FIG Working Week 2012
Rome, Italy 6–10 May

Application of non-parametric tests of significance to the market analyses

Anna BARAŃSKA, D.Sc. Ph.D.
AGH University of Science and Technology
Faculty of Mine Surveying and Environmental Engineering
Department of Geomatics

FIG Working Week 2012
Rome, Italy 6–10 May

INTRODUCTION

Common using of parametric tests to elaborate research results is limited by predetermined assumptions, which must be fulfilled. Parametric tests are useless also in the case of the quality data and the data of a purely ordinal nature. In such situations, we use the tests non-parametric. These tests are not dependent on the parameters of population distribution. The calculation formulas are simple, and the calculations do not take much time. Moreover, we use them, when our data may be arranged according to determined criteria and for some random samples of small size. The power of the non-parametric tests (equal to one minus the magnitude of the type 2 error) is however lower than the power of the parametric tests. Then, they are to be applied only in the cases, where we cannot use a parametric test.

Platinum sponsors: esri Trimble

TESTS OF SIGNIFICANCE

Tests of significance, both parametric and non-parametric run in following stages:

- formulation of a zero hypothesis H_0,
- selection of statistics (test function), according to the content of the zero hypothesis H_0 and to the conditions fulfilled by the random sample,
- determination of the significance level of the test α,
- determination of the alternative hypothesis H_1 on the basis of the random test results,
- determination of the limits of a so-called critical area, according to the content of the alternative hypothesis H_1 (its area is equal to the significance level α),
- drawing conclusions based on the position of the statistic value in relation to the critical area.

FIG Working Week 2012
Rome, Italy 6–10 May
Knowing to: Manage the territory, Protect the environment, Evaluate the cultural heritage

During a statistical inference, on the grounds of performed test we can make two types of errors:

- **type 1 error**, α – rejection of the zero hypothesis H_0, while, in fact, it is true,
- **type 2 error**, β – lack of grounds for rejection (or acceptance) of the zero hypothesis H_0, while, in fact, it is false.

Platinum sponsors: esri, Trimble

right-hand critical area

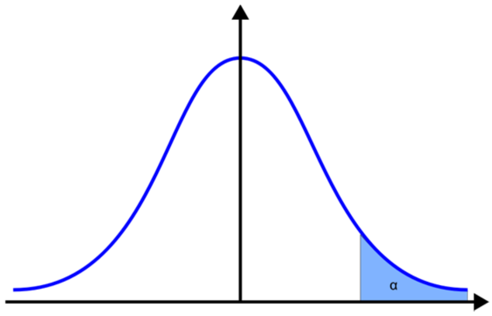

FIG Working Week 2012
Rome, Italy 6–10 May

Among non-parametric tests of significance, we differentiate three types:

- **tests of goodness of fit** – verification of the random variable distribution type (shape), for example, Shapiro-Wilk test or Kołmogorow-Smirnow test,
- **tests of randomness** – verification of the elements randomness in a sample, for example, series test,
- **independence tests** – verification of the independence of two random variables, for example, chi-square independence test.

Platinum sponsors: esri Trimble

TESTS OF GOODNESS OF FIT

The circumstances of applying non-parametric tests of goodness of fit can be as follows:

- as a start point for applying some specific models of parametric tests,
- as one of the elements of the verification of a mathematical model structure, correctness, for example in the case of modelling a real estate market (verification of the model remainders normality distribution),
- a comparison of distributions in two different populations in order to draw conclusions on their similarity,
- other practical issues, like verification of the dice symmetry☺ (does the dice cheat the players at the game?).

FIG Working Week 2012
Rome, Italy 6–10 May
Knowing to: Manage the territory
Protect the environment
Evaluate the cultural heritage

Between non-parametric tests of goodness of fit, we distinguish, the following:

- **Chi-square Pearson test**,
- Kołomogorow test,
- **Kołomogorow-Smirnow test**,
- Kołomogorow-Lillieforse test,
- **Shapiro-Wilk test**,
- Wilcoxon test.

Hypotheses for verification in these tests could be:

H_0: feature X has a distribution F

or: H_0: features X and Y have the same distribution

where:

F - arbitrary determined distribution of probability

Platinum sponsors: esri Trimble

Chi-square test of goodness of fit – run of the test

- classification of the values of the feature X: $x_1, x_2, x_3, \ldots, x_n$ gathered in a random sample (creation of a distributive series:),

No class i	Class boundaries	Class center	Class cardinality n_i
1	(g_0, g_1)	$\bar{x}_1$	n_1
2	(g_1, g_2)	$\bar{x}_2$	n_2
…	…	…	…
k	(g_{k-1}, g_k)	$\bar{x}_k$	n_k
			$\sum_{i=1}^{k} n_i = n$

- formulation of the zero hypothesis H_0: cumulative distribution function of the examined feature is the function $F_0(x)$;

if the hypothesis H_0 is true, the probability p_i that the variable X would take a value belonging to the i-th class (g_{i-1}, g_i) is: $p_i = F_0(g_i) - F_0(g_{i-1})$.

- statistics in this test has the form:

$$\chi_d^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \qquad (1)$$

FIG Working Week 2012
Rome, Italy 6–10 May
Knowing to: Manage the territory, Protect the environment, Evaluate the cultural heritage

Kołomogorow-Smirnow test of goodness of fit

The run of this test in the case of comparing distributions in two random samples is as follows:

- we take random samples from two given populations; we arrange the values from the samples in non-decreasing sequence: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$,
- the zero hypothesis H_0 is: cumulative distribution function in two populations is the same,
- test statistics has the form:

$$D_n = \sqrt{n} \cdot \sup_{-\infty < x < +\infty} \left| F_{n_A}(x) - F_{n_B}(x) \right| \qquad (2)$$

where: $F_{n_{A(B)}}(x) = P_{A(B)}(x_i < x) = \frac{Card\,\{i: x_i < x;\ i = 1,2,..,n_{A(B)}\}}{n_{A(B)}}$, $\quad n = \frac{n_A \cdot n_B}{n_A + n_B}$

(3)

Shapiro-Wilk test of goodness of fit

There are following stages of the test:

- we set the values from the random sample in non-decreasing sequence: $x_{(1)} \leq \ldots \leq x_{(n)}$,
- the zero hypothesis is always H_0: the feature X has a normal distribution,
- the test function has the form:

$$W = \frac{\left(\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i(n)(x_{n-i+1} - x_i)\right)^2}{\sum_{i=1}^{n}(x_i - \hat{x})^2} = \frac{\left(\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i(n)(x_{n-i+1} - x_i)\right)^2}{n \cdot V(X)} \quad (4)$$

where: $\left[\frac{n}{2}\right] = \begin{cases} n/2 & \text{for } n \text{ even} \\ (n-1)/2 & \text{for } n \text{ odd} \end{cases}$

$a_i(n)$ – coefficients from the statistical tables,

$x_{n-i+1} - x_i$ – quasi- intervals of the rank i.

FIG Working Week 2012
Rome, Italy 6–10 May
Knowing to: Manage the territory
Protect the environment
Evaluate the cultural heritage

EXAMPLE – verification of significant differences between the distributions of dwelling prices in districts A and E

Table 1
Unit prices of dwellings in different Krakow districts

District / Number of dwelling	District A Olsza [€/m²]	District B Dębniki [€/m²]	District C Kurdwanów [€/m²]	District D Ruczaj [€/m²]	District E Krowodrza [€/m²]
1	1939	1795	1335	1410	1600
2	1501	1923	1652	1676	1828
3	1245	1724	1271	2070	1114
4	1364	2004	1454	1341	1277
5	1561	2024	1758	1243	1543
6	1420	1756	1572	1393	1470
7	1201	2067	1253	1102	1575
8	1579	1752	1449	1258	1412
9	1350	1825	1305	1204	1431
10	1420	1956	1425	1358	1220
11	1354	2011	1458	1426	1354
12	1510	1780	1589	1208	1750
13	1412	1842	1654	1520	1654
14	1600	1954	1687	1620	1541
15	1735	1820	1420	1820	1412
Mean value	**1479**	1882	1485	1443	**1479**
Standard deviation	**182**	110	154	252	**187**

Platinum sponsors: esri Trimble

Table 2
Calculations in Kołomogorow-Smirnow test

$x_{(i)}$ [€/m²]	District A			District E			$\lvert P_A - P_E \rvert$
	$n_{i(A)}$	cumulative $N_{i(A)}$	$P_A(x_i<x)= =N_{i(A)}/n_A$	$n_{i(E)}$	cumulative $N_{i(E)}$	$P_E(x_i<x)= =N_{i(E)}/n_E$	
1114	0	0	0,000	1	1	0,067	0,067
1201	1	1	0,067	0	1	0,067	0,000
1220	0	1	0,067	1	2	0,133	0,067
1245	1	2	0,133	0	2	0,133	0,000
1277	0	2	0,133	1	3	0,200	0,067
1350	1	3	0,200	0	3	0,200	0,000
1354	1	4	0,267	1	4	0,267	0,000
1364	1	5	0,333	0	4	0,267	0,067
1412	1	6	0,400	2	6	0,400	0,000
1420	2	8	0,533	0	6	0,400	0,133
1431	0	8	0,533	1	7	0,467	0,067
1470	0	8	0,533	1	8	0,533	0,000
1501	1	9	0,600	0	8	0,533	0,067
1510	1	10	0,667	0	8	0,533	**0,133**
1541	0	10	0,667	1	9	0,600	0,067
1543	0	10	0,667	1	10	0,667	0,000
1561	1	11	0,733	0	10	0,667	0,067
1575	0	11	0,733	1	11	0,733	0,000
1579	1	12	0,800	0	11	0,733	0,067
1600	1	13	0,867	1	12	0,800	0,067
1654	0	13	0,867	1	13	0,867	0,000
1735	1	14	0,933	0	13	0,867	0,067
1750	0	14	0,933	1	14	0,933	0,000
1828	0	14	0,933	1	15	1,000	0,067
1939	1	15	1,000	0	15	1,000	0,000
						max=	**0,133**

FIG Working Week 2012

On the basis of the calculation above, we calculate a final statistic value (2):

$$D_n = 0{,}133 \cdot \sqrt{n} = 0{,}133 \cdot \frac{15 \cdot 15}{15 + 15} = 0{,}365$$

which, in comparison with appropriate critical test value, satisfies the condition: $D_n < 1{,}36$. It allows concluding that there are no grounds for rejecting the zero hypothesis, thus we can admit that both compared local dwelling markets have the same price distribution in analysed time. They are therefore actually very similar.

CONCLUSION

- Non-parametric tests are applied to investigate or compare the shape of the random variable distribution.
- They are especially useful in the case of variables expressed in ordinal scale.
- We distinguish among them tests of conformity, tests of randomness and independence tests.
- They have lower power than the parametric tests, so, they facilitate the acceptance of the zero hypothesis, which, in fact, is false. Therefore, they need generally more data (larger random sample) than the parametric tests.
- Non-parametric tests constitute often a preliminary stage of applying parametric tests.
- The stages of a non-parametric test of significance are usually equivalent for the stages of a parametric test of significance.

FIG Working Week 2012
Rome, Italy 6–10 May
Knowing to: Manage the territory, Protect the environment, Evaluate the cultural heritage

Thank you for your attention ☺

Platinum sponsors: esri Trimble