# LAS 6292: Data Types and Data Organization

updated: 2021-01-27

## Pre-Class Preparation (Instructor):

### Send in an email to students:

- Post the pre-class announcement regarding computer and spreadsheets.

### Bring to Class:

- Snacks
- Flip charts and markers
- Dry write markers
- Tent cards for student names

## Objectives and Competencies:

- Be able to identify different categories of data
- Learn best practices for data entry
- Recognize and avoid common problems with data entry and formatting in spreadsheets
- Learn and be able to implement 'Tidy' format for data tables in spreadsheets
- Identify problems with and approaches for proper handling of dates in spreadsheets
- Learn how to export data from spreadsheets in open format

## Pre-class Preparation (Students):

### Readings

1. Tesi, W. 2020. An Outdated Version of Excel Led the U.K. to Undercount COVID-19 Cases. Slate. [read online] [download pdf]
2. Stolberg et al. 2020. CDC Test Counting Error Leaves Epidemiologists 'Really Baffled'. NY Times.
   [read online] [download pdf]
3. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10.
   [read online] [download pdf]
4. Johnson, B. D., Dunlap, E., & Benoit, E. (2010). Organizing "mountains of words" for data analysis, both qualitative and quantitative. Substance Use & Misuse, 45(5), 648-70.

[read online] [download pdf]

**Online Lectures:**

- None

## Class Outline

### Topic 1 Overview: Types of Data (10 min)

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`]? Though we usually treat it as Nominal, it is actually Ordinal because 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 these data types you can watch this video*

### Topic 2 Overview: Spreadsheets (10 min)

Spreadsheets are ok for data entry, but they have some features that make it easy to do terrible, terrible things. People also end up using them for much more: creating tables for publications, calculations, calculations & statistics, figures.

**Don't.**

**I implore you.**

***Please* don't.**

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 anyone else's. 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's 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.

Much of your future as a researcher will be spent cleaning and correcting data, but you can reduce the time spent on this task (and the associated stress) considerably by implementing some good practices from the start. To start developing these good habits we will to take a look at some spreadsheets, identify the things that people should ***not*** be doing, and then determining what they should be doing instead.

**Breakout & Return Results**

**Breakout (15 min):** Take a look at these these spreadsheet files and how the data are organized.

- SAFI_messy.xlsx
- unity-portal-data.xlsx

Now write down answers to the following, bearing in mind the `tidy` principles about which you read in Broman and Woo (2018):

1. What problems can you identify with the way these data are entered/organized?
2. How would you correct each of these issues? Could these data easily be imported into a programming language or a database in its current form?
3. Dates (or things that look like dates) are especially problematic in Excel. Open the file `dates.xlsx` and do the following:
    a. enter the following dates into the column labeled `date_1`. Be sure to type them in exactly as they are written:
    - `7-2-21`
    - `2 july 2021`
    - `july 2, 2021`
    - `july 2,2021` [no space between the comma and 2021]
    - `07-02-21`
    - `7/2/21`
    - Jan 5, 1900
    - Dec 5, 1899

        i) How did these look in the cell?
        ii) Was this the same as what you typed in?
        iii) Why would these issues be a problem for data organization and analysis?

    b. Next enter the dates above into the column labeled `date_2`. Again, be sure to type them in exactly as they are written.

i) what ewas different about the way the data are recorded?
ii) can you figure out why?

c. What would you do to enter dates into Excel in a way that avoids the issues observed above?

4. Export the as a .csv file ('SAFI_messy.csv) after clicking the "OK" when warning box pops up. Now reopen it. What happened? You can find a guide to saving your file in .csv format and why that is a good idea on this website.

**Returning results & Take-home message (35 min)**

1. Alternating between groups, see if the two groups can come up with these ways to :

    - 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.
    - Don't combine multiple pieces of information in one cell.
    - Place comments in a separate column.
    - Record units in column headers.
    - Include only one piece of information in a cell.
    - Avoid spaces, numbers and special characters in column headers.
    - Avoid special characters in your data.
    - Dates
        - The reason dates in Excel are so weird is that it is accounting software. It counts the days from a default of December 31, 1899, and thus stores July 2, 2014 as the serial number 41822. This is so one can can easily calclulate "days from a given date" for accounting purposes (like invoicing) by adding "date+XX days". * Furthermore, 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.
    - Finally, remember that data format and excel defaults can vary by region. Forexample, 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. Use either ISO standard dates like `20180702` or better still split month, day, and year into distinct columns.

2. **Take-home message:** Make your data `tidy`.

    - They should be a rectangle, with only rows and columns.
    - Each column is a different variable (a thing you're measuring, like 'weight' or 'temperature').
    - One row per observation. Each cell has only one value.
    - Use short meaningful column names with no spaces

- 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 YYYY-MM-DD. Better still have separate columns for Year, Month, and Day
- Use one table for each category of data to avoid duplicated chunks of data and to simplify corrections (e.g., taxonomy). Including all data can lead to errors and takes up more space. It is also faster to enter (no repretitive typing or cut-paste).
- Once you are done with data entry, leave the raw data raw - don't change it! Save it as 'read only' and make corrections using scripting!

  **Tidy data are ideally suited to working with computers.**

  **Entrering data in tidy format will make it much easier to analyze.**

  **Collecting data in tidy format makes it easier to enter data in tidy format.**

3. Set up the importance of scripting and reproducibility with a live example of going from dirty to clean data with R.

**Break (10 min)**

**Data Project Overview (15 min)**

Introduce the semester data Project.

**Breakout 2 (45 min)**

**The goal of this breakout** is to learn some ways to minimize the number of mistakes when entering data. **First**, watch the following video (11 min) on 'Data Validation in Excel'. **Second**, open this webpage on 'Quality Assurance and Control in Excel'. It covers the same material, so it's a handy reference to have open during the exercise.

Once you have watched the video, set up a tidy data entry sheet for the Portal data used in Breakout 1:

1. Create a spreadsheet in Excel for data entry. It should have five columns, in which you will be recording (1) the date of observations, (2) the site in which the observations were conducted, (3) the species captured, (4) the mass of each animal, and (5) the length of each animal.
2. Set the following data validation criteria to prevent invalid data from getting entered:
   a. The Date column should be set so that it does *not* convert dates to other formats.
   b. Use data validation so that Site can only be one of the following A1, A2, B1, B2.
   c. Set the error message on this validation criteria to provide information on what the valid values are.

d. Use data validation so that Species can only be one of the following: *Dipodomys spectabilis*, *Dipodomys ordii*, *Dipodomys merriami*.
e. Set the error message on this validation criteria to provide information on what the valid values are.
f. Use data validation so that Mass can only be a decimal greater than or equal to zero but less than or equal to 500.
g. Set the error message on this validation criteria to provide information on what the valid values are.
h. Length should be an integer (i.e., a whole number) between 1 and 10.
i. Set the error message on this validation criteria to provide information on what the valid values are.

3. Check that the validation rules and data formatting are working by entering some data in the cells
4. Save this file as data_entry_form.csv and submit it via the Canvas website as Homework wk3.

**Free Time**

There are 40 min remaining that can be used to install R, meet with students about their data sets for the semester projects, etc.

## Tools & Resources

1. Data Validation in Google Sheets: blog post and video tutorial

## Sources

The information above is drawn from a number of sources, including:


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