

Spreadsheet Best Practices

Spreadsheets are useful for quick and dirty analyses, but...

- Not reproducible, difficult to follow logic buried in formulas
- One mistake can damage reputations:
 - E.g., landmark paper in economics (Reinhart & Rogoff, 2011). Could end up in retraction.

Sample Day	Oxic 1	Oxic 2	Anoxic 1	Anoxic 2
0.6	442	452	417	342
4.8	572	582	327	362
8.1	592	602	357	347
10.8	567	577	362	312
14.8	592	602	317	212
21.8	602	582	372	272
25.0	577	572	317	247
29.0	582	577	327	267
32.0	592	582	332	282
35.7	577	587	312	297
43.0	597	582	312	252
46.7			332	202
51.7			297	247
55.8			242	202

Annotations:

- 1 row per record (points to the first row of data)
- Clear column names – include a readme, header should be 1 row (points to the column headers)
- 1 value per cell (points to the data cells)
- Document what blank means or Explicitly indicate missing data – “NULL”, “N/A”, etc. Don’t use 0, -9999 etc. (points to the empty cells)
- What do values correspond to? Is it a abbreviation? – include codebook (points to the data cells)
- 1 column per var. (points to the column headers)

• See handout for more

< (./01-format-data/)

Data Organization in Spreadsheets (../)

Formatting problems

> (./03-dates-as-data/)

1 Overview

Teaching: 20 min

Exercises: 0 min

Questions

- What are some common challenges with formatting data in spreadsheets and how can we avoid them?

Objectives

- Recognize and resolve common spreadsheet formatting problems.

Authors: Christie Bahlai, Aleksandra Pawlik

Common Spreadsheet Errors

This lesson is meant to be used as a reference for discussion as learners identify issues with the messy dataset discussed in the previous lesson. Instructors: don't go through this lesson except to refer to responses to the exercise in the previous lesson.

There are a few potential errors to be on the lookout for in your own data as well as data from collaborators or the Internet. If you are aware of the errors and the possible negative effect on downstream data analysis and result interpretation, it might motivate yourself and your project members to try and avoid them. Making small changes to the way you format your data in spreadsheets, can have a great impact on efficiency and reliability when it comes to data cleaning and analysis.

- Using multiple tables
- Using multiple tabs
- Not filling in zeros
- Using problematic null values
- Using formatting to convey information
- Using formatting to make the data sheet look pretty
- Placing comments or units in cells
- Entering more than one piece of information in a cell
- Using problematic field names
- Using special characters in data
- Inclusion of metadata in data table
- Date formatting (./03-dates-as-data/)

Using multiple tables

A common strategy is creating multiple data tables within one spreadsheet. This confuses the computer, so don't do this! When you create multiple tables within one spreadsheet, you're drawing false associations between things for the computer, which sees each row as an observation. You're also potentially using the same field name in multiple places, which will make it harder to clean your data up into a usable form. The example below depicts the problem:

Christie Bahlai and Tracy Teal (eds): "Data Carpentry: Data Organization in Spreadsheets Ecology lesson." Version 2017.04.0, April 2017, <http://www.datacarpentry.org/spreadsheet-ecology-lesson/>

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AI			
2	Lake site May 29, 2012				29-May				Lake site Jun 12, 2012				12-Jun				Lake site Jun 19, 2012				19-Jun				Lake site Jun 26, 2012				26-Jun						
3					avr	SEM		plot	bug1	bug2	avr	SEM		plot	bug1	bug2	gen eral	avr	SEM		plot	bug1	bug2	gen eral	avr	SEM		plot	bug1	bug2	gen eral	avr	SEM		
4	1	T1	1	1	2	T1	2.6	0.51	1	T1	6	85	91	T1	30.4	15.47126	1	T1	17	80	97	avr	SEM	1	T1	52	191	243	T1	141.6	60.313				
5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	13	21	T2	0.2	0.2	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T2	0.2	0.2			
6	3	T1	1	3	4	control	0.2	0.2	3	T1	11	0	11	control	0.6	0.6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	control	0	0			
7	4	T1	1	0	1				4	T1	0	6	6				4	T1	0	14	14	control	0.4	0.244949	4	T1	0	39	39	control	0	0			
8	5	T1	0	5	3				5	T1	3	20	23				5	T1	10	70	80			5	T1	4	96	100							
9	6	T2	1	0	1				6	T2	0	0	0				6	T2	1	7	8			6	T2	0	1	1							
10	7	T2	0	0	0				7	T2	0	0	0				7	T2	0	1	1			7	T2	0	0	0							
11	8	T2	0	0	0				8	T2	1	0	1				8	T2	0	0	0			8	T2	0	0	0							
12	9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0			9	T2	0	0	0							
13	10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0			10	T2	0	0	0							
14	11	control	0	0	0				11	control	0	0	0				11	control	0	0	0			11	control	0	0	0							
15	12	control	0	0	0				12	control	0	0	0				12	control	0	0	0			12	control	0	0	0							
16	13	control	0	0	0				13	control	0	0	0				13	control	0	0	0			13	control	0	0	0							
17	14	control	0	0	0				14	control	0	0	0				14	control	0	1	1			14	control	0	0	0							
18	15	control	1	0	1				15	control	3	0	3				15	control	0	1	1			15	control	0	0	0							
19																																			
20																																			
21	Barn site May 29, 2012				29-May				Barn site Jun 12, 2012				12-Jun				Barn site Jun 19, 2012				19-Jun				Barn Site Jun 26, 2012				26-Jun						
22		plot	bug1	bug2	gen eral				plot	bug1	bug2	gen eral				plot	bug1	bug2	gen eral			plot	bug1	bug2	gen eral			avr	SEM						
23	1	T1	3	8	6				1	T1	21	0	21				1	T1	5	0	5			1	T1	0	0	0							
24	2	T1	1	4	5				2	T1	36	74	110				2	T1	65	502	567			2	T1	44	2057	2101			T1	431.8	417.33		
25	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	30.6	20.10124	3	T1	10	7	17	T1	119.4	111.92882	3	T1	12	20	32	T2	0.4	0.4			
26	4	T1	0	0	0	T2	0.4	0.245	4	T1	7	0	7	T2	1	0.774597	4	T1	0	6	6	T2	5	2.1908902	4	T1	0	16	16	control	1.2	0.5831			
27	5	T1	0	1	1	control	1	0.316	5	T1	2	0	2	control	2.2	1.714643	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10						
28	6	T2	0	0	0				6	T2	1	0	1				6	T2	0	8	8			6	T2	0	0	0							
29	7	T2	0	0	0				7	T2	0	4	4				7	T2	0	12	12			7	T2	0	0	0							
30	8	T2	0	1	1				8	T2	0	0	0				8	T2	0	0	0			8	T2	0	0	0							
31	9	T2	0	1	1				9	T2	0	0	0				9	T2	3	0	3			9	T2	0	0	0							
32	10	T2	0	0	0				10	T2	0	0	0				10	T2	2	0	2			10	T2	0	2	2							
33	11	control	0	0	0				11	control	1	0	1				11	control	0	5	5			11	control	0	2	2							
34	12	control	0	1	1				12	control	0	0	0				12	control	1	1	2			12	control	1	0	1							
35	13	control	0	1	1				13	control	0	0	0				13	control	0	0	0			13	control	0	0	0							
36	14	control	1	1	1				14	control	8	1	9				14	control	0	5	5			14	control	0	3	3							
37	15	control	2	2	2				15	control	0	1	1				15	control	0	2	2			15	control	1	0	0							
38																																			
39																																			

In the example above, the computer will see (for example) row 4 and assume that all columns A-AF refer to the same sample. This row actually represents four distinct samples (sample 1 for each of four different collection dates - May 29th, June 12th, June 19th, and June 26th), as well as some calculated summary statistics (an average (avr) and standard error of measurement (SEM)) for two of those samples. Other rows are similarly problematic.

Using multiple tabs

But what about workbook tabs? That seems like an easy way to organize data, right? Well, yes and no. When you create extra tabs, you fail to allow the computer to see connections in the data that are there (you have to introduce spreadsheet application-specific functions or scripting to ensure this connection). Say, for instance, you make a separate tab for each day you take a measurement.

This isn't good practice for two reasons: 1) you are more likely to accidentally add inconsistencies to your data if each time you take a measurement, you start recording data in a new tab, and 2) even if you manage to prevent all inconsistencies from creeping in, you will add an extra step for yourself before you analyze the data because you will have to combine these data into a single datatable. You will have to explicitly tell the computer how to combine tabs - and if the tabs are inconsistently formatted, you might even have to do it manually.

The next time you're entering data, and you go to create another tab or table, ask yourself if you could avoid adding this tab by adding another column to your original spreadsheet.

Your data sheet might get very long over the course of the experiment. This makes it harder to enter data if you can't see your headers at the top of the spreadsheet. But don't repeat your header row. These can easily get mixed into the data, leading to problems down the road.

Instead you can freeze the column headers so that they remain visible even when you have a spreadsheet with many rows.

Documentation on how to freeze column headers (<https://support.office.com/en-ca/article/Freeze-column-headings-for-easy-scrolling-57ccce0c-cf85-4725-9579-c5d13106ca6a>)

Not filling in zeros

It might be that when you're measuring something, it's usually a zero, say the number of times a rabbit is observed in the survey. Why bother writing in the number zero in that column, when it's mostly zeros?

However, there's a difference between a zero and a blank cell in a spreadsheet. To the computer, a zero is actually data. You measured or counted it. A blank cell means that it wasn't measured and the computer will interpret it as an unknown value (otherwise known as a null value).

The spreadsheets or statistical programs will likely mis-interpret blank cells that you intend to be zeros. By not entering the value of your observation, you are telling your computer to represent that data as unknown or missing (null). This can cause problems with subsequent calculations or analyses. For example, the average of a set of numbers which includes a single null value is always null (because the computer can't guess the value of the missing observations). Because of this, it's very important to record zeros as zeros and truly missing data as nulls.

Using problematic null values

Example: using -999 or other numerical values (or zero) to represent missing data.

Solution: One common practice is to record unknown or missing data as -999, 999, or 0. Many statistical programs will not recognize that these are intended to represent missing (null) values. How these values are interpreted will depend on the software you use to analyze your data. It is essential to use a clearly defined and consistent null indicator. Blanks (most applications) and NA (for R) are good choices. White et al, 2013, explain good choices for indicating null values for different software applications in their article: Nine simple ways to make it easier to (re)use your data. (<https://peerj.com/preprints/7/>) Ideas in Ecology and Evolution.

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-+, ..	Uncommon. Can cause problems with data type		Avoid

Using formatting to convey information

Example: highlighting cells, rows or columns that should be excluded from an analysis, leaving blank rows to indicate separations in data.

Plot: 2			
Date collected	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

Solution: create a new field to encode which data should be excluded.

Date collect	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Using formatting to make the data sheet look pretty

Example: merging cells.

Solution: If you're not careful, formatting a worksheet to be more aesthetically pleasing can compromise your computer's ability to see associations in the data. Merged cells will make your data unreadable by statistics software. Consider restructuring your data in such a way that you will not need to merge cells to organize your data.

Placing comments or units in cells

Example: Your data was collected, in part, by a summer student who you later found out was mis-identifying some of your species, some of the time. You want a way to note these data are suspect.

Solution: Most analysis software can't see Excel or LibreOffice comments, and would be confused by comments placed within your data cells. As described above for formatting, create another field if you need to add notes to cells. Similarly, don't include units in cells: ideally, all the measurements you place in one column should be in the same unit, but if for some reason they aren't, create another field and specify the units the cell is in.

Entering more than one piece of information in a cell

Example: You find one male, and one female of the same species. You enter this as 1M, 1F.

Solution: Don't include more than one piece of information in a cell. This will limit the ways in which you can analyze your data. If you need both these measurements, design your data sheet to include this information. For example, include one column for number of individuals and a separate column for sex.

Using problematic field names

Choose descriptive field names, but be careful not to include spaces, numbers, or special characters of any kind. Spaces can be misinterpreted by parsers that use whitespace as delimiters and some programs don't like field names that are text strings that start with numbers.

Underscores (_) are a good alternative to spaces. Consider writing names in camel case (like this: ExampleFileName) to improve readability. Remember that abbreviations that make sense at the moment may not be so obvious in 6 months, but don't overdo it with names that are excessively long. Including the units in the field names avoids confusion and enables others to readily interpret your fields.

Examples

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year

sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Using special characters in data

Example: You treat your spreadsheet program as a word processor when writing notes, for example copying data directly from Word or other applications.

Solution: This is a common strategy. For example, when writing longer text in a cell, people often include line breaks, em-dashes, etc in their spreadsheet. Also, when copying data in from applications such as Word, formatting and fancy non-standard characters (such as left- and right-aligned quotation marks) are included. When exporting this data into a coding/statistical environment or into a relational database, dangerous things may occur, such as lines being cut in half and encoding errors being thrown.

General best practice is to avoid adding characters such as newlines, tabs, and vertical tabs. In other words, treat a text cell as if it were a simple web form that can only contain text and spaces.

Inclusion of metadata in data table

Example: You add a legend at the top or bottom of your data table explaining column meaning, units, exceptions, etc.

Solution: Recording data about your data ("metadata") is essential. You may be on intimate terms with your dataset while you are collecting and analysing it, but the chances that you will still remember that the variable "sglmemgp" means single member of group, for example, or the exact algorithm you used to transform a variable or create a derived one, after a few months, a year, or more are slim.

As well, there are many reasons other people may want to examine or use your data - to understand your findings, to verify your findings, to review your submitted publication, to replicate your results, to design a similar study, or even to archive your data for access and re-use by others. While digital data by definition are machine-readable, understanding their meaning is a job for human beings. The importance of documenting your data during the collection and analysis phase of your research cannot be overestimated, especially if your research is going to be part of the scholarly record.

However, metadata should not be contained in the data file itself. Unlike a table in a paper or a supplemental file, metadata (in the form of legends) should not be included in a data file since this information is not data, and including it can disrupt how computer programs interpret your data file. Rather, metadata should be stored as a separate file in the same directory as your data file, preferably in plain text format with a name that clearly associates it with your data file. Because metadata files are free text format, they also allow you to encode comments, units, information about how null values are encoded, etc. that are important to document but can disrupt the formatting of your data file.

Additionally, file or database level metadata describes how files that make up the dataset relate to each other; what format are they in; and whether they supercede or are superceded by previous files. A folder-level readme.txt file is the classic way of accounting for all the files and folders in a project.

(Text on metadata adapted from the online course Research Data MANTRA (<http://datalib.edina.ac.uk/mantra>) by EDINA and Data Library, University of Edinburgh. MANTRA is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>.)

Key Points

- 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.
- 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.
- Record metadata in a separate plain text file.

