Summarizing data

When we do a research, we will usually a big dataset collected from hundreds or may be thousands of research subjects. Data of each subject is called an observation. For each observation we will have many variables. Understanding this amount of data will be difficult unless we summarize them. We cannot only list each observation with the respective variables and their values because:

1. This needs plenty of space and;
2. This will be meaningless and difficult to interpret

Therefore we have to shorten or summarize the data in some way in order to

1. Present the data briefly and in a small space;
2. Make the data more meaningful and easy to understand and;
3. See how values of each variable are distributed

1) Summarizing categorical data

The following imaginary data is collected from an outbreak of gastroenteritis in a village. In this data we have 30 observations and for each observation we have 5 variables i.e. sex, water source, education level, presence of diarrhea and presence of vomiting. ID is the identification number for the observations. Note that all the variables are categorical. Can you identify type of each categorical variable in the table?

Table 1. Data from an imaginary outbreak of gastroenteritis in a village

ID	Sex	Water source	Education	Diarrhea	Vomiting
1	M	Piped	None	Yes	No
2	M	Well	Primary	Yes	No
3	M	Public tap	Primary	Yes	No
4	F	Piped	None	Yes	No
5	F	Well	Secondary	No	Yes
6	M	Public tap	None	Yes	Yes
7	F	Piped	Primary	No	No
8	M	Well	Primary	No	No
9	F	Public tap	Secondary	Yes	No
10	M	Piped	Primary	Yes	No
11	M	Public tap	Primary	No	No
12	M	Well	University	Yes	No
13	M	Public tap	None	Yes	Yes
14	M	Public tap	None	Yes	No
15	F	Public tap	Secondary	Yes	No
16	M	Piped	Primary	No	Yes
17	F	Public tap	Primary	No	Yes
18	F	Public tap	Primary	Yes	No
19	M	Piped	Primary	Yes	No
20	M	Piped	None	Yes	No
21	F	Public tap	Primary	Yes	No
22	F	Public tap	University	No	No
23	M	Public tap	Secondary	No	No
24	F	Public tap	Primary	Yes	Yes
25	M	Piped	Primary	Yes	No
26	F	Well	None	Yes	Yes
27	M	Well	Secondary	Yes	Yes
28	M	Public tap	Primary	No	Yes
29	M	Public tap	None	Yes	No
30	M	Piped	None	Yes	No

This table displays all the information we have collected from the research subjects. Although it contains all the information, yet it is doesn't make a lot of sense in terms of understanding the data and individual variables. This is the raw data. In order to understand it better we have to manipulate it, to change it to a more meaningful shape. To make categorical data more meaningful we summarize it using the following methods.

1) Frequency distribution

Frequency simply is the count of the values of each variable in all observations. In other words how the values of each variable are distributed across the sample. To calculate frequency of a variable, we just count the observations which have the same value of that variable. For example if frequency distribution of sex tells us how many males are there in the sample and how many females are there in the sample.

We can summarize each variable of table 1 into a frequency distribution of that variable. Table 2 shows frequency distribution of sex and water source. To draw a frequency distribution for sex, we count how many times male is repeated and how many times female is repeated in all observations and report it as in table 2. Sex is a binary variable with 2 values, male and female. Male is repeated in 19 observations and female is repeated in 11 observations. A frequency table tells us how many males and how many females are there in the sample. Water source is a nominal variable with 3 values, piped water, public tap and well. There are 9 observations with piped water, 15 with public tap and 6 with well.

Table 2. Frequency distribution of sex and water source

Variable	Frequency
Sex	
Male	19
Female	11
Water source	
Piped	9
Public tap	15
Well	6

With this presentation of the categorical data we get a clearer picture of the distribution of sex and water source in the sample.

2) Relative frequency

Table 2 shows frequency of the values in absolute numbers. We can calculate frequency of the values as a percentage of all observations. This is called relative frequency which means frequency of each value (category) of a variable in relation to the number of observations as a percentage. We calculate relative frequency by dividing the frequency of the particular value (category) by total frequency i.e. frequency of all categories of that variable.

$$\text{Relative frequency} = \frac{\text{frequency in category}}{\text{Total frequency}} \times 100$$

Table 3 shows relative frequency of sex and water source. In the first column we have the variables and their value categories, in the second column we have frequency and in the third column we have relative frequency.

Table3. Relative frequency for sex and water source

Variable	Frequency	Relative frequency
Sex		
Male	19	63%
Female	11	37%
Water source		
Piped	9	30%
Public tap	15	50%
Well	6	20%
Total observations	30	100%

We can see from this table that:

- 63% of the cases were males and 37% of them were females;
- 50% of the cases used public tap, 30% used piped water and 20% used wells.

Can you tell how these percentages are calculated?

Table 4 shows education level of mother of 925 children less than 6 years of age who visited the burns centre in Sulaimani for a new burn injury in 2008.

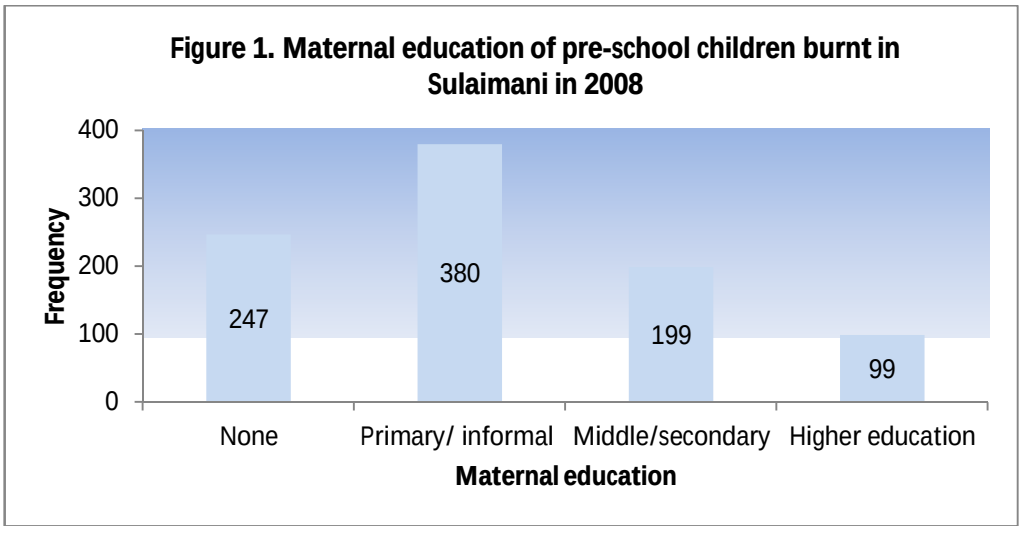
Table 4. Maternal education of pre-school children burnt in Sulaimani in 2008

Mother's education	Frequency	Relative frequency (percent)
None (unable to read and write)	247	26.7
Primary/ informal	380	41.1
Middle/secondary	199	21.5
Higher education	99	10.7
Total	925	100

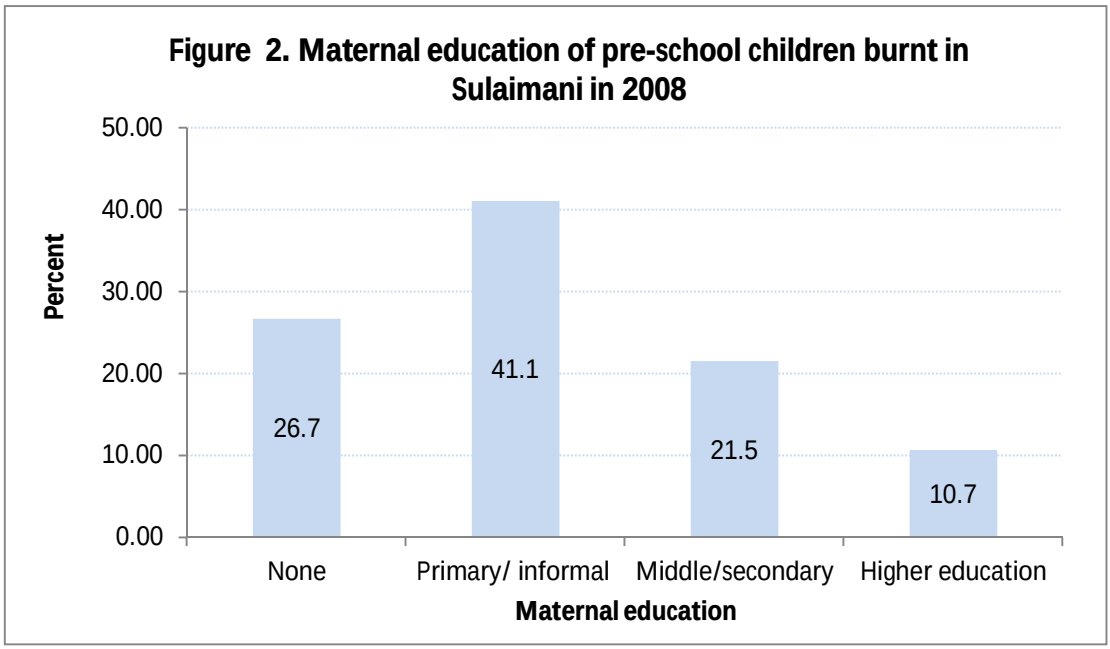
We see that the biggest percentage of mothers (41%) have primacy education and only 10.7% have higher education. What else can you tell from this table?

3) Bar charts

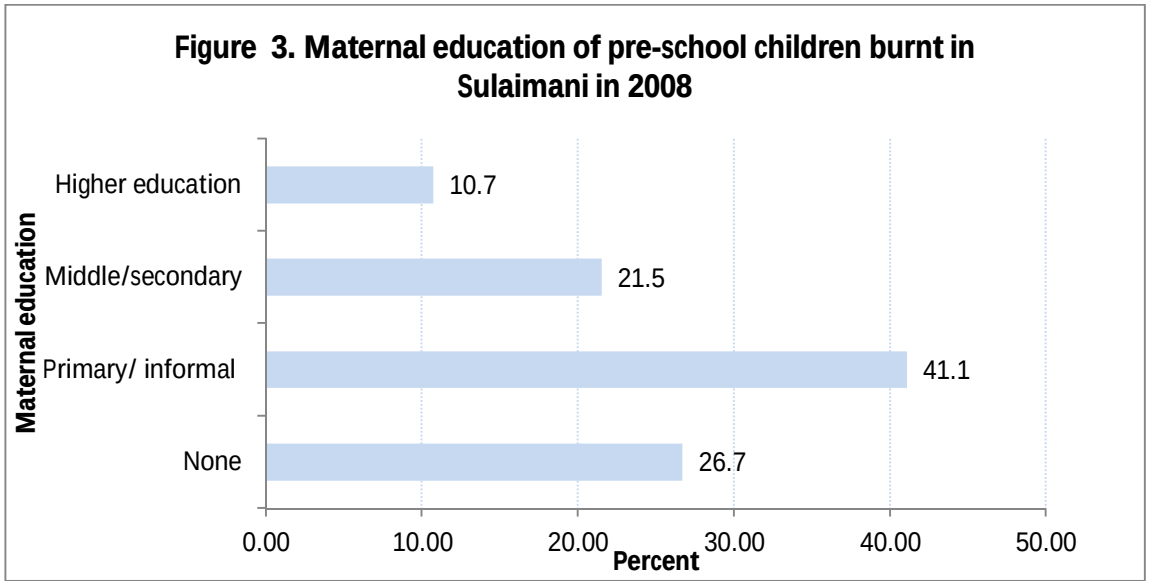
Frequencies and relative frequencies can be presented graphically in bar charts. In a bar chart each category is drawn as a separate bar and the length of the bars is proportional to the size of the frequencies or relative frequencies. Figure 1 below shows a bar diagram of the data in table 4.



This chart shows frequency (number) of each category. We could draw the bar chart with the relative frequency (percent) of each category as in figure 2.



A distribution presented as a bar chart gives a more immediate visual message than a frequency table. You can instantly see which category is more and which is less. Bar charts can be drawn horizontally. Figure 3 is the same bar chart as figure 2 but in a horizontal direction.



Remember the following key points about bar charts:

1. They are used to display qualitative (or discrete quantitative) data.

2. One bar represents one category of the variable
3. The height of the bar equals frequency (or relative frequency) of the category.
4. The bars are of the same width.
5. Bars should have a space between them and the space should be equal.
6. The position of each category is arbitrary if the variable is unordered.
7. The vertical axis of a bar chart starts at zero.

4) Clustered bar chart

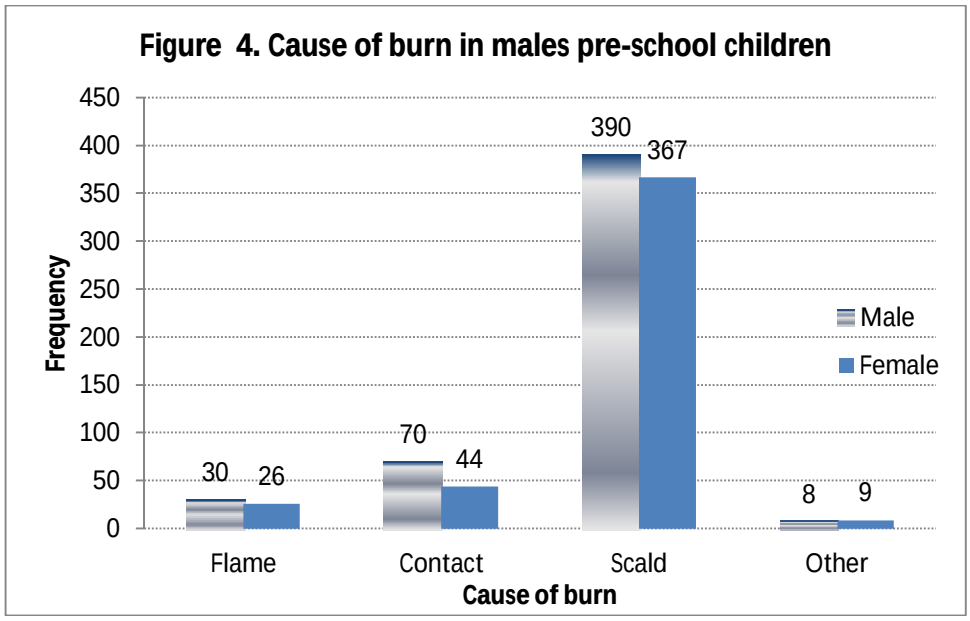
When we have two groups of observations, e.g. males and females, and we want to draw a bar chart for one variable, we use clustered bar charts to display the data in one chart. This clustered chart allows easy comparison between the two groups. The following data shows cause of burn in pre-school children of Sulaimani in 2008.

Table 5. Cause of burn in males pre-school children

Causes of burn	Frequency in males	Frequency in females
Flame	30	26
Contact	70	44
Scald	390	367
Other	8	9
Total	498	446

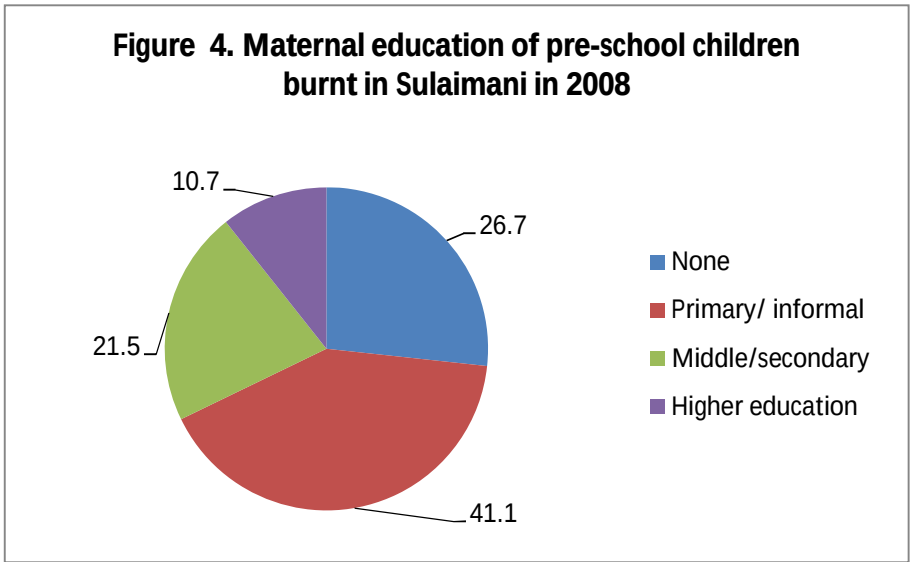
Can you explain this table?

The same data can be drawn in a bar chart. The bar chart below shows frequency distribution. Can you draw a bar chart for the relative frequency of the same data in table 5?



5) Pie charts

In a *pie chart* all values of the variable is represented as a circular area which is then divided into portions each corresponding to one category of the variable. Each portion is similar to one bar in a bar chart. Each portion could either represent the frequency of the category or the relative frequency of the category. The area occupied by the portion should be proportional to the relative frequency of the category. The pie chart in figure 4 displays data in table 4 as relative frequency. Read the chart and compare it with figure 2.



Practical Note: Bar charts and pie charts could be drawn manually but it is better to draw them using computer programmes. All the above figures are created by Microsoft Excel from the data tables.