Exploratory data analysis (EDA)
Part 1: Overview of statistics and scales of measurement

By: Noureddin Sadawi, PhD

University of London

What is statistics?

- Statistics may be viewed as the branch of mathematical science that deals with the collection, presentation, analysis and interpretation of data.
- It has applications in many disciplines and can be used to predict the results in an election, forecast the weather, forecast sales, investigate customer satisfaction levels and much more.
- Two main branches: descriptive and inferential statistics.

Descriptive statistics and Inferential statistics

- Descriptive statistics is the branch of statistics that deals with the collection, summary, presentation and analysis of data in order to transform data into information which can be easily understood and interpreted.
- Inferential statistics is the branch of statistics that involves using data from a sample to make claims, predictions, estimates and test hypotheses about the characteristics of a population.

Population and sample

- Population is the set of all possible individuals (or elements) of interest in a particular study.
- A sample is a subset (or portion) of the population that is representative of the population from which it was selected.

Population and sample

- A specific characteristic of a population is called a parameter.
- A parameter is fixed for a population.
- A specific characteristic of a sample is called a statistic.
- A sample statistic varies from sample to sample.

Example: Population and sample

- Population: all the students at the University of London.
- Variable of interest is age.
- Parameter: the average age of all students at the University of London.
- Statistic: the average age of sampled students at the University of London.

Data

- Data refers to the facts and figures collected, summarised and analysed for presentation and interpretation.
- It consists of individuals, variables and observations.

Individuals, variables and observations

- Individuals are the objects of interest in a study on which data are collected.
- Examples of individuals could be objects such as cars, cities, animals or even people.
- Variables are the attributes or characteristics of interest for an individual.

Individuals, variables and observations

- A few examples of variables are height, weight, blood type and eye colour.
- An observation refers to the set of all outcomes or measurements collected for a particular individual in a study.
- The set of all data collected in a particular study is referred to as data set.

Scales of measurement

- A variable has one of four scales or levels of data measurement: nominal, ordinal, interval or ratio.
- The level of measurement determines the amount of information the data contains and is an indicator of the most suitable data summary and statistical analysis.

Nominal scale

- In the nominal scale of measurement, arbitrary labels or names are used to identify a characteristic of the individual.
- For example, suppose we have data of people belonging to four different gender categories: male, female, transgender or non-binary. The scale of measurement here is nominal, because labels are used to identify the four gender types.

Nominal scale

- Numeric codes as well as non-numeric labels may be used. When numeric codes are used, the numbers in the variables are used only to identify the data and no ranking is allowed.
- For example, we may use numeric code by letting 1 denote the male gender, 2, the female gender, 3, the transgender and 4, for non-binary. Then the numeric values 1, 2, 3, 4 provide the labels used to identify the gender types.

Ordinal scale

- In this scale of measurement, the data shows all the properties of nominal data and there is an order or rank among the variable’s observations.
- When numeric codes are used, there is no measurable meaning to the number differences.
- For example, variables could be educational level: high school, undergraduate degree, Master’s degree or doctorate degree.

Ordinal scale

- The data have the properties of nominal data because labels are used to identify the educational levels. In addition, the data can be ranked with respect to the educational level.
- Thus, doctorate degree is given the highest rank 1, Master’s degree is given the next rank 2, undergraduate degree is given the third rank 3 and high school is given the lowest rank 4.
- However, the difference between the ranks 1, 2, 3, 4 give no meaningful results.

Interval scale

- Interval scales are numeric scales in which data exhibits the properties of ordinal scale and the differences between the values is meaningful and specified on a fixed unit of measure.
- SATs test scores are an example of interval data. For example, suppose three students have SAT scores 500, 600 and 750. Then the test score can be ranked in terms of performance, that is, from best performance to poorest performance.

Interval scale

- Furthermore, the differences between the scores are meaningful. That is, student 1 scored 750 - 600 = 150 more points than student 2 who scored 600 - 500 = 100 more points than student 3, while student 1 scored 750 - 500 = 250 more points than student 3.
- Interval scale has no true zero point. That is, zero does not mean the absence of value but simply indicates another value on the scale.
- For example, there is no such thing as no temperature when temperature is measured in degrees Celsius. Rather, 0 degrees Celsius is another number used on the scale.

Ratio scale

- This exhibits all the properties of interval data, and the ratio of any two values is meaningful. This scale also has a true zero point, that is, when the variable equals 0.0, there is none of that variable.
- For example, consider the yearly income earned by employees at an IT firm. An income of £0 would indicate the employee did no work that year and so earned no money. Furthermore, if we compare the income £30,000 of one employee to the income of £36,000 of another employee, then the ratio property shows that the second employee earns 36k/30k = 1.2 times the first employee salary.

Ratio scale

- Other examples of variables which use the ratio scale of measurement are weight, height and time.