Очень краткое введение в математическую статистику, основанную на частотном подходе, для оценщиков

К. А. Мурашев

22 октября 2021 г.

Какую бы работу не выполнял оценщик, во всех случаях он имеет дело с информацией и данными. Часто эти данные представляют собой числа либо могут быть формализованы иным образом. В любом случае требуется алгоритмическая обработка входных данных и преобразование их в информацию, а в некоторых случаях — в знания. Целью данного фрагмента является формирование общих представлений об основных понятиях и методах математической статистики, необходимых современному оценщику. Автор постарался прибегать к минимальному числу формул и сложных определений, хотя это и не вполне получилось. Поскольку конечной целью всей работы является цифровизация оценочной деятельности, в тексте приводятся короткие листинги на языках R и Python, позволяющие реализовать то, о чём говорится в тексте. В данном фрагменте рассматривается частотный подход к вероятности [52]. Вопросы байесовской вероятности [51] рассмотрены в отдельном материале.

Содержание

1. Что есть математическая статистика?

У термина «статистика» существует несколько определений. Статистикой называют:

- данные, количественно описывающие тот или иной аспект окружающего мира: например данные об уровне безработицы, заболеваемости коронавирусом или доходах граждан, т. е. такие данные, которые описывают явление целиком;
- количественные данные, относящиеся к какому-либо одному субъекту либо результатам его деятельности: например количество выполненных оценщиком отчётов об оценке за календарный год;
- результаты исследования отдельных выборок: например итоги социологических опросов или результаты анализа рынка недвижимого имущества;
- конкретные методы анализа данных с помощью математичсеских методов;

- т. н. статистики критерия, т. е. конкретные числовые значения отдельных вычислений, например статистика критерия Шапиро—Уилка;
- область знаний, которая разрабатывает и использует математические методы для описания данных и формирования суждений о них.

Первый тип статистики как правило не имеет прямого отношения к деятельности оценщика. Подобные сведения чаще всего могут быть получены из открытых источников. Кроме того, даже в случае недоверия к ним, у оценщика всё равно отсутствуют инструменты для получения подобных данных самостоятельно. **Второй** тип — скорее всего также не является особенным предметом интереса оценщика. Определение стоимости объекта осуществляется методом аналогии путём сравнения с наблюдениями (предложениями либо сделками), тогда как погружение в свойства только самого объекта не позволяют определить его стоимость. **Третий** тип статистики отсылает нас к фундаментальному принципу: исследовать генеральную совокупность путём изучения выборки из неё. Собственно это и является предметом данной работы, а также профессиональной деятельности оценщика: формирование предсказания свойства объекта (его стоимости) на основе изучения выборки. **Четвёртый** — является ключевым с узко практической точки зрения. Процесс предсказывания неизвестных свойств объектов на основе известных с учётом знаний, полученных при изучении выборки, по мнению многих, и есть статистика. Данный подход не является ошибочным, однако его вряд ли можно считать полным. Умение применять конкретные методы является необходимым, но недостаточным условием успешной работы. В настоящее время практически все вычисления выполняются программными средствами.[1] В связи с этим важность навыков ручного применения тех или иных методов сведена к минимуму. Вместо этого на первый план выходят навыки планирования оценочного статистического эксперимента, постановки задачи, поиск источников данных, их сбор и предобработка, общее понимание применяемых методов, выбор между ними, а также интерпретация полученного результата. **Пятый** тип означает результаты применения конкретных методов. В настоящее время чаще используются не сами статистики критериев, а универсальный показатель — p-значение (p-value). И, наконец, **шестое** определение означает обширную область человеческих знаний, в рамках которой существуют конкретные методы и результаты их применения. Данное определение может быть заменено термином *«математическая статистика»*, подчёркивающим отличие от других значений общего термина «статистика». Именно в этом значении мы и будем использовать данный термин на протяжении всей работы по ознакомлению со статистическими основами оценки стоимости. Таким образом, если в тексте прямо не указано иное слово «статистика» следует понимать как «математическая статистика».

Как уже было сказано выше, в данном материале рассматривается только частотный подход к понятию вероятности. Это та самая статистика, которую изучают

[1] Автор в своей работе использует языки программирования Python и R и рекомендует поступать также, однако существуют и иные средства: Julia, SPSS, PSPP, Stata и многие другие вплоть до табличных процессоров.

в вузах на нематематических специальностях. В последние 20–25 лет всё бóльшую популярность набирает байесовский подход к вероятности. О различиях между этими подходами можно прочесть, например в [37].

2. Генеральная совокупность и выборка

Генеральной совокупностью называется всё множество объектов, в отношении которого необходимо сделать те или иные выводы. В случае оценки, например торгово-развлекательного центра, расположенного на проспекте Просвещения в Санкт-Петербурге, генеральной совокупностью будут являться все торговые центры, расположенные на территории Санкт-Петербургской городской агломерации, независимо от того, выставлены они в данный момент на продажу или нет. Для того, чтобы понять, что является генеральной совокупностью, необходимо ответить на один простой вопрос: «на какое множество объектов можно обобщить полученные результаты исследования?». Очевидно, что независимо от того, выставлен ли какой-либо из существующих в агломерации ТРЦ на продажу или нет, совершались ли с ним в последнее время сделки или нет, все выводы, сделанные относительно множества «торгово-развлекательные центры, расположенные в СПБГА»,[2] могут быть распространены на него равно как и на вообще любой объект данного типа, находящийся в указанных границах. При этом можно провести разграничение между понятиями «генеральная совокупность объектов на рынке» и «генеральная совокупность аналогов». Первое множество соответствует данному ранее определению, необходимость обозначения второго множества возникла в рамках дискуссий со специалистами оценщиками (А. А. Слуцкий, А. Я. Пичукан, Н. П. Баринов).[3] Можно сказать, что предметом интереса исследования оценщика почти всегда является первое множество. Действительно, объект оценки не входит в выборку аналогов, он может не быть выставлен на продажу (аренду). Таким образом, для того, чтобы говорить о возможности применения к нему результатов анализа выборки аналогов, необходимо реализовать три этапа:

1) статистический анализ выборки, формирование выводов о свойствах всего открытого рынка;

2) обобщение результатов, полученных на предыдущем этапе, на всю совокупность объектов на рынке — *принцип «от частного к общему»*;

3) применение выявленных закономерностей рынка к конкретному объекту оценки, формирование вывода о его стоимости на основе установленных свойств открытого рынка, к которому он относится — *принцип «от общего к частному»*.

[2] Санкт-Петербургская городская агломерация. Включает в себя Санкт-Петербург, бóльшую часть Всеволожского, части Выборгского, Кировского, Тосненского, Гатчинского и Ломоносовского районов Ленинградской области.

[3] Здесь и далее порядок перечисления авторов строго алфавитный: сортировка «**И**мя, **О**тчество, **Ф**амилия.

Таким образом, можно говорить о том, что связь между выборкой аналогов и объектом оценки осуществляется исключительно вследствие их принадлежности к одному множеству — генеральной совокупности всех объектов на рынке. При этом, с узко практической точки зрения оценщики могут говорить о том, что они осуществляют анализ выборки, взятой из генеральной совокупности «все объекты, представленные к продаже (аренде)». Данный вопрос не является принципиально важным, однако верное понимание логических схем никогда не бывает лишним. Прежде всего следует помнить о том, что связь между выборкой аналогов и объектом оценки существует только в первом случае.

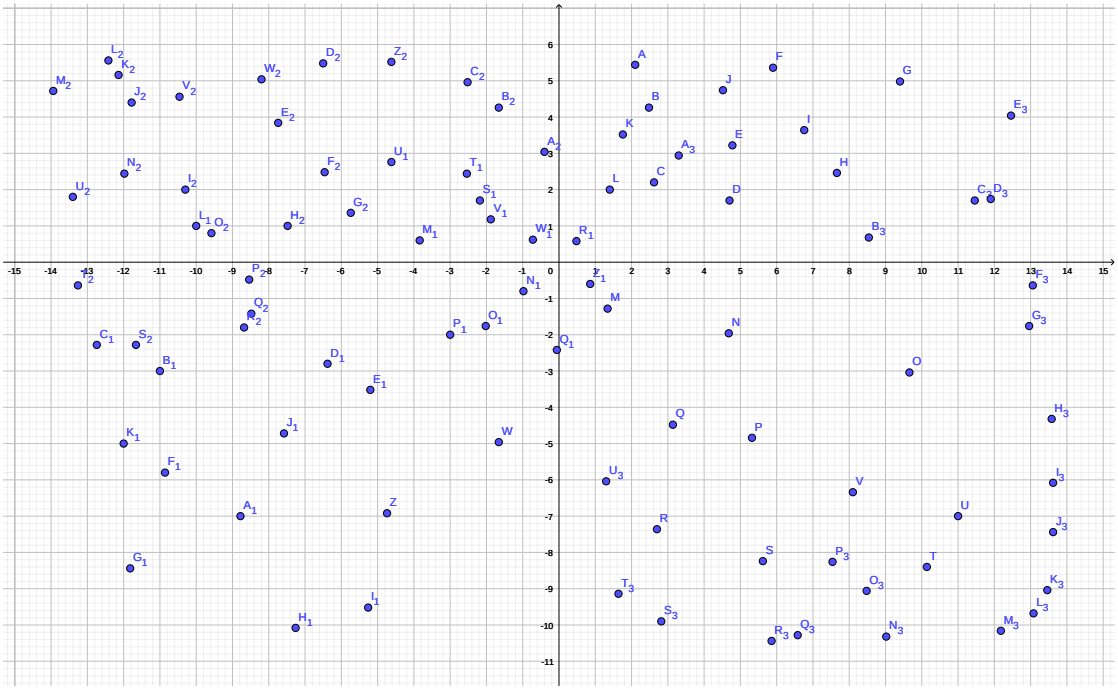

Рис. 1. Генеральная совокупность

Конечно же лучшим способом исследования генеральной совокупности является изучение всех входящих в неё объектов. Иногда это не представляет никакой сложности. Генеральная совокупность «международные аэропорты, расположенные на территории СПБГА» состоит из единственного объекта. Однако чаще всего, генеральная совокупность, во-первых достаточно велика, во-вторых, сведения о существенной, а чаще бóльшей её части недоступны для исследования. В таких случаях целесообразным и единственным способом получения знаний о её свойствах является формирование выборки из данной совокупности и её дальнейшее изучение. Важным условием применимости выводов, полученных на основе анализа выборки, ко всей генеральной совокупности является её репрезентативность, т. е. свойство отражать закономерности, присущие всему множеству.

2.1. Формирование репрезентативной выборки

Предположим, что на рисунке 1 изображена генеральная совокупность объектов, относящихся к какому-либо сегменту рынка, всего 72 объекта. При этом по оси x отложена площадь объектов, по оси y — их стоимость. Простым и корректным способом формирования выборки является т. н. *простая случайная выборка (simple random sample)*. При обеспечении достаточной меры случайности отбора наблюдений и значительного числа, выборка будет схожа с генеральной совокупностью. На рисунке 2 показан пример такой случайной выборки. На первый взгляд выборка выглядит достаточно репрезентативной. Подвидом данного метода формирования выборки является *механическая выборка*. Её отличие заключается в том, что первый элемент отбирается случайным образом, а все последующие — на основе какого-либо правила. Например, случайным образом из списка предложений выбирается одно из них, после чего выбираются два наблюдения, имеющие порядковые номера, отличающиеся по модулю от номера первого наблюдения на какое-либо число, после чего цикл продолжается до получения нужного количества наблюдений. Другим вариантом является формирование *стратифицированной выборки (stratified sample)*. Предположим, мы хотим взять по 4 объекта из каждого сектора, ограниченного осями. Если ввести предположение о том, что оси характеризуют среднее значение показателя, это будет означать, что мы хотим взять четыре объекта, имеющих площадь и стоимость выше средних, четыре — цену выше и площадь ниже средней и т. д. Пример такой выборки показан на рисунке 3. Как видно, данная выборка также выглядит репрезентативной. Следующим вариантом является формирование *групповой выборки (cluster sample)*. В этом случае на первом этапе осуществляется предобработка данных, в ходе которой формируются группы (кластеры), включающие в себя наиболее близкие по тем или иным признакам объекты. Далее из каждой группы отбирается определённое число наблюдений, например два. Пример групповой выборки изображён на рисунке 4.

На первый взгляд может показаться, что стратифицированная и групповая выборка представляют собой один и тот же тип выборки. Однако, это не так. В общем, можно сказать, что примером стратифицированной выборки является отбор по критериям цены и площади, как было описано выше. Примером групповой выборки является отбор наблюдений по-критерию расположения объектов в различных районах. В таблице 1 приведены сводные данные о различиях между этими двумя способами формирования выборки.

Все рассмотренные выше способы формирования выборки относятся к вероятностным. Существуют также детерминированные способы, с которыми можно ознакомиться например в [46].

3. Измерения

Для осуществления статистического анализа рынка необходимо преобразовать массив входящей информации в данные. На практике это означает необходимость присвоения значений ключевым объектам, понятиям и характеристикам. При этом

Таблица 1. Различия между групповой и стратифицированной выборками [50]

	Групповая выборка	Стратифицированная выборка
Охват	Выборка формируется только из некоторых групп (кластеров)	Выборка формируется из всего множества
Требования к однородности и различиям	В пределах группы (кластера) элементы должны отличаться (быть разнородными), при этом существует требование к однородности или схожести между различными кластерами	В пределах страты элементы должны быть однородными, тогда как между стратами должны быть различия
Схема выборки	Нужна только для групп (кластеров), попавших в выборку	Формируется полностью для всех стратифицированных подмножеств
Назначение	Повышает эффективность выборки, уменьшая стоимость исследования	Повышает точность

такое присвоение не может носить случайный либо субъективный характер. Необходима система.

Чаще всего оценщики работают с числовыми данными, однако это не единственный тип данных, подлежащих анализу. Например разделения объектов, пригодных и предназначенных для постоянного проживания в них людей и домашних животных, на квартиры и апартаменты само по себе не несёт никакой числовой информации.

Измерение — это процесс систематического присвоения количественных значений характеристикам объектов и их свойствам для облегчения использования математического аппарата при изучении и описании объектов и их взаимосвязей [27].

Некоторые типы измерений носят вполне определённый характер: площадь земельного участка в квадратных метрах, возраст здания в годах, цена предложения в рублях, пробег автомобиля в километрах. Работа с такими данными очень удобна: они однозначны, могут переводиться из одних единиц в другую, возможно установление связи между различными переменными такого типа, например простое деление стоимости на площадь даёт показатель стоимости за единицу площади. Однако не все данные изначально имеют количественную оценку. О типах данных мы и поговорим далее.

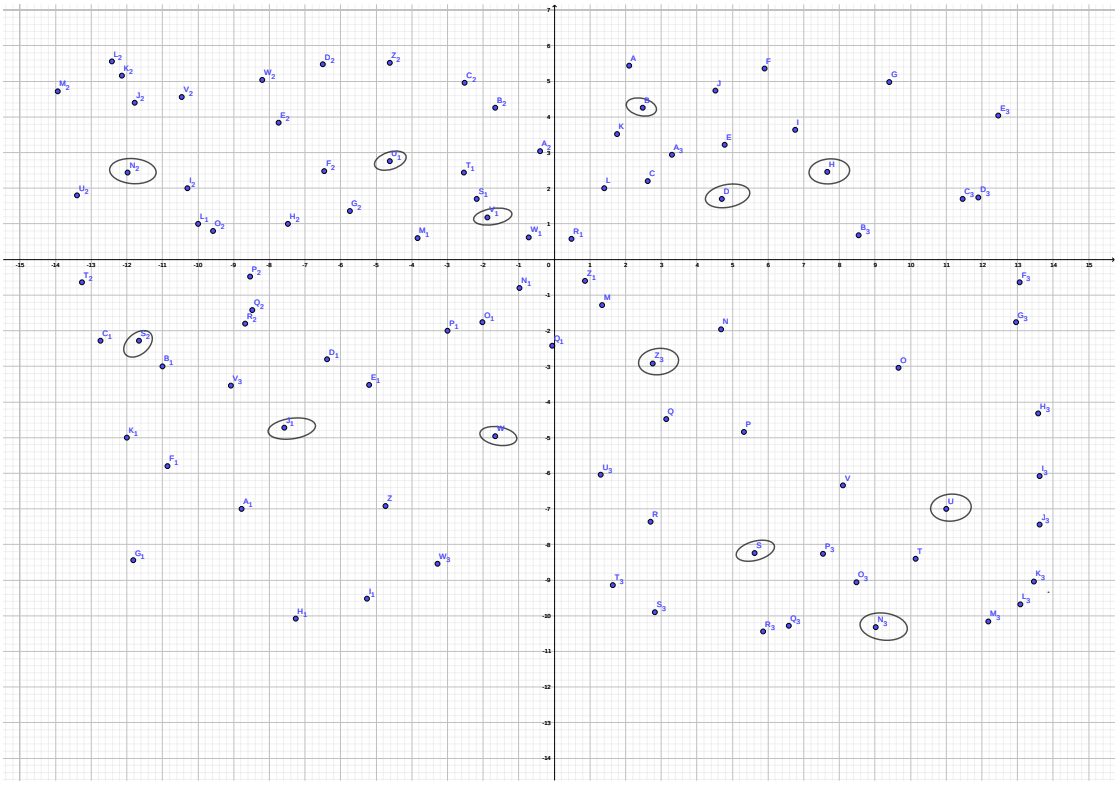

Рис. 2. Простая случайная выборка

3.1. Типы данных

Существует несколько типов данных. В первую очередь следует сказать о том, что данные бывают **количественными**, **порядковыми** и **номинативными**. Третий тип данных также называют номинальными, однако последний термин в практике оценки прочно закрепился за вариантом денежного потока, вследствие этого во избежание двусмысленности оценщикам лучше придерживаться первого варианта обозначения этого типа данных.

3.1.1. Номинативные данные

Номинативные данные представляют собой метку, имя и т. п. качественные характеристики, не имеющие смысла в виде числа. К таким данным относятся, например адрес объекта недвижимости (как целиком, так и, например в виде метки, обозначающей муниципальное образование), его тип, наименование завода-изготовителя промышленной установки, применяемая хозяйственным обществом система налогообложения. Мы можем закодировать такие данные в виде чисел, например присвоив значение «0» квартирам, и «1» апартаментам, либо «0» — предприятиям, применяющим ОСНО, «1» — УСН «доходы», «2» — УСН «доходы, уменьшенные на величину расходов» и т. д., однако такое присвоение является не более чем способом кодирова-

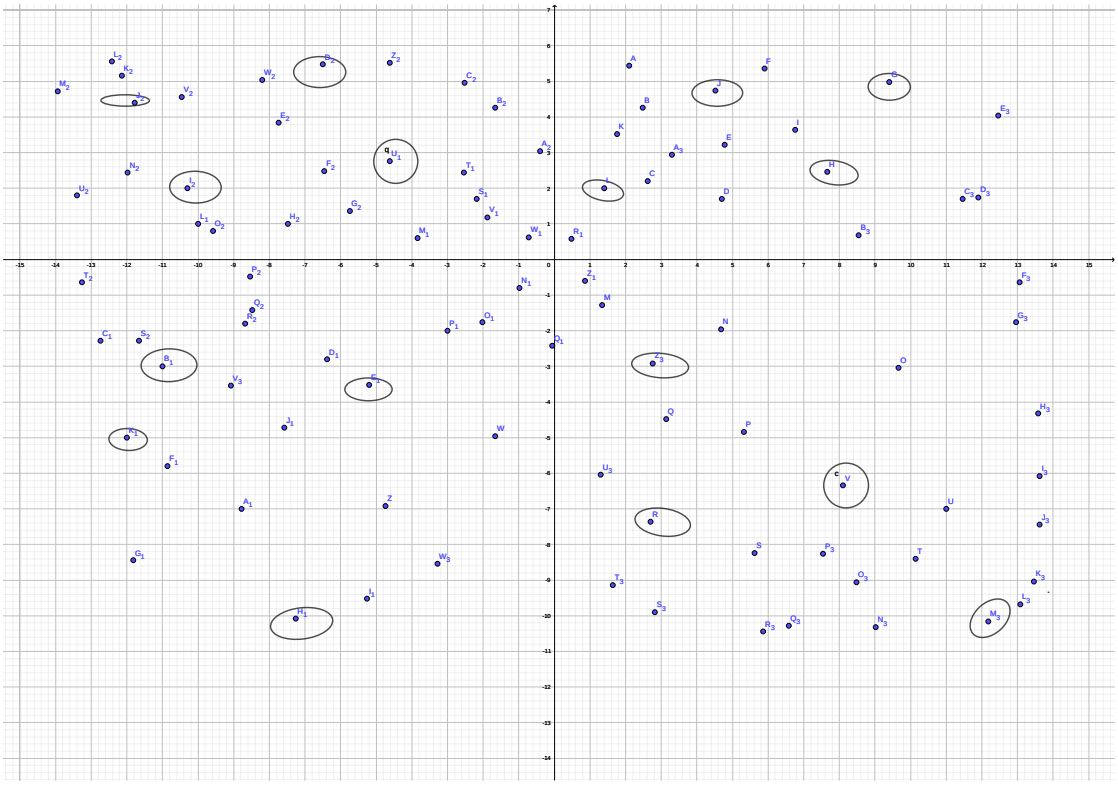

Рис. 3. Стратифицированная выборка

ния данных без какого-либо количественного смысла. Кроме того можно переставить значения кодов и начать присваивать «0» апартаментам, а «1» квартирам без потери либо искажения смысла кодирования. В случае сомнения насчёт того, являются ли те или иные данные номинативными, исследователю необходимо задать себе вопрос «отражают ли числа некоторое свойство так, что более высокое значение означает наличие большего количества этого свойства?». В случае отрицательного ответа можно с уверенностью говорить о том, что имеют место номинативные данные. Следующим критерием отнесения данных к номинативному типу является полная бессмысленность осуществление каких-либо арифметических действий с числовыми метками. Несложно догадаться, что сложение и вычитание, не говоря уже об умножении или делении, например значений кодов ОКТМО являются по меньшей мере бессмысленными. Другое название данного типа — **категориальные данные**, что говорит о том, что наличие того или иного значения отражает принадлежность к определённой категории, а не количественное измерение какого-либо свойства. Вопреки некоторым представления, такие данные вполне поддаются анализу методами частотной статистики. Особым случаем является ситуация, когда возможны лишь два значения. Такие данные называются **бинарными**. Данный случай настолько распространён и важен, что для таких данных существуют особые методы анализа, например *логистическая регрессия*, а также *отношение шансов* и *отношение рисков*.

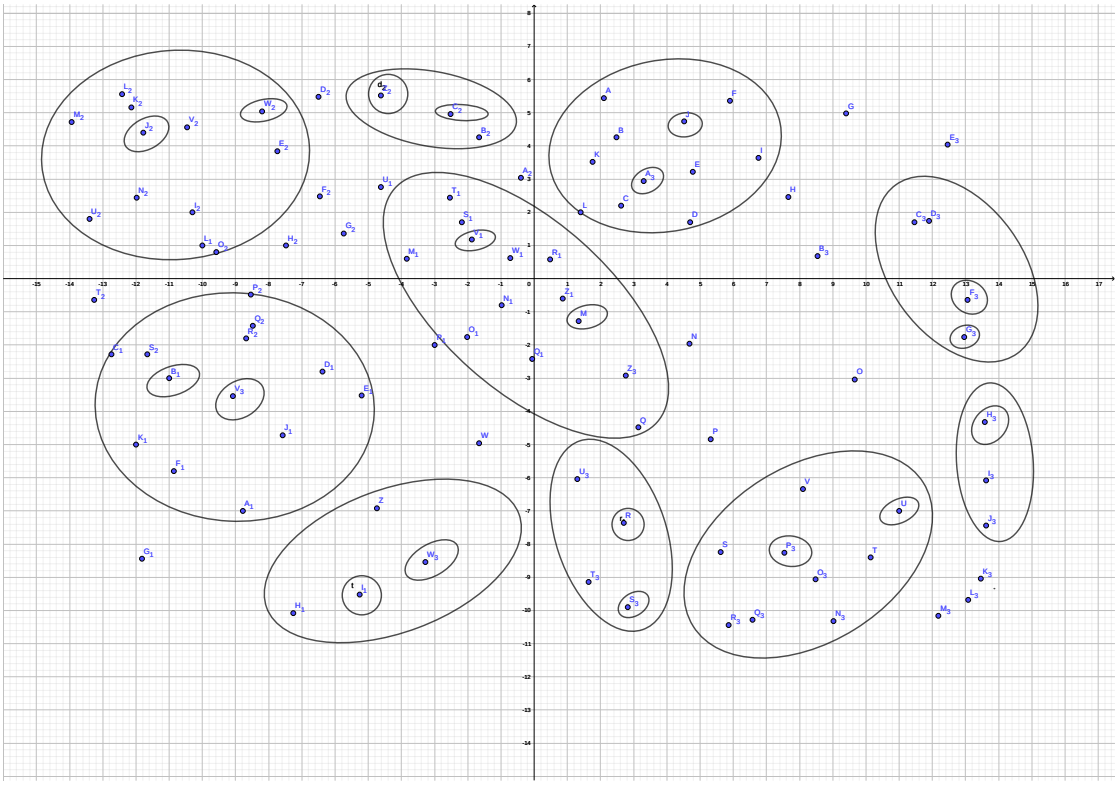

Рис. 4. Групповая выборка

3.1.2. Порядковые данные

Ранговые (порядковые) данные — представляют собой данные, значения которых можно расположить в каком-либо осмысленном порядке, таким образом что бóльшие значения соответствуют большему проявлению какого-либо признака относительно меньших. Примером таких данных можно считать звёздность гостиницы. Большее число звёзд означает более высокий класс качества обслуживания гостей и является косвенным признаком большей типичной стоимости номеров или единицы площади самой гостиницы. При этом категории гостиниц можно расположить в логичной последовательности. Также оценщики часто используют порядковую шкалу для кодирования качества и состояние отделки объектов недвижимости.При этом не существует какой-либо измерительной шкалы, позволяющей определить расстояние между рангами либо ответить на вопрос, «является ли оно одинаковым между всеми соседними рангами или нет?». Сами числа в порядковых данных в отличие от номинативных имеют определённый смысл, существуют методы извлечения полезной информации из таких данных. При этом, всегда следует проявлять большую осторожность при работе с таким типом данных. Если в случае с номинативными данными вряд ли кому-то придёт в голову вычислять среднее значение кода ОКТМО, то нельзя исключать вариант того, что кто-то сочтёт возможным посчитать

среднее значение для порядковых данных, чего делать ни в коем случае нельзя. Вычисление среднего предполагает операцию деления, применимую только к одному типу данных — данным, характеризующим отношения, о которых будет сказано далее в 3.1.3.1. При этом вычисление медианы для порядковых данных является допустимым.

3.1.3. Количественные данные

Количественные данные в отличие от номинативных либо порядковых могут быть получены непосредственно в результате измерения.

3.1.3.1. Интервальные данные и данные, характеризующие отношения **Интервальные данные** характеризуются осмысленным порядком и равными интервалами между измерениями, отражающими равновеликие изменения количества любой измеренной величины. Примером таких данных является температура, измеренная по шкале Цельсия. Различие температуры между 20 и 40 °C такое же, какое оно между 40 и 60 °C. Операции сложения и вычитания являются осмысленными для этого типа данных, поскольку разница на единицу характеризует одинаковое изменение признака на всём протяжении шкалы. Однако, поскольку у шкалы Цельсия (равно как и у шкалы Фаренгейта) нет естественного нуля. Значение температуры 0 °C не означает полное отсутствие тепловой энергии, а является лишь удобной с точки зрения повседневной деятельности точкой отсчёта. В связи с этим некорректно говорить о том, что 40 °C означает состояние «в два раза теплее» относительно 20 °C. Оценщики редко могут встретиться с таким типом данных. Единственным примером можно считать данные, описывающие параметры оборудования или материала с точки зрения их жаропрочности. Для полноценного оперирования такими значениями целесообразно перевести показатели в градусы по шкале Кельвина, поскольку данная шкала оперирует данными, характеризующими отношения, о которых пойдёт речь ниже.

Данные, характеризующие отношения обладают всеми полезными свойствами интервальных данных (осмысленный порядок, равные интервалы), но имеют при этом естественным ноль. Площадь, масса, хронологический возраст и наконец сама стоимость — это данные характеризующие отношения. Любые арифметические действия с такими данными являются осмысленными.

С учётом относительной редкости интервальных данных в данной работе в дальнейшем всегда будут подразумеваться данные, характеризующие отношения, если прямо не указано обратное.

3.1.3.2. Дискретные и непрерывные данные **Дискретные данные** могут принимать только определённые значения, при этом между этими значениями существуют чёткие границы и равные интервалы. Количество комнат и санузлов в квартире, этажей в торговом центре, входов в помещение, число двигателей летательного аппарата или лотков подачи сырья в установку — всё это дискретные данные.

Непрерывные данные могут принимать любые значения в принципе либо внутри определённого диапазона. Расстояние, мощность, выручка — непрерывные данные. В строгом общенаучном смысле можно сказать, что почти любые непрерывные данные на самом деле дискретные. Действительно, выручка может быть измерена с точностью не выше чем до копеек, расстояние не более чем с точностью до $\approx 3 \times 10^{-15}$ м. То есть в любом случае можно говорить о существовании некоего «кванта», дробление ниже которого является невозможным. Однако на практике чаще всего можно и нужно пренебречь такой строгостью и говорить о непрерывности данных. При этом возникает логичный вопрос, каким образом разграничить дискретные и непрерывные данные, если между неми нет однозначной разницы, и в определённом смысле практически любые данные дискретны по своей физической природе. Не сущесвтует однозначного и единственно верного способа провести разграничение. Следует придерживаться стандартов анализа конкретных величин. Существует рекомендация, согласно которой данные можно считать непрерывными, если для них возможными являются 16 и более значений [27].

3.1.4. Основные выводы

Оценщикам не следует забывать о различиях между типами данных и применять к ним только те арифметические действия и методы анализа, которые являются допустимыми. Следует отметить, что полный набор методов частотной статистики применим только в случае непрерывных данных. В остальных случаях следует проявлять осторожность и использовать специальные методы, предназначенные для других типов данных. В отдельных случаях, проявляя осторожность, понимая смысл выполняемых операций и осознавая ответственность за результат, допустимо отступление от стандартных правил. Например, можно рассчитать среднюю этажность домов в населённом пунктке, несмотря на то, что число этажей не является непрерывными данными. Распространённым примером некорректного применения статистических методов является ситуация, когда оценщики кодируют состояние объекта по шкале, например от 1 до 5, а затем применяют такую переменную в качестве предиктора в регрессионной модели, используя её значения как числа. Компьютер само собой не знает о том, что на самом деле ему в обработку была передана не количественная, а порядковая переменная и начинает честно считать среднее значение, сравнивать с ним значения каждого наблюдения и т. д. В итоге, такая модель, хотя и выглядит наукообразно, на самом деле несёт в себе искажение информации и вряд ли может использоваться в серьёзном анализе. В целом, данная проблема не является специфичной для оценки. Известная всем со школы система расчёта среднего балла также является антинаучной, равно как и система потолка оценок, при которой талантливые ученики не могут получить балл выше «5», даже если их уровень намного превышает тот, который по умолчанию соответствует высшей оценке. К сожалению, в мире по-прежнему много пережитков мрачного прошлого, а Знание не всегда побеждает архаику тёмных времён. На данном этапе развития оценочной деятельности особенно важно придерживаться научных принципов, не использовать сомнительные приёмы, лишь на первый взгляд облегчающие

Листинг 1. Преобразование переменной в фактор на языке R

```
df$col <- as.factor(df$col)
```

Листинг 2. Преобразование переменной в фактор на языке Python

```
df[col] = df[col].astype('category')
```

работу. На самом деле нет никакой сложности в том, чтобы корректно включить, например ранговую переменную, в регрессионную модель. Достаточно преобразовать переменную в фактор, после чего компьютер корректно учтёт её. Языки программирования R и Python позволяют осуществлять подобные преобразования путём написания кода длиной в одну четвёртую строки, см. листинги 1, 2. Общие сведения о допустимых и недопустимых действиях в отношении данных, относящихся к одному из рассмотренный выше типов, приведены в таблице 2.

Таблица 2. Допустимые действия и вычисляемые значения для различных типов данных

	Частотный анализ	$+, -$	$\times, /$	Мода	Медиана	Среднее
Номинативные	да	нет	нет	да	нет	нет
Порядковые	да	вычитание	нет	да	да	нет
Дискретные	да	да	нет	да	да	нет
Непрерывные	да	да	да	да	да	да

3.2. Некоторые аспекты и проблемы измерений

Практикам в области анализа данных хорошо известно, что чаще всего само моделирование связанные с ним действия, занимают примерно 20 % времени, тогда как 80 % уходят на планирование исследования, сбор и предобработку данных. Одним из этапов планирования является *операционализация* — процесс определения способа описания и измерения признаков. Операционализация необходима тогда, когда какие-либо признаки не являются количественными либо по иным причинам не могут быть измерены напрямую. Определение способа кодирования, выбор числа возможных значений — всё это является важной составляющей планирования сбора данных и его реализации.

Во многих случаях необходимый признак не может быть измерен напрямую. В таких случаях можно использовать *опосредованное измерение*, т. е. заменить одном измерение другим. Например, чаще всего нет возможности определить износ станка инструментальными методами. Вместо этого можно провести измерения его наработки в моточасах.

4. Описательные статистики

Важным этапом анализа данных рынка является их первичная интерпретация, позволяющая сделать выводы и разработать план исследования. Форма распределения, наличие выбросов, асимметрия, присутствие выраженных центров плотности — всё это является важной информацией, позволяющей опытному аналитику сразу же сделать выводы, необходимые для быстрой оценки свойств изучаемого явления либо сегмента открытого рынка. Эволюционно мозг человека устроен таким образом, что от 60 до 80 процентов информации поступает в него по визуальному каналу и лишь около 10 процентов по каналу, который можно назвать «смысловым». Подробнее с этим и другими выводами, описывающими основы мышления можно ознакомиться, например в работе бывшего руководителя Нидерладского института головного мозга Дика Свааба «Мы — это наш мозг: от матки до Альцгеймера» [26]. Для целей данной работы это означает, что не следует пренебрегать визуализацией анализируемых информации. Менее наглядными, но ещё более важными являются составление описательных статистик и проверка гипотезы нормальности распределения данных. Эти три операции лежат в основе любого первичного анализа данных открытых рынков.

В данном фрагменте в качестве примера были использованы данные из набора *«almaty-apts-2019-1»* [32], содержащего сведения о 2355 предложения в г. Алматы. Данный набор был любезно предоставлен казахстанской коллегой Г. Шуленбаевой, за что автор выражает ей признательность. Выбор этого набора данных обусловлен тем, что предметом интереса настоящей работы является частотная статистика, основанная на предпосылках о случайности как объективном свойстве объектов, процессов и явлений, а также об отсутствии априорно известных свойств. Автор никогда не был в г. Алматы и ничего не знает о рынке квартир там, что соответствует вышеуказанной предпосылке.

4.1. Визуализация данных

В данном подразделе будут рассмотрены следующие способы визуализации данных:

- гистограмма;
- ядерная оценка плотности;
- ящик с усами (boxplot);
- диаграмма рассеяния.

4.1.1. Построение гистограмм

4.1.1.1. Основные сведения Гистограмма — способ визуального отображения функции плотности вероятности вектора данных. Из этого следует, что гистограмма является способом графического отображения распределения данных (значений случайной величины). Частотность событий откладывается по вертикальной оси, группы данных — по горизонтальной. Полосы столбцов имеют одинаковую ширину.

Рассмотрим алгоритм её построения. Предположим, что мы имеем n наблюдений, содержащих числовые значения $x_1 \dots x_n$. Возьмём интервал $[a, b]$, содержащий все эти числа. Разбиваем его на k частей. Вопрос нахождения оптимального значения k до сих пор является научной проблемой. Некоторые её аспекты и предлагаемые пути решения приведены в 4.1.1.2. Второй проблемой является проблема выбора: должны ли эти части содержать строго одинаковое число наблюдений. На основании эмпирических данных и собственного опыта автор пришёл к выводу о том, что на стадии предварительного анализа данных и их визуализации строгое выдерживание одинаковости количества наблюдений в одном интервале не носит обязательный характер, вследствие чего нет необходимости применять процедуры, обеспечивающие выполнение такого требования. Однако по мере возможности следует придерживаться одинаковости срединных интервалов, при необходимости уменьшая число наблюдений в двух крайних так, чтобы в них оставалось не менее 5 наблюдений [15]. Созданные части обязательно должны быть непересекающимися. Обозначим их как $\Delta_1 \dots \Delta_k$. Вследствие этого возникает числовая ось, имеющая границы $[a, b]$ и состоящая из k непересекающихся частей. Затем путём простого деления

$$\frac{n}{k}, \tag{1}$$

где n — число наблюдений,

k — количество частей, на которые разбиты наблюдения,

определяется число наблюдений, попавших в каждую часть. Далее на прямой $a \dots b$ строят прямоугольники, высота h которых может быть пропорциональна количеству наблюдений, попавших в отрезок.

$$h_i \propto n_i, \tag{2}$$

где n — число наблюдений,

i — номер части.

Данный подход к построению гистограммы является интуитивным, однако он не в полной мере отражает вероятностную суть данного способа визуализации.

Более научный вероятностный подход гласит, что

$$\frac{n_i}{n \times |\Delta i|}, \tag{3}$$

где n_i — число наблюдений в i-той части,

i — номер части,

n — общее число наблюдений,

Δ_i — шаг разбиения.

Использование данной формулы обеспечивает выполнение условия равенства суммы площадей всех прямоугольников единице и является необходимым с точки зрения обеспечения возможности сравнивать гистограммы, построенные на выборках с разным количеством наблюдений, тогда как в случае определения высоты столбцов по формуле 2 она будет зависеть от свойств конкретного единичного набора данных. Второй довод в пользу использования формулы 3 заключается в следующем. В силу равенства суммы площадей прямоугольников единице, данный способ обеспечивает отображение плотности распределения вероятности значений наблюдений. В таком случае можно, хоть и не в явном виде, говорить о возможности определения вероятности того, что измеряемые нами значения наблюдений попадут в интервал $[a, b]$. В этом случае получается, что данная вероятность может быть рассчитана как интеграл от плотности вероятности.

$$P(x \in [a, b]) = \int_a^b f(t)dt \tag{4}$$

По мере увеличения числа наблюдений и в случае правильного выбора k гистограмма будет всё меньше отличаться от функции плотности.

Ключевым моментом при восприятии гистограммы должно быть то, что при её анализе мы смотрим на площади. Если вычислить сумму площадей прямоугольников, расположенных на отрезке оси абсцисс $\alpha \dots \beta$, это будет означать, что мы определили вероятность того, что значение случайной величины будет находиться в диапазоне $\alpha \dots \beta$. Разумеется данное утверждение является справедливым только в случае использование формулы 3 для определения высоты столбцов.

При этом следует отметить, что использование формулы 2 может быть оправдано в случаях подготовки гистограмм в презентационных и маркетинговых целях. В целях анализа данных осмысленным является использование исключительно формулы 3.

4.1.1.2. Выбор рационального числа интервалов Число интервалов (k), используемое при вычислении оценок параметров и построении гистограмм, колеблется в широких пределах. Большинство рекомендуемых формул носят эмпирический характер и часто дают завышенные значения. В общем случае можно говорить о том, что количество интервалов k связано с количеством наблюдений переменной. При слишком малом k гистограмма будет отличаться от действительной кривой плотности вследствие слишком крупной ступенчатости, из-за чего будет потеряна информация об особенностях распределения. Так, если представить ситуацию, при которой интервал группировки равен размаху, гистограмма любого распределения будет выглядеть так, как будто имеет место равномерное распределение. В случае с тремя интервалами любое колоколообразное распределение будет сведено к треугольному. При этом в случае слишком большого значения k гистограмма будет содержать не только полезную информацию, но и значительную долю «шума».

Таким образом, необходимо найти баланс, который обеспечит сохранение «сигнала», отсеяв при этом «шум». Таким образом, задача поиска k при построении гистограммы сводится к *задаче оптимальной фильтрации*, а оптимальным значением k является такое, при котором максимально возможное сохранение «сигнала» сочетается с максимальной возможной фильтрацией «шума», иными словами, максимальное сглаживание флуктуаций должно сочетаться с минимальным искажением кривой плотности распределения. Одним из практических признаков приближения к оптимальному k является исчезновение «пиков» и «провалов», а ближайшим к оптимальному может считаться такое максимальное значение k, при котором гистограмма ещё сохраняет плавный характер. На практике можно сформулировать просто правило: при $n > 200$ каждый столбец должен содержать не менее 10 наблюдений. При $n \leq 200$ крайние столбцы могут содержать не менее 5 наблюдений. Значительное число рекомендаций по выбору числа k из различных источников содержится в [15].

В целом можно сказать, что существует пять групп методов определения оптимального значения k:

- обобщённые;
- основанные на значении n;
- основанные на критерии согласия χ^2;
- основанные на энтропийном коэффициенте;
- основанные на четвёртом центральном моменте.

4.1.1.2.1. Обобщённые методы определения числа k В первом приближении можно использовать рекомендацию, данную в [6], и говорить о том, что

$$6 \leq k \leq 20. \tag{5}$$

Также существует ещё более обобщённая рекомендация

$$k = 12 \pm 2 \ldots 3. \tag{6}$$

Данные методы являются оценочными и могут служить лишь ориентиром при первом знакомстве с данными.

4.1.1.2.2. Методы определения числа k на основе n Первая формула, основанная на числе n была предложена Гербертом Стёрджессом в 1926 году в работе [47]

$$k = \frac{range}{1 + \log_2 n}, \tag{7}$$

где range — размах,

n — число наблюдений.

В той же работе были предложены формулы:

$$k = 1 + \log_2 n, \tag{8}$$

$$k = 1 + 3.3 \lg n, \tag{9}$$

Следует отметить, что значения, получаемые на основе формул 8 и 9 приблизительно равны между собой

$$1 + \log_2 n \approx 1 + 3.3 \lg n. \tag{10}$$

Некоторые рассуждения, касающиеся данных формул, приведены в [48]. В работе [6] без ссылки на первоисточник предлагается формула Брукса и Каррузера

$$k = 5 \lg n. \tag{11}$$

В 1965 году в книге [3] И. Хайнхольд и К. Гаеде указывают соотношение

$$k = \sqrt{n}. \tag{12}$$

На рисунках 5, 6 показан прирост k при $n < 400$ и при $n < 50000$ соответственно.

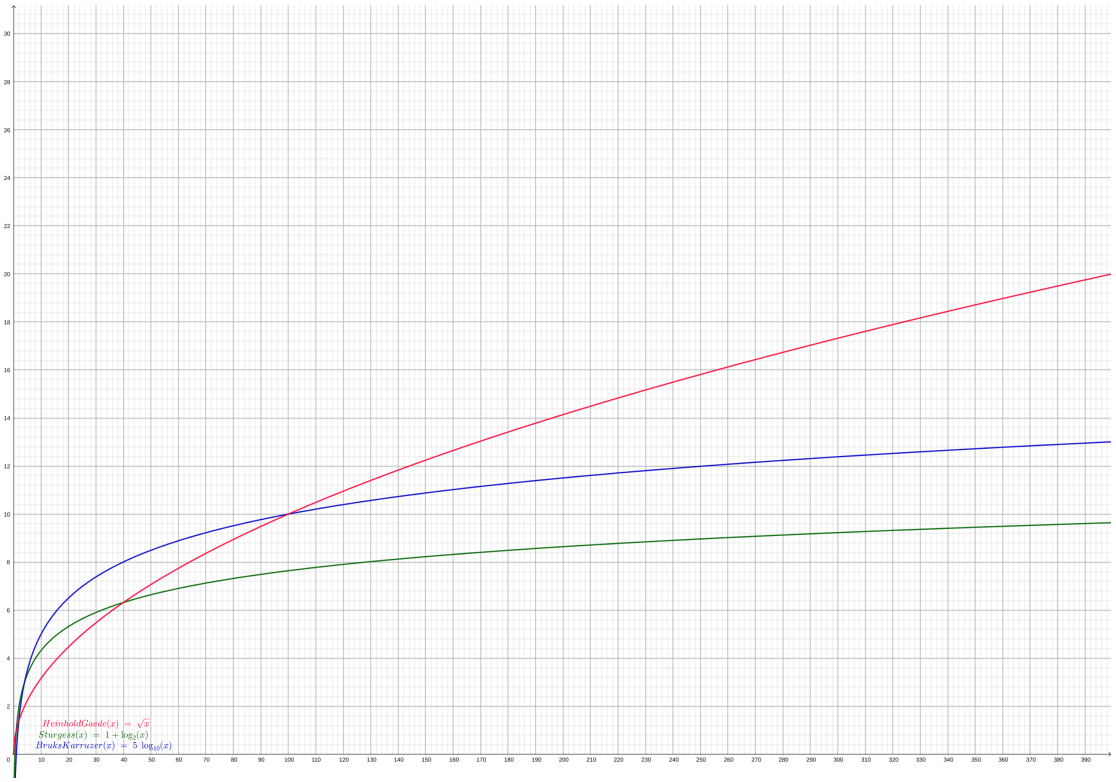

Рис. 5. Визуализация числа k при $n < 400$ в случае использования методов, основанных на значении n

В настоящее время наибольшее распространение получила формула Стёрджесса (формулы 8, 9).

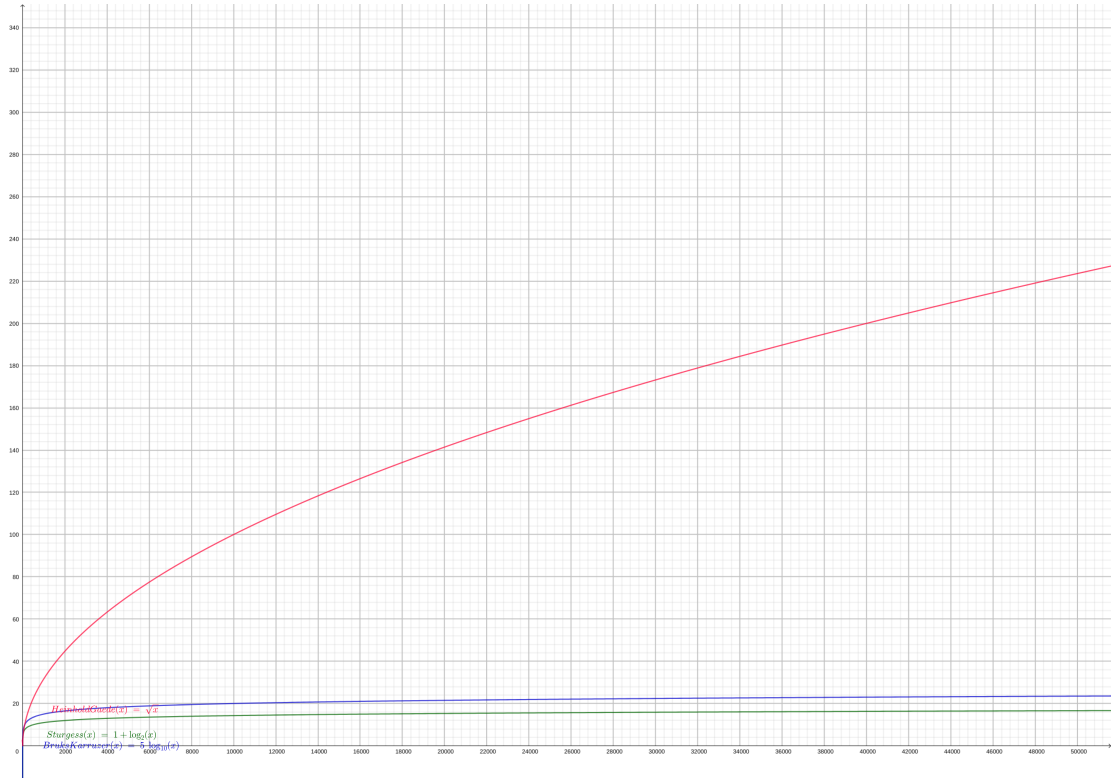

Рис. 6. Визуализация числа k при $n < 50000$ в случае использования методов, основанных на значении n

4.1.1.2.3. Методы определения числа k на основе критерия согласия χ^2 Сам по себе критерий χ^2 не может быть использован без разбиения выборки на интервалы, в которых производится вычисление частных разностей между теоретической моделью и эмпирической выборкой [15]. Исходной предпосылкой эффективности применения критерия χ^2 является использование интервалов, имеющих не равную длину, а равную вероятность в соответствии с теоретической моделью. Следует сказать, что число m интервалов равной длины и число k равной вероятности могут существенно отличаться.

В 1942 году Х. Манном и А. Вальдом было установлено, что при $n \longrightarrow \infty$ оптимальное число равновероятных интервалов k может быть определено по формуле

$$k \approx 4\sqrt[5]{2}(\frac{n}{t})^{0.4}, \tag{13}$$

где t — квантиль нормального распределения, соответствующий заданной вероятности

$$P = 1 - p, \tag{14}$$

где p — принятый уровень значимости [1].

В 1950 году К. Уильямс в работе [2] показал, что коэффициент 4 может быть заменён на 2 без существенной потери информации. Таким образом, формула приобретает вид

$$k \approx 2\sqrt[5]{2}(\frac{n}{t})^{0.4}, \tag{15}$$

В 1973 году М. Кендалом и А. Стюартом в работе [7] было предложено дальнейшее развитие формулы 13

$$k \leq b[\sqrt{2}\frac{(n-1)}{t_1+t_2}]^{0.4}, \tag{16}$$

где $2 \leq b \leq 4$

t_1, t_2 — некоторые задаваемые квантили

В 1967 году Г. Хан и С. Шапиро в работе [5] предложили упрощённый вариант формулы

$$k = 4[0.75(n-1)^2]^{0.2}, \tag{17}$$

что может быть получено при подстановке в формулу 16 $b = 4, t = 1.645$. В случае $b = 2$ выражение принимает вид

$$k = 1.9n^{0.4} \tag{18}$$

На рисунках 7, 8 показана зависимость k от n, определяемая на основе критерия согласия χ^2.

4.1.1.2.4. Методы определения числа k на основе энтропийного коэффициента В 1970–1980-е годы был проведён ряд исследований, основанных на использовании критерия близости в виде энтропийного коэффициента k_∂. Понятий энтропийного коэффициента как числовой характеристики формы распределения впервые было предложено П. В. Новицким в [4]. В случае с гистограммой в соответствии с [10] эта оценка вычисляется как

$$k_\partial = \frac{dn}{2\sigma}10^{-\frac{1}{n}\sum\limits_{j=1}^{n} n_j \log_{10} n_j}, \tag{19}$$

где d — ширина столбца гистограммы,

σ — стандартное отклонение,

m — число столбцов гистограммы,

n_j — число наблюдений в j-ном столбце.

В результате работ под руководством З. Таушанова, основанных на вышеуказанном принципе, в 1973 году в работе [8] было выведено соотношение

$$k = 4\log_{10} n \tag{20}$$

D 1981 году дальнейшее развитие формулы 20 со стороны Е. Тоневой [11] привело к разработке соотношения

$$k = 5\lg n - 5 \equiv 5\lg(\frac{n}{10}). \tag{21}$$

На рисунках 9, 10 реализована визуализация функции зависимости k от n в случае использования методов, основанных на энтропийном коэффициенте.

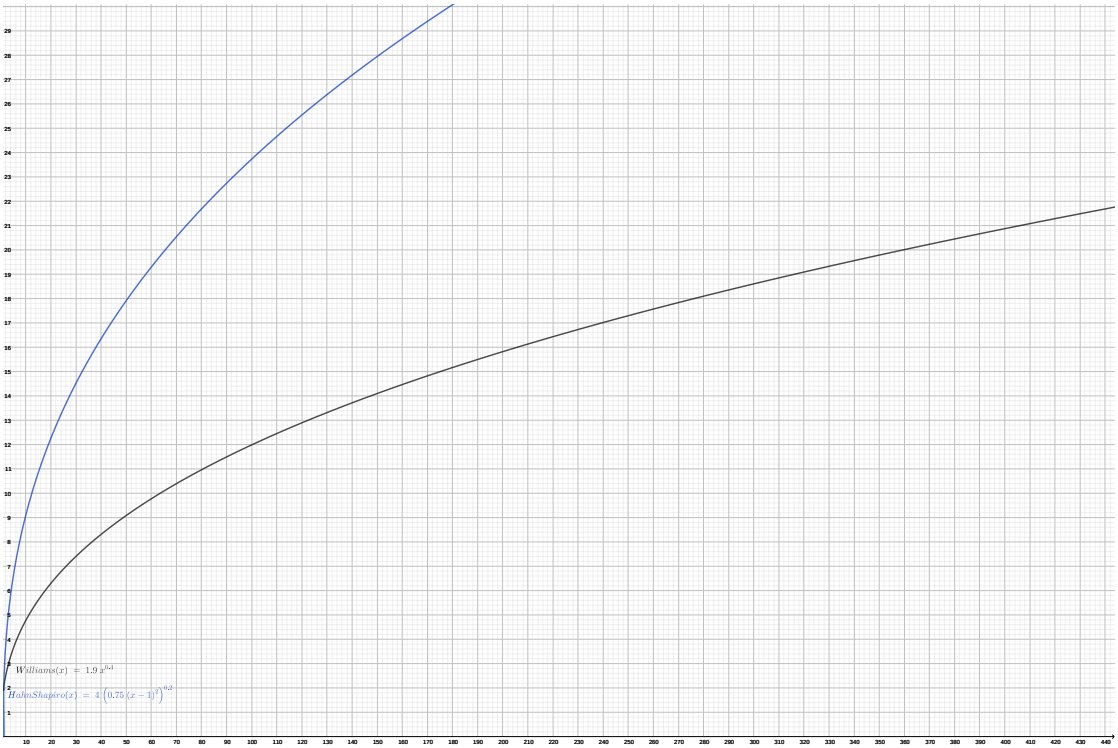

Рис. 7. Визуализация числа k при $n < 400$ в случае использования методов, основанных на критерии согласия χ^2

4.1.1.2.5. Методы определения числа k на основе четвёртого центрального момента Следует отметить, что все исследования, описанные в 4.1.1.2.1–4.1.1.2.4, начиная с работ Стёрджесса, закачивая исследованиями энтропийного коэффициента, рассматривают k исключительно как функцию n, расходясь только в выборе вида этой функции.

Принципиально новый подход был предложен в 1975 году И. У. Алексеевой в работе [9]. Согласно выводам, сделанным автором данной работы, k существенно зависит от значения *коэффициента контрэксцесса*, являющегося величиной обратной *коэффициенту эксцесса*, представляющего в свою очередь относительную меру **четвёртого центрального момента**. Вопросы *центральных моментов* рассмотрены в 5. Следует отметить, что все центральные моменты взаимосвязаны между собой, таким образом, значение четвёртого содержит в себе информацию о первых трёх, что означает учёт обширной информации о об изучаемых данных.

$$k = \frac{4}{\xi} \lg \frac{n}{10}, \tag{22}$$

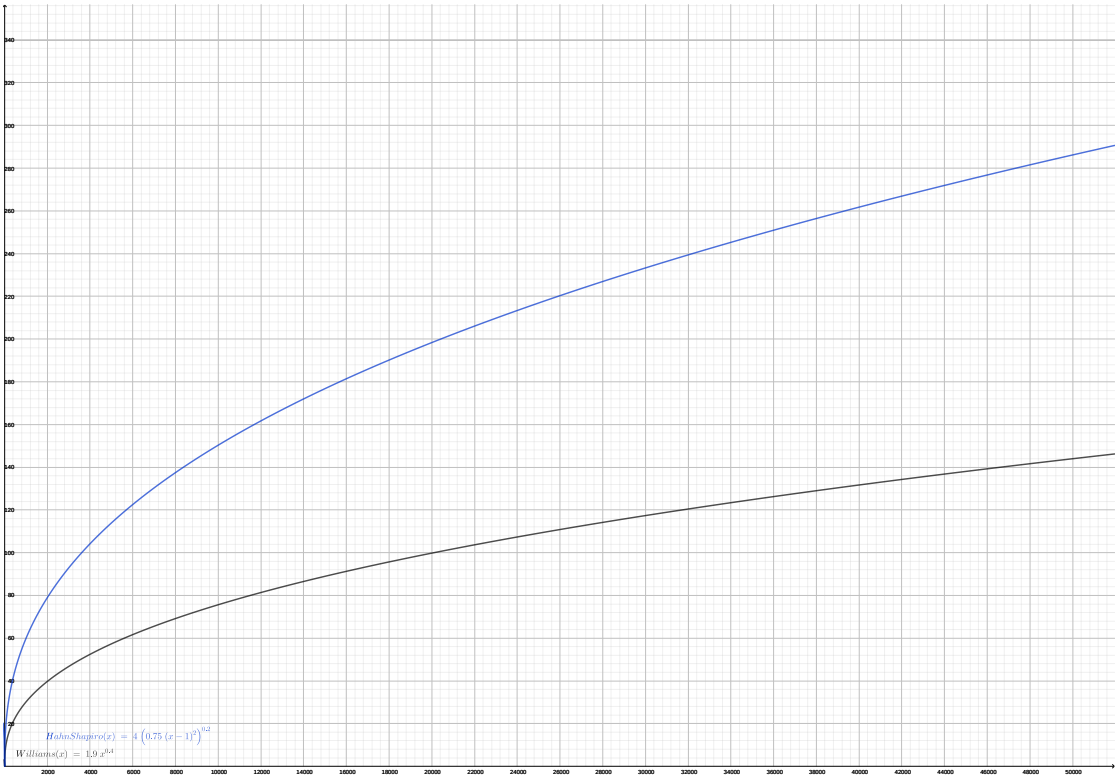

Рис. 8. Визуализация числа k при $n < 50000$ в случае использования методов, основанных на критерии согласия χ^2

где ξ — значение контрэксцесса

$$\xi = \frac{1}{\sqrt{\varepsilon}} == \frac{S^2}{\sqrt{\mu_4}}, \tag{23}$$

, где ε — значение эксцесса,

S — дисперсия,

μ_4 — четвёртый центральный момент.

Как было сказано ранее, детальное рассмотрение вопросов центральных моментов проводится в 5.

Формула 22 открывает новые возможности анализа при первичной интерпретации и визуализации данных. Полученный в [9] сдвиг функции $k = f(n)$ может быть аппроксимирован в зависимости от значения эксцесса выражениями

$$A(\varepsilon) = \frac{\varepsilon + 1.5}{6} \tag{24}$$

либо

$$A(\varepsilon) = \frac{\varepsilon^{0.8}}{3}. \tag{25}$$

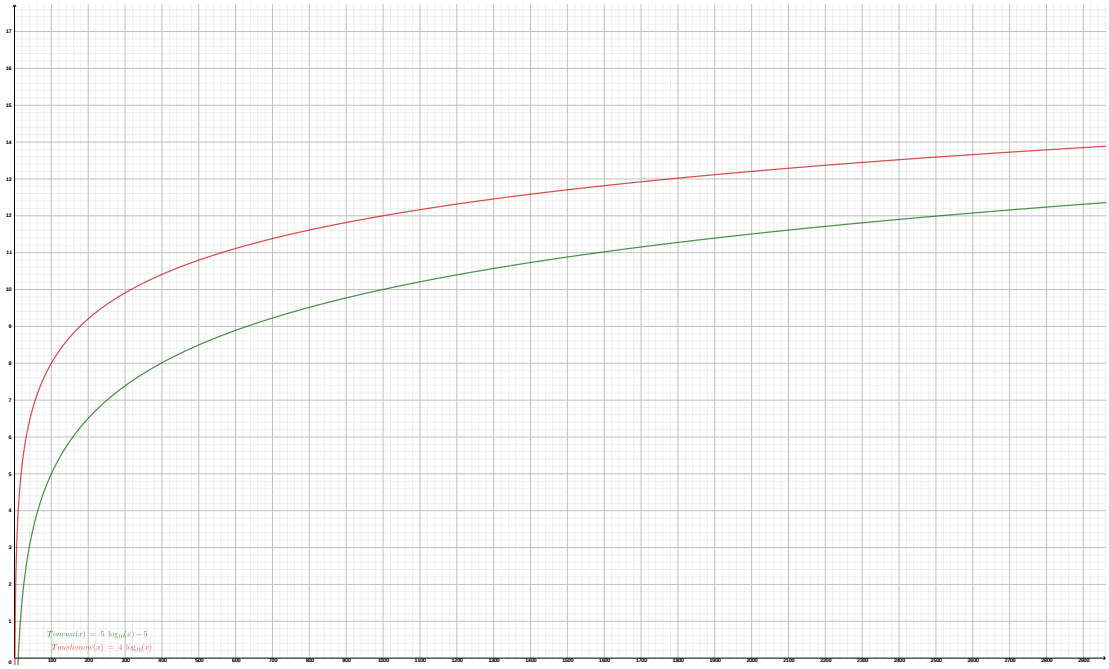

Рис. 9. Визуализация числа k при $n < 3000$ в случае использования методов, основанных на энтропийном коэффициенте

Вследствие этого дальнейшее развитие формулы 22 привело к разработке соотношений

$$k = \frac{\varepsilon + 1.5}{6} n^{0.4}, \tag{26}$$

а также

$$k = \frac{1}{3} \sqrt[5]{\varepsilon^4 n^2} \equiv \frac{1}{3} \sqrt[5]{\frac{n^2}{\xi^8}}. \tag{27}$$

Далее на основе формулы 22 было выведено правило, согласно которому

$$0.55 n^{0.4} \leq k \leq 1.25 n^{0.4} \tag{28}$$

При этом рекомендуется выбирать нечётный k, поскольку в противном случае возникает эффект принудительного уплощения центра гистограммы [15].

4.1.1.2.6. Выводы по итогам анализа методов определения k На основе изложенного выше можно сделать следующие выводы:

1) задача выбора числа интервалов группировки данных выборки представляет собой задачу оптимизации отклонений гистограммы от кривой плотности распределения, соответствующей генеральной совокупности;

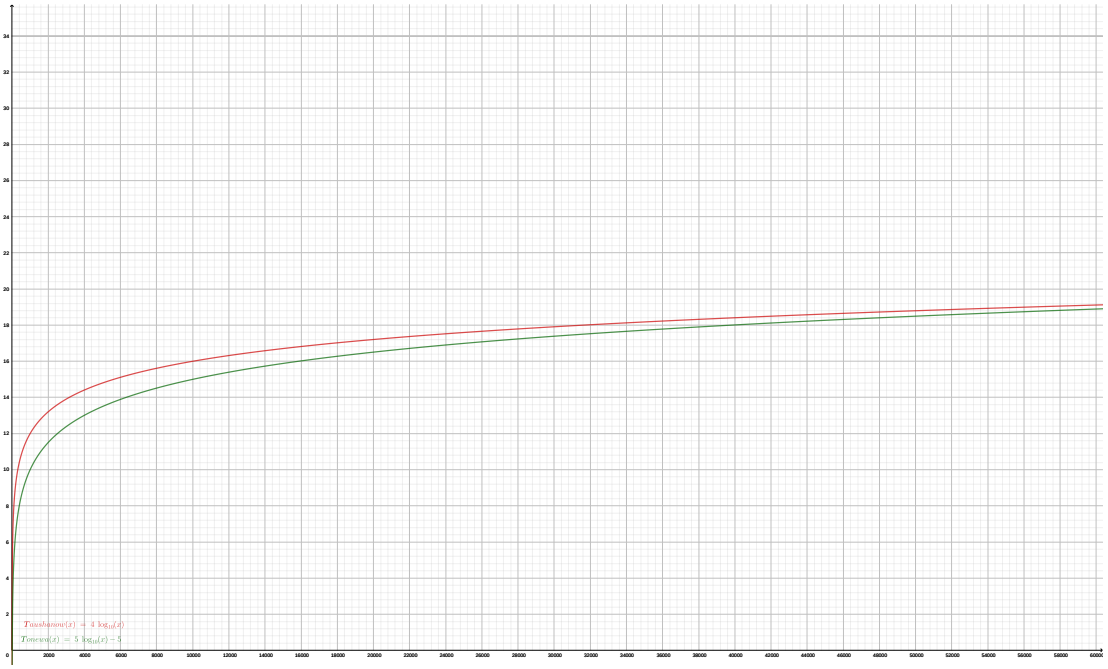

Рис. 10. Визуализация числа k при $n < 50000$ в случае использования методов, основанных на энтропийном коэффициенте

2) существует такое значение k, которое является оптимальным с точки зрения близости полигона гистограммы к плавной кривой плотности распределения генеральной совокупности при данном законе распределения и данном размере выборки;

3) использование интервалов равной вероятности вместо равно длины позволяет учитывать форму закона распределения (см. формулу **??**);

4) в случае использования интервалов равной длины их число в большей степени зависит от коэффициента эксцесса ε, а не объёма выборки n;

5) зависимость k от n лучше всего аппроксимируется выражением $k = An^{0.4}$, где A описывается выражениями 24, 25.

Далее в таблице 3 приводятся сводные данные по всем методам.

4.1.1.3. Реализация Создадим функции, рассчитывающие число k в соответствии с рассмотренными в 4.1.1.2 методами. Данное действие не является обязательным. По умолчанию R строит гистограммы с числом k, определённым на основе формулы Стёрджесса (формулы 9). Вопросы программирования на R и Python рассматриваются в соответствующих фрагментах. Здесь следует сказать лишь то, можно написать формулу для конкретного случая и набора данных, однако хорошей практикой является создание функций, их сохранение и многократное использование

Таблица 3. Обобщение сведений о методах определения числа интервалов при построении гистограмм

№	Автор	Год	Формула	Номер	Источник
			Основанные только на количестве n		
1	Г. Стёрджесс	1926	$k = \frac{range}{1+\log_2 n}$	7	[47], [48]
2	Г. Стёрджесс	1926	$k = 1 + \log_2 n$	8	[47],[48]
3	Г. Стёрджесс	1926	$k = 1 + 3.3 \lg n$	8	[47],[48]
4	Брукс, Каррузер	?	$k = 5 \lg n$	11	[6]
5	И Хайнхольд, К Гаеде	1965	$k = \sqrt{n}$	12	[3]
			Основанные на критерии согласия χ^2		
6	Х. Манн, А. Вальд	1942	$k \approx 4\sqrt[5]{2}(\frac{n}{t})^{0.4}$	13	[1]
7	К. Уильямс	1950	$k \approx 2\sqrt[5]{2}(\frac{n}{t})^{0.4}$	15	[2]
8	Г. Хан, С Шапиро	1967	$k = 4[0.75(n-1)^2]^{0.2}$	17	[5]
9	Г. Хан, С Шапиро	1967	$k = 1.9n^{0.4}$	17	[5]
10	М. Кендалл, А Стюарт	1970	$k \leq b[\sqrt{2}\frac{(n-1)}{t_1+t_2}]^{0.4}$	16	[7]
			Основанные на энтропийном коэффициенте		
11	З. Таушанов	1973	$k = 4 \log_{10} n$	20	[8]
12	Е. Тонева	1981	$k = 5 \lg n - 5 \equiv 5 \lg(\frac{n}{10})$	21	[11]
			Основанные на четвёртом центральном моменте		
13	И Алексеева	1975	$k = \frac{4}{\xi} \lg \frac{n}{10}$	22	[9]
14	П Новицкий	1991	$k = \frac{\varepsilon+1.5}{6} n^{0.4}$	26	[15]
15	П Новицкий	1991	$k = \frac{1}{3}\sqrt[5]{\varepsilon^4 n^2} \equiv \frac{1}{3}\sqrt[5]{\frac{n^2}{\xi^8}}$	27	[15]
16	П Новицкий	1991	$0.55n^{0.4} \leq k \leq 1.25n^{0.4}$	28	[15]

в дальнейшей работе. Важно помнить, что переменные, создаваемые внутри тела функции, не возникают в глобальном окружении, вследствие этого им можно присваивать любые имена. В скрипте 3 приведён код, позволяющий определить k всеми 15-ю рассмотренными выше способами. Файл со скриптом доступен в репозитории по ссылке [49].

Листинг 3. Создание функций для автоматизированного определения оптимального значения k

```
kHistSturgess0 <- function(x, na.omit = FALSE){
n    <- NROW(x)
ks0 = (max(x)-min(x))/(1+log2(n))
return(c(ks0))
}

# создатьфункциюдляпервойверсииформулыСтёрджесса
kHistSturgess1 <- function(x, na.omit = FALSE){
n    <- NROW(x)
```

```
10 ks1 = (1+log2(n))
   return(ks1)
12 }

14 # создатьфункциюдлявторойверсииформулыСтёрджесса
   kHistSturgess2 <- function(x, na.omit = FALSE){
16 n    <- NROW(x)
   ks2 = (1+(3.3*log10(n)))
18 return(ks2)
   }
20
   # создатьфункциюдляформулыБруксаиКаррузера
22 kHistBruksKarruzer <- function(x, na.omit = FALSE){
   n    <- NROW(x)
24 kbk = 5*log10(n)
   return(kbk)
26 }

28 # создатьфункциюдляформулыХайнхольдаиГаеде
   kHistHeinholdGaede <- function(x, na.omit = FALSE){
30 n    <- NROW(x)
   khg = sqrt(n)
32 return(khg)
   }
34
   # создатьфункциюдляформулыМаннаиВальда
36 kHistMannWald <- function(x, na.omit = FALSE){
   n    <- NROW(x)
38 kmw = (4*(2^(1/5)))*((n/qnorm(0.95))^0.4)
   return(kmw)
40 }

42 # создатьфункциюдляформулыУильямса
   kHistWilliams <- function(x, na.omit = FALSE){
44 n    <- NROW(x)
   kwi = (2*(2^(1/5)))*((n/qnorm(0.95))^0.4)
46 return(kwi)
   }
48
   # создатьфункциюдляпервойформулыХанаиШапиро
50 kHistHahnShapiro <- function(x, na.omit = FALSE){
   n    <- NROW(x)
52 khs = 4*(0.75*((n-1)^2)^0.2)
   return(khs)
```

```
54 }

56 # создатьфункциюдлявторойформулыХанаиШапиро
   kHistShapiroHahn <- function(x, na.omit = FALSE){
58 n    <- NROW(x)
   ksh = 1.9*(n^0.4)
60 return(ksh)
   }
62
   # создатьфункциюдляформулыКендаллаиСтюарта
64 kHistKendallStuart <- function(x, na.omit = FALSE){
   n    <- NROW(x)
66 b = 2
   t1 = qnorm(0.95)
68 t2 = 0
   kks = b*(sqrt(2)*(((n-1)/(t1+t2))^0.4))
70 return(kks)
   }
72
   # создатьфункциюдляформулыТаушанова
74 kHistTaushanow <- function(x, na.omit = FALSE){
   n    <- NROW(x)
76 kta = 4*log10(n)
   return(kta)
78 }

80 # создатьфункциюдляформулыТоневой
   kHistTonewa <- function(x, na.omit = FALSE){
82 n    <- NROW(x)
   kto = 5*log10(n)-5
84 return(kto)
   }
86
   # создатьфункциюдляформулыАлексеевой
88 kHistAlekseewa <- function(x, na.omit = FALSE){
   n    <- NROW(x)
90 kurt = kurtosis(x)
   counterkurt = 1/(sqrt(kurt))
92 kal = (4/counterkurt)*(log10(n/10))
   return(kal)
94 }

96 # создатьфункциюдляпервойформулыНовицкого
   kHistNowiczkij1 <- function(x, na.omit = FALSE){
```

```
98 n    <- NROW(x)
   kurt = kurtosis(x)
100 kn1 = ((kurt+1.5)/6)*(n^0.4)
   return(kn1)
102 }

104 # создатьфункциюдлявторойформулыНовицкого
   kHistNowiczkij2 <- function(x, na.omit = FALSE){
106 n    <- NROW(x)
   kurt = kurtosis(x)
108 kn2 = (((kurt^4)*(n^2))^(1/5))*(1/3)
   return(kn2)
110 }

112 # создатьфункциюдлятретьейформулыНовицкого
   kHistNowiczkij3_min <- function(x, na.omit = FALSE){
114 n    <- NROW(x)
   kurt = kurtosis(x)
116 kn3min = 0.55*(n^0.4)
   return(kn3min)
118 }

120 # создатьфункциюдлятретьейформулыНовицкого
   kHistNowiczkij3_max <- function(x, na.omit = FALSE){
122 n    <- NROW(x)
   kurt = kurtosis(x)
124 kn3max = 1.25*(n^0.4)
   return(kn3max)
126 }

128 optimalK <- function(x, na.omit = FALSE){
   ks0 <- kHistSturgess0(x)
130 ks1 <- kHistSturgess1(x)
   ks2 <- kHistSturgess2(x)
132 kbk <- kHistBruksKarruzer(x)
   khg <- kHistHeinholdGaede(x)
134 kmw <- kHistMannWald(x)
   kwi <- kHistWilliams(x)
136 khs <- kHistHahnShapiro(x)
   ksh <- kHistShapiroHahn(x)
138 kks <- kHistKendallStuart(x)
   kta <- kHistTaushanow(x)
140 kto <- kHistTonewa(x)
   kal <- kHistAlekseewa(x)
```

Листинг 4. Создание функций для автоматизированного определения оптимального значения k

```
optimal_k <- optimalK(almatyFlats$price.m)
function_name<- c("Sturgess0",
          "Sturgess1",
          "Sturgess2",
          "BruksKarruzer",
          "HeinholdGaede",
          "MannWald",
          "Williams",
          "HahnShapiro",
          "ShapiroHahn",
          "KendallStuart",
          "Taushanow",
          "Tonewa",
          "Alekseewa",
          "Nowiczkij1",
          "Nowiczkij2",
          "Nowiczkij3_min",
          "Nowiczkij3_max")
k_optimal <- tibble(function_name, optimal_k)
```

```
kn1 <- kHistNowiczkij1(x)
kn2 <- kHistNowiczkij2(x)
kn3_min <- kHistNowiczkij3_min(x)
kn3_max <- kHistNowiczkij3_max(x)
optimal_k <- return(c(ks0, ks1, ks2, kbk, khg, kmw, kwi,
          khs, ksh, kks, kta, kto, kal, kn1, kn2, kn3_min, kn3_max))
return(k)
}
```

Далее применим функции к набору данных «almatyFlats» (см. скрипт4).

В результате были получены следующие значения k (см. таблицу 4).

Для сравнения также был проведён анализ оптимального k для рынка квартир городской агломерации Санкт-Петербург. Для этого взят набор данных «spba_flats_210928» [36], собранный автором 28 сентября 2021 года путём парсинга сайта cian.ru и содержащий сведения о 34821 предложениях. Все наборы данных, использованные при составлении данного фрагмента, доступны в репозитории [49]. Результаты расчётов k приведены в таблице 5.

Далее выставим число интервалов по умолчанию и построим гистограмму для рынка Алматы, см. рисунок 11.

Затем применим формулу 24 для определения числа k и добавим на график

Таблица 4. Оптимальное значение k для переменной, содержащей сведения об удельной цене предложения на квартиры в г. Алматы, полученное различными методами

№	Метод	Значение k
1	Sturgess0	66513.973
2	Sturgess1	12.202
3	Sturgess2	12.128
4	BruksKarruzer	16.860
5	HeinholdGaede	48.528
6	MannWald	84.064
7	Williams	42.032
8	HahnShapiro	66.964
9	ShapiroHahn	42.418
10	KendallStuart	51.739
11	Taushanow	13.488
12	Tonewa	11.860
13	Alekseewa	25.219
14	Nowiczkij1	31.869
15	Nowiczkij2	35.560
16	Nowiczkij3_min	12.279
17	Nowiczkij3_max	27.907

Листинг 5. Построение простой гистограммы для г. Алматы

```
hist(almatyFlats$price.m,
     freq = FALSE,
     xlab = ''price, kaz rubles'',
     ylab = ''probability'',
     main = ''Price per meter histogram, Almaty, summer 2019'')
```

кривые плотности теоретического нормального и эмпирического распределений, см. рисунок 12.

Для сравнения построим гистограмму для аналогичной переменной набора данных [36], см. рисунок 13.

Подытоживая вышесказанное, следует повторить, что в первую очередь надо смотреть на площади столбцов, поскольку они содержат информацию о вероятности. Скрипты для построения приведённых гистограмм приведены в листингах 5, 6, 7.

Таблица 5. Оптимальное значение k для переменной, содержащей сведения об удельной цене предложения на квартиры в городской агломерации Санкт-Петербург, полученное различными методами

№	Метод	Значение k
1	Sturgess0	100989.27
2	Sturgess1	16.09
3	Sturgess2	15.99
4	BruksKarruzer	22.71
5	HeinholdGaede	186.60
6	MannWald	246.92
7	Williams	123.46
8	HahnShapiro	196.72
9	ShapiroHahn	124.59
10	KendallStuart	151.99
11	Taushanow	18.17
12	Tonewa	17.71
13	Alekseewa	97.56
14	Nowiczkij1	534.62
15	Nowiczkij2	479.03
16	Nowiczkij3_min	36.07
17	Nowiczkij3_max	81.97

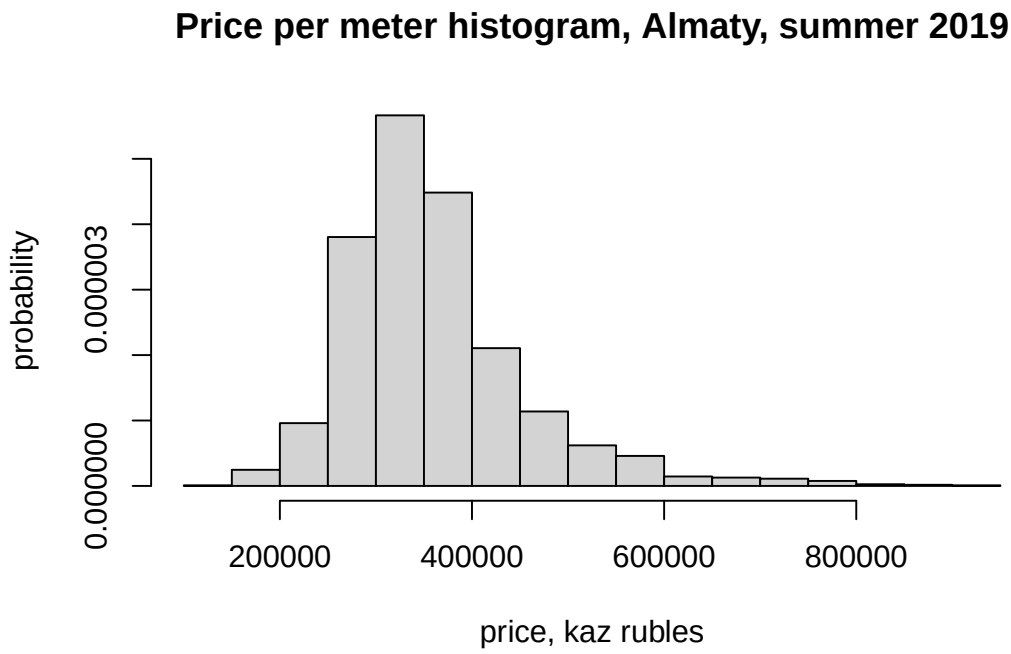

Рис. 11. Гистограмма удельных цен на рынке города Алматы (лето 2019 г.)

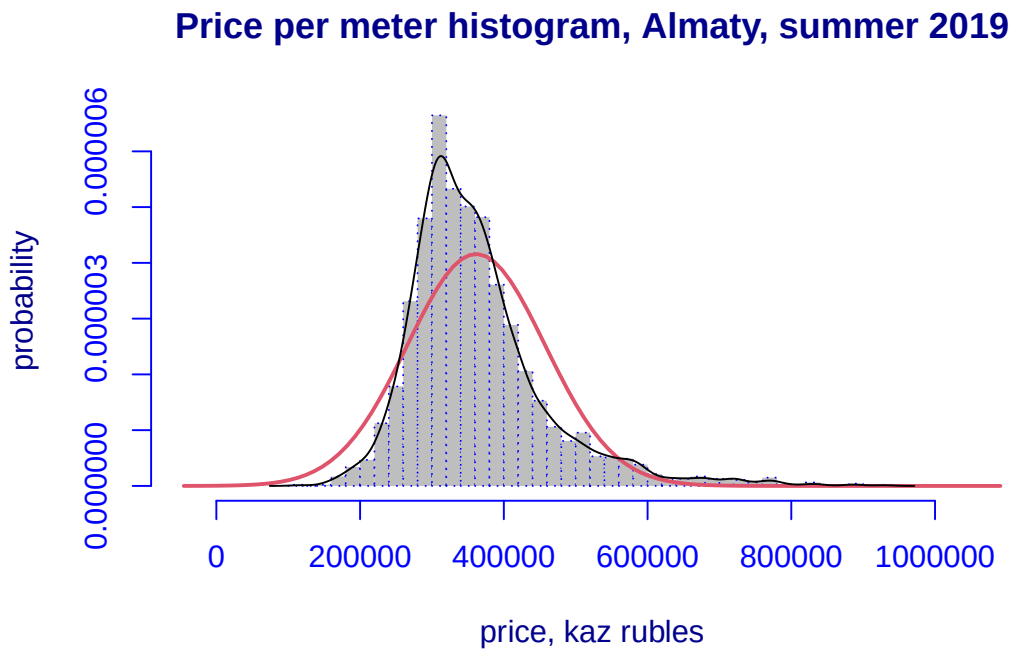

Рис. 12. Гистограмма удельных цен на рынке города Алматы (лето 2019 г.), совмещённая с кривыми плотности теоретического нормального и эмпирического распределений

лишь весьма ограниченный круг вопросов. В списке источников информации дана некоторая подборка материалов, которые, по мнению автора, могут быть полезными в дальнейшем освоении вопросов математической статистики. Поскольку оценщики, как правило, очень занятые люди, освоение вряд ли можно ожидать, что кто-то захочет сразу изучать десятки материалов. В связи с этим автор рекомендует ознакомление в первую очередь со следующими работами:

- В. Савельев. *Статистика и котики* [31]. Прекрасная книга для тех, кто только начинает погружение в область математической статистики.
- С. Бослаф. *Статистика для всех* [27]. Одна из лучших книг по статистике для тех, кто не учился по профилю и хочет освоить её методы на уровне, достаточном для применения в профессиональной деятельности.
- А. Кобзарь. *Прикладная математическая статистика* [23]. Данный материал представляет собой обширное пособие, предназначенное для тех, кто уже владеет некоторой базой и хочет внедрить применение методов математической статистики на профессиональном уровне.

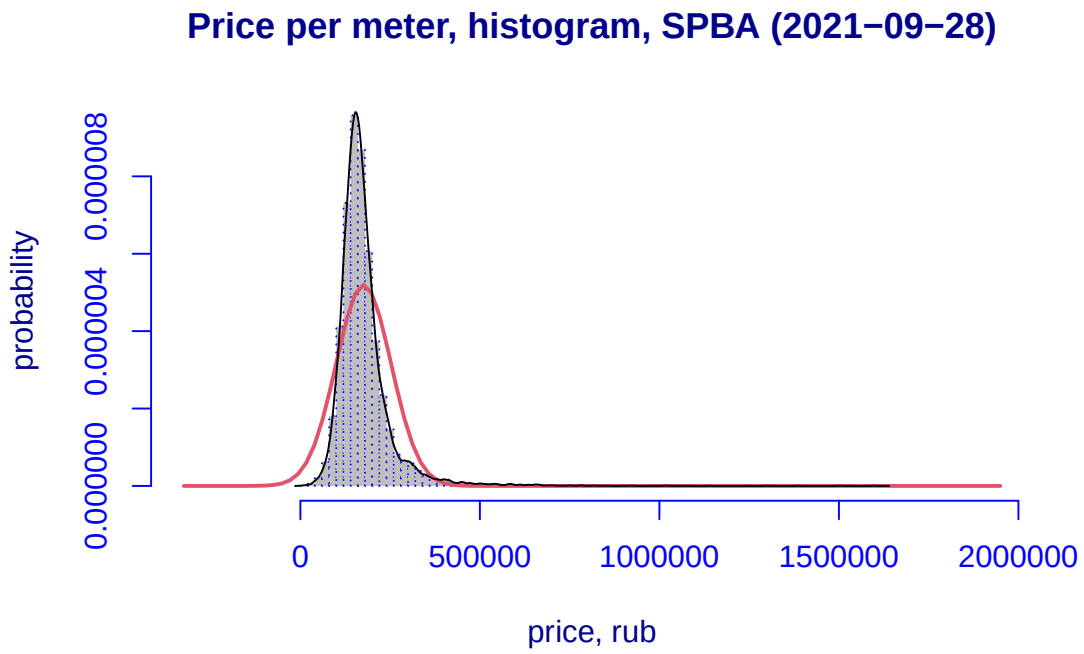

Рис. 13. Гистограмма удельных цен на рынке города Алматы (лето 2019 г.), совмещённая с кривыми плотности теоретического нормального и эмпирического распределений

Учиться следует всю жизнь. Оценочная деятельность трансформируется, и, спустя несколько лет, данное пособие будет казаться детской раскраской. Автор желает читателям непрерывного совершенствования и становления в качестве настоящих специалистов в сфере цифровой оценки XXI века.

Источники информации

[1] H. B. Mann и A. Wald. «On the Choice of the Number of Class Intervals in the Application of the Chi Square Test». В: 13.3 (1942-09), с. 306—317. DOI: 10.1214/aoms/1177731569.

[2] C. A. Williams. «On the choice of the number and midth of classes of chi-square test of goodness of fit». В: *Journal of the Amercian Statistical Association* 45 (1950), с. 77—78.

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

[4] П. В. Новицкий. «Понятие энтропийного значения погрешости». В: *Измерительная техника* 7 (1966), с. 11—14.

Листинг 6. Построение гистограммы совмещённой с кривыми плотности теоретического нормального и эмпирического распределений для г. Алматы

```
histDist(almatyFlats$price.m,
         nbins = kHistNowiczkij1(almatyFlats$price.m),
         xlab = "price, kaz rubles",
         ylab = "probability",
         main = "Price per meter histogram, Almaty, summer 2019")
lines(density(almatyFlats$price.m))
```

Листинг 7. Построение гистограммы совмещённой с кривыми плотности теоретического нормального и эмпирического распределений для городской агломерации Санкт-Петербург

```
histDist(spbaFlats$price_m,
         nbins = kHistAlekseewa(spbaFlats$price_m),
         xlab = "price, rub",
         ylab = "probability",
         main = "Price per meter, histogram, SPBA (2021-09-28)")
lines(density(spbaFlats$price_m))
```


[5] G. I. Hahn и S. Shapiro. *Statistical models in engineering*. New York, London, Sydnei: GE Company, 1967, с. 396.

[6] Р. Шторм. *Теория вероятности. Математическая статистика. Статистический контроль качества*. 1970, с. 368. URL: https://www.studmed.ru/shtorm-r-teoriya-veroyatnosti-matematicheskaya-statistika-statisticheskiy-kontrol-kachestva_26ba1b67977.html (дата обр. 15.10.2021).

[7] Maurice G. Kendall и Alan Stuart. *Статистические выводы и связи. Перевод с английского*. Л. И. Гальчук, А. Т. Терёхин под ред. А. Н. Колмогорова. Пер. А. Т. Терёхин Л. И. Гальчук. Примеч. А. Н. Колмогоров. 1973.

[8] З. Таушанов, Е. Тонева и Р. Пенова. «Вычисление энтропийного коэф-фи-ци-ента при малых выборках». В: *Изобретательство, стандартизация и качество* 5 (1973).

[9] И. У. Алексеева. «Теоретическое и экспериментальное исследование за-конов распределения погрешностей, их классификация и методы оценки их па-раметров. Автореф. дис. на соиск. учен. степени кан. техн. наук.» Дис. ... док. Ленинград, 1975.

[10] П. В. Новицкий. *Электрические измерения неэлектрических величин*. Ленинград: Энергия, 1975, с. 576.

[11] Е. Тонева. «Аппроксимация распределений погрешности средств из-ме-ре-ний». В: *Измерительная техника* 6 (1981), с. 15—16.

[12] С. А. Айвазян. *Прикладная статистика: исследование зависимостей*. 1985.

[13] Н. И. Портенко / А. В. Скороход / А. Ф. Турбин В. С. Королюк. *Справочник по теории вероятностей и математической статистике*. Москва: Наука, 1985, с. 640.

[14] Г. Смит Н. Дрейпер. *Прикладной регрессионный анализ*. 1987.

[15] 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).

[16] Ю. В. Прохоров. *Вероятность и математическая статистика. Энциклопедия*. Moskva: Nauch. izd-vo "Bolqshaya Rossijskaya Enczіklopediya, 1999. ISBN: 5-85270-265-X.

[17] Госстандарт России. *ГОСТ Р 50779.10-2000 Статистические методы. Вероятность и основы статистики. Термины и определения*. Москва, 2000.

[18] Госстандарт России. *ГОСТ Р 50779.11-2000 (ИСО 3534.2-93) Статистические методы. Статистическое управление качеством. Термины и определения (Переиздание)*. Москва, 2000.

[19] Госстандарт России. *ГОСТ Р ИСО 5725-1-2002 Точность (правильность и прецизионность) методов и результатов измерений*. Часть 1. Основные положения и определения. Москва, 2002.

[20] Е. В. Чимитова Б. Ю. Лемешко. «О выборе числе интервалов в критериях согласия типа Хи-квадрат». В: *Заводская лаборатория. Диагностика материалов* (2003).

[21] В. Р. Бараз. *Корреляционно-регрессионный анализ связи показателей коммерческой деятельности с использованием программы Excel*. Федеральное агентство по образованию. ГОУ ВПО «Уральский государственный технический университет — УПИ», 2005.

[22] Т. М. Сизова. *Статистика. Учебное пособие*. 2005.

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

[24] Т. А. Лёзина В. Л. Аббакумов. *Бизнес-анали информации. Статистические методы*. 2009. ISBN: 978-5-282-02918-5.

[25] О. Р. Никитин. *Статистические методы обработки параметров радиосигналов*. 2012.

[26] Д. Ф. Свааб. *Мы — это наш мозг. От матки до Альцгеймера*. Санкт-Петербург: Издательство Ивана Лимбаха, 2014. ISBN: 9785890592026.

[27] Сара Бослаф. *Статистика для всех*. 2015.

[28] А. М. Шихалёв. *Регрессионный анализ. Парная линейная регрессия*. Науч. отч. Казан. ун-т, 2015.

[29] Михаил Хальман. «Регрессионный анализ.» В: *Прикладной статистический анализ данных*. 2017.

[30] Н. В. Казанцева. *Математическое моделирование в программных пакетах Excel и MathCad : учеб.-метод. пособие*. 2018.

[31] Владимир Савельев. *Статистика и котики*. Русский. Москва: Издательство АСТ, 2018, с. 122. ISBN: 978-5-17-106143-2.

[32] Г. Шуленбаева. *almaty-apts-2019-1*. Под ред. К. А. Мурашев. 2019. URL: https://github.com/Kirill-Murashev/AI_for_valuers_R_source/tree/main/datasets/almaty_apts_2019_1.csv.

[33] Институт биоинформатики. *Основы статистики. Часть 1*. Русский. 2020. URL: https://stepik.org/course/76/info (дата обр. 07.10.2021).

[34] Институт биоинформатики. *Основы статистики. Часть 2*. Русский. 2020. URL: https://stepik.org/course/524/info (дата обр. 07.10.2021).

[35] Институт биоинформатики. *Основы статистики. Часть 3*. Русский. 2020. URL: https://stepik.org/course/2152/info (дата обр. 07.10.2021).

[36] К. А. Мурашев. *spba-flats-210928*. 2021-09-28.

[37] К. А. Мурашев. «Краткое введение в различия между часотным и байесовским подходом к вероятности в оценке стоимости». В: (2021-10-10).

[38] Computer Science Center. *Введение в математический анализ*. URL: https://stepik.org/course/95/info.

[39] Computer Science Center. *Ликбез по дискретной математике*. URL: https://stepik.org/course/91/info.

[40] Computer Science Center. *Линейная алгебра*. URL: https://stepik.org/course/2461/info.

[41] Computer Science Center. *Математическая статистика*. URL: https://stepik.org/course/326/info.

[42] Computer Science Center. *Математический анализ (часть 1)*. URL: https://stepik.org/course/716/info.

[43] Computer Science Center. *Математический анализ (часть 2)*. URL: https://stepik.org/course/711/info.

[44] Computer Science Center. *Основы теории графов*. URL: https://stepik.org/course/126/info.

[45] Computer Science Center. *Теория вероятностей*. URL: https://stepik.org/course/3089/info.

[46] FDFGroup. *Выборка. Типы выборок. Расчет ошибки выборки*. URL: https://fdfgroup.ru/poleznaya-informatsiya/stati/vyborka-tipy-vyborok-raschet-oshibki-vyborki/ (дата обр. 12.10.2021).

[47] Herbert A. Sturgess. «The Choice of a Class Intervals». В: *Journal of the American Statistical Association* 21.153 (), с. 65—66. URL: http://www2.esalq.usp.br/departamentos/lce/arquivos/aulas/2013/LCE0216/Sturges1926.pdf (дата обр. 15.10.2021).

[48] M. P. Wand. «Data-Based Choice of Histogram Bin Widt». В: *The American Statistician* 51.1 (), с. 59. DOI: 10.2307/2684697. URL: http://www.stat.cmu.edu/~rnugent/PCMI2016/papers/WandBinWidth.pdf (дата обр. 15.10.2021).

[49] Kirill A. Murashev. R. URL: https://github.com/Kirill-Murashev/AI_for_valuers_R_source.

[50] studfile.net. *Отличия между кластерной и стратифицированной выборкой*. URL: https://studfile.net/preview/1669802/page:5/.

[51] Wikipedia. *Байесовская вероятность*. URL: https://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B9%D0%B5%D1%81%D0%BE%D0%B2%D1%81%D0%BA%D0%B0%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C (дата обр. 09.09.2021).

[52] Wikipedia. *Частотная вероятность*. URL: https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%BD%D0%B0%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C (дата обр. 09.09.2021).

[53] Татьяна Кабанова. *Статистика для гуманитариев*. URL: https://stepik.org/course/83603/info.

[54] Игорь Клейнер. *Теория вероятностей в удовольствие (курс 1 основы)*. URL: https://stepik.org/course/102057/info.