UNIVERSITATEA DE MEDICINĂ ŞI FARMACIE
„VICTOR BABEŞ" DIN TIMIŞOARA

PREPARATION OF BACHELOR DEGREE

PRACTICAL COURSES

Șef.lucr.dr. Ruxandra SAVA-ROȘIANU

STATISTICAL ANALYSIS OF VARIABILITY

The characterization of a statistical community through the indicators of the central tendency helps us to detect what is common, essential in the manifestation of a phenomenon. Every community has a certain internal organization, defined by the way in which the individual values are scattered or concentrated around the central value. Thus it can happen that two communities analysed by the same variable are different by the central tendency (a), by dispersion (b) or by both (c). In this way one central value may be credible, another not. For this reason it is necessary that the analysis by the central tendency indicators be supplemented with the indicators of the variation and the distribution form.

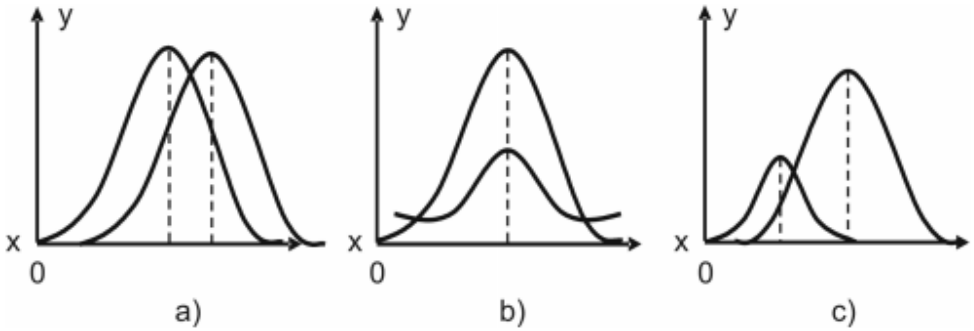

Fig. 1 a) Distributions with different central tendency; b) Distributions with different variability; c) Distributions with different central tendency and variability.

The calculation and analysis of the indicators of the variation of the individual values compared to the average offers the possibility to solve some problems of statistical knowledge. These include:

• analyzing the degree of homogeneity of the data from which the central tendency indicators were calculated and verifying their representativeness;

• comparing in time and (or) spaces of several distribution series according to independent and (or) interdependent characteristics;

• objective selection of significant influence factors, after which the units of a statistical collectivity are structured;

• separating the action of the essential factors from the random factors;

• concentration of the individual values of the characteristics and their displacement to the typical values;

• applying different tests of mathematical statistics.

The variation indicators used in the statistical analyzes are classified according to several criteria:

- according to the number of variants taken into account there are simple indicators and synthetic indicators;

- according to the systematization of the primary data, there are variation indicators calculated for the one-dimensional distribution series and indicators calculated for the multidimensional series;

- after the calculation and expression mode there are indicators of the variation calculated as absolute and relative quantities.

Simple indicators:

- the amplitude of the variation;

- deviation of the individual values from the average; inter-quantile deviation.

Synthetic indicators:

- dispersion;

- mean square deviation;

- coefficient of variation;

Quantilele or the interquartile range deviation - separates the statistical series into "n" parts, comprising the same number, equal to 1 / n of the total number.

a) Quartiles - Q1, Q2, Q3 n = 4 Q2 = Me

b) Deciles - D1 .... D9 n = 10 D5 = Me

c) Centiles - C1 .... C99 n = 100 C50 = Me

d) Promile - P1 .... P999 n = 1000 P500 = Me

❖ **Indicators of dispersion of quantitative characteristics**

The amplitude is the difference between the maximum value and the minimum value in the series. It shows the character of the dispersion, its maximum and minimum limits.

A = Xmax - Xmin

For. series variation with classes takes the upper limit of the class with the highest values for Xmax and the lower limit of the class with the lowest values for Xmin.

Relative amplitude (A%) =

The variance (dispersion) is the arithmetic mean of the squares of deviations between variants and their mean

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot f_i}{n} \qquad n > 30 \qquad n = \Sigma f_i$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot f_i}{n-1} \qquad n \leq 30$$

So for the variance calculation, the weighted arithmetic mean is calculated first

$$\overline{x}_p = \frac{\sum x_i f_i}{\sum f_i}$$

x_i	f_i	$x_i \cdot f_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 f_i$
Total	Σ=	Σ=			Σ=

Tabel nr. 1 Table for calculating variance

x_i	f_i	$x_i - x_0$	$(x_i - x_0) f_i$	$(x_i - x_0)^2 f_i$
Total			$\Sigma=$	$\Sigma=$

Tabel nr. 2 Table for calculating the variance by the moments method

$$s_x^2 = \frac{\sum (x_i - x_0)^2 f_i}{n} - \left[\frac{\sum (x_i - x_0) f_i}{n}\right]^2$$

$$\frac{\sum (x_i - x_0)^2 f_i}{n} \rightarrow M_2 = \text{II}^{\text{nd}} \text{ class moment}$$

$$s_x^2 = M_2 - M_1^2$$

Standard deviation *(s_x)*

$$s_x = \pm\sqrt{s_x^2} \qquad s_x = \sqrt{M_2 - M_1^2}$$

The standard deviation preserves the unit of measure of that characteristic. This fact is also a drawback, because it cannot be used to compare the dispersion for two phenomena expressed by characteristics with different units of measurement.

The standard deviation highlights the value range, around the average, in which the frequencies of the investigated phenomenon were distributed. The lower the values, the closer the values are around the average and the sample is more homogeneous.

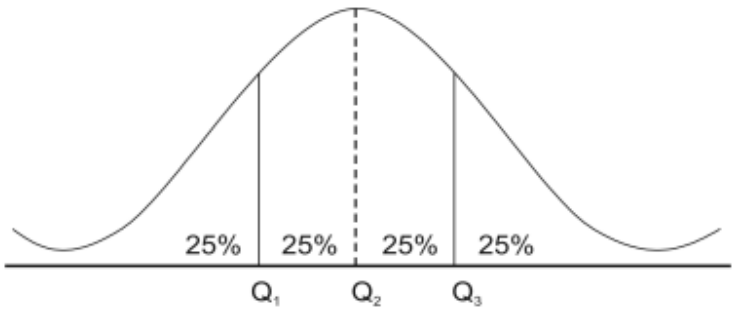

Fig. 2. Curba Gauss- Laplace

The coefficient of variation (C.V.%) is the percentage ratio between the standard deviation and the average value of the string.

The value of the coefficient of variation does not have a unit of measure.

It gives the degree of homogeneity:

CV <10% => small dispersion

10% <CV <30% => medium dispersion

CV> 30% => high dispersion

Observation:

It is accepted that the average is less representative for a range of values, the higher the dispersion.

❖ **Central tendency indicators of qualitative characteristics**

- **Intensive or frequency indices**

$$\text{I.I.} = \frac{\text{nr.cazuri} \cdot 10^2 (10^3)(10^4)(10^5)}{\text{nr.persoane din colectivitatea respectiva}}$$

As an example, the frequency of caries disease in a community can be evaluated.

- **Extensive or structural (or specific weight) indices**

It shows the specific weight of a part of the community in relation to the totality of the community.

$$\text{I.E.} = \frac{\text{nr.caz dintr - o componentă}}{\text{nr. total de cazuri}} \cdot 100$$

Example: $M_{prop} = \frac{D_{cd}}{D} \cdot 100$

M_{prop} – lethality;

D_{cd} – deaths caused by dsease;

D - total deaths.

As an example, lethality can be calculated for squamocellular cancer, or for other malignancies in the oro-maxillofacial sphere in the county of Timiș.

- **Indicators of dispersion of qualitative characteristics**

p - the proportion of a possible state of the qualitative characteristic

- alternative $\sigma_p^2 = p \cdot q$

- standard deviation $\sigma_p = \pm\sqrt{\sigma_p^2}$

❖ **Appreciation of the truthfulness of the average and relative values**

The constants (mean or dispersion) obtained by us on the samples are called "statistics" and their values are more or less close to the values of the constants of the general collectivities (parameters) - depending on the degree of representativeness (qualitative and quantitative) of the samples, but they are never identical.

With the help of these sample constants, the "statistics", the general community constants, the "parameters" are estimated. The estimation result based on the sampling constants is random. As such, the exact values of the parameters of the general community remain unknown, instead we can specify a value interval in which the average of the general community (MA = absolute average) will be located, around which, within that value range, the sample averages will be distributed, with a certain probability.

If we extract several samples from a general community, the average values of these samples will be very close to the value of the absolute average, their distribution, around the absolute average, being made according to the same Gauss-Laplace curve.

The standard error or the average error of the means is the constant that allows us to determine the value range in which the absolute average is found and around which the average sample values are distributed, with a certain probability, it is called standard error (ES) or average error of the values sample averages versus universe averages.

❖ **Safety range**

The value range, determined by means of the standard error, in which the absolute average is estimated, is called the confidence interval or statistical confidence, in this safety interval, determined by the average of the sample plus / minus the standard error. The absolute average will be found in a proportion of 68.26%, so the probability that the absolute average is within this range is 68.26%.

M-ES> l <(M + ES) = 68.26% P-ES> lIÎ <(P + ES)

In our example, the weighted average being 5.84 teeth and the standard error 0.14, the safety interval will be 5.84 ± 0.14, so between 5.70 and 5.98. 2 × ES 3 × ES

❖ **Significance threshold**

The counterprobability or probability that the sample averages exceed the limits - maximum and minimum - of the safety interval, being outside them, is called the significance threshold, if the safety interval is determined by M ± ES, then the counter-probability (the threshold) of significance) is obtained by subtracting from 100 the probability value, so:

q = 100 - 68.26% = 31.74%

In our example the weighted average being 5.84 teeth affected, the safety interval will be: 5.84 ± 0.14, so it will be between 5.70 -5.98.

If we want the value of the counter-probability, the possibility of failure, to be lower, then we must increase the safety interval. This interval is increased by adding and subtracting from the average value twice the value of the standard error.

M ± 2.ES = 5.84 + 2.0.14 = 5.84 ± 0.28

The safety interval will therefore be between the limits of 5.56 and 6.12. In this interval, slightly higher, the value of the absolute average will be found with a higher probability than in the first case, of 95.45%, and the possibility of being outside this safety interval is reduced to 4.55% (100% - 95.45% = 4.55%). If we want to further reduce the probability of error, then we increase the safety interval by adding and subtracting from the average value three times the value of the standard error. The safety range will be:

M ± 3.ES = 5.84 ± 3.0.14 = 5.84 ± 0.42

The safety interval in this case will be between 5.42 and 6.26. In this safety interval, much higher, the absolute average will be found with a probability of 99.73%, and the significance threshold will be 0.27% (100% - 99.73% = 0.27%).

It follows that the safety interval does not have fixed limits, but they change according to the desire with which we want to ensure our results, if we accept a higher threshold of significance, so a higher probability of error, then the safety range is lower. The more we want to work more precisely, so the less we make mistakes, the more the safety interval increases.

❖ **Significance test**

The size of the safety interval depends on the fact that we take only once, twice or three times the value of the standard error around the average. The multiple of the standard error (1, 2 or 3), which determines the size of the safety interval, is called significance test and is noted with the letter "t". As such, at a probability of 68.26% and a significance threshold of 31.74% the value of t = 1; at 95.45% probability and 4.55% significance threshold, the value of t = 2; at a probability of 99.73% and a significance threshold of 0.27% the value of t = 3. Usually, we do not guarantee the results or conclusions, obtained on the sample, with the probability that they are within the safety interval and with the counter probability, with the probability of failing, so with the significance threshold.

In medicine and biology the significance thresholds of 31.74%, 4.55% and 0.27%, corresponding to t values of 1.2 or 3, are not used very much, however, the results are guaranteed with the thresholds of significance of 0.05 (5%), 0.01 (1%) and 0.001 (0.1%). Based on calculations it was established that for these significance thresholds the corresponding values of t are 1.96, 2.58 and 3.29. As such, at the 5% significance threshold, the value of t will be 1.96, and the safety interval will be: M ± 1.96.ES. At 1% significance level the value of t will be 2.58, and the size of the safety interval will be given by M ± 2,58.ES. At 0.1% significance threshold the value of t will be 3.29, and the safety interval will be given by M: 3.29.ES.

These values of t remain unchanged if we work on samples whose number of frequencies is greater than 120, if we work on samples with a number of frequencies less than 120, then the significance test value changes and are taken from the table of the test t, which we find in the specialized literature of statistics and biostatistics.