This article is Chapter I from the author's book Statistics and Probability Flashcards.

Sommaire

Definitions

Individuals and variables

In a dataset, the individuals are the items with one or more properties, called variables. Individuals can be events, cases, objects, people, etc.

Student (individuals)	Height (variables)
John	190 cm
Ali	175 cm
Paul	165 cm
Clara	160 cm

Table 1.1. example of a data set with items and variables.

Individuals and variables are called data. Table 1.1 is called a data table.

Here's another example of a data table containing other variables:

Student	Height	Weight	Likes football
John	190 cm	100 kg	Yes
Ali	175 cm	90 kg	No
Paul	165 cm	60 kg	No

Clara	160 cm	63 kg	Yes

Table 1.2. example of a data set with items and more than 1 variable category.

Variables can be categorical or quantitative. In table 1.1 there's one quantitative variable: the height whereas in table 1.2 there are two quantitative variables (height and weight), and one categorical variable (likes football).

Quantitative variables are numerical variables: counts, percents, or numbers.

Categorical variables are non-numerical variables. Their values aren't represented with numbers: words, not numbers.

This data set presented in table 1.1 and table 1.2 is called one-way data because we have just a single individual (item) that has one or many properties attached to it.

How to build a data table?

When you build a data table, it is important to think about whether you have more individuals or more variables.

In tables 1.1 and 1.2 the number of individuals listed was greater than the number of variables. If we have many variables but only a few individuals, it is advisable to list the individuals across the top and the variables down the left side.

	John	Ali
Height	190 cm	175 cm
Weight	90 kg	75 kg
Likes football	Yes	No
Likes pizza	Yes	Yes

Table 1.3. Since the number of variables is bigger than individuals, listing the variables vertically would make the data table more appropriate than if we had tried to list all the variables horizontally.

Data visualization

Bar graphs and pie charts

Two of the simplest ways to summarize and graphically represent data are bar graphs and pie charts.

Bar graphs apply a series of rectangular bars to show absolute values or proportions for each of the data categories whereas pie charts show how substantial each data category represents as a part or proportion of the whole, by using a circular format with different-sized "slices" for different percentages of the total.

Rank	Country	Oil production (bbl/day)
01	USA	15,043,000
02	Saudi Arabia (OPEC)	12,000,000
03	Russia	10,800,000
04	Iraq (OPEC)	4,451,516
05	Iran (OPEC)	3,990,956
06	China	3,980,650
07	Canada	3,662,694
08	United Arab Emirates (OPEC)	3,106,077
09	Kuwait (OPEC)	2,923,825
10	Brazil	2,515,459

Table 1.4. Top 10 world Oil producers ("Production of Crude Oil including Lease Condensate 2019" U.S. Energy Information Administration)

Figure 1. Bar chart – Top 10 world Oil producers ("Production of Crude Oil including Lease Condensate 2019" U.S. Energy Information Administration)

Notice that we have a list of the Oil producers (countries) across the bottom of the bar graph, with the count of the Oil production (bbl/day) up the left side.

The countries are the individuals, and the count is a quantitative variable because it represents the numeric property of each of the individuals. The bar graph is one of the best ways to represent this data because it is possible to get quickly an overview of which countries produce the most oil.

Figure 2. Pie chart – Top 10 world Oil producers ("Production of Crude Oil including Lease Condensate 2019" U.S. Energy Information Administration)

Now we can quickly see that the United States produces the most of the total oil daily, biggest than any other country, Saudi Arabia occupies second place, and Brazil is the 10th world's biggest oil producer.

Venn diagrams

A Venn diagram is a diagram that shows all possible logical relations between a finite collection of different sets from a two-way table.

	Good	Cheap	Fast	Total
Expensive	10	0	10	20
Low quality	0	10	10	20
Slow delivery	10	10	0	20
Best choice	10	10	10	30
Other	20	20	20	60
Total	50	50	50	150

Table 1.5. two-way data table

Figure 3. Venn diagram

Box-and-whisker plots

Box-and-whisker plots (also called box plots) are a great method for graphically depicting groups of numerical data through their quartiles. It is very useful when you want to show the median and spread of the data (see chapter IV) at the same time.

Assuming that we have the following data set: [1, 2,2, 2, 3, 3, 4, 6, 8,8, 10, 11, 11, 16]:

Figure 4. Box-and-whisker chart

The horizontal line in the center of the box is the median of the data set, so the median of the data set represented in the chart above is 5.

The dot at the end of the bottom whisker is the minimum of the data set, and the dot at the top of the right whisker is the maximum of the data set. So in this plot, we can say that the minimum is 1, that the maximum is 16, so the range would be $16 - 1 = 15$.

The IQR (interquartile range) is given by the ends of the box. Since the box above extends from 2 to 10.25, the IQR

is 10.25 − 2 = 8.25.

We can summarize the information given by the Box-and-whisker chart above in the following table:

Min	Q1	Median	Q3	Max
1	2	5	10.25	16