

Keeping data tidy: Guidelines for research projects

Duncan Golder

2019-11-22

Contents

1	Introduction	5
1.1	Why write these guidelines?	5
1.2	Tidy vs untidy data	5
1.3	The bottom line: Problems with spreadsheets	5
2	Types of variables	9
2.1	One variable, one column rules	9
3	A simple example	11

Chapter 1

Introduction

1.1 Why write these guidelines?

The guidelines have been written as a result of years of experience helping students to analyse data for research projects and dissertations. Many courses on data analysis and statistics assume that the data is already available in a well designed, conventional format, all ready for analysis. However this is rarely the case in the real world. Supervisors and tutors often assume that students intuitively know how to collect and maintain data. However relying on common sense and intuition alone is just not enough.



The use of spreadsheets for data analysis causes the problem. There is nothing intrinsically wrong with keeping data in a spreadsheet. However there are many potential hazards involved. Assuming spreadsheets can be used without any training is dangerous.

1.2 Tidy vs untidy data

Large quantities of useful data are stored in a naturally tidy format. All electronic devices designed to generate and store data will produce tidy data. A GPS produces tidy data in the form of a .gpx file. A data logger or automated weather station produces tidy data. However, as more and more quantitative data is generated and stored on the internet the ratio of the number of files consisting of untidy data to the number of files containing tidy data increases. There is a reason for this. When data values are generated in a tidy format they can simply be added to the end of a pre-existing table in a data base. However untidy data that does not follow any conventional format and so cannot be combined easily with pre-existing data files.



Untidy data leads to a feeling of data overload. It is bad for your mental health!

1.3 The bottom line: Problems with spreadsheets

Spreadsheets were designed originally as a tool for accounting, not scientific data management. When spreadsheets were first developed no one imagined that they would be used for managing large amounts of data. Excel was sold bundled with the MSAccess data base program under the assumption that data would be held in Access and only analysed and presented in Excel. However times have changed. Spreadsheets can now hold many more rows of data than in the past. They have become the tools of choice for most data



Figure 1.1: The problem with spreadsheets! The assumption that there is a bottom line.

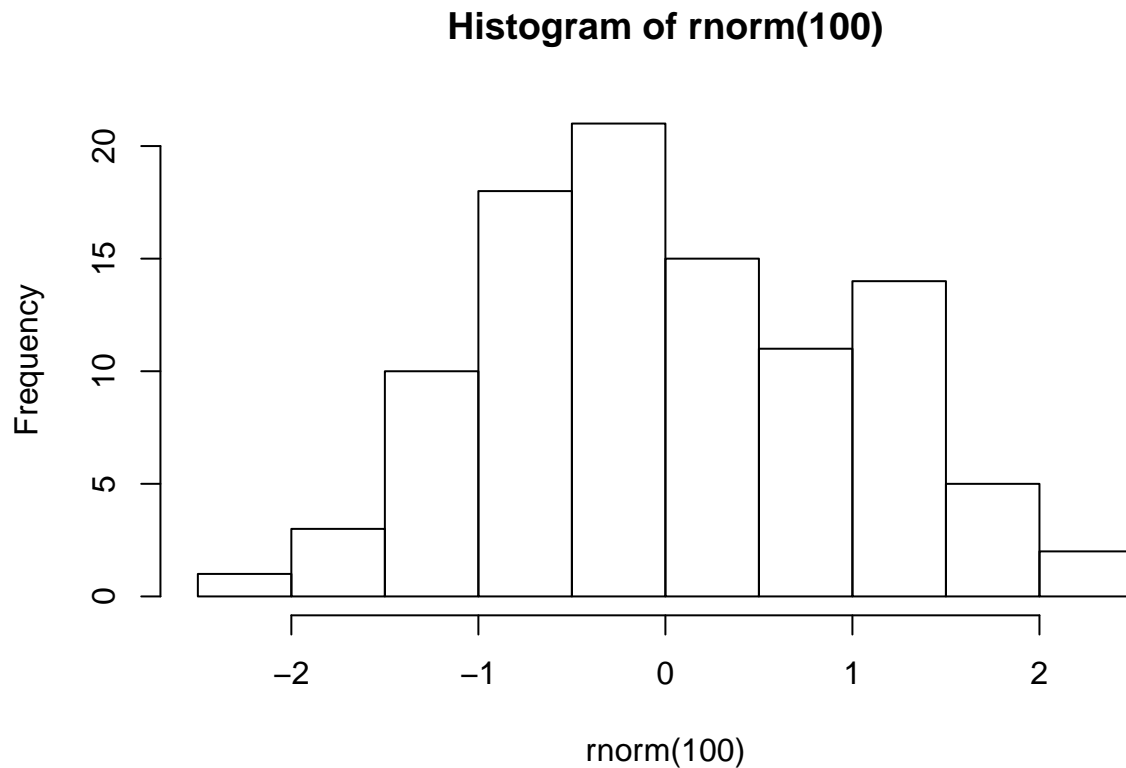


Figure 1.2: A figure caption

capture and management. This does not have to lead to problems, but in reality it often does. The key difference between a spreadsheet and a traditional data base lies in the fact that spreadsheets do not impose any structure on the data format. This is why they became popular in the first place. It is very easy to type data into a spreadsheet, or cut and paste data from a range of sources straight into a table. That is not possible in a traditional data base. A traditional data base imposes a rigid structure on the data. The type of data held in each field is pre-defined. The first data bases even insisted on the user predefining the maximum number of characters or bytes used in order to ensure disk space was not wasted.

Figure 1.1 is a humorous demonstration of the potential problems with a spreadsheet.

Figure 1.1 is clearly a joke, but it does make a serious point. When data is entered into a spreadsheet there is no built in sanity check to ensure that the units of measurement are compatible. There is not even any check to ensure that numerical values are not mixed with text. Calculations add a bottom line to a column of data, but these derived values represent a

```
hist(rnorm(100))
```



http://r.bournemouth.ac.uk:82/AQM/AQM_2018/Crib_sheets/Classical_statistical_tests.html



Chapter 2

Types of variables

The key to keeping data in a tidy format is to properly define the variables you will be using **before** collecting any data. This is a natural process if a traditional data base is used for data storage. However when using a spreadsheet it is not necessarily as straight forward.

There are two broad classes of variable.

1. Categorical variables
2. Numerical variables.

Within these classes some further distinctions can be made. Categorical variables can be binary (e.g. presence or absence, true or false), ordinal (e.g strength of feeling) or non-ordinal (e.g vegetation types)

Numerical variables can be subdivided into integers, numbers that represent an interval scale (e.g temperature) and those that represent a ratio scale. For data management purposes these distinctions are not particularly important. The key difference is that between categorical variables and numerical variables.

2.1 One variable, one column rules

If you follow three simple rules you can usually guarantee that your data will be tidy.

1. All the values in any column must be the result of exactly the same measurement process.
2. Aggregated data must not be mixed with raw data.
3. There should be one (and only one) column for each measured variable.

These rules lead to the formation of a data frame. The best way to explain this is to look at what happens when the rules are violated.

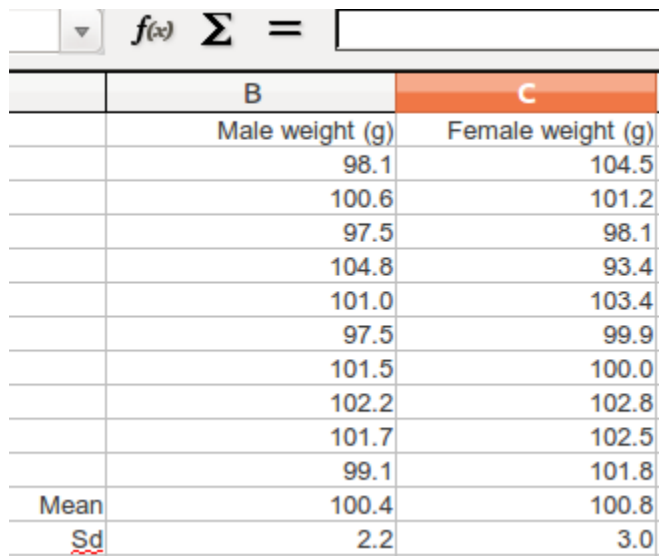
```
d %>% pivot_wider(names_from =gender,values_from=wt)%>% select(-1) %>% datatable()
```


Chapter 3

A simple example



This is an example of how **not** to capture data. If you have been handling data like this, now is the time to think again.



The image shows a screenshot of an Excel spreadsheet. At the top, there is a formula bar with a dropdown arrow, the text $f(x)$, a summation symbol Σ , and an equals sign $=$. Below the formula bar is a table with two columns, B and C. Column B is labeled 'Male weight (g)' and column C is labeled 'Female weight (g)'. The table contains 11 rows of data, followed by two summary rows: 'Mean' and 'Sd' (Standard Deviation). The 'Sd' row has a red underline under the 'Sd' label.

	B	C
	Male weight (g)	Female weight (g)
	98.1	104.5
	100.6	101.2
	97.5	98.1
	104.8	93.4
	101.0	103.4
	97.5	99.9
	101.5	100.0
	102.2	102.8
	101.7	102.5
	99.1	101.8
Mean	100.4	100.8
<u>Sd</u>	2.2	3.0

Figure 3.1: Typical use of Excel. This approach works perfectly well with small amounts of data. However it will not scale up.