07 Measures of variability

As we already know, the second central moment is called a **variance**. Given a sequence y, its variance is denoted by var(y), or $s^2(y)$, so

$$\mathrm{var}(y) := \frac{1}{N}\cdot \sum_{k=1}^{N} (y_k - \bar{y})^2 .$$

It is easy to see that

$$\mathrm{var}(y) = \frac{1}{N}\cdot \sum_{k=1}^{N} y_k^2 - \bar{y}^2 .$$

The square root of the variance is called a **standard deviation**. A standard deviation of the sample s is denoted by std(y) or $s(y)$, so

$$\mathrm{std}(y) := \sqrt{\mathrm{var}(y)} .$$

The interval $< \bar{y} - \mathrm{std}(y),\ \bar{y} + \mathrm{std}(y) >$ is referred to as an **interval of (classical) variability.**

If the mean $\bar{y} \neq 0$, the formula

$$\mathrm{dispersion}(y) := \frac{\mathrm{std}(y)}{\bar{y}}$$

defines a (**classical**) **variability coefficient**, shortly called a **variability**, a **variation**, a (**classical coefficient of**) **dispersion**.

For brevity let's denote d = dispertion(y).

If $d < 0.2$, we say that a sample y has a weak (or faint) dispertion,
if $0.2 < d < 0.4$, we say y varies moderately,
if $0.4 < d < 0.6$, we say y exhibits a strong variation,
if $0.6 < d$, we say y varies very strongly,
(and the limiting values can be included arbitrarily).

The variance, the standard deviation and the (classical) dispersion are classified among **classical measures of variability**. Another measures included to this class is an **average deviation**, a **mean deviation**, and a **relative mean deviation**, defined by formulas

$$d_1(y) := \frac{1}{N}\cdot \sum_{k=1}^{N} | y_k - \bar{y} | ,$$

$$H_1(y) := \frac{d_1(y)}{\bar{y}} ,$$

resp.

Positional measures of variability are

- the range of the sample,

– the interquartile range, aka a **quartile deviation**, $\mathrm{IQR}(y) = Q_3 - Q_1$,

– a **positional dispersion**, $\mathrm{pod}(y) := \dfrac{\mathrm{IQR}(y)}{\mathrm{median}(y)} = \dfrac{Q_3 - Q_1}{Q_2}$.

Analoguously as the interval of (classical) variability is formed, the positional dispertion and the median produce an **interval of positional variability**

$$< \mathrm{median}(y) - \mathrm{IQR}(y),\ \mathrm{median}(y) + \mathrm{IQR}(y) >.$$

Example–12. In previous examples we found that the ordeence

$$z = (\, 2.0, 2.0, 2.1, 2.2, 2.2, 2.9, 2.9, 2.9, 2.9, 3.1, 3.3, 3.3, 3.3, 3.5, 3.5, 3.8, 4.3, 6.4, 7.0, 10)$$

has classical measures of position and variablity

the (arithmetic) mean: $\bar{z} = \mathrm{mean}(z) = 3.68$,

the variance: $\mathrm{var}(z) = \dfrac{74.552}{20} = 3.7276$,

the standard deviation: $\mathrm{std}(z) = \sqrt{3.7276} = 1.93069$,

the dispersion: $\mathrm{dispersion}(z) = \dfrac{1.93069}{3.2} = 1.16487$,

the average deviation: $d_1(z) = \dfrac{19.76}{20} = 0.988$,

the relative mean deviation: $H_1(z) = \dfrac{0.988}{3.68} = 0.268478$,

and positional ones:

quartiles: $Q_1 = 2.55$, $Q_2 = \mathrm{median}(z) = 3.2$, $Q_3 = 3.65$,
deviation: $IQR = 3.65 - 2.55 = 1.10$,
dispersion: $\dfrac{1.10}{3.2} = 0.34375$

□ *Example–12*.

Obviously, classical measures of position and variability of the ordeence $z = \mathrm{ord}(y)$ are the same as that of y. This property doe not hold true when the condensation is done, and we illustrate it in the example below.

Example–13. In Example–10 we produced the condence

$$(c, q) = (3.0, 16; 5.0; 7.0, 2; 9.0)$$

assigned to the sequence z. This condence is the multence we dealt with in Example-5. The we found that its mean $a = 3.8$, and its variance (i.e., the second central moment) $M_2 = 2.96$. In consequence, the standard deviation of considered condence (c, q) is $\sqrt{M_2} = 1.77046$. These three quantities differ from that produced for the multence (x, m) which is nothing else than a special record of the sequence z. Relative errors of these quantities are

$$\frac{3.8-3.68}{3.8} = 0.0316,\ \frac{2.96-3.7279}{3.7279} = -0.206,\ \frac{1.77046.8-1.93069}{1.93069} = -0.083.$$

□ *Example–13.*

Properties of the variance

a) variance of the constant sequence is 0,

b) $\mathrm{var}(\beta\, y) = \beta^2\, \mathrm{var}(y)$ for any constans β and arbitrary sequence y,

c) $\mathrm{var}(x + y) = \mathrm{var}(x) + 2\, \mathrm{cov}(x, y) + \mathrm{var}(y)$,
where x and y are arbitrary sequences of the same size,

$$\mathrm{cov}(x, y) := \mathrm{E}(\{x - \mathrm{E}(x)\}\{y - \mathrm{E}(y)\}).$$

The just introduced quantity, cov(x, y), is called a **covariance** of sequences x and y.

The formula for the variance of the sum of two sequences follows immediately:

$$\begin{aligned}\mathrm{var}(x + y) &= \mathrm{E}(\{\{x + y\} - \{\mathrm{E}(x) + \mathrm{E}(y)\}\}^2) = \\ &\mathrm{E}(\{\{x - \mathrm{E}(x)\} + \{y - \mathrm{E}(y)\}\}^2) = \\ &\mathrm{E}(\{x - \mathrm{E}(x)\}^2 + 2\{x - \mathrm{E}(x)\}\{y - \mathrm{E}(y)\} + \{y - \mathrm{E}(y)\}^2) = \\ &\mathrm{E}(\{x - \mathrm{E}(x)\}^2) + 2\, \mathrm{E}(\{x - \mathrm{E}(x)\}\{y - \mathrm{E}(y)\}) + \mathrm{E}(\{y - \mathrm{E}(y)\}^2) = \\ &\mathrm{var}(x) + 2\, \mathrm{cov}(x, y) + \mathrm{var}(y).\end{aligned}$$

The covariance is an extension of the variance: for $x = y$ there is

$$\mathrm{cov}(x, x) = \mathrm{var}(x, x).$$

Often properties a) and b) are notified together:

$$\mathrm{var}(\alpha + \beta\, y) = \beta^2\, \mathrm{var}(y)$$

for any constant α, β, and arbitrary sequence y

Notice that if for two given sequences x and y there exist a constant β such that

$$y = \beta\, x,$$

then $\mathrm{E}(y) = \beta\, \mathrm{E}(x)$ and

$$\begin{aligned}\mathrm{cov}(x, y) = \mathrm{cov}(x, \beta x) &= \mathrm{E}(\{x - \mathrm{E}(x)\}\{\beta\, x - \beta\, \mathrm{E}(x)\}) = \\ &\beta\, \mathrm{E}(\{x - \mathrm{E}(x)\}^2) = \beta\, \mathrm{var}(x).\end{aligned}$$

Two non-zero vectors x and y satisfying the relation $y = \beta\, x$ are said to be linearly dependent. In statistics, two sequences, x and y, are said to be (**statistically**) **independent** if their covariance is 0,

$$\mathrm{cov}(x, y) = 0,$$

or, equivalently, if the variance of their sum is the sum of their variances,

$$\mathrm{var}(x + y) = \mathrm{var}(x) + \mathrm{var}(y).$$

In statistics there is used a notion which is broader than the independence. This notion is called a correlation, and a linear correlation is numerically expressed via so-called Pearson correlation coefficient (cocoP): in its definition there is involved the covariance, it is defined is sensitive to linear relationship between sequences, and we will discuss it later.