Practical application of the Wilcoxon-Mann-Whitney test in valuation.

Selection of attributes as pricing factors based on the principle of unbiased estimates

K. A. Murashev

August 3, 2022

In their practice appraisers often face the need to take into account differences in quantitative and qualitative characteristics of objects. In particular, one of the standard tasks is to determine the attributes that influence the cost (so-called "pricing factors") and to separate them from the attributes that do not or cannot be determined.

Subjective selection of attributes taken into account in determining the value is widespread in valuation practice. In this case, specific quantitative indicators of the impact of these attributes on the cost are often taken from the so-called "reference books". While not denying the speed and low cost of this approach, it should be recognized that only data directly observed in the open markets is a reliable basis for a value judgment. The priority of such data over other data, in particular those obtained by expert survey, is enshrined, among others, in RICS Valuation — Global Standards 2022 [7], International Valuation Standards 2022 [8], as well as in IFRS 13 "Fair Value Measurement" [6]. Therefore, we can say that mathematical methods for analyzing data from the open market are the most reliable means of interpreting market information used in market research and predicting the value of individual objects.

The aim of this work is to justify the necessity and possibility of using a rigorous mathematical Wilcoxon–Mann–Whitney test, which allows us to answer the question about the necessity of taking into account the binary attribute as a price-generating factor. Instead of the judgmental approach, which is most commonly used by appraisers in selecting the attributes to be considered in appraisal, this paper proposes the idea of prioritizing the measuring approach based on the results of a mathematical test that allows to draw a conclusion about the importance or otherwise of the binary attribute influence on the value. It should be noted that despite the fact that the statistical test under consideration belongs to frequentist statistics, it, through its connection to ROC analysis and AUC, is related to modern machine learning methods, which will be discussed later in the text of this material. The presence of this relationship and elements of Bayesian statistics seems particularly interesting and promising from the point of view of introducing machine learning and data analysis methods into the everyday practice of appraisers.

Users should have some general math background and basic Python and R programming skills to understand and practice all of the material in the text, but lack of that knowledge and skill is not a barrier to learning most of the material and implementing the test in the spreadsheet that comes with it.

The material consists of four blocks:

- a description of the Wilcoxon–Mann–Whitney test (hereafter "U-test"), its probabilistic meaning, and its relationship to other mathematical methods;
- a practical implementation of the U-test in a spreadsheet on an example of test random data;

- practical implementation of the U-test on the real data of the residential real estate market of St. Petersburg agglomeration by means of Python programming language, the purpose of the analysis was to check the significance of the difference in the unit price between the objects located in the urban and suburban parts of the agglomeration;
- practical implementation of the U-test on real data of residential real estate market of Almaty by means of R programming language, the purpose of the analysis was to check the significance of difference in unit price between the objects sold without demountable improvements and the objects sold with them.

The current version of this material, its source code, Python and R scripts, and the spreadsheet are in the repository on the GitHub portal and are available at the permanent link [9].

This material and all of its appendices are distributed under the terms of the cc-by-sa-4.0 license [10].

Contents

List of Tables

List of Figures

Listings

1 Technical details

This material, as well as the appendices to it, are available at permanent link [9]. The source code for this work was created using the language TEX [24] with a set of macro extensions LATEX 2_ε [25], distribution TeXLive [26] and Editor TeXstudio [34]. The spreadsheet calculation was done with LibreOffice Calc [12] (Version: 7.3.4. 2 / LibreOffice Community Build ID: 30(Build:2); CPU threads: 4; OS: Linux 5.11; UI render: default; VCL: kf5 (cairo+xcb) Locale: en-US (en_US.UTF-8); UI: en-US Ubuntu package version: 1:7.3.4 rc2-0ubuntu0.20.04.1 lo1; Calc: threaded). The calculation in R [27] (version 4.2.1 (2022-06-23) – "Funny-Looking Kid") was done using an IDE RStudio (RStudio 2022. 02.3+492 "Prairie Trillium" Release (1db809b8, 2022-05-20) for Ubuntu Bionic; Mozilla/5.0 (X11; Linux x86_64); AppleWebKit/537.36 (KHTML, like Gecko); QtWebEngine/5.12.8; Chrome/69.0.3497.128; Safari/537.36) [23]. The calculation in Python (Version 3.9.12) [11] was performed using the development environment Jupyter Lab (Version 3.4.2) [18] and IDE Spyder (Spyder version: 5.1.5 None* Python version: 3.9.12 64-bit * Qt version: 5.9.7 * PyQt5 version: 5.9.2 * Operating System: Linux 5.11.0-37-generic) [22]. The graphics used in the subsection **??** were prepared using Geogebra (Version 6.0.666.0-202109211234) [13]. The following values were used in this material as well as in most of the works in the series:

- significance level: $\alpha = 0.05$;
- confidence interval: $Pr = 0.95$;
- initial position of the pseudo-random number generator: $seed = 19190709$.

A dot is used as a decimal point. Most of the mathematical notations are written as they are used in English-speaking circles. For example, a tangent is written as tan, not tg. The results of statistical tests are considered significant when

$$p \leq \alpha. \tag{1.1}$$

This decision is based, in part, on the results of the discussion that took place on researchgate.net [17].

2 Subject of research

When working with market data, the appraiser is often faced with the task of testing the hypothesis of whether a quantitative, ordinal or nominal attribute has a significant effect on the price. Real estate market analysts, developers, realtors, employees of collateral departments of banks, leasing and insurance companies, tax inspectors and other specialists have a similar task. At the same time, it is often impossible to collect large amounts of data that would allow a wide range of machine learning methods to be applied. In some cases appraisers consciously narrow the area of data collection to the narrow market segment, resulting in only very small samples of less than thirty observations at their disposal. In this case, the price data most often has a distribution that differs from the normal one. In this case, a rational solution is to use U-test. Let us formulate the problem:

- suppose that we have two samples of unit prices for commercial premises, some of which have some attribute (e. g., having a separate entrance) and some of which do not;
- it is necessary to determine whether the presence of this feature has a significant impact on the unit value of this type of real estate or not.

At first glance, according to established practice, an appraiser can simply subjectively recognize some attributes as significant and others as not, and then accept the adjustment values for differences in these attributes from the reference books. However, as mentioned above, this approach is hardly considered best practice because it lacks any market analysis. Also, in that case, it is unlikely that such work is of any serious value at all.

Instead, it is possible to use random samples of market data and apply mathematical analysis to them, allowing scientific and evidence-based conclusions to be drawn about the significance of a particular attribute's impact on value. The data used in this paper to perform the U-test using Python and R are real market data, some of which were collected by the author through web scraping and some provided by colleagues for the analysis. The attached spreadsheet is set up so that test raw data can be generated in a pseudo-random fashion.

The subject of this paper is the nonparametric Wilcoxon-Mann-Whitney test, specifically designed for samples that have a distribution other than normal. This circumstance is important because the price data that appraisers deal with most often have this distribution, which excludes the possibility of applying the parametric t-criterion and z-criterion. In addition, the test under consideration is of great interest because it has a connection to machine learning methods through AUC, the calculation of which through the formula provided in the test framework gives a value equal to that calculated by ROC analysis. Thus, the study of the U-test paves the way for a further dive into the world of machine learning, which is entering many areas of human activity and will significantly change the field of value estimation in the foreseeable future.

The material contains a description of the test and instructions for performing it, sufficient in the author's opinion for its demonstrable use in the estimation process.

3 Basic information about the test

3.1 Assumptions and formalization of hypotheses

First of all, it should be said that, in spite of the stated common name, it is more correct to speak of two tests:

- Wilcoxon rank-sum test developed by Frank Wilcoxon in 1945 [21];
- Mann–Whitney U-test which is a further development of the aforementioned criterion developed by Henry Mann and Donald Whitney in 1947 [19].

Looking ahead we can say that the statistics of these criteria are linearly related and their p-values are almost the same which from a practical point of view allows us to talk about variations of one test rather than two separate tests. This paper uses the common name throughout the text, as well as a shortened version of "U-test" which historically refers to the Mann-Whitney test. Some authors[4] recommend using the Wilcoxon rank-sum test when there are no assumptions about variance, and the Mann-Whitney U-test when variance of the two samples are equal. However, the experimental data indicate that the Wilcoxon rank-sum test and Mann-Whitney U-test values are essentially the same when the variance of the samples is significantly different. Adhering to the KISS principle [29] underlying the entire series of publications, the author concludes that a unified approach is possible. Also remember that the Wilcoxon signed-rank test is a separate test designed to analyze differences between two matched samples, whereas the Mann-Whitney U-test discussed in this paper is designed to work with two independent samples.

Suppose that there are two samples:

$$x^m = (x_1, x_2, \ldots, x_m), x_i \in \mathbb{R}; \quad y^n = (y_1, y_2, \ldots, y_n), y_i \in \mathbb{R} \quad : m \leq n.$$

- Both samples are simple and random (i.e., SRS [32]), the combined sample is independent.
- The samples are taken from unknown continuous distributions *F(x)* and *G(y)*, respectively.

Simple random sample (SRS) — is a subset of individuals (*a sample*) chosen from a larger set (*a population*) in which a subset of individuals are chosen randomly, all with the same probability. It is a process of selecting a sample in a random way. In **SRS**, each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals.A simple random sample is an unbiased sampling technique. Equivalent definition: a sample

$x^m = (x_1, x_2, \ldots, x_m)$ is simple if the values $(x_1, x_2, \ldots, x_m)$ are realizations of m independent equally distributed random variables. In other words, the selection of observations is not only random but also does not imply any special selection rules (e.g., choosing every 10th observation).

The U-test — is a nonparametric criterion to test the null hypothesis that for randomly chosen from two samples of observations $x \in X$ and $y \in Y$ the probability that x is greater than y is equal to the probability that y is greater than x. In mathematical language, the null hypothesis is written as follows

$$H_0 : P\{x < y = \frac{1}{2}\}. \tag{3.1}$$

For the test's own consistency, an alternative hypothesis is required, which is that the probability that the value of a characteristic of observation from X is greater than that of observation from Y differs upward or downward from the probability that the value of a characteristic of observation from Y is greater than that of observation from X. In mathematical language, the alternative hypothesis is written as follows

$$H_1 : P\{x < y\} \neq P\{y < x\} \vee P\{x < y\} + 0.5 \cdot P\{x = y\} \neq 0.5. \tag{3.2}$$

According to the basic concept of the U-test, if the null hypothesis is true, the distribution of the two samples is continuous; if the alternative hypothesis is true, the distribution of one sample is stochastically greater than the distribution of the other. In this case, it is possible to formulate a number of null and alternative hypotheses for which this test will give a correct result. His most extensive generalization lies in the following assumptions:

- the observations in both samples are independent;
- the data type is at least ranked, i. e., with respect to any two observations you can tell which one is greater;
- the null hypothesis assumes that the distributions of the two samples are equal;
- the alternative hypothesis assumes that the distributions of the two samples are unequal.

With a stricter set of assumptions than those given above, for example the assumption that the distribution of the two samples is continuous if the null hypothesis is valid and that the distribution of the two samples has a shift in the distribution if the alternative one is valid i. e. $f_1(x) = f_2(x + \sigma)$, we can say that the U-test is a test for the hypothesis of equality of medians. In this case, the U-test can be interpreted as a test of whether Hodges–Lehman's estimate of the difference in central tendency measures differs from zero. In this situation, the Hodges–Lehman estimate is the median of all possible values of differences between the observations in the first and second samples. However, if both the variance and the shape of the distribution of the two samples differ, the U-test cannot correctly test the medians. Examples can be shown where the medians are numerically equal and the test rejects the null hypothesis because of the small p-value. Thus, a more correct interpretation of the U-test is to use it to test the shift hypothesis [20].

Shift hypothesis — is a statistical hypothesis often considered as an alternative to the hypothesis of complete homogeneity of samples. Let us have two samples of data. Let us also give two random variables X and Y, which are distributed as elements of these samples and have distribution functions *F(x)* and *G(y)*, respectively. In these terms, the shift hypothesis can be written as follows

$$H : F(x) = G(x + \sigma) \ : \forall x, \ \sigma \neq 0. \tag{3.3}$$

In this case, the U-criterion is valid regardless of the characteristics of the samples.

Simply put, the essence of the U-test is that it allows us to answer the question of whether there is a significant difference in the value of the quantitative attribute of the two samples. With regard to valuation, we can say that the use of this test helps to answer the question of whether it is necessary to take into account one or another attribute as a price-generating factor. It follows from the above that by default we are talking about a two-sided test. In practice, this means that the test does not give a direct answer to the question, for example: "Is there a significant excess of the unit value of premises with a separate entrance to the premises that do not have it. At the same time, there are also one-sided realizations that allow us to answer the question about the sign of the difference in the value of the attribute in the two samples.

In addition to the above requirements for the samples themselves, the conditions for applying the U-test are:

- the distribution of quantitative attribute values of samples is different from normal (otherwise it is advisable to use parametric Student's t-test or z-test for independent samples).
- at least three observations in each sample, it is allowed to have two observations in one of the samples, provided that there are at least five in the other sample.

To summarize the above, there are three variants of the null hypothesis, depending on the level of rigor outlined in the table below 3.1.

3.2 Test implementation

3.2.1 Test statistic

Suppose that the elements $x_1, \ldots, x_n$ represent a simple independent sample from $X \in \mathbb{R}$, and the elements $y_1, \ldots, y_n$ represent a simple independent sample from $Y \in \mathbb{R}$ and the samples are independent of each other. Then the relevant U-statistic is defined

Table 3.1: Variants of the null hypothesis when using the U-test in valuation.

Type of hypothesis	Formulation
Scientific	The two samples are completely homogeneous, i. e. they belong to the same distribution, there is no shift and the estimate made for the first sample is unbiased for the second one.
Practical	The medians of the two samples are equal to each other.
Set forth in terms of valuation	The difference in the attribute between the two samples of object-analogues is not significant, its accounting is not required and this attribute is not a pricing factor.

as follows:

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} S(x_i, y_j),$$

$$S(x,y) = \begin{cases} 1, & x > y, \\ \frac{1}{2}, & x = y, \\ 0, & x < y. \end{cases} \tag{3.4}$$

3.2.2 Calculation methods

The test involves calculating a statistic usually called the U-statistic whose distribution is known if the null hypothesis is true. When working with very small samples, the distribution is specified tabularly; when the sample size is more than twenty observations, it is approximated quite well by the normal distribution. There are two methods of calculating U-statistics: manual calculation using the formula 3.4 or using a special algorithm. The first method, due to its labor-intensive nature, is only suitable for very small samples. The second method can be formalized as a step-by-step set of instructions and will be described below.

1. You must construct a common variation series for the two samples and then assign a rank to each observation, starting with one for the smallest of them. If there are ties, i. e. groups of repeating values (such a group can be, e. g., only two equal values), each observation from such a group is assigned a value equal to the median of the group ranks before adjustment (for example, in the case of a variation series (*3, 5, 5, 5, 5, 8*) the ranks before adjustment are (*1, 2, 3, 4, 5, 6*) after — (*1, 3.5, 3.5, 3.5, 3.5, 6*).

2. It is necessary to calculate the sums of the ranks of the observations of each sample, denoted as R_1, R_2 respectively. In this case, the total sum of ranks can be calculated by the formula

$$R = \frac{N(N+1)}{2}, \tag{3.5}$$

where N —the total number of observations in both samples.

3. Next, we calculate the U-value for the first sample:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \tag{3.6}$$

where R_1 —the sum of ranks of the first sample, n_1 — the number of observations in the first sample.

4. The U-value for the second sample is calculated in the same way:

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}, \tag{3.7}$$

where R_2 —the sum of ranks of the second sample, n_2 — the number of observations in the second sample.

From the above formulas it follows that

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}. \tag{3.8}$$

It is also known that

$$\begin{cases} R_1 + R_2 = \dfrac{N(N+1)}{2} \\ N = n_1 + n_2. \end{cases} \tag{3.9}$$

Then

$$U_1 + U_2 = n_1 n2. \tag{3.10}$$

Using this formula as a control ratio can be useful for checking the correctness of calculations in a spreadsheet processor.

5. From the two values of U_1, U_2 in all cases we choose the smaller which will be the U-statistic and used in further calculations. Let us denote it as U.

3.2.3 Interpretation of the result

For a correct interpretation of the test result it is necessary to specify:

- size of each sample;
- values of the measure of central tendency for each sample (given the nonparametric nature of the test, the median appears to be the appropriate measure of central tendency);
- the value of the U-statistic itself;
- the CLES index [28] the value of which is equivalent to the AUC and ρ-statistic;
- rank-biserial correlation coefficient (RBC) [30];

- the accepted level of significance (usually 0.05);
- the calculated p-value.

The concept of U-statistic was discussed earlier and most of the other indicators are widely known and do not require any particular consideration.

3.2.3.1 CLES = ρ-statistic = AUC

First of all, it must be said that all of these indicators are equivalent to each other. Thus

$$CLES = f = AUC_1 = \rho. \tag{3.11}$$

3.2.3.1.1 Common language effect size (CLES)

Common language effect size (CLES) — is the probability that the value of a randomly chosen observation from the first sample is greater than the value of a randomly chosen observation from the second sample. This indicator is calculated by the formula

$$CLES = \frac{U_1}{n_1 n_2}. \tag{3.12}$$

The designation *f (favorable)* is often used instead of *CLES*. This sample value is an unbiased estimate of the value for the entire population of objects belonging to the set.

It should be noted that the value and meaning of this indicator is equivalent to the value and meaning of the AUC[31]. Thus, we can say that this indicator characterizes the quality of the U-test as a binary classifier.

$$CLES = f = AUC_1 = \frac{U_1}{n_1 n_2}. \tag{3.13}$$

The relationship between the U-statistic and AUC is discussed in **??**.

3.2.3.1.2 ρ-statistic A statistic called ρ that is linearly related to U and widely used in studies of categorization (discrimination learning involving concepts), and elsewhere, is calculated by dividing U by its maximum value for the given sample sizes, which is simply $n1 \times n2$. Thus, ρ is a non-parametric measure of the overlap between two distributions; it can take values between 0 and 1, and it is an estimate of $P(Y > X) + 0.5P(Y = X)$, where X and Y are randomly chosen observations from the two distributions. Both extreme values represent complete separation of the distributions, while a $= 0.5$ represents complete overlap. This statistic is useful in particular when despite a large p-value the medians of the two samples are essentially equal to each other.

3.2.3.2 Rank-biserial correlation (RBC)

The method of representing the measure of impact for the U-test is to use a measure of rank correlation known as rank-biserial correlation (hereafter RBC). As in the case of other measures of correlation, the value of the RBC coefficient has a range of values [-1;1], with a zero value indicating the absence of any relationship. The RBC coefficient is usually denoted as r. A simple formula based on the CLES (AUC, t, ρ) value is used to calculate it. Let us state the hypothesis that in a pair of random observations, one of which is taken from the first sample and the other from the second, the value of the first is greater.Let's write it down in mathematical language:

$$H : x_i > y_j, \quad x \ \in X, \ y \in Y. \tag{3.14}$$

Then the value of the RBC coefficient is the difference between the proportion of random pairs of observations that are favorable (f) to the hypothesis and the complementary proportion of random pairs that are unfavorable to the hypothesis. Thus, this formula is a formula for the difference between the CLES scores for each of the groups.

$$r = f - u = CLES_1 - CLES_2 = f - (1 - f) \tag{3.15}$$

There are also a number of alternative formulas that give identical results:

$$r = 2f - 1 = \frac{2U_1}{n_1 n_2} - 1 = 1 - \frac{2U_2}{n_1 n_2}. \tag{3.16}$$

3.2.4 Calculation of the p-value and the final test of the null hypothesis

If the number of observations in both samples is large enough, the U-statistic has an approximately normal distribution. Then its standardized value (z-score) can be calculated by the formula

$$z = \frac{U - m_U}{\sigma_U}, \tag{3.17}$$

where m_U is mean for U and σ_U is its standard deviation. A visualization of the concept of *standardized value for a normal distribution* is shown in Figure 3.1. The mean for the U is calculated by the formula

$$m_U = \frac{n_1 n_2}{2}. \tag{3.18}$$

The formula for the standard deviation in the case of no ties is as follows:

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}. \tag{3.19}$$

In case of the presence of tied ranks, a different formula is used:

$$\sigma_{U_{ties}} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum_{k=1}^{K} ({t_k}^3 - t_k)}{12 n (n-1)}} = \sqrt{\frac{n_1 n_2}{12}\left((n+1) - \frac{\sum_{k=1}^{K} ({t_k}^3 - t_k)}{n(n-1)}\right)}, \tag{3.20}$$

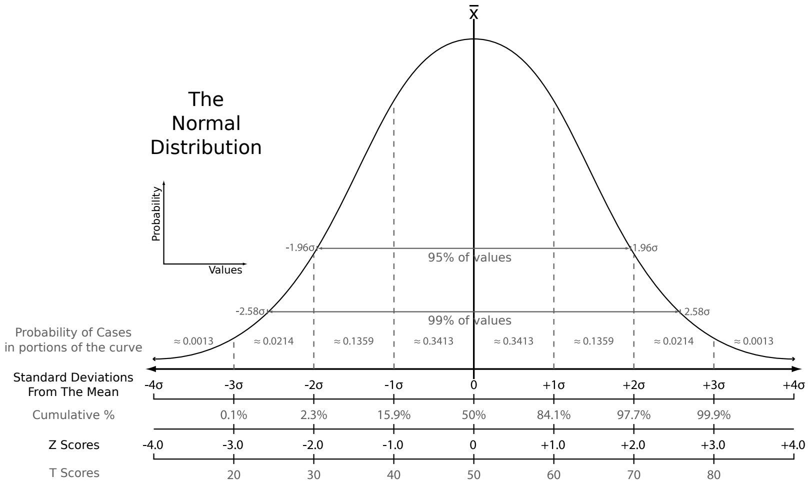

Figure 3.1: A visualization of the concept of standardized value for a normal distribution [33]

where t_k is the number of observations with rank k and K is the total number of tied ranks. Then, by obtaining a standardized value (z-score) and using an approximation of the standard normal distribution, the p-value for a given level of significance (usually 0.05) is calculated. The interpretation of the result is as follows:

$$\begin{aligned} &p \leq 0.05 \Rightarrow \text{the null hypothesis is rejected} \\ &p > 0.05 \Rightarrow \text{the null hypothesis can not be rejected.} \end{aligned} \tag{3.21}$$

However, there is also an alternative interpretation:

$$\begin{aligned} &p < 0.05 \Rightarrow \text{the null hypothesis is rejected} \\ &p \geq 0.05 \Rightarrow \text{the null hypothesis can not be rejected.} \end{aligned} \tag{3.22}$$

To date, there is no unambiguous position on how the situation when $p = \alpha$ should be interpreted. This paper uses the version described in 3.21.

Bibliography

[1] Donald R. Mann Henry B.and Whitney. “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics* (1947), pp. 50–60.

[2] Michael C Mozer. *Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic*. English. 2003.

[3] Corinna Cortes and Mehryar Mohri. “AUC optimization vs. error rate minimization”. In: *Advances in neural information processing systems* 16.16 (2004), pp. 313–320.

[4] . . . . 2006.

[5] Dennis D. Boos and L. A. Stefanski. *Essential Statistical Inference*. English. Springer, 2013.

[6] The IFRS Foundation. *IFRS 13 Fair Value Measurement*. UK, London: The IFRS Foundation, 2016-01-31. URL: `http://eifrs.ifrs.org/eifrs/bnstandards/en/IFRS13.pdf` (visited on 06/10/2020).

[7] Royal Institution of Chartered Surveyors (RICS). *RICS Valuation — Global Standards*. English. UK, London: RICS, 2021-11-30. URL: `https://www.rics.org/uk/upholding-professional-standards/sector-standards/valuation/red-book/red-book-global/` (visited on 05/11/2022).

[8] International Valuation Standards Council. *International Valuation Standards*. 2022-01-31. URL: `https://www.rics.org/uk/upholding-professional-standards/sector-standards/valuation/red-book/international-valuation-standards/`.

[9] K. A. Murashev. “Practical application of the Mann-Whitney-Wilcoxon test (U-test) in valuation”. English, Spanish, Russian, Interslavic. In: (2022-05-15). URL: `https://github.com/Kirill-Murashev/AI_for_valuers_book/tree/main/Parts&Chapters/Mann-Whitney-Wilcoxon` (visited on 05/15/2022).

[10] Creative Commons. *cc-by-sa-4.0*. URL: `https://creativecommons.org/licenses/by-sa/4.0/` (visited on 01/27/2021).

[11] Python Software Foundation. *Python site*. . Python Software Foundation. URL: `https://www.python.org/` (visited on 08/17/2021).

[12] The Document Foundation. *LibreOffice Calc*. . URL: `https://www.libreoffice.org/discover/calc/` (visited on 08/20/2021).

[13] *GeoGebra official site*. URL: `https://www.geogebra.org/` (visited on 08/26/2021).

[14] Bob Horton. *AUC Meets the Wilcoxon-Mann-Whitney U-Statistic*. URL: https://blog.revolutionanalytics.com/2017/03/auc-meets-u-stat.html (visited on 07/14/2022).

[15] Bob Horton. *Calculating AUC: the area under a ROC Curve*. URL: https://blog.revolutionanalytics.com/2016/11/calculating-auc.html (visited on 07/14/2022).

[16] Bob Horton. *ROC Curves in Two Lines of R Code*. URL: https://blog.revolutionanalytics.com/2016/08/roc-curves-in-two-lines-of-code.html (visited on 07/14/2022).

[17] *If p-value is exactly equal to 0.05, is that significant or insignificant?* URL: https://www.researchgate.net/post/If_p-value_is_exactly_equal_to_005_is_that_significant_or_insignificant.

[18] *Jupyter site*. URL: https://jupyter.org (visited on 05/13/2022).

[19] Machinelearning.ru. *U- -* . Russian. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%A3%D0%B8%D0%BB%D0%BA%D0%BE%D0%BA%D1%81%D0%BE%D0%BD%D0%B0-%D0%9C%D0%B0%D0%BD%D0%BD%D0%B0-%D0%A3%D0%B8%D1%82%D0%BD%D0%B8 (visited on 05/14/2022).

[20] Machinelearning.ru. . URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%93%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0_%D1%81%D0%B4%D0%B2%D0%B8%D0%B3%D0%B0 (visited on 05/15/2022).

[21] Machinelearning.ru. . Russian. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%A3%D0%B8%D0%BB%D0%BA%D0%BE%D0%BA%D1%81%D0%BE%D0%BD%D0%B0_%D0%B4%D0%B2%D1%83%D1%85%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BE%D1%87%D0%BD%D1%8B%D0%B9 (visited on 05/14/2022).

[22] *Spyder IDE site*. URL: https://www.spyder-ide.org/.

[23] PBC Studio. *RStudio official site*. . URL: https://www.rstudio.com/ (visited on 08/19/2021).

[24] CTAN team. *TeX official site*. English. CTAN Team. URL: https://www.ctan.org/ (visited on 11/15/2020).

[25] LaTeX team. *LaTeX official site*. English. URL: https://www.latex-project.org/ (visited on 11/15/2020).

[26] *TeXLive official site*. URL: https://www.tug.org/texlive/ (visited on 11/15/2020).

[27] The R Foundation. *The R Project for Statistical Computing*. . The R Foundation. URL: https://www.r-project.org/ (visited on 08/17/2021).

[28] Wikipedia. *Commom Language Effect Size*. English. URL: https://en.wikipedia.org/wiki/Effect_size#Common_language_effect_size (visited on 05/16/2022).

[29] Wikipedia. *KISS principle*. URL: https://en.wikipedia.org/wiki/KISS_principle (visited on 11/06/2020).

[30] Wikipedia. *Rank-biserial correlation*. English. URL: https://en.wikipedia.org/wiki/Effect_size#Rank-biserial_correlation (visited on 05/16/2022).

[31] Wikipedia. *Receiver operating characteristic*. URL: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (visited on 05/17/2022).

[32] Wikipedia. *Simple random sample*. URL: https://en.wikipedia.org/wiki/Simple_random_sample (visited on 07/26/2022).

[33] Wikipedia. *Standard score*. URL: https://en.wikipedia.org/wiki/Standard_score (visited on 05/17/2022).

[34] Benito van der Zander. *TeXstusio official site*. URL: https://www.texstudio.org/ (visited on 11/15/2020).