Early Human Development xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

WASP (Write A Scientific Paper) using Excel — 1: Data entry and validation

Victor Grech

Academic Department of Paediatrics, Mater Dei Hospital, Malta

ARTICLE INFO

Keywords:
Software
Computers
Statistics
Biostatistics

ABSTRACT

Data collection for the purposes of analysis, after the planning and execution of a research study, commences with data input and validation. The process of data entry and analysis may appear daunting to the uninitiated, but as pointed out in the 1970s in a series of papers by *British Medical Journal* Deputy Editor TDV Swinscow, modern hardware and software (he was then referring to the availability of hand calculators) permits the performance of statistical testing outside a computer laboratory. In this day and age, modern software, such as the ubiquitous and almost universally familiar Microsoft Excel™ greatly facilitates this process. This first paper comprises the first of a collection of papers which will emulate Swinscow's series, in his own words, "addressed to readers who want to start at the beginning, not to those who are already skilled statisticians." These papers will have less focus on the actual arithmetic, and more emphasis on how to actually implement simple statistics, step by step, using Excel, thereby constituting the equivalent of Swinscow's papers in the personal computer age. Data entry can be facilitated by several underutilised features in Excel. This paper will explain Excel's little-known form function, data validation implementation at input stage, simple coding tips and data cleaning tools.

1. Introduction

The publication of one's research is critical for career advancement. However, even the most basic research or simplest attempt at data collection involves the gathering of datasets with different types of variables. Their analyses then allows the user to derive descriptive statistics. The latter define the properties of a given dataset, and includes tabulations of summary data as well as the creation of graphs. After introspection on the data's properties, it is then possible to compare these results with those obtained from other datasets by the processes of inferential statistics, i.e. significance testing.

This first paper will focus on data entry and forms. Real datasets utilised by this author in publications will be utilised, as well as examples used in Swinscow's original series [1].

2. Types of data

Datasets consist of variable/s that are collected for analysis. Variables may be of different types and these categories as well as specific examples are shown in Table 1 [2].

Whatever the data, since it is quantitative in nature, for the purposes of analysis, it must be tabulated, that is, entered in software in column/s. It is customary to use the first row of such a column as a header, that is, the name of the variable being input. While Excel copes with header variable names that start with a number, some packages do not and it is worth keeping this in mind, should the data need to be analysed in a different software package. For example, questionnaire answers should be entered under headers "Q1", "Q2" etc. and not as "1", "2" etc.

3. Coding

It is crucial to plan the spreadsheet so as to facilitate analysis later on. Overall, variables are best entered numerically as this is the least likely option to generate error. A perfect dataset has no blank cells and no subtotals or totals.

A coding key must be kept so as to keep track of the way in which data is coded. Such a key or keys for different variables could be entered in a separate sheet in the same workbook lest the researcher forgets what the codes stood for when the data is eventually analysed. For example, a field such as marital status: single/married/divorced/widowed could be coded as:

Single = 1
Married = 2
Divorced = 3
Widowed = 4

This will facilitate data validation implementation as will be ex-

E-mail address: victor.e.grech@gov.mt.

https://doi.org/10.1016/j.earlhumdev.2018.01.002

Table 1
Examples of types of data [1].

Quantitative	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Male or female, alive or dead (the above two are special cases: binary datasets) Blood group O, A, B, AB

Table 2
US live births for 2007–2013 by state, gender, and month and year of birth (subsection of dataset) [4].

State	Year	Month	Gender	Births
Alabama (01)	2007	January	Female	2622
Alabama (01)	2007	January	Male	2693
Alabama (01)	2007	February	Female	2496
Alabama (01)	2007	February	Male	2440
Alabama (01)	2007	March	Female	2621
Alabama (01)	2007	March	Male	2722
Alabama (01)	2007	April	Female	2462
Alabama (01)	2007	April	Male	2646

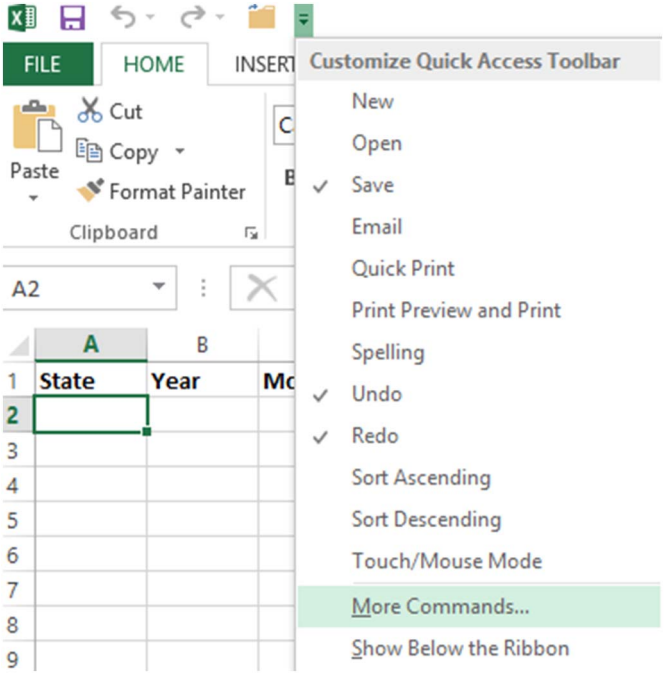

Fig. 1. Activating the Quick Access customisation options.

plained later. If this was not done but needs to be done for the purposes of data analysis, the Excel REPLACE() function can be used which has the following format:

REPLACE(old_text, start_num, num_chars, new_text)		
Old_text	Required	Target cell
Start_num	Required	Starting number in above cell (from the left) of the text to be replaced
Num_chars	Required	The number of characters in the cell that you want REPLACE to replace with new text
New_text	Required	The text that will replace characters in the cell

For example, if the value in cell A1 is "M" for male, the formula = REPLACE(A1,1,1,"1") will replace this with the number "1". The SUBSTITUTE() command may also be found useful in this respect and has the format SUBSTITUTE(text, old_text, new_text, [instance]). Instance is optional. If not supplied, all instances of old_text are replaced with new_text. SUBSTITUTE is useful to replace text based on content. REPLACE is useful to replace text based on its location.

4. Forms

Table 2 displays a subset of columns that constitute part of a large dataset, United States live births for 2007–2013 by state, gender, and month and year of birth [3]. Excel has a form entry feature which greatly facilitates the entry of data if such data were to be manually input (for example data from a new study).

The form feature is not easily accessible. The easiest place to make it accessible is to add it in Excel's Quick Access Toolbar, a one-time procedure:

- Click down arrow at the right end of the Quick Access Toolbar (Fig. 1).
- Choose More Commands (Fig. 1).
- This opens the Customize the Quick Access Toolbar dialog box.
- Instead of Popular Commands select All Commands (Fig. 2).
- Scroll down the alphabetical list to Form command (Fig. 3).
- Click on Add button then click OK (Fig. 3).

The Form button will now appear in the Quick Access Toolbar at the top left hand corner of the screen (Fig. 4).

In order to enter data as per Table 2, the following should be entered as headers, from A1 to E1: State, Year, Month, Gender, Births (Fig. 5). Click on cell A2 and click the Form button (that was obtained per bullet points above and is now present at the top left of the screen) and the following dialog will be displayed: "Excel cannot determine which row … contains column labels which are required for this command. If you want the first row … used as labels and not as data click OK." Do click OK as this is the correct choice. The form will appear (Fig. 6). As data is input in this form, Excel appropriately and automatically fills the sheet. The form is also useful for quick navigation within long columns of data.

5. Data validation

Garbage in-garbage out is a database mantra. It is crucial to ensure that any data input adheres to preset criteria, or analysis will be impossible or invalid. One way to reduce error is to use validation rules. These may be accessed from the data tab, data validation, data validation (Fig. 7).

For example, highlighting column B (year) and then restricting the