FINDING STORIES WITH SPREADSHEETS

By Paul Bradshaw
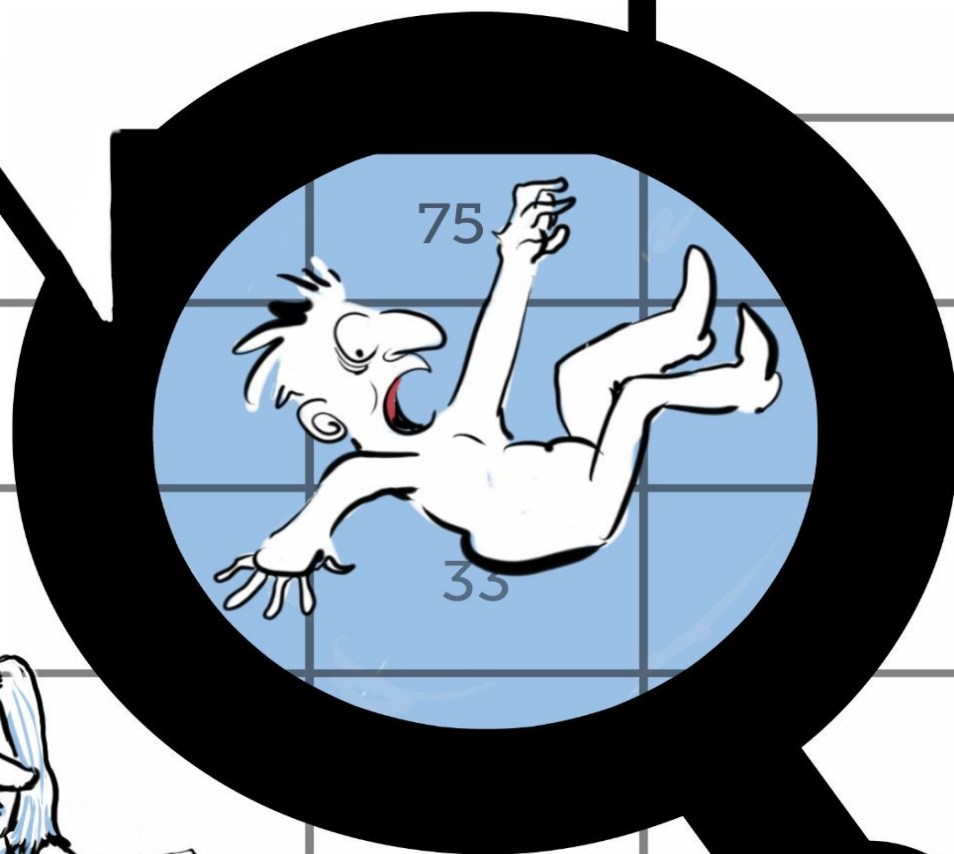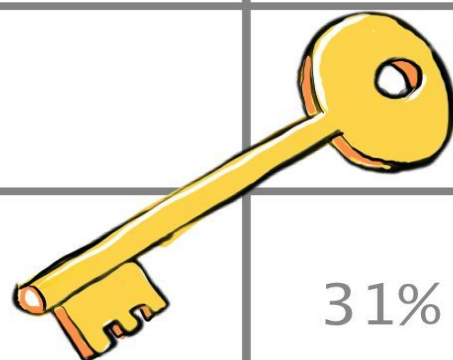
2222.31   12%

31%

44%

2007.02   75   47%

1095.44   53%

1500.12   33   64%

1430.0   56%

1200.   22   45%

1125.   13   %

1090.70   9

Finding Stories in Spreadsheets

Recipes for interviewing data - and getting answers

Paul Bradshaw

This book is for sale at http://leanpub.com/spreadsheetstories

This version was published on 2021-05-29

Tweet This Book!

Please help Paul Bradshaw by spreading the word about this book on Twitter!

The suggested hashtag for this book is #spreadsheetstories.

Find out what other people are saying about the book by clicking on this link to search for this hashtag on Twitter:

#spreadsheetstories

Also By **Paul Bradshaw**

Scraping for Journalists (2nd edition)

8000 Holes: How the 2012 Olympic Torch Relay Lost its Way

Model for the 21st Century Newsroom - Redux

Stories and Streams

Organising an Online Investigation Team

Data Journalism Heist

Excel para periodistas

Periodismo de datos: Un golpe rápido

Learning HTML and CSS by making tweetable quotes

Snapchat for Journalists

Snapchat para periodistas

For Orrell, Entwistle, Howard and Sewart, who did something noble I will always be grateful for.

Contents

1. Introduction

I've yet to meet a spreadsheet that didn't have a story to tell.

More often than not, they are bursting with those stories.

Some have stories of success and (often more interestingly) failure. Others can hold a mirror up to ourselves, and tell stories about where we've been, and where we're going; how we've changed, and how we've stayed the same.

Some spreadsheets will shout stories of alarm, while others whisper about curiosities we might have otherwise missed. Some will tell stories about secrets. Some will tell stories about lies.

The key to all these stories, as with all sources, is knowing *what questions to ask.*

This book is about those questions.

In fact, it is *full* of those questions, in the language that spreadsheets understand. They include questions like:

- How much has something changed?
- Who or where experienced the most change? Who changed the least?
- Which company or person got the most money?
- What is typical - and who stands out the most as being atypical?
- How often does a name appear? Where?
- Can you show me just the parts I'm interested in?
- How much was spent in total on something?
- Have I got things right?
- Where is information missing?
- What is unusual in this data?
- Where should I be looking?

Some techniques in this book are about helping the spreadsheet get into a position where it can answer questions it couldn't answer before: breaking dates down into years or months, extracting regions or names, for example, or combining it with other sets of data.

If you are new to spreadsheets you can read the book in the order it's written, with each chapter building on concepts and techniques introduced in the chapter before, plus a number of exercises and questions to give you spreadsheet experience alongside exercising your editorial muscles - the latter is just as important as the technical skills if you're going to be telling stories.

But the book is also written for **those who already have spreadsheet experience**: you can dip into chapters like a recipe book to find the chapter that tackles the question you want to ask. And once you've read it the first time, it's designed to be used again and again that way.

For that reason, please don't approach the book as something to be 'memorised' from beginning to end - it is better to remember what's possible *in general* with spreadsheets, and play with the techniques that are useful to you.

Then, when you need to do something particular, you will know you can find it in these pages.

The more often you use particular techniques, the more they will be retained in your memory, and the less you will need to refer to this book.

Oh, and if you have a question which isn't answered here, let me know and I'll try to update the book.

Later I will tackle some useful tips and techniques when getting to know a spreadsheet. But first, let's get stuck in with some basic calculations.

Starter jargon: cells, sheets and workbooks

I'll be using a number of terms from the start which I should explain now:

People often talk about 'spreadsheets' to mean one of two different things: either a document containing a number of different *sheets* of data, or just *one* of those sheets.

To avoid confusion I'm going to use the phrase **workbook** to refer to a spreadsheet *document* as a whole, and **sheet** or **spreadsheet** to refer to one of those individual sheets.

I'll also talk about **cells** - these are the individual boxes on a spreadsheet containing a piece of data.

2. Stories about change, stories in context: basic calculations

Questions asked as 95% of Birmingham events budget spent on Conservative Party Conference

Working out a proportion can generate surprising results - and raise more questions

At its most basic, a spreadsheet acts as a calculator. They can add, subtract, divide, and multiply, and through a combination of those can calculate averages, proportions, rates, and various other things which can provide an insight into the data.

Calculations in Excel are normally referred to as **formulae**, and every formula begins with an equals sign:

=

This tells Excel that you are not directly entering information, but rather are expecting it to do some *work.*

After the equals sign comes the calculation. Try the following in any cell in an empty spreadsheet:

=15+145

As soon as you press Enter, the calculation is performed and the cell shows the result (160 hopefully).

Most of the time, however, Excel is used to perform calculations on values that are in cells elsewhere in the spreadsheet.

Cell references

Every cell in a spreadsheet has its own reference, based on the row and column it is in. The very first cell, for example, is in column A and row 1, which makes it cell **A1** (the letter can be found above at the top of the column; the row number can be found to the far left).

The cell next to that in column B is **B1**, (different column, same row); and the cell *underneath* A1 is **A2** (same column, different row).

If in doubt, you can find out the **cell reference** (the location of the cell - in this case A1) by selecting it and looking along the bar just above your column letters.

To the left of this bar is a box that shows you the cell reference, and after that is a longer box which shows what that cell has in it. That longer box is called the **Formula bar**.

Note that if the cell contains a formula, you will see the formula and *not* the results of the formula - which is what you see in the sheet itself.

These cell references make it possible to perform calculations on data in your sheet.

For example, try typing 15 in cell A1, and 145 in cell B1.

Now in cell C1, type the following:

=A1+B1

The formula in C1 - note that the formula bar above shows the formula while the cell shows the result

That formula adds whatever value is in A1, to whatever is in B1.

There are two advantages to this:

- If the value in A1 changes, your formula automatically re-calculates. Try it and see (make sure you press Enter after changing the value).
- If you have two or more columns of numbers and want to create a new column full of the same calculation for each row, it's very easy to do so.

We'll see the second in action in the next chapter.

Calculating a change

Before then, it's worth exploring some commonly-used calculations you'll need to tell stories from your data.

One of the most common of those stories is about *change.* Change is, almost by definition, newsworthy, and an integral part of storytelling. When things change, it makes the news.

But how big a change was it - and from what?

A calculation to find out how much something has changed is a straightforward subtraction:

What something is now MINUS what something was before.

For example: if this year there were 1000 assaults in Aberdeen, and last year there were 600, the calculation is:

```
1000-600
```

The result of that is 400. In other words, assaults have gone up by 400.

Sometimes you will get a negative result, which means something has dropped.

Let's change our example to show that: this time let's say this year there were 1000 assaults in Aberdeen, and last year there were actually 1100. Now the calculation is:

```
1000-1100
```

And the result is -100. In other words, assaults have gone down, by 100.

Unfortunately, when we look at change in terms of an absolute number: up by 400, or down by 100, that doesn't tell us whether 400 or 100 is a lot, or not very much. And that's when *proportions* come in handy.

Almondst a third of criminals given restorative disposals reoffend

Almost a third of offenders given restorative disposals went on to commit another crime.

Data obtained under the Freedom of Information Act shows that 1,202 people in Windsor and Maidenhead were given restorative disposals between 2009, when they were first introduced by Thames Valley Police, and 2013.

Of these 339 reoffended.

Restorative disposals are a technique used by police to open communication between victims and perpetrators.

The aggrieved is given a direct involvement in the rehabilitation process. This could include meeting the perpetrator to discuss the consequences of the crime and getting a face-to-face apology.

According to Victim Support, the method can offer 'real benefits' to everybody involved.

This story is based on calculating a proportion - although it needs the context of what the proportions are for techniques other than restorative disposals

Calculating a proportion

An equally common calculation used in stories about numbers is **what proportion one thing is of another**. Here are just a few examples:

- 95% of Birmingham events budget spent on Conservative Party Conference[1]

---

[1] http://www.birminghampost.co.uk/news/regional-affairs/questions-over-95-birmingham-events-7028090

- US spends 17.9% of its GDP on healthcare[2]
- The most successful UK universities get 11% of their alumni to donate money[3]
- Almost one in three criminals given treatment which involves speaking to their victims reoffend[4]

Calculating a proportion is a simple division: **dividing a part by the whole**.

If the US spends $2.6 trillion on health, and its total GDP is $14.5 trillion, then to get the proportion above you divide 2,600,000,000,000 by 14,500,000,000,000.

The result is 0.179 (rounded down to three decimal places) - or 17.9%.

How is 0.179 the same as 17.9%? A useful approach is to think of it like this: 1 is the same as 100% - anything below 1 is a *fraction* of 1: 0.5 is *half* of 1; 0.25 is a *quarter* of 1; and so on.

0.5, then, can also be expressed as 50%, and 0.25 can be expressed as 25%.

To see those decimals as percentages we can do one of two things: either multiply them by 100 (0.5 x 100 = 50) to bring them to the other side of the decimal place, or simply change the formatting of the cell containing the result.

The latter is probably the better practice, as it retains the 'true' number and we can still perform calculations with it (100 times 50% is 50, for example, whereas 100 times 50 is 5000).

To do this, right click on the cell and select `Format cells...` - or use the keyboard shortcut CTRL+1. On the window that appears change the formatting to **Percentage** - you can also specify how detailed that percentage is.
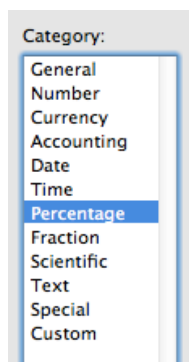
Ratios: calculating a proportion as '1 in 10'

One of the examples above doesn't present the proportion as a percentage, but rather a *ratio: 1 in 3.* This is a particularly clear way of presenting a proportion: it's much easier for readers to visualise

[2]http://www.theguardian.com/news/datablog/2012/jun/30/healthcare-spending-world-country
[3]http://www.nytimes.com/2007/04/15/weekinreview/15lyall.html?pagewanted=print&_r=0
[4]http://www.maidenhead-advertiser.co.uk/News/Areas/Maidenhead/Almost-a-third-of-criminals-given-restorative-disposals-reoffend-05022014.htm

the proportion 'one in three' than the rather abstract figure of '33%' (which you can get if you divide one by three - 0.33, or 33%).

So how do you express a percentage as a ratio? The most straightforward way is to look at the clear markers along the way.

For example:

- 0.05 is one in twenty (5%)
- 0.10 is one in ten (10%)
- 0.20 is one in five (20%)
- 0.25 is one in four - or a quarter (25%)
- 0.33 is one in three (33%)
- 0.40 is two in every five (40%)
- 0.5 is one in two (50%)
- 0.6 is three in every five (60%)
- 0.66 is two in every three (66%)
- 0.75 is three-quarters, or three in every four (75%)
- 0.8 is four out of every five (80%)
- 0.9 is nine out of every ten (90%)
- 0.95 is nineteen out of every twenty (95%)

Whatever your figure is, look at the proportion closest to it, and use that as a basis with the qualifier 'more than' or 'almost'.

For example:

- 0.06 is 'over one in twenty'
- 0.09 is 'almost one in ten'
- 0.23 is 'more than one in five'
- 0.27 is 'over a quarter'
- 0.36 is 'more than one in three'
- 0.39 is 'just under two in every five'
- 0.49 is 'almost half'

...and so on.

You can also use multipliers of some of the proportions in the main list: 0.15, for example, is three times 'one in twenty' and so could be expressed as 'three in every twenty' and 0.3 is 'six in every twenty'.

The more you do this, the more it becomes second nature. Until then, keep the list above as a useful reminder.

Combining both: calculating what proportion a change is

Welfare spending jumps 32% during Obama's presidency

💬 703 | SIZE: + / − | 🖨 PRINT | REPRINTS

This Washington Times story on welfare spending is all about calculating change as a proportion of the previous figures

So back to our change in assaults. Now we know how to calculate a proportion, we can calculate just how significant that rise of 400 assaults, or fall of 200, is.

The question is: what do we divide it by? The latest number, or the older one?

The answer lies in the way we would express it: 'Assaults have risen' or 'assaults have fallen'. Implicit in that is this: '*from what it used to be*'.

So: **you must divide your change by the original figure it has *changed from***

If assaults were 600 and have gone up to 1000, we need to work out how big a change that represents from the original figure of 600, i.e.

=400/600

The result is 0.66 - 66% higher, or two-thirds higher, *than its previous figure.*

The fall of 100 means a calculation like this:

=-100/1100

That's -0.09, or a drop of 9% *from its previous figure.*

Now when it comes to increases you need to be very careful about your language. You might talk about something being 66% *higher*, but you can also talk about figures being 166% what they were (the 66% increase plus the original figure - 100%). Make sure you are clear yourself about what it is you are saying.

Beware also of confusing *percent* with *percentage points*: the latter is used more often in politics, where a party's lead on a rival party might increase from 3% to 6% - that's an increase of three *percentage points* but also a 'doubling of their lead'. Try to avoid using "percentage points" outside of these very specific areas where usage has been better established, as they can cause confusion.

Another issue is whether you use the word 'percent' or the symbol '%'. If in doubt, check the style guides available online such as the APA grammar of percentage[5]

To see this process in practice, look at just one story in The Washington Times, *'Welfare spending jumps 32% during Obama's presidency*[6] *'*. The numbers behind the headline are given low down in the article:

> "Welfare spending as measured by obligations stood at \$563 billion in fiscal year 2008, but reached \$746 billion in fiscal year 2011, a jump of 32 percent."

So, 746 minus 563 is 183. 183 divided by 563 (the starting figure) is 0.325 - or 32.5%.

Recap

- Calculations are called **formulae** in Excel and other spreadsheet software
- A formula begins with an equals operator like so: =
- Formulae can work with numbers or strings (indicated by quotation marks), but are most likely to use **cell references** like A2 or B300.
- To calculate how much something has changed, take what something is now minus what it was before.
- To calculate what proportion the change represents, take that amount of change and divide it by the earlier figure (that it's changed from).
- Try to express proportions as a ratio such as 'almost one in ten' or 'over a quarter' - often this is easier for the reader to take in and understand than a dry, precise percentage - unless the precision is important.
- Beware of confusing *percent* with *percentage points* - the latter is used much less and has a different meaning.

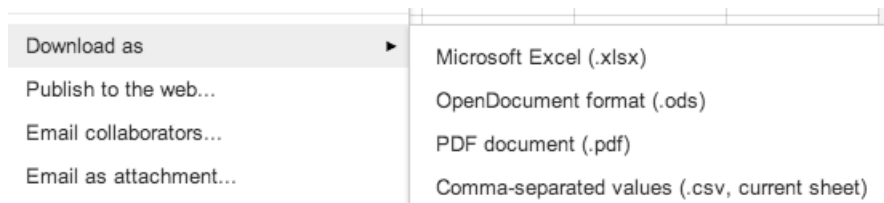[5]http://blog.apastyle.org/apastyle/2011/11/the-grammar-of-mathematics-percentage.html
[6]http://www.washingtontimes.com/news/2012/oct/18/welfare-spending-jumps-32-percent-four-years/?page=all

✏️ Find the story: are there more drunk and disorderly arrests?

fx	Month		
	A	B	C
1	Month	2011	2012
2	Jan	291	213
3	Feb	229	145
4	Mar	224	223
5	Apr	266	175
6	May	179	202
7	Jun	196	184
8	Jul	275	235
9	Aug	259	251
10	Sep	231	184
11	Oct	249	211
12	Nov	218	164
13	Dec	251	224
14			

At this link[7] you'll find a simple spreadsheet showing the number of people arrested for being drunk and disorderly in every month for two years. Download it by selecting **File > Download as > Microsoft Excel (xlsx)** or **Comma-separated values (CSV)**.

Download as	▶	Microsoft Excel (.xlsx)
Publish to the web…		OpenDocument format (.ods)
Email collaborators…		PDF document (.pdf)
Email as attachment…		Comma-separated values (.csv, current sheet)

You want to know *if arrests are going up or down*. Here's how:

1. Open the spreadsheet. Create a simple formula in D13, at the end of December's figures, to calculate the change in arrests that month between 2011 and 2012
2. Did it go up or down? By how much?
3. Can you write a new calculation to work out how much that is as a percentage change? Remember that you're looking at that change as a percentage of arrests in the *first* year.
4. Can you express that calculation as a proportion, e.g. 'one in ten' or 'half of'? What proportion would it be?
5. Why might it be important to have arrest numbers for each month of the last two years, rather than just looking at how arrests have changed in the last two months?
6. What headline might you use for this story - and what terms might you avoid?
7. Apart from whether arrests have gone up or down between the two years, what other stories might we look for in this data?
8. We'll cover how to repeat this calculation for all months in the next chapter, but see if you can copy the calculations for all the months.

I'll give the answers at the end of the next chapter.

---

[7] https://docs.google.com/spreadsheets/d/1iUMiNs7P5P1mEQXwsFDgIn8dF5ZnO4YhRcAEMsmciVI/edit?usp=sharing

# 3. How much did it cost? How many people were affected? The first function: adding a series of cells with SUM

NEWS ENGLAND

Home | World | UK | England | N. Ireland | Scotland | Wales | Business | Politics | Health | Education | Sci/Env

7 May 2014 Last updated at 05:58

Share

South West motorists receive £800,000 in pothole compensation

By Claire Jones
BBC News Online

Financial totals can make attention-grabbing headlines - although it isn't always made clear why it's significant
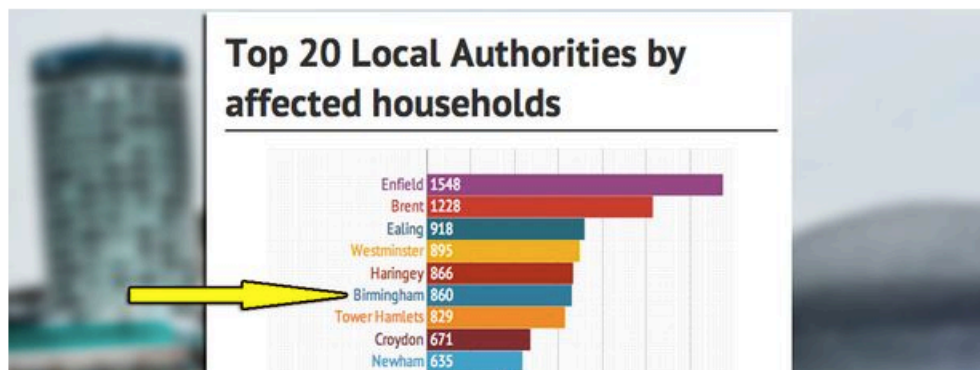
When it comes to numbers of people, however, the significance is often clearer. This story from the Birmingham Mail has a simple premise: how many people have been affected by a new policy

One of the most common stories you'll want to tell with a spreadsheet is 'How much did it cost?' These are *grand total* stories, and involve one of the most basic calculations you will want to perform in your spreadsheet: **adding up the values across a number of cells**.

Now, you could do this with a very long formula like so:

=A1+A2+A3+A4+A5

…And so on. But that would be laborious. It is for this reason that spreadsheets have what are called **functions**.

Functions are words that have special meanings in Excel and other spreadsheet software. They are a shortcut for a series of instructions.

The function for adding up a series of cells, for example, is called SUM. All you need to do to use SUM is tell it what cells you want to add up, in parentheses after the function.

Functions in different languages

Throughout this book I will be using the English language names for functions. However, if your spreadsheet software is not set up to use English then many functions will have different names. SUM, for example, is still SUM in Danish, but SOMMA in Italian and SUMA in Spanish and SOMA in Portuguese.

Many translators of Excel functions are available, including Dolf Trieschnigg's website[1], or you can search for 'Excel function SUM in Russian', replacing the function and language names with your own.

To add the values in cells A1 to A5 using that function, then, you might write the following:

```
=SUM(A1:A5)
```

It is easiest to think of functions as being like recipes: instead of having to write out 'add this and then add that and then add that' you can simply say: 'use the CAKE recipe with these ingredients', or 'use the PIE recipe with these ingredients'.

Once you know that you can start to guess which sets of common instructions are going to be recipes - functions - in your spreadsheet software. Here are just a few instructions which have their own one-word functions:

- 'Add all these figures up'
- 'Calculate an average for these figures'
- 'Tell me if a cell contains an error'
- 'Count how many numbers I have in this column'
- 'Extract the month from this date'
- 'Replace the full stops in this cell with commas'

There are dozens and dozens more, too.

Once you know this, you can start to *search for useful functions whenever you come across a task or problem which you think others may have already solved*, adding the word "function". For example, you might search Google for 'Excel function extract day from date' or 'Google spreadsheets function count blank cells'.

---

[1]http://dolf.trieschnigg.nl/excel/index.php

Why functions and not buttons

You can perform sums and other basic calculations using some of the buttons in Excel, but I'll be focusing on functions in this book for a few reasons:

- Firstly, typing functions is often quicker than using buttons. It's especially quick when you want to combine results - for example using a sum as part of a bigger calculation. You also have more control over where the results go.
- Secondly, the placements and availability of buttons varies widely between spreadsheet software, and between different versions of Excel, so it's difficult to give a consistent guide to their use. Knowing functions means you can use the same techniques whatever version of spreadsheet software you end up on (although there are a few which are specific to Google spreadsheets or Excel - I'll mention this at the time).
- Finally, once you get beyond basic calculations there won't always be buttons available to do what you need. You might as well start as you mean to go on...

Another SUM story - this time adding up how much empty housing there is across a number of countries

This is how much West Yorkshire Police spent on informants in the past five years

West Yorkshire is the UK police force with the seventh highest spending on informants, a new analysis has revealed.

By Georgina Morris
Wednesday, 29th January 2020, 6:00 am

Using Freedom of Information to find out how much was spent on particular things, or their costs, is a well-worn approach - it's always best to ask for amounts in different years, and breakdowns where possible to provide more concrete details and context

If functions are recipes, arguments are the ingredients

You can't cook a recipe without ingredients, and so (just as with our SUM example) every function is followed by at least one ingredient, in parentheses. These are called **arguments**.

In our first example above, SUM was the function and A1:A5 was the argument.

Sometimes you will get an error message because you didn't 'supply the right arguments'. For example, you may have submitted too many or too few (or none at all). Normally the error message will give you clues, and even a link to more information. And of course you can search for more information on that error. But knowing the jargon helps.

Argument or parameter?

You may see the term **parameter** also used to refer to a function's ingredients. This is the name for what type(s) of ingredients a function needs *in general*. So, for example, the SUM function has one parameter: a range of cells to add up. But the *formula* =SUM(A1:A5) supplied one **argument**: the *specific* range of cells A1 to A5.

In other words, when you read about a function you will hear about the parameters it uses; but when you actually use them, you will be supplying your own specific arguments. The same jargon is used in programming (for example, advanced scraping or data visualisation), where functions are common, so if you understand this you'll have a head start if you decide to try out programming.

Back to our formula, then. You'll notice that we specify a range of cells using the colon symbol:

=SUM(A1:A5)

So if we wanted to add all the cells from A1 to A500 we could adapt it like so:

=SUM(A1:A500)

You can also select cells across more than one column. If we wanted to add all the cells from A1 in the top left of our range to B500 as the bottom right (last) cell we could select them with this range:

```
=SUM(A1:B500)
```

Crucially, functions like SUM which work with numbers will **ignore any words in the specified cell range**. In other words, text does not cause it any problems.

For that reason, if all you want to do is add up all the numbers in a column you can often do so much more quickly by just using the column *letters* like so:

```
=SUM(A:A)
```

This simply means: *use the SUM function on all the numbers in column A.*

The same principle applies to rows:

```
=SUM(1:1)
```

Would mean *add the values in all cells in row 1.*

⚠ Checks to make when specifying a column- or row-only range

If you are going to specify a column or row range such as A:A or 1:1, check the following:

- The heading is not numerical, e.g. years or dates. If they are, they will be included in the sum.
- There is no grand total at the bottom of your column or the end of your row. This will also be included in the sum, giving you a total twice as large as it should be.
- There are no further tables beneath the one you are working with (the formula would add all numbers in column A in that table too).

You can check these issues quickly using the keyboard shortcuts explained in the chapter on getting to know the spreadsheet - CTRL or CMD and the cursor keys - to explore the edges of your table.

When there's more than one ingredient: commas and semicolons

The SUM function is a simple recipe, and so only needs one ingredient. But other functions take more than one - as we'll see in further chapters.

In these cases, each ingredient - argument - is separated by a comma (in English language versions of software) or semicolon (in Spanish, Portuguese, German and other language versions).

In fact, SUM is unusual in that it can take one or *more than one* argument: you can add more than one range of cells by simply adding a comma or semicolon between each one like so:

```
=SUM(A1:A5,A7:A9)
```

```
=SUM(A1:A5;A7:A9)
```

This adds together the values in cells A1 to A5, and from A7 to A9.

But before we come on to other functions with multiple ingredients, there are some other simple ones to explore.

> For the rest of this book I'll be using commas, but remember to use semicolons instead if your software uses those.

Recap

- **Functions** are words that have special meanings in Excel and other spreadsheet software: a shortcut for a series of instructions.
- Functions are always followed by parentheses containing the ingredients it needs. These are called **arguments** (specifically) or **parameters** (in general).
- If a function uses more than one parameter, each one is **separated by a comma or semi-colon**, depending on the language of the software you're using.
- You can **search for useful functions** whenever you come across a task or problem which you think the designers of the spreadsheet software may have already solved, adding the word "function".
- Specify a range of cells by putting a colon between the first and last cell in that range, e.g. A2:A300.
- You can *more quickly* **select a whole column or row** by only using the letter or number of that column or row, e.g. A:A or 1:1

✒ Find the story: drunk and disorderly arrest totals

To finish with our drunk and disorderly arrests data[2], type a calculation using the SUM function that **works out the totals for each year**.

1. What formula will add up all the numbers for 2011, and what formula will do the same for 2012? (You can always copy the calculation into a cell to the right and it will apply relevant cells for the next year)
2. Why might using a column range like B:B not be a good idea in this case?
3. What formula would calculate the change from 2011 to 2012?
4. Based on the result of that, what formula would calculate that change as a percentage?
5. What simple stories might you tell about these totals?
6. What other data could you look for to make a stronger story?

The answers will be at the end of the next chapter.

---

[2]https://docs.google.com/spreadsheets/d/1iUMiNs7P5P1mEQXwsFDgIn8dF5ZnO4YhRcAEMsmciVI/edit?usp=sharing