) - 0.5w(t))$, which is a quite complicated formula for students at this level. In fact, only one team could give this formula in Excel without bracketing errors; the other teams described it in words and actions, and seemingly entered the numerical values one by one in their sheet. One team explained their choice as follows: "Point of view (C) must be exactly in the middle, so we take the formula $s = d/(n - 0,5w)$. This formula means that censored data are subtracted midyear, we assume that in the middle of the year censored data are halfway." Two teams came up with the formula $s(t) = \left(d(t)/n(t) + d(t)/(n(t) - w(t))\right)/2$, which was easier to use in Excel and was also easier to understand as lying between the given points of view. One team came up with a really original idea: they decided to use the average number of censored persons per year, and this average being six persons for the given data set, proposed the formula $s(t) = d(t)/(n(t) - 6)$. Their explanation was: "We think that this is a better solution because it balances between the two extremes of not including some things at all or counting very accurately." This kind of statistical reasoning exceeded all expectations that we had beforehand.

Conclusions

In summary, we are very pleased with the work climate and performance of the students in the classroom experiment. We have never seen them so engaged in their work. The students were interested in the topic of survival analysis. Although the tasks were very different from the textbook exercises that they usually do, they were able to get results. They also enjoyed working with the spreadsheet program because they thought it useful (also for future work) and they could avoid number-crunching. The learning material and the instructional setting let the students overcome their math anxiety, in particular their fear of using formulas and their lack of confidence in their own answers to mathematical questions. They could discuss problems with fellow students or the teacher in a non-intimidating setting. The mental blockade that the students cannot overcome when they work with formulas on paper turned out to be absent when they work with formulas in a spreadsheet program. They were also able to make their own choices for handling censored observations, to explain the methods they had chosen, to compute survival probabilities, and to draw conclusions from the computations. They were less successful in expressing their results algebraically. We may say that they perform better than they reveal in their written reports.

References

[1] Zwaneveld, B. (1999) *Kennisgrafen in het wiskundeonderwijs* [Knowledge Graphs in Mathematics Education]. PhD thesis, Open University of the Netherlands. Chapter 2.

[2] Garfield, J. & Gal, I. (1999) Assessment and Statistics Education: Current Challenges and Directions. *International Statistics Review, 67*, 1-12.

[3] Shaughnessy, J.M, Garfield, J & Greer, B. Data Handling. In: A.J. Bishop et al (Eds.) *International Handbook of Mathematics Education* (pp. 205-237). Dordrecht: Kluwer Academic Publishers..

[4] Camp, van den W. & Heck, A. (2003). Analysee van overlevingsdata [Analysis of Survival Data]. www.science.uva.nl/~heck/research/survivalanalysis.

[5] Borovkova, S. (2002). Analysis of Survival Data. *Nieuw Archief voor Wiskunde* 5/3, nr. 4, 302 – 307.

[6] Neuwirth, E. (1995). Visualizing Formal and Structural Relationships with Spreadsheets. In A. diSessa, C. Hoyles & R. Noss (Eds.). *Computers and Exploratory Learning* (pp. 155-173). Heidelberg: Springer-Verlag.

[7] Tall, D. et al (2001). Symbols and the Bifurcation between Procedural and Conceptual Thinking. *Canadian Journal of Science, Mathematics and Technology Education, 1*(1), 81-104.