Практическое применение критерия Манна”– Уитни”– Уилкоксона в оценочной деятельности

К. А. Мурашев

31 мая 2022 г.

В своей практике оценщики часто сталкиваются с необходимостью учёта различий количественных характеристик объектов. В частности, одной из стандартных задач является установление признаков, влияющих на стоимость (т. н. ценообразующих факторов) и их отделение от признаков, влияние которых на стоимость отсутствует либо не может быть установлено. В практике оценки широкое распространение получил субъективный отбор признаков, учитываемых при определении стоимости. При этом конкретные количественные показатели влияния этих признаков на стоимость зачастую берутся из т. н. «справочников». Не отказывая такому подходу в быстроте и невысокой стоимости его реализации, нельзя не признать, что только данные, непосредственно наблюдаемые на открытом рынке, являются надёжной основой суждения о стоимости. Приоритет таких данных над прочими, в частности, полученными путём опроса экспертов, закреплён, в том числе в Стандартах оценки RICS [8], Международных стандартах оценки 2022 [9], а также МСФО 13 «Оценка справедливой стоимости [5]. Поэтому можно говорить о том, что математические методы анализа данных, полученных на открытом рынке, являются наиболее надёжным средством интерпретации рыночной информации, применяемой при исследованиях рынка и предсказании стоимости конкретных объектов. В данном материале будут рассмотрены основные теоретические вопросы, касающиеся теста Манна—Уитни—Уилкоксона (далее U-тест), а также проведён пошаговый разбор применения данного теста к реальным данным. Материал содержит строки кода, необходимые для проведения U-теста с использованием языков программирования Python и R, а также приложение в виде электронной таблицы, содержащей тестовые данные и формулы для проведения рассматриваемого теста и полностью готовой для её применения на любых иных данных. Данный материал и все приложения к нему распространяются на условиях лицензии cc-by-sa-4.0 [13].

Содержание

Список таблиц

Список диаграмм

Листинги

1. Технические данные

Данный материал, а также приложения к нему доступны по постоянной ссылке [10]. Исходный код данной работы был создан с использованием языка TEX [24] с набором макрорасширений LATEX [25], дистрибутива TeXLive [26] и редактора TeXstudio [38]. Расчёт в форме электронной таблицы был выполнен с помощью LibreOffice Calc [15]

(Version: 7.3.3.2, Ubuntu package version: 1:7.3.3 rc2-0ubuntu0.20.04.1 lo1 Calc: threaded). Расчёт на языке R [27] (version 4.2.0 (2022-04-22) – "Vigorous Calisthenics") был выполнен с использованием IDE RStudio (RStudio 2022.02.2+485 "Prairie Trillium"Release (8acbd38b0d4ca3c86c570cf4112a8180c48cc6fb, 2022-04-19) 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]. Расчёт на языке Python (Version 3.9.12) [14] был выполнен с использованием среды разработки Jupyter Lab (Version 3.4.2) [17] и 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]. Графические материалы, использованные в подсекции 4.1, были подготовлены с использованием Geogebra (Version 6.0.666.0-202109211234) [16].

2. Предмет исследования

В случае работы с рыночными данными перед оценщиком часто встаёт задача проверки гипотезы о существенности влияния того или иного признака, измеренного в количественной или порядковой шкале, на стоимость. Аналогичная задача возникает у аналитиков рынка недвижимости, специалистов компаний-застройщиков, риелторов. При этом зачастую отсутствует возможность сбора больших массивов данных, позволяющих применить широкий спектр методов машинного обучения. В ряде случаев оценщики осознанно сужают область сбора данных до узкого сегмента рынка, в результате чего в их распоряжении оказываются лишь сверхмалые выборки объёмом менее тридцати наблюдений. При этом, ценовые данные чаще всего имеют распределение отличное от нормального. В данном случае рациональным решением является применение U-теста. Сформулируем задачу:

- предположим, что у нас существуют две выборки удельных цен коммерческих помещений, часть из которых обладает некоторым признаком (например, имеет отдельный вход), часть — нет;
- необходимо установить: оказывает ли наличие этого признака существенное влияние на удельную стоимость недвижимости данного типа или нет.

На первый взгляд, согласно сложившейся практике, оценщик может просто субъективно признать те или иные признаки значимыми, а прочие нет, после чего принять значения корректировок на различия в этих признаках из справочников. Однако, как было сказано выше, такой подход вряд ли может считаться лучшей практикой, поскольку в этом случае отсутствует какой-либо серьёзный анализ рынка. Кроме того, в таком случае вряд ли можно говорить о какой-либо ценности такой работы в принципе.

Вместо этого возможно использовать случайные выборки рыночных данных и применять к ним математические методы анализа, позволяющие делать доказательные с научной точки зрения выводы о значимости влияния того или иного признака на стоимость. Данные, используемые в настоящей работе при проведении U-теста средствами Python и R, представляют собой реальные рыночные данные, часть

из которых была собрана автором путём парсинга, часть — предоставлена коллегами для анализа. Прилагаемая электронная таблица настроена таким образом, что исходные данные могут быть сгенерированы случайным образом.

3. Основные сведения о тесте

3.1. Предпосылки и формализация гипотез

В первую очередь необходимо сказать, что, несмотря на заявленное общее название, правильнее всё же говорить о двух тестах:

- двухвыборочный критерий Уилкоксона, разработанный Фрэнком Уилкоксоном в 1945 году [20];
- U-критерий Манна"– Уитни, являющийся дальнейшим развитием вышеуказанном критерия, разработанный Генри Манном и Дональдом Уитни в 1947 году [18].

Забегая вперёд, можно сказать о том, что статистики данных критериев линейно связаны, а сами p-значения практически одинаковы, что с практической точки зрения позволяет скорее говорить о вариациях одного теста, а не о двух отдельных [20]. В данной работе по всему тексту используется общее название, а также его сокращённый вариант — U-тест, исторический относимый к критерию Манна-Уитни. Некоторые авторы [4] рекомендуют использовать двухвыборочный критерий Уилкоксона в случаях, когда нет предположений о дисперсиях, а в случае равных дисперсий применять U-критерий Манна-Уитни. Однако экспериментальные данные указывают, что p-значения критериев Уилкоксона и Манна—Уитни практически совпадают, в том числе и в случае, когда дисперсии выборок существенно различаются. Придерживаясь принципа KISS [30], лежащего в основе всего данного цикла публикаций, автор приходит к выводу о возможности применения единого подхода.

Также следует помнить о том, что существует Критерий Уилкоксона для связных выборок [21], представляющий собой отдельный тест, предназначенный для анализа различий между связанными выборками, тогда как рассматриваемый в данной работе U-тест предназначен для работы с двумя независимыми выборками.

Предположим, что заданы две выборки:

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

- Обе выборки являются простыми, объединённая выборка независима.
- Выборки взяты из неизвестных непрерывных распределений *F(x)* и *G(y)* соответственно.

Простая выборка — это случайная, однородная, независимая выборка. Эквивалентное определение: выборка $x^m = (x_1,x_2,\ldots,x_m)$ является простой, если значения $(x_1,x_2,\ldots,x_m)$ являются реализациями m независимых одинаково распределённых случайных величин. Иными словами, отбор наблюдений является

не только случайным, но и не предполагает наличия каких-либо специальных правил (например, выбор каждого 10-го наблюдения).

U-тест — это непараметрический тест для проверки нулевой гипотезы, заключающейся в том, что для случайно выбранных из двух выборок наблюдений $x, x \in X$ и $y, y \in Y$ вероятность того, что x больше y, равна вероятности того, что y больше x. На математической языке запись нулевой гипотезы выглядит следующим образом:

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

Для целостности теста требуется альтернативная гипотеза, которая заключается в том, что вероятность того, что значение признака наблюдения из выборки X превышает его у наблюдения из выборки Y, отличается (больше или меньше) от вероятности того, что значение признака у наблюдения из Y превышает значение у наблюдения из X. На математическом языке запись альтернативной гипотезы выглядит следующим образом:

$$H_1 : P\{x < y\} \neq P\{y < x\} \vee P\{x < y\} + 0.5 \cdot P\{x = y\} \neq 0.5. \tag{2}$$

Согласно базовой концепции U-теста, при справедливости нулевой гипотезы распределение двух выборок непрерывно, при справедливости альтернативной распределение одной из них стохастически больше распределения другой. При этом, можно сформулировать целый ряд нулевых и альтернативных гипотез, для которых данный тест будет давать корректный результат. Его самое широкое обобщение заключается в следующих предположениях:

- наблюдения в обеих выборках независимы;
- тип данных является как минимум ранговым, т. е. в отношении любых двух наблюдений можно сказать, какое из них больше;
- нулевая гипотеза предполагает, что распределения двух выборок равны;
- альтернативная гипотеза предполагает, что распределения двух выборок не равны.

В случае более строгого набора допущений, чем приведённые выше, например, в случае допущения о том, что распределение двух выборок в случае справедливости нулевой гипотезы непрерывно, альтернативной — имеет сдвиг расположения двух распределений, т. е. $f_1 x = f_2(x + \sigma)$, можно сказать, что U-тест представляет собой тест на проверку гипотезы о равенстве медиан. В этом случае, U-тест можно интерпретировать как проверку того, отличается ли от нуля оценка Ходжеса—Лемана разницы значений мер центральной тенденции. В данной ситуации оценка Ходжеса—Лемана представляет собой медиану всех возможных значений различий между наблюдениями в первой и второй выборках. Вместе с тем, если и дисперсии, и формы распределения обеих выборок различаются, U-тест не может корректно проверить

медианы. Можно показать примеры, когда медианы численно равны, при этом тест отвергает нулевую гипотезу с вследствие малого p-значения.

Таким образом, более корректной интерпретацией U-теста является его использование для проверки именно гипотезы сдвига [19].

Гипотеза сдвига — статистическая гипотеза, часто рассматривающаяся как альтернатива гипотезе о полной однородности выборок. Пусть даны две выборки данных. Пусть также даны две случайные величины X и Y, которые распределены как элементы этих выборок и имеют функции распределения $F(x)$ и $G(y)$ соответственно. В этих терминах гипотезу сдвига можно записать следующим образом:

$$H : F(x) = G(x + \sigma) \quad |\forall x,\, \sigma \neq 0. \tag{3}$$

В этом случае U-критерий является состоятельным независимо от особенностей выборок.

Простыми словами, суть U-теста заключается в том, что он позволяет ответить на вопрос, является ли существенным различие значения количественного признака двух выборок. Применительно к оценке можно сказать, что применение данного теста помогает ответить на вопрос, является ли необходимым учёт того или иного признака в качестве ценообразующего фактора. Из сказанного выше следует, что речь идёт о двухстороннем тесте. На практике это означает, что тест не даёт прямой ответ, например на такой вопрос: «имеет ли место значимое превышение удельной стоимости помещений, имеющих отдельный вход, относительно помещений, не обладающих им». Вместо этого корректно говорить о том, «существует ли существенное различие в значении стоимости между помещениями двух типов: с отдельным входом и без такового».

Условиями применения U-теста помимо вышеуказанных требований к самим выборкам являются:

- распределение значений количественного признака выборок отлично от нормального (в противном случае целесообразно использование параметрического t-критерия Стьюдента для независимых выборок);
- не менее трёх значений признака в каждой выборке, допускается наличие двух значений в одной из выборок, при условии наличия в другой не менее пяти.

Подытоживая вышесказанное, можно сказать, что существуют три варианта нулевой гипотезы, в зависимости от уровня строгости.

3.2. Реализация теста

3.2.1. Статистика критерия

Допустим, что элементы $x_1, \ldots, x_n$ представляют собой простую независимую выборку из множества $X \in \mathbb{R}$, а элементы $y_1, \ldots, y_n$ представляют собой простую

Таблица 1. Варианты нулевой гипотезы при использовании U-теста при оценке стоимости

Тип гипотезы	Формулировка
Научная	Наблюдения из двух выборок полностью однородны, т. е. принадлежат одному распределению, сдвиг отсутствует, оценка, сделанная для первой выборки, является несмещённой и для второй
Практическая	Медианы двух выборок равны между собой
Изложенная в терминах оценки	Различие признака между двумя выборками объектов-аналогов не является существенным, его учёт не требуется, данный признак не является ценообразующим фактором

независимую выборку из множества $Y \in \mathbb{R}$, при этом выборки являются независимыми относительно друг друга. Тогда соответствующая U-статистика определяется следующим образом:

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

при

$$S(x,y) = \begin{cases} 1, & \text{если } x > y, \\ \frac{1}{2}, & \text{если } x = y, \\ 0, & \text{если } x < y. \end{cases} \tag{4}$$

3.2.2. Методы вычисления

Тест предполагает вычисление статистики, обычно называемой U-статистикой, распределение которой известно в случае справедливости нулевой гипотезы. При работе со сверхмалыми выборками распределение задаётся таблично, при размерах выборки более двадцати наблюдений оно достаточно хорошо аппроксимируется нормальным распределением. Существуют два методы вычисления U-статистики: подсчёт вручную по формуле 4, применение специального алгоритма. Первый способ подходит только для сверхмалых выборок в силу трудоёмкости. Второй способ может быть формализован в виде пошагового набора инструкций и будет описан далее.

1) Необходимо построить общий вариационный ряд для двух выборок, а затем присвоить каждому наблюдению ранг, начиная с 1 для наименьшего из них. В случае наличия связок, т. е. групп повторяющихся значений (такой группой могут являться в т. ч. только два равных значения), каждому наблюдению из такой группы присваивается значение, равное медиане значений рангов группы до корректировки (например, в случае вариационного ряда (*3, 5, 5, 5, 5, 8*) ранги до корректировки имеют вид (*1, 2, 3, 4, 5, 6*) после — (*1, 3.5, 3.5, 3.5, 3.5, 6*)).

2) Необходимо провести подсчёт сумм рангов наблюдений каждой из выборок, обозначаемых как R_1, R_2 соответственно. При этом, общая сумма рангов R может быть вычислена по формуле

$$R = \frac{N(N+1)}{2}, \tag{5}$$

где N — общее число наблюдений в обеих выборках.

3) Далее вычисляем U-значение для первой выборки:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \tag{6}$$

где R_1 — сумма рангов первой выборки, n_1 — число наблюдений в первой выборке.

Аналогичным образом вычисляется U-значения для второй выборки:

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}, \tag{7}$$

где R_2 — сумма рангов второй выборки, n_2 — число наблюдений во второй выборке.

Из вышеприведённых формул следует, что

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}. \tag{8}$$

Также известно, что

$$\begin{cases} R_1 + R_2 = \dfrac{N(N+1)}{2} \\ N = n_1 + n_2. \end{cases} \tag{9}$$

Тогда

$$U_1 + U_2 = n_1 n2. \tag{10}$$

Использование данной формулы в качестве контрольного соотношения может быть полезно для проверки корректности вычислений при расчёте в табличном процессоре.

4) Из двух значений U_1, U_2 во всех случаях выбираем меньшее, которое и будет являться U-статистикой и использоваться в дальнейших расчётах. Обозначим его как U.

3.2.3. Интерпретация результата

Для корректной интерпретации результата теста необходимо указать:

- размеры выборок;

- значения меры центральной тенденции для каждой выборки (с учётом непараметрического характера теста, подходящей мерой центральной тенденции представляется медиана);
- значение самой U-статистики;
- показатель CLES [29];
- рангово-бисериальный коэффициент корреляции (RBC) [34];
- принятый уровень значимости (как правило 0.05).

Понятие U-статистики было рассмотрено ранее, большинство других показателей широко известны и не требуют какого-либо отдельного рассмотрения. Остановимся на показателях CLES и RBC.

3.2.3.1. Показатель CLES

Common language effect size (CLES) — вероятность того, что значение случайно выбранного наблюдения из первой группы больше значения случайно выбранного наблюдения из второй группы. Данный показатель вычисляется по формуле

$$CLES = \frac{U_1}{n_1 n_2}. \tag{11}$$

Вместо обозначения *CLES* часто используется обозначение *f (favorable)*. Данное выборочное значение является несмещённой оценкой значения для всей совокупности объектов, принадлежащих множеству.

Следует отметить, что значение и смысл данного показателя эквивалентны значению и смыслу показателя AUC[35]. Таким образом, можно говорить о том, что

$$CLES = f = AUC_1 = f = \frac{U_1}{n_1 n_2}. \tag{12}$$

3.2.3.2. Рангово-бисериальная корреляция Метод представления степени влияния для U-теста заключается в использовании меры ранговой корреляции, известной как рангово-бисериальная корреляция. Как и в случае с иными мерами корреляции значение коэффициента рангово-бисериальной корреляции может принимать значения в диапазоне $[-1; 1]$, при этом нулевое значение означает отсутствие какой-либо связи. Коэффициент рангово-бисериальной корреляции обычно обозначает как r. Для его вычисления используется простая формула, основанная на значении CLES. Выдвинем гипотезу о том, что в паре случайных наблюдений, одно из которых взято из первой выборки, другое — из второй, значение первого больше. Запишем её на математическом языке:

$$H : x_i > y_j \quad x \ \in X, y \in Y. \tag{13}$$

Тогда значение коэффициента рангово-бисериальной корреляции представляет собой разницу между долей случайных пар наблюдений, удовлетворяющей (favorable) гипотезе — f, и комплементарной ей доле случайных пар, не удовлетворяющих (unfavorable) гипотезе — u. По сути, данная формула представляет собой формулу разности между показателями CLES для каждой из групп.

$$r = f - u = CLES_1 - CLES_2 = f - (1 - f) \tag{14}$$

Существует также ряд альтернативных формул, дающих идентичный результат:

$$r = 2f - 1 = \frac{2U_{min}}{n_1 n_2} - 1 = 1 - \frac{2U_{max}}{n_1 n_2}. \tag{15}$$

3.2.4. Вычисление p-значения и итоговая проверка нулевой гипотезы

При достаточном большом числе наблюдений в каждой выборке, значение U-статистики имеет приблизительно нормальное распределение. Тогда её стандартизированное значение (z-метка, z-score) [37] может быть вычислено по формуле

$$z = \frac{U - m_U}{\sigma_U}, \tag{16}$$

где m_U — среднее арифметическое U, σ_U — её стандартное отклонение. Визуализация понятия стандартизированное значения для нормального распределения приведена на рисунке 1. Среднее для U вычисляется по формуле

$$m_U = \frac{n_1 n_2}{2}. \tag{17}$$

Формула стандартного отклонения в случае отсутствия связок выглядит следующим образом:

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}. \tag{18}$$

В случае наличия связок используется другая формула:

$$\sigma_{U_{ties}} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum_{k=1}^{K} ({t_k}^3 - t_k)}{12n(n-1)}} = \sqrt{\frac{n_1 n_2}{12}\left((n+1) - \frac{\sum_{k=1}^{K} ({t_k}^3 - t_k)}{n(n-1)}\right)}, \tag{19}$$

где t_k — количество наблюдений, имеющих ранг k, K — общее число рангов, имеющих связки. Далее, получив стандартизированное значение (z-score), и используя аппроксимацию стандартного нормального распределения, вычисляется *p-значение* для заданного уровня значимости (как правило 0.05). Интерпретация результата осуществляется следующим образом:

$$\begin{aligned} &p < 0.05 \Rightarrow \text{нулевая гипотеза отклоняется} \\ &p \geq 0.05 \Rightarrow \text{нулевая гипотеза не может быть отклонена.} \end{aligned} \tag{20}$$

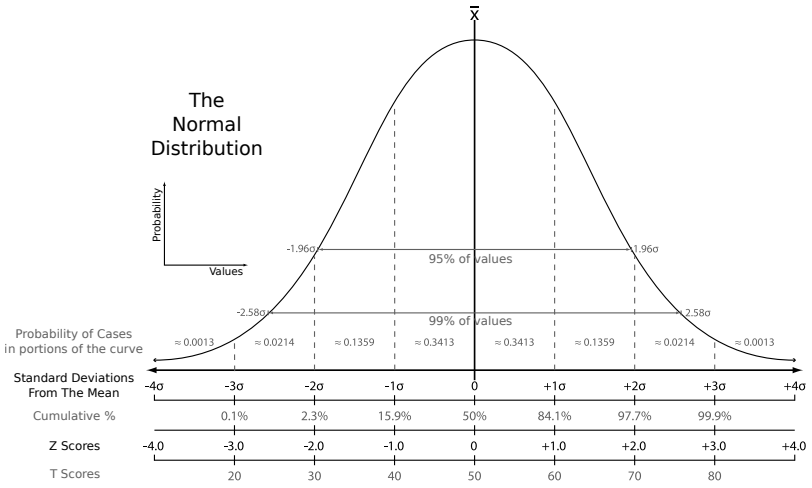

Рис. 1. Визуализация понятия стандартизированного значения (z-score) для нормального распределения [37]

4. Практическая реализация

4.1. Реализация в табличном процессоре LibreOffice Calc

На данный момент можно с уверенностью сказать, что табличные процессоры являются стандартом для расчётов оценщиков. Проникновение средств разработки, например на языке Python либо R, в профессиональную деятельность оценщиков идёт достаточно медленно. Кроме того, самостоятельный пошаговый расчёт позволяет лучше понять методику U-теста. Поэтому было принято решение создать пошаговую инструкцию для проведения U-теста в электронной таблице. Для этого был использован программный продукт LibreOffice Calc (Version: 7.3.3.2, Ubuntu package version: 1:7.3.3 rc2-0ubuntu0.20.04.1 lo1 Calc: threaded), существенная часть функционала которого имеется также и в наиболее распространённом приложении такого рода Microsoft Excel. Отсутствуют основания полагать, что сделанные расчёты не будут корректно работать в приложениях, отличных от LibreOffice Calc. Однако гарантировать это также невозможно. Для однозначно корректного проведения теста рекомендуется использовать именно данное приложение, имеющее версии для всех основных операционных систем. Актуальная версия файла U-test.ods находится в репозитории вместе с остальными материалами данной работы.

Данные, рассматриваемые в данной подсекции, являются вымышленными и были созданы алгоритмом генерации псевдослучайных чисел LibreOffice Calc. Для повтор-

ной генерации необходимо использовать сочетание клавиш *ctrl+shift+F9*.

Рассмотрим учебную задачу. В ячейках I3:J30 содержатся данные значений некоторого количественного признака для двух выборок, взятых из множеств I и J соответственно. Различие между элементами этих множеств заключается в наличии некоторого признака у элементов множества I и его отсутствия у элементов множества J. Задача заключается в проверке гипотезы о том, что различие в данном признаке следует признать существенным, а сам признак является ценообразующим фактором. Выдвинем нулевую гипотезу, сформулировав её в трёх вариантах, соответствующих трём уровням строгости, описанным ранее в таблице 1. Следует отметить, что U-тест основан на т. н. *частотном подходе к вероятности* (о различиях между *частотным* и *байесовским* подходом к вероятности применительно к оценке стоимости можно прочитать, в частности в [7]). Как известно, частотный подход базируется на предпосылке о том, что случайность является следствием объективной неопределённости, которая может быть уменьшена только путём проведения серии экспериментов. В частотном подходе существует чёткое разделение на случайные и неслучайные параметры. Типичной задачей является оценка тех или иных параметров генеральной совокупности, представляющей собой набор случайных величин на основе детерминированных параметров выборки, например: среднее, мода, дисперсия и т. д. Последние представляют собой конкретные значения, в которых уже нет никакой случайности. Таким образом, принимая фундаментальное предположение о случайном характере изучаемых величин, мы применяем те или иные методы математической статистики, позволяющие получить конкретные значения оценок параметров. Из это следует, что нулевая гипотеза чаще всего «пессимистична», т. е. гласит о том, что в основе исследуемого явления, процесса или объекта лежит случайность, вследствие чего мы не имеем возможность делать надёжные выводы. С учётом всего вышесказанного сформулируем нулевую и альтернативную гипотезы в трёх вариантах, согласно уровням строгости, показанным в таблице 1. Ячейки C2:C19 содержат некоторые описательные статистики. Для удобства первичного анализа бывает полезно показать свойства выборок графически. На рисунке 2 изображена диаграмма «ящик с усами» (Boxplot), позволяющая сделать некоторые выводы на основе одного взгляда. Как видно, значения средних и медиан двух выборок различны. При этом также отличаются минимальные значения. При этом максимальное значение одинаково. Также следует обратить внимание, что несмотря на то, что среднее и медиана первой выборки превышают аналогичные показатели второй, минимальное значение первой меньше чем у второй. В таких условиях ещё сложнее сделать вывод о том, является ли различие в признаке существенным или же разница в показателе стоимости носит случайный характер. Следующим подготовительным этапом является проверка нормальности распределения значений количественного признака (в данном случае условного показателя удельной стоимости). Существует ряд строгих тестов, позволяющих провести такую проверку численными методами. В подсекциях 4.2 и 4.3 будут показаны соответствующие способы проведения такого теста. В данном разделе ограничимся графическим способом. На рисунках 3, 4 изображены гистограммы распределения частот для первой и второй выборок соответственно, совмещённые с кривыми функции плотности

Таблица 2. Нулевая и альтернативная гипотезы при анализа тестовых данных

Тип гипотезы	Нулевая гипотеза (H0)	Альтернативная гипотеза (H1)
Научная	Распределение удельных показателей стоимости одинаково для объектов-аналогов, обладающих признаком «X» (множество объектов I), и не обладающих им (множество объектов J), сдвиг между ними отсутствует, статистические оценки, сделанные для одного множества объектов-аналогов, являются несмещёнными для другого.	Распределение удельных показателей стоимости для объектов из множества I отличается от распределения, имеющего место у множества J, существует сдвиг, оценка, сделанная для объектов, принадлежащих множеству I будет смещённой для объектов, принадлежащих множеству J.
Практическая	Медианное значение удельного показателя стоимости объектов, обладающих признаком «X», не отличается от медианного значения удельного показателя стоимости объектов, не обладающих признаком «X» — их медианы равны.	Медианное значение удельного показателя стоимости объектов, обладающих признаком «X», отличается от медианного значения удельного показателя стоимости объектов, не обладающих признаком «X» — их медианы не равны.
Изложенная в терминах оценки	Наличие или отсутствие признака «X» не оказывает сколько-нибудь заметного влияния на стоимость — признак «X» не является ценообразующим фактором.	Наличие или отсутствие признака «X» оказывает влияние на стоимость — признак «X» является ценообразующим фактором.

вероятности для нормального распределения. Как видно на обеих диаграммах, форма распределения обеих выборок существенно отличается от формы кривой функции плотности вероятности нормального распределения. При работе с реальными данными лучше всё же проводить количественные тесты, однако на данном этапе остановимся на интерпретации диаграмм и сделаем вывод о том, что распределения обеих выборок отличаются от нормального, что позволяет сделать вывод о неприменимости параметрических методов статистического оценивания и необходимости использования непараметрических, к числу которых относится и U-тест.

При работе с электронной таблицей потребность в отдельном построении общего вариационного ряда для двух выборок отсутствует. Вместо этого можно сразу перейти к вычислению рангов наблюдений. С учётом возможного наличия связок (повторяющихся значений) следует использовать функцию RANK.AVG, последовательно указав при этом три аргумента: наблюдение, для которого вычисляется ранг, диапазон всех значений общего ряда, тип сортировки: 0 — по убыванию, 1 — по возрастанию, в нашем случае необходимо указать 1. Столбцы L, N содержат

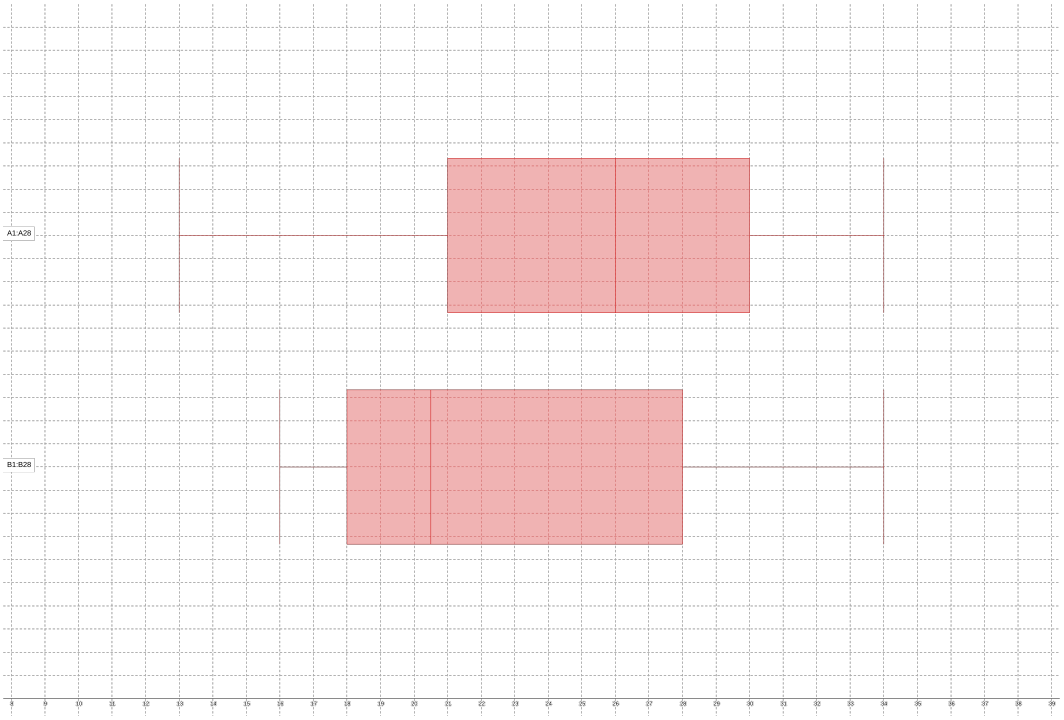

Рис. 2. Диаграмма «ящик с усами» (Boxplot) для обеих выборок

дублирующие значения, столбцы M и O — ранги соответствующих наблюдений.

После этого проведём подсчёт сумм рангов для каждой из выборок в ячейках C20:C21. В ячейке C22 проведём подсчёт общей суммы рангов обеих выборок. Для проверки рассчитаем тот же показатель согласно формуле 5.

Далее в ячейках C25, C26 по формулам 6, 7 вычислим соответственно значения U_1, U_2. После чего проверим корректность контрольного соотношения 10 в ячейке D27. В C28 выбираем меньшее значение, которое и будет использоваться в дальнейшем в качестве U-статистики. В нашем случае меньшее значения U-статистики у выборки из множества *J*.

Рассчитаем показатель CLES. Для этого используем формулу 11. Результат содержится в C29. В рассматриваемом примере значение показателя составляет 0.39477, что следует интерпретировать следующим образом: «вероятность того, что значение показателя удельной стоимости случайно выбранного наблюдения из первой выборки превышает аналогичный показатель случайно выбранного наблюдения из второй выборки составляет 0.39477 (39.48 %)».

Далее рассчитаем значение коэффициента рангово-бисериальной корреляции по формуле 14, 15, разместив его в ячейке C36. В рассматриваемом случае значение составило -0.21, что означает, что говорит об обратном влиянии отсутствия признака «X» на стоимость. Говоря более понятным языком, можно сделать вывод о том, что сила корреляционной связи между наличием этого признака и показателем

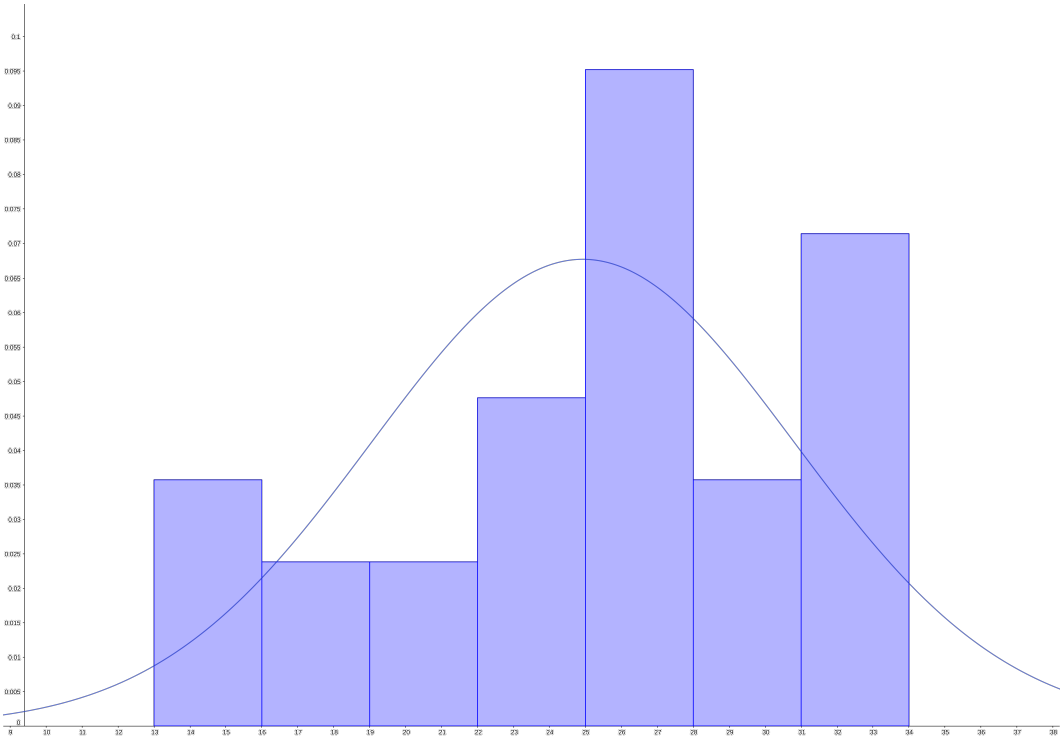

Рис. 3. Гистограмма первой выборки, совмещённая с кривой функции плотности вероятности для нормального распределения

стоимости составляет 0.21.

После этого перейдём к расчёту стандартизированного значения согласно формуле 16. Для этого в ячейке C37 рассчитаем среднее по формуле 17, а затем перейдём к вопросу расчёта стандартного отклонения. Следует отметить, что для этого существуют две формулы: одна (18) применяется в случае отсутствия связок (ячейка C38), вторая (19) — при их наличии (ячейка C39). В рассматриваемом случае связки имели место. Их обработка осуществлялась в столбцах P и Q, а также в ячейках E35:E49. В результате было получено два значения, отличие между которыми составило менее одного процента. Учёт фактора связок необходим с точки зрения максимальной научной корректности результата, однако в повседневной практической деятельности некоторые оценщики могут столкнуться со сложностями с корректным учётом фактора связок, а также не иметь достаточно времени для дополнительных расчётов. Практический опыт говорит о том, что сколько-нибудь существенное отличие значений стандартного отклонения, полученных с помощью формулы 19 от значений, полученных согласно 18, бывает в случаях большого числа связок, а также наличия крупных групп. В остальных ситуациях более простая формула, автоматически вычисляющая показатель σ, даёт корректный результат, достаточный для практического применения в оценке. В любом случае, решение об использовании строгих либо простых методов принимает сам‘оценщик. В рассматриваемом примере учёт

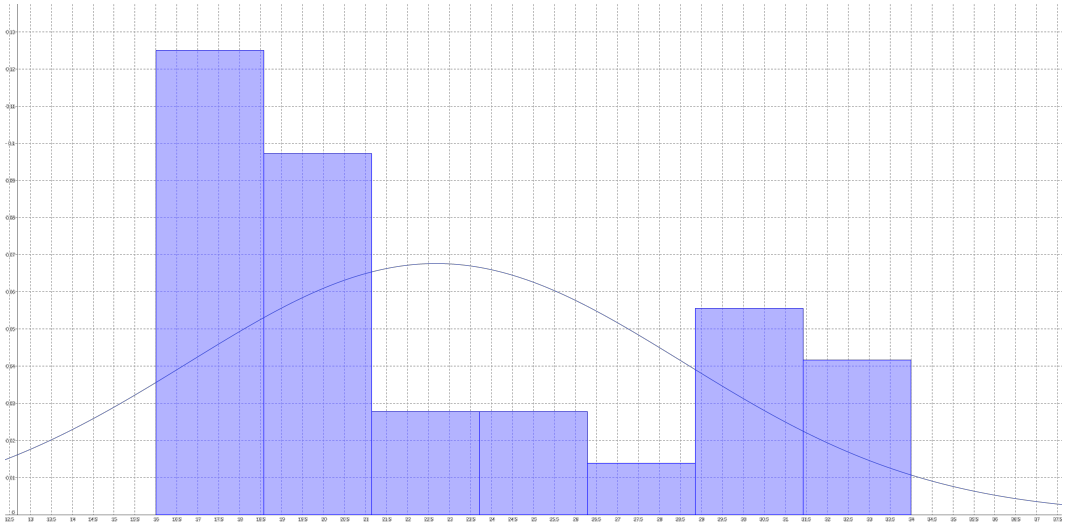

Рис. 4. Гистограмма второй выборки, совмещённая с кривой функции плотности вероятности для нормального распределения

фактора связок был осуществлён.

Зная среднее и стандартное отклонение, вычисляем z-метку в ячейке C44, а затем, используя аппроксимацию стандартного нормального распределения, — p-значение. В рассматриваемом примере оно составило 0.1757. Используя правило 20, приходим к выводу о невозможности отклонить нулевую гипотезу. Таким образом, используя формулировку, наиболее близкую к оценочной деятельности (см. таблицу 2), можно придти к следующему выводу: Наличие или отсутствие признака «X» не оказывает сколько-нибудь заметного влияния на стоимость — признак «X» не является ценообразующим фактором.

В данной подсекции мы рассмотрели пошаговый расчёт статистики критерия, а также осуществили интерпретацию результата. Следует отметить, что, несмотря на возможность и даже относительное удобство такого варианта проведения U-теста, предпочтение всё же следует отдавать профессиональным средствам разработки в области машинного обучения и статистического вывода, например, языкам программирования Python или R, о которых и пойдёт речь ниже.

4.2. Реализация на Python

В сфере машинного обучения и, в особенности, в ряде областей таких как deep learning язык Python уже стал де факто стандартом. Кроме того, он универсален и прекрасно подходит для разработки тех или иных экспертных систем. Его популярность означает в т. ч. наличие огромного количества обучающих материалов по всем аспектам разработки в области анализа данных, предназначенных для пользователей любого уровня подготовки. При этом, большая часть необходимых оценщику вычислений можно провести путём вызова готовых функций из подключаемых

библиотек, предназначенных для анализа данных, без необходимости написания большого объёма кода и без глубоких знаний в области программирования. По мнению автора данной работы, будущее оценки заключается именно в применении экспертных систем, основанных на обучении моделей на основе наборов данных открытых рынков. Как будет показано ниже, применение Python существенно сокращает время проведения U-теста, позволяет создавать визуализации исследуемого рынка не прибегая к сторонним средствам. Кроме того, использование готовых функций практически исключает вероятность возникновения ошибок в расчётах. При написании кода была использована версия языка Python 3.9.12, а также IDE Spyder (5.1.5). Код в формате скрипта доступен по ссылке [11], код в формате Python Notebook доступен по ссылке [12].

Рассмотрим реальный набор данных, содержащий сведения об удельных показателях стоимости квартир в Санкт-Петербургской агломерации. Данные были собраны 28 сентября 2021 года с сайта cian.ru и доступны по ссылке. Рассматриваемый набор данных содержит 34821 наблюдение. При этом Санкт-Петербургская агломерация включает в себя как территории, входящие в состав города федерального значения, так и те, которые формально относятся к Ленинградской области. При этом разделение на город и область носит чисто юридический характер. С социально-экономической точки зрения ближайшие территории Ленинградской области неразрывно связаны с Санкт-Петербургом и являются частью одной агломерации, к слову, крупнейшей в мире на такой широте. При формировании запросов, использованных в процессе скрепинга, южная граница агломерации была установлена примерно по оси автодороги А-120, северная — автодороги 41А-189. При этом в её состав были включены некоторые населённые пункты за пределами этих границ, например, города Кировск и Шлиссельбург.

Сформулируем задачу. Необходимо установить наличие либо отсутствие статистически значимого различия в ценах объектов, расположенных в границах самого Санкт-Петербурга, и объектов, формально расположенных в Ленинградской области. Аналогично предыдущему случаю, сформулируем нулевую и альтернативную гипотезы, имеющие на этот раз практический смысл.

Язык Python изначально не был создан специально для анализа данных. Поэтому в его базовой версии могут отсутствовать многие функции, необходимые для проведения расчётов. К счстью, для решения задач в области машинного обучения и анализа данных существует ряд подключаемых библиотек, содержащих множество необходимых функций. Их количество и широта решаемых задач не столь велики, как, например, у языка R, однако они являются исчерпывающими для тех задач, которые стоят перед 95 % оценщиков. Для решения задач, рассматриваемых в данном материале, потребуются следующие библиотеки: numpy, pandas, math, matplotlib.pyplot, scipy.stats. Для их подключения потребуется код, представленный в листинге 1. Установим уровень значимости α, принимаемый для всей дальнейшей работы. Выбор его значения остаётся за исследователем, однако в работах по эконометрике и исследованию операций чаще всего встречается значение *0.05*, которое и будет использовано в данной подсекции. Для задания значения уровня значимости используется код 2.

Листинг 1. Подключение необходимых библиотек

```
# import libraries
import numpy as np
import pandas as pd
import math
import matplotlib.pyplot as plt
from scipy.stats import norm
import scipy.stats as stats
from scipy.stats import normaltest
from scipy.stats import shapiro
from scipy.stats import anderson
from scipy.stats import mannwhitneyu
```

Листинг 2. Задание применяемого уровня значимости

```
# set significance level
alpha = 0.05
```

После этого всё готово для начала работы. Создадим датафрейм на основе текстового файла, содержащего изучаемый набор данных (листинг 3). Датафрейм в точности повторяет содержимое исходного файла и содержит 34821 наблюдения и 4 переменные: порядковый номер, ссылку на объявление, показатель стоимости 1 кв. м, а также код местоположения, состоящий из четырёх букв: первая из которых означает регион (s — Санкт-Петербург, l — Ленинградская область), вторая и третья — административный район, три последних — муниципальное образование либо территорию. При этом Python добавил собственную переменную, содержащую номера наблюдений. Следует обратить внимание на то, что нумерация в Python как и в большинстве языков программирования начинается, не с единицы, а с нуля. Поскольку переменные, содержащие номера наблюдений и ссылки на объявления из исходного файла, не будут использоваться в дальнейшем, создадим новый датафрейм, содержащий только необходимые переменные, а также выгрузим из виртуальной памяти первый датафрейм для оптимизации ресурсов компьютера (листинг 4). В рассматриваемом случае такая микрооптимизация не играет большой роли, одна-

Листинг 3. Загрузка данных и создание датафрейма

```
# import dataset
df = pd.read_csv('spba-flats-210928.csv')
print(df)
type(df['price_m'])
```

Листинг 4. Создание датафрейма содержащего только необходимые переменные и выгрузка из памяти неиспользуемых данных

```
# get only prices and counties, release RAM
df1 = df[['price_m', 'county']]
del [[df]]
```

Листинг 5. Создание датафрейма содержащего только необходимые переменные и выгрузка из памяти неиспользуемых данных

```
# get only prices and counties, release RAM
df1 = df[['price_m', 'county']]
del [[df]]
```

ко в целях выработки навыков написания хорошего кода, лучше всё же написать одну дополнительную строку. Теперь, в распоряжении оценщика в удобном виде есть рабочий датафрейм, содержащий данные о рынке квартир всей агломерации Санкт-Петербурга. Для формирования первого представления о распределении построим гистограмму, совмещённую с кривой плотности для нормального распределения. Для определения рационального числа интервалов (столбцов гистограммы) k используем формулу Heinhold-Gaede cite[2]:

$$k = \sqrt{n}, \tag{21}$$

где n — число наблюдений. Используем для этого скрипт из листинга 5. Рассмотрим полученную гистограмму 5. Ось x содержит значения цен за 1 кв. м, ось y — значения вероятностей интервалов. Обе оси представлены в стандартном виде. Также показаны значения матожидания и стандартного отклонения. Как видно, распределение имеет тяжёлый правый хвост, что позволяет сделать предварительный вывод о том, что оно отличается от нормального. В дальнейшем будет проведён строгий тест на нормальность, пока же можно ограничиться первичной субъективной интерпретацией гистограммы. Поскольку предметом исследования является различие между объектами, расположенными в двух частях агломерации, а исходный набор данных содержит сведения о наблюдениях из обеих, потребуется создание двух отдельный датафреймов. Здесь следует сделать небольшое отступление: практический опыт говорит о том, что сам анализ данных и построение моделей занимают только 20 % времени, тогда как 80 % уходит на сбор и предобработку данных. Одним из важных элементов этих процессов является правильная разметка данных. В случае с рассматриваемым набором данных их анализ в разрезе отдельных территорий вплоть до уровня муниципалитетов был предусмотрен изначально путём указания индекса территории для каждого наблюдения. Как было сказано выше, первая буква индекса содержит указание на то, в какой части агломерации расположено наблюдение. В этом случае, для создания двух раздельных датафреймов достаточно двух строк,

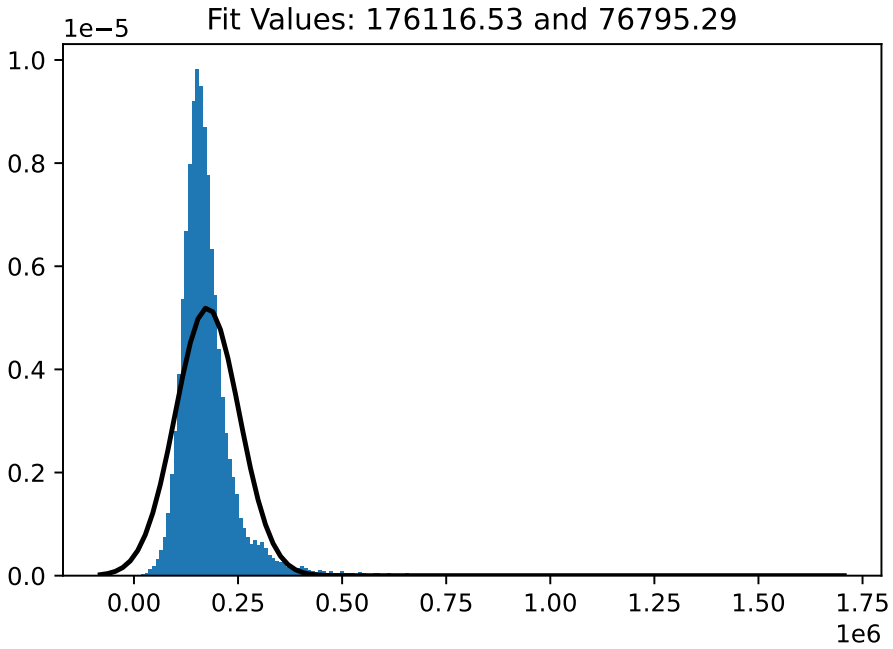

Рис. 5. Гистограмма плотности распределения цен за 1 кв. м квартир в Санкт-Петербургской агломерации, совмещённая с кривой функции плотности вероятности для нормального распределения

содержащих несложные регулярные выражения. Датафрейм *dfs* содержит данные для наблюдений из Санкт-Петербурга (28643 наблюдения), *dfl* — Ленинградской области (6178 наблюдений). Построим гистограммы для обеих частей агломерации. Гистограмму иногда путают со столбчатой диаграммой. Следует напомнить, что правильно построенная гистограмма является отображением вероятностных свойств данных, сумма площадей всех её прямоугольников равна единице, а по оси y отложены значения вероятностей диапазонов (столбцов гистограммы), а не число наблюдений в каждом диапазоне. Как видно из гистограммы 6, распределение удельных цен в Санкт-Петербурге также как и в случае с распределением цен для всей агломерации, имеет тяжёлый правый хвост. При этом распределение цен для объектов агломерации, находящихся за пределами границ Санкт-Петербурга, показанное

Листинг 6. Создание отдельных датафреймов для Санкт-Петербурга и Ленинградской области

```
# create separate data frames for city and suburbs
dfs = df1[df1['county'].str.startswith('s')] # Saint-Petersburg
dfl = df1[df1['county'].str.startswith('l')] # Leningradskaja oblastq
```

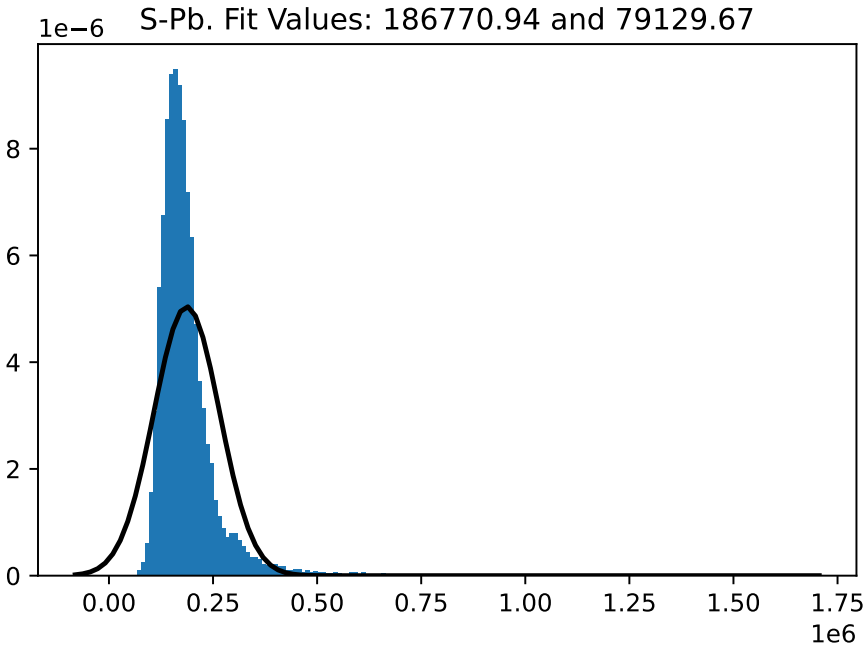

Рис. 6. Гистограмма плотности распределения цен за 1 кв. м квартир в Санкт-Петербурге, совмещённая с кривой функции плотности вероятности для нормального распределения

на гистограмме 7 выглядит относительно симметрично.

Также построим график «ящик с усами» для обоих датафреймов (см. диаграмму 8). Как видно, значение медианы цен объектов, расположенных в Санкт-Петербурге, выше значения третьего квартиля цен объектов, расположенных на прилегающих территориях Ленинградской области. Данные обстоятельства позволяет сделать субъективное предположение о том, что нулевую гипотезу следует отклонить. Однако графические методы анализа подходят только д, для формирования объективного доказательного суждения потребуется проведение самого U-теста.

Для проверки применимости U-теста следует провести тест на нормальность распределения для обоих датафреймов (*dfs*, *dfl*). Существует множество критериев для проверки гипотезы о нормальности распределения выборки. В данном случае были использованы три теста:

- тест Шапиро—Уилка [3];
- тест K^2 Д'Агостино [**Agostino-test**];
- тест Андерсона—Дарлинга [1].

Тест Шапиро—Уилка оценивает выборку данных и вычисляет, насколько вероятно, что она была взята из генеральной совокупности, имеющей нормальное распреде-

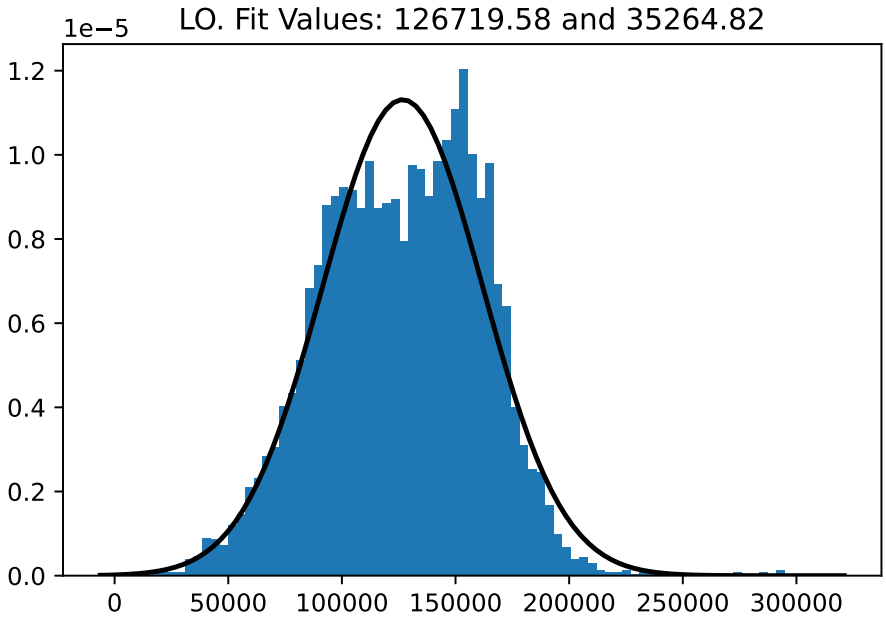

Рис. 7. Гистограмма плотности распределения цен за 1 кв. м квартир в Ленинградской области, расположенных в границах агломерации Санкт-Петербурга, совмещённая с кривой функции плотности вероятности для нормального распределения

ление. Данный тест считается одним из наиболее мощных тестов проверки на нормальность [4]. При этом существуют некоторые предпосылки, указывающие на то, что он лучше всего работает на выборках среднего размера, не превышающих пяти тысяч наблюдений.

Тест K^2 Д'Агостино основывается на анализе показателей асимметрии [36] и эксцесса [33], представляющих собой третий и четвёртый центральные моменты [28] соответственно. Данный тест также считается одним из наиболее мощных и не имеет ограничений по максимальному числу наблюдений.

Тест Андерсона—Дарлинга представляет собой модифицированную версию критерия согласия Колмогорова"– Смирнова [31] и используется для проверки гипотезы о том, что эмпирическое распределения согласуется с одним из известных теоретических. В отличие от двух предыдущих тестов, его результатом является не p-значение, а статистика критерия, что требует более сложной интерпретации результата, которая однако легко автоматизируется.

Сформулируем нулевые гипотезы:

- H0(SPb): распределение значений удельных цен предложений квартир в Санкт-Петербурге не отличается от нормального;
- H0(LO): распределение значений удельных цен предложений квартир на терри-

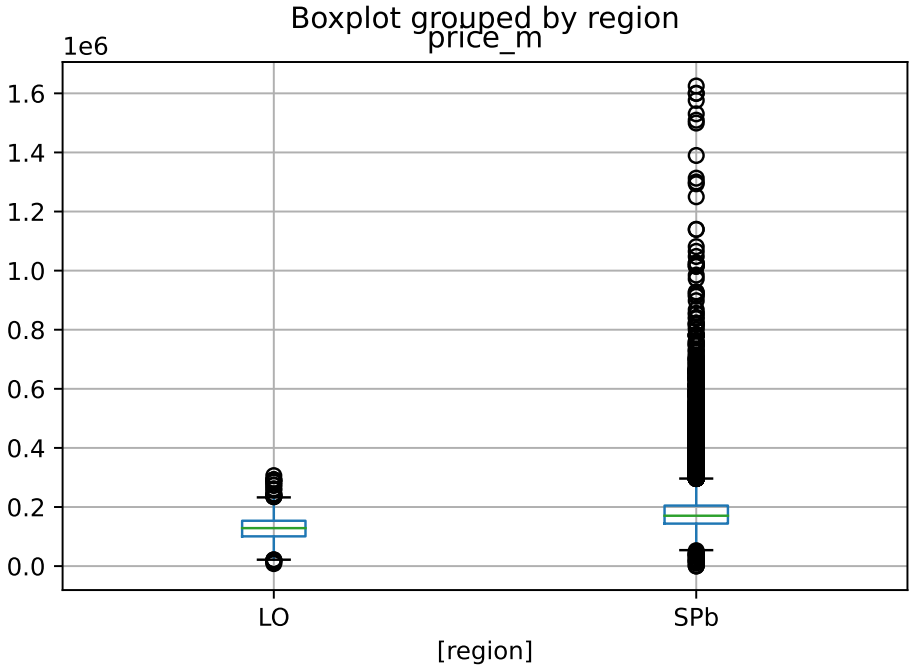

Рис. 8. Диаграмма «ящик с усами» для цен предложений квартир в Санкт-Петербургской агломерации в разрезе региональной принадлежности

ториях Ленинградской области, входящих в агломерацию Санкт-Петербурга, не отличается от нормального;

Таким образом всего будет выполнено 6 тестов, результаты которых сведены в таблицу 4. В отношении данных по Санкт-Петербургу все три теста позволили отклонить *H0(SPb)*, два из трёх тестов также позволили отклонить *H0(LO)*. На основании данных результатов можно сделать вывод о том, что распределение, одной из выборок однозначно отличается от нормального, второй — отличается от нормального с высокой вероятностью. В связи с этим, применение параметрических тестов для сравнения двух выборок является неуместным, вследствие чего следует использовать рассмотренный выше U-тест. Теперь остаётся только провести сам U-тест. Для этого используем скрипт 13. Его результаты представлены в таблице 5. Поскольку p-значение меньше заданного уровня значимости, можно сделать практический вывод о том, что различия в показателях стоимости объектов, расположенных в границах Санкт-Петербурга, и объектов, расположенных на территориях его агломерации, расположенных в Ленинградской области, являются существенными и требуют соответствующий учёт. Другие интерпретации результата могут быть получены из столбца «Альтернативная гипотеза (H1)» таблицы 3.

Листинг 7. Тест Шапиро-Уилка для данных по Санкт-Петербургу

```
stat, p = shapiro(dfs['price_m'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
if p < alpha:
print('Sample does not look Gaussian (reject H0)')
else:
print('Sample looks Gaussian (fail to reject H0)')
```

Листинг 8. Тест Шапиро-Уилка для данных по Ленинградской области

```
stat, p = shapiro(dfl['price_m'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
if p < alpha:
print('Sample does not look Gaussian (reject H0)')
else:
print('Sample looks Gaussian (fail to reject H0)')
```

Листинг 9. Тест K2 Агостино для данных по Санкт-Петербургу

```
stat, p = normaltest(dfs['price_m'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
if p < alpha:
print('Sample does not look Gaussian (reject H0)')
else:
print('Sample looks Gaussian (fail to reject H0)')
```

Листинг 10. Тест K2 Агостино для данных по Ленинградской области

```
stat, p = normaltest(dfl['price_m'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
if p < alpha:
print('Sample does not look Gaussian (reject H0)')
else:
print('Sample looks Gaussian (fail to reject H0)')
```

Листинг 11. Тест Андерсона-Дарлинга для данных по Санкт-Петербургу

```
result = anderson(dfs['price_m'])
print('Statistic: %.3f' % result.statistic)
p = 0
for i in range(len(result.critical_values)):
sl, cv = result.significance_level[i], result.critical_values[i]
if result.statistic < result.critical_values[i]:
print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))
else:
print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))
```

Листинг 12. Тест Андерсона-Дарлинга для данных по Ленинградской области

```
result = anderson(dfl['price_m'])
print('Statistic: %.3f' % result.statistic)
p = 0
for i in range(len(result.critical_values)):
sl, cv = result.significance_level[i], result.critical_values[i]
if result.statistic < result.critical_values[i]:
print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))
else:
print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))
```

Листинг 13. Проведение теста Манна-Уитни-Уилкоксона для данных удельных цен предложения квартир в агломерации Санкт-Петербурга

```
stat, p = mannwhitneyu(dfs['price_m'], dfl['price_m'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p < 0.05:
print('Probably different distributions')
else:
print('Probably the same distribution')
```

4.3. Реализация на R

Язык программирования R не столь распространён как Python, хотя и пользуется достаточной популярностью в развитых странах. В Северной Евразии область его применения является достаточно нишевой и, чаще всего, он используется в научной деятельности, в особенности в области биологии и химии. Для специалиста по машинному обучению знание данного языка является скорее бонусом, но не основным навыком. Тем не менее, следует отметить достоинства R, к которым можно отнести:

- большой набор библиотек и функций, существенное превосходящий набор средств Python;
- очень хорошие средства визуализации результата;
- удобные инструменты разработки веб-приложений, например, Shiny;
- язык является не компилируемым, а интерпретируемым, что зачастую удобнее в случае решения конкретных задач.

Последнее обстоятельство является, пожалуй, главным аргументом в пользу включения языка R в цикл публикаций по искусственному интеллекту для оценщиков. Если Python как язык общего назначения изначально предназначен для создания компилируемых исполняемых приложений, R разработан для пошагового анализа данных и представления всех промежуточных результатов.

Выбор основного языка программирования, используемого оценщиком, зависит от конкретной задачи: в случае разработки крупных комплексных решений предпочтительнее использование Python. В ситуациях, когда целью является решение частной задачи, в особенности требующей серьёзной визуализации результата, есть смысл обратить внимание на R. В любом случае, оба этих языка обладают достаточным набором средств для решения всего спектра задач по анализу данных, возникающих в процессе оценки стоимости.

При написании кода на R была использована его версия 4.2.0 (2022-04-22) "—"Vigorous Calisthenics", а также IDE RStudio (RStudio 2022.02.2+485 "Prairie Trillium" Release (8acbd38b0d4ca3c86c570cf4112a8180c48cc6fb, 2022-04-19) for Ubuntu Bionic).

Рассмотрим ещё одну практическую задачу на примере набора данных о рынке жилья города Алматы, предоставленный профессором университета «Нархоз» G. Shoulenbaeva. Файл с данными доступен по ссылке[6]. Рассматриваемый набор данных содержит 2355 наблюдений, а также 12 переменных, содержащих сведения о значениях признаков наблюдений. Одна из переменных содержит сведения о том, предлагается ли квартира к продаже вместе с мебелью и бытовой техников или без них. Возможны три варианта значения переменной:

- продажа квартиры без мебели и техники;
- продажа квартиры с частичным оснащением предметами интерьера и техники;

- продажа полностью оснащённой квартиры.

Сформулируем задачу: необходимо установить наличие либо отсутствие влияния оснащения квартиры предметами движимого имущества на её стоимость. Данная задача, по мнению автора, представляет определённый теоретический и практический интерес. Во-первых, теория оценки гласит, что при определении стоимости объекта недвижимости, следует учитывать стоимость только неотделимых улучшений объекта, тогда как стоимость элементов, являющихся движимым имуществом, следует исключать из стоимости самого объекта. При этом, на практике, зачастую невозможно точно определить принадлежность того или иного элемента к отделимым либо неотделимым улучшениям, а также определить их наличие у объектов-аналогов. Математический анализ данных рынка позволит ответить на вопрос, существует ли данная проблема в принципе, либо влияние фактора наличия улучшений, имеющих признаки отделимых, слишком несущественно и в любом случае не может быть корректно учтено при проведении оценки. Во вторых, решение данной задачи даст новые знания о конкретном рынке недвижимого имущества. Для дальнейшего анализа будем считать, что существуют только два варианта:

- продажа без потенциально отделимых улучшений и движимого имущества;
- продажа вместе с потенциально отделимыми улучшениями и движимым имуществом.

Решение объединить две категории в одну продиктовано, во-первых, математическими ограничениями U-теста, предназначенного для сравнения только двух выборок (для анализа более чем двух выборок существует непараметрический тест Краскела”– Уоллиса также известный как односторонний ранговый ANOVA [32]), во-вторых, с точки зрения обозначенной выше теоретической проблемы важно понять, оказывает ли влияние на стоимость факт наличия каких-либо отделимых улучшений как таковых, в третьих, деление объектов на частично и полностью оснащённые могло носить несколько субъективный характер.

При написании кода на R автор использовал его версию 4.2.0, а также IDE RStudio (version 2022.02.2 Build 485). При начале работы следует подключить необходимые библиотеки, задать некоторые константы, а также установить адрес рабочего каталога, например так, как это показано в скрипте 14.

Далее необходимо создать датафрейм на основе существующего текстового файла с данными. Затем в целях оптимизации использования ресурсов желательно оставить только необходимые переменные ”price.m”furniture”и преобразовать его в более удобный и современный формат ”tibble”(скрипт 15).

5. Выводы

Листинг 14. Подключение библиотек и задание значений констант и адреса рабочего каталога

```
# activate libraries
library(tidyverse)
library(moments)
library(RCurl)

# set constants
options(''scipen''=999, ''digits''=3)
set.seed(19190709)

# set work catalog
setwd(''~/.../Mann-Whitney-Wilcoxon/'')
```

Листинг 15. Подключение библиотек и задание значений констант и адреса рабочего каталога

```
# create data set from file, create subset with needed variables,
# change the type of object to a more convenient and modern one
almatyFlats <- read.csv('almaty-apts-2019-1.csv', header = TRUE, sep =
'','', dec = ''.'')
myvars <- c(''price.m'', ''furniture'')
almatyFlats <- almatyFlats[myvars]
as_tibble(almatyFlats)
```

Таблица 3. Нулевая и альтернативная гипотезы при анализе данных Санкт-Петербургской городской агломерации

Тип гипотезы	Нулевая гипотеза (H0)	Альтернативная гипотеза (H1)
Научная	Распределение удельных показателей стоимости квартир, расположенных в границах Санкт-Петербурга, и квартир, расположенных на прилегающих к нему территориях Ленинградской области, одинаково, сдвиг между ними отсутствует, статистические оценки, сделанные для множества объектов-аналогов, расположенных в одной части агломерации, являются несмещёнными для объектов, расположенных в другой.	Распределение удельных показателей стоимости квартир, расположенных в границах Санкт-Петербурга отличается от распределения удельных показателей стоимости квартир, расположенных на прилегающих к нему территориях Ленинградской области, существует сдвиг, оценка, сделанная для объектов, расположенных в одной части агломерации, будет смещённой для объектов, расположенных в другой её части.
Практическая	Медиана удельного показателя стоимости квартир, расположенных в границах Санкт-Петербурга равна медиане удельного показателя стоимости квартир, расположенных на прилегающих территориях Ленинградской области	Медиана удельного показателя стоимости квартир, расположенных в границах Санкт-Петербурга не равна медиане удельного показателя стоимости квартир, расположенных на прилегающих территориях Ленинградской области.
Изложенная в терминах оценки	Расположение квартиры в границах Санкт-Петербурга либо на прилегающих к нему территориях Ленинградской области не является существенным различием и не требует какого-либо специального учёта.	Расположение квартиры в границах Санкт-Петербурга либо на прилегающих к нему территориях Ленинградской области является существенным различием и требует отдельный учёт.

Таблица 4. Нулевая и альтернативная гипотезы при анализе данных Санкт-Петербургской городской агломерации ($\alpha = 0.05$)

Тест	Санкт-Петербург	Ленинградская область
Шапиро—Уилка:	7	8
статистика критерия (W)	0.689	0.991
p-значение	0.000	0.000
H0	отклоняется	отклоняется
K^2 Д'Агостино:	9	10
статистика критерия (K^2)	28166.251	4.067
p-значение	0.000	0.131
HO	отклоняется	не может быть отклонена
Андерсона—Дарлинга:	11	12
статистика критерия (A^2)	1688.671	15.795
H0:	отклоняется	отклоняется
Итоговый вывод:		
H0	отклоняется	отклоняется

Таблица 5. Нулевая и альтернативная гипотезы при анализе данных Санкт-Петербургской городской агломерации ($\alpha = 0.05$)

Показатель	Значение
Статистика критерия	142555441.000
p-значение	0.000
Нулевая гипотеза (см. таблицу 3)	отклоняется

Источники информации

[1] T. W. Anderson; D. A. Darling. «Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes». English. В: *Annals of Mathematical Statistics* 23.2 (1952), с. 193—212. DOI: 10.1214/aoms/1177729437. URL: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-23/issue-2/Asymptotic-Theory-of-Certain-Goodness-of-Fit-Criteria-Based-on/10.1214/aoms/1177729437.full (дата обр. 29.05.2022).

[2] J. Heinhold и K. W. Gaede. *Ingenieur-Statistik*. 1965, с. 327.

[3] S. S. Shapiro и M. B. Wilk. «An analysis of variance test for normality (complete samples)». English. В: *Biometrika* 52 (3-4 1965-12-01), с. 591—611.

[4] А. И. Кобзарь. *Прикладная математическая статистика*. 2006.

[5] Министерство финансов России. *Международный стандарт финансовой отчётности (IFRS) 13 «Оценка справедливой стоимости»*. с изменениями на 11 июля 2016 г. Russian. Russia, Moscow: Минфин России, 2015-12-28. URL: https://normativ.kontur.ru/document?moduleId=1&documentId=326168#l0 (дата обр. 10.06.2020).

[6] G. Shulenbaeva. *almaty-apts-2019-1*. Под ред. K. A. Murashev. 2019. URL: https://github.com/Kirill-Murashev/AI_for_valuers_R_source/tree/main/datasets/almaty_apts_2019_1.csv.

[7] K. A. Murashev. «Short Introduction to the differences between Frequentist and Bayesian approaches to probability in valuation». В: (2021-10-10). URL: https://github.com/Kirill-Murashev/AI_for_valuers_book/tree/main/Parts-Chapters/Frequentist-and-Bayesian-probability (дата обр. 23.05.2022).

[8] 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/ (дата обр. 11.05.2022).

[9] 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/.

[10] K. A. Murashev. «Practical application of the Mann-Whitney-Wilcoxon test (U-test) in valuation». English, Spanish, Russian, Interslavic. В: (2022-05-15). URL: https://github.com/Kirill-Murashev/AI_for_valuers_book/tree/main/Parts&Chapters/Mann-Whitney-Wilcoxon (дата обр. 15.05.2022).

[11] URL: https://github.com/Kirill-Murashev/AI_for_valuers_book/blob/main/Parts-Chapters/Mann-Whitney-Wilcoxon/U-test.py.

[12] URL: https://github.com/Kirill-Murashev/AI_for_valuers_book/blob/main/Parts-Chapters/Mann-Whitney-Wilcoxon/U-test.ipynb.

[13] Creative Commons. *cc-by-sa-4.0*. URL: https://creativecommons.org/licenses/by-sa/4.0/ (дата обр. 27.01.2021).

[14] Python Software Foundation. *Python site*. Английский. Python Software Foundation. URL: https://www.python.org/ (дата обр. 17.08.2021).

[15] The Document Foundation. *LibreOffice Calc*. Английский. URL: https://www.libreoffice.org/discover/calc/ (дата обр. 20.08.2021).

[16] *GeoGebra official site*. URL: https://www.geogebra.org/ (дата обр. 26.08.2021).

[17] *Jupyter site*. URL: https://jupyter.org (дата обр. 13.05.2022).

[18] Machinelearning.ru. *U-критерий Манна-Уитни*. 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 (дата обр. 14.05.2022).

[19] Machinelearning.ru. *Гипотеза сдвига*. 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 (дата обр. 15.05.2022).

[20] Machinelearning.ru. *Критерий Уилкоксона двухвыборочный*. 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 (дата обр. 14.05.2022).

[21] Machineleraning.ru. *Критерий Уилкоксона для связных выборок*. 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%BB%D1%8F_%D1%81%D0%B2%D1%8F%D0%B7%D0%BD%D1%8B%D1%85_%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BE%D0%BA (дата обр. 14.05.2022).

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

[23] PBC Studio. *RStudio official site*. Английский. URL: https://www.rstudio.com/ (дата обр. 19.08.2021).

[24] CTAN team. *TeX official site*. English. CTAN Team. URL: https://www.ctan.org/ (дата обр. 15.11.2020).

[25] LaTeX team. *LaTeX official site*. English. URL: https://www.latex-project.org/ (дата обр. 15.11.2020).

[26] *TeXLive official site*. URL: https://www.tug.org/texlive/ (дата обр. 15.11.2020).

[27] The R Foundation. *The R Project for Statistical Computing*. Английский. The R Foundation. URL: https://www.r-project.org/ (дата обр. 17.08.2021).

[28] Wikipedia. *Central Moment*. URL: https://en.wikipedia.org/wiki/Central_moment (дата обр. 29.05.2022).

[29] Wikipedia. *Commom Language Effect Size*. English. URL: https://en.wikipedia.org/wiki/Effect_size#Common_language_effect_size (дата обр. 16.05.2022).

[30] Wikipedia. *KISS principle*. URL: https://en.wikipedia.org/wiki/KISS_principle (дата обр. 06.11.2020).

[31] Wikipedia. *Kolmogorov–Smirnov test*. URL: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test (дата обр. 29.05.2022).

[32] Wikipedia. *Kruskal–Wallis one-way analysis of variance*. URL: https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance (дата обр. 31.05.2022).

[33] Wikipedia. *Kurtosis*. URL: https://en.wikipedia.org/wiki/Kurtosis (дата обр. 29.05.2022).

[34] Wikipedia. *Rank-biserial correlation*. English. URL: https://en.wikipedia.org/wiki/Effect_size#Rank-biserial_correlation (дата обр. 16.05.2022).

[35] Wikipedia. *Receiver operating characteristic*. URL: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (дата обр. 17.05.2022).

[36] Wikipedia. *Skewness*. URL: https://en.wikipedia.org/wiki/Skewness (дата обр. 29.05.2022).

[37] Wikipedia. *Standard score*. URL: https://en.wikipedia.org/wiki/Standard_score (дата обр. 17.05.2022).

[38] Benito van der Zander. *TeXstusio official site*. URL: https://www.texstudio.org/ (дата обр. 15.11.2020).