
# UNIVERSITATEA DE MEDICINĂ ŞI FARMACIE „VICTOR BABEŞ” DIN TIMIŞOARA

## PREPARATION OF BACHELOR DEGREE

## PRACTICAL COURSES

**Șef.lucr.dr. Ruxandra SAVA-ROȘIANU**

# ELEMENTS OF PROBABILITY THEORY

The basis of probability theory is the notion of a random (random) event. The occurrence of an event depends on the conditions that generate it. If under the given conditions an event must necessarily take place, it is called a sure event; if the event cannot necessarily be called an impossible event. An event that, under the given conditions, may or may not occur, is called a probable or accidental event.

**Definition:** The probability of an event is the ratio between the average number of favorable cases and the total number of possible cases, provided that all cases are equally possible.

$$P(A) = \frac{m}{n}$$

m – number of favourable cases

n – number of possible cases

Example: Zarul. The chance to get one of the dice faces is equal for each of the dice. There are 6 possible events. So we have the probability of 1/6 for each part of the dice, which represents the favorable event.

The number of favorable cases cannot be greater than that of the possible cases, so the value of P cannot exceed the number 1, equivalent to 100%. If a certain dice face does not appear at any throw, the probability of this case is 0 (zero). So the probability of an event can be between 0 and 1, respectively between 0% and 100%, probability 0 being equivalent to an impossible, unattainable event, and the probability of 100% with a certain, absolutely safe event.

The opposite probability, or the probability of the opposite event, is noted with q = 1 - p

If it is an event A, we can write the probability of the opposite event $\overline{A}$ so: $P(\overline{A}) = q = 1 - p$

In the case of the dice: P(A) = 1/6 (to get a face)

$P(\overline{A}) = q = 1 - p = 1 - 1/6 = 5/6$ (probability of the opposite event)

If we add the chance (1/6) with the chance (5/6) we notice that it results:

1/6 + 5/6 = 6/6 = 1. We come to the same above relationship:

p + q = 1 => p = 1 - q

$$q = 1 - p$$

If we want to make a percentual expression we multiply by 100, $P = \frac{m}{n} \cdot 100$

Ex.: $P = \frac{1}{6} \cdot 100$ = 16%

$q = \frac{5}{6} \cdot 100$ = 84% => p + q = 16 + 84 = 100%

An application of this theory is BAYS' FORMULA.

Usually the doctor knows the probability of a symptom or a positive test, conditioned by a certain disease. $P_S(B)$ or $P_+(B)$, but it is important to know the probability of the disease when the test is (+).

Eg: prevalence of a disease = 0.002 (so if we take a person by chance, the probability of having the disease is 0.002)

If we have a specific laboratory test for this disease, our appreciation that the person has the disease may change.

Suppose the test is real (+) in 80% of the situations - that is, in sick people. In 10% of situations it is (+) in non-ill persons (false positive test).

We can ascertain all these probabilities as follows:

P (B) - probability of the disease being present at a randomly chosen person (corresponding to the prevalence of 0.002)

P () - probability that the disease is not present in a randomly chosen person = 0.998

P + (B) - the probability that a test (+) will confirm the disease

TEST SENSITIVITY = 0.80

P + () - probability that a test (+) corresponds to the absence of the disease, false (+) = 0.10

TEST SPECIFICITY = 1- 0.10 = 0.90

PB (+) - probability of disease, when the test is (+)

BAYES' FORMULA:

$$P_B(+) = \frac{P(B) \cdot P_+(B)}{P(B) \cdot P_+(B) + P(\overline{B}) \cdot P_+(\overline{B})}$$

$$P_B(+) = \frac{(0{,}002) \cdot (0{,}80)}{(0{,}002) \cdot (0{,}80) + (0{,}998) \cdot (0{,}10)} = 0{,}158 \text{ or } 1{,}6\%$$

higher value, compared to 0,002

## CORRELATION AND REGRESSION

### CORRELATION

When we want to know whether or not there is any dependency relation between the series of variation of two or more phenomena, we resort to the calculation of correlations.

Definition: Correlation is the study of the connection between 2 random variables.

The intensity of the statistical links is given by the correlation coefficient "r". It also gives the sense of connection not only the intensity.

$-1 \leq r \leq +1$

Interpretation of "r"

- r (+) and close to 1 => strong direct correlation (both variables increase);

- r (-) and close to 1 => strong inverse correlation (one variable increases and one decreases);

- r (±) and close to 0 => poor correlation;

- r = 0 => variables are independent - there is no correlation.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) \cdot f_i}{\sqrt{\sum (x_i - \bar{x})^2 \cdot f_i \cdot \sum (y_i - \bar{y})^2 \cdot f_i}}$$

x = the independent variable (on the abscissa)

y = dependent variable (by ordinate)

**The graphical method** of studying the connection allows to form an image about the direction, form and intensity of the connection between phenomena. The Cartesian system of rectangular axes is used, in which each unit observed is represented by a point, having as coordinates the values of the independent (x) and dependent (y) variables. The result is a cloud of points that illustrates the intensity and shape of the connection.

The stronger the connection between the two variables, the more points will be grouped around a certain line, which may be straight (linear correlation) or curve (curvilinear correlation).

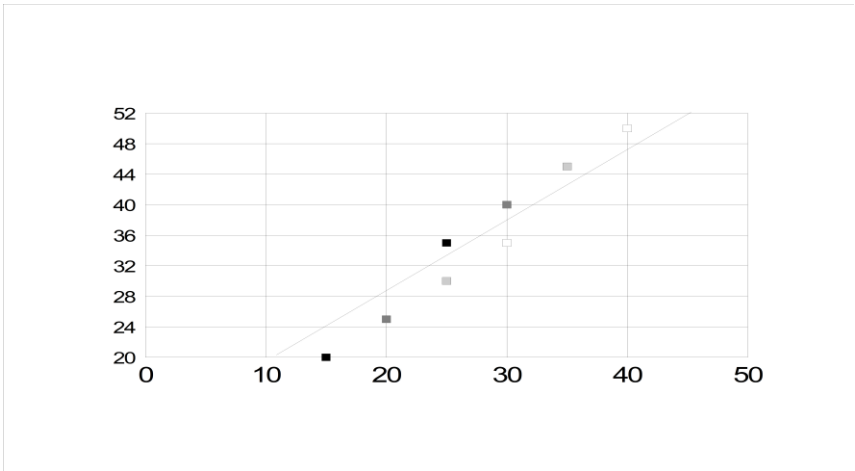


Fig.37. Strong and direct correlation


Fig.38. Strong and reverse correlation

The correlation problem covers any kind of statistical connection, be it quantitative or qualitative. Apart from linear correlation and curvilinearity, we have:

- partial correlation
- multiple correlation
- rank correlation
- temporal correlations - self-correlation
  - synchronous
  - asynchronous

**Interpretation of the correlation coefficient**

The correlation coefficient can be between minus one, zero and plus one. When the value of the correlation coefficient approaches +1, it means that there is a very strong connection between the two phenomena that are correlated. The + sign of the correlation coefficient denotes that the dependence link between phenomena is direct. So both phenomena evolve in the same direction, in the same direction. When the correlation coefficient value approaches -1, it means that between the two phenomena there is a very strong but inverse connection, in the opposite direction: a phenomenon increases, the one with which it correlates decreases.

In medicine, we usually find values of the correlation coefficient intermediate to values -1 and +1. In order to interpret the intensity of the dependency link between phenomena, the following

**CRITERIA** is used:

> the value of the correlation coefficient between ± 1 denotes a very strong correlation between phenomena;

> the value of the correlation coefficient between ± 0.99 and ± 0.70 denotes a strong correlation;

> the value of the correlation coefficient between ± 0.69 and ± 0.30 denotes an average correlation between phenomena;

> the value of the correlation coefficient between ± 0.01 and ± 0.29 expresses the existence of a weak correlation between phenomena;

> the value of the correlation coefficient 0 denotes that the connection between phenomena is practically considered non-existent. The two phenomena thus evolve independently of each other.

The correlation coefficient between phenomena can be correctly interpreted if the following **ASPECTS** are taken into account:

> between phenomena that are logically correlated there is a connection;

> the two phenomena should be investigated on homogeneous samples;

> the choice or selection of the frequency of the samples to be done at random.

The correlation coefficient, which expresses the dependency link between two phenomena, is usually obtained on samples and not on the universe. Its values differ more or less from the value of the correlation coefficient that we would have obtained by studying the phenomena in the entire population.

**Example:**

The level of insurance with dentists (x) and the share of children treated within 5 dental units (y)

| Nr. | Asigurare medici, x | Copii sanatosi, y | $d_X$ | $d_Y$ | $d^2_X$ | $d^2_Y$ | $d_X d_Y$ |
|---|---|---|---|---|---|---|---|
| 1 | 2,3 | 65,7 | -1 | -19,7 | 1 0,01 | 388,1 | 19,7 |
| 2 | 3,2 | 91,7 | -0,1 | 6,3 | 0,1 | 39,7 | -0,63 |
| 3 | 3,4 | 88,0 | 0,2 | 2,6 | 0,04 | 6,8 | 0,52 |
| 4 | 3,6 | 91,4 | 0,3 | 6,0 | 0,09 | 36,0 | 1,8 |
| 5 | 3,9 | 90,3 | 0,6 | 4,9 | 0,36 | 24,7 | 2,94 |
| Total | 16,4 | 427,1 | | | 1,5 | 494,6 | 24,3 |
| | 3,3 | 85,4 | | | 27,2 | | |

$r_{\underline{XY}} = 0,89$

In the grouped statistical series the correlation coefficient is obtained quite complicated, we need the help of the biostatisticians (reporting the sum of the products between the deviations of the values of the variants from the weighted average of the two phenomena that are correlated and the frequencies of the pairs of values of the variants at the square root of the sum of the products between the squares of deviations of the values of the variants of the weighted average and the frequencies corresponding to each variant of the first phenomenon, multiplied by the sum of the products between the squares of the deviations of the values of the variants of the

weighted average and the frequencies corresponding to each variant of the second phenomenon, with which they correlate).

**REGRESSION**

In case of the linear link, the dependent variable changes uniformly under the influence of the independent variable. This change can be expressed using the straight linear equation:

$y = a + bx$

x - the independent variable

y - the dependent variable

a, b - the coefficients of the regression line

b - represents the slope of the straight line geometrically - and is called the regression coefficient. It represents the measure by which the variable y is changed when x is changed by one unit. It takes a negative value in the case of the inverse correlation.

$y = a + bx$

We are in the situation of simple linear regression, in which if:

$b > 0$ => right is ascending

$b = 0$ => right is parallel to Ox

$b < 0$ => right is ascending

$b = tg\alpha$ = slope of the right = coefficient of regression

a = ordered by origin - indicates the value of y when x = 0

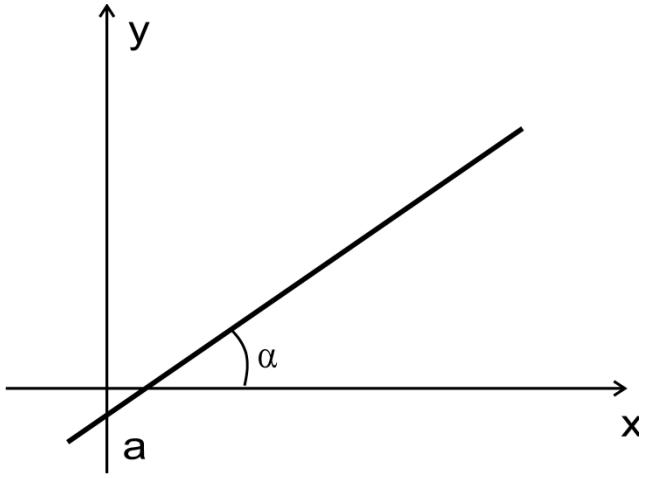


Fig.39. Simple linear regression

The regression coefficient is also expressed by the formula:

$$R_{x/y} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$$R_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x}$$

σx - the standard deviation of x

σy - the standard deviation of y

Regression analysis can be:

A. By the number of variables - simple (1x, 1y)

- multiple (more indep. x, 1 y)

B. By the kind of relationship - linear - simple

- multiple

- nonlinear - simple

- multiple

**Example:**

Height and weight of children aged Z (n = 22)

| Height | Frecvency | Weight | Frecvency |
|---|---|---|---|
| x | fx | y | fy |
| 130 | 3 | 29 | 3 |
| 132 | 5 | 30 | 4 |
| 135 | 7 | 31 | 3 |
| 136 | 4 | 32 | 4 |
| 137 | 3 | 33 | 4 |
| | | 34 | 2 |
| | | 35 | 1 |
| | | 36 | 1 |

The distribution of 22 children according to height (x) and weight (y) gave the following data: average height - 134 cm; average weight - 31.8 kg; standard deviation for height (x) - 2.37; standard deviation for weight (y) - 1.97; correlation coefficient - 0.82.

How much will the body weight change in the children in question, if their height is created by 1 cm?

Substituting the data in the regression coefficient formula, we obtain:

Rxy = 0,82 (1,97/ 2,37)= 0,82* 0, 83= 0,68 kg/cm

**Conclusion**

Increasing the average height in children studied by 1 cm will result in their weight increase by 0.68 kg.

Using the regression coefficient one can find the magnitude of the phenomenon y (in the case analyzed above - the weight), without resorting to its measurement, using for this purpose only the phenomenon x (the weight). The following regression equation is used:

$y = Y + Rg\ \ (x - X)$

wherein: y - weight investigated; x - known height height;

$Rg_{yx}$ - height regression coefficient in relation to weight; Y- the average weight of the investigated community; X - the average height of the investigated community. In this case the average height - 134 cm; average weight - 31.8 kg; Rg - 0.68. It is evaluated what will be the weight of the children who have the height of 135cm.

Substituting the data in the formula, we obtain: $y = Y + Rg\ (x - X) = 31.8 + 0.68\ (135-134) = 31.8 + 0.68 \times 1 = 32.5$ kg

Thus, the height of 135 cm corresponds to the weight of 32.5 kg.

**Regression scale**

In the field of somatometric research of children and adolescents, the method of estimating height, body weight, thoracic perimeter indicators is very important. The individual values of these values sometimes differ quite clearly. In people who have the same height the body weight can vary within quite large limits.