Практическое применение критерия Уилкоксона– Манна– Уитни в оценочной деятельности

Отбор признаков в качестве ценообразующх факторов на основе принципа не смещённых оценок

К. А. Мурашев

15 июня 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 [17], Международных стандартах оценки 2022 [19], а также в МСФО 13 «Оценка справедливой стоимости [13]. 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.

Основной материал состоит из четырёх блоков:

- описание теста Манна—Уитни—Уилкоксона (далее *U-mecm*), его вероятностного смысла и связи с другими математическими методами;
- практическая реализация U-теста в электронной таблице на примере тестовых случайных данных;
- практическая реализация U-теста на реальных данных рынка жилой недвижимости Санкт-Петербургской агломерации средствами языка программирования Python, целью анализа являлась проверка существенности различия удельной стоимости между объектами, расположенными в городской и пригородной частях агломерации;
- практическая реализация U-теста на реальных данных рынка жилой недвижимости города Алматы средствами языка программирования R, целью анализа являлась проверка существенности различия удельной стоимости между объектами, продаваемыми без отделимых улучшений и объектами, продаваемыми вместе с ними.

Актуальная версия данного материала, её исходный код, скрипты на Python и R, а также электронная таблица находятся в репозитории на портале GitHub и доступны по постоянной ссылке [20].

Данный материал и все приложения к нему распространяются на условиях лицензии cc-by-sa-4.0 [24].

Оглавление

Список таблиц

Список диаграмм

Листинги

Глава 1.

Технические данные

Данный материал, а также приложения к нему доступны по постоянной ссылке [20]. Исходный код данной работы был создан с использованием языка TEX [40] с набором макрорасширений LATEX 2_ε [41], дистрибутива TeXLive [42] и редактора TeXstudio [56]. Расчёт в форме электронной таблицы был выполнен с помощью LibreOffice Calc [28] (Version: 7.3.3.2, Ubuntu package version: 1:7.3.3 rc2-0ubuntu0.20.04.1 lo1 Calc: threaded). Расчёт на языке R [43] (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) [39]. Расчёт на языке Python (Version 3.9.12) [27] был выполнен с использованием среды разработки Jupyter Lab (Version 3.4.2) [31] 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) [38]. Графические материалы, использованные в подсекции 4.1, были подготовлены с использованием Geogebra (Version 6.0.666.0-202109211234) [29]. В данном материале как и в большинстве работ цикла были использованы следующие значения:

- уровень значимости — $\alpha = 0.05$;
- доверительный интервал — $Pr = 0.95$;
- начальное положение датчика псевдослучайных чисел — $seed = 19190709$.

В качестве десятичного знака используется точка. Большинство математических обозначений записаны так, как это принято в англоязычной среде. Например, тангенс обозначается как tan, а не tg. Результаты статистических тестов признаются значимыми в случае, когда

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

Данное решение основано, в частности на результатахдискуссии, имевшей место на портале researchgate.net [30].

Глава 2.

Предмет исследования

В случае работы с рыночными данными перед оценщиком часто встаёт задача проверки гипотезы о существенности влияния на стоимость того или иного признака, измеренного в количественной или порядковой шкале. 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 анализ рынка. Кроме того, в таком случае вряд ли можно говорить о серьёзной 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 на стоимость со стороны того или иного признака. 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 веб-скрепинга, часть — предоставлена

коллегами для анализа. Прилагаемая электронная таблица настроена таким образом, что тестовые исходные данные могут быть сгенерированы случайно.

Глава 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 [34];
- U-критерий Манна– Уитни, являющийся дальнейшим развитием вышеуказанном критерия, разработанный Генри Манном и Дональдом Уитни в 1947 году [32].

Забегая вперёд, можно сказать о том, что статистики данных критериев линейно связаны, а их p-значения почти одинаковы, что с практической точки зрения позволяет говорить о вариациях одного теста, а не о двух отдельных [34]. В данной работе по всему тексту используется общее название, а также его сокращённый вариант — U-тест, исторически относимый к критерию Манна—Уитни. Некоторые авторы [12] 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 [34]. Придерживаясь принципа KISS [47], 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 [35], 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 наличие каких-либо специальных правил отбора (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 $x \in X$ и $y \in Y$ вероятность того, что x больше y, равна вероятности того, что y больше x. На математической языке запись нулевой гипотезы выглядит следующим образом:

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

Для целостности теста требуется альтернативная гипотеза, которая заключается в том, что вероятность того, что значение признака у наблюдения из выборки X превышает его у наблюдения из выборки Y, отличается (в большую или меньшую сторону) 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:

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

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.

При более строгом наборе 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 вследствие малого p-значения.

Таким образом, более корректной интерпретацией U-теста является его использование для проверки именно гипотезы сдвига [33].

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:

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

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, по умолчанию, 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». При этом существуют и односторонние реализации, позволяющие ответить на вопрос о знаке различия значения признака у двух выборок.

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

- распределение значений количественного признака выборок отлично от нормального (в противном случае целесообразно использование параметрических t-критерия Стьюдента либо z-критерия 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, изложенные далее в таблице 3.1.

Таблица 3.1. Варианты нулевой гипотезы при использовании U-теста при оценке стоимости

Тип гипотезы	Формулировка
Научная	Две выборки 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

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 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

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

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 метода вычисления U-статистики: подсчёт вручную по формуле 3.4 либо применение специального алгоритма. Первый способ в силу трудоёмкости подходит только для сверхмалых выборок. 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 единицы для наименьшего из них. В случае наличия связок, т. е. групп повторяющихся значений (такой

группой могут являться в т. ч. и 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

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

где N — общее число наблюдений в обеих выборках.

3) Далее вычисляем U-значение для первой выборки:

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

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

Аналогичным образом вычисляется U-значение для второй выборки:

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

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

Из вышеприведённых формул следует, что

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

Также известно, что

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

Тогда

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

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 [46];
- рангово-бисериальный коэффициент корреляции (RBC) [51];
- принятый уровень значимости (как правило 0.05);
- расчётное p-значение.

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

3.2.3.0.1. 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

$$CLES = \frac{U_1}{n_1 n_2}. \tag{3.11}$$

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[52]. Таким образом, можно говорить о том, что данный показатель характеризует качество U-теста как бинарного классификатора.

$$CLES = f = AUC_1 = \frac{U_1}{n_1 n_2}. \tag{3.12}$$

Вопросы связи между U-статистикой и показателем (AUC) рассмотрены в 3.4.

3.2.3.0.2. 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 имеет область значения; $[-1; 1]$, при этом нулевое значение означает отсутствие какой-либо связи. Коэффициент рангово-бисериальной корреляции обычно обозначается 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:

$$H : x_i > y_j, \quad x \ \in X, y \in Y. \tag{3.13}$$

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. Таким образом, данная формула представляет собой формулу разности между показателями CLES для каждой из групп.

$$r = f - u = CLES_1 - CLES_2 = f - (1 - f) \tag{3.14}$$

Существует также ряд альтернативных формул, дающих идентичный результат:

$$r = 2f - 1 = \frac{2U_1}{n_1 n_2} - 1 = 1 - \frac{2U_2}{n_1 n_2}. \tag{3.15}$$

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) [54] может быть вычислено по формуле

$$z = \frac{U - m_U}{\sigma_U}, \tag{3.16}$$

где m_U — среднее арифметическое U, σ_U — её стандартное отклонение. Визуализация понятия *стандартизированное значения для нормального распределения* приведена на рисунке 3.1. Среднее для U вычисляется по формуле:

$$m_U = \frac{n_1 n_2}{2}. \tag{3.17}$$

Формула стандартного отклонения в случае отсутствия связок выглядит следующим образом:

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}. \tag{3.18}$$

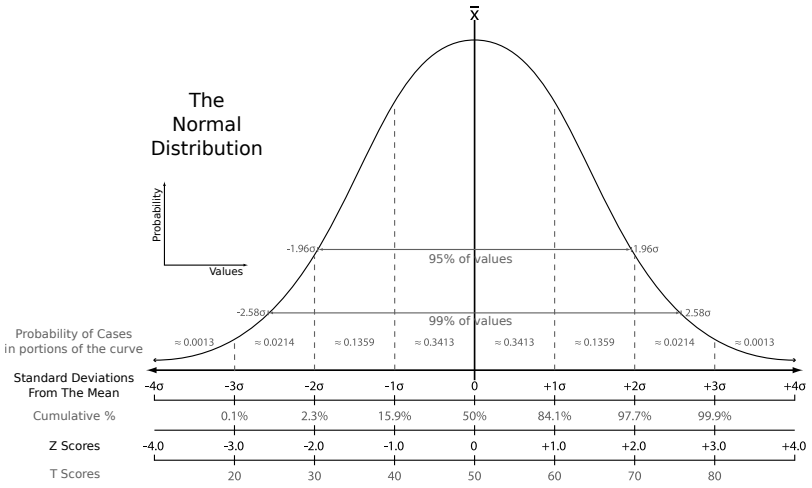

Рис. 3.1. Визуализация понятия стандартизированного значения (z-score) для нормального распределения [54]

В случае наличия связок используется другая формула:

$$\sigma_{U_{ties}} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum_{k=1}^{K} ({t_k}^3 - t_k)}{12n(n-1)}} = \sqrt{\frac{n_1 n_2}{12}\left((n+1) - \frac{\sum_{k=1}^{K} ({t_k}^3 - t_k)}{n(n-1)}\right)}, \tag{3.19}$$

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:

$$\begin{aligned} &p \leq 0.05 \Rightarrow \text{нулевая гипотеза отклоняется} \\ &p > 0.05 \Rightarrow \text{нулевая гипотеза не может быть отклонена.} \end{aligned} \tag{3.20}$$

При этом существует и альтернативный вариант интерпретации:

$$\begin{aligned} &p < 0.05 \Rightarrow \text{нулевая гипотеза отклоняется} \\ &p \geq 0.05 \Rightarrow \text{нулевая гипотеза не может быть отклонена.} \end{aligned} \tag{3.21}$$

На сегодняшний день нет однозначной позиции, как следует интерпретировать ситуацию, когда $p = \alpha$. В данной работе используется вариант, описанный в 3.20.

3.3. Соотношение с другими статистическими тестами

3.3.1. Сравнение U-теста Манна-Уитни-Уилкоксона с t-тестом Стьюдента

Часто можно услышать о том, что U-тест является непараметрическим аналогом t-теста, предназначенным для данных, чьё распределение отличается от нормального. С чисто практической точки зрения действительно можно сказать, что в случае нормального распределения определение наличия существенной разницы между двумя выборками целесообразно проводить посредством t-теста, в случае отличия распределения от нормального — U-теста. Таким образом, можно сказать, что эти тесты используются в одной и той же цели.

При этом, математический смысл U-теста и t-теста существенно отличается. Как уже было сказано ранее U-тест предназначен для проверки нулевой гипотезы, заключающейся в том, что для случайно выбранных из двух выборок наблюдений $x \in X$ и $y \in Y$ вероятность того, что x больше y, равна вероятности того, что y больше x, альтернативная гипотеза несёт утверждение о том, что эти вероятности не равны. В то же время t-тест предназначен для проверки нулевой гипотезы о равенстве средних двух выборок, при этом альтернативная гипотеза заключается в том, что средние двух выборок не равны. В связи с этим, при сравнении этих тестов следует иметь ввиду, что в общем случае, U-тест и t-тест проверяют разные нулевые гипотезы, хотя и имеющие отчасти схожий практический смысл. Результат U-теста чаще всего очень близок к результату двухвыборочного t-теста для ранжированных данных. Далее в таблице 3.2 проводится общее сравнение U-теста с t-тестом.

3.3.2. Альтернативные тесты в случае неравенства распределений

В случае необходимости проверки стохастического упорядочивания двух выборок (т. е альтернативной гипотезы $P(Y > X) + 0.5P(Y = X) \neq 0.5$) без предположения о равенстве их распределений (т. е. когда нулевая гипотеза имеет вид: $P(Y > X) + 0.5P(Y = X) = 0.5$, но не $F(X) = G(Y)$), следует использовать более подходящие тесты. К ним относятся, в т. ч. тесты Брунера-Мунцеля [**Bruner-Munzel-test**], представляющий собой устойчивый к гетероскедастичности аналог U-теста и Флигнера-Поличелло [26], представляющий собой тест на равенство медиан. В частности, в случае использовании более общей нулевой гипотезы $P(Y > X) + 0.5P(Y = X) = 0.5$ U-тест может достаточно часто приводить к возникновению ошибки первого рода даже в случае работы с большими выборками (в особенности в случае неравенства дисперсий и существенно различающегося объёма выборок), вследствие чего в таких случаях использование альтернативных тестов будет предпочтительным [15]. Таким образом, в случае отсутствия предположения о равенстве распределений в случае справедливости нулевой гипотезы, использование альтернативных тестов будет являться предпочтительным.

В случае проверки гипотезы сдвига при существенно отличающихся распределениях U-тест может дать ошибочную интерпретацию значимости [11], вследствие

Таблица 3.2. Свойства U-теста относительно t-теста

Свойство	Описание
Применимость к порядковым данным	В случае работы с порядковыми (ранговыми), а не количественными данными применение U-теста является предпочтительным относительно применения t-теста, при этом следует помнить, что расстояние между соседними значениями вариационного ряда нельзя считать постоянным.
Робастность	Поскольку U-тест работает с суммой рангов, а не значений признаков, он реже чем, t-тест ошибочно указывает на значимость вследствие наличия выбросов. Однако, в целом U-тест больше подвержен ошибке первого рода в случае, когда данные одновременно обладают свойством гетероскедастичности и имеют распределение отличное от нормального.
Эффективность	В случае нормального распределения асимптотическая эффективность U-теста составляет $\frac{3}{4}\pi \approx 0.95$ от показателя t-теста [10]. В случае существенного отличия распределения от нормального и достаточно большого числа наблюдений эффективность U-теста существенно превышает эффективность t-теста [6]. Однако такое сравнение эффективности следует интерпретировать с осторожностью, поскольку U-тест и t-тест проверяют разные гипотезы и оценивают разные величины. В случае, например, потребности в сравнении средних значений применение U-теста не является оправданным в принципе.

чего в таких условиях предпочтительным будет использование варианта t-теста [44], предназначенного для случаев неравных дисперсий [11].

В ряде случаев может быть оправданным преобразование количественных данных в ранги и последующее проведение t-теста в том или ином его варианте в зависимости от предположений о равенстве дисперсий. При преобразовании количественных данных в порядковые исходные дисперсии не будут сохранены, их следует пересчитать для самих рангов. В случае равенства дисперсий подходящей непараметрической заменой F-теста [25] может являться тест Брауна-Форсайта.

3.3.3. Связь между U-тестом и задачами классификации

U-тест представляет собой частный случай модели упорядоченного выбора (ordered logit model) [36].

3.4. Связь между U-тестом и понятиями Receiver operating characteristic (ROC), Area under curve (AUC)

3.4.1. Основные сведения о ROC

Основываясь на сказанном в 3.3.3, можно сделать вывод о том, что U-тест является не только тестом для проверки гипотезы сдвига (либо иной аналогичной по смыслу), но и представляет собой некий классификатор. Забегая вперёд можно сказать, что смысл U-теста как классификатора заключается в следующем:

- существует «позитивный» исход сравнения двух случайных наблюдений, заключающийся в том, что наблюдение из X больше наблюдения из Y;
- проводится оценка доли суммы рангов «позитивных» элементов;
- как и в целом с ROC, в случае, если значение доли «позитивных» элементов превышает 0.5, это говорит о том, что классификатор в целом выполняет свою функцию, в случае равенства 0.5 — его эффективность равнозначна угадыванию с помощью подбрасывания монеты, в случае значения менее 0.5 — использование такого классификатора даёт обратный результат.

На первый взгляд, связь между U-тестом и ROC не выглядит очевидной. В данной секции будет предпринята попытка разобраться в том, почему эти понятия всё же имеют связь, и в чём заключается суть U-теста как классификатора.

Сам ROC-анализ не входит в периметр данного материала. Поэтому рассмотрим лишь его основные моменты.

ROC-кривая (ROC-curve: Receiver Operator Characteristic) — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак (true positive rate (*TPR*), называемой *чувствительностью алгоритма классификации*), и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак (false positive rate (*FPR*), величина *1-FPR* называется *специфичностью алгоритма классификации*) при варьировании порога решающего правила. Также известна как **кривая ошибок**. Анализ классификаций с применением ROC-кривых называется **ROC-анализом**.

Количественная интерпретация ROC даёт показатель AUC (Area Under Curve, площадь под кривой). AUC — это площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций (ось абсцисс). Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0.5 демонстрирует непригодность выбранного метода классификации (соответствует случайному угадыванию с помощью монеты). Значение менее 0.5 говорит, что классификатор действует с точностью до наоборот: если положительные результаты назвать отрицательными и наоборот, классификатор будет работать лучше [52].

Введём некоторые термины.

P — количество объектов в выборке, обладающих некоторым признаком (Condition positive).

N — количество объектов в выборке, не обладающих некоторым признаком (Condition negative).

TP — результат теста, корректно определивший наличие существующего в действительности (Positive) признака (True positive, истинно положительный).

TN — результат теста, корректно определивший отсутствие несуществующего в действительности (Negative) признака (True negative, истинно отрицательный).

FP — результата теста, ошибочно определивший наличие несуществующего в действительности (Negative) признака (False positive, ложно положительный).

FN — результат теста, ошибочно определивший отсутствие существующего в действительности (Positive) признака (False negative, ложно отрицательный).

На основании вышесказанного можно создать таблицу 3.3 сопряжённости результатов применения бинарного классификатора. Строки содержат данные о фактическом наличии либо отсутствии признака, столбцы — предсказанном (predicted) с помощью классификатора. Как видно из таблицы 3.4 бинарный классификатор может приво-

Таблица 3.3. Таблица сопряжённости результатов работы бинарного классификатора

Всего $P+N$	Predicted Positive (PP)	Predicted negative (PN)
Positive (P)	TP	FN, ошибка второго рода [55]
Negative (N)	FP, ошибка первого рода [55]	TN

дить к возникновению ошибок двух типов. Введём ещё несколько определений и определим формулы для расчёта вероятностей исходов его работы. Вероятность TPR может быть записана как

$$P_{TPR} = \mathbb{P}(1|x \in C_1), \tag{3.26}$$

что означает, что если объект x принадлежит классу C_1, то данный показатель оценивает вероятность того, что бинарный классификатор отнесёт объект x к этому классу. Вероятность FPR записывается как

$$P_{FPR} = \mathbb{P}(1|x \in C_0), \tag{3.27}$$

что означает вероятность того, что объект, принадлежащий классу C_0 будет ошибочно отнесён к классу C_1.

Как правило, принцип работы бинарного классификатора основан на сравнения измерения x с некоторым фиксированным порогом c. Из этого следует, что два предыдущих выражения можно переписать и объединить в систему.

$$\begin{cases} P_{TPR} = \mathbb{P}(x > c|x \in C_1) \\ P_{FPR} = \mathbb{P}(x > c|x \in C_0) \end{cases} \tag{3.28}$$

Таблица 3.4. Возможные исходы применения бинарного классификатора

Показатель	Формула	Альтернативные названия
TPR (SEN)	$TPR = \frac{TP}{P} = 1 - FNR \quad (3.22)$	True positive rate, Recall, Sensitivity, Probability of detection, hit rate, power, чувствительность
FPR	$FPR = \frac{FP}{N} = 1 - TNR \quad (3.23)$	False positive rate, probability of false alarm, fall-alarm
FNR	$FNR = \frac{FN}{P} = 1 - TPR \quad (3.24)$	False negative rate, miss rate
TNR (SPC)	$TNR = \frac{TN}{N} = 1 - FPR \quad (3.25)$	True negative rate, specifitity, selectivity, специфичность

Из этого следует, что ROC-кривая представляет собой диаграмму

$$P_{FPR}(c), P_{TPR}(c), \tag{3.29}$$

таким образом, построение кривой означает изменение значения порога c.

Рассмотрим пример [23]. Возьмём $f(x|C_0) = \mathcal{N}(0,1)$ и $f(x|C_1) = \mathcal{N}(2,1)$ в качестве функций плотности вероятностей C_0 и C_1 соответственно. Далее поэтапно построим ROC-кривую средствами языка Python. Далее рассмотрим диаграмму 3.2, построенную с помощью кода, приведённого в скрипте 3.1. Область, закрашенная синим цветом, показывает вероятность FPR, т. е. ложно-положительного обнаружения значимости, тогда как область, закрашенная зелёным цветом — плотность вероятности TPR, т. е. корректного обнаружения значимости. ROC-кривая показывает значения именно этих показателей. Вертикальная прерывистая линия — порог чувствительности c. В данной ситуации он находится на отметке 0 по оси абсцисс. В случае его смещения на отметку 1 площадь под кривой FPR (синий цвет) существенно уменьшится, т. е. снизится вероятность ложно-положительного обнаружения признака, однако, вместе с этим уменьшится и площадь TPR (зелёный цвет), что означает увеличение ложно-отрицательных результатов. Данная ситуация рассмотрена № на диаграмме 3.3.

Листинг 3.1. Построение диаграммы плотностей распределения вероятностей TPR и FPR

```
# Import Libraries
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Plot
f0 = stats.norm(0, 1)
f1 = stats.norm(2, 1)
fig, ax = plt.subplots()
xi = np.linspace(-2, 5, 100)
ax.plot(xi, f0.pdf(xi), label=r'$f(x|C_0)$')
ax.plot(xi, f1.pdf(xi), label=r'$f(x|C_1)$')
ax.legend(fontsize=16, loc=(1, 0))
ax.set_xlabel(r'$x$', fontsize=18)
ax.vlines(0, 0, ax.axis()[-1] * 1.1, linestyles='--', lw=3.)
ax.fill_between(xi, f1.pdf(xi), where=xi > 0, alpha=.3, color='g')
ax.fill_between(xi, f0.pdf(xi), where=xi > 0, alpha=.3, color='b')

# Save to .pdf
plt.savefig('Plot-ROC-step-1.pdf', bbox_inches='tight')
```

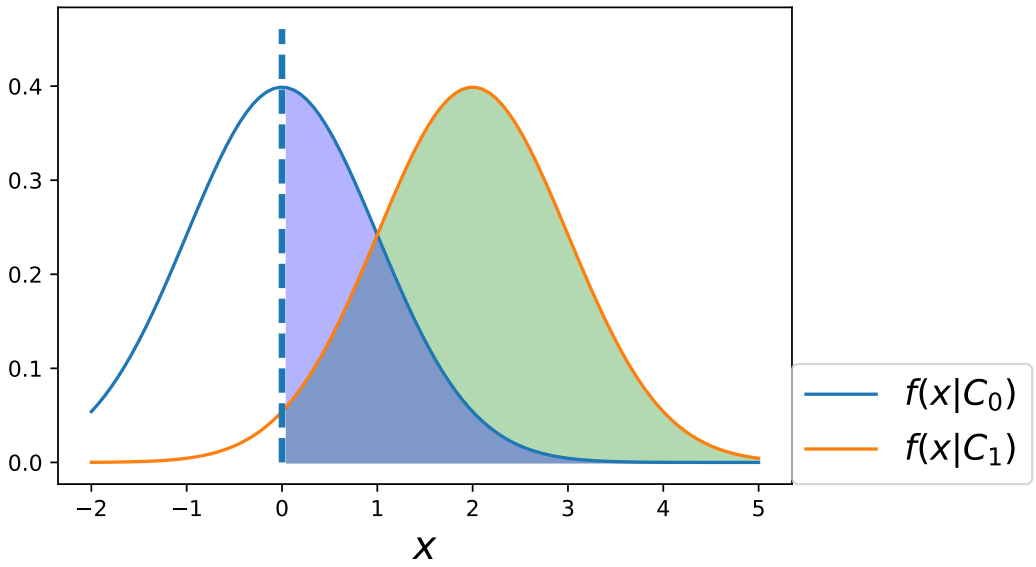

Рис. 3.2. Диаграмма плотностей распределения вероятностей TPR и FPR при пороговом значении 0

Как видно из приведённых диаграмм, повышение порога приводит к потере части как истинно-положительных, так и ложно-положительных результатов, уменьшение — к увеличению числа фиксации наличия признака в принципе. В предельных случаях, слишком низкое пороговое значение приведёт к тому, что все результаты будут интерпретированы как положительные, слишком высокое — к нулевому количеству наблюдений, у которых был обнаружен признак. Задача ROC-анализа заключается в выборе рационального порогового значения.

Добавим к уже имеющимся диаграмма ROC-кривые, соответствующие пороговым значениям 0 и 1. А также создадим интерактивную диаграмму с помощью кода, представленного в скрипте 3.2. Формат PDF не позволяет добавлять подобные интерактивные элементы, поэтому рассмотрим случаи с фиксированными значениями 0 и 1, представленные на диаграммах 3.4, 3.5 соответственно. В левой части каждой из них показаны уже знакомые графики функций плотности вероятности распределений TPR, FPR. В правой части показана ROC-кривая и точка, соответствующая заданному пороговому значению. Несложно догадаться, что координата x точки соответствует площади под кривой FPR, координата y — площади под кривой TPR. Увеличение порогового значения влечёт за собой смещение точки влево, уменьшение — вправо.

Чем лучше сам бинарный классификатор, тем ближе к левому верхнему углу будет проходить соответствующая ему ROC-кривая, поскольку в этом случае высокое значение TPR будет сочетаться с низким значением FPR. Бинарный классификатор, работающий также хорошо (на самом деле плохо) как алгоритм угадывания с помощью подбрасывания монеты (в том случае, если монета является «честной»)

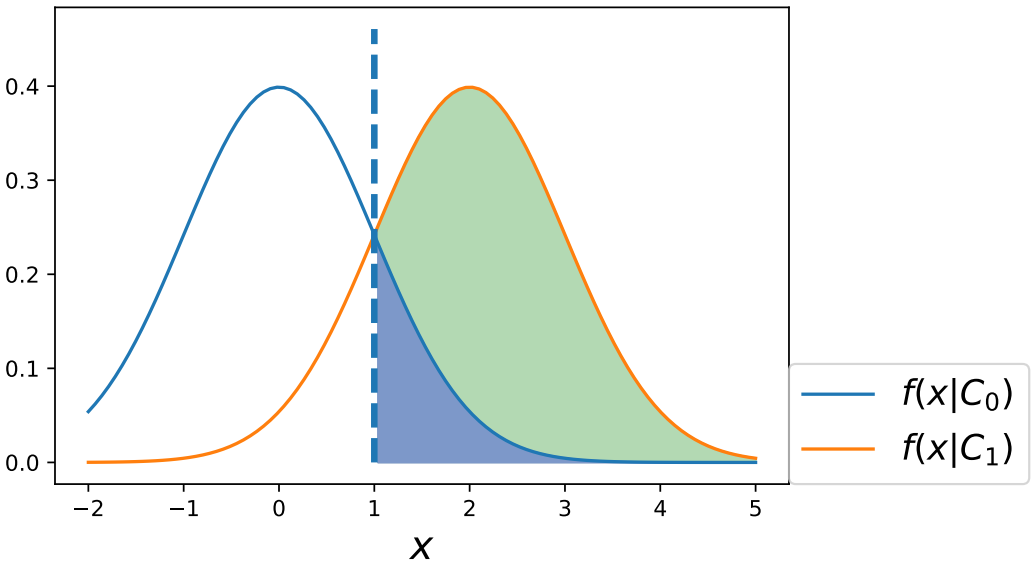

Рис. 3.3. Диаграмма плотностей распределения вероятностей TPR и FPR при пороговом значении 1

даёт ROC-кривую, представляющую собой прямой отрезок между точками (0,0) и (1,1). В таком случае левая часть диаграммы покажет полное совпадение кривых функций плотности вероятностей TPR, FPR. Такой случай показан на диаграмме 3.4. Для самостоятельной практики можно использовать скрипт 3.2, запуская его в среде Jupyter Lab, позволяющей использовать интерактивные возможности браузера.

Листинг 3.2. Построение интерактивной диаграммы плотности распределения TPR и FPR и соответствующей ей ROC кривой для заданного порогового значения

```
# Import Libraries
%matplotlib inline
from ipywidgets import interact
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Plot
f0 = stats.norm(0, 1)
f1 = stats.norm(2, 1)
fig, ax = plt.subplots()
xi = np.linspace(-2, 5, 100)
ax.plot(xi, f0.pdf(xi), label=r'$f(x|C_0)$')
ax.plot(xi, f1.pdf(xi), label=r'$f(x|C_1)$')
ax.legend(fontsize=16, loc=(1, 0))
ax.set_xlabel(r'$x$', fontsize=18)
ax.vlines(0, 0, ax.axis()[-1] * 1.1, linestyles='--', lw=3.)
ax.fill_between(xi, f1.pdf(xi), where=xi > 0, alpha=.3, color='g')
ax.fill_between(xi, f0.pdf(xi), where=xi > 0, alpha=.3, color='b')

# Plot ROC-curve and make all interactive
def plot_roc_interact(c=0):
xi = np.linspace(-3,5,100)
fig,axs = plt.subplots(1,2)
fig.set_size_inches((10,3))
ax = axs[0]
ax.plot(xi,f0.pdf(xi),label=r'$f(x|C_0)$')
ax.plot(xi,f1.pdf(xi),label=r'$f(x|C_1)$')
ax.set_xlabel(r'$x$',fontsize=18)
ax.vlines(c,0,ax.axis()[-1]*1.1,linestyles='--',lw=3.)
ax.fill_between(xi,f1.pdf(xi),where=xi>c,alpha=.3,color='g')
ax.fill_between(xi,f0.pdf(xi),where=xi>c,alpha=.3,color='b')
ax.axis(xmin=-3,xmax=5)
crange = np.linspace(-3,5,50)
ax=axs[1]
ax.plot(1-f0.cdf(crange),1-f1.cdf(crange))
ax.plot(1-f0.cdf(c),1-f1.cdf(c),'o',ms=15.)
ax.set_xlabel('False-alarm probability')
ax.set_ylabel('Detection probability')

interact(plot_roc_interact,c=(-3,5,.05))
%
```

Рис. 3.4. Диаграмма плотностей распределения вероятностей TPR и FPR при пороговом значении 0

3.4.2. Понятие AUC и её вычисление

Как следует из названия, AUC представляет собой площадь под ROC-кривой, ограниченную точкой, соответствующей заданному пороговому значению. В нормированном пространстве, в котором обычно и строится ROC-кривая, значение AUC эквивалентно вероятности того, что классификатор присвоит больший вес случайно выбранной положительной сущности, чем случайно выбранной отрицательной. AUC не зависит от конкретного порогового значения, поскольку ROC-кривая строится путем его перебора. Это означает, что AUC вычисляется путём интегрирования по пороговым значениям. AUC задаётся выражением:

$$AUC = \int P_{TPR}(P_{FPR})dP_{FPR}. \tag{3.30}$$

Пошаговый расчёт AUC выполняется следующим образом.

$$P_{TPR}(c) = 1 - F_1(c), \tag{3.31}$$

где F_1 — кумулятивная функция плотности для C_1. Аналогичным образом вычисляется

$$P_{FPR}(c) = 1 - F_0(c), \tag{3.32}$$

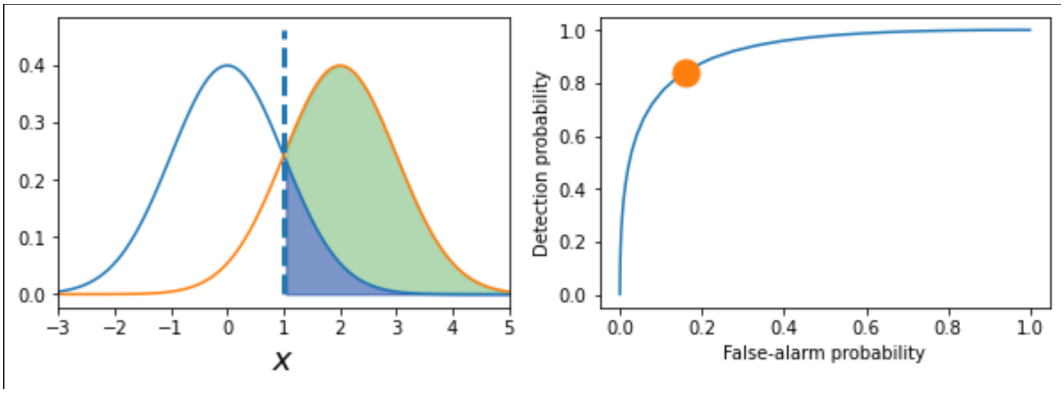

Рис. 3.5. Диаграмма плотностей распределения вероятностей TPR и FPR при пороговом значении 1

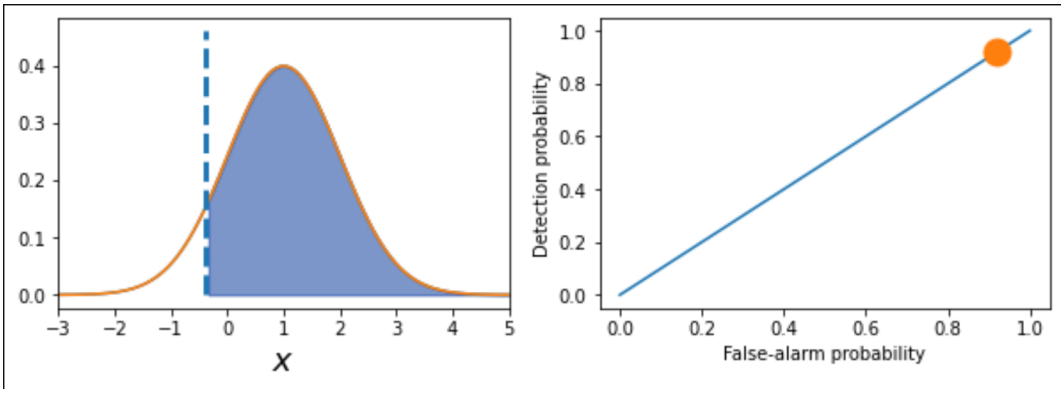

Рис. 3.6. Диаграмма плотностей распределения вероятностей TPR и FPR при равном среднем

где F_0 — кумулятивная функция плотности для C_0.

Возьмём некоторое конкретное значение c^*, которому соответствует определённая $P_{FPR}(c^*)$. Иными словами, ему соответствует вероятность того, что случайный элемент x_0, принадлежащий классу C_0 больше порогового значения c^*, т. е.

$$P_{FPR}(c^*) = \mathbb{P}(x_0 > c^* | x_0 \in C_0). \tag{3.33}$$

Тогда, рассуждая аналогичным образом относительно TPR, получим

$$P_{TPR}(c^*) = \mathbb{P}(x_1 > c^* | x_1 \in C_1). \tag{3.34}$$

Далее, опираясь на то, что AUC реализуется через интеграл, и подбирая таким образом, что распределение c^* соответствует распределению F_0. В таком случае P_{TPR} является самостоятельной случайной величиной, с соответствующим ожиданием в виде

$$\mathbb{E}(P_{TPR}) = \int P_{TPR} dP_{FPR} = AUC. \tag{3.35}$$

Теперь возможно сформулировать определение для AUC.

AUC — ожидаемая вероятность того, что элемент $x_1 \in C_1$ будет отнесён к C_1 с большей вероятностью, чем элемент $x_0 \in C_0$. Таким образом,

$$1 - F_1(t) > 1 - F_0(t) \forall t. \tag{3.36}$$

Формулировка «для любых t» означает, что $1 - F_1(t)$ *стохастически* больше $1 - F_0(t)$. Последнее обстоятельство является ключевым с точки зрения связи AUC с U-тестом, которая будет показана далее.

3.4.3. Связь между U-тестом и AUC

Ранее было приведено достаточно подробное описание U-теста. Данный параграф содержит только краткие сведения о нём, имеющие непосредственное отношение к вопросу его связи с AUC.

U-тест представляет собой непараметрический тест, позволяющий проверить принадлежность двух выборок одному распределению. Его основная идея заключается в том, что если между двумя классами отсутствует различие, то их объединение в один больший класс (множество) и последующее вычисление статистики (любой) для нового большего класса даст несмещённую оценку для любого из начальных классов. Иными словами, в случае отсутствия разницы в распределении у двух выборок, их объединение и предположение о том, что реально наблюдаемые данные двух выборок представляют собой лишь один из равнозначных вариантов перемещения наблюдений, означают отсутствие различия любой статистической оценки для любого варианта перемещения относительно другого, а также относительно объединённого множества.

Предположим, что нам нужно сравнить выборки посредством медианы, среднего или какой-либо иной меры центральной тенденции. С точки зрения кумулятивных функций распределения для двух популяций, в случае $H0$ мы имеем следующее:

$$H_0 : F_X(t) = F_Y(t) \quad \forall t, \tag{3.37}$$

что указывает на то, что все наблюдения принадлежат одному распределению. Тогда альтернативная гипотеза заключается в том, что

$$H_1 : F_X(t) < F_Y(t) \quad \forall t, \tag{3.38}$$

что возможно, в частности, в случае существования *сдвига* одного распределения относительно другого. В этом случае выборки $X_{ii=1}^{n}$, $X_{jj=1}^{m}$ представляют собой независимые группы наблюдений. При этом размер выборок может отличаться.

Методика теста заключается в объединении двух выборок в одно множество и присвоения рангов каждому элементу внутри него. U-статистика представляет собой сумму рангов для множества X. Если значение статистики достаточно мало, это означает, что распределение множества X стохастически смещено влево относительно распределения множества Y, т. е. $F_X t < F_Y t$.

Листинг 3.3. Вычисление p-значения для тестовых данных

```
print('p-value:',stats.wilcoxon(f1.rvs(30), f0.rvs(30))[1])
```

Поскольку при достаточно большом числе наблюдений (20 и более) распределение U-статистики хорошо аппроксимируется нормальным распределением, для оценки значимости подходит p-значение. Вычислим его с помощью языка Python согласно скрипту 3.3. p-значение составило $1.9729484515803686e-05$, что меньше уровня значимости 0.05, вследствие чего мы можем отклонить нулевую гипотезу 3.37. Поскольку данные были сгенерированы случайным образом, в случае повторения эксперимента, конкретное p-значение будет отличаться от полученного при написании данной работы. Однако оно всегда будет ниже порогового значения в силу заданных в алгоритме параметров.

U-статистика, в частности, может быть записана следующим образом:

$$U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(Y_j > X_i), \tag{3.39}$$

где $\mathbb{1}(Y_j > X_i)$ — характеристическая функция, показывающая, что статистика (для дискретного случая) оценивает вероятность того, что Y стохастически больше X. Таким образом, это соответствие означает, что её значение равно AUC. Связь между AUC и U-тестом заключается в схожей сути: проверке стохастического превышения значения наблюдений, принадлежащих одной выборке над наблюдениями, принадлежащими другой выборке.

Глава 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 в профессиональную деятельность, например языков программирования Python и R, идёт достаточно медленно. Кроме того, самостоятельный поэтапный 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 повторной 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 отсутствии 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 3.1. Следует отметить, что U-тест входит в периметр т. н. *частотного подхода к вероятности* (о различиях между *частотным* и *байесовским* подходами к вероятности приме-

нительно к оценке стоимости можно прочитать, в частности в [16]). 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 этого следует, что нулевая гипотеза чаще всего «пессимистична», т. е. несёт утверждение, что в основе исследуемого явления или процесса лежит случайность, вследствие чего мы не имеем возможность делать надёжные выводы. С учётом всего вышесказанного, сформулируем нулевую и альтернативную гипотезы (таблица 4.1) в трёх вариантах, согласно уровням строгости, показанным в таблице 3.1. Ячейки C2:C19 содержат некоторые описательные статистики. Для удобства первичного анализа бывает полезно показать свойства выборок графически. На рисунке 4.1 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 4.2, 4.3 изображены гистограммы распределения частот для первой и второй выборок соответственно, совмещённые с кривыми функции плотности вероятности для нормального распределения.

Как видно на диаграммах, форма распределения обеих выборок существенно отличается от формы кривой функции плотности вероятности нормального распределения. При работе с реальными данными в любом случае необходимо проводить количественные тесты проверки на нормальность распределений, 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 отсутствует потребность в отдельном построении общего вариационного ряда для двух выборок. Вместо этого можно сразу

Таблица 4.1. Нулевая и альтернативная гипотезы при анализе 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 одному множеству, будет смещённой для объектов, принадлежащих другому.
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.

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 вариационного ряда, тип сортировки: 0 — по убыванию, 1 — по возрастанию, в нашем случае необходимо указать 1. Столбцы L, N содержат 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 3.5.

Далее в ячейках C25, C26 по формулам 3.6, 3.7 вычислим соответственно значения U_1, U_2. После чего проверим корректность контрольного соотношения 3.10 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 3.11. Результат

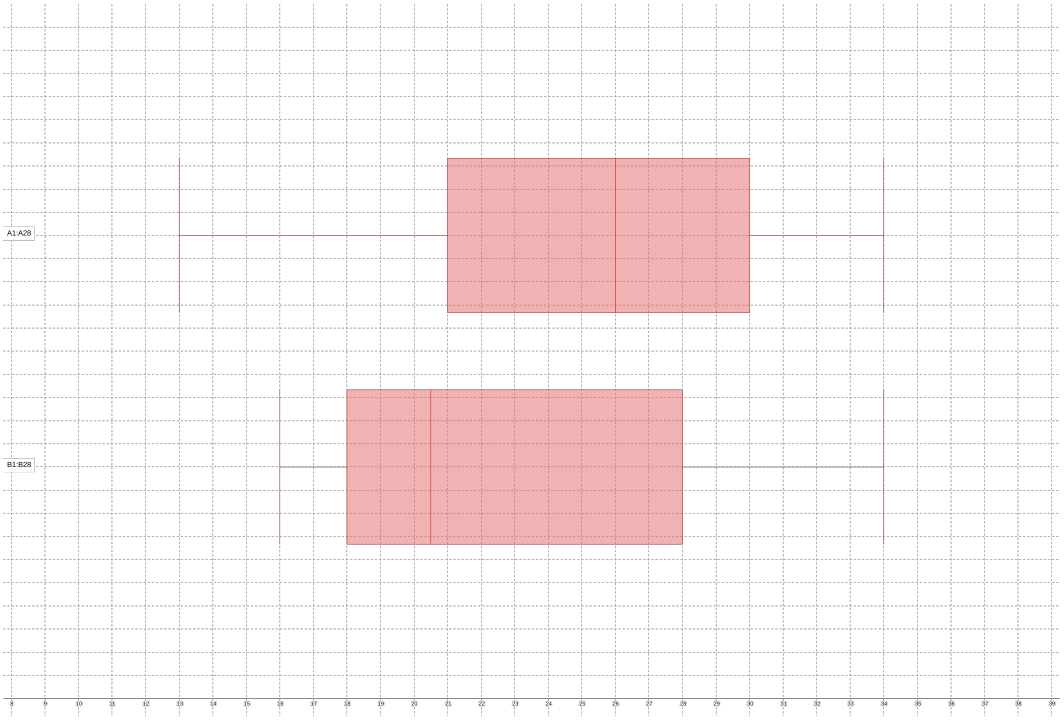

Рис. 4.1. Диаграмма «ящик с усами» (Boxplot) для обеих выборок

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 множества J превышает аналогичный показатель случайно выбранного наблюдения из множества I составляет 0.39477 (39.48 %)».

Далее рассчитаем значение коэффициента рангово-бисериальной корреляции по формулам 3.14, 3.15, разместив его в ячейке C36. В рассматриваемом случае значение составило 0.21, что говорит о том, что сила корреляционной связи между наличием у объекта признака «X» и удельным показателем его стоимости составляет 0.21.

После этого перейдём к расчёту стандартизированного значения согласно формуле 3.16. Для этого в ячейке C37 рассчитаем среднее по формуле 3.17, а затем перейдём к вопросу расчёта стандартного отклонения. Следует отметить, что для этого существуют две формулы: одна (3.18) применяется в случае отсутствия связок (ячейка C38), вторая (3.19) — 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

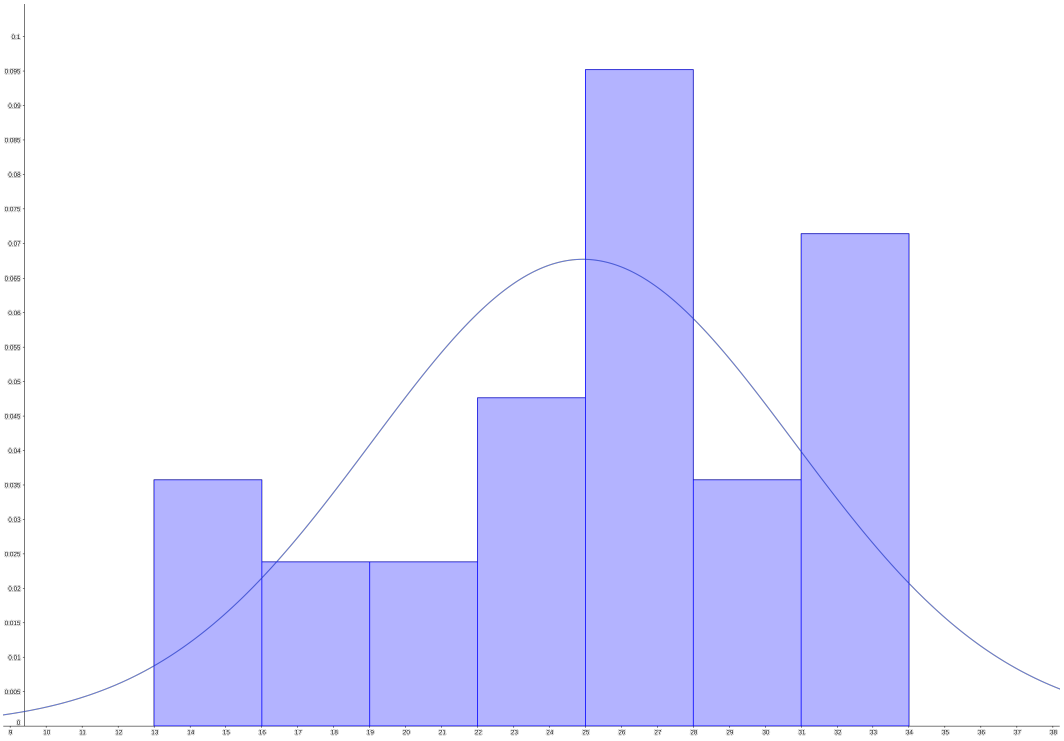

Рис. 4.2. Гистограмма первой выборки, совмещённая с кривой функции плотности вероятности для нормального распределения

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 3.19 от значений, полученных согласно 3.18, 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 фактора связок был осуществлён.

Зная среднее арифметическое и стандартное отклонение, вычисляем z-метку в ячейке C44 (поскольку одной из предпосылок U-теста является непрерывность распределения, а эмпирические данные имеют дискретное, при вычислении z-метки используется поправка), а затем, используя аппроксимацию стандартного нормального распределения, — p-значение. В рассматриваемом примере оно составило 0.173. Используя правило 3.20, приходим к выводу о невозможности отклонить нулевую гипотезу. Таким образом, используя формулировку, наиболее близкую к оценочной деятельности (см. таблицу 4.1), можно придти к следующему выводу: наличие или отсутствие признака «X» не оказывает сколько-нибудь заметного влияния на стоимость — признак «X» не является ценообразующим фактором.

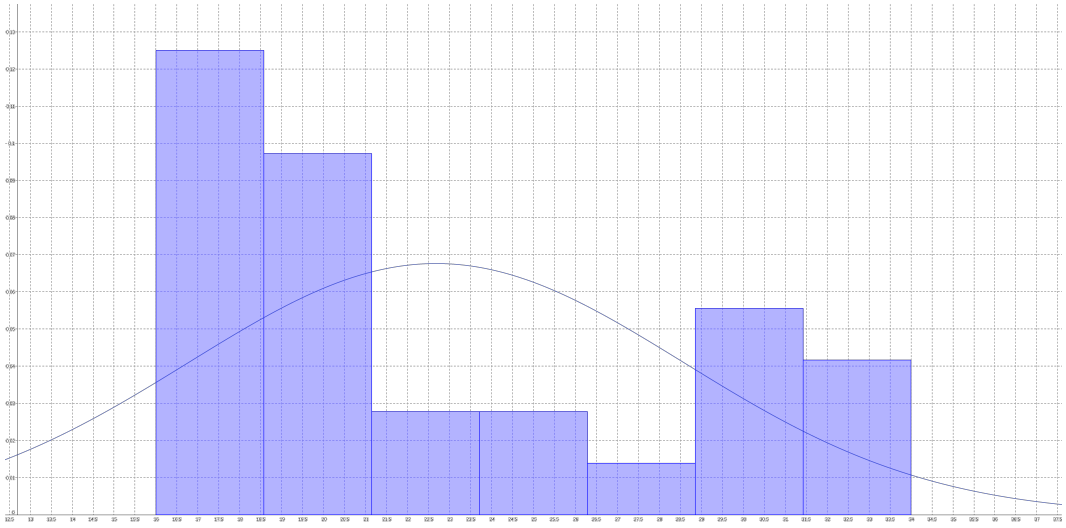

Рис. 4.3. Гистограмма второй выборки, совмещённая с кривой функции плотности вероятности для нормального распределения

Для лучшей интерпретации результата построим ROC кривую (диаграмма 4.4). Графические средства табличного процессора сильно уступают возможностям средств разработки, поэтому качество диаграммы оставляет желать лучшего, однако она всё же даёт возможность сделать некоторые интересные выводы. Следует отметить, что построение ROC кривой и её дальнейшая интерпретация носят лишь приблизительный характер и дают точность приближения в пределах нескольких процентов. Для точного анализа следует использовать профессиональные средства разработки, в частности описанные в подсекциях sections 4.2 and 4.3.

LibreOffice Calc не имеет штатных средств расчёта площади под кривой. Поэтому используем её аппроксимацию полиномом второй степени, получив в результате следующее выражение:

$$f(x) = -0.955836070665159x^2 + 1.83487119435351x + 0.021486995647532,$$

округляемое до

$$f(x) = -0.95584x^2 + 1.83487x + 0.02149.$$

Для нахождения площади под аппроксимирующей кривой решим определённый интеграл вида

$$\int_0^1 (-0.95584x^2 + 1.83487x + 0.02149)\ dx.$$

Преобразуем выражение в более удобное:

$$\int_0^1 (-\frac{2987x^2}{3125} + \frac{183487x}{100000} + \frac{2149}{100000})\ dx.$$

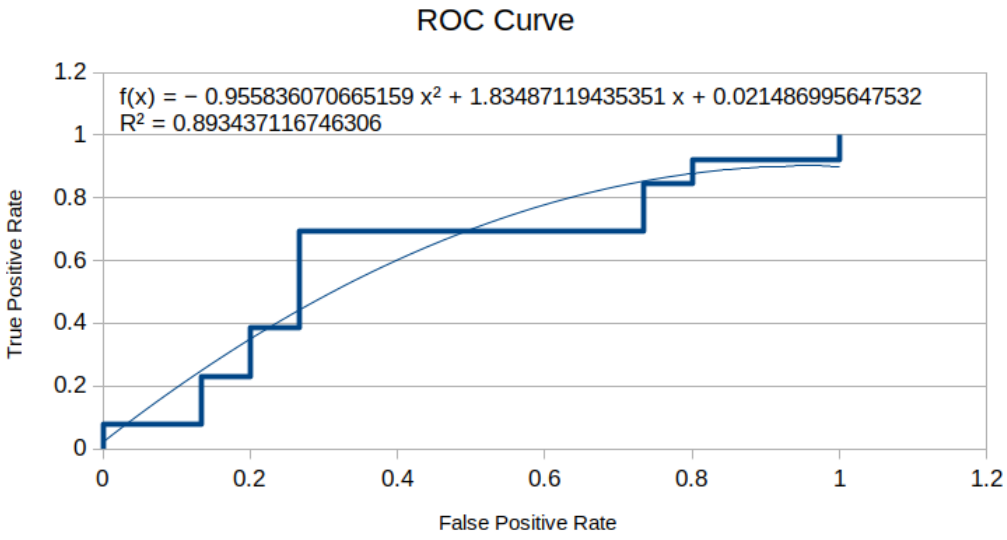

Рис. 4.4. ROC кривая для тестовых данных

Далее вычислим значение определённого интеграла.

Проинтегрируем выражение почленно и выделим константу:

$$-\frac{2987}{3125}\int_0^1 x^2\,dx+\frac{183487x}{100000}\int_0^1 x\,dx+\frac{2149}{100000}\times\int_0^1 1\,dx.$$

Известно, что первообразной функции x^2 является $\frac{x^3}{3}$,

тогда $\left(-\frac{2987x^3}{9375}\right)_0^1+\frac{183487x}{100000}\int_0^1 x\,dx+\frac{2149}{100000}\times\int_0^1 1\,dx.$

Рассчитаем антипроизводные первого члена в пределах и вычислим разность:

$$\left(-\frac{2987x^3}{9375}\right)_0^1=\left(-\frac{2987\times1^3}{9375}\right)-\left(-\frac{2987\times0^3}{9375}\right)=-\frac{2897}{9375},$$

подставив затем в общее выражение: $-\frac{2897}{9375}+\frac{183487x}{100000}\int_0^1 x\,dx+\frac{2149}{100000}\times\int_0^1 1\,dx.$

Таже известно, что первообразной функции x является $\frac{x^2}{2}$,

тогда $-\frac{2897}{9375}+\left(\frac{183487x}{100000}\right)_0^1+\frac{2149}{100000}\times\int_0^1 1\,dx.$

Рассчитаем антипроизводные второго члена в пределах и вычислим разность:

$$\left(\frac{183487x}{100000}\right)_0^1=\left(\frac{183487\times1^2}{100000}\right)-\left(\frac{183487\times0^2}{100000}\right)=\left(\frac{183487x}{100000}\right),$$

подставив затем в общее выражение:$\frac{359293}{600000}+\frac{2149}{100000}\times\int_0^1 1\,dx.$

Известно, что первообразной 1 является x,

тогда $\frac{359293}{600000}+\left(\frac{2149}{100000}\right)_0^1$

Рассчитаем антипроизводные третьего члена в пределах и вычислим разность:

$$\left(\frac{2149}{100000}\right)_0^1=\frac{2149\times1}{100000}-\frac{2149\times0}{100000}=\frac{2149}{100000},$$

подставив затем в общее выражение: $\frac{372187}{600000}=0.621312$

Полученное значение приблизительно соответствует значению показателя CLES, по мере увеличения степени аппроксимирующего полинома абсолютная разница между значениями AUC и CLES будет уменьшаться, стремясь к нулю при приближении степени полинома к $\frac{N}{2} - 1$.

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, а также 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 [21], код в формате Python Notebook доступен по ссылке [22].

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 [18]. Рассматриваемый набор данных содержит 34821 наблюдение. Как известно, 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 (см. таблицу 4.2).

Язык Python изначально не был создан специально для анализа данных. Поэтому в его базовой версии могут отсутствовать многие функции, необходимые для проведения расчётов. К счастью, 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 вели-

Листинг 4.1. Подключение необходимых библиотек

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

Листинг 4.2. Задание применяемого уровня значимости

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

ки как, например, у языка R, однако они являются исчерпывающими для более чем 95 % задач, стоящих перед оценщиками. Для решения задач, рассматриваемых в данном материале, потребуются следующие библиотеки: numpy, pandas, math, matplotlib.pyplot, scipy.stats. Для их подключения используем код, представленный в скрипте 4.1.

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. Для задания значения уровня значимости используется код 4.2.

После этого всё готово для начала работы. Создадим датафрейм на основе текстового файла, содержащего изучаемый набор данных (скрипт 4.3). 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

Листинг 4.3. Загрузка данных и создание датафрейма

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

Листинг 4.4. 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]]
```

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 4.4). В рассматриваемом случае такая микрооптимизация не играет большой роли, однако 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]:

$$k = \sqrt{n}, \tag{4.1}$$

где n — число наблюдений. Для построения гистограммы используем скрипт 4.5.

Рассмотрим полученную гистограмму 4.5. 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 для теоретического нормального распределения. 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 отдельных 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

Листинг 4.5. Построение гистограммы для агломерации Санкт-Петербурга

```
# calculate the number of observations on data frame
spbaLenR = round(math.sqrt(len(df1.index)))

# fit a normal distribution to the data: mean and standard deviation
mu, std = norm.fit(df1['price_m'])

# plot the histogram
plt.hist(df1['price_m'], bins=spbaLenR, density=True)

# plot the PDF
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)

plt.plot(x, p, 'k', linewidth=2)
title = "Fit Values: {:.2f} and {:.2f}".format(mu, std)
plt.title(title)

# save to .pdf
plt.savefig('spba-price-histogram-py.pdf')
```

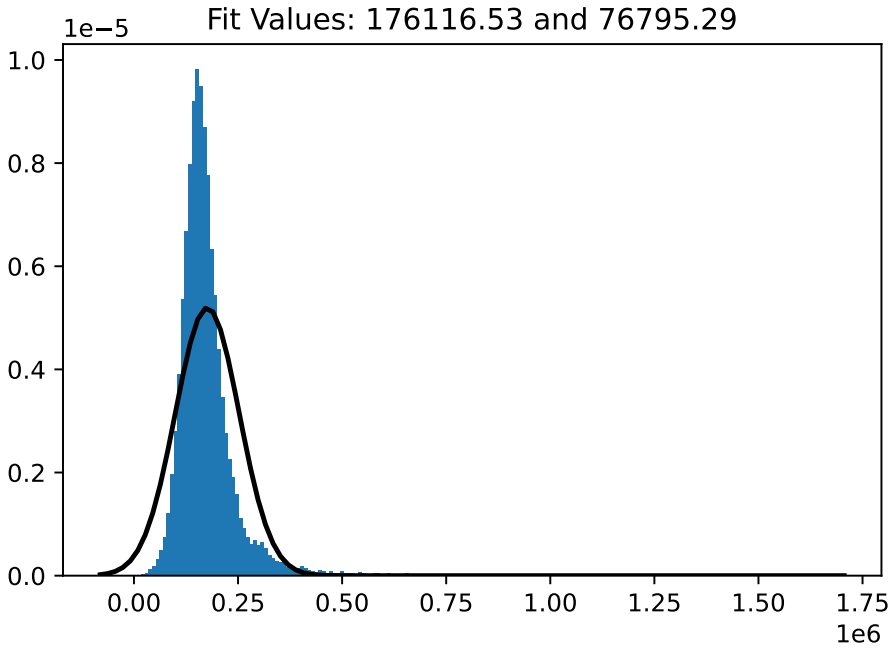

Рис. 4.5. Гистограмма плотности распределения цен за 1 кв. м квартир в Санкт-Петербургской агломерации, совмещённая с кривой функции плотности вероятности для нормального распределения

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, 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 (скрипты 4.7, 4.8).

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

Листинг 4.6. Создание отдельных датафреймов для Санкт-Петербурга и Ленинградской области

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

Листинг 4.7. Построение гистограммы для Санкт-Петербурга

```
# Saint-Petersburg
# calculate the number of observations on data frame
spbLenR = round(math.sqrt(len(dfs.index)))

# fit a normal distribution to the data: mean and standard deviation
muS, stdS = norm.fit(dfs['price_m'])

# plot the histogram
plt.hist(dfs['price_m'], bins=spbLenR, density=True)

# plot the PDF
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
ps = norm.pdf(x, muS, stdS)

plt.plot(x, ps, 'k', linewidth=2)
title = "S-Pb. Fit Values: {:.2f} and {:.2f}".format(muS, stdS)
plt.title(title)

# save to .pdf
plt.savefig('spb-price-histogram-py.pdf')
```

Листинг 4.8. Построение гистограммы для Ленингадской области

```
# LO
# calculate the number of observations on data frame
loLenR = round(math.sqrt(len(dfl.index)))

# fit a normal distribution to the data: mean and standard deviation
muL, stdL = norm.fit(dfl['price_m'])

# plot the histogram
plt.hist(dfl['price_m'], bins=loLenR, density=True)

# plot the PDF
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
pl = norm.pdf(x, muL, stdL)

plt.plot(x, pl, 'k', linewidth=2)
title = "LO. Fit Values: {:.2f} and {:.2f}".format(muL, stdL)
plt.title(title)

# save to .pdf
plt.savefig('lo-price-histogram-py.pdf')
```

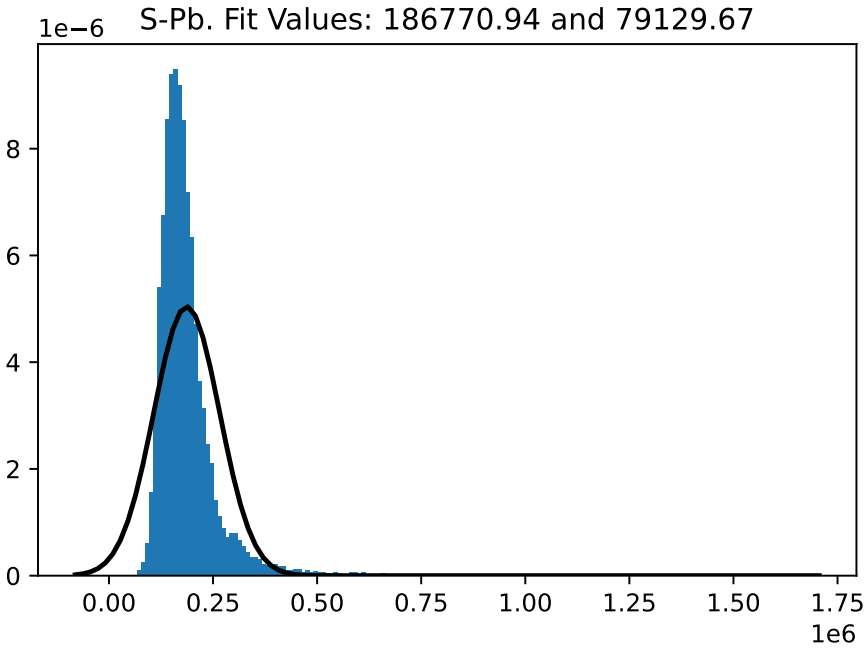

Рис. 4.6. Гистограмма плотности распределения цен за 1 кв. м квартир в Санкт-Петербурге, совмещённая с кривой функции плотности вероятности для нормального распределения

наблюдений в каждом диапазоне. Как видно из гистограммы 4.6, 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 на гистограмме 4.7 выглядит относительно симметрично.

Также построим график «ящик с усами» для обоих датафреймов с помощью скрипта 4.9 (см. диаграмму 4.8). 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 для быстрой первичной интерпретации, а также для презентационных целей. 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:

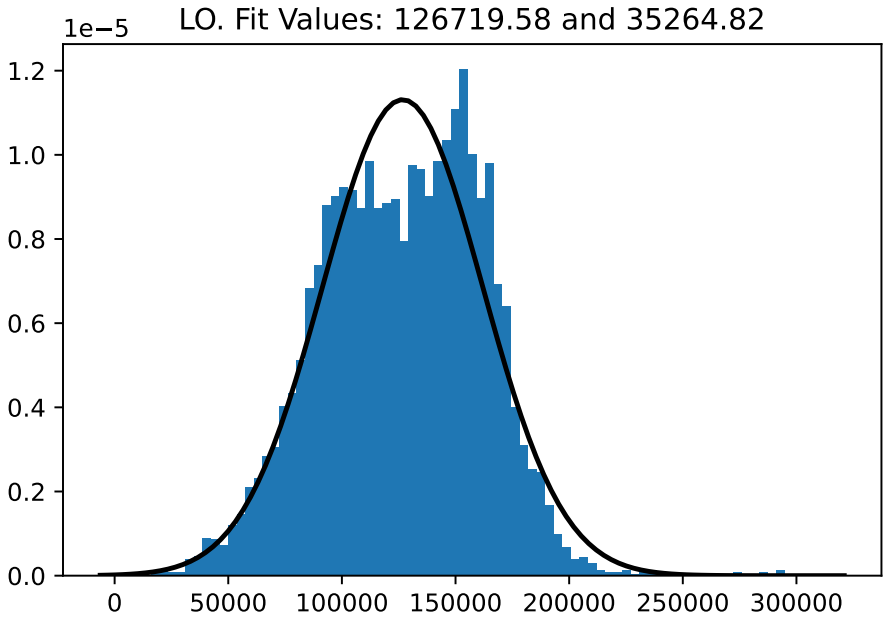

Рис. 4.7. Гистограмма плотности распределения цен за 1 кв. м квартир в Ленинградской области, расположенных в границах агломерации Санкт-Петербурга, совмещённая с кривой функции плотности вероятности для нормального распределения

- тест Шапиро—Франчиа [3];
- тест K^2 Д'Агостино [8];
- тест Андерсона—Дарлинга [1].

Тест Шапиро—Франчиа оценивает выборку данных и вычисляет, насколько вероятно, что она была взята из генеральной совокупности, имеющей нормальное распределение. Данный тест считается одним из наиболее мощных тестов проверки на нормальность [12]. При этом существуют некоторые предпосылки, указывающие на то, что он хорошо работает на выборках среднего размера, не превышающих пяти тысяч наблюдений (минимальное количество должно быть не менее пяти).

Тест K^2 Д'Агостино основывается на анализе показателей асимметрии [53] и эксцесса [50], представляющих собой третий и четвёртый центральные моменты [45] 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 [48] 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,

Листинг 4.9. Построение диаграммы «ящик с усами» (boxplot) для обеих подвыборок

```
# add labels to data
dfs[''region''] = ''SPb''
dfl[''region''] = ''LO''

# plot boxplot
prices = [dfs, dfl]
allPrices = pd.concat(prices)
plt.figure()
allPrices.boxplot(by=''region'')

# save to .pdf
plt.savefig('spb-lo-boxplot-py.pdf')
```

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 территориях Ленинградской области, входящих в агломерацию Санкт-Петербурга, не отличается от нормального.

Таким образом всего будет выполнено 6 тестов, результаты которых сведены в таблицу 4.3. 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 4.16. Его результаты представлены в таблице 4.4. 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 4.2.

Листинг 4.10. 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 4.11. 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 4.12. Тест К2 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 4.13. Тест К2 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 4.14. 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 4.15. 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 4.16. 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')
```

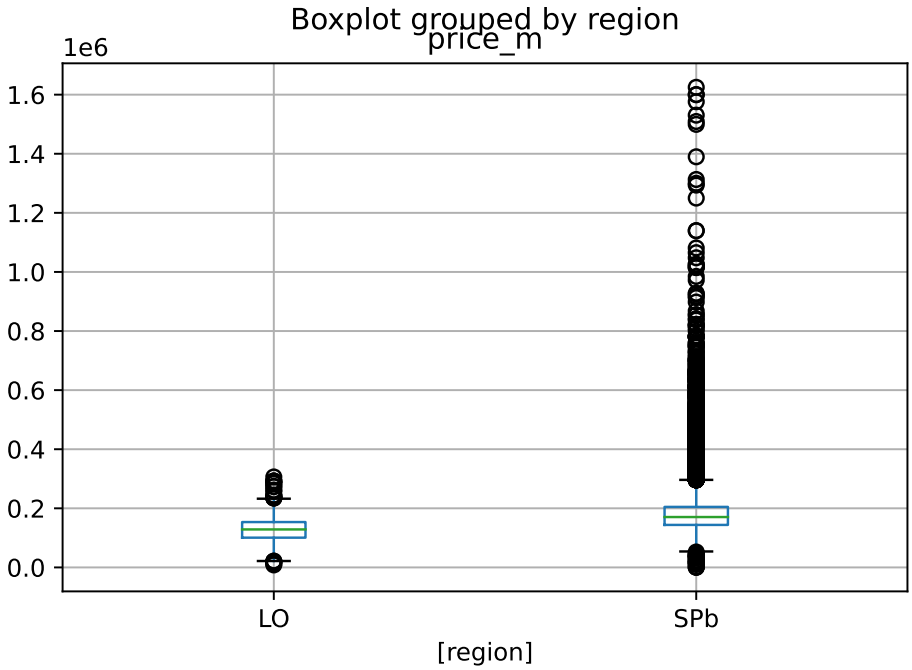

Рис. 4.8. Диаграмма «ящик с усами» для цен предложений квартир в Санкт-Петербургской агломерации в разрезе региональной принадлежности

Таблица 4.2. 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 4.3. Результаты проведения тестов проверки данных по Санкт-Петербургской агломерации на нормальность ($\alpha = 0.05$)

Тест	Санкт-Петербург	Ленинградская область
Шапиро—Уилка:	4.10	4.11
статистика критерия (W)	0.689	0.991
p-значение	0.000	0.000
H0	отклоняется	отклоняется
K^2 Д'Агостино:	4.12	4.13
статистика критерия (K^2)	28166.251	4.067
p-значение	0.000	0.131
HO	отклоняется	не может быть отклонена
Андерсона—Дарлинга:	4.14	4.15
статистика критерия (A^2)	1688.671	15.795
H0:	отклоняется	отклоняется
Итоговый вывод:		
H0	отклоняется	отклоняется

Таблица 4.4. Результаты проведения U-теста для данных Санкт-Петербургской агломерации ($\alpha = 0.05$)

Показатель	Значение
Статистика критерия	142555441.000
p-значение	0.000
Нулевая гипотеза (см. таблицу 4.2)	отклоняется

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, существенно 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[14]. 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 техникой 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 [49]), 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. Варианты нулевой и альтернативной гипотез приведены в таблице 4.5.

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 4.17.

Далее необходимо создать датафрейм на основе существующего текстового файла с данными. Затем в целях оптимизации использования ресурсов желательно оставить только необходимые переменные 'price.m' и 'furniture', а затем преобразовать датафрейм в более удобный и современный формат 'tibble' (скрипт 4.18). После этого необходимо рассчитать общее число наблюдений, а также каждого типа в зависимости от наличия отделимых улучшений (скрипт 4.19). Результаты подсчёта представлены в таблице 4.6.

В целях первичной визуализации данных построим гистограммы для всех наблюдений. Число столбцов на этот раз будет определено по формуле, разработанной

Листинг 4.17. Подключение библиотек и задание значений констант и адреса рабочего каталога

```
# activate libraries
library(tidyverse)
library(moments)
library(ggplot2)
library(gamlss)
library(normtest)
library(nortest)

# set constants
options('scipen'=2, 'digits'=3)
set.seed(19190709)

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

Листинг 4.18. Создание датафрейма и его настройка

```
# 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)
```

Листинг 4.19. Подсчёт количества наблюдений

```
# calculation of the total number of observations,
# as well as depending on the equipment
n.total <- nrow(almatyFlats)
n.non.equip <- NROW(almatyFlats$furniture[ which(almatyFlats$furniture
== 0)])
n.equip <- NROW(almatyFlats$furniture[ which(almatyFlats$furniture >
0)]))
```

Листинг 4.20. Создание функции для расчёта k по формуле P. W. Nowiczkij

```
# create function for second Nowiczkij formula
kHistNowiczkij2 <- function(x, na.omit = FALSE){ # create function,
ignore missed values
n <- NROW(x) # calculate n
kurt = kurtosis(x) # calculate kurtosis
kn2 = (((kurt^4)*(n^2))^(1/5))*(1/3) # calculate k
return(kn2) # return k
} # end of function
```

Листинг 4.21. Расчёт k по формуле P. W. Nowiczkij для наблюдений различных типов

```
# calculation numbers of k for different types of observations
k.all.data <- kHistNowiczkij2(almatyFlats$price.m)
k.non.equip <- kHistNowiczkij2(almatyFlats$price.m[
which(almatyFlats$furniture == 0)])
k.equip <-kHistNowiczkij2(almatyFlats$price.m[
which(almatyFlats$furniture > 0)])
```

в 1991 году P. W. Nowiczkij [9]:

$$k = \frac{1}{3}\sqrt[5]{\varepsilon^4 n^2} \equiv \frac{1}{3}\sqrt[5]{\frac{n^2}{\xi^8}}, \tag{4.2}$$

где ε — коэффициент эксцесса, ξ — коэффициент контрэксцесса. Для удобства, сначала создадим соответствующую функцию (скрипт 4.20). а затем рассчитаем рациональное число интервалов для всего набора данных (скрипт 4.21), а затем отдельно для квартир, предлагаемых к продаже без отделимых улучшений и с ними. Результаты расчёта приведены в таблице 4.6.

Построим гистограммы для всех наблюдений (диаграмма 4.9), тех, которые имеют оснащение 4.10, и тех, которые продаются без каких-либо отделимых улучшений 4.11. Из диаграмм следует, что распределения имеют тяжёлые правые хвосты, что косвенно указывает на то, что распределение во всех случаях отличается от нормального. Для обоснованного суждения в дальнейшем будут выполнены количественные тесты проверки нормальности. Код для построения гистограмм приведён в скрипте 4.22. В таблице 4.7 содержатся базовые описательные статистики для каждого из трёх типов наблюдений, построенные при помощи кода 4.23. Для лучшего восприятия с помощью скрипта также была построена диаграмма «ящик с усами» (boxplot) (скрипт 4.24), приведённая на рисунке 4.12. Как видно, медиана цен объектов, предлагаемых к продаже вместе с отделимыми улучшениями, выше медианы объектов, предлагаемых к продаже без них.

Листинг 4.22. Построение гистограмм для наблюдений различных типов

```
# plot the histogram, combined with the density curve of the theoretical
# normal distribution for all observations
histDist(almatyFlats$price.m,
     density = TRUE,
     nbins = kHistNowiczkij2(almatyFlats$price.m),
     xlab = 'price per meter, kaz tenge',
     ylab = 'probability',
     main = 'Price per meter histogram, all observations')

# plot the histogram, combined with the density curve of the theoretical
# normal distribution for observations without equipment
histDist(almatyFlats$price.m[ which(almatyFlats$furniture == 0)],
     density = TRUE,
     nbins = kHistNowiczkij2(almatyFlats$price.m[
     which(almatyFlats$furniture == 0)]),
     xlab = 'price per meter, kaz tenge',
     ylab = 'probability',
     main = 'Price per meter histogram, observations witout equipment')

# plot the histogram, combined with the density curve of the theoretical
# normal distribution for observations with equipment
histDist(almatyFlats$price.m[ which(almatyFlats$furniture > 0)],
     density = TRUE,
     nbins = kHistNowiczkij2(almatyFlats$price.m[
     which(almatyFlats$furniture > 0)]),
     xlab = 'price per meter, kaz tenge',
     ylab = 'probability',
     main = 'Price per meter histogram, observations witout equipment')
```

Листинг 4.23. Построение базовых описательных статистик для наблюдений различных типов

```
# summaries
summary(almatyFlats$price.m)
summary(almatyFlats$price.m[ which(almatyFlats$furniture == 0)])
summary(almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
```

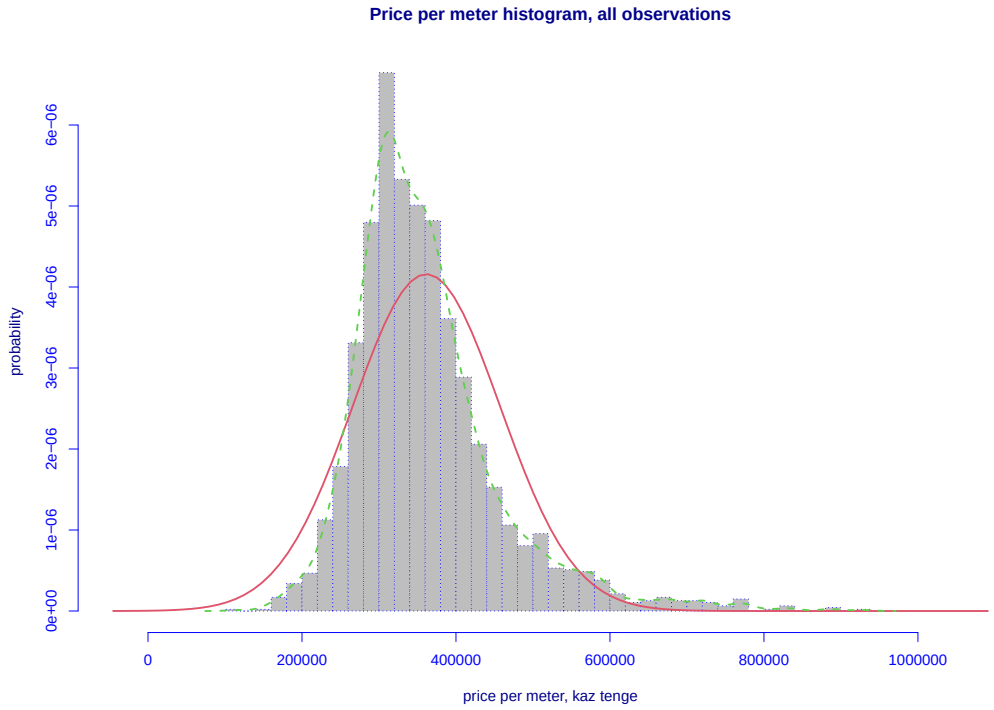

Рис. 4.9. Гистограмма цен предложения для всех объектов, совмещённая с кривой функции плотности эмпирического распределения, а также кривой функции плотности теоретического нормального распределения.

Следующий этап анализа заключается в проверке нормальности распределения удельных цен предложений. Как уже было сказано выше, язык R обладает очень богатым набором инструментов. Существует множество библиотек, предлагающих в общей сложности несколько десятков тестов. В данной работе были применены следующие тесты:

- критерий Шапиро—Уилка [3];
- критерий Шапиро—Франчиа [5];
- критерий Андерсона—Дарлинга [1];
- скорректированный критерий Харка—Бера [7];
- критерий Колмогорова—Смирнова с поправкой Лилиефорса [4].

Код для проведения тестов для объектов, продаваемых без отделимых улучшений, приведён в скрипте 4.25, вместе с ними — 4.26. Результаты тестов для обеих групп сведены в таблицу 4.8. Результаты всех тестов позволяют сделать однозначный

Листинг 4.24. Построение диграммы «ящик с усами» для рынка Алматы

```
# plot boxplots
nequiped <- subset(almatyFlats, furniture == 0)
equiped <- subset(almatyFlats, furniture > 0)
boxplot(nequiped$price.m, equiped$price.m,
ylab = 'price per meter',
names =c('not equiped', 'equiped'))
rm(nequiped)
rm(equiped)
```

Листинг 4.25. Проведение тестов на нормальность для наблюдений без отделимых улучшений

```
# normality tests for non equipped observations
# Shapiro-Wilk test for normality
shapiro.test(almatyFlats$price.m[ which(almatyFlats$furniture == 0)])
# Shapiro-Francia test for normality
sf.test(almatyFlats$price.m[ which(almatyFlats$furniture == 0)])
# Anderson-Darling test for normality
ad.test(almatyFlats$price.m[ which(almatyFlats$furniture == 0)])
# Adjusted Jarque-Bera test for normality
ajb.norm.test(almatyFlats$price.m[ which(almatyFlats$furniture == 0)])
# Lilliefors (Kolmogorov-Smirnov) test for normality
lillie.test(almatyFlats$price.m[ which(almatyFlats$furniture == 0)])
```

Листинг 4.26. Проведение тестов на нормальность для наблюдений с отделимыми улучшениями

```
# normality tests for non equipped observations
# Shapiro-Wilk test for normality
shapiro.test(almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
# Shapiro-Francia test for normality
sf.test(almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
# Anderson-Darling test for normality
ad.test(almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
# Adjusted Jarque-Bera test for normality
ajb.norm.test(almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
# Lilliefors (Kolmogorov-Smirnov) test for normality
lillie.test(almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
```

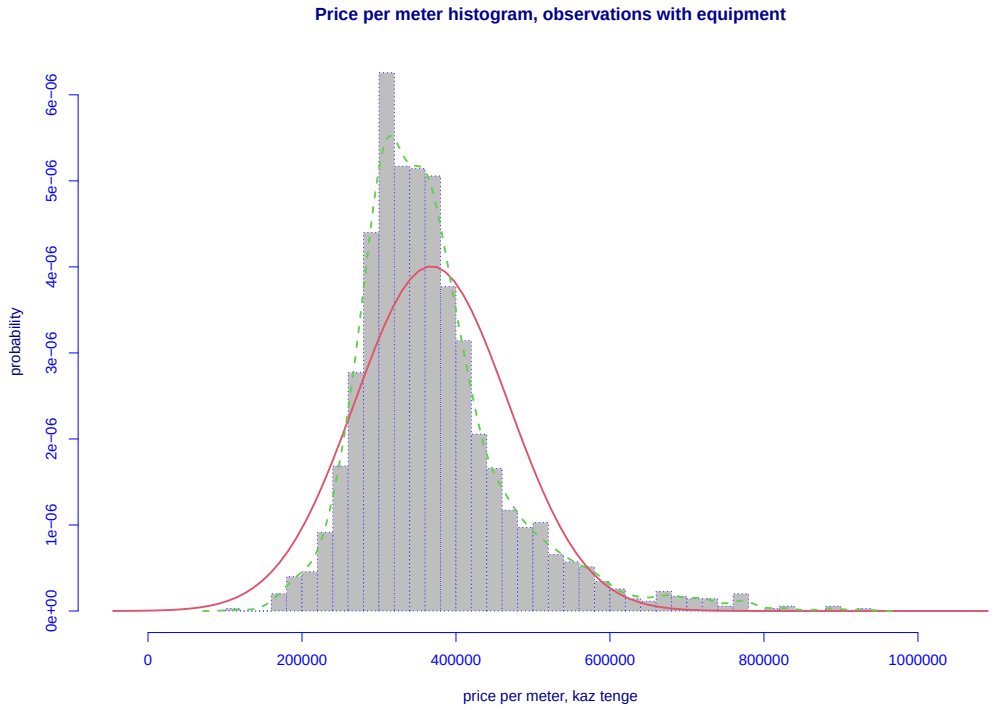

Рис. 4.10. Гистограмма цен предложения для объектов, предлагаемых к продаже предлагаемых к продаже совместно с отделимыми улучшениями, совмещённая с кривой функции плотности эмпирического распределения, а также кривой функции плотности теоретического нормального распределения.

вывод: распределения обеих подвыборок отличаются от нормального, что указывает на необходимость применения непараметрических критериев.

Для проведения U-теста достаточно выполнить код, содержащийся в скрипте 4.27. Результаты теста приведены в таблице 4.9. На основании полученного результата можно сделать вывод о том, что учёт фактора наличия неотделимых улучшений и движимого имущества должен быть произведён.

Листинг 4.27. Проведение U-теста для данных города Алматы

```
# perform Mann-Whitney U-test
wilcox.test(almatyFlats$price.m[ which(almatyFlats$furniture == 0)],
almatyFlats$price.m[ which(almatyFlats$furniture > 0)])
```

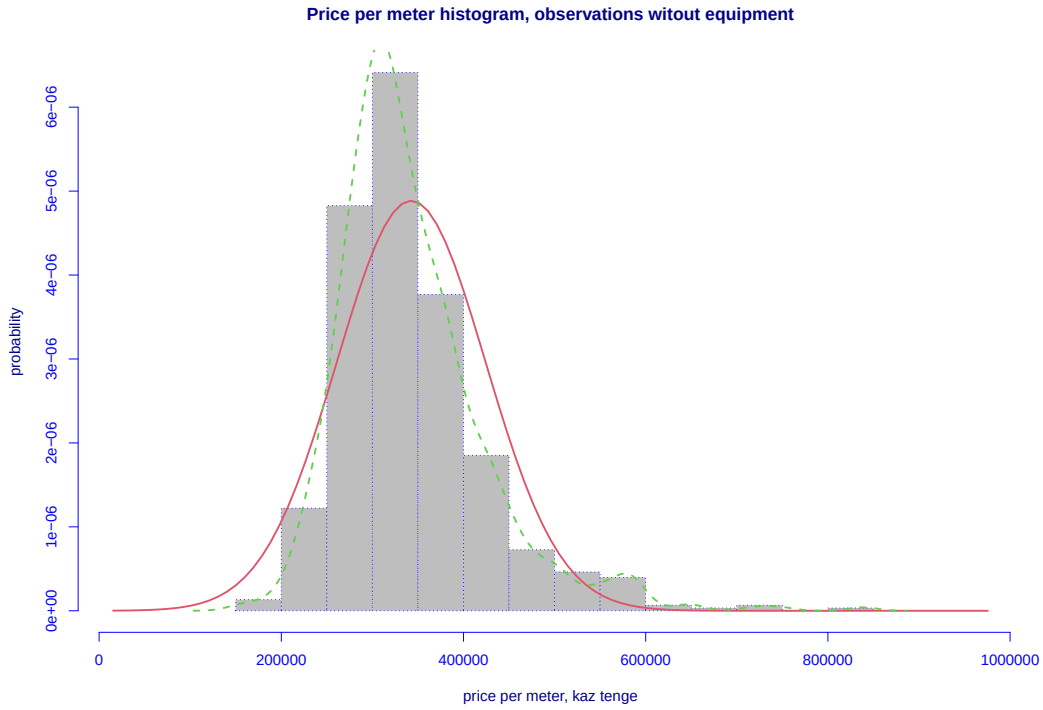

Рис. 4.11. Гистограмма цен предложения для объектов, предлагаемых к продаже без отделимых улучшений, совмещённая с кривой функции плотности эмпирического распределения, а также кривой функции плотности теоретического нормального распределения.

Таблица 4.5. Нулевая и альтернативная гипотезы при анализе данных Алматы

Тип гипотезы	Нулевая гипотеза (H0)	Альтернативная гипотеза (H1)
Научная	Распределения удельных показателей стоимости квартир, предлагаемых к продаже вместе с отделимыми улучшениями и движимым имуществом, и квартир, продаваемых без них, одинаковы, сдвиг между ними отсутствует, статистические оценки, сделанные для множества объектов-аналогов, продаваемых вместо с отделимыми улучшениями и движимым имуществом, являются несмещёнными для объектов, продаваемых без них (справедливо и обратное утверждение).	Распределение удельных показателей стоимости квартир, предлагаемых к продаже вместе с отделимыми улучшениями и движимым имуществом, и квартир, продаваемых без них, различается, существует сдвиг, оценка, сделанная для объектов, предлагаемых к продаже вместе с отделимыми улучшениями и движимым имуществом, будет смещённой для объектов, предлагаемых к продаже без них (справедливо и обратное утверждение).
Практическая	Медиана удельного показателя стоимости квартир, предлагаемых к продаже вместе с отделимыми улучшениями и движимым имуществом, равна медиане удельного показателя стоимости квартир, предлагаемых к продаже без них	Медиана удельного показателя стоимости квартир, предлагаемых к продаже вместе с отделимыми улучшениями и движимым имуществом, не равна медиане удельного показателя стоимости квартир, предлагаемых к продаже без них.
Изложенная в терминах оценки	Наличие либо отсутствие отделимых улучшений и движимого имущества в составе продаваемой квартиры не является существенным различием и не требует какой-либо специальный учёт, т. е. не является ценообразующим фактором.	Наличие либо отсутствие отделимых улучшений и движимого имущества в составе продаваемой квартиры является существенным различием и требует специальный учёт, т. е. является ценообразующим фактором.

Таблица 4.6. Сведения о количестве наблюдений различных типов на рынке города Алматы

Тип наблюдений	Количество	Рациональное число интервалов (k)
Все наблюдения	2355	36
Наблюдения без отделимых улучшений	605	22
Наблюдения с отделимыми улучшениями	1750	31

Таблица 4.7. Базовые описательные статистики наблюдений различных типов на рынке города Алматы (единица — казахстанский тенге)

Тип наблюдений	Min	1Q	Медиана	Среднее	3Q	Max
Все наблюдения	117000	300000	344432	361554	400000	928571
Наблюдения без отделимых улучшений	152542	291803	325581	342581	378788	838462
Наблюдения с отделимыми улучшениями	117000	305446	350331	368113	406183	928571

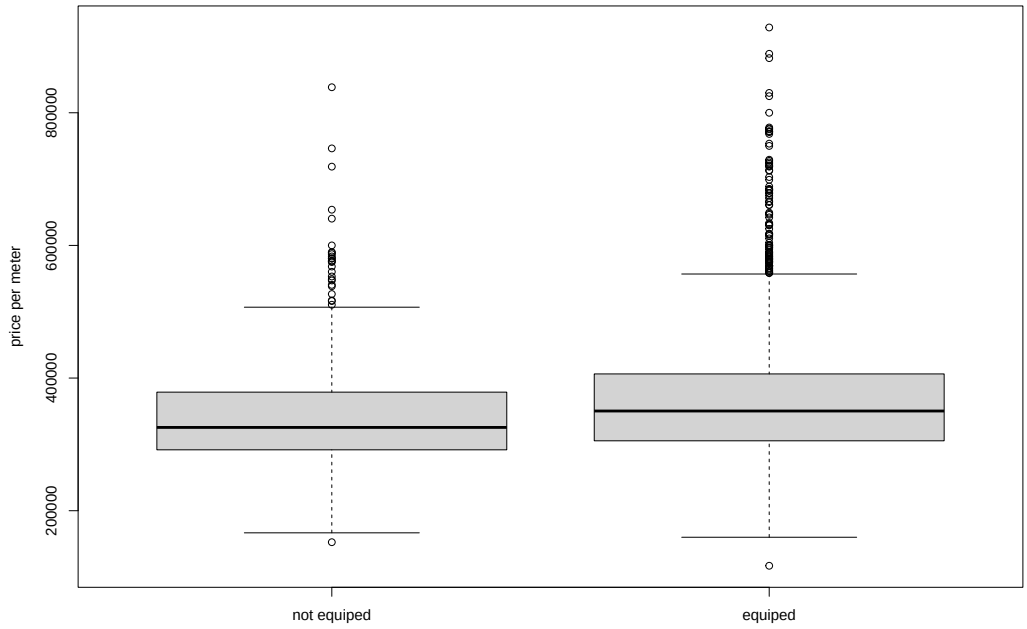

Рис. 4.12. Диаграмма «ящик с усами для рынка Алматы».

Таблица 4.8. Результаты проведения тестов проверки данных по г. Алматы на нормальность ($\alpha = 0.05$)

Тест	Без отделимых улучшений	С отделимыми улучшениями
Шапиро—Уилка:	4.25	4.26
статистика критерия (W)	0.9	0.9
p-значение	<2e-16	<2e-16
H0	отклоняется	отклоняется
Шапиро—Франчиа:	4.25	4.26
статистика критерия (W)	0.9	0.9
p-значение	<2e-16	<2e-16
H0	отклоняется	отклоняется
Андерсона—Дарлинга	4.25	4.26
статистика критерия (A)	13	42
p-значение	<2e-16	<2e-16
H0	отклоняется	отклоняется
Жарка—Бера (скорр.)	4.25	4.26
статистика критерия (AJB)	757	1720
p-значение	<2e-16	<2e-16
H0	отклоняется	отклоняется
Лиллиефорса (K-C)	4.25	4.26
статистика критерия (D)	0.1	0.1
p-значение	<2e-16	<2e-16
H0	отклоняется	отклоняется
Итоговый вывод:		
H0	отклоняется	отклоняется

Таблица 4.9. Результаты проведения U-теста для данных Алматы($\alpha = 0.05$)

Показатель	Значение
Статистика критерия	441360
p-значение	1e-09
Нулевая гипотеза (см. таблицу 4.5)	отклоняется

Глава 5.

Выводы

В данной работе были рассмотрены теоретические и практические аспекты применения критерия Манна—Уитни—Уилкоксона в повседневной практике оценщика. Данный тест может стать надёжным критерием проверки значимости тех или иных признаков объектов-аналогов, наблюдаемых на открытом рынке. Самостоятельный анализ рыночных данных является лучшим и, вероятно, единственным подлинно доказательным инструментом оценщика, стремящегося создавать ценность для заказчика, а также новые знания о рынках. Несмотря на некоторый начальный порог входа, анализ данных средствами языков программирования является достаточно простым. Строго говоря, часть сделанных шагов является избыточной и была сделана с целью продемонстрировать малую часть возможностей языков Python и R, а также показать важность визуализации данных. Минимальный набор действий, необходимых для проведения U-теста включает следующие этапы:

- загрузка данных и создание датафрейма;
- создание двух подвыборок на основе значения изучаемого признака;
- проведение тестов на нормальность распределений (в большинстве случаев достаточно проведения теста Шапиро—Франчиа либо K2 Д'Агостино);
- проведение U-теста.

С учётом наличия готовых скриптов проведение теста занимает всего несколько минут.

Конечно же наука не стоит на месте.

В завершение материала хочется дать следующую рекомендацию: современному оценщику в первую очередь необходимо изучать Python: скорее всего, его инструменты исчерпывающе закроют все потребности по анализу данных. В случае стремления к совершенству и наличия желания быть на уровень выше, знание R может помочь в решении особенно сложных и нетривиальных задач. Впрочем, существуют и альтернативные мнения [37]. Эпоха массового применения методов машинного обучения и математической статистики только начинается, и сейчас сложно предсказать, какие средства станут стандартом оценщика через 5 или 10 лет.

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] Hubert W Lilliefors. «On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown». В: *Journal of the American Statistical Association* 62.318 (1967-06-01), с. 399—402. ISSN: 0162-1459. DOI: 10.1080/01621459.1967.10482916. URL: https://en.wikipedia.org/wiki/Lilliefors_test (дата обр. 04.06.2022).

[5] S. S. Shapiro и R. S. Francia. «An Approximate Analysis of Variance Test for Normality». В: *Journal of the American Statistical Association* 67.337 (1972), с. 215—216. DOI: 10.1080/01621459.1972.10481232. URL: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Francia_test (дата обр. 04.06.2022).

[6] William J. Conover. *Practical Nonparametric Statistics*. John Wiley и Sons, 1980.

[7] Carlos M. Jarque и Anil K. Bera. «Efficient tests for normality, homoscedasticity and serial independence of regression residuals». В: *Economics Letters* 6.3 (1980), с. 255—259. URL: https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test (дата обр. 04.06.2022).

[8] Ralph B. Agostino, Albert Belanger и Jr. Ralph B. Agostino. «A suggestion for using powerful and informative tests of normality». English. В: *The American Statistician* 44.4 (1990), с. 316—321. DOI: 10.2307/2684359. URL: https://web.archive.org/web/20120325140006/http://www.cee.mtu.edu/~vgriffis/CE%205620%20materials/CE5620%20Reading/DAgostino%20et%20al%20-%20normaility%20tests.pdf (дата обр. 29.05.2022).

[9] П. В. Новицкий и И. А. Зограф. *Оценка погрешностей результатов измерений*. 2-е изд. Ленинград: Энергоатомиздат, 1991, с. 304. ISBN: 5283045137. URL: http://kepstr.eltech.ru/tor/mri/Literatura/Novitzkij%201991.pdf (дата обр. 15.10.2021).

[10] Erich L Lehamnn. «Elements of Large Sample Theory». В: *Springer* (1999), с. 176.

[11] Eiiti Kasuya. «Mann–Whitney U test when variances are unequal». В: *Animal Behaviour* 6.61 (2001), с. 1247—1249. DOI: 10.1006/anbe.2001.1691. URL: https://www.sciencedirect.com/science/article/abs/pii/S0003347201916914 (дата обр. 06.06.2022).

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

[13] 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).

[14] 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.

[15] Julian D. Karch. «Psychologists Should use Brunner-Munzel's instead of Mann-Whitney's U-test as the Default Nonparametric Procedure». В: *Advances in Methods and Practices in Psychological Science* 2.4 (2021). ISSN: 2515-2459. DOI: 10.1177/2515245921999602. URL: https://journals.sagepub.com/doi/10.1177/2515245921999602 (дата обр. 06.06.2022).

[16] 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).

[17] 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).

[18] К. А. Мурашев. *spba-flats-210928*. 2021-09-28. URL: https://github.com/Kirill-Murashev/datasets/blob/main/Saint-Petersburg/flats/spba_flats_210928.csv.

[19] 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/.

[20] 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).

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

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

[23] *AUC-Derivation.ipynb*. URL: https://colab.research.google.com/github/unpingco/Python-for-Signal-Processing/blob/master/AUC_Derivation.ipynb#scrollTo=BgrH5C49LqMx (дата обр. 14.06.2022).

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

[25] *F-test*. URL: https://en.wikipedia.org/wiki/F-test (дата обр. 06.06.2022).

[26] *Fligner-Policello test in R*. URL: https://search.r-project.org/CRAN/refmans/RVAideMemoire/html/fp.test.html (дата обр. 06.06.2022).

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

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

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

[30] *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.

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

[32] 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).

[33] 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).

[34] 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).

[35] 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).

[36] *Ordered logit*. URL: https://en.wikipedia.org/wiki/Ordered_logit (дата обр. 06.06.2022).

[37] I. Shutov. *Who decided for everyone that Python is convenient for commercial analytics?* Russian. URL: https://habr.com/ru/post/670250/ (дата обр. 08.06.2022).

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

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

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

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

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

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

[44] *Welch-t-test*. URL: https://en.wikipedia.org/wiki/Welch's_t-test (дата обр. 06.06.2022).

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

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

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

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

[49] 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).

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

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

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

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

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

[55] Wikipedia. *Type I and type II errors*. URL: https://en.wikipedia.org/wiki/Type_I_and_type_II_errors (дата обр. 08.06.2022).

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