The culture of open scholarship

Karl Broman

Biostatistics & Medical Informatics Univ. Wisconsin–Madison

kbroman.org

Slides: bit.ly/broman2022



About me

- Applied statistician working in genetics
- ► Co-author on 170 papers and 1 book
- ► Reviewer for 89 different journals
- ► Formerly
 - Associate editor and Senior editor at Genetics
 - Associate editor at Biostatistics
 - Associate editor at Journal of the Americal Statistical Association
 - Academic editor at PeerJ
 - Editorial Board member of BMC Biology







Data Organization in Spreadsheets

Karl W Broman^a and Kara H Woo^b

^aDepartment of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; ^bInformation School, University of Washington, Seattle, WA

ARSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization, Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

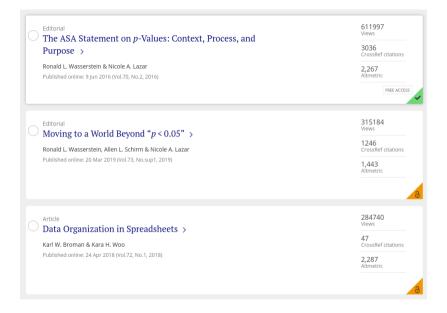
ARTICLE HISTORY

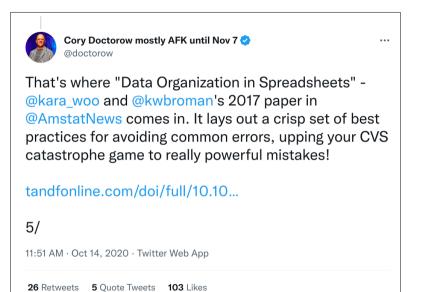
Received June 2017 Revised August 2017

KEYWORDS

Data management: Data organization: Microsoft Excel: Spreadsheets

	A	В	С	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17





bit.ly/3TRWuYj

data organization organizing data in spreadsheets

My collaborators sometimes ask me, "In what form would you like the data?" My response is always, "In its current form!" If the data need to be reformatted, it's much better for me to write a script than for them to do a bunch of cut-and-paste. I'm a strong proponent of data analysts being able to handle any data files they might receive

But in many cases. I have to spend a lot of time writing scripts to rearrange the layout of the data. And how would you like your data analysts to spend their time? Reorganizing data, or really analyzing data?

Most of my collaborators enter and store their data in spreadsheets, and mostly Microsoft Excel. Before starting to enter data into a spreadsheet, it's good to spend some time thinking about the layout. The way that you organize the data in spreadsheets can have a big impact on your data analyst's quality of life.

This is a tutorial on that topic: how to organize data in spreadsheets. For complex, high-dimensional data, it may be better to use a formal database. But for many projects, spreadsheets are perfectly fine. But data in spreadsheets can be pretty and easy to work with, or they can be a sloppy mess requiring serious downstream reorganization efforts. We want to avoid the latter.

I don't think these ideas come naturally to anyone. So if you're not happy with the structure of your current data files, don't despair! And also don't apply tedious and potentially error-prone hand-editing to revise the arrangement. Rather, apply these principles when designing the layout for your next dataset, to help make analyses easier.

- Re consistent
- Write dates as YYYY-MM-DD.
- . Fill in all of the cells.



Data organization in spreadsheets

 ${\rm Karl~W.~Broman~^*}$ Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison and

Kara H. Woo

Information School, University of Washington

September 11, 2018

Abstract

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this paper offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, don't leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header

INVOICE

INVOICE NUMBER: 943345712

INVOICE DATE: 10/06/2017 TAX INVOICE CUSTOMER NUMBER: 3551015

Please quote your customer number on all correspondence

TERMS: Payable in 30 Davs Taylor & Francis
Taylor & Francis Group

PAID

INVOICE TO: Biostatistics & Medical Informatics University of Wisconsin-Madison Biostatistics & Medical Informatics 2126 Genetics-Biotechnology Center 425 Henry Mall MADISON WI 53706 LINITED STATES OF AMERICA DESPATCH TO: Mr Karl Broman Biostatistics & Medical Informatics University of Wisconsin-Madison 2126 Genetics-Biotechnology Center 425 Henry Mall MADISON WI 53706 UNITED STATES OF AMERICA

OUR REF:

YOUR TAX REF:

OUR TAX REF: 04-3801744

ORDER NUMBER: 4490118 CUSTOMER ORDER: 10.1080/00031305.2017.137598

ORDER REF.	QTY	ISBN/ISSN	TITLE	UNIT PRICE	DISC	NET VALUE	TAX	TAX %
T&F iOpen Access Fee	1	1537-2731	The American Statistician Online	2,950.00	0.00%	2,950.00	0.00	



bit.ly/3sIRtVY

Lessons

- ► Continued attachment to journal articles
- lacktriangle Open access \longrightarrow more readers

What has changed?

- ► Rise of preprints in biology
- ► Rise of *Nature Communications*
- ► PubMed Central: expansion, with no embargo

What hasn't changed?

- ► Attachment to Impact Factor
- ► Researchers don't read much

Barriers to OA

- ► Most scientists don't seem to care
- ► Cost