STILL COMING DOWN FROM THE MOUNTAINS

Roger Stern[1,2], Ric Coe[2,3] , and David Stern[1]
[1]University of Reading
[2]Statistics for Sustainable Development
[3]ICRAF
r.d.stern@reading.ac.uk

Statistics has changed in many ways since the 1960s, when many African countries became independent. Among these changes are an increasing emphasis on data and a set of unifying principles that can simplify the teaching of statistical modelling. These changes have yet to impact training in statistics in many countries. Access to technology is needed if these changes are to be incorporated into statistics teaching, and this is now feasible in many African universities.

INTRODUCTION

We are in the middle of a "Data Revolution" according to the United Nations (2014). Statistical skills are needed to make sense of data and this means that training in statistics should include components that are more data based, using a range of real examples. This is reflected in the 2016 GAISE report, which recommends that (introductory) statistics courses should "integrate real data with a context and purpose". The report also notes that the "rapid increase in available data has made the field of statistics more salient".

Prior to the data revolution there was a statistical modelling revolution that gathered pace following the seminal paper by Nelder and Wedderburn (1972) together with the accompanying GLIM software. The changes in modelling have made more advanced statistical skills accessible to a wider range of students.

The teaching of statistics in many African universities has been unaffected by these major developments in the subject. One reason has been the lack of technology for students. This has now changed. Many African universities are open to change and this provides exciting opportunities in statistics education for both staff and students.

This paper summarises the convergence of some statistical modeling concepts and methods. It then describes new challenges and how curricula have responded or can now respond.

COMING DOWN FROM THE MOUNTAINS

Up in the mountains

Statistics was a piecemeal subject in the 1960s, the decade when many African countries became independent. It was still "at the top of the mountain". A standard textbook at the time often had three main parts:

- *Descriptive statistics*, with summaries like means and standard deviations and bar charts and histograms, not box plots which had yet to be "invented" by Tukey (1977).
- *Statistical modelling for small problems.* This was particularly for probability ideas and the method of maximum likelihood for estimating the one or two unknown parameters from a wide range of distributions. These distributions included the normal, log-normal, exponential, gamma, binomial, Poisson and negative binomial.
- *More general statistical modelling*, particularly multiple regression and ANOVA.

What was puzzling at this time was the absence of these distributions in the last part of the book. The more general modelling opened the door to being able to analyse realistic data sets with complex structures. But why did students learn about all the different distributions, when they could only be used for relatively trivial problems?

The last part of the course assumed normally distributed residuals. If this assumption was not satisfied then data were transformed, perhaps by square roots or logs. The work of Box and Cox (1964) could help to define an appropriate transformation. This paper has, so far, been cited over 10,000 times.

Statistical methods were truly like little streams at the top of a mountain. There were even separate chapters for regression models, where the x-variables were all numeric and for ANOVA

where factors (categorical variables) were part of the model. The analysis of covariance sat slightly uncomfortably between these two situations. Then the chi-square test was another "little stream" to analyse frequencies in 2-way tables, with its own special methods depending on whether it was a 2 by 2 table or some of the factors had more than 2 levels.

If none of these ideas appealed, then there was a set of distribution-free (non-parametric) methods with a wide range of "tricks" to process different types of data.

On top of another mountain was a group that advocated a totally different approach, namely to use Bayesian statistical methods, such as Lindley (1965). This was fun philosophically but Bayesian analyses were (at this time) intractable for all but relatively trivial problems and hence not usable for the interesting problems of multiple regression or ANOVA.

Coming down

Nelder and Wedderburn (1972) started the trip down the mountain through the methods in their paper titled "Generalised Linear Models". This was accompanied by an interactive statistical package, called GLIM, which was released in 1974. Regression modelling and ANOVA were now combined into linear models and these could be for a wide range of distributions. Contingency tables that had previously been restricted to 2 dimensions could now also be analysed within this framework as log-linear models. GLIM is no longer available, but these methods have been incorporated in most standard statistical packages and in Genstat in particular (VSN, 2016).

This simplification of statistical methods had one important limitation. It was essentially for fixed effects models at a single "level". Analysing a split plot experiment or a survey with data at multiple levels was not within this framework. Work on this area proceeded in the 1970s, but it was only easily accessible in the statistical packages in the 1980s and 1990s, with REML in Genstat, PROC Mixed in SAS and the Mixed command in SPSS and Stata. Analysing data at multiple levels includes the idea that responses are (usually) more correlated if they are at "close together", e.g. in the same household, or within the same person, compared to items that that are at a higher level. The same idea applies to spatial data for which the same class of models can therefore also be used. So we have now descended a long way from the top of the mountains in the 1960s. This unity in methods should simplify both teaching and learning.

The changes in modelling did not stop there. They continue all the time. For example the statistical models mentioned above all had to be in the "exponential family" of nice distributions. For example, Cole (2001) showed how the same ideas could be applied to extreme value distribution, though they are not in this family, and these ideas are now in the R package called ExtRemes (Gilleland & Katz, 2016).

Meanwhile GLMs became Hierarchical GLMs (Lee & Nelder, 2006) to permit models where any parameter of the distribution has its own dependency on other measurements. In some cases this has been facilitated by the integration of Bayesian methods that make the estimation feasible in practice.

Other descents

In the 1960s there was also usually a course called "The Design of Experiments". This was an optimistic title which was often taught more as "The Analysis of Designed Experiments". As with much of statistics teaching, it concentrated more on analysis than on planning and design. This was usually separate from a course on "Survey Design and Analysis".

The design of experiments has a long history in agriculture. Research in this area has changed, and much is now undertaken "on farm" and is now in what might be called large-N trials, such as the one described by the Collaborative Crop Research Program (2016). Here the number of farmers is similar to the sample size of a survey, and the methods of survey analysis and modelling are more relevant than the traditional analysis of experimental data.

Meanwhile the standard teaching of experimental design often remains largely appropriate only for on-station trials (or for industrial experiments) and for situations where surveys and experiments remain separate topics. Graduates are therefore progressively less able to contribute to research questions for agricultural applications and in other important areas. This includes the possible roles of randomized control trials in educational and other research, and the potential value of controls in monitoring and evaluation studies.

Statistics courses in the 1960s did not discuss the design and analysis of participatory studies or the richer types of data that are now often collected. Hence, some statisticians, and also some social scientists, consider this to be a different area and hence not relevant for statisticians. Holland et al (2013) show otherwise. Kagugube et al (2009) show the value of this mixed approach for the National Household Survey conducted by the Uganda Bureau of Statistics (UBOS), while Barahona and Levy (2003) provide an example from Malawi.

Statistical Software

Statistical software has a long history. SPSS is perhaps the most used package now and it was first released in 1968. The other giant of statistical packages, SAS, was first developed in 1966, while Genstat was first released in 1971 and Minitab in 1972. This is a long way back in the history of computers that only really started in the 1940s.

These software systems have each added components as new methods appeared. However they have not always been quick to abandon methods developed earlier. They sometimes show their age, with components still available from the top of the mountains. This provides users with a rich, but also sometimes confusing mix of methods for their data analysis.

Microcomputers appeared in the 1980s, together with many new statistics packages. By the 1990s, once the microcomputers were large enough, the original systems dominated again, joined by a few newcomers, particularly Stata and R.

For universities access to most of this software is through annual licenses, except Stata which is a one-off purchase for each version and R which is free and open source. All have powerful language capabilities, though the language in some of them reflects the age of the package. All, except R, can alternatively be used through a menu-driven front end and various front-end menus have been written for R. This includes the recent R-Instat from the African Data Initiative (Stern, 2017) which is designed to provide an introduction to R that is as gentle as any of the other existing packages. R-Instat, like R, is free and open source.

LIFE IN THE VALLEY

Statistics is still expanding

New mountains are emerging, or perhaps they are islands or whole landscapes of statistical methods that are not yet integrated into the main stream. Important examples include methods used in bioinformatics, image analysis, machine learning or financial mathematics. Like participatory methods, mentioned above, these might be seen as the preserve of other disciplines or application areas. However the smart applied statistician needs to be aware of these areas so that they can use concepts or tools from them that might help. In addition they might eventually contribute to the further convergence of methods used in these areas. Hence curricula should provide eye-openers and, as far as possible, the software should extend to embrace some of them.

The Broader Picture

Table 1 summarises the way a statistician works (or should work) to analyse data as part of a research team. Much the same is true if the 'clients' in the team would not call themselves “scientists” (social or biophysical). Sometimes scientists learn enough statistics or vice versa for the same person to take both roles but that cannot be assumed or used when teaching applied statisticians.

Steps	Work done by: Scientists	Scientists and statisticians together	Statisticians
1. Prepare	Assemble relevant datasets	Understand sources, definitions, structures Choose responses	Quality checking Reshaping, merging
2. Describe	Describe analysis questions or objectives	Visualise potential answers in terms of the data available	Calculate and display
3. Model	Understand and describe the theory that supports potential models for the data	Determination of appropriate model types and structures	Estimate models Determine quality of fits Quantify evidence in support of alternative models
4. Interpret and use	'Usual science' – reflect and next steps	Determine implications	Explain nature and limitations of statistical evidence

Table 1: Components of an applied statisticians toolbox.

Table 1 remains incomplete in the range of topics needed by a statistician, in that it starts with data so important areas concerned with the stages involved in the design and collection of data need to be added. The 'down from the mountains' story largely turned the grey shaded part in Table 1 from little pieces into a single box. An applied statistics curriculum is then often largely descriptive statistics plus this modelling component, i.e. the blue (dashed) box in Table 1. However, an effective applied statistician needs concepts, tools, skills and experience in all the cells of the (much larger) solid green box.

Curricula are changing

There have been substantial efforts at reforming the way statistics is taught. In the US there are the 2016 GAISE guidelines, which push for an education that includes an emphasis on creating understanding and using real world examples. This is an important response to the challenge facing statistics educators. Despite these efforts, statistics education in most of the world is still mostly traditional.

New Zealand is an exception to the norm in this regard. They have been dynamic in their design of a new approach (Forbes, 2014). This has emphasized bootstrapping and integrated technology including customized software and other free resources such as CAST (Stirling, 2017). It is probably the most revolutionary statistics curriculum in the world and has been built from extensive collaboration with practitioners.

Data science training is another response to the extensive need for people with skills for working with data which statisticians are not able to satisfy (Finzer, 2013). This has become popular and in the UK alone there are currently over 50 MSc degrees offered in data science.

AFRICAN DATA INITIATIVE

The African context

This unified structure to statistical modelling, together with the increasing importance of data should enhance and also, to some extent simplify statistics teaching and learning. However, in many African universities much is still taught as though relatively little has changed since the 1960s. Some lecturers and hence their students still remain near the top of the mountain.

Until recently universities in Africa suffered from students having little access to technology, particularly to computers or tablets, and this has now changed (Kurji et al., 2010). This should now be assumed, i.e. don't teach or study statistics unless students can gain some access to technology.

This access to technology is not just for statistics specialists, but also for students who need to take statistics courses to serve their main subjects. Whether this access is largely the student's responsibility (like assuming school uniforms), or the university, or is shared, will differ between universities. But lecturers need to be able to assume access to computers for their courses.

A second issue is sometimes the heavy lecture loads of staff in many African universities. This can act as a block to change. Where staff welcome change it is therefore important that innovations are easy, and not too time consuming, to implement. "Making it easy" needs to be a component of suggested reforms, otherwise they are not practical.

Creating access

The African Data Initiative (Stern, 2017) was designed to overcome some perceived barriers to change. The first was the lack of availability of suitable statistical software, and this has led to the construction of a front-end to R, called R-Instat. One solution to the lack of easily accessible textbooks is CAST, Manyalla et al (2014), which has been used successfully in Western Kenya. The moves, worldwide, towards open data should also provide impetus to support some courses becoming more problem, rather than method, based.

CONCLUSIONS

The directions of change in statistics teaching have been clear for a long time. What is more recent is the practicalities of implementing these changes in universities in Africa. The feasibility of access to the appropriate technology by students is crucial. To that we now add the resources, such as the new R-Instat software, together with resources, such as CAST and ILRI that have been available for some time.

Although classes are large the experience from of much larger groups from MOOCS shows this is not an impediment, again if the technology is available so a learning management system can used, to support the teaching.

The staff must be excited by the scope for making changes in their teaching. This is consistent with the agenda in some of the many new universities in Africa that are embarrassing innovation partly as a way to make their name. Finally many of the MSc graduates from the exciting AIMS (African Institute for Mathematical Sciences) courses have been exposed to aspects that are needed in their own universities. There are over 250 AIMS graduates in universities now, and the number is growing each year.

Once universities in Africa decide on change, there is little or no delay in its implementation. The challenge is therefore for some of these universities to "leapfrog" other centres, and then hopefully to "infect" others through the results.

REFERENCES

Barahona, C. E. & Levy, S. (2003). How to generate statistics and influence policy using participatory methods in research: reflections on work in Malawi 1999–2002. IDS Working Paper 212.

Box, G. E. P. & Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*(2), 211–252.

Collaborative Crop Research Program. (2016). http://www.ccrp.org/projects/large-n-trials

Cole, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London, England: Springer.

Finzer, William. (2013). The Data Science Education Dilemma. *Technology Innovations in Statistics Education, 7*(2).

Forbes, S. (2014). The coming of age of statistics education in New Zealand, and its influence internationally. *Journal of Statistics Education, 22*(2).

GAISE College Report ASA Revision Committee. (2016). Guidelines for Assessment and Instruction for Statistics Education (GAISE) Report. Washington, DC: American Statistical Association. http://www.amstat.org/education/gaise

Gilleland, E. & Katz, R. W. (2016). extRemes 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software, 72*(8), 1–39.

Holland, J. (Ed.) (2013). *Who Counts: The Power of Participatory Statistics*. Rugby, United Kingdom: Practical Action Publishing.

Kagugube, J., Ssewakiryanga, R., Barahona, C., & Levy, S. (2009). Integrating qualitative dimensions of poverty into the third Uganda National Household Survey (UNHS III). *Le Journal statistique africain, 8.*

Kurji, P., McDermott, B., Stern, D. A., Stern, R. D. (2010) The Growing Role of Computers for Teaching Statistics in Kenya. In C. Reading (Ed.), *Proceedings of 8th International Conference on Teaching Statistics,* Ljubljana, Slovenia. Voorburg, the Netherlands: ISI.

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models with discussion. *Applied Statistics, 55*, 139–185.

Lindley, D. (1965) *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge, England: Cambridge.

Manyalla, B., Zachariah, M. Stern, D. A., Stern, R. D. (2014). Measuring the Effectiveness of Using Computer Assisted Statistics Textbooks in Kenya. In K. Makar, B. de Sousa, & R. Gould (Eds.), Sustainability in Statistics Education. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Association for Statistical Education.

Nelder, J.A & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A, 135*(3), 370–384.

Stern, D. A. (2017) The African Data Initiative. In A. Molnar (Ed.), *Proceedings of the IASE Satellite Conference, Rabat, Morocco.*

Stirling, W. D. (2017). CAST release 6.1 [Computer software]. http://cast.massey.ac.nz

Tukey, J. W. (1977). *Exploratory Data Analysis.* London, England: Addison–Wesley.

United Nations. (2014). UN Data revolution. http://www.undatarevolution.org/

VSN International (2016). GenStat for Windows (18th Edition). [Computer software]. VSN International, Hemel Hempstead, UK. http://GenStat.co.uk