Discovery of data analysis on Excel

This training module is provided under the terms of the Creative Commons License: Attribution-NonCommercial-ShareAlike 4.0 International
This training module was developed by CartONG. Created in 2006, CartONG is a French H2H/Support NGO specialised in Information Management. Our goal is to put data at the service of humanitarian, development and social action projects.

We are dedicated to improving the quality and accountability of field activities, in particular through better needs assessments and monitoring and evaluation.

We act as a multidisciplinary resource and expertise centre, accompanying our partners’ strategies and operations.

Our staff and volunteers also support the community as a whole by producing documentation, building capacities and raising awareness on the technical, strategic and ethical challenges of digital technologies.
This training module was developed by CartONG, as part of the project “Strengthening Information Management within francophone CSOs”. This project, led by CartONG, is co-financed by the French Development Agency (AFD) over the 2019-2022 period.

For more information, please refer to the article published on our website or contact us by email.

This training module is supported by the French Development Agency (Agence Française de Développement - AFD). Nevertheless, the ideas and opinions presented in this toolbox do not necessarily represent those of AFD.
TABLE DES MATIÈRES

• Creating a database in Excel

• Data cleaning in Excel
  → Basic formulas and conditional formatting

• Data analysis in Excel
  → Count, minimum, maximum
  → Central tendency: mean, median
  → Variability: standard deviation, variance

• Data visualisation

• Useful resources
Creating a database (DB) in Excel
Creating a DB - Definition

What is a database?

**Database:** “an organised collection of structured quantitative or qualitative data, or information, stored so that it can be easily accessed, managed, filtered, manipulated, and/or updated (usually electronically)” (ACAPS, 2013, P.3)

- Usually, the database is the **Excel file** that teams use to follow their **program data**, but it can take many forms (from paper to an Excel sheet, or more complex purpose-built software).
Creating a DB - Definition

Why do we need a database?

- To foster greater **access** to data
  - Large volumes of data can be stored in a **single location**.

- To quickly and easily **locate** data.
  - **Search** feature (filters, conditional formatting...)
  - Data **sorting**
If you are not familiar with Excel, do not hesitate to consult the following sections: “2.1 Discovering Excel features” and “2.2 Creating your spreadsheet” in the Excel Toolbox developed by CartONG.

You will discover:

• The Excel Interface
• The different types of data
• How to manage spreadsheet columns and rows
Creating a DB - 3 steps

- 3 steps to designing a database:
  1. Define the rows
  2. Define the columns
  3. Design the columns

Set the unit for each record

Define the data field/variables

Set the data values to constrain/facilitate data entry
1. Define the rows
Each row corresponds to a record/an observation.

Each record constitutes the basic unit of the report, for example an individual, a household, an infrastructure, etc.

This will likely correspond to a set of data collection responses.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td>2</td>
<td>001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Set the unit for each record

Examples: Household, individual, community, ...
2. Define the columns
A column corresponds to a variable/discrete information unit. In the cell at the top of each column, specify the variable name.

Example: below, the variable name is name_head hh.

Be concise

Define the data field
Example: Q1 is the name of the head of household.
### Creating a DB - Step 3

#### 3. Design the columns

**Unique and short column headers** - reference to the column's contents and to the questionnaire

**Additional column headings** - can be used to connect answers to the questionnaire, for example to section headings

**Unique ID in the first column** - unique identifier for each row in the database

**Stratification data** at the beginning - the main factors that will be used to stratify the analysis (comparison elements defined according to the needs mentioned in the analysis plan).

*Examples: displaced/host, urban/rural, male/female, ...*

**Geographic information** - accurately recorded to ensure data is accurate for mapping and comparison

**Other questions** - in the same order as in the data collection tool to allow for easier navigation

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ID</td>
<td>status</td>
<td>sex</td>
<td>camp</td>
<td>lat</td>
<td>long</td>
<td>Q1</td>
<td>Q2.a</td>
<td>Q2.b</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>idp</td>
<td>male</td>
<td>A</td>
<td>3,201</td>
<td>48,55</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>host</td>
<td>male</td>
<td>B</td>
<td>3,5</td>
<td>49,24</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>idp</td>
<td>female</td>
<td>B</td>
<td>3,499</td>
<td>49,25</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Here is an example of a database in Excel:
You may want **to progressively add data to your database**. You can enter the data directly into the Excel sheet. To do this:

- **Try not to modify/move/delete the database structure**
  - This prevents reference breaks/errors in the database, especially for a PivotTable and other advanced analyses/formulas.

- **Protect your workbook**
  - This helps avoid mistakenly deleting sheets or unwanted rearrangements.
Check the size of the database

This DB has 367 rows and 244 columns.

Average: 3786770.11  Count: 6903  Sum: 23326503875
To navigate the database more easily, **filter** your data:
• When you use formulas in Excel, you refer to cells.
• An **absolute reference** is a reference that will not be modified when you copy the formula.
  → To do this, add a $ sign before the column and/or row reference.

**Without absolute reference**

**With absolute reference**

Extend the formula
Naming your cells is an effective means of **saving time on formula entry and readability**.

*In this example, the range of cells from B2 to B5 is named “price_cfa”.*

To do this:
1. **Select** the range of cells to rename
2. **Enter** the desired name in this **range**.
Data cleaning in Excel
• **Database cleaning** is predominantly a logical process that notably entails **data consistency analyses** and **triangulation** with other sources of available information.

• Cleaning your data will allow you to always have an **up-to-date database that can be analysed without the risk of being misleading to those meant to use the information provided by the analysis.**
Data cleaning in Excel - Concept

As such, we will seek to identify:

- Missing values
- Empty cells filled/vice versa
- Typos
- 0 and N/A
- Errors
- Percentages >100
- Calculations
- Export (special characters)
- Duplicates
- Units
- Outliers
- Texts in numbers and vice versa
- Duplicate
- Unnecessary
- Outliers
Data cleaning in Excel - Format

- One common error is for numbers to be saved as text. Select the entire data range concerned → right-click → “Convert to number”
It is common to find a dot (instead of a comma) between numbers - especially in files shared between Anglophones and Francophones. Use the “Find and Replace” function to change this:
The **TRIM()** formula removes all spaces from a text string except for simple spaces between words. **This function is used to clean text with irregular spacing.**

**Example:**

In B2, we have text that was entered with extra spaces at the beginning. **The TRIM()** formula enables you to clean this up by eliminating all these extra spaces as can be seen in the result displayed in cell C2.
• An outlier is an extreme value, aberrantly different from the spread of a variable.

• In other words, the value of this observation greatly differs from the other values of the same variable.

• Just as with the search for duplicates, the detection of these extreme values (or outliers) is an essential step in data cleaning, since outliers can excessively influence upward or downward certain statistics produced - particularly those relating to means - and thus lead to erroneous analyses.
Data cleaning in Excel - Identifying duplicates or outliers

• This is hence one of the first things to check in a database (because it is the first source of error, especially if the survey format is poorly constrained).

• You can use **conditional formatting**, located under the **Home** tab.
Data cleaning in Excel - Identifying duplicates or outliers

- Using the various features:

You can quickly find outliers with Greater than... and Less than...

You can highlight duplicates directly.
Data cleaning - Identifying duplicates or outliers

You can also apply **conditional formatting** using formulas.

You can format according to the value of another cell.
For more data cleaning features, check out the Excel toolkit, developed by CartONG.

Data are unorganised raw facts or figures that need to be processed and analysed. Variables are a type of data that can change. They form the basis of most analyses performed to understand situations, trends and linkages.

Data and variables can take different forms: Simple and random in appearance, or statistical and complex in appearance. In any case, data and variables form the basis of the analysis but are of no use until they are processed, analysed and finally converted into information. Before being analysed, the data must first be checked for possible errors. Thus, database cleaning is primarily a logical process, including data consistency analysis and triangulation with other available information.

Some errors are difficult to detect before the analysis begins; for example, some outliers are identifiable only when the data is better known. However, it is preferable to detect as many errors as possible in order to avoid having to backtrack when analysing the data.

Make sure that all changes made to your dataset have been recorded in a “change log”.

This module consists of 7 sub-parts:
Data analysis in Excel
• **Data analysis** refers to the process of reviewing and interpreting data so as to develop responses to questions.

• For a variable (i.e. a column in the Excel database), it is thus possible to calculate:
  → The **frequency** for the different response modalities
  → The **minimum**, the **maximum**
  → The **mean**, the **median**
  → The **standard deviation**, the **variance**
  → The **sum**, under certain conditions
Data analysis in Excel - 3 types of descriptive statistics

• There are 3 main types of descriptive statistics:

→ **Distribution (or spread)** refers to the relative **frequency** of each value.
→ The **central tendency** concerns **average values**.
→ The **variability** or dispersion describes how **values are distributed** (concentrated or scattered).

• You can apply these operations to assess a variable (univariate analysis) or compare 2 or more in bi- and multivariate analyses.

⚠️ Bi/multivariate analyses are more complex than they at first appear, establishing causality being mathematically demanding.
Count, minimum, maximum
Objective: Calculating the frequency of a data item within a statistical series, i.e., the number of times a response modality appears for a given variable.

The COUNTIF function is used to count the number of cells that meet a criterion:

\[ \text{COUNTIF}(\text{range}; \text{criterion}) \]

- It is used to count the number of cells that meet a criterion.

  - **Range:** The range of cells on which to apply the count on if the criterion is met.
  - **Criterion:** Condition to be checked (<, >, =, <>, ...)
Example: Use the COUNTIF function to calculate the number of respondents over the age of 30.
Missing data is automatically excluded from analysis with the COUNTIF function.

- You can count the number of filled cells (other than 0) in each column using `=COUNT`. You will then know how many cells are empty (subtracting this result from the total number of observations) in the database.

- It is crucial to keep this in mind, as the exclusion of missing values reduces the representativeness of the sample and can therefore lead to biased inferences about the entire population.
Objective: Gives the largest/smallest value in a data range.

This function is not necessarily useful as such, but can be combined with others (for example, conditionalities) for dynamic analyses.

MAX(number1, [number2], ...)
MIN(number1, [number2], ...)

- **MAX**: Gives the largest value in a data range.
- **MIN**: Gives the smallest value in a data range.

✓ **Number1, number2, ...** the numbers 1 to 255 for which you want to find the maximum or minimum value.
Central tendency: mean, median
Central tendency - MEAN

Objective: Calculates the **average** of a range of data, i.e., the **sum of all values divided by the number of values**.

→ This is the most widely used central tendency measure.

**Example: What is the average age of respondents?**

<table>
<thead>
<tr>
<th>Respondent</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondent 1</td>
<td>35</td>
</tr>
<tr>
<td>Respondent 2</td>
<td>25</td>
</tr>
<tr>
<td>Respondent 3</td>
<td>20</td>
</tr>
<tr>
<td>Respondent 4</td>
<td>28</td>
</tr>
</tbody>
</table>

\[
\text{Average} = \frac{35 + 25 + 20 + 28}{4} = 27 \text{ years}
\]
Central tendency - MEAN

Objective: Calculates the average of a range of data, i.e., the sum of all values divided by the number of values.

→ This is the most widely used central tendency measure.

In Excel, use AVERAGE(...)
Always keep in mind that the mean does not reflect the shape of your distribution, so it is important to supplement this measure with others or you may have misleading representations.

Age of the 4 respondents in 2 different surveys

Survey 1

Respondent 1: 35
Respondent 2: 25
Respondent 3: 20
Respondent 4: 28

Survey 2

Respondent 1: 45
Respondent 2: 48
Respondent 3: 8
Respondent 4: 7

Average = 27
Moyenne = 27

Note to reader: In 2020, the average age of respondents was 27.
Source: CartONG test survey, 2019, 2021
Objective: Calculates the median of a range of data, i.e., the value that lies in the middle of the other values (middle value) when the series is ordered (usually in ascending order).

Example: What is the median of this data set?

<table>
<thead>
<tr>
<th>Respondent</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondent 1</td>
<td>20</td>
</tr>
<tr>
<td>Respondent 2</td>
<td>25</td>
</tr>
<tr>
<td>Respondent 3</td>
<td>28</td>
</tr>
<tr>
<td>Respondent 4</td>
<td>35</td>
</tr>
</tbody>
</table>

Here, there is an even number in the list, the median is therefore the average of the two middle values:

\[
\text{Median} = \frac{25 + 28}{2} = 26.5 \text{ years}
\]
Central tendency - **MEDIAN**

**Objective:** Calculates the median of a range of data, i.e., the value that lies in the middle of the other values (middle value) when the series is ordered (usually in ascending order).

In Excel, use **MEDIAN(...)**
Central tendency - Summary

- The mean is very sensitive to the existence of outliers.
- The median is much less sensitive to outliers.
Variability: standard deviation, variance
Measures of variability

• Measures of variability are essential because they provide a picture of the **distribution** of values in your statistical series.

• They are an essential complement to **central tendency** measures.

• Variability can also be referred to as **dispersion**.

Below are various measures of variability:

• **Range** (combination of **=max()-min()**): The difference between the lowest value and the highest value.

• **Standard deviation** (**=STDEV.S**): the distance to the mean.

• **Variance** (**=VAR.S**): the average of squared deviations from the mean.
Objectives: The range (or span) gives you an overview of the distance between the most extreme values in the series. To calculate it, simply take the difference between the largest and the smallest value.

In Excel, use: \( \text{MAX}(...) - \text{MIN}(...) \)
Objective: The **standard deviation** allows you to measure the dispersion of your series. This tells you, on average, how far each value lies from the mean in your sample. The higher the standard deviation, the more your data set is dispersed/variable.

For example, if the mean is 0, it is estimated that:

- 68.2% of the values are between: $[-1\sigma ; 1\sigma]$
- 95.4% of the values are between: $[-2\sigma ; 2\sigma]$

$\sigma = \text{standard deviation}$
Objective: The **standard deviation** allows you to measure the dispersion of your series, this tells you, on average, how far each value lies from the mean in your sample. The higher the standard deviation, the more your data set is dispersed/variable.

There **are six steps** for finding the **standard deviation**:  
1) Find the mean.  
2) Subtract the mean from each score/value in the series.  
3) Square each deviation from the mean.  
4) Add up all of the squares.  
5) Divide this sum by \( N-1 \).  
6) Find the square root of this number.
Variability - the Standard deviation - Examples

Example 1: Standard deviation of the age of respondents in 2019
In the table below, you see calculation steps 1 through 4.

<table>
<thead>
<tr>
<th>Data</th>
<th>Deviation from mean</th>
<th>Squared deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>35-27=8</td>
<td>8*8=64</td>
</tr>
<tr>
<td>25</td>
<td>25-27=-2</td>
<td>-2*-2=4</td>
</tr>
<tr>
<td>20</td>
<td>20-27=-7</td>
<td>-7*-7=49</td>
</tr>
<tr>
<td>28</td>
<td>28-27=1</td>
<td>1*1=1</td>
</tr>
<tr>
<td>Mean=27</td>
<td>Sum=0</td>
<td>Sum of squares=118</td>
</tr>
</tbody>
</table>

Step 5: 118/3=39.3
Step 6: √39.3=6.3

Standard deviation = 6.3 which means that on average, each age will “deviate from the mean” by 6.3 years.

Example 2: Standard deviation of the respondents in 2020

<table>
<thead>
<tr>
<th>Data</th>
<th>Deviation from mean</th>
<th>Squared deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>18-42=-24</td>
<td>24*24=576</td>
</tr>
<tr>
<td>65</td>
<td>65-42=23</td>
<td>23*23=529</td>
</tr>
<tr>
<td>27</td>
<td>27-42=-15</td>
<td>-15*-15=225</td>
</tr>
<tr>
<td>58</td>
<td>58-42=16</td>
<td>16*16=256</td>
</tr>
<tr>
<td>Mean=42</td>
<td>Sum=0</td>
<td>Sum of squares=1586</td>
</tr>
</tbody>
</table>

Step 5: 1586/3=396.4
Step 6: √396.4=22.9

Standard deviation = 22.9 which means that on average, each age will “deviate from the mean” by 22.9 years.
Variability - the Standard deviation - Why?

Average = 100
Variability - The standard deviation in Excel

Rest assured, in Excel, you only need a simple formula to calculate the standard deviation:

=STDEV.S
Objective: The variance refers to the average of squared deviations from the mean (that is, the squared standard deviation). The variance also reflects a degree of spread in the series: the higher the variance, the higher the overall “distance to the mean”.

- To find the variance, you simply have to have the standard deviation squared. The symbol of the variance is $s^2$.

Example - Age of respondents in 2019: $s=6.9$, $s^2=6.9*6.9=47.6$
In concrete terms, the dispersion is low.

Example - Age of respondents in 2020: $s=22.9$, $s^2=22.9*22.9=524$
In concrete terms, this means that the age of respondents is far more dispersed than in 2019.

- The greater the standard deviation, the more spread out the values in the series will be.
The formula used in Excel to calculate the variance is as follows:

=VAR.S
Data visualisation
Data visualisation affects **decision-making processes** by allowing more people to understand and **interpret** data more quickly. It provides insight into what information means through a clear **visual representation**. This helps the human brain to apprehend data more “naturally”, thereby identifying **patterns, trends, outliers**, etc.

Some data can be more complicated to represent than others; to help you through this, ask yourself:

- **What am I trying to say?**
- **Who am I trying to say it to?**

To learn more about the principles of data visualisation, refer to the training module on Data Management.
Data visualization - Keep it simple

In a nutshell, thanks to this short animation.

Before

After
How to represent **quantitative data**?

**Multiple statistical representations** are possible (graphs, chart, table, etc.). However, it is important to carefully select a chart that best applies to your data.

**Comparison**

- Simple multiple variables
- Simple over time: visualise a tendency
- Multiple (series and categories) and over time
Data visualisation

**Composition**

- Part of a whole
- Several parts of a whole

**Distribution**

- Single variable: Classes
  
  *Examples: number of children, age groups*

- Correlation: Multiple variables
Example of a column chart:

Characteristics of respondents by age category

<table>
<thead>
<tr>
<th>Age Category</th>
<th>2019</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 20</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20-40</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td>41-60</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

Note to reader: In 2020, 45% of survey respondents are in the 20-40 age group.
Source: Solidarité Internationale, KAP surveys 2019, 2020

- The main graphs used for frequencies are column charts, histograms and pie charts.
A stacked histogram is a variant that usually groups population subgroups/strata that make up a whole (which the entire bar represents).

Note to reader: In 2019, 90% of survey respondents were women. Source: Solidarités International, KAP surveys 2019
A few general principles for data visualisation:

**Stacked Pie Charts and Histograms:**
- Used for single-choice questions: the total must be 100%.
- Avoid representing modalities with less than 5%.
  → Group into an “Other” category.

For multiple-choice questions:
- Preferably use a column chart.
Data visualisation - in Excel

• Example in Excel - We want to represent the answers to this survey question:
  *Does your household receive food assistance (general in-kind food distribution and/or cash grants and/or food vouchers)*?

• Using a PivotTable, the following results were obtained:

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>Receive food assistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>40%</td>
</tr>
<tr>
<td>No</td>
<td>60%</td>
</tr>
<tr>
<td>Grand Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

→ For more information about PivotTables, refer to the Excel toolkit.
1) Select the PivotTable you want to represent.
2) From the top bar, select “Insert”

3) Then select the graphic representation of your choice (in this example, the *pie chart*).
4) Format the chart:
   • Add a title, data labels, and caption.
   • Change the colours for each category (right-click on a coloured part of the pie).
• Here's what you can get!
Useful resources
Resources developed by CartONG

- Excel Toolkit, developed by CartONG: https://cartong.pages.gitlab.cartong.org/learning-corner/en/3_nettoyage_page

- Data Analysis Toolkit, developed by CartONG: https://cartong.pages.gitlab.cartong.org/learning-corner/en/7_case_study_data_analysis_page

- Remove to Improve (the data-ink ratio) - Speaker Deck https://speakerdeck.com/player/87bb9f00ec1e01308020727faa1f9e72