Institute for Health and Society, Medical College of Wisconsin

MEDICAL COLLEGE OF WISCONSIN

Datum

NEWSLETTER

Division of Biostatistics

Data entry in Excel

Features and tips to reduce errors

Aniko Szabo, PhD

Introduction

Microsoft Excel remains a popular data entry tool, despite the availability of more reliable alternatives, such as RedCAP (MCW is a RedCAP partner). The main difference between dedicated data collection tools and Excel is that the latter does not help the user to enter the data consistently and in a form that is convenient for future analysis. For example, RedCAP always shows the name of the variable that is being entered right next to the entry box, and it would not let the user enter characters for a numeric outcome or an invalid date. It will also export data in a format that can be conveniently imported into statistical software. However, with self-restraint and use of some of less-known Excel features, the reliability and ease of use of Excel can be greatly improved. In this article, we will provide some tips on doing that. The screenshots use Excel 2010, but other versions work similarly.

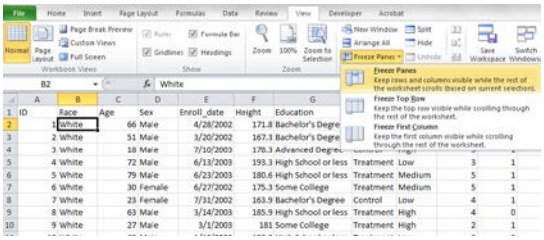
Data arrangement

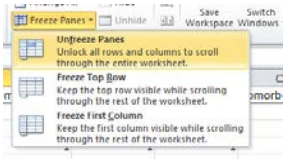
In the *Database Basics* article (Datum Vol 20 No.1), Dan Eastwood described the general principles of setting up a data collection instrument in Excel. We always recommend that data be set up in a rectangular format, with rows representing separate observations/subjects and the columns representing variables. The first row should contain short, but descriptive variable names; more detailed explanations about each variable, including units, coding conventions, etc. can be placed on a separate "Data Dictionary" worksheet. Each row should be self-contained, so rearranging the order of the rows (by sorting, for example) should not result in any loss of information. As an example, information about the treatment group should be repeated in every row and not just once as a heading or a merged cell. Cell shading and text formatting (colors, bolding, etc.) should not be used to convey information as these attributes

are lost after importing the data into statistical software; instead use a separate column with a text or numeric indicator.

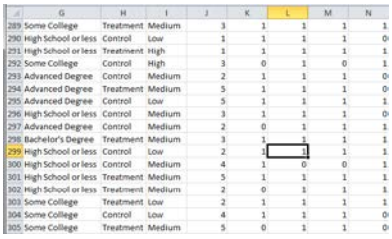
Seeing the variable names and subject IDs

If the data set has more than a few subjects and variables, it is often difficult to tell which subject and which variable does any given cell refer to, especially if the first row or the first column have "scrolled" out of view (as in Figure 1a). The "freeze panes" feature allows the user to specify that some rows and/or columns should remain visible when scrolling. In Figure 1b





this feature has been activated, so that the first column with patient IDs and the first row with variable names are always visible (separated by a thin black line), so it is easy to see that the selected cell refers to the Comorbidity_2 variable of subject #298.



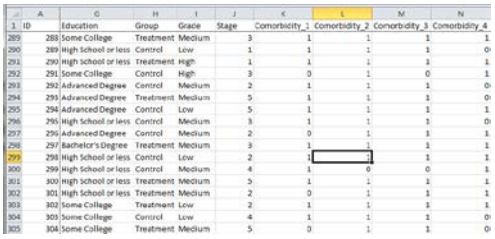


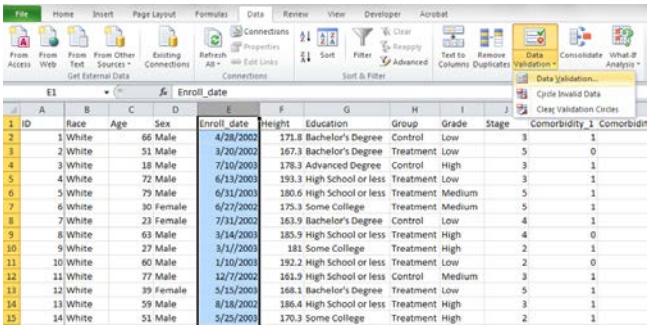
Figure 2 shows how to "freeze panes". The first step is to select the anchor cell for the freezing action: the top left cell in the area of that should not be frozen. So to freeze one row and one column, the second cell in the second row (B2) is selected in Figure 2a. The "Freeze Panes" button is located on the "View" menu tab, and should be clicked only when the appropriate anchor cell has been selected. It is also possible to freeze just the top row or first column by using the appropriate options. Once the desired rows/columns have been locked, the "Freeze Panes" menu item becomes "Unfreeze Panes", which can be used to deactivate this setup (Figure 2b).

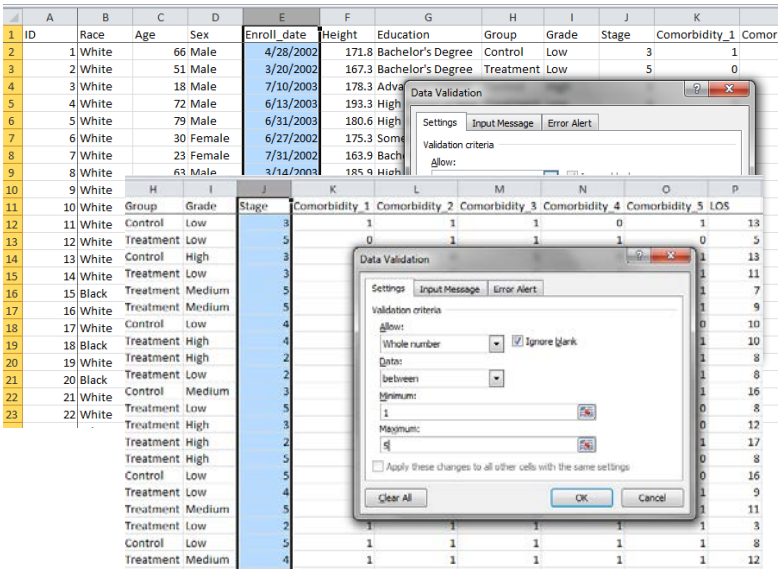
Data types

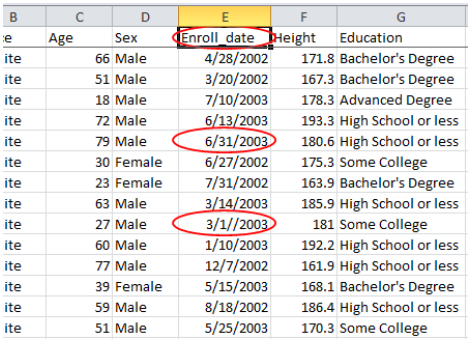
A common occurrence in Excel spreadsheets is the presence of invalid or inconsistent data, whether due to a typo, column shift, or user misunderstanding. Common examples include text notes among numbers, impossible dates or codes, and inconsistent spelling. Excel has data

validation tools that allow checking of values that have already been entered, or stop the user from entering invalid values. Filtering is another feature to check quickly for inconsistent entries.

Consider the example in Figure 3. The highlighted column should contain dates, but there are two subtle typos. They can be easily detected by telling Excel that the values in this column should be dates. In the "Data Validation" menu of the Data tab one can specify the type of possible values for the highlighted area (here the entire column E has been highlighted). By selecting "Date" (Figure 3b) and then selecting the "Circle Invalid Data" menu item, invalid values are highlighted (Figure 3c). Now we can easily see the invalid entries: there are only 30 days in June, so 6/31/2003 is invalid, while 3/1//2003 had an extra slash. After setting this validation, attempt to enter non-date values in column E will result in an error message.





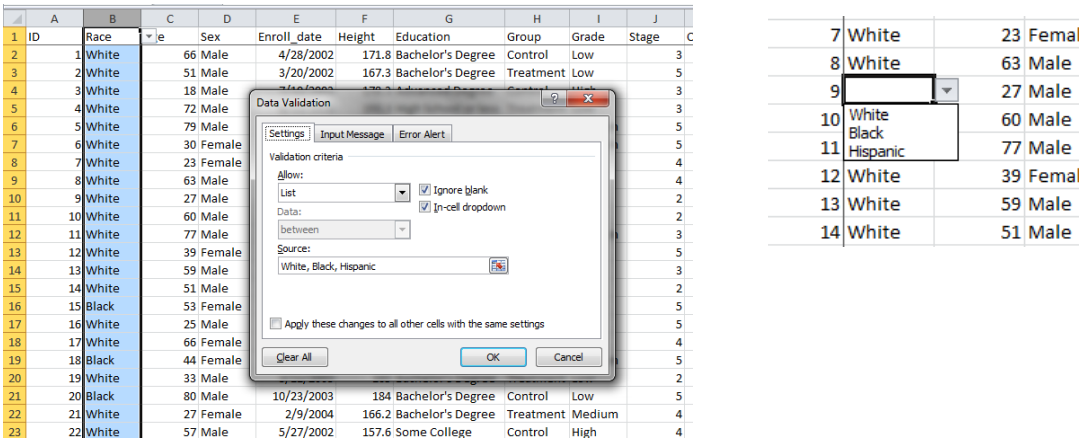


Using the same "Data Validation" menu we can specify that the values in column J, which represent disease stage, should be whole numbers between 1 and 5, avoiding illegal entries such as "2b" or "0" (assuming that is the intention).

Multiple choice dropdown

The data validation feature can also be used to ensure consistent entry of categorical variables such as race or sex. A common approach is to use integer coding for such variables (eg 1=male, 2=female), however such conventions are error-prone as they are difficult to remember during data entry. In fact, modern statistical software has no difficulty handling character variables as long as they are entered consistently with the same spelling and capitalization.

Figure 4 shows how to set up a dropdown box of possible options for data entry for the race/ethnicity variable. Such a list makes it easy to remember the possible options and avoid typos. In the same "Data Validation" menu that we saw before, we would select the "List" validation option, and either enter the possible values separated by comma in the "Source" box, or use the red arrow to highlight a range of cells (on another worksheet) in which all the possible values are listed.



Summary

As this article shows, Excel provides a variety of tools to simplify data entry. We recommend showing the planned data collection sheet to a statistician before starting data entry, perhaps with a few filled rows, to ensure that the data can be easily utilized for statistical analyses at the end of the project.