Data collection and spreadsheet management

Ristan Greer

June 8, 2020

1 What data do you want to collect?

Data should be collected prospectively, that is, before you commence your study. Review your aims, hypotheses and your research questions. Examine your study design carefully. Ensure that you collect data on the primary outcome measures, and causes or risk factors you propose to evaluate, including potential confounding factors and any factors which may modify your outcome measure (effect modifiers).

2 Make a causal diagram

A causal diagram is very helpful when designing your study and planning the actual data you need to collect (Figure 1).

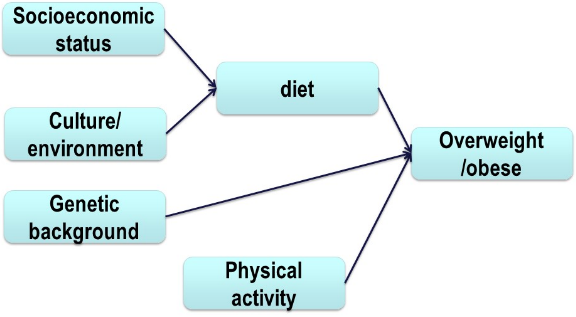


Figure 1: A basic causal diagram showing risk factors for overweight or obesity

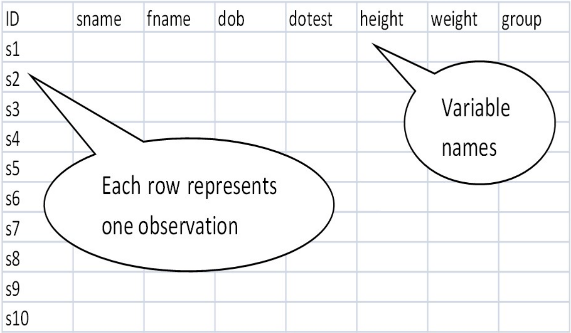
3 Setting up a spreadsheet

Setting up a spreadsheet for data collection: you need to consider

- Variables and observations
- Variable format - variables (columns) should be either numeric or character variables, not a mixture of both
- One observation per subject or many (repeated) observations per subject?
- Longitudinal or panel data
- Coding data

The aim of setting up your spreadsheet the right way in the first place is to save time re-coding data later on. Essentially all analysis will require that the data be read into some sort of statistics software package, such as R, Stata, SAS or SPSS. These software packages will only read data that is set up correctly. Variables are the items you wish to collect. For example, some kind of identifier for each subject, date of birth, date of test, sex, and then the variables pertaining to your study. For example, you might want to collect anthropometric data such as height, weight, ethnic background, vaccination status, treatment group, response to treatment, disease severity, etc.

Data is collected into a spreadsheet. Each row of the spreadsheet represents a single observation. Each column of the spreadsheet is headed by a 'variable name', and contains data on a particular variable. Variable names are sometimes called 'fields'. There should be ONLY ONE ROW of variable names. Each row represents one observation (Figure 2).



ID	sname	fname	dob	dotest	height	weight	group
s1							
s2							
s3							
s4							
s5							
s6							
s7							
s8							
s9							
s10							

Figure 2: A typical spreadsheet layout

4 Variables and observations

Setting up a spreadsheet for data collection: you need to consider

1. Variables and observations
2. Variable format - variables (columns) should be either numeric or character variables, not a mixture of both
3. One observation per subject or many (repeated) observations per subject?
4. Longitudinal or panel data
5. Coding data

It is useful to code the variable names in 'shorthand', as a short word of less than eight characters. Do not use spaces or special characters such as '&', '*', etc. Some researchers prefer to use all lower case letters. Do not use variable names which start with a number or a space. Many modern statistical packages recognise and handle long variable names, spaces, and distinguish upper case from lower case, but some packages do not, so it is safer to use these simple coding rules. It is helpful to create a 'key', or 'data dictionary', to remind yourself what the abbreviated variable names represent. For the above example, you might have a key as follows:

1. ID = The ID variable identifies each participant, animal or element of the study. S1 might represent study participant number 1.
2. sname = surname
3. fname = first name
4. dob = date of birth

5. dotest = date of test

6. height = height in centimetres

7. weight = weight in kilograms

8. group = study group where 0=control, 1 = treatment (the reference or comparison group is usually coded as 0, this makes any statistical analysis easier to interpret).

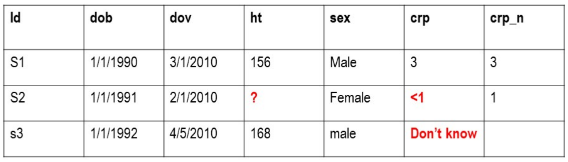
5 Character and numeric variables

You must make sure that all observations under a particular variable heading are uniformly 'character' or 'numeric'. A character variable is read as text and usually contains letters. Eg 's1', 'Smith',' U21347', are examples of observations in a character variable. Numeric variables contain ONLY numbers, eg '123', '41.7', '53.200098' are numeric variables. DO NOT put notes or annotations into a numeric variable column, because statistical packages cannot read these and will return an 'error' message. For example, DO NOT put 'not done', N/A, ?, or similar, into a numeric variable column.

Character variables are also known as 'string' variables.

Statistics packages cannot read characters such as < or >. These are commonly used in pathology reports. When this occurs, you should enter the raw data into one column and then enter the value you want to use for statistical analysis in another column. For example, if a value is given as <1.0, you might elect, from your clinical knowledge, to enter the value as 1.0 in the second column.

If you have missing data, or data which is not available, and you wish to indicate this in a numeric variable column, one option is to use a numeric code, eg '999' = 'missing data', '998' = 'not available', etc. Then the statistical package will still read your data sheet. You may choose the option of leaving blank cells for missing data.



Id	dob	dov	ht	sex	crp	crp_n
S1	1/1/1990	3/1/2010	156	Male	3	3
S2	1/1/1991	2/1/2010	?	Female	<1	1
s3	1/1/1992	4/5/2010	168	male	Don't know	

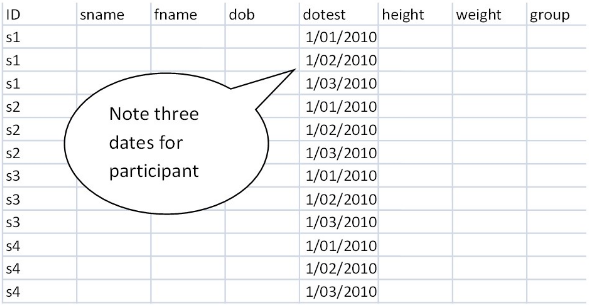
Figure 3: Mixed variables types - don't do this

6 Longitudinal or panel data

For some studies, for example those which measure a person, animal or other unit of study at a number of timepoints, the ID variable appears more than once in the spreadsheet, once for each time point. For example, Figure 4 shows a spreadsheet layout which could be used in a study where participants are measured every month for three months. Such a layout is known as the 'long' form of the data.

While it is possible to arrange your data for each visit to be entered 'across' the spreadsheet, this is practical only if there are very few visits or times of observation. If you have a large number of time points, a 'wide' layout quickly becomes very unwieldy.

All statistical packages will convert this to 'wide' form if you need to calculate measures such as, for example, the difference in weight between two time points or visits.



ID	sname	fname	dob	dotest	height	weight	group
s1				1/01/2010			
s1				1/02/2010			
s1				1/03/2010			
s2				1/01/2010			
s2				1/02/2010			
s2				1/03/2010			
s3				1/01/2010			
s3				1/02/2010			
s3				1/03/2010			
s4				1/01/2010			
s4				1/02/2010			
s4				1/03/2010			

Note three dates for participant

Figure 4: Layout for longitudinal data

7 Coding data

If you want to use categorical variables (ordinal or nominal variables) in your statistical analysis, you may find it saves time to code the data as it is entered in your spreadsheet. While some statistical packages can cope with uncoded data (i.e. character or text format), some cannot. Eg STATA requires most variables to be in numeric format. For example, if you have the following breeds of cows you could code them numerically as follows:

1. Angus = 0
2. Shorthorn = 1
3. Droughtmaster = 2

8 Categorical variables - nominal vs ordinal

Categorical variables can be nominal or ordinal. Nominal variables have no arithmetic meaning. Examples of nominal categorical variables are eye colour, ethnic background, mode of birth, breed of cow. With ordinal variables, the number has an ordered meaning, for example disease severity might be coded as health = 0, mild disease = 1, moderate disease = 2 and severe disease = 3.

9 Recoding variables

Most statistical packages include tools to enable you to recode variables at a later stage during the actual analysis. It's your choice whether you perform coding at data entry or later, during the analysis.

10 Text formatting and colour

Tip: many researchers use text formatting and colour in their data collection spreadsheets. This will not be recognised by statistical packages and may make it harder to import your data into the package. If you find formatting and colour helpful, use these tools sparingly and remember you'll probably have to get rid of them before doing any serious analysis.

10.1 Data management

1. Protecting original data
2. Collating and keeping track of data; data log

10.2 Protecting original data

Your data is very valuable. It is usually the result of a long, expensive process, starting with conception of the project, through design, ethics and governance approval, planning and data collection.

Your original data should be kept. Whether it is in hard copy or electronic format (eg entered directly into a laptop in the field), original paper and disc copies should be stored in a safe, secure location. Several copies of any discs should be made in case of loss or damage. It is important to be able to go back and check the original data is any question arises as to the existence or accuracy of any particular piece of information. Raw data for analysis should be created from the original data, and this should be backed up too.

Your raw data set, derived from the original data set, is often changed and edited. It is a useful practice to keep the raw data set in its first format, and make a copy to use for analysis. This way you can manipulate the data, perhaps using recoding, or newly derived (calculated) variables, or subset the data, without danger of losing any information or perhaps making an error and not being able to 'start again'. It may be useful to designate your raw data set with a special name, eg 'master data', or 'original data'.

In excel, you can easily make a copy of a worksheet: hold down the 'control' key and drag the worksheet tab to the right. A copy of the first worksheet will be created. You can rename this is you wish by double clicking on the tab and typing in a new name.

Figure 3. Copying worksheets in excel. The 'copied' worksheet is called 'master data(2)'.

11 Collating and keeping track of data: data log

You will find that as you delve deeper into your topic, your data analysis will become quite complex as you explore different questions and develop more sophisticated analyses. It is very important to keep a data log. This should include:

1. the date of analysis
2. the aim of the analysis
3. name of the data set you are using for analysis
4. usually the name of the program you have developed or written to do that particular analysis
5. which method of statistical analysis you used and any options associated with the particular method
6. details such as any observations you may have excluded for any reason and any insights from your analysis.

(i.e. What did you do? When did you do it? Why did you do it this way?) A standard laboratory log book is a good starting point to record your work, but you can keep it in any format you like - eg an excel spreadsheet or word document, or a hard-copy laboratory notebook.

It is good policy to print out or electronically save the computer output from key data analyses. Most statistics packages have some kind of automated logging which can be saved as a text file, you need this when writing up results because you won't remember a few days or weeks later what you did.

Keeping a data log will be of great assistance when you are writing up your results and conclusions. Your data log supports reproducible research, is a memory jogger when you are addressing reviewer's comments, and an invaluable tool for research integrity.

12 Recommended reading

- An Introduction to Stata for Health Researchers. Svend Juul & Morten Frydenberg. 2014 4th Edition. Chapter 10. Taking good care of your data.
- Veterinary Epidemiologic Research. Ian Dohoo, Wayne Martin & Henrik Stryhn. Chapter 30 'A structured approach to data analysis'.3