ILLUMINATE 03

ANALYZING DATA: A SHORT INTRODUCTION TO WORKING WITH SPREADSHEETS

In this third note of the *Decoding Injustice* Illuminate module, we explore how to start giving meaning to the data we have collected. Here, activists and changemakers will find tools to start crunching the numbers that will build evidence for change.

Key Questions

How can spreadsheets help us analyse the data we have collected?

What are some of the basic features of spreadsheets and how are they used?

What are some of the key mathematical concepts to consider when analysing data?

CENTER for ECONOMIC and SOCIAL RIGHTS

ILLUMINATE 03

ANALYZING DATA: A SHORT INTRODUCTION TO WORKING WITH SPREADSHEETS

Introduction

There's a difference between data and evidence. To become evidence, data needs to be analyzed and interpreted. For our purposes, this means asking: What does the data tell us about the indicators we've identified for each of the dimensions of the OPERA Framework? Sometimes data can be read and interpreted quite easily. Other times, the dataset (the primary or secondary data that has been gathered for an analysis) is too large to be read easily in one or two pages. When this is the case, a spreadsheet can be a useful tool. This is particularly the case when working with budgetary data.

Working with spreadsheets can be a new skill for many activists. This note introduces basic skills for using spreadsheets. It also outlines some of the key mathematical concepts that can be used to analyze data in a spreadsheet. It is intended to be introductory, so feel free to skip the parts that you're already familiar with!

Feeling confident in the ability to work with numbers is important. It can better equip us as activists to either analyze data ourselves, or to collaborate with economists, statisticians and others to undertake such an analysis. We encourage you to take the time to give yourself a refresher where needed.

Spreadsheet Basics

Entering data into a spreadsheet makes it possible to organize, rearrange and analyze it in more detail, in order to test hypotheses about how different pieces of data relate to one another. **A spreadsheet is essentially a table of cells arranged into rows and columns**. Columns are normally represented by letters, while rows are represented by numbers. A single cell can be referred to by its row and column; "C10", for instance.

There are a variety of spreadsheet programs and applications available, such as Microsoft Excel, Calc from OpenOffice, and Google Sheets. **For this note, we will use Microsoft Excel.**

IMPORTING SECONDARY DATA

Unless you're entering primary data into the spreadsheet itself, the second step after creating a spreadsheet is to import your data. This could be primary data you have recorded in another document, or secondary data from an external source such as the World Bank or a government ministry. If you are downloading secondary data from an online source, make sure that it is in the correct format. You will have the option to choose the format of the data you wish to download from the data source. One of the most common formats for downloading data is CSV, which stands for "comma separated values". This is a format that can be read easily by Excel.

To import data, select the "Data" tab at the top of the spreadsheet. This will then show you different options for importing data (e.g., "From Access", "From Web", "From Text"). If you select "From Text", this will open a dialog box, from which you will select the file you have previously downloaded.

This document is organized according to an innovative method for collecting, analyzing and presenting evidence around three steps:

INTERROGATE
Map the problem in depth using OPERA to identify indicators and benchmarks.

ILLUMINATE
Spotlight the underlying issues by collecting, analyzing and visualizing data.

INSPIRE
Take action to build power and hold decision-makers accountable.

HELPFUL TIP

Whenever you download a dataset, the first thing you should do is to make a copy of it. Any changes you make should be done in this copy. This means that you can go back and check the original data at any time. It is also good practice to note where you sourced your data from, and when and how it was retrieved.

If the "Text Import Wizard" dialog box pops up, you can leave the selections set to their defaults:

- Step 1 (data type): Delimited
- Step 2 (delimiters): Tab
- Step 3 (column data format): General

Some useful keyboard shortcuts include:

- Copy [Ctrl + c]: copies the selected cells into the clipboard
- Paste [Ctrl + v]: pastes the clipboard
- Cut [Ctrl + x]: copies the selected cells into the clipboard and removes them from their original position
- Undo [Ctrl + z]: undoes the last change you made
- Redo [Ctrl + y]: undoes an undo

ENTERING PRIMARY DATA

If you are entering your own data into a spreadsheet, it is crucial to do so in a consistent, standardized way. Data that is entered inconsistently is harder to search, count, sort and filter. As an example, here is demographic information that has been entered into a spreadsheet:

Name	Sex
Arun	M
Hyun	Male
Eleanor	F

The problem here is easy to spot: the sex of the individuals has been recorded in different ways. Entering the data in a standardized way avoids these errors. This means making choices about how the data can be represented consistently. Data that tends to be recorded inconsistently includes:

- **Dates and times:** 01 July 2005, July 1st 2005, 1/7/2005 or 20050701 are all ways of representing the same date.
- **Names:** For example, will you use "United Nations" or "UN"? Will you put a person's first and second names in different columns or in the same column?
- **Places:** How specific do you need to be when describing geographical data? Should you use a country's administrative geography (e.g., town, city, district), electoral geography (ward, constituency) or operational geography?

The key point is to consider the different ways of recording data, make a choice and then apply it consistently. It will save an enormous amount of time and frustration.

The second issue is how to "code" (or classify) the data. This is particularly important if you are recording qualitative data that you will subject to quantitative analysis. For example, say you want to record information about schools you have visited, classifying them as either small, medium or large. When you and your colleagues review the spreadsheet, you need to be confident that each time you see a school described as "small", it means the same thing. To ensure the data is coded consistently, you could design a set of coding rules to let everyone working on the data know that:

- Small = between 0 and 99 students
- Medium = between 100 and 499 students
- Large = between 500 and 999 students.

It is crucial to make sure that everyone entering the data follows the same rules each time.

LOCKING ROWS AND COLUMNS

Before you start sorting your data, it can be useful to "lock" the top row and/or the first column of your spreadsheet, particularly if the spreadsheet you are working with is large. This means you can always tell what each row and/or column refers to. For example, you may have downloaded data on government spending on education as a percentage of GDP for all countries for the years 1990-2014. This will include many rows and many columns. To keep the years and the country names visible, these rows and columns can be locked so that they do not disappear as you move around your spreadsheet.

SORTING DATA

The first thing to do when looking at a new dataset is to become familiar with it. This involves sorting the data so that it makes the most sense. Say, for example, you have a spreadsheet with data on GDP, healthcare expenditure and life expectancy for all countries, over a number of years. To familiarize yourself with the data, you might want to look at the data first in order by country, and then in order by date. To sort the whole spreadsheet, use the following steps:

1. Select the "Data" tab and then the "Sort" button. This will open an additional dialog box.
2. Select the column you would like to sort by in the drop-down menu; you can sort ascending or descending (or A-to-Z or Z-to-A, if alphabetical).
3. You can also choose multiple criteria to sort by (e.g., first by country, then by date), by clicking the "Add level" button.

FILTERING DATA

It is common to filter out the values you do not want to see in the dataset. In the dataset mentioned above, for example, you might only be interested in reviewing the most up to date values.

To filter the whole spreadsheet, use the following steps:

1. Under the "Data" tab, select the "Filter" button. You now should see triangles next to the column names in the first row.
2. Click on the triangle next to the column you are interested in. A dialog box should appear, from which you can select and deselect the options you want to include or exclude.

Analyzing Data With Spreadsheets

Once you have sorted and filtered your spreadsheet, it is time to begin your analysis. Analyzing data helps you extract answers to your questions. Say you have data on GDP, population, life expectancy and healthcare expenditure as a percentage of GDP, for all countries over a number of years. There are many questions concerning the right to health you could answer using this information. For example:

- How much does each country spend in total on healthcare (in USD)?
- How does a specific country compare to others?
- How much is spent per capita in each country (in USD)?
- How has spending in each country changed over time?

It is important to remember that the data in your spreadsheet will not always provide all the information you need to answer your research questions. For example, based on the data we have, we do not know the amount of money spent on healthcare in each country. However, we can calculate this figure by using data that we do have, namely GDP per country and health expenditure as a percentage of GDP. You can calculate new values such as these using spreadsheet **formulae**.

USING FORMULAE

Formulae are basic mathematical functions; the kind of things you would find on a simple calculator. By typing these formulae into a spreadsheet cell, you can make calculations in your spreadsheet. The most common symbols that you will use in these formulae include:

Symbol	Description
=	This is the first thing that should go in your formula cell, because it tells the spreadsheet that you are writing a formula.
+	Add
-	Subtract
*	Multiply
/	Divide

Basic mathematical rules about the order of functions also apply when working with spreadsheets. For example, the formula =3+5*2 will equal 13, **not** 16. If you want to change the order of function, you must include parentheses. Formulas inside parentheses will be calculated before any other formula. For example, if you want the formula above to result in 16, it should be entered as: =(3+5)*2

Addition means adding up values in a range of cells in a spreadsheet. A quick way to do this is to use the symbol **=SUM** (as opposed to linking each cell individually with a "+" symbol). For example, if you want to calculate the total GDP for a group of 10 countries, use the following steps:

Data on GDP is in column B. In the first empty row in column B, at the end of the list of countries' GDP in 2010, type =SUM and then an open bracket; =SUM(.

Then select the cell with the first country in the group. While you hold the SHIFT key down, select the last country in the group. The formula cell should automatically add the cell numbers.

Complete the formula by adding a close bracket at the end; for example, =SUM(B2:B11).This will add the GDP values in the given range.

Multiplication and division is useful if you want to convert a numerical value into a percentage, or vice versa. For example, say you want to find out how much (in USD) is spent on healthcare, in total, in your country. You know that its GDP is USD 100,357,000,000 and that government spending on health is 3.48% of GDP. You can answer your question by using the following formula:

GDP * (health spending/100)
Or in a spreadsheet =100357000000*(3.48/100)

You do not need to manually type in the numbers; clicking in the cell will do this for you. The formula looks something like: =C2*(D2/100), where cell C2 is your country's GDP and cell D2 is health expenditure as a percentage of GDP in your country.

Copying formulae is useful when you want to apply the same calculation across rows or across columns. To do this, simply copy the formula you have just written (using Ctrl + c) and paste it into the cell below (using Ctrl + v). Alternatively, click on the lower right corner of the cell (the blue square) and drag the formula down to the end of the column.

Empty cells can be problematic when you start to analyze your data, because they can end up producing zero values. Properly handling missing values is an important step in data cleaning and analysis. Large datasets are rarely complete, and so you should have a strategy to deal with missing parts. One way to find errors is to filter the data, select all "0" values and delete them.

Useful Mathematical Concepts For Analysing Data

Ordering data to help answer your research questions can be daunting if you are trying to compare multiple variables. For instance, if you are looking at data that relates to the right to food, you might want to look at: (a) per capita availability; (b) of major food items; (c) across municipalities. You might be interested in finding out the *average* availability for these items, the *minimum and maximum* available, and the difference, or the *variation*, between the minimum and maximum. Finding out the answers to these questions is commonly referred to as descriptive statistics.

RANGE

The first piece of information you might want to find from your data is the range; in other words, from where to where does your data stretch? Does it start with small numbers? Large numbers? Does it run from negative to positive? This is all essential information that will help you to deal with your data.

Looking at the range will also help you to find errors in your data. For example, say you are looking at data related to the indicator "Average years of schooling in the adult population". You might find that your data ranges from 4 to 58. There is clearly a mistake; the likelihood that any adult in any country has 58 years of schooling is very small. You should go back to your data and check it.

So how do you find your range? Simply go through your data and find the minimum and maximum values: the lowest and the highest, respectively. For example, say you have the following data on average years of schooling in the adult population for your own country, as well as for several neighboring countries:

8.5, 5.8, 6.5, 7.6, 10.2, 8.4, 7.3, 7.2, 9.2, 9.3

- Question: What is the range of your dataset?
- Answer: The lowest number (minimum) is 5.8 and the highest number (maximum) is 10.2. Thus, the range is from 5.8 to 10.2.

In a spreadsheet, you can do this by sorting the data from smallest to largest, or with the formulae **=MIN** and **=MAX**, using brackets to select the cells you want to include in the calculation; for example, =MIN(D12:D84).

COUNT

The next important piece of information you might want to determine is how many *things* do you have data for. How many countries? How many households? And so on.

How do you get this information from your data? Simply count it. In the dataset above, for example, there are 10 observations.

If the dataset is too large to count, you can use the formulae **=COUNT** when you have numbers in your cells or **=COUNTA** when you do not have numbers in your cells. Again, use brackets to select the cells you want to include in the calculation; for example =COUNTA(A5:A2089).

This might seem simple. However, when it comes to analysis and interpretation of the data, it is very important. For instance, if you are comparing your data on hospitals in your country, is data for 10 hospitals sufficient to make that comparison?

AVERAGES

The next piece of information you might want to look at is the **central value** and how the data is distributed in relation to that value. Does the central value give a good indication of the whole dataset, with an equal distribution of data points above and below it? Or is the distribution "skewed", meaning there is a peak at one end of the data range with a long tail towards the other? The distribution tells you what kind of further descriptors are practical to use. There are a number of different ways to answer these questions.

The mean (or "average") is the most common way of looking at the "central value". It will be familiar from reports; for example, average unemployment has increased in country X or average literacy rates have decreased for country Y.

So how is the mean calculated? The mean is the sum of all the values in the dataset divided by the number of values there are. For example, say you are interested in determining the average household income. You have the following data on average annual household income, measured in dollars, for a number of households: 1120, 241, 876, 201, 112, 345, 567, 156, 154, 1345

- Question: What is the mean average of your dataset?
- Answer: Add up the incomes you have for all households (1120 + 241 + 876 + 201 + 112 + 345 + 567 + 156 + 154 + 1345 = 5117). Then divide that number by the number of households you have (5117/10). Your answer is 511.7.

In a spreadsheet, you can calculate this using the formula **=AVERAGE.**

The mean can give you a good estimate of what is "normal" when the rest of your data is distributed evenly above and below it. However, if the rest of your data is "skewed" to either side, a different measure of the average may be more appropriate. For example, if you have a small number of extremely high earners in a population group, using the mean would make average per capita income appear higher than it actually is.

The median is the numerical value separating the higher half of values in the dataset from the lower half. It is useful when the rest of your data are not evenly distributed on either side of the mean. In the example of average household income above, the mean income value is 511.7. However, this is quite a large number in relation to most of the values in the dataset. It results because there are a few households with very high incomes, which skew the data. In this case, the median may provide a better estimate of what is "normal" than the mean.

So how is the median calculated? Firstly, sort the data (ascending or descending, it does not matter) and the value in the middle of the dataset is the median. If there is an even number of values in the dataset, take the average of the two middle values.

- Question: What is the median household income?
- Answer: First, sort the data: 112, 154, 156, 201, 241, 345, 567, 876, 1120, 1345. There are 10 values and the middle two values are 241 and 345. Find the average between these two numbers (241 + 345 / 2 = 293). The median is 293.

In a spreadsheet, you can calculate this using the formula **=MEDIAN.**

The mode is the value that appears most often in a set of data. Sometimes neither the mean nor median really tells us what we want to know. For example, if you would like to know the average number of children per household enrolled in school, you might have the following dataset:

0, 1, 1, 1, 1, 2, 2, 2, 3, 5

The mean number of children enrolled in school per household is 1.8, and the median is 1.5. But what you really want to find out is how many children are enrolled in school, in the majority of households. You can see that "1" child is the most frequent answer. This is the mode.

In a spreadsheet, you can calculate this using the formula **=MODE.**

In the case where more than one value is the most frequent, the dataset can be bimodal (two average values) or multimodal (more than two average values).

VARIATION

The next important piece of information you might want to identify is the size of the variation in the dataset. This is crucial when looking at aggregated and disaggregated data. For instance, say you are investigating the realization of the right to work. You want to find out the average unemployment rate. However, you might also want to test how representative the average is of different municipalities within your country. There are two common measures for doing this.

The standard deviation is a measure of by how much, on average, data values are off the mean. The following three steps show how to calculate it:

1. Sum the square of the differences between the values and the mean.
2. Divide that sum by the number of values minus one.
3. Take the square root.

In a spreadsheet, you can do this using the formula **=STDEV.**

Say you have the following data on unemployment rates as a percentage of the total active population for four different municipalities:

1, 2, 3, 4

- What is the standard deviation of your dataset?
- The mean is 2.5, so following steps 1 to 3:

Value	Difference to mean	Squared difference
1	-1.5	2.25
2	-0.5	0.25
3	0.5	0.25
4	1.5	2.25

1. Sum the square of the differences between the values and the mean = 5.
2. Divide that sum by the number of values minus one = 5/(4-1) = 5/3.
3. Take the square root = 1.291.

If the data is normally distributed (all data points are evenly distributed either side of the mean) then 68.27% of the data points will fall within one standard deviation from the mean and 95.45% of the data points will fall within two standard deviations from the mean. The bigger the standard deviation value, the more variation there is in the dataset.

Using the unemployment example, a standard deviation of 1.291 means that 68.27% of the data points will fall within this

distance from the mean (assuming a normal distribution). This suggests that there may be reasonable geographical variation in unemployment, so reporting the mean alone could be misleading.

The median absolute deviation is similar to the standard deviation, but is used with the median rather than the mean. The following two steps show how to calculate it:

1. Calculate the median and the absolute differences between each value and the median.
2. Calculate the median of the differences.

To take the same unemployment example as above, but with data on one additional municipality, your data is 1, 2, 3, 4, 5.

1. The median is 3 and the absolute differences are 2, 1, 0, 1, 2.
2. Order the differences 0, 1, 1, 2, 2 and the median absolute deviation is 1.

NORMALIZATION

Once you have an idea of the data you are dealing with, you might begin to make comparisons. For instance, in an investigation of the extent to which the right to health is being fulfilled, you might want to compare government spending on health in your country with that of another country that differs in a number of aspects. Comparing the total value of government spending on healthcare in this case will only tell you so much, if for example your country is very large and the country you are comparing it to is very small. The bigger country will most likely spend much more on health in total than the smaller country. Does this mean the bigger country is meeting its obligation to fulfil the right to health better than the smaller country? Not necessarily. To answer that question, the two countries have to be compared on an equal basis. This is usually done using an indicator that tells us how big a country is; often the size of its population. To compare government spending on health, you can divide total spending by the population. This is called **normalization**.

HELPFUL TIP

If you end up working with a very large dataset, there may come a point when you "outgrow" your spreadsheet. If this happens, consider using other database software such as SPSS, Stata or R. Unlike a spreadsheet, which is designed to be able to be "read" on the screen, the way data is stored in a database is often completely hidden from the user. This enables abstract, complex ways to store larger amounts of data and gives the user more flexibility in how to use it. That said, databases are a much more technical way to store and analyse data, and if you use them, you may need to work with statisticians, programmers and designers.

Applying These Concepts When Analyzing Budgetary Data

Analyzing a government's budget through a human rights lens involves some specific calculations. This is because the figures in a budget are always relative. Analyzing whether budgetary figures are high or low, for example, involves asking high or low *relative to what*? Taking a hypothetical example, let's say a country's social housing budget goes from 100 million USD in 2015 to 200 million USD in 2020. Doubling the social housing budget may seem quite significant. However, what if the government's overall budget tripled in that same period? As a percentage share, the social housing budget actually shrinks. What if, due to inflation, the cost of constructing social housing increases 150% over the same period? In this scenario, the government's purchasing power decreases, so the budget does not stretch as far. Because of this relativity, it is often necessary to convert budgetary figures.

CONVERTING NUMBERS INTO PERCENTAGES OR RATIOS

Using percentages is one way of making data comparable. For example, you might want to compare two countries' tax revenues. One way to do this is to normalize tax revenue as a percentage of both countries' GDP, which may show that the first country collects a relatively high amount of revenue, whereas the second country does not.

CALCULATING PER CAPITA ALLOCATIONS

To uncover discriminatory patterns of allocations, the budget must also be evaluated in terms of how it distributes benefits among households and individuals. To do this, you might need to calculate per capita spending; for example, by region or municipality. You can do this simply by dividing the total allocation to the region or municipality by the total population of that region (or by the total population group that you are interested in (such as school-aged children or women of reproductive age).

ADJUSTING FOR INFLATION

It is important to compare budget allocations over time and assess whether there has been an increase or decrease in the amounts allocated to different sectors. This is important because it tells us something about whether a government is taking action to "progressively" realize these rights.

HELPFUL TIP

Inflation calculators can help you to make these adjustments. There are several available online, you just need to enter the nominal amounts for each year and the Consumer Price Index (CPI) for each year and calculate. See: https://www.bls.gov/data/inflation_calculator.htm

However, this is not as simple as just looking at budget figures from different years. Budget figures are reported in "nominal" terms, which means they do not take inflation into account. For this reason, it is necessary to convert allocations from "nominal" into "real" amounts. This makes budget figures from different years "equivalent" to the current values of one of the years, and enable valid comparisons to be made over time. In other words, real value = nominal value adjusted for inflation.

FORMULA FOR ADJUSTING FOR INFLATION

$$\text{Real Value} = \frac{\text{Target year's nominal value} \times \text{base year's consumer price index (CPI)}}{\text{Target year's CPI}}$$

For example, 2010 money in 2000 values would be calculated as

$$\text{Real Value} = \frac{\text{2010 value} \times \text{2000 CPI}}{\text{2010 CPI}}$$

Interpretation And How To Avoid Common Misconceptions

While simplification is required to understand what the data means, as when you are presenting evidence in a graphical format, it is crucial to stay as close as possible to the "full story".

CORRELATION IS NOT CAUSATION

In general, it is extremely difficult to establish causality between two correlated observations. There are several reasons why common sense conclusions about cause and effect may be wrong. For example, you might want to report on the relationship between education and health. Using a scatter plot, you have average years of schooling on one axis and average life expectancy on the other. The scatter plot reveals that the two variables are highly positively correlated: higher average years of schooling are associated with higher average life expectancy. But can it be said that higher average years of schooling *causes* higher average life expectancy? No. There could be a number of reasons for this association:

- Higher average years of schooling may cause higher average life expectancy.
- Higher average life expectancy may cause higher average years of schooling.
- Higher average years of schooling and higher average life expectancy are consequences of a common cause, but do not cause each other. The common cause may be higher average income, for instance, or a more equitable distribution of income.
- There is no connection between average years of schooling and life expectancy. The correlation is coincidental.

PERCENTAGE CHANGE AND PERCENTAGE POINT CHANGE

Percentage change and percentage point change can sometimes be confused in the interpretation of data. For example, if a value changes from 5% to 10%, what is the percentage change? A common mistake would be to answer 5%. This is incorrect. The percentage change in question is actually 100%. It is, however, a change in five percentage points. The choice on whether to report percentage change or percentage point change will depend on your question. Often it is useful to report both.

CONCLUDING THOUGHTS

Building the confidence needed to work with numbers helps to equip us as activists either to analyze data ourselves or to collaborate with economists, statisticians and others to undertake such an analysis. Doing so can take time. It involves being experimental and exploratory, which means playing around with the data we have. For example, we might compare variables, disaggregate, look at it over time, etc. This helps us to see what is most relevant to our research.

That said, it is very important to keep a log of any changes made to the data. This is especially the case if different people are working on the dataset. This makes it possible to keep track of any errors that are made, which makes it easier to go back and correct them.

It's also important to remember that number-crunching isn't the only step in interpreting data. There's often still more to dig into, in order to answer why a situation is the way it is. Doing so is crucial in deciding what conclusions to draw from your research. This topic will be addressed in the fifth note in this module.