Data Basics

By Tanya Hoskin, a statistician in the Mayo Clinic Department of Health Sciences Research who provides consultations through the Mayo Clinic CTSA BERD Resource.

Any statistician will tell you that a large percentage of our time is spent cleaning and preparing data for analysis. It is a tedious, time-consuming, and unglamorous part of the job, but it is absolutely essential. The quality of the data determines the quality of our results and the conclusions we draw from them. For the researcher who is collecting, entering, or manipulating data, mastery of fundamental concepts about data is important and can make your and/or your consulting statistician's job easier.

We start with the most fundamental concepts. These ideas are probably best conceptualized by thinking of a spreadsheet, such as Microsoft Excel. Two spreadsheets appear below. Upon first glance, we see that they contain essentially the same information. The structure and quality of the second spreadsheet is far superior to the first, however. From an analysis perspective, the first spreadsheet is unusable in its current form. These two spreadsheets will set the stage for an explanation of the basics of data.

Bad Spreadsheet

ID	Gender	DOB	Height (cm)	Mass (kg)	Dx
1	M	1/1/1980	183	88	1
2	M	16/1/1981	187	80	2,1
3	F	2/1/25	188	unknown	2
4	MALE	2/15/1983	172cm	82	2
4					3
6	male	March 1, 1984	180	87	2
6	m	3/16/1985	184	84	2 (dx 6/2/00)
7	m	4-1-1988	185 ???	88	1
8	female	April, 1987	188	69	1
9	F	5/1/1958	162	65kg	diabetes
10	f	1989	164	64	2
			average=186		

Good Spreadsheet

ID	Gender (1=M,2=F)	DOB	Height (cm)	Mass (kg)	Dx1	Dx2
1	1	01/01/1980	183	88	1	
2	1	01/16/1981	187	80	2	1
3	2	02/01/1925	188		2	
4	1	02/16/1983	172	82	2	3
6	1	03/01/1984	180	87	2	
6	1	03/16/1985	184	84	2	
7	1	04/01/1988	185	88	1	
8	2	04/16/1987	188	69	1	
9	2	05/01/1958	162	65	3	
10	2	07/01/1989	164	64	2	

Two fundamental concepts: variables and observational units

- We use the structure of the spreadsheet (columns and rows) to organize our data and enforce consistency. The columns of a spreadsheet represent variables. You could think of a variable as a single piece of information collected on every individual (e.g., height, blood type).
- The rows in the spreadsheet represent the individuals on whom we are collecting data. Typically, it is desirable to have the data for each unique experimental/observational unit (e.g., patient) in a single row of the spreadsheet.

Rules for spreadsheet cells

- Each cell of a spreadsheet (or each variable for an individual) should contain a single piece of information.
- This follows from the fundamental concepts that each column contains a variable (a single piece of information collected on individuals) and each row contains information for a single individual. A cell is the intersection of a particular column and row; therefore, a cell should contain the information for one variable for one individual.
- For example, don't enter gender and date of birth in the same cell. It may seem like a good idea to combine demographic variables or other similar variables, but it's not. Remember that although you can still see the information, software programs cannot easily separate those two pieces of information. You must keep each distinct piece of information in a distinct cell. If you use a comma at any time during your data entry, it's likely a problem!
- See the last column of the "bad spreadsheet" for an illustration of this point. Two different diagnosis codes are entered in a single cell and separated by a comma. See the "good spreadsheet" for one possible solution to this type of situation.

Consistency

For data to be useful, it must be recorded in a consistent format. One of the most important tips related to this concept is to avoid using text (letters, symbols, etc.) in your spreadsheet whenever possible. Why? The use of text makes it very difficult to maintain consistency and can result in a spreadsheet that is very difficult to use for even the simplest analyses.

Whenever possible, you should use numeric coding. For example, if the variable is gender, it is better to develop a numeric coding system such as 1 = male, 2 = female rather than using text such as male/female or M/F. Although this method may seem more cumbersome, think of it this way – there is only one way to write a "1" but many ways to write a text description of gender ("M", "m", "male", etc.). To the software program, "M" and "m" are two different character values. Look at the gender column of the "bad spreadsheet". If you tried to use this spreadsheet for analysis, the software package would tell you that there are seven different categories for the gender variable. Make certain to carefully document what your numeric codes mean.

- Text added to an otherwise numeric column (for example, including "kg" and entering 50kg for mass) will also make that column very difficult to work with; enter

the number only and document the units (in the column name, for example). Refer to the mass column of the “bad spreadsheet” for an example.

- If you need to capture a text description in addition to the variable of interest, feel free to add and use a notes column. Just remember that it is unlikely that you will be able to use this comment field for any type of analysis. Record only numeric values in the column that contains the variable you want to analyze.

Insider information: miscellaneous tips

- There is a difference between a missing value and a zero. Zero is a number and should only be used when a value of zero is observed for the variable of interest. Procedures for entering missing values differ among projects. However, if you are working with a spreadsheet for data entry, leaving the cell blank is often the best choice. Remember that using text such as “N/A” or “unknown” will make an otherwise numeric column difficult to use.
- Use four-digit years for any date variables. Enter dates using a consistent format. The format of date variables should be MM/DD/YYYY. In the “bad spreadsheet”, notice that the use of seven different date formats creates a mess.
- Keep extraneous information, such as data summaries, out of the spreadsheet. The ideal spreadsheet should contain only raw data with a single row of column headings, which contains the variable names. Notice that the last row of the “bad spreadsheet” contains the average height. This is not part of the raw data and should be recorded elsewhere.
- Statistical software packages recognize two variable types: numeric and character. For a variable to be considered numeric, which is desirable in most situations, the entire column must contain only numbers. Statistical software packages read the entire column, and any text in any cell of that column will cause that variable to be treated as character data. For example, the height column in the “bad spreadsheet” would be read in as a character variable. This would make it impossible to get even simple descriptive statistics, such as the mean and standard deviation, without going through special procedures to clean the data.

Although this list cannot cover everything you need to think about when using a spreadsheet to collect data, these concepts and tips should provide a good starting point. If you go back to the “bad spreadsheet” and “good spreadsheet” examples we started with, it should now be clear why the latter is superior.

More information

The Mayo Clinic CTSA provides a biostatistical consulting service through its BERD Resource. More information can be found on the BERD home page.