DESCRIPTIVE STATISTICS

Dr Alina Gleska

Institute of Mathematics, PUT

April 20, 2018

Measures of statistical dispersion (the variation) - in addition to locating the center of the observed values of the variable in the data, another important aspect of a descriptive study of the variable is numerically measuring the extent of variation around the center. Two data sets of the same variable may exhibit similar positions of center but may be remarkably different with respect to variability. We distinguish:

- classical measures - depending on all observations;
- positional measures - depending on the position in the series.

All measures of dispersion we can divide with the respect to another criterium:

- absolute - they have the same units as variables;
- relative - they have no units (or are presented in percentage).

If we want to compare the variables with different units we can use only relative measures.

The sample range of the variable is the difference between its maximum and minimum values in a data set:

$$R = x_{max} - x_{min}.$$

The very simple measure (advantage), but it depends on outliers (disadvantage). Used only for the preparatory analysis.

The sample interquartile range of the variable, denoted R_0 (or IQR), is the difference between the first and third quartiles of the variable, that is,

$$R_0 = Q_3 - Q_1 .$$

Roughly speaking, the R_0 gives the range of the middle 50% of the observed values.

The quartile deviation it is a half of the sample interquartile range:

$$Q = \frac{R_0}{2} = \frac{Q_3 - Q_1}{2}.$$

It informs how much is the average deviation of the middle 50% of the observed values from the median and it is mainly used when the distribution is highgly skewed.

Positional typical range of variables

Typical units are those observations that belong to the interval $(Me - Q, Me + Q)$.

Remark

We have to distinguish the positional typical range of variables from the mode or from the mode interval. There are two different concepts.

The average deviation allows to determine how much the concrete observations differ from the arithmetic mean:

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|,$$

where x_i is the value of the i-th observation, $\overline{x}$ is the arithmetic mean, and n denotes the total number of observations.

Properties:

- the average deviation is always non-negative: $d \geq 0$,
- $d = 0$ only if all observations are the same,
- the bigger average deviation, the higher diversity of the population,
- the average deviation has the same units as variables.

The variance is the average squared deviation from the mean. Its usefulness is limited because the units are squared and not the same as the original data. The sample variance is denoted by

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

where x_i is the value of the i-th observation, $\overline{x}$ is the arithmetic mean, and n – the total number of observations.

We can modify the previous formula to the more convenient one:

$$s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2.$$

Properties:

- the variance is always non-negative: $s^2 \geq 0$,
- $s^2 = 0$ only if all observations are the same,
- the bigger variance, the higher diversity of the population.

The standard deviation determines how much observations differ from the arithmetic mean:

$$s = \sqrt{s^2}.$$

The standard deviation has the same units as variables.

Properties:

- the standard deviation is always non-negative: $s \geq 0$,
- $s = 0$ only if all observations are the same,
- the bigger standard deviation, the higher diversity of the population,
- the standard deviation is always greater than the average deviation $s > d$,
- there is a relation between the standard deviation, the average deviation and the quartile deviation: $s > d > Q$.

Remark

We can calculate the variance using the different formula:

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

Both methods are correct. We use $n-1$ in formula when our data come from the small sample ($n < 30$), and we want to estimate the variance of the whole big population. It can be mathematically proven that this variance, calculated with $n-1$ is a better estimator ot the real variance.

Classical typical range of variables

Typical units are those observations that belong to the interval $(\overline{x} - s, \overline{x} + s)$.

Remark

We have to distinguish the classical typical range of variables from the mode or from the mode interval. There are two different concepts.

The quartile coefficient of dispersion is defined as:

$$V_Q = \frac{Q}{Me} \cdot 100\%.$$

Properties:

- $V_Q \geq 0\%$,
- $V_Q = 0\%$, if there is no diversity in the population,
- the higher value of the quartile coefficient of dispersion, the higher diversity of the population.

Classification of the value of the quartile coefficient of dispersion V_Q:

- $0\% - 20\%$ – a weak diversity,
- $20\% - 40\%$ – a moderate diversity,
- $40\% - 60\%$ – a strong diversity,
- more than 60% – a very strong diversity.

The classical coefficient of variation (V_s) is defined as the ratio of the standard deviation s to the arithmetic mean $\overline{x}$. It shows the extent of variability in relation to the mean of the population.

$$V_s = \frac{s}{\overline{x}} \cdot 100\%.$$

Properties:

- $V_s \geq 0\%$,
- $V_s = 0\%$, if there is no diversity in the population,
- the higher value of the classical coefficient of variation, the higher diversity of the population.

Classification for the classical coefficient of variation V_s:

- $0\% - 20\%$ – a weak diversity,
- $20\% - 40\%$ – a moderate diversity,
- $40\% - 60\%$ – a strong diversity,
- more than 60% – a very strong diversity.

The variance and the standard deviations for discrete grouped series:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i \quad \text{or} \quad s^2 = \left(\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i\right) - \overline{x}^2,$$

The variance and the standard deviations for continuous grouped series (in the intervals):

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i^0 - \overline{x})^2 n_i \quad \text{or} \quad s^2 = \left(\frac{1}{n}\sum_{i=1}^{k}(x_i^0)^2 n_i\right) - \overline{x}^2.$$

Gauss (normal) distribution.

Chebyshev's Rule

At least $1 - \frac{1}{k^2}$ of the data will lie within k standard deviations of the mean (in the intervals $(\overline{x} - ks, \overline{x} + ks)$), where $k > 1$. So in the interval:

- $(\overline{x} - 2s, \overline{x} + 2s)$ lies at least $\frac{3}{4}$ (75%),
- $(\overline{x} - 3s, \overline{x} + 3s)$ lies at least $\frac{8}{9}$ (89%),
- $(\overline{x} - 4s, \overline{x} + 4s)$ lies at least $\frac{15}{16}$ (93,75%),
- $(\overline{x} - 5s, \overline{x} + 5s)$ lies at least $\frac{24}{25}$ (96%).

Example