+Spreadsheets

What spreadsheets are for.
Spreadsheets are commonly used to house and analyze small datasets, of all sorts.
Today, we're going to run through some tips and tricks to make spreadsheet use less error prone, and more efficient.
While spreadsheet use should be limited, and not used for all data analysis, programs like Excel and OpenOffice are perfectly fine for performing simple computations on tables of data. And, these programs do have lots of advanced features that are rarely used, but can tell you a lot about the data you're working with.

Today, we're going to run through two realistic situations. The first is that you have to set up a spreadsheet for yourself and others, that may be used for years in the future. The second is that you've inherited a poorly organized spreadsheet from someone else, and you're tasked with cleaning it up.

Our examples are drawn from typical situations in behavioural psychological research, but of course the features we'll point out apply to any sort of data you're working with.

<u>You're making a spreadsheet from scratch</u>

We're going to make this together.

Let's say we're recording some information about study participants. We plan on creating or collecting a few key pieces of information about all the participants we've seen:

- study ID
- classification
- birth date
- sex
- date tested
- test scores for 3 tests

Let's set this simple spreadsheet up together.

First, let's write the column names, so we remember what we're doing.

Next, let's generate the study IDs. For now, let's say we have just 6 participants.

Study ID
Let's say study ID is base on 2 components: The study name (which is the same for all participants), and a sequential number. There's no reason to type these out by hand.

First, say the study name is ‘BRAINS’. We can type this once, and pull it down. (Pulling down is called the “fill handle’.)
Second, let’s get Excel to give us the sequential number. Type 1 and 2, and use the fill handle to generate the next numbers in the series.
Third, let’s remember it’s good practice to separate constants and non-constants with some kind of symbol. Let’s put an underscore ‘_’ between the study name and number. Again, use the fill handle to pull down.

So, now we have the right components that we can stitch together! We can use the inbuilt function in Excel called CONCATENATE. The concatenate function is just another way of saying ‘to combine’.
=CONCATENATE() separate by comma
Here, we simply specify the order in which we want to combine our information.

That looks right! Now, we can pull this formula down (again, using the fill handle) and have our study IDs.

That achieved what we want, and it was pretty fast. But believe it or not, we didn’t ever have to use the fill handle that much. In fact, there is a better way. There’s two parts of our study ID that stay the same, no matter what. That is, the study name ‘BRAINS’, and the underscore separating the study name from the number. So, we didn’t actually have to type this out more than once!

Let’s delete everything but one entry.

Again, we should use the CONCATENATE function, but use ABSOLUTE, instead of RELATIVE, references in the subsequent arguments. By default, Excel uses relative references.
To transform a relative reference into an absolute reference, simply place a $ before the column letter and row number.

Classification
We know we’re only going to have healthy controls, and people with schizophrenia. If you’ve ever cleaned data before, you know that schizophrenia can be spelled 9 different ways, and that there’s even more ways to refer to healthy controls. To make these options constant, let’s use DATA VALIDATION. We can use data validation to restrict users to select from certain, predefined values.

Go to Sheet 2
Type in our options
Name these two items (ParticipantGroup)
Go to Sheet 1
Click the cell

Data > Data validation
Allow > list
= (list name)

Note, we need to use something called CAMEL CASE, which is the practice of writing multiple words as compound words, such that each word or abbreviation begins with a capital letter. We cannot have spaces.

Birth date
Generate whatever dates you like.
Bring up Format Cells with Command 1

Sex
Let's make a validated list.

Date tested
=TODAY()

Test scores
Let's pretend we've already had these participants in, and tested them on three measures.

Let's say the first is a well-known test that is administered using proprietary software. We actually have a spreadsheet of scores elsewhere, that the program spits out. We can get these scores into our spreadsheet, so that everything is in one place.

Open and save the proprietary software workbook (View > Window)
Select the appropriate cells in our new workbook
Type =
Select the cells in the old workbook that we want to share
If you return to the new workbook, you will see that a link has been made.

Or, instead of opening up the workbook, you could type
='[workbookname.xlsx]Sheet1'!$A$2:$A$7

Let's test! When we update our big datasheet, the linked copy also updates! This is cool. It means that if these scores adjust - for some legitimate reason - your data will be automatically updated.

Of course, at some later stage - when you're writing up your results, for example - you want to ensure the data doesn't change, and use whatever is there. This means that you need to break the link.
Data
Edit Links

Break links

Ok, moving on. Let's say the second is an experiment built in our lab, on which participants score between 20 and 60.
=RANDBETWEEN(20, 60)

(Of course, this doesn't mean that participants are getting random scores; we're just demonstrating the RAND function. An important thing to note here is that random numbers are generated every time something updates. To stop this from happening, let's copy > paste special > values.)

Let's say that the third test actually hasn't been administered yet.You happen to know the RA administering the test, and his data entry skills aren't the most reliable. We can minimize his errors by using data validation.

Data > Validate > Data Validation. Allow whole number between [range].

If the RA enters something other than that, a message will pop up.

Great! Now we have a record of a few key pieces of information. This is our data!

Aggregation.
We can use Excel's basic summary statistics functions to describe the data. In spreadsheet lingo, this is called "aggregation". This means we can combine several values into a single result. Basically, this allows us to keep tabs. Clearly, this isn't as pressing because our current dataset is so small (and we glean an understanding at a glance). But, as it gets bigger, the aggregation will become essential.

One good thing about Excel is that the way to write these formulas closely mirrors what we could stay in English.

Ok. The first variable in this chart that we're likely interested in counting is the number of participants that belong to our Schizophrenia and Healthy Control groups, respectively. (It doesn't make sense to have summary statistics for everything.) There's a few ways we can do this:

=COUNTIF(range, "criteria")

Or, instead of typing of the "criteria" in our formula, we can refer to a cell. Remember when we named 'Schizophrenia' and 'Control' on the other sheet? Let's try referencing those cells in our formula.
=COUNTIF(range, cell reference)

Another important thing to note here is that if our data changes, the formulas get to work right away, and immediately adjust our summary statistics. Let's test this out.

Now, let's move on to the test scores. We probably don't care how many people got a particular score. Instead, for the moment, we might be more interested in looking at average.

Minimum, Maximum, Average, Median, Mode, Standard deviation

As you may have noticed, these functions are pretty intuitively named. And additional good news is you don't have to remember every single function name, ever. Simply click the 'Insert Function' button on the Formula toolbar, and scroll through to read descriptions.

We can just pull these formulas over, because the formulas automatically update to relative references.

Awesome. But here's the thing. What we did worked, but it's not actually the best way. Unless you have perfect typing skills, you may have noticed that it was difficult and error prone to write in all these cells references. Instead of writing cell references for all of our formulas, we should use NAMING RANGES.

Select the portion of cells. Go to upper left hand corner NAME BOX, and name.

This is easier to understand, in the same way that giving variables meaningful names makes them easier to understand. Now, instead of typing in cell ranges, we can simply references these names.

But… Let's do a test. Let's add a record in the middle of the rows.
If we go to Formulas > Name Manager, we see that everything updates

But, now let's add a record to the end. This formula doesn't automatically update, and our calculations don't immediately adjust! We have to redefine the parameters of the name.

We can do this in Formulas > Name Manager

Alternatively, DYNAMIC NAMED RANGES automatically adjust their values to include all the cells in a column. That's a little more advanced. Let's make a test 4. Enter any test scores you like, here.

Select cell
Formulas > Define Name
Refers to:

=OFFSET(click where start, how many rows down 1, how many rows over 0, tell how many rows in range - this changes - went all counta(click column) -1 so get rid of heading, 1 because 1 column)

To check: Go to Formulas > Name Manager and then click on content of 'refers to'.

Note that dynamic ranges don't appear in left hand corner dropdown. But it is in the Name Manager. We can add values far down in the column to make sure it's working.

You might be thinking that it's rarely the case we want to look at average scores across all participants - here, we'd want to look at the average of the HC and the SZ group. We can do that too:

AVERAGEIF > (participantgroup, criteria - one of group, test)

Ok, let's take a look at what we've got.

Some of these summary statistics are kind of ugly. For example, there's several decimal points in the average and standard deviation. Since we're not doing any sort of data analysis, this isn't important right now, let's set everything to display 2 decimal places. When we format cells in Excel, we are just changing the appearance of a number, without actually changing the number itself.
Right click. Bring up context menu. 'Number' category. Tell Excel to display just 2 decimal places.

Arguably - depending on what kind of data you're working with - more than 2 (or sometimes 4) decimal points are not needed. Another option is to round the numbers. However, when you round, you lose precision.
=ROUND(A1, 2)

Things are starting to look good! But let's take a moment to look at the data. If we look closely, we see that Test 1 has a minimum score of 0. You'll remember that Test 1 scores were from the proprietary software import. A score of 0 could result from any number of things: perhaps the participant really was very poor at the test, or perhaps he refused to complete the test. Or, it's possible there was a glitch in the software. Let's make a note to check in on this later. But for now, we're really interested in average.

To get a better idea of what the average really is - assuming this entry of 0 is a mistake, or will ultimately be excluded, you can calculate
=AVERAGEIF(range, "<>0")

Great! So, so far, we've got in the demographic characteristics and raw data, and run some summary statistics to make sure that everything is on track.

Now, what to do with the data?

What we've set up might look basic, but it's a strong platform for what comes next. And if your day is like mine, what comes next is people wanting to know lots of things about the data, right away.

Situation1: You get a call from your manager. She's recruiting participants across all studies who are high achieving. (You're part of a Centre, and all participants have consented to data sharing).

We might want to classify the test scores as impaired, or not impaired. Let's set an arbitrary threshold, based on the mean. We should we the IF function. The IF function checks whether a condition is met, and returns one value if TRUE and other value if FALSE. Let's say we're just interested in classifying participants into two grounds, high and low, of our first test.

= IF (A1>Average, "High", "Low")

Remember, this isn't fail safe for data analysis, because as we saw, the mean will fluctuate as we add more participants to the table, which means that these labels might change as well.

However, maybe it's the case that there's some agreed upon cutoff for High Achieving and Low Achieving for Test 2.

= IF (A1>50, "High", "Low")

For text 3, let's pretends we're now interested in splitting people into 3 tiers: Low, Average, and High. We can specify what these ratings are, in a different way. Let's say the lowest value of low is 0 and the lowest value of average is 15, and the lowest value of high is 30. We should write this out. We can now write a formula that will look up the label that is equal to or less than the search value.

=LOOKUP (test3, numbers, labels)

By using the OR function, we can give our manager a list of participants who are classified as High Achieving on at least one of the three tests. The OR function returns TRUE if any of the conditions are TRUE, and returns FALSE if all conditions are FALSE.

= IF(OR(A1="High", B1>"High", C1>"High"), "Eligible", "Not Eligible")

Maybe our criteria isn't stringent enough. Too many are in the upper tier on at least one tests. She really wants to have participants who are in the upper tier on all tests. This is defined here as in the upper ½ across all 3 tests. Again, this is easy. (The only difficult part of this is setting the criteria.)

= IF(AND(A1>8, B1>10, C1>15), "High achieving", "Not high achieving")

As an aside, the AND and OR function can check up to 255 conditions. So, it's quite powerful. Can have 64 nested levels.

Maybe she decides that she really just wants the top ranked participant on test 1.
This takes 3 parameters.
=RANK (specific cell, cell reference, 0)

Situation 3: you get an email from the PI

Imagine your PI emails you, and wants to know what the mean age of participants is in the study, and how many, if any, participants over the age of 30 have a diagnosis of schizophrenia. He's on the phone, and wants to know right now. Don't panic.

The first thing we need to do is calculate current age. In general, it's best not to include redundant info in spreadsheets, as space and size can become a real issue (and it can be a source of confusion). Age in particular is a source of confusion because it's not always clear if we're talking about current age, or age at assessment. Make sure to clearly label variable names.

=DATEDIF(E7, TODAY(), "Y")
Need to format as a number
Can choose to remove decimals.

The next step is to use the COUNTIFS function. This is very similar to COUNTIF that we used earlier, but the 's' on the end indicates that we are counting cells based on multiple criteria.

e.g., COUNTIFS (A1:A5, ">30", B1:B5, "Schizophrenia")

You give your PI the answer. He reminds you that the study must have a mean age of whatever. He tells you to get the average closer to what it should be by next week. You have just one person coming in next week.

Instead of crying, use a Goal Seek analysis, which tells you how old the new participant must be to meet your goal.

Data > What if > Goal Seek

You've inherited a messy spreadsheet

Ok, on to part 2!

You take a look at it, and you're not exactly sure what's going on. But you have been told that it's actually related to the data you've been collecting, and it's from the same participants. This means no matter how much you want to, you can't ignore it.

Here's some important first steps.

Cleaning it up, aesthetically.
The first thing I like to do is make what's there easy to read. This might seem superficial, but we're human, so it's essential to make sure the spreadsheet is human readable.

Freeze panes

Colour, bold, italics.
We can make values more readable by changing the way values are displayed.

Now, on to the more crucial stuff. Let's take a look at what we have.

Name
On our new spreadsheet, we have names! We didn't have this information before, but the common Participant IDs tell us we're talking about the same people. However, there's some problems with the way names appear here.

The first problem is that for a couple of names, the first and last name appear in the same cell. It's typically considered a best practice to have each component of information belong to its own cell. We can see that we'll have a bit of a problem with 'Tommy', so let's manually fix that one.

Data > Text to columns
Space as the delimiter

That looks better, but actually, it's often par for the course to no longer use full names. We use first initial of the first name, and first three initials of the last name. Let's fix this up
=LEFT (A1,1)
=LEFT(B1, 3)
CONCATENATE

Also, we'll notice that because this is coming from an old study, the participants' classification (i.e., case or control) is apparent in their Study ID. Technically, this shouldn't be the case. Let's fix this, by extracting just the number, and removing 'SZ' and 'HC' from each individual. This means that we simply need to extract the 1 or 2 rightmost characters from the string.

This is a good opportunity to use Find and Replace

Later, if we wanted to get just the last name (3 characters), we could remove the leftmost character in the way we've seen, or we could use
MID(cell, startpoint, number of characters)

Let's sort by subject ID number.

Adding validation rules after the fact
I told you there was as many different ways to spell Schizophrenia as there are cells. Here, we see that Schizophrenia is spelled correctly in full, as well as abbreviated as SZ and SSD. Even though we know this, Excel won't. It will treat these three notations as separate groups. To make sure we have accurate data, we need to decide on one notation, and stick with it.

Let's clean this up.
We can insert data validation rules after the fact.
First, make a rule like we did before. Make a list, then Data Validation
Data> Validate > Circle Invalid Data

Let's reformat the dates. One is resisting reformatting. Does anyone know why?

Moving on to Test A. We see something is highlighted, but we don't know why. An important fact to note here is that this highlighting will be lost if this Excel file is saved as a csv, which is likely will be. Formatting is stripped. We should add a comments box instead.

We can clean up the formatting in Test B.

Now, we see a column of percentages. But what does this mean?
Deciphering previous calculations
Formula > trace precedents, trace dependents
We can see that this first column is just Test A turned into a percentage point.

Dealing with missing data
Some of our data is missing. Real world data always has gaps like this. For example, it might be that a participant refused to complete a particular task, or that a new task was introduced midway through a longitudinal study. It's very important to understand that Excel treats gaps and 0s differently. Namely, 0s count as values, but gaps do not. We can also add a string, i.e., a

list of characters to an empty cell to reduce ambiguity. This string could say “NA” or “no data”. Excel will treat this as if it is empty.

Here, we see that some of the missing data simply isn’t really missing raw data. It’s a missing calculation. We can go ahead and safely add this in.

Now, on to the second percentage column. We can Formula > Show Formula

We don’t really know what’s going on in test c; seems like raw schores.

Text that contains.
Let’s take a look at the comments box. You want to highlight all the data that needs to be reviewed by your supervisor, as it perhaps shouldn’t be included in the analysis. One way that RAs have indicated this is by writing “consider removing” or a close variant in the comments box.

Using Conditional Formatting, you can send these cases over to your supervisor for review. Use wildcards (*) whenever possible.

Or, may it’s the case that you only care about “remove” if this is the first word in the box, because of an existing convention
=FIND(“remove”, range) will tell you the start position of the string in the cell.

Another thing to note is that, while not fail safe, there are default ‘Cell Styles’ in Excel. For instance, we might want to mark something with the ‘Check Cell’ formatting. But be aware that not everyone will recognize this styling, and it will also be stripped in csv.

The data is starting to look better. It’s probably as good as it’s going to get.

There’s lots of fancy things we can do to help us make sense of the data quickly.
Conditional Formatting > Data Bars
Charts.

It gets better.
Working this this messy spreadsheet has made you think about how you can avoid confusion in the future.

Comments sheet
Add a sheet at the beginning of the workbook that has information about the data contained therein.

Add header and footer

When printing, this will tell everyone important info about the spreadsheet (that you specify, such as date you created it, edited it, etc).
Insert> Header and Footer

Track Changes
Review > Track Changes

Sharing with other users
Tools > Share workbook > Editing

Locking certain cells
Select cells > Format > Cells > Protection > Unlock
Then, Protect Worksheet

Read only
File > Save As > Options > Password to Modify

Protecting Worksheet
Tools> Protect Worksheet

In sum:

Most of what you want to do in a spreadsheet can - and should be - automated. Basically, the only information you should ever be typing in is the rawest form of the data - things like names, dates, raw scores.

If you want to compute something from the data, or organize the data in a new way, rest assured the spreadsheet program can do it for you. You just have to find out how to tell the program what you want it to do.

Remember, even if you think you're the only one who's ever going to see or use your spreadsheet, operate as if you will be sharing it with others. It needs to be easy to read for the next person in line, and also for you, in the distant future, when you've forgotten all about what you were doing on that spreadsheet, anyways.

When to move beyond spreadsheets.
It's important to remember that it's not always best to work within a spreadsheet. For example, when your data gets too big, it's outgrown a spreadsheet.

And, Excel is not inherently multi-user software, though we've seen some workarounds. It's only meant to be used by you on your machine. It's not a database.

Lastly, you'll sometimes be doing something so complicated that you need to keep track of your steps, potentially give others a record of them, and run the same steps again. To record exactly what you've done, and to write a script to that others can do it at the touch of a button, you should move to programs like R, instead of using VBA.