EMBEDDING DATA MANIPULATION IN STATISTICS EDUCATION

James Musyoka[1], John Lunalo[2], Cathy Garlick[4], Steven Ndung'u[2], David Stern[4], Danny Parsons[5] and Roger Stern[3,4]

[1]Maseno University, Department of Statistics and Actuarial Science, Kenya
[2]African Maths Initiative, Kenya
[3]Statistics for Sustainable Development, UK
[4]Statistical Services Centre, University of Reading, UK
[5]Mathematical 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 R. , 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 Fig 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).

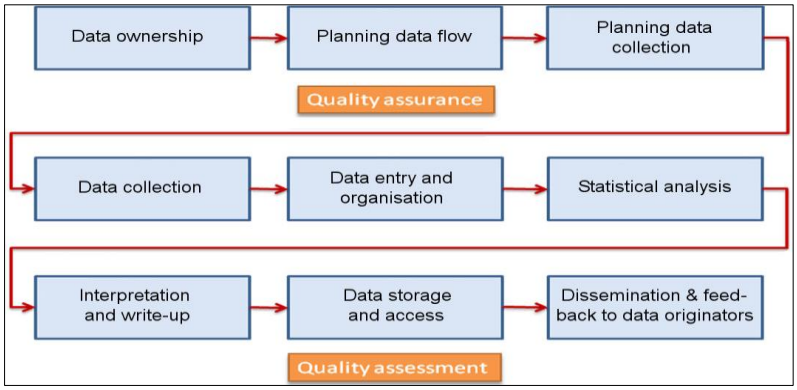

Fig. 1: The Data Flow Process

Source: (SSC, 2009)

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, see Fig. 2.
3. Entry into CSPRO that has similarly been specially prepared, see Fig. 3.

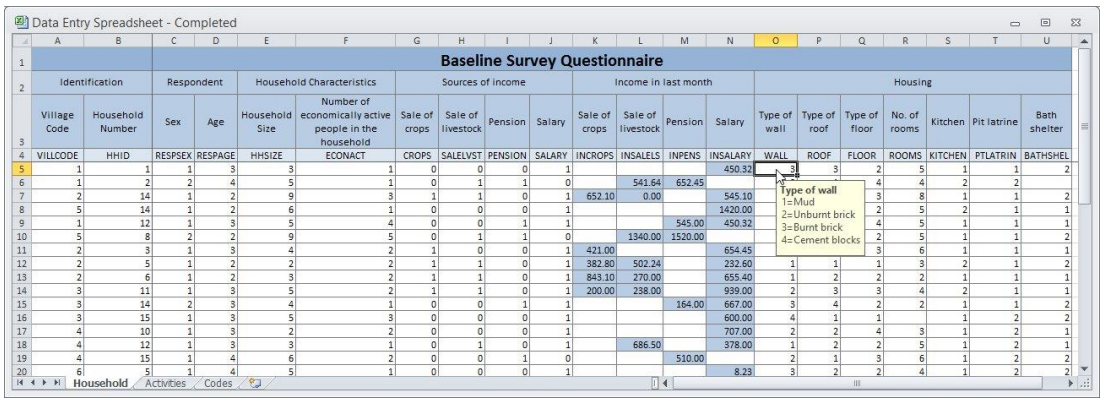

Fig. 2 Data in a specially prepared Excel sheet

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.

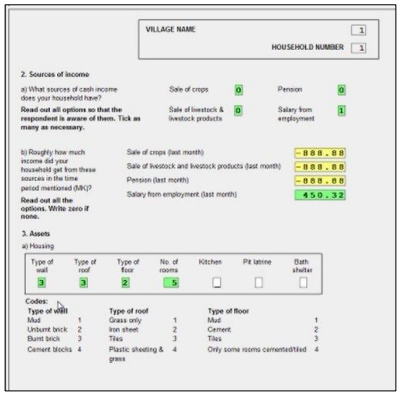

Fig. 3 Data entry into CSPro

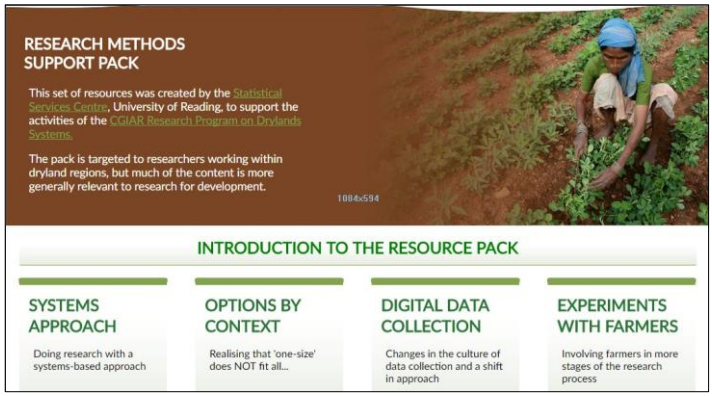

Fig. 4 Videos on digital data collection

Source: (Statistical Service Centre, n.d.)

In a consultancy project on survey data collection, 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 Fig. 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 Fig. 2 and Fig. 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. Fig. 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.

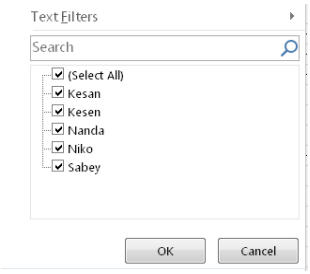

Fig. 5a The filter shows more categories than expected

Row Labels	Count
Kesan	1
Kesen	8
Nanda	9
Niko	9
Sabey	9
Grand Total	36

Fig. 5b So does a pivot table

Pivot tables are an excellent feature of spreadsheet packages and are useful both at the data checking and the analysis stages. Fig. 5b shows a simple one-way table that illustrates the same issue. The data in Fig. 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 Fig. 5a. It is important for students sometimes to see different ways of resolving the same problem.

	A	B	C	D	E	F	G	H	I
1	Village	Field	Size	Fertiliser	Fertlevel	Variety	Yield	infected	insects
2	Kesan	4	8	0	1: none	Trad	27.6	yes	107
3									

Fig. 6 Selecting particular rows in a pivot table

Fig. 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).

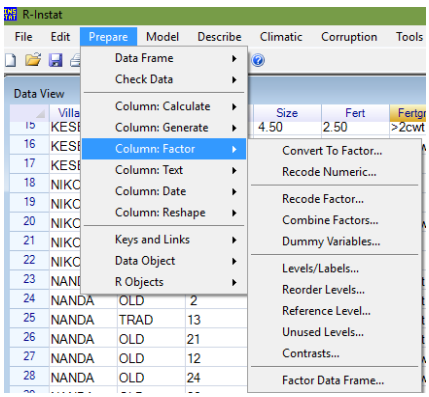

Fig. 7a The Factor Section of prepare menu in R-Instat

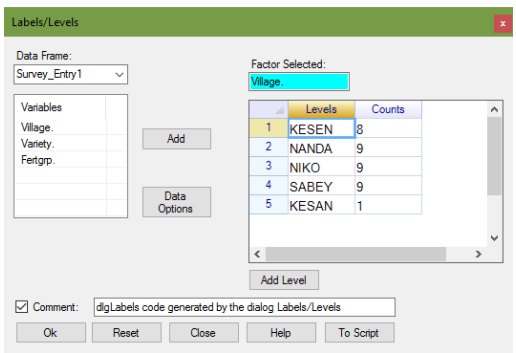

Fig. 7b Finding an extra level of the categorical column

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 Fig. 8 and 9. In Fig. 8 the data have been transferred to the statistics package, but the "shape" is incorrect for analysis. In Fig. 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 Fig. 8b. All statistics packages have simple procedures for stacking data.

station_id	yyyy	mm	Day1	Day2	Day3	Day4	Day5	Day6	Day7
KUND0002	1980	7	0	0	13	7	143	116	29
KUND0002	1980	8	0	0	0	0	0	0	0
KUND0002	1980	9	93	51	64	0	6	33	21
KUND0002	1980	10	6	0	211	113	40	412	87
KUND0002	1980	11	115	145	80	89	10	10	65
KUND0002	1980	12	0	0	0	0	0	0	0
KUND0002	1981	1	0	0	0	0	0	0	0
KUND0002	1981	2	0	0	0	0	0	0	0
KUND0002	1981	3	0	0	0	0	0	0	89
KUND0002	1981	4	145	0	64	158	0	0	56
KUND0002	1981	5	0	0	0	5	0	76	0
KUND0002	1981	6	0	0	51	10	10	145	0
KUND0002	1981	7	180	137	99	69	99	18	0

Fig. 8a Daily data, unstacked

station_id	yyyy	mm	dd	rainfall
KUND0002	1980	7	Day1	0
KUND0002	1980	7	Day2	0
KUND0002	1980	7	Day3	13
KUND0002	1980	7	Day4	7
KUND0002	1980	7	Day5	143
KUND0002	1980	7	Day6	116
KUND0002	1980	7	Day7	29
KUND0002	1980	7	Day8	63
KUND0002	1980	7	Day9	24
KUND0002	1980	7	Day10	386
KUND0002	1980	7	Day11	72
KUND0002	1980	7	Day12	240
KUND0002	1980	7	Day13	36

Fig. 8b Daily data stacked

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), Fig. 9a. R has simple commands, used by R-Instat, to *join* these data together in preparation for the analysis across villages, shown in Fig. 9b.

Data View

	Variety. (f)	Field	Size	Fert	Fertgrp. (f)	Yield
1	OLD	9	1.50	2.00	>2cwt	40.4
2	TRAD	8	7.00	0.00	0cwt	25.8
3	OLD	1	2.50	1.50	.5-2cwt	40.7
4	TRAD	4	8.00	0.00	0cwt	27.6
5	OLD	6	4.50	2.50	>2cwt	48.7
6	TRAD	3	2.00	0.50	.5-2cwt	27.0
7	TRAD	5	4.00	0.00	0cwt	19.1

Data View

	Variety. (f)	Field	Size	Fert	Fertgrp. (f)	Yield
1	TRAD	4	4.50	2.00	>2cwt	36.6
2	OLD	2	3.50	2.50	>2cwt	57.4
3	TRAD	13	20.00	3.00	>2cwt	42.7
4	OLD	21	4.00	3.00	>2cwt	49.3
5	OLD	12	3.00	1.00	.5-2cwt	46.2
6	OLD	24	6.00	1.50	.5-2cwt	42.2
7	OLD	26	4.00	2.00	>2cwt	41.3
8	TRAD	8	3.00	0.00	0cwt	37.6
9	NEW	20	2.50	2.50	>2cwt	58.1

Fig. 9a Data in separate data frames; KESEN and NANDA villages

Data View

	id (c)	Variety. (c)	Field	Size	Fert	Fertgrp. (f)	Yield
1	KESEN	OLD	9	1.50	2.00	>2cwt	40.4
2	KESEN	TRAD	8	7.00	0.00	0cwt	25.8
3	KESEN	OLD	1	2.50	1.50	.5-2cwt	40.7
4	KESEN	TRAD	4	8.00	0.00	0cwt	27.6
5	KESEN	OLD	6	4.50	2.50	>2cwt	48.7
6	KESEN	TRAD	3	2.00	0.50	.5-2cwt	27.0
7	KESEN	TRAD	5	4.00	0.00	0cwt	19.1
8	NANDA	TRAD	4	4.50	2.00	>2cwt	36.6
9	NANDA	OLD	2	3.50	2.50	>2cwt	57.4
10	NANDA	TRAD	13	20.00	3.00	>2cwt	42.7
11	NANDA	OLD	21	4.00	3.00	>2cwt	49.3
12	NANDA	OLD	12	3.00	1.00	.5-2cwt	46.2
13	NANDA	OLD	24	6.00	1.50	.5-2cwt	42.2
14	NANDA	OLD	26	4.00	2.00	>2cwt	41.3
15	NANDA	TRAD	8	3.00	0.00	0cwt	37.6
16	NANDA	NEW	20	2.50	2.50	>2cwt	58.1
17	NANDA	OLD	14	2.50	1.50	.5-2cwt	45.8
18	NANDA	TRAD	28	2.00	2.00	>2cwt	38.7
19	NANDA	OLD	25	4.00	2.00	>2cwt	42.4
20	NANDA	TRAD	5	4.50	1.50	.5-2cwt	25.8
21	NANDA	NEW	9	1.50	2.00	>2cwt	61.4

Fig. 9b Data joined into a single "long" data frame

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

African Maths Initiative. (2017, 03 22). Retrieved from http://www.africanmathsinitiative.net/blog/initiatives/african-data-initiative-part-1-missing-tool/

Fundi, M., Clements, L., Stern, D., Stern, R., & Sananka, A. (2017). Describing Data Well in R-Instat. *IASE*.

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing . *SAGE journals* .

OECD. (2011). *OECD Principles and Guidelines for Access to Research Data from Public Funding.* Paris: OECD (Organisation for Economic Co-operation and Development) PUBLICATIONS.

Parsons, D., & Stern, D. (2017). Enabling Multilevel Data to be Accessible to African Students. *IASE Satellite.*

SSC. (2009). *Data Flow Organising action on Research Methods and Data Management.* Reading: Statistical Service Centre, University of Reading.

Statistical Service Centre. (n.d.). Retrieved from http://www.reading.ac.uk/ssc/resourcepage/packs.php

Statistical Services Centre. (2013). *Data Quality Checking.* Reading: University of Reading.

Stern, D. (2017). Seeding the African Data Initiative. *IASE 2017 Satellite.*

Stern, D., Ongati, N., Agure, J., & Ogange, B. (2010). Incremental Modernization of Statistics Teaching and Curriculum at Maseno University, Kenya. *8th International Conference on Teaching Statistics.* Retrieved from http://iase-web.org/documents/papers/icots8/ICOTS8_C174_STERN.pdf

Stern, R., Coe, R., & Stern, D. (2017). Still Coming Down from the Mountain. *IASE 2017 Satellite.*

Stern, R., Coe, R., Stern, D., & McDermott, B. (2013). MSc Training in Research Methods Support. *Technology Innovations in Statistics Education*. Retrieved from http://escholarship.org/uc/item/8hp227qb

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*.