ERASMUS: Course Probability and Statistics
Part II: Statistics - summary

collection, analysis and interpretation of data of experiments

statistical population		random sample
a set of entities concerning which statistical inferences are to be drawn; totality of the observations with which we are concerned	⇒	finite subset taken from the population for investigation; must be representative of the population

Descriptive statistics: summarize the population data by describing what was observed in the sample (collected data) numerically or graphically.

Inferential statistics: find conclusion (inferences) about the population using sample data. These inferences may take the form of:

- answering yes/no questions about the data (hypothesis testing),
- estimating numerical characteristics of the data (estimation),
- describing associations within the data (correlation),
- modeling relationships within the data (regression),

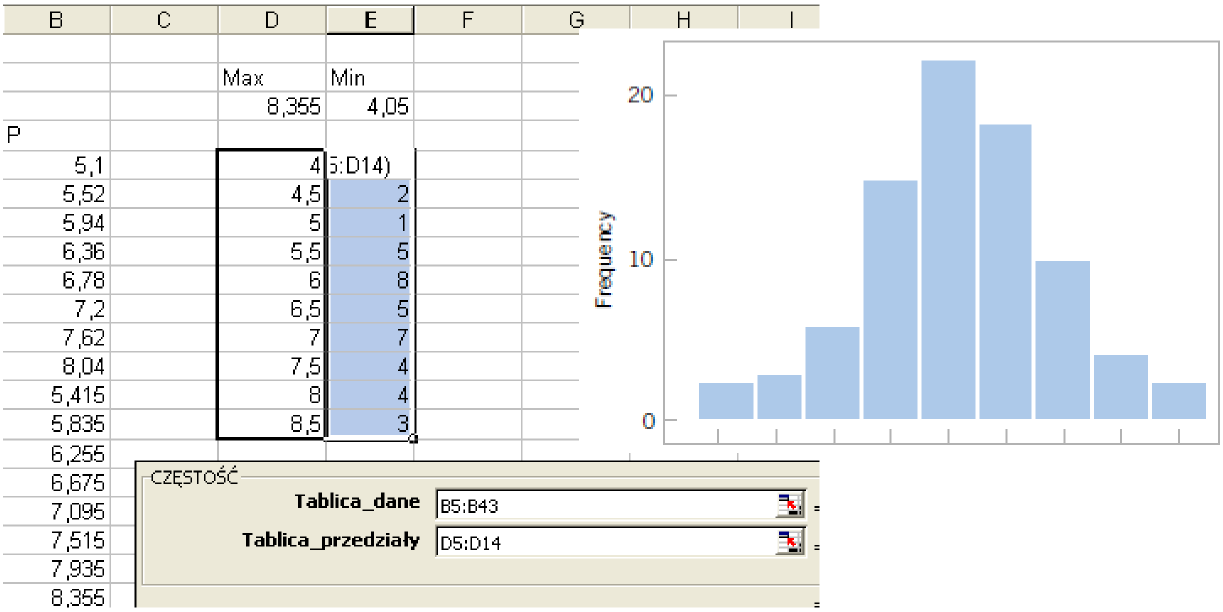
Basic numerical descriptors:

- measures the location or central tendency in the data:
  - **sample mean** (arithmetic, geometric, harmonic) (arithmetic: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$)
  - **median** (numeric value separating the higher half of a sample from the lower half), **quartiles**
  - **mode** (the value that occurs the most frequently in a data set)
- measures of the variability or spread:
  - **sample variance** ($s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$)
  - **standard deviation** (the positive square root of the sample variance)
  - **sample range** ($Max(x_1,...,x_n) - Min(x_1,...,x_n)$)

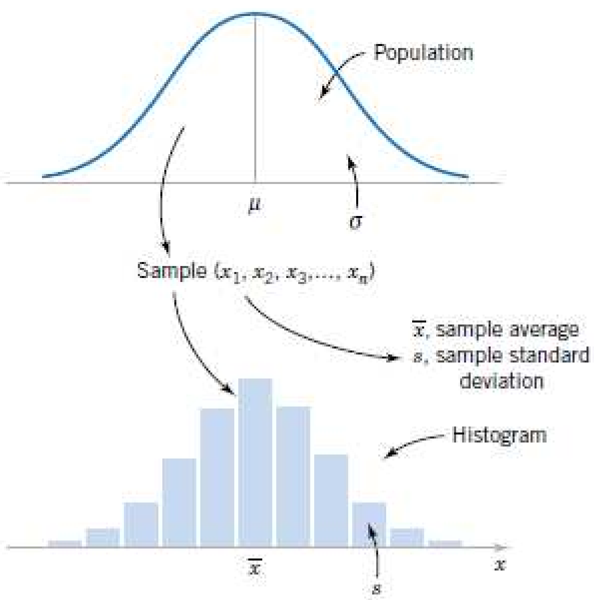
data are a **sample** of observations that have been selected from some larger **population** characterized by **probability distribution** (with population parameters **μ, σ,** coefficient of variation $= \frac{\mu}{\sigma}$)

Presentation of data

- **series, stem-and-leaf, tabular frequencies**
- **histogram** (a graphical display of tabular frequencies, shown as adjacent rectangles. Each rectangle is erected over an interval, with an area equal to the frequency of the interval)



EXCEL: **CZĘSTOŚĆ - frequency :** Tablica_dane – data; Tablica_przedziały – intervals (class); (turk. SIKLIK)



Relationship between a **population** and a **sample**

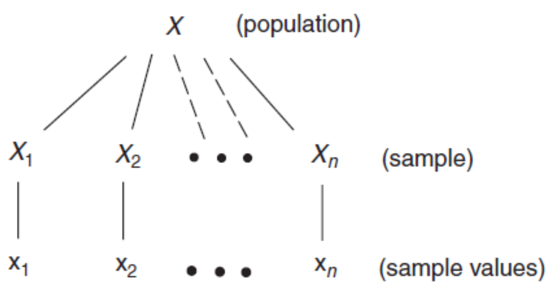
Example;

For a given data (Erasmus-data1.xls) calculate sample mean, median, mode, variance, standard deviation, range. Form a tabular frequencies and histogram. Make a graph of a density function of normal distribution which estimates the sample

Point estimation

θ – a population parameter of a random variable X with a given density function f(x) (ex. mean or standard deviation).

$x_1,\dots,x_n$ - values of observations of a random variable X – a sample from a population $\Rightarrow$ we can consider these values as random variables $X_1,\dots,X_n$ (random sample of size n from X)



Definition:

A function of a random sample of size n $\theta(X_1,\dots,X_n)$ is called *statistic.* If statistic $\theta(X_1,\dots,X_n)$ is used for estimation of some population parameter (single value) is called a *point estimator*. A *point estimate* of some population parameter is a single numerical value of a statistic $\theta(X_1,\dots,X_n)$.

(one parameter can have more estimators)

Reasonable point estimates of basic parameters are as follows:

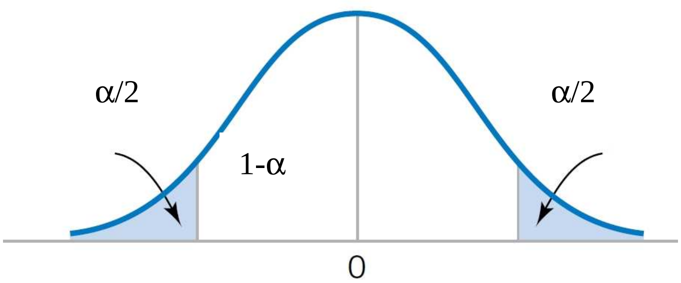
- the **mean** μ of a single population: $\bar{x}$ (sample mean), median
- the **variance** (standard deviation) of a single population: sample variance
- the **proportion** p of items in a population that belong to a class of interest: the sample proportion $\frac{x}{n}$, where x is the number of items in a random sample of size n that belong to the class of interest (Bernoulli trials)

Example;
Suppose that the random variable X is normally distributed with an unknown mean . Calculate a point estimator of the unknown population **mean** using estimators: sample mean, median and a point estimator of the unknown population **standard deviation** using sample standard deviation
Sample: 25, 30, 29, 31, 33
Solution: $\bar{x} = 29.6$, median = 29, $\sigma = 8.8$

Statistical intervals. Interval estimation.

Definition:
Let θ be a population parameter to be estimated. The interval (L_1, L_2) is called a $[100\cdot(1-\alpha)]\%$ **confidence interval** for θ if $P(L_1 < \theta < L_2) = 1 - \alpha$. $1 - \alpha$ is called a **confidence level** (usually expressed as a percentage)

Example;
Let $\sigma = 2$ be a standard deviation of a normal distribution $N(\mu, \sigma^2)$ with unknown μ.
Let $\bar{x} = 34.1$ be a sample mean (n = 16). Then



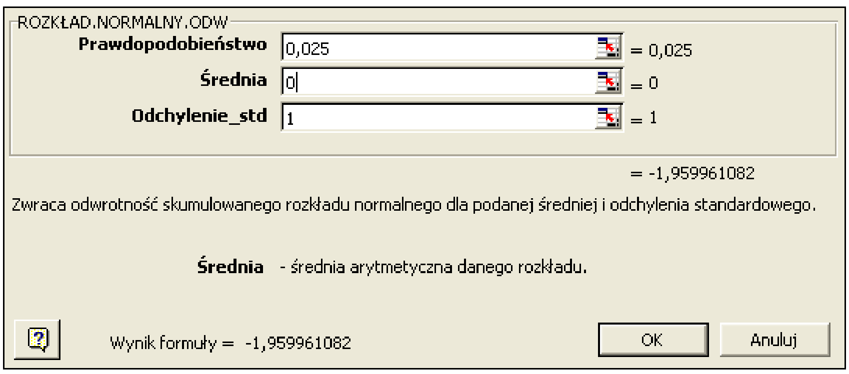
$$P\left[u(\frac{\alpha}{2}) < \frac{\bar{x}-\mu}{\sigma}\sqrt{n} < u(1-\frac{\alpha}{2})\right] = 1-\alpha,$$

thus

$$\bar{x} - u(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} - u(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}$$, where u – density function of **N(0,1)**

Calculate limits of confidence interval assuming α=0.05.

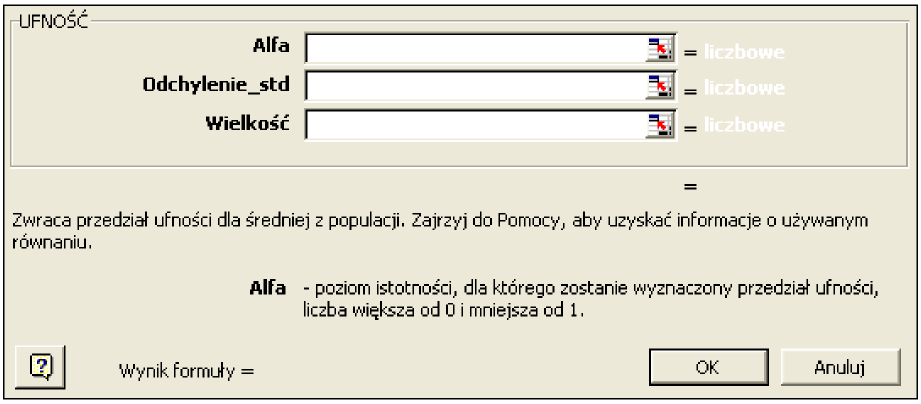
Solution:
u(0.025) = - 1.96,
u(0.975) = - u(0.025) = 1.96
EXCEL: **ROZKŁAD.NORMALNY.ODW – X of normal distribution N(μ, σ^2)** (turk. NORMTERS)
prawdopodobieństwo – probability; średnia (μ)– mean; odchylenie_std (σ) – standard deviation.



c.interval = [34.1-0.98, 34.1+0.98]
= [33.12, 35.08]

EXCEL: **UFNOŚĆ – limits of confidence interval** (turk. GURENIRLIK)

alfa – (α); odchylenie_std (σ) – standard deviation; wielkość – size of a sample



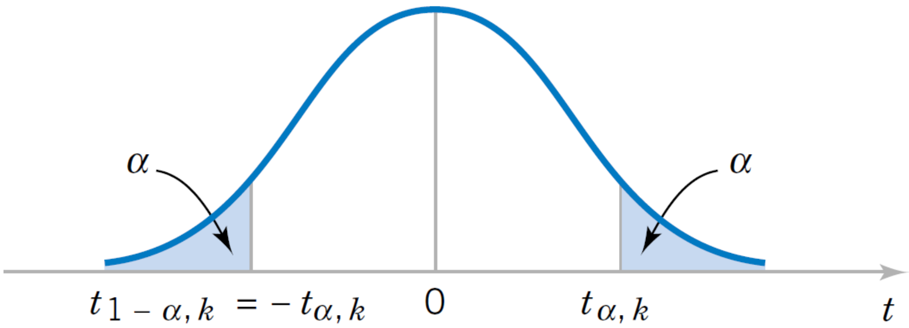
UFNOŚĆ(0.05, 2, 16) = 0.98

Example;

Let $\bar{x}$ = 344 be a sample mean and s =31.13 be a sample standard deviation (n = 10) of a normal distribution **N(μ, σ^2)** with unknown μ and σ. Then, confidence interval is given by limits:

$\bar{x} \pm t(\alpha, n-1)\frac{s}{\sqrt{n-1}}$, where $t(\alpha, n-1)$ is a quartile of Student's distribution.

Calculate limits of confidence interval assuming α=0.05.

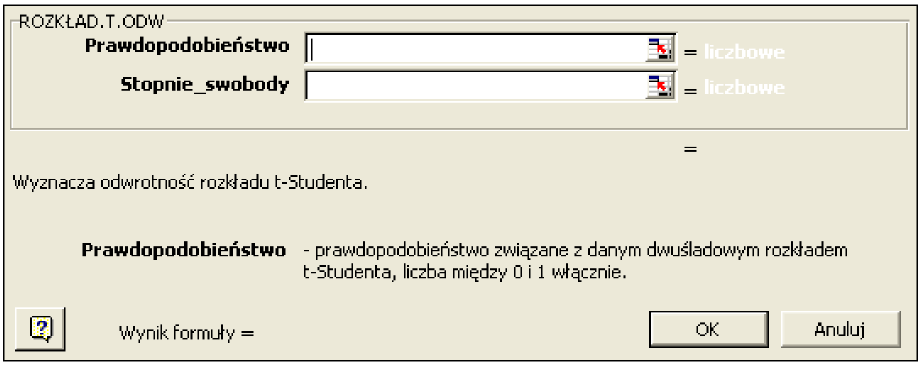


t(0.05, 9) = 2.26

EXCEL: **ROZKŁAD.T.ODW – X of Student's distribution** (turk. TTERS)

prawdopodobieństwo – probability;

stopnie swobody (n-1) – degree of freedeom.



ROZKŁAD.T.ODW(0.05, 9) = = 2.26

Problem:

Parameter **X** has a **N(μ, σ^2)** distribution. How many elements should have a sample for length of a confidential interval = 2L ?

$$2u(1-\frac{\alpha}{2}) < 2L \quad \Rightarrow \quad n > \left(\frac{u(1-\frac{\alpha}{2})\cdot\sigma}{2}\right)^2$$

, where u – density function of **N(0,1).**

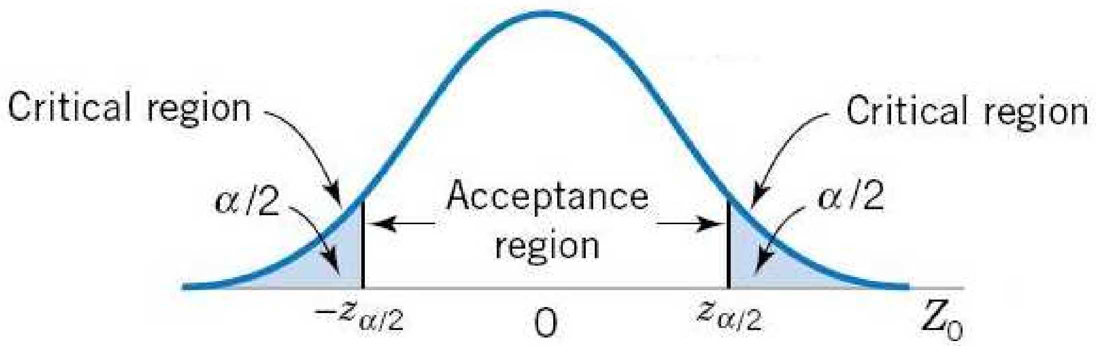
Example;

Let σ = 2, u(0.975) = 1.96. assuming L = 0.5 we get n > 61.

<u>Statistical hypothesis</u>: assumption about some aspects of the statistical behavior of population, related to values of statistical parameters or properties of distribution

Statistical hypothesis testing: methods for making statistical decision using experimented data

The hypothesis testing procedure:
- formulation a null hypothesis $\mathbf{H_0}$ and alternative $\mathbf{H_1}$ hypothesis
- deciding which test is appropriate, and stating the relevant test statistic *T* (usually with known distribution)
- calculate from the observations the observed value of the test statistic *T*
- decide to either **fail to reject** the null hypothesis or **reject** it in favor of the alternative
- the decision rule is to reject the null hypothesis $\mathbf{H_0}$ if the observed is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.



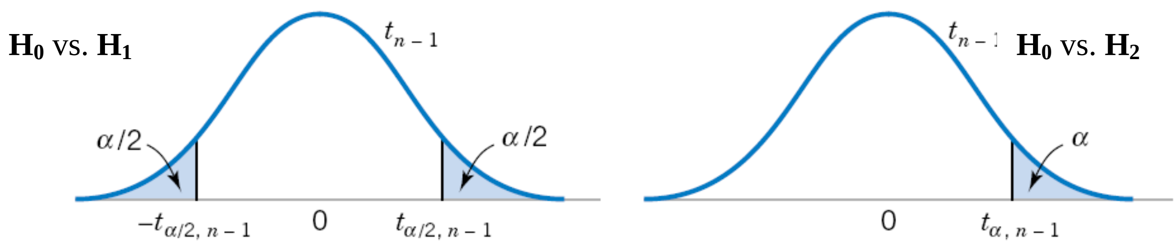
- either the null hypothesis is rejected, or the null hypothesis cannot be rejected at that significance level (which however does not imply that the null hypothesis is *true*).

α (*significance level of a test*): probability of incorrectly rejecting the null hypothesis.

p-value : probability of incorrectly rejecting the null hypothesis (the smallest level of significance that would lead to rejection of the null hypothesis with the given data). One often rejects a null hypothesis if the *p-value* is less than α

<u>Hypothesis Tests on the Mean</u>

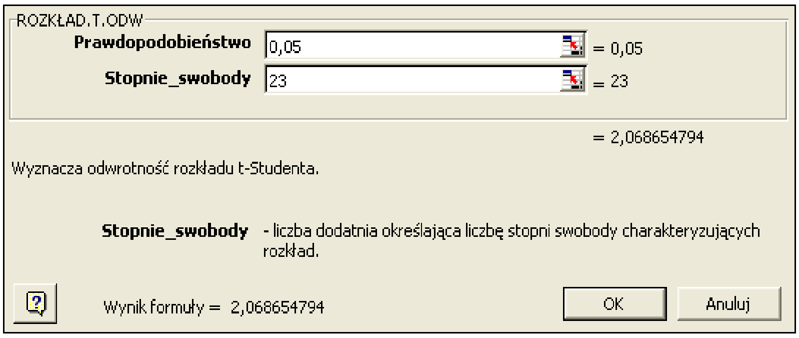
Assume that X has a normal distribution $\mathbf{N(\mu, \sigma^2)}$ with unknown μ, σ and μ_0 is a specified constant. We wish to test the hypotheses:
- null hypothesis $\mathbf{H_0} : \mu = \mu_0$
- alternative hypotheses $\mathbf{H_1} : \mu \neq \mu_0$, $\mathbf{H_2} : \mu > \mu_0$

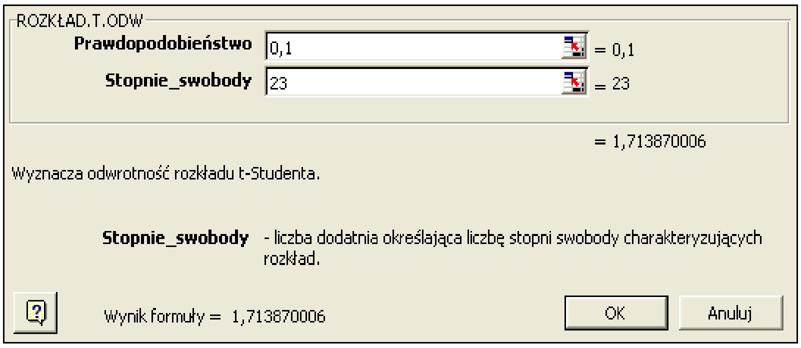


We use a test statistic $t = \frac{\bar{x} - \mu_0}{s}\sqrt{n-1}$ which has a t-Student distribution with *(n-1)* degree of freedom

<u>Example</u> (<u>*hypothesis Tests on the Mean*</u>):

X has a normal distribution **N(μ, σ²).** Let $\bar{x}$ = **13.83** be a sample mean and **s =3.348** be a sample standard deviation (**n = 24**). Assume **μ_0 = 15.5, α=0.05.** We examine null hypothesis **H_0 : μ = μ_0** with alternative hypotheses **H_1 : μ ≠ μ_0** and **H_2 : μ > μ_0**





<u>Solution</u>:

test t = -2.387

- **H_0 : μ = μ_0** vs. **H_1 : μ ≠ μ_0** ⇒

$t(1-\frac{\alpha}{2}, n-1) = 2.068 \Rightarrow$

accept. region = [-2.068, 2.068] ⇒ null hypothesis **is rejected,** the alternative **H_1 : μ ≠ μ_0** hypothesis **is accepted**

- **H_0 : μ = μ_0** vs. **H_1 : μ > μ_0** ⇒

$t(1-\alpha, n-1) = 1.7138 \Rightarrow$

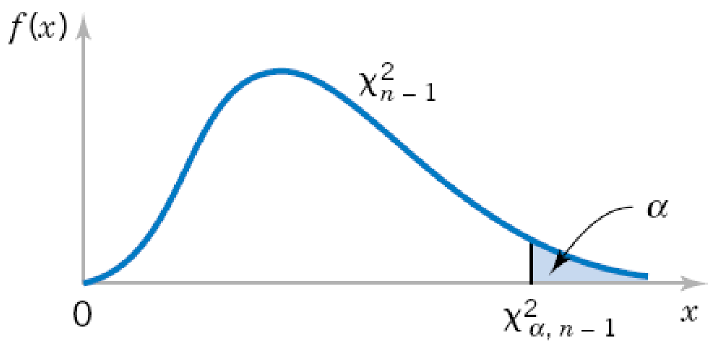
accept. region = [-∞, 1.7138] ⇒ null hypothesis **cannot be rejected**

EXCEL: **ROZKŁAD.T.ODW – X of Student's distribution** (turk. TTERS)
prawdopodobieństwo – probability;
stopnie swobody (n-1) – degree of freedeom.

Pearson's chi-square test: is used to establish whether or not an observed frequency distribution differs from a theoretical distribution. A null hypothesis is formulated that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.

The test statistic asymptotically approaches a χ^2 distribution. The value of the test-statistic is given by $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

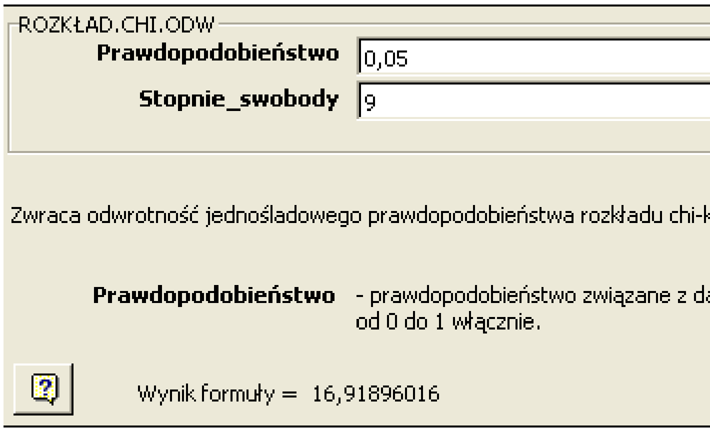
- O_i : an observed frequency;
- E_i : an expected (theoretical) frequency;
- n : the number of outcomes of each event.
- $\chi^2{}_{\alpha,n-1}$: a critical value



Example:

classes	midpoints of classes	observed data	theoretical data	chi-square
4	3,75	0	0,317593	0,317593
4,5	4,25	2	0,932705	1,221308
5	4,75	1	2,184458	0,642238
5,5	5,25	5	4,08009	0,207406
6	5,75	8	6,077459	0,608176
6,5	6,25	5	7,219394	0,682288
7	6,75	7	6,839207	0,00378
7,5	7,25	3	5,166985	0,908813
8	7,75	4	3,113113	0,252663
8,5	8,25	3	1,495819	1,51259
9	8,75	0	0,573178	0,573178
		38	sum :	6,930033

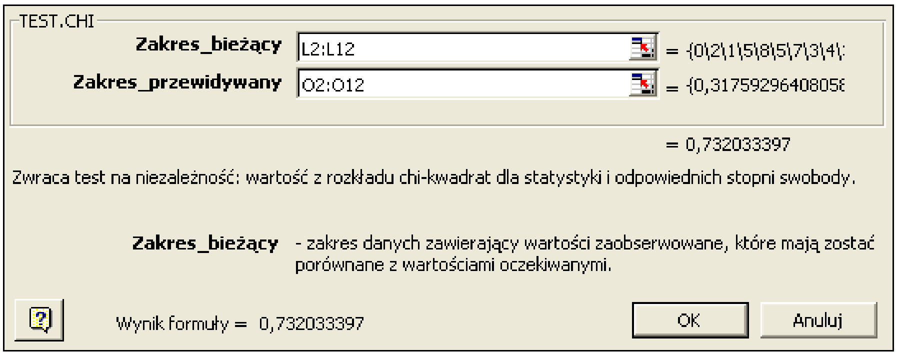
critical value $\chi^2_{\alpha,n-1}$ =16.918



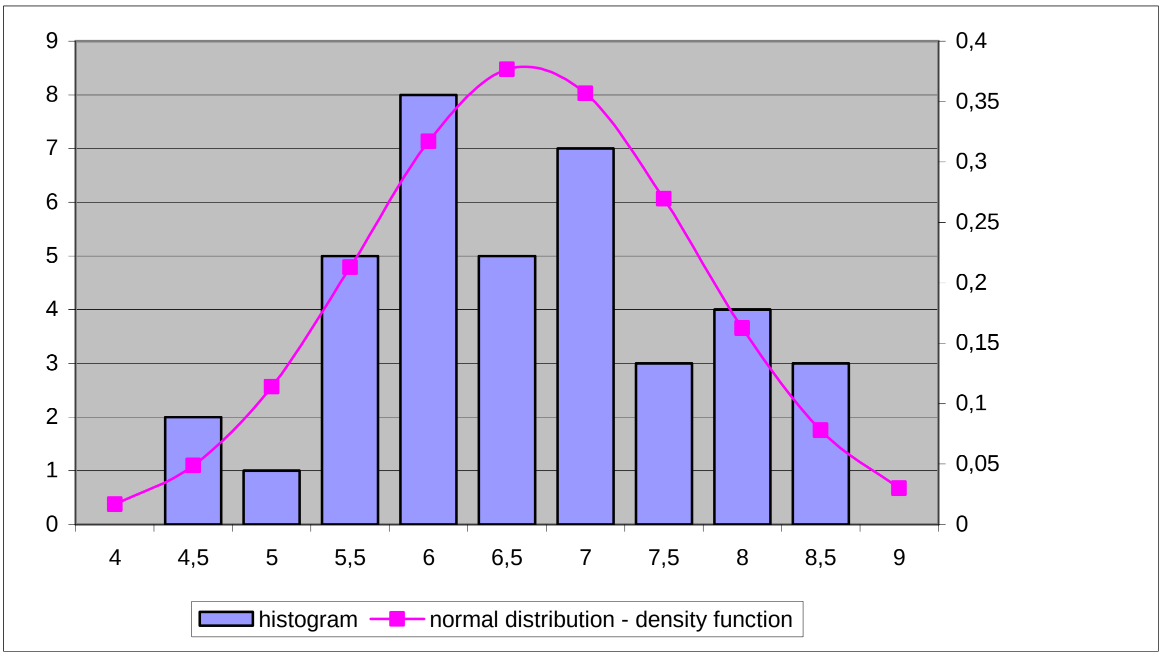
EXCEL: **ROZKŁAD.CHI.ODW – X of χ^2 distribution** (turk.)
prawdopodobieństwo – probability; stopnie swobody (n-1) – degree of freedeom.

Conclusion: the value of the statistic (6,93) is not in a critical region ([16.918, ∞]). we accept a null hypothesis that frequency distribution of observed data in a sample is consistent with a theoretical distribution **N(μ, σ²)** (μ = 6.38, σ = 1.05).

EXCEL: **TEST.CHI – p-value of Person- χ^2 test** (turk.)
zakres bieżący – observed frequency;
zakres przewidywany – theoretical frequency.



p-value = 0.732 (> 0.05)



Regression analysis

- modeling and analysis several variables
- relationship between a dependent variable and independent variables (how the typical value of the dependent variable changes when any one of the independent variable is varied)
- used in prediction
- understand which among the independent variables are related to the dependent variable

Regression model $y = f(x, \beta)$. Method of least squares: $\sum_i (y_i - f(x_i, \beta))^2 \to \min$

R^2 – coefficient of determination : information about the goodness of fit of a regression model . $R^2 \in [0, 1]$.