USING A PROBLEM-DRIVEN APPROACH FOR TEACHING STATISTICS AT THE AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

Emanuele Giorgi
Lancaster Medical School, Lancaster University, UK
e.giorgi@lancaster.ac.uk

The African Institute for Mathematical Sciences (AIMS) is a research and education institute with multiple centers across Africa that offers Master-level programmes in statistics and applied mathematics. In this paper, I first describe the teaching context and pedagogical objectives of AIMS. I then propose a problem-driven approach for teaching statistics which emphasizes the role of statistical inference as integral to scientific investigation. Finally, I illustrate a sample activity where a case study in public health is used to introduce geo-statistical modelling to students. I argue that the proposed approach helps students at AIMS to develop critical thinking skills and gain a deeper understanding of the links between statistical modelling and science.

BACKGROUND

The African Institute for Mathematical Sciences (AIMS) is a research and education institute that was established in 2003 in South Africa with the ultimate goal of enhancing capacity building for science in Africa. Since then, AIMS has considerably expanded its outreach in different parts of the African continent and, in addition to South Africa, has nowadays centers in Senegal, Ghana, Cameroon, Rwanda and Tanzania. Each centre offers Master-level programmes in applied mathematics and statistics, enrolling about 40 students annually from across Africa. The curriculum is reliant upon lecturers from both international and local universities who volunteer to teach courses for the duration of three weeks.

An academic year consists of the three following phases.

1. Skills courses. In this first phase, the students attend a set of compulsory courses that introduce them to foundational topics in mathematics, physics and computing science.
2. Review courses. In the second phase, a total of 11 out 18 optional courses are chosen by each student. These courses cover specific topics that span a wide range of applications (e.g. bioinformatics, financial mathematics, agricultural science) that are of relevance for the development of Africa.
3. Research phase. The students undertake a research project for the duration of 10 weeks and are supervised by one of the visiting lecturers.

In the remainder of this paper I shall focus my attention on the teaching context relevant to the first and second phases by drawing from my personal experience as a lecturer for AIMS Ghana and Tanzania. I then propose a problem-driven approach for teaching statistics and argue that this is one of the most effective ways of presenting statistical ideas to AIMS students. To this end, I then illustrate a sample activity where this approach is used to introduce geostatistical modelling of tropical diseases in Africa.

TEACHING CONTEXT

As outlined by the AIMS teaching guidelines (AIMS Ghana, 2017), the objective of the Master programme is to provide students with a broad overview of the modern problems and methodologies of applied mathematics and statistics. Most AIMS students indeed see this experience as a stepping stone to further studies in the applied sciences.

The schedule of classes is based on an intensive block system. Table 1 shows an example of the weekly schedule for a three-weeks block delivered at AIMS Ghana during the academic year 2016-17. A single course has a total of 10 contact hours per week. However, lecturers and students meet outside class with varying frequency during the week but more intensively than in most Western universities. This is also facilitated by the fact that students and lecturers live next to each other for the entire duration of a course. As a result, the actual number of contact hours is difficult to quantify if we factor in the additional office hours. Lecturers also have the option of delivering additional evening classes in order to fill any knowledge gap identified during the daytime lectures.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Table 1. Weekly schedule of a three-weeks block at the African Institute for Mathematical Sciences in Ghana.

TIME	Monday	Tuesday	Wednesday	Thursday	Friday
08:30-10:30	Functional analysis	Complex networks	Statistical modelling	Complex networks	Statistical modelling
11:00-13:00	Complex networks	Statistical modelling	Functional analysis	Statistical modelling	Functional analysis
14:00-16:00	Statistical modelling	Functional analysis	Complex networks	Functional analysis	Complex networks

The size of the class can range from 10 to 50 students who usually come from 14 different African countries. Every AIMS student is required to hold a 4-year Bachelor degree with a strong mathematical component in order to be admitted to the programme; see AIMS South Africa (2017). However, due to the quality-variation in higher education across Africa and because of the different educational backgrounds of the students, their level of preparation in statistical subjects ranges from none to good knowledge of basic statistical methods. This leads to very different learning paces among the students and, therefore, flexibility in the course programme is a fundamental aspect that any lecturer at AIMS should take into consideration. Rigidly sticking to a predefined set of learning outcomes might be harmful to the motivation of the students. On the other hand, an excessively loose control on the actual teaching content might lead to cover a significantly smaller proportion of the course programme than what originally planned.

For these reasons, lecturers are also encouraged to design their courses with a strong formative assessment component by promoting activities that help the students to develop critical thinking and avoid rote learning. Low-stakes assignments and lab practicals are an example of the most commonly used forms of formative assessments by AIMS lecturers, in order to identify individual deficiencies and provide prompt feedback to the students. The final grade given to each student is a weighted average of the scores obtained in the weekly assignments and quizzes, whose weights are defined retrospectively by the lecturer and an AIMS committee which includes tutors and the local academic director.

This brief description highlights the pedagogical challenges which sets the AIMS' teaching context apart from most of those in the Western higher-education systems. However, feasible solutions to such issues can be more easily found when statistics is presented as an ancillary discipline of science rather than as a set of prescribed methods. The centrality of science in the statistics curriculum is indeed essential in order to provide meaningful contexts (Watson, 2014).

STATISTICAL INFERENCE AND ITS RELATION TO PROBLEMS IN SCIENCE

Application of the scientific method to investigate questions about the world can be summed up into three main stages. First, we formulate a theory which is used to understand and predict the behaviour of the process in nature under investigation. In a second stage, guided by our theory, we conceive an experiment to interrogate nature about its validity. The data generated by our experiment are then used in a third stage to find any evidence that might lead us to abandon or revise our theory. To make this final step successful, statistical inference plays a crucial role by providing mathematically and scientifically principled tools for measuring the strength of such evidence. In such description of the relation between statistical inference and science, it is important to highlight that statistics cannot be used to prove a theory to be true but only to disprove it. In my teaching experience at AIMS, this aspect has been often misunderstood by the students, especially by those who were more used to the rigour of pure mathematics. Hence, to successfully convey statistical ideas and how these can be used to solve scientific problems, our teaching framework should take into account the epistemological foundations of statistics.

We can outline the formulation of a statistical model in general terms as follows. We first lay out plausible modelling assumptions that relates to the specific question to address. Let [X] be a shorthand notation meaning "the model for X"; we can define a statistical model as

$$[\text{S and D}] = [\text{S}]\,[\text{D given S}]$$

where S represents a state of nature, our object of scientific interest, and D stands for our data that are obtained either as the result of an experiment or by directly observing the natural world. The right hand-side of the above equation defines the two basic ingredients that make up a statistical model: the model for S; and the model for D given S. This framework provides a fertile ground for discussion in class of case-studies and how our knowledge about the problem can help us to formulate appropriate statistical models. Subject matter knowledge of the scientific context is indeed required in the formulation of [S], showing the students that scientific investigation is a joint endeavour between different disciplines, including statistics. Finally, considerations on the sampling process and how the experiment was carried out give us insights which relate to [D given S].

Following the formulation of a suitable statistical model, we then define the objective of statistical inference, which generally falls under one of the following headings.

- *Parameter estimation.* Our object of scientific interest is the true value of one or more of the parameters in [S]. The goal of statistical inference is to provide the best possible guesses for these parameters through the data and quantify the uncertainty around them.
- *Prediction.* In this case, we are interested in making inference on [S given D], i.e. in predicting the behaviour of S in light of what we have observed.
- *Hypothesis testing.* We want to test whether the data are compatible with a specific set of values for one or more of the parameters in [S].

Which of the these better applies to a specific problem provides additional material for discussion with the students. In the next section I illustrate a sample activity where the case-study presented to the students falls under the class of prediction problems known as "disease mapping".

SAMPLE ACTIVITY: A CASE-STUDY IN *LOIASIS* RISK MAPPING

In this section, I illustrate a case-study in public health that I used to introduce geospatial methods for a course delivered at AIMS Tanzania in March 2017. I also provide key excerpts of my interaction with the students.

The problem presented to the class concerned *Loiasis*, an infectious disease caused by the parasitic worm *Loa loa*, also known as the "eye worm".

I started my class by providing background information on the disease.

*"*Loa loa *is transmitted from human to human through the bite of an infected deerfly of the genus* Chryspos *which is mainly found in the rainforests of West and Central Africa. This disease has become of public health importance in Africa because of the adverse effects that have been observed in individuals who are treated for other infectious diseases, namely river blindness and lymphatic filariasis."*

This was then followed by a description of the life cycle of the *Loa loa* worm as shown in Figure 1. In describing the different stages from 1 to 8, I also gave further explanation on some epidemiological concepts, including "diagnostic stage" and "infective stage".

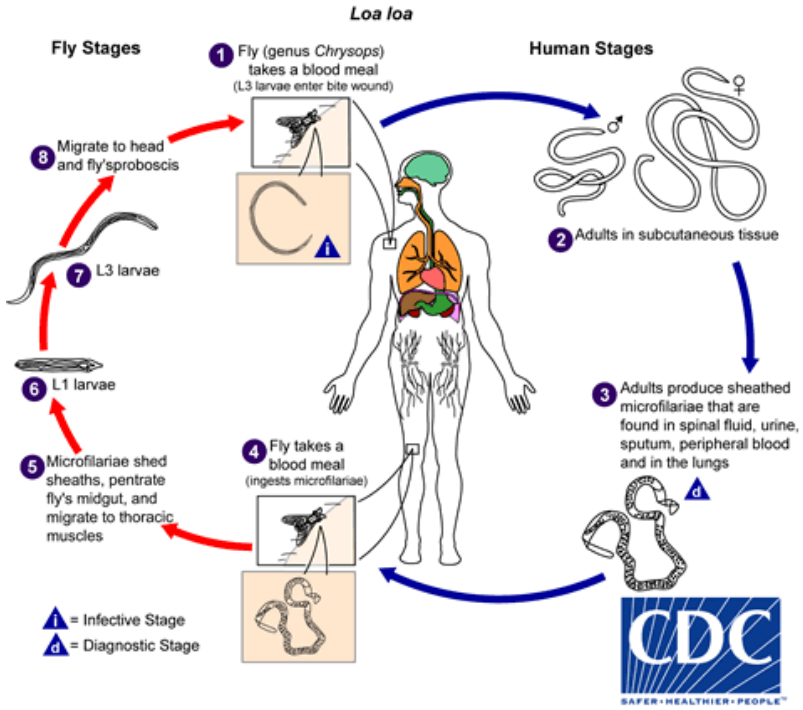

Figure 1. The life cycle of the *Loa loa* worm.

I then defined the public health problem of this study and briefly described the data available for the analysis.

"In places where Loiasis infection exceeds 20% prevalence is now policy to put in place precautionary measures, such as mass drug administration. The goal is to identify areas in Nigeria and Cameroon that are highly likely to exceed a 20% prevalence threshold. The data were collected from 197 randomly sampled villages in the forest and savannah areas of Cameroon and Nigeria. At each sampled village, individuals over the age of 5 years were randomly selected and tested for Loiasis. The total number of those who tested positively was recorded."

This information provided enough material for application of the framework described in the previous section.

"*In this case-study, the state of nature S represents the underlying risk of contracting Loiasis that is experienced by the different communities in Cameroon and Nigeria where such disease is endemic. We also know that environmental and climatic factors affect the successful completion of the life cycle of Loa loa. As these variables vary over space, we would also expect the distribution of Loiasis risk to exhibit a spatially varying distribution.*"

The above remarks were a first attempt to convey the concept that formulation of an appropriate model for Loa loa risk will take into account how this varies over space. The *first law of geography* can also help to explain this in an easy way to understand: "close things are more related than distant things in space". By supplementing this with other examples of the first law of geography, the students were then able to recognize that villages closer to each other are more likely to experience similar levels in Loiasis risk.

I then outlined a possible formulation for [S] based on the available data.

"*The data also include information on the elevation in meters associated with each sampled village. A model for S can be expressed as*

S = Elevation + Unmeasured risk factors

where "Unmeasured risk factors" represent the residual variation in Loiasis risk which is unexplained by elevation. Can you give me an example of other risk factors in addition to elevation?"

By referring to the background information, the students also identified temperature, humidity and vegetation density as variables that could be used in the model, if these were available. For their first analysis, I asked the students to develop a model under the assumption that elevation was the only risk factor for Loiasis. Before introducing more complex statistical methods, it is instructive to show the use and limitations of simpler methods that the students are already familiar with. Hence, in their first attempt, the students analysed the data using standard logistic regression (SLR) that they had already studied in a previous course. In SLR, the model for D given S is assumed to be a set of mutually independent Binomial variables with probability of contracting *Loiasis* given by $p(S)=\exp(S)/\{1+\exp(S)\}$.

Finally, I defined the goal of statistical inference for this case-study as falling under the class of predictions problems.

"Since our interest is in identifying areas that are highly likely to exceed a 20% prevalence threshold, our predictive target is

Probability{p(S) > 20% given D}."

This case study was revisited in the following lectures by questioning the underlying assumptions of SLR. We carried out analysis of the residuals and used the variogram as a tool for testing the presence of residual spatial correlation. I then gave a brief introduction to the theory of Gaussian processes as a way of accounting for the "Unmeasured risk factors" that were ignored in this first analysis.

CONCLUSION

Teaching statistics at the African Institute of Mathematical Sciences (AIMS) offers several pedagogical challenges as a result of a strongly diverse class. The intensive schedule of the classes over a three-weeks period takes up almost all of the time available to a lecturer during the week, also due to the continuous interaction with students that takes place outside class. I have argued that these two features of the teaching context makes AIMS a unique academic environment.

To address these issues, I then developed an approach for teaching statistics that emphasizes the link between statistical models and science. Teaching of statistical ideas to a heterogeneous group of students can be more effective if we focus on how those relate to scientific problems. To achieve this, I have described the formulation of a statistical model by distinguishing two main components: the process of nature, S, our object of scientific interest, and the sampling process that we use to find out about the properties of S. Separating these two aspects of a statistical model allows to bring in more scientific context and also provides the students with a guideline for developing suitable modelling techniques instead of choosing methods from a list. In my experience as a lecturer at AIMS, I found this teaching approach to be effective in helping students develop critical thinking skills and in motivating them by showing examples of how statistics provides feasible solutions to problems that afflict their continent.

In the sample activity, I have shown the application of the proposed teaching framework to introduce geostatistical modelling through a case-study in tropical disease epidemiology. For other case-studies in agricultural, environmental and physical sciences where such framework would find a natural application, see Diggle & Chetwynd (2011) and Weisberg (2014).

REFERENCES

African Institute for Mathematical Sciences - Ghana (2017). Teaching at the AIMS. *Annual report.* Retrieved October 11, 2017, from www.nexteinstein.org/our-annual-report.

African Institute for Mathematical Sciences - South Africa (2017). AIMS structured Master's applications - January intake. Retrieved October 11, 2017, from https://www.aims.ac.za/en/programmes/aims-structured-masters-in-mathematical-sciences/applications-january-intake.

Diggle, P. J., & Chetwynd, G. C. (2011). *Statistics and Scientific Method: An Introduction for Students and Researchers.* New York: Oxford University Press.

Watson, J. (2014). *Curriculum expectations for teaching science and statistics.* In K. Makar, B. de Sousa, & R. Gould (Eds.), Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Weisberg, S. (2014). *Applied Linear Regression* (4th ed.) John Wiley & Sons, Inc.