DESCRIBING DATA WELL IN R-INSTAT

Maxwell Fundi[1], Lily Clements[2], David Stern[3], Roger Stern[2,3] François Renaud[1] and Alex Sananka[1]
[1]African Maths Initiative
[2]Statistics for Sustainable Development
[3]University of Reading
maxwell@africanmathsinitiative.net

In 21st century, there is an increasing need to have skills to derive meaning from the growing data around us. In Africa, too much of statistical teaching is theoretical. This leaves students with a lack of data handling skills, and often unprepared to find meaning in data. The African Data Initiative (ADI) aims to change this. A first step has been to develop R-Instat, an open-source, free software based on the increasingly used statistics software R. This paper explains some of the decisions behind R-Instat's approach to encouraging descriptive analysis. It also proposes how this could support the teaching of good descriptive statistics.

INTRODUCTION

In Kenya the teaching of statistics has been too theoretical and formulae based (Odhiambo, 2002). Despite small pockets of progress the general situation remains largely the same as when this report was written 15 years ago. Students do not encounter interesting datasets and are often taught highly theoretically with little emphasis on practical skills.

We recently administered at the African Institute for Mathematical Sciences (AIMS) in Tanzania. Through Open Data Kit (ODK) (University of Washington, 2008), we asked the 54 MSc students from 16 African countries whether they had previously used data sets, spreadsheets or statistical software in their BSc degree courses. All the students had taken multiple courses in statistics and some had single or joint statistics degrees, however, the lack of use of data in their undergraduate teaching was striking where we saw more than half had never encountered a data set in their teaching. This is explored further in the discussion.

The African Data Initiative (ADI), (AMI, 2015), attempts to change this by creating conditions conducive to spreading innovations in statistical education virally (Stern, 2014). The initial phase is to develop R-Instat, an open source front end to R, to ensure that anyone can have access to an easy to use statistics package.

R-Instat has already started creating impact https://chuffed.org/project/africandatainitiative and has been used in teaching in Maseno University, Kenya, as well as University of Tor Vergata, Rome, Italy and finally at AIMS Tanzania.

Improvements continue to be made to R-Instat in Africa with a big part of development team in Kenya. This initiative came about after collaborative work of AMI with other organisations across Africa that saw the need of better statistical literacy and hence starting with a tool to be used to teaching statistics better was key to developing this capacity for good statisticians. Beta versions of the software are available for download at African Maths website initiatives page http://www.africanmathsinitiative.net/blog/initiatives/african-data-initiative/

This paper discusses three uses of descriptive statistics, namely data exploration, describing data and explaining variability. The exploration phase is used to provide initial insights in what can be learned from the data and serves as a continuation of the quality control process. In particular it involves checking that the data fits current expectations. The process of describing data finds visualisations which relate to study objectives. Finally descriptive statistics also plays an important role in explaining variability of the data.

The paper will build on the methods section where we discuss how these three uses of descriptive statistics are implemented in R-Instat. This will include the structure of the menu system

as well as specific aspects of the implementation. In particular the extensive use of the R package *ggplot2*, (Wickham, 2016) will be presented.

Finally, as mentioned, we include a discussion emphasising the importance of including more practical descriptive analytical skills in African statistical education. We will also give some illustration to demonstrate that it is possible to engage students in statistics.

METHODS

Combining Tables and Graphs to make Describe

At the start of conceptualising the menu system in R-Instat, we considered having separate graphics and statistics menus as this is typical of most statistics packages. However, we wanted to place a high value on descriptive analysis and having a single Describe menu (Figure 1) has enabled a coherent approach by combining both tables and graphs. The menu structure of Prepare, Describe and Model is designed to encourage extensive use of descriptive analysis, before the modelling step.

The first main section of the describe menu (Figure 1) contains three submenus; the One Variable, Two Variable and Three Variable. This is to enable users to quickly obtain descriptive statistics and graphs to explore their data. R makes this very easy. For example, in **Describe > One Variable > Summarise** the default is to use the standard R summary function (figure 3). There are other options to create customised summaries. A second example from **Describe > One Variable > Graph** produces a single display with, by default, a bar chart for each factor and a boxplot for each numerical variable, with other graphs options available. Through this, quick and simple exploration can be performed to see if the results align with expectations.

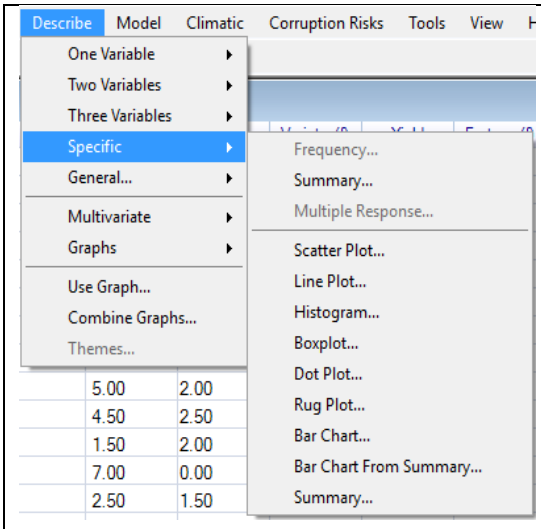

Figure 1 – The Describe menu in R-Instat. This shows the **Specific** description options.

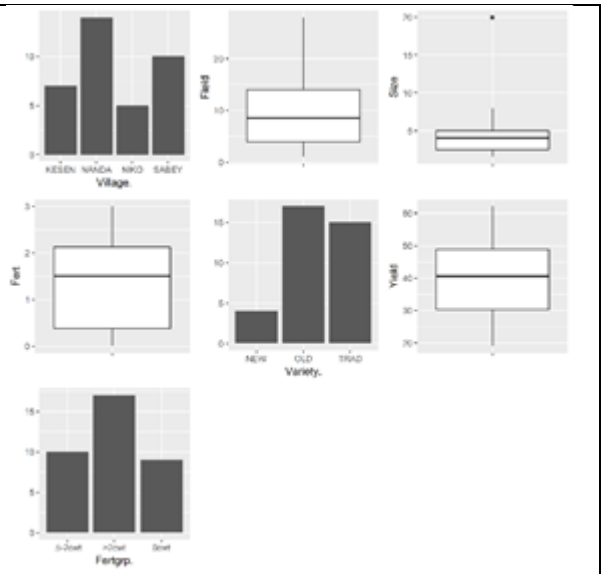

Figure 2 – Applying **One Variable > Graph** to the *Survey* dataset. A dataset available in the R-Instat Library,

```
 Village.      Field            Size            Fert          Variety.      Yield          Fertgrp.
KESEN: 7   Min.   : 1.0   Min.   : 1.50   Min.   :0.000   NEW : 4   Min.   :19.1   .5-2cwt:10
NANDA:14   1st Qu.: 4.0   1st Qu.: 2.50   1st Qu.:0.375   OLD :17   1st Qu.:30.4   >2cwt  :17
NIKO : 5   Median : 8.5   Median : 4.00   Median :1.500   TRAD:15   Median :40.5   0cwt   : 9
SABEY:10   Mean   :10.7   Mean   : 4.49   Mean   :1.444             Mean   :40.6
           3rd Qu.:14.0   3rd Qu.: 5.00   3rd Qu.:2.125             3rd Qu.:48.9
           Max.   :28.0   Max.   :20.00   Max.   :3.000             Max.   :62.1
```

Figure 3 – Applying the default values in One Variable > Summarise to the *Survey* dataset

Exploratory analysis of data with structure can be done using the **Two Variable** or **Three Variable** sub menus. These provide similar simple facilities to look at variables related to other variables.

The next submenu (Figure 1) contains **Specific** graphs and tables. The dialogs enable users to quickly produce a graph of their choice. For example the user can quickly obtain a boxplot to compare yields between villages (Figure 4). If the user is then interested in exploring this relationship further, it is possible to add a scatterplot layer to see how the yield within each village depends on the variety used (Figure 5). This can be either through further plot options in the boxplot dialog or by using the **General** sub menu.

Adding layers to plots is possible because R-Instat uses *ggplot2* (Wickham, 2010) an R package for data visualisation. This package enables complex and informative graphs to be constructed. It is at the heart of almost all graphs obtained through R-Instat.

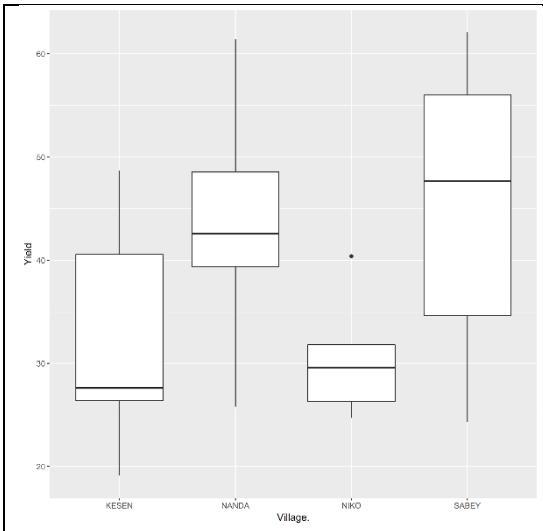

Figure 4 – A Boxplot to demonstrate the "Specific" menu

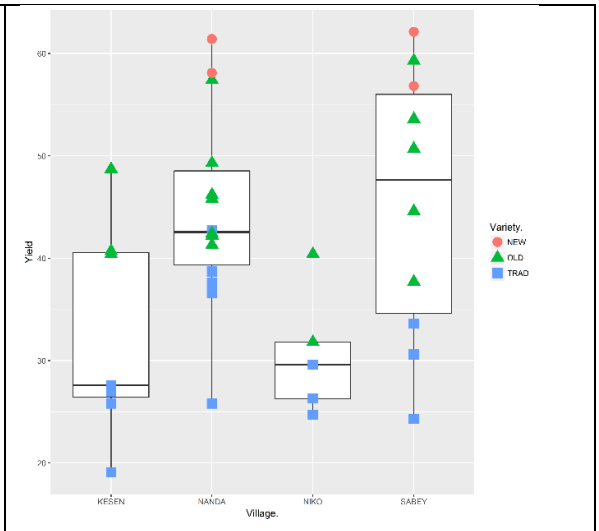

Figure 5 – Using *ggplot2* to add layers to the boxplot in R-Instat

The final section of the Describe menu (Figure 1) continues to use *ggplot2* functionality through the **Use Graph**, **Combine Graph** and **Themes** dialogs. The remaining section includes other descriptive tools, including most notably **Multivariate** from which the correlation example is shown below (Figure 7).

Building from the Grammar of Graphics

The R package *ggplot2* is an implementation of the Grammar of Graphics (Wilkinson, 2005)**.** The grammar enables graphics to be built up by breaking down components of a graphic into "data and aesthetic mappings, geometric objects, scales, and facet specification" (Wickham, 2010).

The Grammar of Graphics is a very powerful tool to describe data through visualisation. However, many people find it confusing at first. The implementation in R-Instat is appropriate for users who may or may not already understand *ggplot2*. Hence, we are aiming to balance close to full access to *ggplot2* functionality while keeping dialogs easy-to-use whenever possible.

In R-Instat we have also created a calculation and summary system, which is similar in philosophy to the Grammar of Graphics. This underlies most of the summaries and tables throughout the **Describe** menu. These two general systems help build consistency throughout the software for graphs and summaries respectively. In the future, we imagine there could be a more coherent grammar of descriptives which includes tables as well as graphs. The next section explains why we feel this is needed.

Blurred Boundaries between Tables and Graphs

In the **Describe** menu, tables and graphs tend to be separate dialogs. However, there are cases which blur this boundary. The first example is from *ggplot2* when producing graphs in a tabular manner using facets. This is essentially producing multiple graphs split by levels of a factor (Wickham, 2016). A scatterplot using facets (Figure 6) displays a table of graphs which includes a margin, this adds a further explanatory dimension to the previous boxplots (Figure 5).

Another example of how the lines between tabulation and graphics are being blurred comes from the **Multivariate > Correlations** menu**.** Here the pairwise plot (Figure 7) gives an output containing both correlation amounts and correlation graphs.

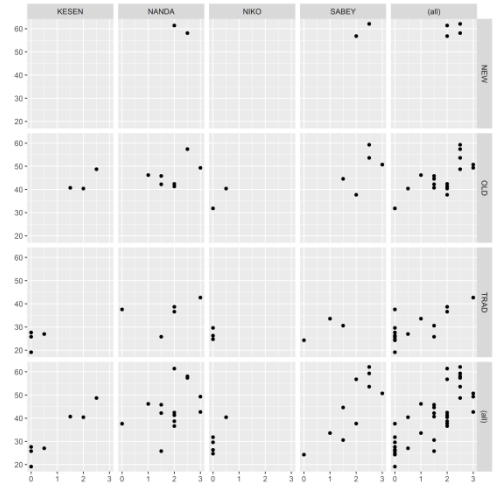

Figure 6 – Using facets with the *Survey* dataset to demonstrate the narrowing distinction between facets and graphs

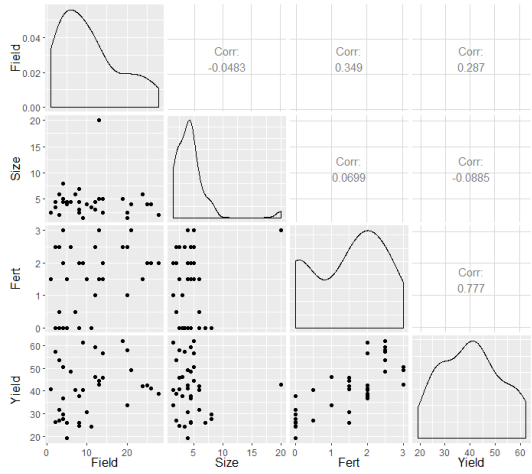

Figure 7 – A Pairwise Plot using the *Survey* Dataset.

```
          Df Sum Sq Mean Sq F value
Village.   3   1391   463.7   4.164
Residuals 32   3564   111.4      NA
          Df Sum Sq Mean Sq F value
Variety.   2   3528 1763.91   40.79
Residuals 33   1427   43.24      NA
```

Figure 8 - Descriptive ANOVA by the **Two Variables > Summarise** dialog

RESULTS

This section explains how the **Describe** menu was conceived to facilitate the processes of exploring data, describing data and explaining variability.

Exploring Data

In the exploratory phase we wish to understand the data that has been collected. Data exploration is sometimes a continuation of the quality control process and can also provide initial insights into what can be learned from the data.

We believe in the principle that for any a dataset, users should look at every variable at least once. This starts through the **One Variable** sub menu. The aim is to enable users to explore their data easily, hence encouraging this step before a more formal analysis.

The **Two** and **Three Variable** sub menus are designed to facilitate this process further.

Describing Data

Description of data should start with a review of study objectives. Some of the objectives may have to be modified based on the knowledge gained up to this point. Following this, the initial analysis provides the appropriate graphs and tables that correspond to these objectives. This is also

a time to check the role of each of the variables that were measured. Sometimes some are found to be irrelevant to the objectives and hence may not need to be collected in a future study of this type.

Descriptive analysis related to study objectives is an essential part of most studies. Often users already know the descriptive tools they would like to use to analyse their data. We have made this stage as familiar and simple as possible through the **Specific** sub menu (figure 1).

Further, good description of data helps with deciding what modelling tools might be appropriate for the data. In fact, to a trained eye, it is often possible to estimate the modelling output from the descriptive analysis.

Explaining Variability

Explaining variability is at the heart of any data analysis. Using layers and facets in *ggplot2* can help to explain variability. For example, layers can enable the user to see further variability within groups (figure 5). Similarly, facets helps the user to see step-by-step how the data alters as you move between levels of a factor variable (figure 6).

Another key component in R-Instat is the inclusion of ANOVA, where we have not only used this in the **Model** menu, but also in the **Describe** menu. This inclusion in the **Describe** menu allows an ANOVA to be used as a descriptive tool. Often users only look at the p-value of an ANOVA table, however, "Data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible" (Wasserstein & Lazar, 2016). Therefore, to encourage ANOVAs to be used to understand and explain variability rather than for hypothesis testing, we have not only included it in the **Describe** menu (figure 8), but also omitted the p-value by default in this case.

DISCUSSION

As mentioned, we surveyed 54 students from 16 countries across Africa at AIMS Tanzania. We found that more than half the students had not used a statistical package in their bachelor's degree; similarly, more than half had not used a spreadsheet and, most surprisingly, more than half had never encountered datasets in their previous teaching. The students' virtually unanimous suggestion was that their training should become more practical and "data-based". This is a suggestion we agree with. Of course, starting with data is only part of the subject of statistics, but it is important to be able to describe data and derive meaning.

Most small data sets are not so interesting, and hence many statistics courses move quickly to statistical modelling. The virtual omission of descriptive statistics is not good preparation for the real world, which is full of large and interesting problems of dealing with data. Many statistics courses need a restructure to be taught better and in turn help the students understand more. For example, it is nearly ten years since the University of Nairobi changed the structure of their teaching of statistics for their agriculture students, (Kurji, McDermott, Stern, & Stern, 2010). This provided "a restructuring of the statistics courses; offering a descriptive course in the second year and inference in the third year followed by a course on practical aspects of scientific investigation in the final year culminating in a final year project. This was as accompanied by a change from a technique based approach to a data based problem solving, approach." This change in the structure prompted some of the agriculture students to state that "Statistics is now our favourite course!".

The availability of R-Instat has the potential to address these shortcomings of teaching with the kind of resources it has and by partnering with institutions of higher learning. It comes with data sets from R-packages and other sources. This gives the opportunity for students to learn data manipulation skills (Musyoka, et al., 2017) alongside the descriptive analysis skills described in this paper.

CONCLUSION

The R-Instat **Describe** menu is structured differently from many other statistical software in a manner that encourages good practices for exploring, describing and explaining variability in data. We have used these three components via *ggplot2* to plan both the layout and content of the dialogs.

While recommending that students spend considerably longer on descriptive statistics than is currently in most Universities in Africa, this is, in no sense a critique of statistical modelling. Those ideas remain crucially important and are considered in (Kogo, Stern, Clements, Parsons, & Stern, 2017).

Having seen that students see few realistic data sets during their courses, we believed that we could change this by having new easy way for students to learn statistics and lecturers teach better. With the development of R-Instat and its inbuilt resources including the datasets from R packages, and collaboration with institutions of higher learning around Africa, we hope to be able to transform statistical teaching and learning. This will be possible by enabling students to interact with data sets well enough to get the data handling skills they need for the real world.

Now that the software can be used, we have started and we will keep working with at lecturers to shift focus from theory to a practical approach of teaching statistics. We also hope to translate the software into many languages and hence have a broad multilingual audience. More information about R-Instat and its journey can be found at https://chuffed.org/project/africandatainitiaitve

REFERENCES

AMI. (2015, November 08). *African Data Initiative*. Retrieved from African Maths Initiative: http://www.africanmathsinitiative.net/blog/initiatives/african-data-initiative-part-1-missing-tool/

Kogo, S., Stern, D., Clements, L., Parsons, D., & Stern, R. (2017). Tuning statistical modelling education to improve understanding. *IASE.* Morocco.

Kurji, P., McDermott, B., Stern, D., & Stern, R. (2010). The growing role of computers for teaching statistics in Kenya. *ICOTS8 Conference.*

Musyoka, J., Lunalo, J., Garlick, C., Ndung'u, S., Stern, D., Parsons, D., & Stern, R. (2017). Embedding Data Manipulation in Statistics Education. *IASE.* Morocco.

Odhiambo, J. W. (2002). Teaching of Statistics in Kenya. Paper presented at the 6th International Conference on Teaching Statistics (*ICOTS 6*). Cape Town, South Africa.

Wasserstein, R., Lazar, N. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician 70:2, pages 129-133.

Stern, D. (2014). Reflections on using technology to teach statistics in Kenya. *ICOTS 9.*Flagstaff. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_3D2_STERN.pdf

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

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*.

Wickham, H. (2016). Elegant graphics for data analysis. *Springer Science+Business Media*.

Wilkinson, L. (2005). *The Grammar of Graphics.* Springer.