Data Organization in Spreadsheets

Types of Data

Down the road when doing data correction, organizing data, and doing analyses it will be essential to classify data according to their 'type'. It can also help with data entry, which is why we will introduce some of these types here:

1. Nominal aka Factor: categories or groups, such as [`apple`, `orange`], [`trumpet`, `flute`, `violin`]
2. Ordinal aka Ordered Factor: groups where there is an order: [`first>second>third`], [`small<medium<large`]. Note that this order doesn't imply quantitative value, e.g., One is not stating that `medium` is twice the size of `small` or that `large` is twice the size of `medium`.
3. Character: [`a`, `gnv`, `mexico`, `Inigo Montoya`]
4. Numeric (real or decimal): `2`, `15.5`
5. Integer: [`1`,`2`,`3`]
6. Logical: [`True`, `False`]
7. Complex: `1+4*i`
8. Interesting case: what category are [`red`, `orange`, `green`, `blue`]? We usually treat it as Nominal, but it is actually Ordinal - the colors represent wavelengths on the visible light spectrum (650, 600, 550, and 450 nm respectively). If you were recording the wavelength itself, it would be Numeric.
9. *For more on different categories of data you can watch this video on LinkedIn Learning*. More information on how to log on to LinkedIn Learning as a UF Affiliate is found here

Spreadsheets

Spreadsheets are ok for data entry, but they have some features that make it easy to do terrible, terrible things. People often use spreadsheets for much more, including calculations, statistical analyses, and creating tables or figures for publications and presentations.

After you have entered data in a spreadsheet:

> **Don't do calculations or data correction.**
>
> **I implore you.**
>
> **Don't. Please**

There are several reasons why. **(A)** The 'drag-and-drop', menu-driven nature of spreadsheet programs makes it very difficult (or impossible) to replicate your steps, much less those of another person. This means you can't easily find where mistakes were made, and if you have to reconstruct an analysis or figure you have to start from from the very beginnng. This is extremely tedious. **(B)** Furthermore, when doing calculations in a spreadsheet it is easy to accidentally apply a slightly different formula to multiple adjacent cells. It is easy to introduce mistakes. **(C)** Finally, at some point during your data correction or analyses - probably without even realizing it - you will make a mistake either 'sorting' or trying to fill in cells with 'copy-drag-drop-paste'.

This will potentially ruin several days of your life (or more) while you try to fix it (assuming you realize you made this mistake, which people often don't).

When using spreadsheets:

- Make your data `tidy`
  - Spreadsheets should be a rectangle, with only rows and columns.
  - Each column is a different variable (a thing you are measuring, like 'weight' or 'temperature').
  - One row per observation. Each cell has only one value.
- Column headers: Use short meaningful column names with no spaces or special characters. Don't start column names with numbers. Record units in column headers.
- Use consistent names, abbreviations/codes, and capitalization.
- Use good null values (not -999, blanks ok, some prefer `NA` or similar but this can be language specific).
- Write dates as YYYYMMDD. Better still have separate columns for Year, Month, and Day.
- don't enter the same data on multiple spreadsheets: Use one for each category of data to avoid duplicated data and to simplify corrections (e.g., taxonomy).
- Avoid using multiple tables within one spreadsheet.
- Avoid spreading data across multiple tabs (but do use a new tab to record data cleaning or manipulations).
- Record zeros as zeros.
- Use an appropriate null value to record missing data.
- Don't use formatting to convey information or to make your spreadsheet look pretty.
- Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.
- Remember that data format and excel defaults can vary by region. For example, depending on the part of the world where a user is based, the default value for the decimal and thousands operator could be a , (comma) or a . (period); some regions use mm-dd for dates while others use dd-mm.

Take-home Messages

1. **Once you are done with data entry, save it as 'read only' and make *all* corrections using scripting!**
2. **Entering data in tidy format will make it much easier to analyze.**
3. **Collecting data in tidy format makes it easier to enter data in tidy format.**

Tools & Resources

1. Data Validation in Google Sheets: blog post and video tutorial. A pdf version is available for download here.

2. Why not bypass spreadsheets lile Excel and use a csv editor like [Comma Chameleon][https://comma-chameleon.io/] instead? CC and other csv editors allow you to enter data in the same way - into cells, by adding and removing rows - and then export your file. But that's about it, which means you can't do many of the things (e.g., calculations, color in cells) that cause problems down the road.
3. More advanced users comfortable with R can also look into Data Curator, with whch you can create and edit tabular data from scratch or from a template, open Microsoft Excel and CSV files, and automatically correct common problems found in these and other file types.

Sources for this lesson

1. DataONE Community Engagement & Outreach Working Group (2017) "Data Quality Control and Assurance". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/05_qaqc/index on Aug 31, 2020
2. DataONE Community Engagement & Outreach Working Group (2017) "Data Entry and Manipulation". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/04_entry/index on Aug 31, 2020
3. Philip Woodhouse, Gert Jan Veldwisch, Daniel Brockington, Hans C. Komakech, Angela Manjichi, Jean-Philippe Venot. 2018. SAFI Survey Results. https://figshare.com/articles/dataset/SAFI_Survey_Results/6262019 doi:10.6084/m9.figshare.6262019.v4
4. Chris Prener, Trevor Burrows (Eds.). Data Carpentry: Data Organization in Spreadsheets for Social Scientists. https://datacarpentry.org/spreadsheets-socialsci/
5. Peter R. Hoyt, Christie Bahlai, Tracy K. Teal (Eds.), Erin Alison Becker, Aleksandra Pawlik, Peter Hoyt, Francois Michonneau, Christie Bahlai, Toby Reiter, et al. (2019, July 5). datacarpentry/spreadsheet-ecology-lesson: Data Carpentry: Data Organization in Spreadsheets for Ecologists, June 2019 (Version v2019.06.2). Zenodo. http://doi.org/10.5281/zenodo.3269869
6. Ernest, Morgan; Brown, James; Valone, Thomas; White, Ethan P. (2017): Portal Project Teaching Database. figshare. https://doi.org/10.6084/m9.figshare.1314459.v6