# Guidance for Chapter 3, Extra Exercise 1

## How to calculate a percentage in Excel

We’re going to start with the *Titanic* dataset, which looks like this:



The first thing we want to do is to calculate what percentage of people on board the *Titanic* survived the disaster. To do this we need two pieces of information:

1. How many people were aboard in total (the **denominator**)
2. How many people survived (the **numerator**)

Calculating the denominator is very easy. We can just scroll to the bottom of the list of names. We can see that the last name on the list (Mr Leo Zimmerman) is in row 1,310. We therefore know that there were 1,309 people aboard (the top row is the variable names) – that’s our denominator.

To calculate the numerator we need to know what the numbers in the ‘survived’ column mean. If we look at the ‘Notes’ sheet in the data we can see that 1 means the person survived and 0 means they didn’t. So to get our numerator we need to add up all the ones in the ‘survived’ column. To do this we use the Excel ‘Sum’ formula.

Find an empty cell and type:

=sum(e2:e1310)

Typing a formula like this into Excel is a sort of ‘command’: you are telling Excel that you want it to do something. In this case you are telling it that you want it to add up (sum) all the numbers in column E, starting from row 2 and ending at row 1310 (the last row of data).

After you’ve typed in the formula, hit enter. You should then see that the formula disappears and is replaced by a number. Excel has gone away and done what we asked and come up with an answer: 500. That’s our numerator.

*Note: Here we told Excel what range of cells we wanted to add up by just typing the numbers into the formula. However, you can also click and drag over the cells you want rather than typing them in. Just click and drag over the relevant cells after you have typed the opening bracket, then, when you’re done, add the closing bracket*.

Now that we have our numerator and denominator, it’s a simple matter to calculate the percentage. You could do this on a calculator if you wanted, but we’ll do it in Excel.

In an empty cell, type the following formula:

=500/1309

This formula tells Excel to divide 500 by 1309 and give us the answer, which (rounded to two decimal places) is 0.38.

To get this as a percentage (rather than a proportion), we just multiply this number by a 100. We can amend the formula above so that Excel does this for us:

=(500/1309)\*100

This formula tells Excel to divide 500 by 1309, then multiply the result by 100 and give us the answer, which (again rounded to two decimal places) is 38.20%. This is the percentage of *Titanic* passengers who survived.

## Selecting a random sample in Excel

The next step in the exercise is to select a random sample of 100 passengers. To do this we will make use of Excel’s ability to generate random numbers.

First create a new column (Column F) called ‘Random’ (you can call it whatever you want). Then, if the first cell below the title (cell F2) type the following formula:

=randbetween(1,1000000)

This formula tells Excel to generate a random number between 1 and 1 million. Each person in the data now has a random number associated with them.

Now select all of the data. You can do this by pressing the Ctrl and A keys together, or by clicking the little triangle at the top left of the spreadsheet.

Once all the data is selected, go the ‘Data’ tab in the Excel ribbon. Then click the ‘Sort’ Button. This dialogue box will appear:



We want to sort the data by our new ‘Random’ column. So click the arrow next to the ‘Sort by’ box, then select ‘Random’ from the list.

*Note: If you don’t see Name, Gender, Age, Random, etc. as options, but instead see Column A, Column B, Column C, etc., then you need to use the tick box in the top right of the dialogue box to tell Excel that the top row of your data is your title row (contains headers) and is not actually part of your data.*

Once you click OK, all the people in the data will be sorted not alphabetically, but according to the random number Excel gave them. The top 100 people in the list will now be a random sample of the passengers.

*Note: You will notice that the numbers don’t seem to be in the right order after you sort. This is because Excel re-generates the random numbers after everything you do. So every time you tell Excel to do something, it will re-generate the numbers in this column. If you want the numbers to stay still, you can copy them into a new column and select ‘Paste as values’.*

## Calculating a percentage for a subsample

Now we want to calculate the percentage of people in our random sample of 100 who survived the disaster. To do this, we again need the numerator and the denominator.

Again, the denominator is easy – it’s 100.

To calculate the numerator we again need to use the Excel ‘Sum’ formula:

=sum(e2:e101)

You can see that this time, instead of calculating the sum of the ‘survived’ column for the whole dataset, we are only calculating it for the top 100 passengers (our randomly selected sample).

Once we have our numerator and denominator we can again calculate the percentage (this is even easier than last time because our denominator is already 100). The first time I did this I got 37%. Your random sample is different so you might get a different number.

## Additional guidance

There is a lot more to learn about Excel formulas. A good place to start is [GCLearnFree](https://www.gcflearnfree.org/excel2016/intro-to-formulas/1/).