EMBEDDING DATA MANIPULATION IN STATISTICS EDUCATION

James Musyoka1, John Lunalo2, Cathy Garlick4, Steven Ndung’u2, David Stern4, Danny Parsons5, and Roger Stern3,5
1Maseno University, Department of Statistics and Actuarial Science, Kenya
2African Maths Initiative, Kenya
3Statistics for Sustainable Development, UK
4Statistical Services Centre, University of Reading, UK
5Mathematical Institute, University of Oxford, UK
johnlunalo95@gmail.com

University courses in statistics in many African countries are dominated by data analysis. This is just one component of the subject and students therefore lack knowledge on many important practical laboratory work components. Here we consider how data collection and data entry can be included in training courses. The next important stage of preparing the data, so it is ready for analysis is also considered. These stages, before data analysis, may use a combination of a spreadsheet, a statistics package and special software. Examples of each are considered.

INTRODUCTION

Over the past decade, Maseno University in Kenya has been implementing innovations aimed at improving its statistics teaching (Stern, Ongati, Agure, & Ogange, 2010). This involved integrating real datasets as well as practical skills into the statistics courses. The curriculum of undergraduate programmes have also included relevant computer based courses with specific components on handling data. The postgraduate research methods program (Stern, Coe, Stern, & McDermott, 2013) also included explicit components related to data handling. This involved changing the mode of teaching from physical attendance of classes to use of online mode by use of technology.

These courses have made an important step in the right direction. We draw on three components of data handling that we have brought into teaching. Firstly, getting the data into a computer, then ensuring that the data entered is of as high quality as possible and finally preparing the data so it is ready to be analysed.

Both spreadsheets and specialised data entry software have already been used in teaching at Maseno University to help students understand these ideas. However, we have struggled to integrate the third component of data preparation and manipulation into statistics courses. The African Data Initiative (Stern D. , 2017) includes work on R-Instat, a front-end to R, which complements our existing efforts in each of these three areas.

R-Instat is equipped with facilities to import data easily from the data entry sources that we use, expand on the quality control and make it easier to manipulate data. This is all part of the effort to improve how we can integrate data handling statistics courses.

This paper demonstrates how the software can be used together with a spreadsheet and a specialised software, CSPro, to teach these data handling ideas. It draws on the Data Flow concept summarised by Figure 1 and the attempt to prepare graduates to be able to support researchers through the data flow process. Although this paper only covers components within the Data Entry and Organisation box, this diagram emphasises the statistician’s role throughout the process. The data flow concept was developed through dealing with data within research projects (SSC, 2009) and reflects the broad role described for applied statisticians in (Stern, Coe, & Stern, 2017).
GETTING DATA INTO A COMPUTER

Here we use an example of data entry where students, working in groups, have to enter data in three different ways. In total, there are 30 records to be entered and 10 records are entered using each of 3 methods:

1. Direct entry into a spreadsheet – this is a common method in practice.
2. Entry into a spreadsheet that has been specially prepared, as shown in Figure 2.
3. Entry into CSPRO that has similarly been specially prepared, as shown in Figure 3.

The practical guide for this exercise was prepared assuming Excel, but alternatives, such as Open Office’s Calc can also be used. The data are adapted and simplified from a survey in Malawi. It includes multiple response questions and has data at two different levels: household (30 records) and activity within the households (88 records). There are three instructional videos, one on the survey itself, one on the entry into the special Excel sheet and one on the entry into CSPRO.

The practical work can depend on the level of maturity of the students. For MSc students, the emphasis was on a comparison of the three methods of entry, in preparation for possible similar tasks in the future. Method 1, namely simple entry into Excel is very simple, but accident-prone, while methods 2 and 3 depend on someone having a sufficient level of expertise to set up the appropriate data entry system.
In a consultancy project on survey data collection, conducted in 2017 by one of the authors, two of the four countries opted to collect data manually and then to use CSPro for the data entry. The other two countries proposed to collect data on tablets using the Open Data Kit (ODK). This is another tool that students could learn how to use. Designing a survey that uses ODK is straightforward but students also need to be aware of the advantages and limitations of the use of mobile phones or tablets for direct data capture. Information on these aspects is in a series of videos shown in the third option in Figure 4, titled “Digital Data Collection”.

There is now a strong move towards “Open Data”. This is the idea that datasets should become freely available at some point, for example once the initial research phase is completed. This is now often a requirement from research funders. Access to primary data is also sometimes demanded for referees to be able to validate results from intended publications (OECD, 2011).

Making data open can raise complications if this aspect has not been considered when the data were collected. Discussions on this aspect are therefore important when teaching data collection and initially this may be considered as a simple task, i.e. just ensure that all names and exact locations are deleted from the file. However, consider the situation where data are supplied at village level and there is just one family with a widow and 3 children. This information immediately identifies the respondent and destroys confidentiality. Fortunately, comprehensive software now exists, for example the R package, sdcMicro (Templ, Kowarik, & Meindl, 2015), to help with data anonymization.

Once data are made available, sites such as Dataverse can be used to store both metadata and the data itself. These therefore also provide an excellent resource for staff and students looking for examples of past surveys and other data collection exercises (King, 2007).

Finally, in discussing getting data, R-Instat has options of importing data from external sources. These include importing data from ODK servers, importing from CSPro and importing from databases for “secondary” data.

**IMPROVING DATA QUALITY**

Our aim is for students to be aware of the importance of data being of high quality i.e. it should be consistent, accurate complete and easy to use and understand. This needs attention to quality at all stages in the data collection and entry process. They should be aware, and practice, examples that emphasise data checking on entry, the possibility of double entry and the importance of checking the data once entered (Statistical Services Centre, 2013).

The example shown above in Figure 2 and Figure 3 can be revisited with a new set of objectives. As used above the emphasis is usually largely on the comparison of the three methods of entry. The special entry into Excel has included a series of checks on entry. CSPro has the option of double entry and this could warrant a further practical.

Students should also learn about checks in data quality that can be done once the data have been entered. These can often be done with a spreadsheet, or using a statistics package. One common problem is that, after data entry, categorical data columns may not have the same number of categories that the producer of the data expected. This is easy to see in a spreadsheet using the
filtering option. Figure 5a shows an example where an extra village has been created by a typing error in the name of the village; a survey took place in the villages of Kesen, Nanda, Niko and Sabey, but a fifth village called “Kesan” appears in the list. Note that this type of error is usually avoided if the data are entered using Methods 2 or 3 of the example above.

Pivot tables are an excellent feature of spreadsheet packages and are useful both at the data checking and the analysis stages. Figure 5b shows a simple one-way table that illustrates the same issue. The data in Figure 6 are from clicking on the appropriate cell in this table, and show the row of data that has the problem. It could equally have been shown through use of the filter in Figure 5a. It is important for students sometimes to see different ways of resolving the same problem.

The two screenshots in Figure 7 shows how the same issue could be examined in a statistics package. In R-Instat this uses the Prepare menu’s dialog to examine the levels of a categorical column (called a factor column by R).

Finally, a combination of graphs and tables in the describe menu in R-Instat makes data exploration easier for students to understand (Fundi, Clements, Stern, Stern, & Sananka, 2017). This also helps in checking data quality especially for categorical variables.

**MANIPULATING DATA**

Often the primary data, as entered, are not in the *shape* that is needed for the subsequent data analysis. Changing the shape of the data and preparing them for analysis is another stage that is relatively easy for students to learn. It may again involve either a spreadsheet or a statistics package. We usually find that this stage is simpler with a statistics package. Two common examples are shown
in Figures 8 and 9. In Figure 8, the data have been transferred to the statistics package, but the “shape” is incorrect for analysis. In Figure 8a, which shows daily data, each row has one month of data and the days of the month are each in a different column. These data need to be “stacked” to put them into the right shape for analysis. The result is shown in Figure 8b. All statistics packages have simple procedures for stacking data.

A second example is where identical data have been collected for a series of villages and have been entered into separate sheets in a spreadsheet package. They are now separate data frames in the R-Instat statistics package (African Maths Initiative, 2017), Figure 9a. R has simple commands, used by R-Instat, to join these data together in preparation for the analysis across villages, shown in Figure 9b.

In addition, some of R’s text manipulation features have also been implemented in R-Instat. These include splitting, combining and transforming text columns. The final example is discussed in (Parsons & Stern, 2017). This is the common situation where the data are not at the right “level” for the analysis. For example, daily climatic data may have been provided and an initial summary step to produce monthly or annual summaries is needed for the analysis.

**CONCLUSIONS**

In the past many statistics courses in African universities have not assumed computers are available. Access to technology is now feasible. They may be added simply to improve the teaching of the current syllabus. We propose here that the introduction of the technology should change both what is taught, as well as how it is taught. We should also be using the technology to move away, at least to some extent from a lecture-based approach to teaching. There is a possibility for education to be more student centred.

Any changes in teaching in Africa have also to reflect the heavy lecture loads for university staff. Hence it is important to make change easy to implement and not too time consuming. This is becoming more feasible, because of the many resources that currently exist. We have illustrated some of these here hence, the development of new course materials should build on existing resources rather than starting with completely new ones.
REFERENCES


