Statistics Corner: Data Cleaning-II

Kamal Kishore¹, Rakesh Kapoor²

REALITY CHECK

Microsoft Excel is one of the most routinely used spreadsheets to enter and clean data by researchers. However, the lack of professional training in Excel hinders its efficient usage. Majority of investigators are aware of basic data entry and cleaning mechanisms in Excel. However, there are numerous advanced easy-to-use functions that increase the efficiency and accuracy of data entry. The present article will be discussing four vital functions to enter and clean data efficiently. These functions are the following:

- Data form
- Data validation
- Go to special
- Protecting specific columns, worksheet, and workbook

Journal of Postgraduate Medicine Education and Research (2019); 10.5005/jp-journals-10028-1339

*"Let's not kid ourselves: the most widely used piece of software for statistics is Excel."*¹

INTRODUCTION

Empirical studies are used to support theories or generate new knowledge. However, the generated information can be flawed and wrong due to the garbage (unclean data) in, garbage (incorrect results) out (GIGO) principle. Therefore, managing an error-free data is crucial. In this regard, researchers undertake data scrubbing or data cleaning as an initial step before proceeding ahead with analysis. The anomalies such as missing data, duplicate, wrong formats, codes, and incorrect data are corrected. Ignoring data cleaning may lead to inaccurate results, delay in analysis, and wastage of resources such as time, money, and human resources. Data cleaning such as spelling corrections, syntax errors, labeling, and coding can be frustrating at times. However, it is a vital initial step before proceeding ahead for analysis.

Various researchers have criticized Excel (Microsoft Corp., Redmond, WA, USA) for its statistical analysis properties.²⁻⁴ Despite that, it remained the most widely used spreadsheets for initial data entry and elementary calculations. The cleaned data are analyzed with advanced analysis programs (such as SPSS, SAS, Stata, and R) for more reliable and advanced statistical analysis. The Excel was initially launched for Mac in 1985 and then subsequently for Windows in 1987. Since then, it has undergone many modifications and improved tremendously in terms of usage and functions. Despite its popularity, most of the researchers are not professionally trained to operate Excel. It may be that many researchers perceive Excel as a mundane tool to enter and store data. Therefore, many of the investigators are not aware of many advanced features of Excel for data entry and cleaning. Thus, in continuation of the previous "Data Cleaning-I" article,⁵ this article discusses four better ways of entering and cleaning data in Excel. The lead author of this article routinely uses these functions to input, code, clean, and protect the data. We demonstrate and discuss all four functions using Excel 2019 in Windows 10 Pro edition. However, these functions, except form, are also available for MAC users.

^{1,2}Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

Corresponding Author: Kamal Kishore, Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India, Phone: +91 9591349768, e-mail: kkishore.pgi@gmail.com

How to cite this article: Kishore K, Kapoor R. Statistics Corner: Data Cleaning-II. *J Postgrad Med Edu Res* 2019;53(4):164-167.

Source of support: Nil

Conflict of interest: None

Data Forms

The traditional way of entering and navigating data in Excel is time-consuming and error-prone. The error rate increases with the increase in the number of columns and records. However, the "forms" option in Excel simplifies and accelerates the data entry process.

A form is a dialogue box to enter data, which systematically arranges the variables in a record. The forms can be used to add, delete, modify, and search the data. Moreover, to facilitate and reduce data entry errors, data validation checks can be applied to the forms option. However, this option is not available by default in the ribbon. It needs to be activated first time by selecting it from "form tool" available under all commands options in Excel.

Let us assume an investigator collected gender, date of birth (DOB), height, systolic blood pressure (SBP), and body mass index (BMI). Subsequently, clicking the form option after selecting the top row with variable names will activate the form. The form with all the variables will appear in a vertically structured format on the top of the Excel worksheet. The newly opened form can be used to fill value at appropriate places. Figure 1 presents the steps to activate and use the form tool. A significant limitation of the form tool is that it can be used to collect a maximum of 32 variables for a record.

Data Validation

Data validation in Excel is a useful tool to control and restrict the data in specific rows or columns. It helps in ensuring the consistency and accuracy of data at the entry stage. It prevents users from entering wrong data. The implementation of this function gives the user a default or custom message at the data entry stage.

Serial	Gender	DOB	Height	SBP	BMI
1	Male	10-12-1985	174	120	22.1
2	Female	09-09-1980	155	125	22.5
3	Male	01-01-1975	174	134	23
4	Male	05-07-1986	161	140	24.7
5	Female	10-10-1979	145	115	23.6
6	Female	02-02-1981	150	127	24.7
7	Male	03-03-1982	179	130	25.8
8	Female	24-04-1976	160	135	21.6
9	Female	05-07-1983	157	139	23.4
10	Male	18-08-1981	162	129	15.2
11	Male	18-07-1978	156	123	22

Fig. 1: Steps to activate and usage of "Form Option" in Excel

Figure 2 illustrates the stepwise process of setting up data validation in Excel for the DOB column. Panel A shows the 'Data Validation' dialog box with 'Date' selected as the validation criteria. Panel B shows the 'Input Message' tab where a message is configured to appear when the cell is selected. Panel C shows the 'Error Alert' tab where a warning message is configured to appear when invalid data is entered. Panel D shows the resulting spreadsheet where the DOB column is highlighted in yellow, and a warning message box is displayed, indicating that the data must be in DD-MM-YYYY format.

Figs 2A to D: Stepwise display of data validation steps in Excel; (A) Select the validation criteria after selecting the respective column; (B) A hint regarding the standard format of data entry; (C) Select an error alert for wrong data entry; (D) Visualizing data display function in action

There are eight options (any value, whole number, decimal, list, date, time, text length, and custom) to select from the data validation tool. The default option during data entry is the any value option. An advantage of this option is to enter any data in any cell. However, Excel is notorious for intermixing formats (such as number, text, and dates).

There are several documents highlighting formatting issues in Excel.^{4,6,7} Therefore, it is desirable to select a specific option such as date, text, number, or currency from various alternatives available in Excel. Data validation has three components, i.e., settings, input message, and error alert for a user to work. A systematic visual framework of data validation for the number is displayed in Figure 2. Similarly, the options such as text, date, formula, and currency can be validated.

Go to Special and Duplicates

The flexibility of storing data in different formats (such as text, date, number, and formula) in Excel is perhaps one of the biggest strengths and challenges. Intermixing of various formats (such as date and number), blank cells, rows, and columns poses a significant threat to data validity. These threats escalate with the increase in the number of records and variables. Manually looking at each cell across rows and columns is not a viable option.

The "Go to Special" function in Excel is a handy tool in these kinds of situations. A single click of a button is sufficient to replace all the blank cells. Similarly, one-click identifies and replaces columns with formulas as having texts or numbers all at once. "Go to special" function is available in the find and select drop-down menu under the home tab option. Figure 3 displays the various options available under the "Go To Special" function.

Fig. 3: Segregating numerals and text with the help of “Go To Special” function

Figs 4A to F: Systematic steps to protect sheet, workbook and selected columns; (A) Protecting a specific sheet in a workbook; (B) Protecting the whole workbook; (C) Click hide option to hide the columns; (D) 1st step in protecting selected columns; (E) 2nd step in protecting selected columns; (F) Final step in protecting selected columns

Data Protection

Many a time the investigators are interested in protecting the entire workbook, worksheet, or specific columns in the Excel sheet. The whole workbook can be protected by first selecting the review tab in the ribbon and then subsequently by selecting the protect workbook option. Similarly, the protect sheet option can be selected under the review tab to protect the sheet.

There are at least two ways to protect a subset of columns in an Excel worksheet. First and foremost is to hide the specific columns. Initially, select the range of columns to be concealed. Subsequently, right-click to activate and select hide options. However, this is a temporary measure. Anybody can make the changes in the columns after unhiding it.

The second choice is to protect the specific columns with a password. In the first step, select the range of columns not to be protected. In the second step, select format cells after right-clicking. In the third step, go to the protection tab and uncheck the "Locked" option. Finally, go to the protect sheet in the "Review tab" and activate the protection with password. This process will protect all the columns except the range of columns selected in the first step. This process will need the password to edit the protected columns. Figure 4 displays the various processes of protection of data.

CONCLUSION

The Excel is widely used to enter, code, and clean the data. However, the lack of professional training in Excel leads to a reduction in the

efficiency, accuracy, and protection of data. Therefore, the routine usage of four advanced easy-to-use Excel options proposed in the article will help researchers to accurately and efficiently capture and protect data.

ACKNOWLEDGMENTS

Authors would like to thank Dr Nipun Verma and Dr Rahul Mahajan for reviewing this article..

REFERENCES

1. Ripley BD. Statistical methods need software: a view of statistical computing. In: Presentation RSS Meeting; 2002.
2. Mlard G. On the accuracy of statistical procedures in Microsoft Excel 2010. *Comput Stat* 2014;29(5):1095–1128.
3. McCullough BD. Special section on Microsoft Excel 2007. *Comput Stat Data Anal* 2008;52(10):4568–4569.
4. Elliott AC, Hynan LS, Reisch JS, et al. Preparing data for analysis using Microsoft Excel. *J Investig Med* 2006;54(6):334–341. DOI: 10.2310/6650.2006.05038.
5. Kishore K, Kapoor R, Singh A. Statistics corner: data cleaning-I. *J Postgrad Med Educ Res* 2019;53(3):130–132.
6. Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. *Genome Biol* 2016;17(1):177. DOI: 10.1186/s13059-016-1044-7.
7. Zeeberg BR, Riss J, Kane DW, et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics* 2004;5:80. DOI: 10.1186/1471-2105-5-80.