Descriptive statistics:

Descriptive statistics is a method of organizing and summarizing each piece of data obtained from an observation.
Observations can yield a lot of data. You can collect data by conducting experiments indoors, or you can collect data by conducting surveys in the city. Observation is an experiment in the field of natural science, and a survey in the field of social science.
The data obtained from the observation is first entered into Excel or something. At that point, it is nothing more than a list of data.
The more data you have, the more difficult it is to understand. The more data you have, the more confusing the situation becomes.
We want to read the data correctly and efficiently. To do this, you can try tables and graphs, or calculate averages and standard deviations.

Mean:

In English, there are two words for averages: average and mean. In general, both average and mean means "arithmetic mean" unless otherwise specified.

Standard deviation:

Variance is a measure of "how much the data varies around the mean. However, it is important to note that "variance can be compared to each other, but variance and mean cannot be added together, nor can variance and mean be compared. This is due to the fact that each data is squared when calculating the variance.

For example, if the heights of 100 people are measured in "cm", the unit of the mean is "cm", but the unit of the variance is the square of that, "cm2", so the values of the mean and variance cannot be compared or calculated as they are.

So, by calculating the "square root" of the variance, the squared units are restored and can be added or subtracted. The positive square root of the variance is called the "standard deviation".

A 95% confidence

A 95% confidence interval means that if you take a sample from the population and calculate a 95%

confidence interval from the mean 100 times, 95 times the mean will be included in the interval.

For example, let's say that the average height of all Hungarians (i.e., the population mean) is 170 cm. In this case, an experiment is conducted 100 times to calculate the 95% confidence interval from the heights of 100 randomly selected people. The results are as follows

The first experiment: 150 < x < 175
The second experiment: 162 <x< 172
...
100th experiment: 145 <x< 180

Suppose that the confidence intervals are calculated as follows.
Of these 100 confidence intervals, 5 or so are in the range not including the population mean of 170. This is what the 95% confidence interval means.

The u-test

Method for detecting differences in ordinal variables (ordinal scale) between two groups Method for determining the difference between ordinal variables (ordinal scale) in two groups
- Equivalent to an uncorrelated t-test (continuous variables only)
- Can be used for continuous as well as ordinal variables, and is said to have high power
It is said to be a test with high power.
- Even for continuous variables, the number of subjects is small and the distribution is not normal.
If the number of subjects is small and the data is not normally distributed (not appropriate to be expressed as mean and variance)
If the number of subjects is small and the data is not normally distributed (not appropriate to be expressed as mean and variance), this method is used.
- When treated as ordinal variables, the median and range (minimum and maximum) are used as descriptive statistics.
The median and range (minimum and maximum) are presented as descriptive statistics.

$$U_k = n_1 n_2 + \frac{n_k(n_k+1)}{2} - R_k$$

- NOTE; n1,n2 is the sample size (for example, if there are 10 mens and 8 womens, the n1 can be 10 and n2 can be 8.Rk is the total of the sample data between two groups, like if 10 mens height mean is 10.8, the Rk is 108 (because the number of mens are 10))

The statistic for the Mann-Whitney U test is the smaller of the values of U1 and U2 obtained above. When the sample size is small, the statistical value table for Mann-Whitney U test is used to determine

the rejection limit. In contrast, when the sample size is large, as a rule of thumb, if n1 > 20 or n2 > 20, the Standardize the statistic U and approximate it with a standard normal distribution.

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$Z = \frac{U - \mu_U}{\sigma_U} \sim \mathcal{N}(0,1)$$