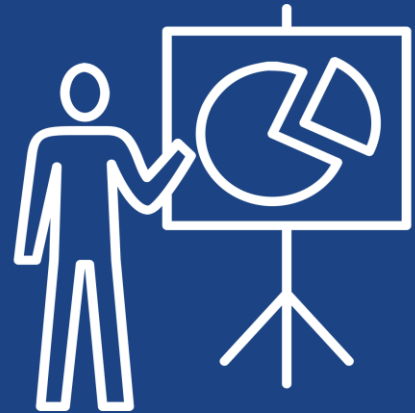




# DATA**SHIFT**

Data for Action Toolkit

## Working with Survey Data in Excel



## ■ Introduction

This toolkit has been developed for CIVICUS-DataShift's Data for Action programme. CIVICUS is a global alliance of 3,600+ civil society organisations (CSOs) and activists dedicated to strengthening citizen action and civil society around the world. DataShift is dedicated to building the capacity and confidence of organisations to produce and use data.



This toolkit was made possible by the European Commission.

[www.thedatashift.org](http://www.thedatashift.org)

## ■ Data for Action

Our vision is an informed civil sector empowered to collect and use data to amplify local narratives. The Data for Action programme provides organisations with facilitated training in using data for evidence-based decision making. By developing action-oriented surveys, we demonstrate how data can support an organisation's work during three key phases of the project lifecycle:



- **Scoping** - What should we do?
- **Programming** - How should we do it?
- **Monitoring** - Is it working?

## ■ Who is this toolkit for?

This toolkit is part of a larger effort by DataShift to strengthen organisational capacity to work with data in an actionable way through the Data for Action programme. It was developed to serve as a guide for organisations in thinking critically about survey data using spreadsheets.



This resource has been developed for small- to medium- sized civil society organisations working with survey data. The instructions contained in this manual are designed specifically for data entry by a single user using one computer at a time. This is not intended to be applicable for large-scale data collection utilising multiple data entry personnel on multiple devices.

## ■ How to use this toolkit

This toolkit has been designed to fit two purposes:

- As a quick reference for common functions and tools when working with survey data in Excel
- As a guide for beginners working with survey data and Excel for the first time





## ■ Toolkit features

### Data security

Best practices for ensuring responsible data and data security are highlighted throughout this toolkit. However, we encourage users to seek additional training resources to build up best practices in this area.

### Table of Contents

Use this tool to ease searching for specific topics within the toolkit.

### Glossary

A glossary of terminology used in this toolkit is can be found in Annex A.

### Checklists

Annex B contains a series of quick checklists listing the main steps for each process.

#### Take note

*All of the screenshots and video demonstrations provided in this toolkit have been recorded using Microsoft® Excel for Mac (2017).*



Our goal is to demonstrate some of the **most useful functions** Excel is capable of performing. While the location of some buttons and tools may vary slightly depending on the version of Excel you are using and the type of computer (Mac or PC), all of the functions will remain the same.

#### Not using a Mac?

Locate a particular tool or function for a different version/operating system using a Google search. Simply type the function you are looking for, the version of Excel you are using, and the operating system. For example, “creating pivot tables in Excel 2007 for Windows.” Performing this search results in a number of tutorial videos to walk you through creating pivot tables using this version of the programme.



## ■ Table of contents



<b>Introduction</b>	2
<b>Data for Action</b>	2
<b>Who is this toolkit for?</b>	2
<b>How to use this toolkit</b>	2
<b>Toolkit features</b>	3
<b>Thinking critically about data</b>	5
Minimising decision-making during research	6
<b>Data entry</b>	7
Quality assurance	7
Creating keys for data entry	8
Data validation	9
Drop-down menus	9
How to create drop-down menus in Excel	11
Single versus multiple response questions	12
<b>Data cleaning</b>	14
Data triangulation	14
Duplicate entries	15
Checking for outliers	16
<b>Data analysis</b>	17
Count if	17
Sum	18
Average	18
<b>Responsible data visualisation and communication</b>	19
<b>Annex A: Glossary of terms</b>	20
<b>Annex B: Survey methods checklist</b>	23

## Thinking critically about data

Before designing a survey, it is important to understand the ***purpose of the research*** and how you intend to ***make decisions with the resulting data***. This does not mean knowing the actual questions that you will put in the survey. Rather, it means understanding exactly what you intend to do with the data once it has been collected. Think critically about how to make your data actionable at each step.



The benefit of ***making an effort to design a strong, actionable survey*** far outweighs the consequences of attempting to turn bad data into something useful.

It is common practice for organisations to conduct research in the following manner:

- 1) Create a research question
- 2) Design a research tool (i.e. survey, interview guidelines, etc.)
- 3) Collect the data
- 4) Input the data into a database
- 5) Conduct data analysis

This linear approach often fails to produce actionable data. Why? This is because if and when issues arise during the later stages of data collection, cleaning, and analysis, it is already too late to fix. A best practice is to establish a data management plan in advance that describes each of the following components:

- 1) Research objectives
- 2) Data collection plan and sampling methodology
- 3) Coding
- 4) Database design
- 5) Data entry and collation
- 6) Cleaning and validation
- 7) Analysis
- 8) Visualisation
- 9) Communication and publishing
- 10) Data security
- 11) End of project plan (data destruction)
- 12) Quality assurance



With a plan already in place, it is easier to test each component ahead of time and develop each component simultaneously to ensure you are getting the results you desire.



For example, after pilot testing your survey with a small sample group, go ahead and practice your data coding scheme, enter the data into your pre-prepared database, and conduct analysis on the test data. This will give you a better idea if your data is:

- 1) In line with your research objectives
- 2) Producing the desired results
- 3) Entered and coded appropriately and logically

If you experience any difficulties entering, coding, or analysing the data, you can proactively make changes before investing time and resources in full-scale data collection.



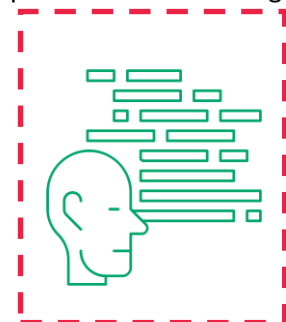
### Minimising decision-making during research

A lot of decisions are made when conducting research. An enumerator senses discomfort and considers skipping a question on a survey. Data entry staff encounter a question with two answers marked instead of one. Situations such as these arise often and sometimes the appropriate course of action is unclear. At the end of the day, the fewer decisions an individual working on the project makes on their own, the better. Clearly lay out how to handle these situations in advance to minimise errors and increase data quality.



## ■ Data entry

Data entry is the process of taking survey data and entering the responses into the spreadsheet. There are many ways to do this, but the ultimate goal is to be as accurate as possible while creating a spreadsheet that is relatively simple to understand and use.



Data coding is the practise of reducing data into a manageable form to make it easier to analyse and visualise. It also refers to the process of taking qualitative data, such as responses to open-ended questions, and assigning a reference, or code, to each type of response.

Data entry may seem like a simple process, but it is necessary to plan out how the data is organised in the spreadsheet, how the data will be represented (i.e. coded), and to put measures in place to reduce errors (i.e. bias).

### Quality assurance

Quality assurance refers to the process of ensuring that your data is of high value and accuracy. Before entering data into a spreadsheet, the surveys should be reviewed for completeness and accuracy. Any inconsistencies should be crosschecked and verified prior to continuing with data entry.

#### Quick tip

Survey data should be reviewed by the enumerator immediately after conducting a survey in the field and again by a supervisor at regular intervals. Thorough reviews should be sure to check for:

- Completeness
- Accuracy
- Legibility
- Existence of unique identifier (survey ID)



One way to minimise errors in data entry is to reduce the number of people actually performing the data entry. Ideally, one person is assigned to this task. This may not be possible, depending on the number and length of the surveys.



A second way to reduce the burden of typing in all individual responses is to create a key of responses using short, simple text. We will review some of the best practices for this, when to do it, and how it will affect the appearance and usability of your data.



Finally, a great way to ease data entry and reduce errors is through the use of the data validation functions in Excel, especially the drop-down menu option. This section will go into more detail about how to utilise these functions to produce high-quality spreadsheets.

## Creating keys for data entry

Consider the survey question below, which comes from a human trafficking survey that was conducted in Nepal.

2) What is the primary use of your data?

☐ For writing reports and publications

☐ For writing of grant proposals

☐ For advocacy purposes

☐ To share with other organisations working on trafficking

☐ To create GIS maps

☐ Other:

Each response item contains multiple words. We encounter challenges when entering these long response items into a spreadsheet for two primary reasons:

- It is time-consuming
- It is highly prone to errors in data entry

We overcome these challenges in two practical ways:

- Creating keys
- Data validation

A key is a document that keeps track of the how the data is entered into the spreadsheet. Using the example question above, we can make our lives easier by assigning a shortened word or phrase to refer to each piece of the questions, as shown in the table below:

Original	Key
2) What is the primary use of your data?	Q2
For writing reports and publications	Reports
For writing grant proposals	Grants
For advocacy purposes	Advocacy
To share with other organisations working on human trafficking issues	Sharing
To create GIS maps	GIS
Other:	Other

The cells in the column containing responses for question two will only contain one of the shortened phrases listed in the right column.

There are a number of different ways to create keys and to code data. Each person you meet may have a different preference. For the purposes of this toolkit, we will avoid traditional coding mechanisms which assigns numbers or letters to response items. Rather, we will focus on easing data entry using shorthand entries for identifying questions and responses, such as in the example above.



## Data validation

Using the methodology from the previous section, we can manipulate the spreadsheet to allow only permitted responses to be entered into a particular cell. To do this, Excel has a function called “data validation,” which limits and defines the numbers and types of data that a cell will accept. A user can create rules to restrict the number and types of values that are acceptable in a particular set of data. When an unacceptable value is entered, the user is notified and must re-enter an approved value. Data validation is a powerful tool for safeguarding data accuracy and we highly recommend using this feature to prevent typos or inaccurate data.

Instructions:

- 1) Select the range of cells that you want to apply the new rule to
- 2) Click on the “data” tab
- 3) Click on the “data validation” button
- 4) Specify the data validation criteria
- 5) Click “ok”
- 6) Test your new rule by entering both a valid and invalid value



### Best Practice

Use data validation and drop-down menu for data entry to reduce errors

## Drop-down menus

Drop-down menus are a specific type of data validation rule and should be used in the case that you have a set list of response items for a question. Drop-down menus define and limit the response items that a user is able to input into a particular cell. When a drop-down menu has been created and assigned to a set of data in a spreadsheet, data entry is limited to the set of responses contained within the drop-down menu. This feature is a great tool for quality assurance, as it reduces data entry errors and mistakes caused by misspelled words, extra spaces, or accidentally entering the data in the wrong cell.



### Security tip

Restrict access to the original database. Data entry personnel should only be able to enter survey data into select cells using drop-down menus and cells with defined data validation rules. Visit the “tools” tab and scroll to “protection” to review options for protecting workbooks and sheets.

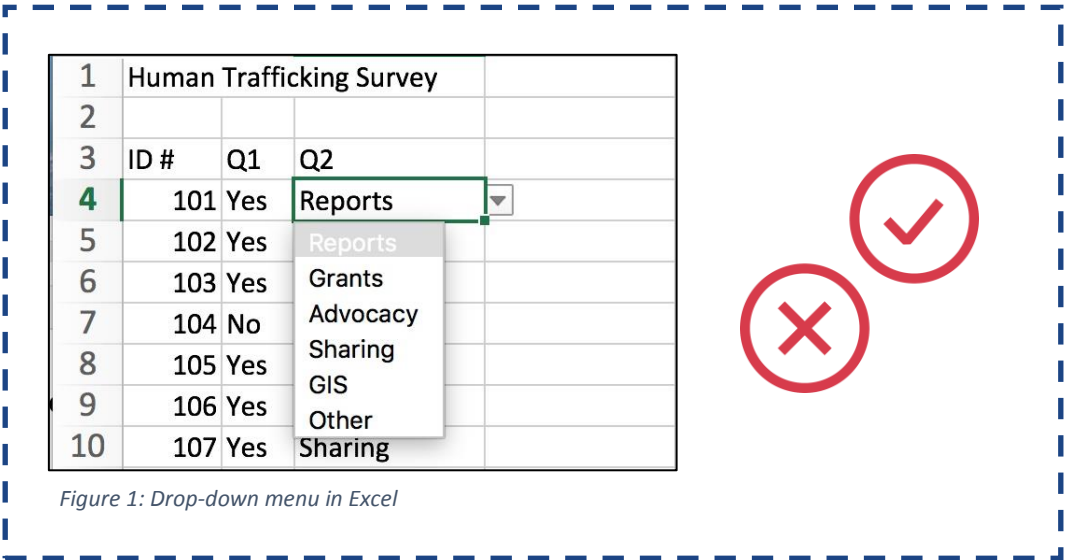
Before you are able to set up a drop-down menu, you must first define the acceptable values that a user will be able to select from. Let’s continue with the same example from above. We have already defined the shorthand code for the response items to the following survey question:

2) What is the primary use of your data?

- ☐ For writing reports and publications
- ☐ For writing of grant proposals
- ☐ For advocacy purposes
- ☐ To share with other organisations working on trafficking
- ☐ To create GIS maps
- ☐ Other:

Original	Key
2) What is the primary use of your data?	Q2
For writing reports and publications	Reports
For writing grant proposals	Grants
For advocacy purposes	Advocacy
To share with other organisations working on human trafficking issues	Sharing
To create GIS maps	GIS
Other:	Other

The cells contained under the column heading “Q2” (question 2) will all contain one of the following coded entries: reports, grants, advocacy, sharing, GIS, or other. Therefore, we need to create a data validation rule that allows for drop-down menu where the user can select one of these six acceptable entries.



## How to create drop-down menus in Excel



Question code must begin with an underscore or a letter and cannot contain any extra spaces. We suggest using the question number in the form of "question1"

\*

\*\* Alternatively, you can access via the insert tab → name → define

Type "=" and then the code used to define the list of responses. In our example, you would enter "=question1"

\*\*\*

### Single versus multiple response questions

So far, we have reviewed data entry for single-response questions. Single response questions are those in which the respondent can only provide one response from a set of response items. The following question comes from the same human trafficking survey from the previous example.

Q1) Do you collect data on human trafficking?

- ☐ Yes
- ☐ No

In this question, the respondent could either give the response “yes” or “no,” but should not mark both responses.

Whereas coding for single response questions is fairly straightforward, it becomes a bit more complicated when working with multiple response questions. Multiple response questions are those in which a respondent can select more than one option from a list of response items. We highly recommend using single-response questions unless you can make a strong argument for the use of multiple response questions. This is because analysis can get complicated for multiple response questions, especially if the analysis plan has not been determined in advance.

### Best Practice

Avoid multiple response questions unless you have a strong argument for why it is necessary and how you plan to analyse the resulting data.

Taking a look at the example question below, we can see that the respondent or enumerator is instructed to check all of the responses that apply:

2) Do you collect data regarding (check all that apply):

- ☐ Trafficking routes
- ☐ Origins of victims
- ☐ Potential destination of victims
- ☐ Border crossings potentially used by trafficker and victim
- ☐ Age of victims
- ☐ Number of individuals being trafficked
- ☐ Other: \_\_\_\_\_

There are a couple of ways to go entering this data into a spreadsheet, but we are going to use the methodology most commonly used for data analysis purposes.

Using our methodology, we could create a key for this question as follows:

Original	Key
3) Do you collect data regarding:	Q3
Trafficking routes	Routes
Origins of victims	Origins
Potential destination of victims	Destination
Border crossings potentially used by trafficker and victim	Border
Age of victims	Age
Number of victims being trafficked	Number
Other	Other

Each cell in your database should only contain **one** piece of data. If a respondent selects more than one response to the question, we cannot enter both responses into a single cell because then we would be entering more than one piece of data into the cell. We overcome this by creating a **separate column** for each response item and then treating each one as if it were a “yes” or “no” question.

- 1) Begin by adding columns to your spreadsheet equal to the number of response items.
- 2) Then label the header row with the question number code and the response item code.
- 3) Then set up a drop-down menu for each column with the response options “yes” and “no”
- 4) Then enter the data, selecting “yes” if the respondent selected that response item and “no” if they did not

	A	B	C	D	E	F	G	H	I	J
1	Human Trafficking Survey									
2										
3	ID #	Q1	Q2	Q3 Routes	Q3 Origins	Q3 Desintation	Q3 Border	Q3 Age	Q3 Number	Q3 Other
4	101	Yes	Reports	Yes	Yes	Yes	Yes	No	No	No
5	102	Yes	Reports	Yes	Yes	Yes	Yes	No	No	No
6	103	Yes	Advocacy	No	No	No	No	Yes	Yes	No
7	104	No								
8	105	Yes	Grants							
9	106	Yes	Reports	Yes						
10	107	Yes	Sharing	No						
11	108	Yes	GIS							
12	109	No								
13	110	Yes	Grants							

Figure 2: Coding and entering multiple response data into Excel

**Note:** In some cases, you it may make more sense to list your response items for multiple response questions in a difference way, such as “selected” and “not selected” or “yes” and “no response” – regardless



of how you decide to label the responses, the goal is to turn the multiple response questions into a series of binary response questions.

## ■ Data cleaning

Before the 2012 London Olympics, thousands of eager fans flooded online to buy tickets. They used an online ticket agent, Ticketmaster, which coordinated with the organising committee to manage the ticket sales. It was soon realised, however, that four events were oversold by 10,000 tickets. Staff was then required to contact customers and explain their error. It was discovered that this mistake was the result of a simple data entry error in their spreadsheets by a staff member that accidentally entered 20,000 instead of 10,000 in the cell containing the number of remaining tickets.<sup>i</sup>



This example highlights how a simple error in data entry can have a massive impact on your data and how it is used. In this section, we will review some basic data cleaning methods to ensure your data is as accurate as possible.

Data cleaning is the process of reviewing databases for inconsistencies or suspicious responses. It also involves the process of modifying and reformatting spreadsheets to optimise them for certain data analysis tools.

In this section, we will review a few useful data cleaning functions including:

1. Data triangulation
2. Conditional formatting for duplicate entries
3. Checking for outliers



### **Data triangulation**

In an ideal world, someone would sit with all of the physical surveys and crosscheck the original responses with those entered into the spreadsheet. Unless you are only collecting a handful of survey, this is usually not feasible.

Data triangulation is the process of crosschecking an original survey with its corresponding entries in Excel.

1. Select a random survey
2. Crosscheck the data in Excel
3. Correct any errors



This methodology for data cleaning is useful to search for random errors and to determine the relative accuracy of your data entry process. Significantly, data triangulation can help identify any systematic errors, or those errors occurring frequently and in similar patterns. These errors can then be investigated more closely and a decision can be made about how to handle the issue.



## Best Practice

Begin data triangulation as a quality assurance practice from the beginning of data entry. This will help catch errors early on.

### Duplicate entries

Duplicate entries can be highlighted using a function called “conditional formatting. There are two ways available to access the conditional formatting window and specify the parameters. Both processes are described below.

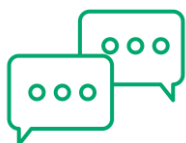
This is a useful function to search for surveys that have been entered multiple times by mistake. Each survey has been assigned a unique identifier, which also appears on the physical survey. By searching the unique identifier column for duplicate entries, we can quickly review our database to ensure each survey has only been entered once.



	A	B	C	D	E	F	G	H	I	J
1	Human Trafficking Survey									
2										
3	ID #	Q1	Q2	Q3 Routes	Q3 Origins	Q3 Desintation	Q3 Border	Q3 Age	Q3 Number	Q3 Other
4	101	Yes	Reports	Yes	Yes	Yes	Yes	No	No	No
5	102	Yes	Reports	Yes	Yes	Yes	Yes	No	No	No
6	103	Yes	Advocacy	No	No	No	No	Yes	Yes	No
7	104	No								
8	105	Yes	Grants	Yes	Yes	Yes		Yes	Yes	No
9	106	Yes	Reports	Yes	Yes	Yes		Yes	Yes	No
10	107	Yes	Sharing	Yes	Yes	No		Yes	Yes	No
11	108	Yes	GIS	No	Yes	Yes		No	Yes	
12	109	No								
13	110	Yes	Grants	Yes	No	Yes	Yes	No	Yes	Yes
14	110	Yes	Grants	Yes	No	Yes	Yes	No	Yes	Yes
15	111	Yes	Reports	Yes	Yes	Yes	No	Yes	Yes	No
16	112	Yes	Other	No	No	No	Yes	No	Yes	No
17	113	No								

Figure 3: Conditional formatting window in Excel

Once the duplicate values have been highlighted, you can now go through the entries and determine how the error occurred. If the survey data has been entered twice, you can simply delete the extra row. If the duplicated unique identifiers appear to contained different data in the rows, the data will need to be triangulated with the original survey to determine the source of the error and clean the data accordingly.



### Excel tip

Conditional formatting makes it easy to visualise patterns in Excel. You can use this function to highlight data based on almost any criteria you require.

B17					
	A	B	C	D	E
1	Human Trafficking Survey				
2					
3	ID #	Q1	Q2	Q3 Routes	Q3 Origins
4	101	Yes	Reports	Yes	Yes
5	102	Yes	Reports	Yes	Yes
6	103	Yes	Advocacy	No	No
7	104	No			
8	105	Yes	Grants	Yes	Yes
9	106	Yes	Reports	Yes	Yes
10	107	Yes	Sharing	Yes	Yes
11	108	Yes	GIS	No	Yes
12	109	No			
13	110	Yes	Grants	Yes	No
14	111	Yes	Reports	Yes	Yes
15	112	Yes	Other	No	No
16	113	No			
17		10			

### Conditional formatting using tabs

1. Select the column containing the unique identifier (survey ID)
2. Go to the “format” tab
3. Select “conditional formatting”
4. Ensure the top tab reads “show formatting rules for: current selection”
5. Select “highlight cell rules”
6. Click the small + on the bottom left
7. Under the style tab select “classic”
8. Select “format only unique or duplicate values” in the following drop-down menu
9. Select “duplicate” in the next drop-down menu
10. Select the colour and formatting of the duplicate values
11. Click “ok”
12. Duplicate entries should appear in the selected formatting



### Conditional formatting using button

1. Select the column containing the unique identifier (survey ID)
2. Click on the “conditional formatting” button
3. Select “highlight cell rules”
4. Select “duplicate values”
5. Click “ok”
6. Duplicate entries should appear in the selected formatting



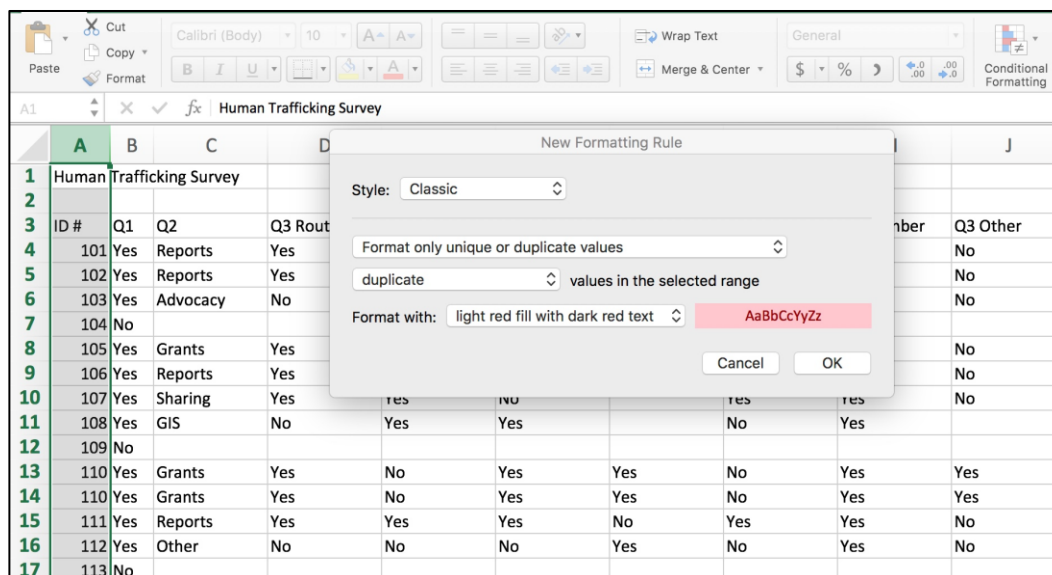


Figure 4: Conditional formatting in Excel

### Checking for outliers

Outliers are data points that fall outside of the normal set of responses. Sometimes these are accurate, but other times they are due to an error in data recording or entry. Use the data validation function (see page 9) or conditional formatting to search for and highlight data lying outside of the typical range. Correct any irregularities using data triangulation.

## Data analysis

Once you've entered all of your survey responses and cleaned your dataset, the next step is data analysis where you will begin the process of learning from your database. In short, data analysis is the process of taking a set of information to understand the broader trends within the population with the goal of making useful conclusions to inform your actions. Understanding what your data is telling you does not have to be a complex process. Using a few of the features and tools in Excel, we can begin to answer basic questions about the sample population.

Formulas are instructions entered into cells in Excel.

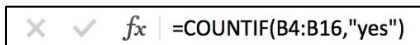
1. All formulas start with an equal sign (=)
2. Pay attention to the placement and use of parentheses
3. A range of cells is denoted by the cell reference for the first cell and the final cell separated by a colon. For example, the range from cell B2 to B30 is written "B2:B30"

## COUNT IF

This formula will count the total number of all cells lying in a specific range that fit specified criteria.

### Formula

=COUNTIF(range, criteria)

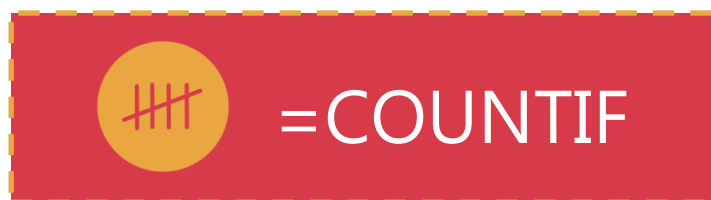


### Example

In this example, the “count if” function is used to add up the total number of respondents that answered “yes” to question one.

As you can see in the formula bar, the range is specified as “B4:B17” and the criteria is “yes” – the criteria must be expressed exactly as it appears in the cell and must be in quotation marks. Now Excel will automatically count the number of cells between B4 and B16 that contain the response “yes.”

Q1
Yes
Yes
Yes
No
Yes
Yes
Yes
Yes
No
Yes
Yes
Yes
No
=COUNTIF(B4:B16, "yes")



## SUM

This formula will add up the total sum of the numbers in the range of cells specified in the formula. **Formula** =SUM(number 1, [number 2],...)

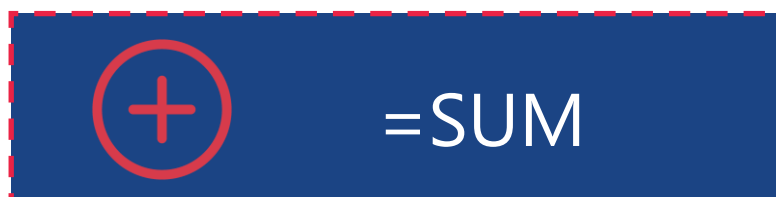
### Example

In this example, the “sum” function is used to add up the total sum of the numbers for question four in column K.

In the formula bar, the range is specified as “K4:K16”

Excel will automatically count the total sum of the numbers contained in cells between K4 and K16.

Q4
3
4
3
5
7
3
2
6
1
2
7
4
0
=sum(K4:K16)



## AVERAGE

This formula will calculate the average of the numbers in the range of cells specified in the formula.

### Formula

=AVERAGE(number 1, [number 2]..)

✗ ✓ *fx* =AVERAGE(K4:K16)

### Example

In this example, the “average” function is used to determine the average, or mean, of the numbers for question four in column K.

In the formula bar, the range is specified as “K4:K16”

Excel will automatically calculate the average of the numbers contained in cells between K4 and K16.






Q4	
3	
4	
3	
5	
7	
3	
2	
6	
1	
2	
7	
4	
4	
=AVERAGE(K4:K16)	



=AVERAGE

## Step 6: Responsible data visualisation and communication

This section is meant to briefly review a few basic considerations for visualising and communicating data in a responsible manner.

-  Always describe the research question or the purpose of the survey
-  Acknowledge any limitations of the study design (questionnaire, sampling, analysis, etc.)
-  Understand if your data is not representative and be sure to use language that informs the reader
-  Correlation does not mean causation! Just because your data indicates a connection, or correlation, between variables and outcomes, does not mean that one caused the other. Be sure to list any reason that correlation may exist beyond causation, such as missing data, low response rates, and other confounders.
-  Do not create graphs or charts that drastically skew the axes for a more dramatic effect



## ANNEX A

### ■ Glossary of Terms

#### **Bias**

Biased data is data that is misleading or influenced by external factors. There are ways to minimize bias through the research design, method of data collection, types of questions that are asked, and training of enumerators. It is important to be able to recognize bias and understand its implications. In research, bias occurs when the data collection methods disproportionately represent the perspectives of a one group more than another. Bias can also occur when the questions designed to collect the information you seek are not successful at learning the truth.

#### **Cell**

A cell refers an individual box within a spreadsheet. Data is entered into individual cells and each cell represents one piece of data.

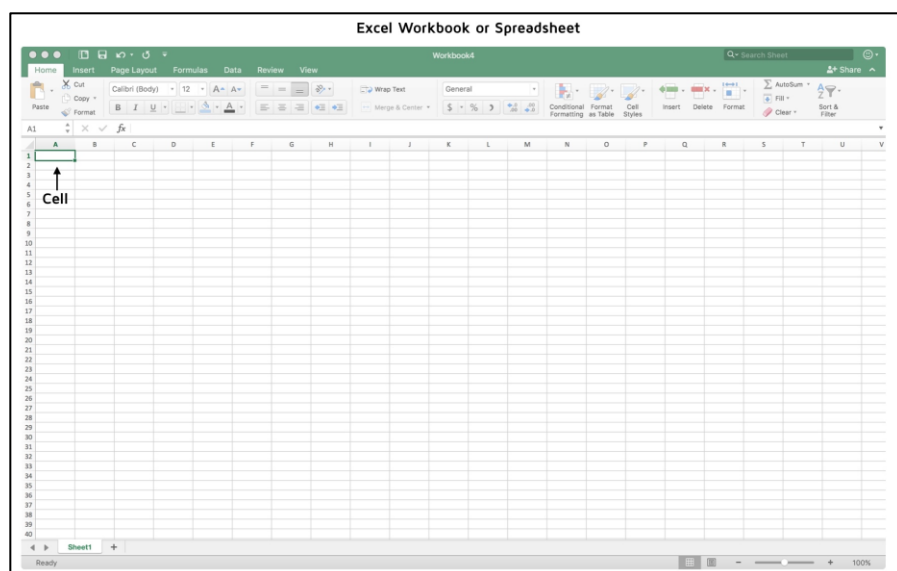


Figure 5: An Excel spreadsheet

### Cell Reference

Cells are typically referenced in the following format: column letter, row number. In the example below, cell B2 has been highlighted. (see image X)

### Column

A column refers to the vertical series of cells within a spreadsheet. Columns are typically referenced with letters, beginning with Column A on the far left and moving to the right. (see Figure 6)

### Enumerator

An enumerator is a person that is trained to administer surveys to respondents. During in-person surveys, they are responsible for recruiting respondents, administering the questionnaire, and filling in responses.

### Formula bar

The formula bar displays the text or formula contained within a cell. When a cell is selected, or highlighted, the data in the cell is displayed both in the cell itself and in the formula bar next to the function symbol (fx). To the left of the formula bar, the selected cell reference is displayed. In the example below, cell K22 has been selected. (see Figure 6)

### Header Row / Column Heading

A header row refers to the horizontal row within a spreadsheet where you will enter the type of information that will be placed into the cells below. Each column will contain *only one type* of information. The headers placed in this row correspond to the questions in the survey. (see Figure 6)

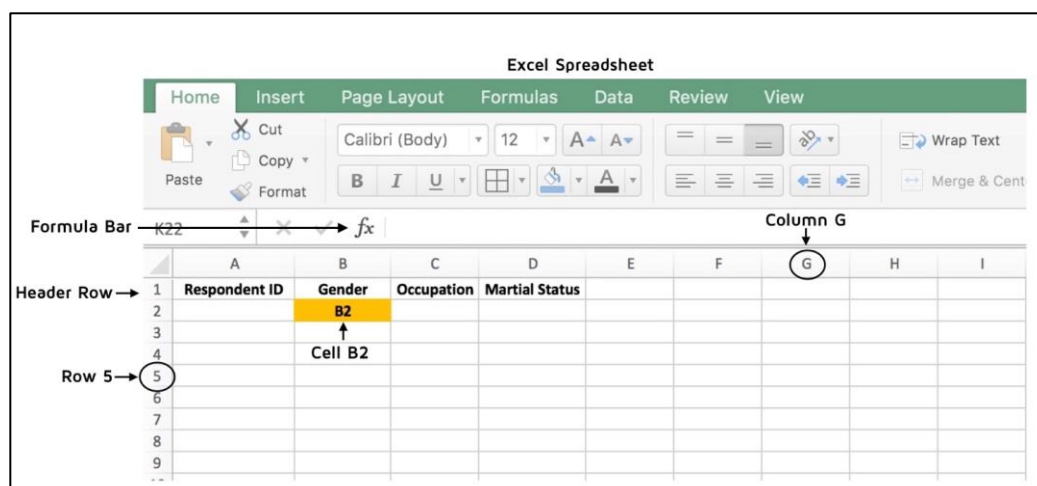


Figure 6: An Excel spreadsheet with labels for a row, column, header row, formula bar, and cell reference.

### Response item

A response item refers to one of the possible answers to a survey question. For example, if you ask people if they voted in the most previous election, “yes” and “no” and “choose not to disclose” would be three potential response items.

### Responsible data

The duty to ensure people’s rights to consent, privacy, security and ownership around the information processes of collection, analysis, storage, presentation and reuse of data, while respecting the values of transparency and openness.

\*Responsible Data Forum, working definition, September 2014

### Row

A row refers to the horizontal series of cells within a spreadsheet. Rows are typically referenced with numbers, beginning with Row 1 on the top left and moving down. (see Figure 6)

### Spreadsheet

A spreadsheet is a matrix divided into row and columns that is used to organise information. Spreadsheets are typically created using a computer-based programme, such as Excel or Google Sheets.

### Unique Identifier

A unique identifier is a number or code that is assigned to each survey and is written (or typed) onto the survey itself. They are used to protect the individual identity of the survey respondent. When the data is entered into the spreadsheet, the unique identifier is most often entered in the first column and labelled as “Respondent ID” or simply “ID Number.” Each row within a spreadsheet contains responses from *one* individual survey. This is done to ease data entry and to protect the respondent’s identity. It is also helpful in the case that you need to match data in the spreadsheet to its original source. For example, during data cleaning, if you discover an error or a potential mistake in data entry, it is possible to look at the respondent ID and search for the corresponding physical survey to review the respondent’s information.

**Community Health Survey**

ID #: 54

1. Occupation: \_\_\_\_\_

2. City: \_\_\_\_\_

3. Country: \_\_\_\_\_

4. Gender    ☐ M   ☐ F

5. Age        \_\_\_\_\_

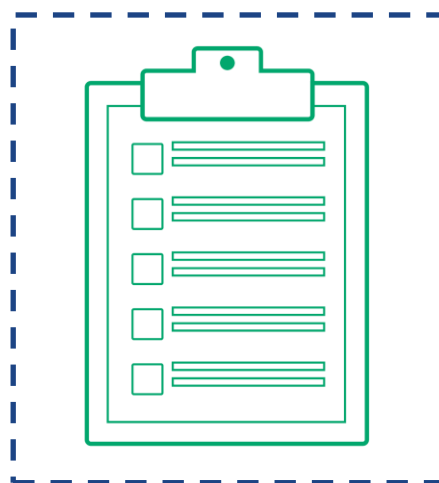
Figure 7: A sample survey with a respondent ID (unique identifier)

## ANNEX B

### ■ Survey methods checklist

#### Pre-survey design

- ✓ Identify target population
- ✓ Select survey method
  - In-Person
  - Phone
  - Online
  - Other



#### Question selection and design

- ✓ Include only necessary questions / eliminate non-useful questions – make the survey as short as possible
- ✓ Use simple, clear language / ensure vocabulary reflects literacy level of target population

- ✓ Use neutral language – do not imply positive or negative associations with specific answers
- ✓ Use structured questions as often as possible
- ✓ Multiple choice questions are comprehensive and all responses are mutually exclusive
- ✓ Review dichotomous questions for inclusion of a “neutral” option
- ✓ Avoid multiple responses questions unless the value has been clearly demonstrated and the analysis plan has been put in place
- ✓ Avoid double-barrelled questions
- ✓ Avoid leading or loaded questions
- ✓ Avoid open-ended questions unless used as opening questions or for exploratory research
- ✓ Avoid ambiguous words (such as “often” or “regularly”)
- ✓ Be specific when asking about dates (frequency, date, number of days/years)

### Survey design and preparation

- ✓ Ensure a logical question order – begin with general questions followed by specific questions
- ✓ Place sensitive / difficult questions at the end
- ✓ Include informed consent
- ✓ For long surveys, divide the questions into logical sections
- ✓ Be mindful of how the user will perceive the survey (too long, too complicated, etc.)
- ✓ Ensure instructions are easy to find and understand
- ✓ Translate survey tools using professional translators
- ✓ Sequentially number the questions
- ✓ Create post-data collection plan



Develop coding scheme  
 Design spreadsheet for data entry  
 Develop analysis plan / identify how each question will contribute towards research objectives

### Pilot testing

- ✓ Perform cognitive interviewing
  - Determine options for close-ended questions
- ✓ Pilot test the survey with at least ten individuals from target population in the delivery method selected for the survey
  - Ensure all questions align with research objectives
  - Ensure survey length is appropriate (i.e. not too long)
  - Add any missing questions
  - Alter any confusing questions
  - Review for unexpected responses
  - Add any missing options for multiple choice questions
- ✓ Pilot all translations (if applicable)
- ✓ Redraft the survey based on feedback from pilot testing
- ✓ Code and analyse the pilot responses to test analysis plan

### Data entry and coding

- ✓ Label all column headers
- ✓ Ensure only one variable appears in each cell
- ✓ Do not merge cells
- ✓ Divide up individual pieces of data into a separate column (e.g. if collecting names, place the first name in one column and last name in a separate column)
- ✓ Set drop-down menus for close-ended questions
- ✓ Enter a separate column for each response item for multiple response questions
- ✓ Keep a key or codebook for all coded or shortened data
- ✓ Manage permissions to the spreadsheet – encrypt and password protect sensitive data

## Data cleaning

- ✓ Check for outliers using data validation
- ✓ Check for missing data
- ✓ Check for duplicate entries in the column containing unique identifier
- ✓ Triangulate any suspicious entries (e.g. if a respondent's age is 101)
- ✓ Ensure logical layout of data
- ✓ Standardise font, font size, borders, column widths, etc.
- ✓ Remove any merged cells

## Data analysis

- ✓ Do not perform data analysis on original dataset
- ✓ Always keep track of the sample size; be mindful of missing responses either due to nonresponse, selection criteria, or other reasons
- ✓ Put a plan for data destruction in place

## Data communication and visualisation

- ✓ Acknowledge potential reasons that result may not be causal
  - Missing data
  - Low response rates
  - Confounders
- ✓ Acknowledge the limitation of the study design (survey, sampling, analysis, etc.)
- ✓ Describe the research question and lay out what the survey was attempting to answer
- ✓ Visual representations of the data are not misleading
- ✓ Graphs and charts are clean, simple, labelled, and easy to understand

## Data destruction

- ✓ Destroy data according to data management plan

---

<sup>i</sup> <http://www.telegraph.co.uk/sport/olympics/8992490/London-2012-Olympics-lucky-few-to-get-100m-final-tickets-after-synchronised-swimming-was-overbooked-by-10000.html>