A PROBLEM-SOLVING COURSE IN STATISTICS FOR MATHEMATICAL SCIENCE STUDENTS

Danny Parsons[1], Roger Stern[2], David Stern[1]
[1]IDEMS International, UK
[2]University of Reading, UK & Statistics for Sustainable Development, UK
danny@idems.international

A "Statistics Problem-Solving" course has been designed for the African Institute for Mathematical Sciences (AIMS) and given to 47 students on a Mathematical Sciences MSc in Cameroon. The course exposed students to problems in statistics ranging from design, collection, manipulation and organisation of data through to analysis and reporting through games and simulated and real data. Students also worked in groups to explore and report on a specific problem. These included climate for agriculture, procurement for corruption, a poverty survey and 5 other topics. The students' evaluations confirmed that the course was an "eye opener" with some students stating a new-found interest despite no previous background in statistics. For others, it was totally different to their past statistics courses. This paper presents this statistics problem-solving course, which was designed to excite and engage students through experiential learning.

INTRODUCTION

There is general agreement within the statistics education community of the need to update statistics curricular to better equip students with skills needed for a 21st century job market that is abundant with data. There is a growing body of work discussing how to achieve such change. Franklin, et al. (2007) outlines a series of recommendations, and Cobb (2015) and Nolan & Temple Lang (2010) argue for a radical overhaul of university curricular.

More concretely, new introductory courses are being developed at the undergraduate level to address some of these issues. Dierker, et al. (2018) discusses an introductory statistics course that uses project based learning for both science and non-science majoring students. The focus is on encouraging students to learn how to select the appropriate methods and tools for their research questions, rather than learning a fixed set of methods and tools in isolation. Hardin, Horton, & Nolan (2015) gives an overview of a range of new introductory data science courses that "all share a goal of having students become proficient in data technologies and programming tools for problem solving with data" and the paper suggests Programming, Data technologies and formats, and Statistical Topics as topics to consider when integrating data science into statistics curricular. This is more prescriptive than the approach taken by Dierker et al. (2018) where the choice of tools and methods varied across implementations and students within a class.

This paper presents a statistics problem solving course at the MSc level, taught for the first time to 47 students at the African Institute for Mathematical Sciences (AIMS), Cameroon as part of an MSc in Mathematical Sciences. There was substantial educational background diversity of students from 12 African countries. Their experience with statistics varied from no undergraduate statistics courses to full statistics degrees. The course included a project-based component in the final third of the course and aimed to expose students to the broad definition of statistics, as is often experienced by practicing statisticians, including design and collection, explored through experiential learning with statistical games, as well as analysis, interpretation and reporting.

BACKGROUND

The African Institute for Mathematical Sciences (AIMS) is a pan-African Network of Centres of Excellence that offers a 10-month intensive Master's in Mathematical Science to African nationals, at each of its six centres across Africa with approximately 300 graduates per year in total. Courses are taught in three week blocks by visiting lecturers recruited internationally. They are highly diverse and include mathematics, statistics, theoretical physics, computer science and other related mathematical areas. Students are from a range of mathematical science and engineering backgrounds. The Master's program is designed to prepare students with relevant skills needed to solve problems that can contribute to the development of Africa. Hence, the program is highly practical, including the use of

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. ©2019 ISI/IASE iase-web.org/Conference_Proceedings.php

statistical and programming software. There are no formal exams and students are assessed through course work, projects and assignments.

Since its inception in 2003, the AIMS Master's program has included a mathematics and a physics problem-solving course at the beginning of the program. These are designed to equip students with core problem-solving skills and prepare them for subsequent mathematics and physics courses. The idea for a statistics problem-solving course was partly inspired by experiences giving the mathematics problem-solving course across multiple centers and an RSS initiative to strengthen the statistical component of AIMS training.

The AIMS MSc program uses R as its standard statistical software. In 2015 we made a case, through crowd-funding, to develop a front-end to R to be support a transformation in the teaching and use of statistics in Africa, and beyond (Stern, 2017). This led to the development of R-Instat, which was used on this course. In later statistics courses these students used R directly. In the past, in an introductory course, we found students confused learning of statistical ideas with their learning of R. The use of R-Instat served to introduce R in a gentle way, while enabling the course to maintain its focus on statistical issue, and particularly data issues.

We have long used statistical games to teach statistical concepts, e.g. (Mead & Stern 1973; Stern, Latham, & Stern, 2009). Two were used in the course. The tomato game simulates an experiment, while paddy is a crop-cutting survey to estimate rice yields. Both computer and paper versions exist which give students experiential learning in the concepts of design, data collection and entry, followed by analysis. They provide good examples of exercises where the questions can remain the same, while different answers can be demanded, depending on prior knowledge.

METHODS

The Course

Students' prior statistics courses, ranged from zero (e.g. maths, physics degree) to 30 (full statistics degree) and most in between (e.g. applied maths, computer science degree). There were also four tutors supporting the lecturers on the course, who are based at AIMS for the full length of the program. They are typically PhD students or recent PhD graduates.

Where statistics has been taught, it was largely formal and lacking in data. Understanding of some basic concepts beyond their formal definitions (mean, median, sd) was low. We therefore emphasised the importance of understanding the concepts behind the formulas (e.g. through estimation examples) and moving student's thinking from looking for single "perfect" answers to problems and instead to understand that problems with real data are complex and often don't have single, neat solutions. Instead they have many aspects that can be explored and considered.

Given this background, we had three broad aims for the course. By the end of the course students should be able to:

- understand the role of statistics in solving real problems with data
- be comfortable with producing and interpreting simple descriptive tools (tables and graphs)
- understand the broader processes within statistics including design, collection, management and analysis of data and interpretation and presentation of results

The course was taught over a three-week period with five two-hour lectures each week. There were also tutorial sessions each week, led by the AIMS tutors, to reinforce some of the concepts from the class. Assessment for the course was through three (individual and group) assignments, supplemented by short in-class quizzes.

Week 1

We defined statistics, using a definition from the Royal Statistical Society (Royal Statistical Society, 2019) together with a diagram of "data flow". Both emphasised that statistics is not only analysis, but a much broader subject that includes design, collection, organisation, interpretation and communication. We posed seven statistical consultancy questions based on real consultancy issues. Most included issues of design. The simplest was a project on the growth of tadpoles in jars of water at three different temperatures. The question was simply whether to have a few jars, with many tadpoles in each, or many jars with just a few. Others included the design of a variety trial and the proposed structure of a PhD to consider climate change and poverty. Students, in small groups, discussed how they would respond and how they would discuss with the client.

The first quiz included a question where students were presented with the following six numbers (in order) of the number of livestock per household: 0, 0, 0, 10, 15, 20. They were asked for the median and mean, and then also whether any other summary was possible. For most, it was a totally new idea that the zeros could be considered apart and the averages then given for those who had livestock. Students were also asked to estimate the standard deviation of a set of roughly symmetric data, between 10 and 20. This was a bewildering question for some, as it relied on an understanding of the concept rather than a calculation.

We gave a survey (on paper and using ODK (University of Washington, 2008)) asking the students how many statistics courses they had taken (the range was 0 to 30) and what was the largest data set they encountered in any of their undergraduate courses (few had used much data, and it was up to 300 cases). The results emphasised that students' past courses were almost all on analysis, and generally students knew many significance tests but little about descriptive methods.

This course introduced R through R-Instat, a menu-driven front end to R. The tutors gave two tutorials sessions on R-Instat, following introductory guides, one using the diamonds data from the ggplot2 R package (Wickham, 2009), with 50,000 cases and the second using daily climatic data from Dodoma, Tanzania (30,000 daily values).

Students collected data (in pairs) from two simulated case studies. One was an experiment, with 8 treatments (a 2*2*2 factorial design) in 2 blocks, each of size 6, and for 2 years. Among the ideas introduced naturally was understanding of factors, treatments and interactions and the distinction between blocks and replicates. Many had an experiment with 1.5 replicates - and unequal replication was a new idea to many who had previously studied the subject. Others insisted on 2 replicates, but then had to omit 2 treatment combinations. The fact of having no obvious "optimal design" was an important message.

The second study was a rice survey in a sample of villages, with farmers in each village and plots in each field. This was a multi-level exercise, where the selected farmers provided information on their variety of rice, fertilizer used, and the size of their field. One objective was to estimate the total production in the district, and a second was to investigate the relationship between the yields and the inputs.

In each case the students had to design the study, collect data, enter data, analyse and then write a report in the style of a short paper (Introduction, Methods, Results, Conclusions). The importance of linking analysis to the objectives of the study was emphasised.

Through this exercise it became clear that students were familiar with spreadsheet software, but had not met pivot tables. Again, this important tool of a practicing statistician is often omitted from undergraduate training. This was presented in another tutorial.

Week 2

A larger version of the experimental game from week 1, that runs in Microsoft Excel, simulates a large-scale on-farm trial and is loosely based on a current study in Niger. A different sample of 1040 plots of simulated data was provided to each student. These data were from different regions, soil types, genders of respondents as well as the three treatment factors. Visualising these data in different ways (summary tables, boxplots) – to understand variability and pattern was discussed. Students were introduced to ANOVA as a descriptive tool, (i.e. without discussing p-values) to promote the idea that a good analysis is one that explains variability.

Students were introduced to CAST (http://cast.massey.ac.nz/), a series of electronic, interactive textbooks, with many examples from Africa. Installation issues limited the use of this resource, other than for self-learning of statistics concepts.

Week 3

In the last lecture of week 2 students were introduced to the 14 week-three projects. Topics were based on real problems, with most from projects the course instructors have been involved with in Africa. Students gave their preferences and the tutors then allocated them in groups of four. Eight of the 14 topics were chosen, some with two groups. The chosen topics were:

- Corruption "red flags" in public procurement. Using open World Bank data (200,000 records from over 140 countries.)
- Cameroon climatic data analysis. Using daily data from Cameroon Met Service for two stations. Each group used one station.
- Analysis of timber trees. Using multi-level data from farms and plots.
- Designing an ODK survey. One group designed a survey to study stress of the AIMS students while the other looked at the possible outreach activities of the students.
- IFAD poverty survey data from 2018 in Lesotho or Kenya. Lesotho was chosen. The survey had about 1300 respondents and over 400 variables.
- Moving from R-Instat to R. This used a guide on how to write R commands from within R-Instat, and how to transfer to use commands in R/RStudio itself.
- Tidy data. Based on a paper by Hadley Wickham and including his data from Mexico (500,000 cases).
- Fuma Gaskiya. Data from a 2017 on-farm trial of low cost fertilisers involving 1,700 mainly women farmers in Niger. (Similar to the simulation game used in week 2.)

The lecture times in week 3 were largely used as working time on projects, with short plenary sessions to discuss common themes arising from the projects. The main task was for students to prepare a written report and a presentation for the class. Two groups also constructed a short video.

Course evaluation

Students completed a course evaluation form online through a Google Form. The responses were anonymous, and all 47 students completed the evaluation. Students rated the course on various aspects including the content, teaching style and assignments. They were also asked about their experience and views of statistics prior to the course and to what extent their views had been changed.

RESULTS

Lecturers evaluation

This was a "skills course" within the MSc and hence had the dual aims of preparing the students for their further statistics courses, as well as for their future. The varied previous experience of the students constrained what could be achieved. Many statistical problems are (at least partially) solved through good data skills, followed by descriptive summaries. Here we interpret descriptive methods as including producing the appropriate graphs and tables that correspond to the objectives of a given study.

The course was designed as an "eye opener" as most students had no previous ideas of the varied work undertaken by a statistician. The statistical journal *Significance*, which was donated by the RSS for the AIMS library, could be used to follow this up.

Individual discussions with students confirmed that the course had been an "eye opener". Some students stated a new-found interest despite no previous background in statistics. For others, it was totally different to their past statistics courses and provided practical contexts for many concepts which they only know as a mathematical formulation. For many it seemed they gained an appreciation for what statistics is and how it is applies to real world problems, and they also gained interest in studying statistics.

The reports and presentations also demonstrated that students had developed an understanding of the importance of linking their data analysis to the objectives of the study, and were able to interpret results (graphs and tables) in the context of a problem.

In contrast to their lack of knowledge of some basic concepts, by week 3 they attacked some difficult problems, many involving complex, real data sets, and they did so with great enthusiasm and considerable skill, resulting in some impressive presentations in the final session. This demonstrated that a lack of knowledge of statistical theory is not a barrier to introducing interesting, complex data problems to students.

Within descriptive statistics we concentration more on good graphs, than good tables. The production of good multiway tables for frequencies, and other summaries was not given enough time as needed, as students often struggle at choosing the appropriate percentages. Only one student was totally negative about the course. This student had done no previous statistics courses and felt unable to understand the broad ideas in the course due to its lack of theoretical content as it was "just practical". Thus, some students get "lost" in a practical course when their comfort zone, with a mathematical background, is theory heavy. This could be addressed by providing more traditional lecture notes on the content in a more familiar way, possibly as reading material.

Student Course Evaluation

Almost all students rated the course as "Great" or "Excellent" in terms of coherence and relevance indicating that the structure and topics covered were deemed appropriate for the students' needs. Just one student rated it as "Weak" for both, as mentioned above. Almost all students agreed or strongly agreed that the assignments enhanced their learning

Most students (29/47) said the difficulty was "Just right" with 8 saying it was "Quite easy" and 10 saying it was "Quite hard". This is consistent with the large variation in the background of the students. Nine of the 10 students who rated the course "Quite hard" had less than three statistics courses previously. The response to other questions of those 10 students indicated they had a positive experience with comments such as "I didn't know anything about statistic before but now I can do some analysis of statistics data. I'm very very happy for this." and "Since then I had very little interest in statistics because I had no idea about its practice in life, but now I find that it is a promising field that will help Africa to solve several problems".

40 out of 47 students said the course prepared them "A lot" to use statistics to solve problems, with just one response of "Not at all". 38 of the students stated that their view of statistics has changed because of the course. This was independent of the students' previous statistics experience with 4 out of the 5 students who had more than 10 previous statistics courses stating that their view had changed. This confirms that the course presented statistics in a way that is not commonly done, even to students doing a full statistics degree.

Those who stated that their view of statistics had changed were asked how it had changed. 18 commented that they now view statistics as a practical subject, with comments such as "Before this course the little amount of knowledge that I had in statistics was related to the computation of some basic formulas.", "Computers can make calculations, but it is for the statistician to give to those calculations a real signification." and "As applied mathematics, this course has changed my point of view in data analysis and the application of exploratory statistics." and "I now know that statistics is more practical than theoretical as taught in schools". Eight of the students commented on seeing how statistics relates to the real world and its importance in helping to solve problems: "I now understand what a typical statistician should do", "now I find that it is a promising field that will help Africa to solve several problems" and "it is useful in solving the problems of society.". Five students made comments about a new-found interest in statistics and an enthusiasm for the subject: "this course gave me the interest in statistics, I didn't know before that statistic is a very good field" and "statistics has never been more fun". Three students made comments relating to understanding statistics in a broader way "That statistics is not all about Data Analysis but its starts even before the problem, data collection until the final conclusion" and two made specific comments about better understanding of data analysis.

There were suggestions for improvements such as having time to understand all the projects in more detail and to generally have longer for the course. There were also two comments requesting more theory in the course "I do not know the purpose of the course and is not theory, but just practical.".

CONCLUSIONS

The approach taken in the course had strong overlaps with that of (Dierker, et al., 2018) in their project-based learning course. Both courses took the approach of presenting methods and tools in the context of (research) questions and realistic statistical problems. While Dierker, et al. was at the undergraduate level and for students from science and non-science disciplines, there are parallels in our MSc course with some students having no previous experience or interest in statistics, alongside those with a full statistics degree. There are also many overlaps with the data science courses summarised in (Hardin, Horton, & Nolan, 2015), particularly the strong emphasis on data manipulation in many of those courses as a recognition to it often being completely omitted from the curriculum.

There are also differences between the approach of the data science courses and our statistics problem-solving course. The data science courses focussed on well defined tools and methods within data science, demonstrated by the strong emphasis on programming. The approach we took was primarily on students gaining experiences with problem solving in statistics. Thus, as in (Dierker, et al., 2018), lecturers and students select the tools as appropriate, whether that is a statistical programming language such as R, or a menu driven interface such as Minitab or R-Instat. This approach means the exact learning outcomes are less predictable, which has benefits and limitations. We see a role in new statistics curricula for a broad problem-solving course that complements more focused data science courses.

The concept of a statistics problem-solving course need not be restricted to the postgraduate level, and a next step is to adapt this for other academic levels and contexts. This relates to ongoing discussions on redesigning undergraduate statistics curricula, but may also relate to engaging students in statistics at an earlier level, as has been done for mathematics through extra-curricular maths camps to inspire students in mathematics of the 21st century. We hypothesise that the experiential learning through exploration of complex, real world problem and presentation of research related projects that have genuine complexity without single "correct" answers are key components and can be adapted to the level and experience of the students involved.

ACKNOWLEDGMENTS

We thank the RSS for supporting the series of statistics courses delivered at AIMS, and the staff and students of AIMS Cameroon for giving us an enjoyable teaching experience.

REFERENCES

Cobb, G. W. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician, 69*(4), 266–282.

Dierker, L., Evia, J. R., Singer-Freeman, K., Woods, K., Zupkus, J., Arnholt, A., . . . Rose, J. (2018). Project-Based Learning in Introductory Statistics: Comparing Course Experiences and Predicting Positive Outcomes for Students from Diverse Educational Settings. *International Journal of Educational Technology and Learning*, 52–64.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE) report. *American Statistical Association*.

Hardin, J., R., H., Horton, N. J., & Nolan, D. (2015). Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*, *69*(4), 343–353.

Mead, R., & Stern, R. D. (1973). The Use of a Computer in the Teaching of Statistics. *Journal of the Royal Statistical Society. Series A (General)*, *136*(2), 191–205.

Nolan, D., & Temple Lang, D. (2010). Computing in the Statistics Curricula. *The American Statistician*, *64*(2), 97–107.

Royal Statistical Society. (2019). *Stats Careers*. Retrieved from StatsLife: https://www.statslife.org.uk/stats-careers-11-16

Stern, D. (2017). Seeding the African Data Initiative. *Proceedings of the IASE Satellite Conference "Teaching Statistics in a Data Rich World"*. Rabat, Morocco.

Stern, D., Latham, S., & Stern, R. (2009). Statistical Games to support problem-based learning. *IBS 2009 SUSAN Conference proceeding.*

Univerity of Washington. (2008). Retrieved from Open Data Kit: https://opendatakit.org/

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer-Verlag.