Chapter 2 Descriptive Statistics

2.1 Introduction to statistics

Statistics refers to the scientific method by which data is collected, organized, analyzed and interpreted for the purpose of description and decision making. It is therefore the foundation knowledge of data science.

Statistical methods can be further divided into two subdivisions:

- **Descriptive statistics** deals with the presentation of numerical facts, or data, in either tables or graphs form, and with the methodology of analyzing the data.
- **Inferential statistics** involves techniques for making inferences about the whole population on the basis of observations obtained from samples.

In this course, we will focus on descriptive statistics which can be applied to data analytics.

2.2 Central tendency

It seems apparent that in most set of numerical data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as **central tendency**.

For a given set of data, the measure of location we use depends on what we mean by middle; different definitions give rise to different results. We shall consider some more commonly used measures, namely arithmetic mean, median and mode. The formulas in finding these values depend on whether they are ungrouped data or grouped data.

Arithmetic mean (or simply mean) is obtained by adding together all of the measurements and dividing by the total number of measurements taken. The population mean μ is the mean of all the N measurements from the whole population. Mathematically it is given by the following formula:

$$\mu = \frac{\sum x_i}{N}$$

where $x_1, x_2, x_3, \ldots$ are the measurements.

Arithmetic mean can be used to calculate any numerical data and it is always unique. It is obvious that extreme values affect the mean. Also, arithmetic mean ignores the degree of importance in different categories of data.

Median is defined as the middle item (or 50^{th} percentile) of all given observations arranged in order. In case of the number of measurements is even, the median is obtained by taking the average of the middle. Median unique and it is not affected by a few extreme values.

Mode is the value which occurs most frequently. Given a set of data, we simply count the frequency of each value. If more than one values have the same largest frequency, then the mode is not unique.

Consider the following set of data:

$$2,2,2,2,8,10,15,17,20,99$$

The mean is:

$$\frac{2+2+2+2+8+10+15+17+20+99}{10} = \frac{177}{10} = 17.7$$

The median is:

$$\frac{8+10}{2} = 9$$

The mode is:

$$2\ (freq: 4)$$

The example above illustrates using different measures of central tendency might give different results. Depending on the situation, we need to choose a suitable method. For example, to reflect the salary level of employees in Hong Kong, median is more suitable than mean and mode. The reason is that mean salary is greatly affected by a small fraction of employees with extremely high salary (e.g. CEO of big firms), while the mode might be just equal to the minimum wage.

2.3 Dispersion

Central tendency including mean, median and mode may not be able to reflect the true picture of some data. Compare the two sets of data below:

$$A: 28,29,30,31,32$$

$$B: 10,10,30,50,50$$

Both the mean and median of A are 30, which is equal to that of B. However, it is obvious that the structure of A and B are quite different in the sense that the values in A are close, meanwhile those in B are more extreme. It is necessary to set up some measures to study the **dispersion** or variability in a set of data.

Range is the difference between the maximum and the minimum values.

Quartiles are the most commonly used values of position which divides distribution into four equal parts such that 25% of the data are $\leq Q_1$; 50% of the data are $\leq Q_2$ (or median); 75% of the data are $\leq Q_3$. It is also denoted the value $(Q_3 - Q_1)/2$ is the quartile deviation, QD, or the semi-interquartile range.

Mean absolute deviation (MAD) is the mean of the absolute values of all deviations from the mean. Therefore it takes every item into account. Mathematically it is given as:

$$\frac{\sum|x_i - \mu|}{N}$$

where μ is the population mean.

Variance and **standard deviation** can also be used to measure variation. The population variance σ^2 is the mean of the square of all deviations from the mean. Mathematically it is given as:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

And the population standard deviation σ is defined as the square root of variance:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Notice that σ has the same unit as the data values x_i but σ^2 does not.

Coefficient of variation (CV) is a measure of relative importance. It does not depend on unit and can be used to make comparison even two samples differ in means or relate to different types of measurements. The coefficient of variation gives:

$$\frac{\sigma}{\mu}$$

Now we can revisit the example in the beginning of the section. Given two sets of data A, B with the same mean $\mu_A = \mu_B = 30$. Compare their dispersion.

$$MAD_A = \frac{|28-30| + |29-30| + |30-30| + |31-30| + |32-30|}{5} = 1.2$$

$$MAD_B = \frac{|10-30| + |10-30| + |30-30| + |50-30| + |50-30|}{5} = 16$$

$$\sigma_A^2 = \frac{(28-30)^2 + (29-30)^2 + (30-30)^2 + (31-30)^2 + (32-30)^2}{5} = 2$$

$$\sigma_B^2 = \frac{(10-30)^2 + (10-30)^2 + (30-30)^2 + (50-30)^2 + (50-30)^2}{5} = 320$$

$$\sigma_A = \sqrt{2} \approx 1.414$$

$$\sigma_B = \sqrt{320} \approx 17.89$$

$$CV_A = \sqrt{2}/30 \approx 0.0471$$

$$CV_B = \sqrt{320}/30 \approx 0.596$$

We can see that the dispersion in B is much greater than that of A with either method.

2.4 Standardization

Due to the nature of the measurement or the unit used, different sets of data might have different scales. Since many machine learning algorithms require feature scaling for optimal performance, there is a need to transform the data sets with a uniform scale. This process is called **standardization**. Suppose a set of data $x_1, x_2, \dots, x_N$ has mean μ and standard deviation σ, we can subtract μ from each data and then divide the difference by σ. In other words, the standardized data is given by the standardization formula:

$$z = \frac{x - \mu}{\sigma}$$

This value of z is also called **standard score** of the data. This process is invertible. In other words, if we are given a standard score z, we can retrieve the original data value by:

$$x = \mu + z\sigma$$

if mean μ and standard deviation σ is known to us.

Notice that the standardized set of data $z_1, z_2, \dots, z_N$ has mean 0 and standard deviation 1, despite of the mean, standard deviation or even the unit used in the original set of data.

Consider the set of data A in the previous example. We evaluated have $\mu = 30, \sigma = \sqrt{2}$. Using standardization formula, we can evaluate the standardized data as follows:

$$z_1 = \frac{28 - 30}{\sqrt{2}} = -\frac{2}{\sqrt{2}} \approx -1.414$$
$$z_2 = \frac{29 - 30}{\sqrt{2}} = -\frac{1}{\sqrt{2}} \approx -0.707$$
$$z_3 = \frac{30 - 30}{\sqrt{2}} = 0$$
$$z_4 = \frac{31 - 30}{\sqrt{2}} = \frac{1}{\sqrt{2}} \approx 0.707$$
$$z_5 = \frac{32 - 30}{\sqrt{2}} = \frac{2}{\sqrt{2}} \approx 1.414$$

2.5 Sampling

The population mean μ and population variance σ^2 introduced before can be evaluated from the whole set of population data. However, if the population is too large, data collection and evaluation would be practically difficult. Instead, we can use a sample with relatively small size to estimate the population parameters. This process is called **sampling**. Imagine if we ask to find the population mean height of all citizen in Hong Kong, we can randomly choose 100 citizen and measure their height. The estimation however, will be dependent on the random sample. The concept of estimation and confidence interval will be further discussed in some statistics courses. In this course, we will just introduce the formula for sample mean and sample variance.

Consider a set of sample data $x_1, x_2, \dots, x_n$ with sample size n. We can use sample mean $\overline{x}$ to estimate the population mean.

$$\overline{x} = \frac{\sum x_i}{n}$$

We can also use sample variance s^2 to estimate population variance.

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

Sample standard deviation s is just the square root of sample variance.

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

Notice that sample variance involves division by $n-1$, while population variance involves division by N. We can formulate this into $N - ddof$ where $ddof$ refers to "**delta degree of freedom**". For sample variance, $ddof = 1$. For population variance, $ddof = 0$.