МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ»

Факультет физико-математических и естественных наук

Кафедра прикладной информатики и теории вероятностей

«Допустить к защите»

Заведующий кафедрой
прикладной информатики
и теории вероятностей
д.т.н., профессор
_____ К.Е. Самуйлов

«____»_____ 20__г.

Выпускная квалификационная работа бакалавра

Направление 02.03.02 «Фундаментальная информатика и информационные технологии»

ТЕМА «Статистический анализ выборок малого объема»

Выполнил студент **Логинов Сергей Андреевич**

(Фамилия, имя, отчество)

Группа НФИбд-01-18

Студ. билет № 1032182520

Руководитель выпускной квалификационной работы

Хохлов А. А., к.ф.-м.н., доцент, доцент кафедры прикладной информатики и теории вероятностей

(Ф.И.О., степень, звание, должность)

(Подпись)

Автор

(Подпись)

г. Москва

2022 г.

Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»

АННОТАЦИЯ
выпускной квалификационной работы

Логинов Сергей Андреевич
(фамилия, имя, отчество)

на тему: «Статистический анализ выборок малого объема»

Данная выпускная квалификационная работа бакалавра имеет объем 75 страниц, содержит 101 формулу, 2 таблицы, 4 примера и 36 графических объектов. К работе прилагается 6 страниц приложений, дополняющих ее. При написании работы использовалось 32 источника информации. Работа состоит из списка сокращений, введения, основной части из трех разделов и заключения.

Целью работы является поиск и реализация статистических методов, подходящих для анализа малых и очень малых выборок. Рассматривается три больших группы методов: точечного и интервального оценивания, а также методы проверки различных статистических гипотез. Все выбранные методы реализованы программным образом.

В ходе выполнения работы был проведен анализ источников, содержащих информацию о статистических методах. Обоснование их использования содержится в каждом пункте, описывающем конкретный метод. Ограничения и рекомендации по использованию методов также приведены в работе. В результате проведенной работы были получены программные решения для задачи анализа малых выборок, которые могут использоваться для работы с различными подзадачами.

Автор ВКР _____ _____

(Подпись) (ФИО)

Оглавление

Список используемых сокращений

ГС – генеральная совокупность

СВ – случайная величина

МО – математическое ожидание

СКО – среднеквадратическое отклонение

КК – коэффициент корреляции

НСВ – непрерывная случайная величина

ФР – функция распределения

ММП – метод максимального правдоподобия

ММ – метод моментов

МНК – метод наименьших квадратов

ТО – точечное оценивание

ИО – интервальное оценивание

ДИ – доверительный интервал

ТИ – толерантный интервал

ЭФР – эмпирическая функция распределения

ДА – дисперсионный анализ

ОДА – однофакторный дисперсионный анализ

РА – регрессионный анализ

ОЛР – одномерная линейная регрессия

МЛР – многомерная линейная регрессия

КД – коэффициент детерминации

СКД – скорректированный коэффициент детерминации

НП – независимая переменная

ЗП – зависимая переменная

Введение

В данной работе рассматривается задача статистического анализа выборок малого объема. Здесь под выборкой понимается какой-либо набор числовых и качественных показателей, полученный в результате наблюдения, исследования или эксперимента. Исследователи изучают эти данные с целью получения информации, которую можно будет распространять на явления или величины, из которых были получены выборки. И здесь возникает следующая проблема: далеко не всегда исследователи имеют большие объемы данных для изучения. Эта проблема может иметь разные причины. Например, сбор большого объема какой-либо социальной статистики требует достаточно большого количества времени, а сбор достаточного объема наблюдений о каких-либо испытаниях, допустим, технических средств или систем, может требовать больших финансовых затрат. Также могут возникать проблемы с нехваткой людей, способных осуществить сбор большого количества информации, а также технического оснащения, которое помогает в данном вопросе. И именно по причине того, что далеко не каждый исследователь в своей работе будет иметь качественный и достаточно большой объем данных, вопрос способов проведения качественного анализа на относительно малых, малых или очень малых данных является актуальным. Более того, возможность работы с малыми объемами данных позволяет достаточно хорошо оптимизировать рабочие процессы, если говорить об исследованиях в компаниях. Это снизит рабочую нагрузку на персонал, позволит предотвратить лишние и, вероятно, большие траты различных ресурсов, как денежных, так и временных. Но в то же время необходимо, чтобы уменьшение необходимого объема для анализа и исследований не сказывалось на качестве результатов, как минимум имело допустимую погрешность. Исходя из этого очень важно знать и уметь применять методы, которые позволят проводить качественные исследования даже ограниченного набора данных. В большей степени речь идет о статистических методах, которые в своей основе имеют понятия из области теории вероятностей и математической статистики и следуют принципам данных дисциплин. Исследователь должен понимать, что он почти никогда не получит полностью достоверных результатов. Однако статистические методы позволяют делать очень точные выводы, которые, в свою очередь, подкреплены вычислениями, известными принципами и теоремами. Такие результаты будут являться

статистически значимыми. Существует огромное множество статистических методов. Однако, далеко не все могут быть применены к малым выборкам, как по причине чисто принципиальных соображений, так и из-за больших погрешностей и неточных результатов. Поэтому в данной работе собраны, реализованы и протестированы методы, которые возможно использовать при анализе малых объемов данных. Более того, некоторые из методов разработаны специально для работы с малыми выборками. Осталось численно обозначить объем выборок, которые считаются малыми. Проводя анализ источников по теме, отмечаем, что авторы не всегда сходятся в мнении о том, какие же выборки считают малыми. В большей части литературы объем малых выборок находится в пределах от 50 до 200 элементов. Выборки, размер которых не превышает 30 элементов, называют очень малыми. Поэтому назовем выборку малой, если ее объем не превышает 200 элементов, и очень малой в случае объема менее 30 элементов. Далее в работе рассматриваются методы, которые можно применять при работе с выборками озвученных объемов. При описании каждого метода будет сказано об ограничениях в размере выборок (если такое ограничение присутствует). Используемые методы протестированы и рекомендованы многими исследователями и авторами. В нашей работы мы также проверим на практике их работоспособность и пригодность для анализа малых выборок. Совокупность рассматриваемых в работе методов позволит нам качественно и быстро решать широкий список задач, многие из которых имеют большое практическое значение для исследований или бизнеса. Важно отметить, что в работе большое внимание уделяется работе с нормальным распределением. Это обосновано тем, что оно лучше всего изучено и имеет наибольшее практическое и жизненное распространение.

Поговорим об объекте и предмете исследования, а также о целях и задачах работы. Объектом исследования в данной работе можно считать статистические методы, а предметом – статистические методы для анализа малых выборок. Целью работы является определение методов, с помощью которых можно быстро, удобно и качественно решать задачи в работе с малыми выборками. Наравне с определением и пониманием данных методов, еще одной целью является практическая реализация выбранных методов.

Список задач, которые решаются в работе, а также методов, позволяющих решать поставленные задачи, приведен ниже:

Таблица 1. Задачи работы и методы их решения

Задача	Методы
Точечное оценивание параметров распределения	МНК, ММП, ММ
Интервальное оценивание	Критерии t, χ^2, построение толерантного интервала
Проверка гипотез о параметрах распределения	Критерии $z, t,$ Уилкоксона
Проверка гипотез о типе распределения	Критерии χ^2, Крамера-Мизеса-Смирнова, модифицированный χ^2, модифицированный Крамера-Мизеса-Смирнова, Шапиро-Уилка
Проверка зависимости количественной переменной от одной или нескольких качественных внутри одной малой выборки	Дисперсионный анализ
Анализ взаимосвязи и влияния одной или нескольких количественных переменных на одну количественную внутри одной малой выборки	Линейная регрессия

1. Исторический и литературный обзор проблемы статистического анализа малых выборок

Проблема анализа малых выборок впервые была затронута при исследовании задачи оценивания различных характеристик случайных величин и методах ее решения в трудах Р. Фишера и Стьюдента. Дальнейшее развитие проблема получила в работах многих исследователей: А. Н. Колмогорова, А. А. Петрова, И. Н. Володина и других [9-11, 17, 20]. Тем не менее, даже после публикаций данных работ, тема анализа малых выборок нуждалась в дополнительных исследованиях и в систематизации большого количества информации для решения большого количества практических задач. Основой сбора информации, проверки и отбора методов и погружения в задачу стало два литературных источника. В данных работах авторы попытались собрать воедино многие задачи и методы анализа малых выборок, основываясь на уже проведенных исследованиях. Примечательно, что между этими двумя работами имеется промежуток в несколько десятков лет, поэтому мы имеем возможность проследить изменения в подходе решения поставленных задач при увеличивающихся вычислительных мощностях и на фоне общего научно-технического прогресса.

Первый из этих источников – книга Гаскарова Д. В. И Шаповалова В. И. «Малая выборка» [2]. Несмотря на то, что с момента издания книги прошло уже более 40 лет, некоторая информация, которая в ней содержится, до сих пор актуальна и способна помочь в решении различных задач. В работе авторы не приводят четкое значение объема, начиная с которого выборку можно считать малой, упоминая о том, что многие исследователи оценивают это значение в 50 элементов, а некоторые – в 200. Авторы смогли представить методы решения различных задач в одной работе, которая получилась достаточно комплексной и разносторонней. В ней изучаются задачи оценивания закона распределения и моментов случайной величины, учитывая как точечные, так и интервальные оценки. Также исследуется задача проверки статистических гипотез, например, о типе и параметрах распределения или однородности распределений. Ближе к концу работы затрагивается задача проверки статистических зависимостей. Все эти задачи исследуются в условиях малых выборок. Методы, предложенные авторами для решения этих задач, также нацелены на работу с ограниченным количеством

данных, некоторые из которых, в немного измененном виде, приведены в работе. Например, критерий Уилкоксона для проверки гипотез об однородности распределений случайных величин. Важно отметить, что работа создавалась достаточно давно и авторы имели соответствующие времени возможности для вычислений и экспериментов. Поэтому в наши дни данная работа не выглядит лучшим вариантом для поиска информации о проблемы. Хотя нельзя не отметить, что она содержит большое количество полезной информации и в свое время она могла являться отличным источником. К сожалению, сейчас эта информация выглядит полезной больше в теоретическом, чем в практическом смысле. За 40 лет изменилось очень многое и это стоит понимать. Поэтому необходимо было найти достаточно современный, но такой же полезный источник как теоретической, так и практической информации, речь о котором идет далее.

Вторым и самым главным источником информации о проблеме малых выборок стала книга Б. И. Сухорученкова «Анализ малой выборки [16]. Прикладные статистические методы». Из самого названия мы можем понять, что автор делает упор именно на практических аспектах задачи и методов ее решения. В целом, данная книга отчасти напоминает вышеописанную, но она была написана на 30 лет позже, поэтому ощущается и, на мой взгляд, является более современной версией источника информации об анализе малых выборок. В ней автор неоднократно упоминает о компьютерных и программных вычислениях, во многих моментах уделяет внимание удобству и легкости программной реализации методов. Малыми автор называет выборки, объем которых не превышает 30 элементов, а также упоминает о том, что многие исследователи определяют верхнюю границу объема малой выборки как 50 или 200 элементов. Б. И. Сухорученков в своей работе рассматривает более широкий список тем, чем авторы прошлой книги. В работу включены различные методы точечного оценивания как моментов, так и параметров распределения. Приведены различные методы построения доверительных интервалов для моментов и параметров распределения. Большое внимание уделено теме проверки статистических гипотез, в которой автор приводит методы проверки гипотез о параметрах и типах распределения. В книге имеется очень понятная и простая в реализации версия регрессионного анализа. Все методы, приведенные автором, проверялись в условиях малых выборок, а результаты схожих методов

сравнивались на предмет отличий, которые объяснялись при выявлении. В работе присутствует достаточно понятная и современная версия визуальных дополнений в виде графиков, четко демонстрирующих принципы работы или результаты того или иного метода. Важно отметить, что абсолютно каждый метод в работе обязательно сопровождается примером, что очень сильно облегчает понимание его работы. В целом можно сказать, что данная книга позволит читателю достаточно сильно погрузиться в тему статистического анализа малых выборок, получить необходимую информацию как теоретическую, так и практическую, а также повысить эффективность своих исследований малого объема данных, познакомившись с задачами и методами их решения. Подача материала и язык повествования автора сохраняет тонкую грань между практическими рекомендациями и научным трудом, что воспринимается исключительно положительно. Источник из прошлого пункта напротив, больше воспринимался как научная работа, которую не так легко изучать. Но для полного понимания и более простого изучения и усвоения материала книга все же требует знаний и подготовки в области математики и математической статистики. Некоторые пробелы в знаниях и понимании помог заполнить следующий источник, содержащий общую, но обширную информацию по математической статистике.

Основой общих статистических сведений, а также помощником в понимании некоторых аспектов математической статистики стала книга «Теория вероятностей и математическая статистика» авторов Лебедева А. В. и Фадеевой Л. Н. [8]. Данная работа помогла восстановить в памяти некоторые понятия и формулы, а также более подробно изучить некоторые области математической статистики. Книга охватывает широкий список базовых тем изучаемой области, достаточно удобно и понятно написана. Важно отметить, что книга имеет хорошие практические примеры почти по каждой теме, а различная направленность этих примеров (экономика, финансы) повышает интерес и упрощает понимание. Работа оставила исключительно положительные эмоции, помогла вспомнить некоторые моменты из области математической статистики, а какие-то вовсе помогла понять.

Также во время написания работы использовалась книга о дисперсионном анализе В. А. Юденкова «Дисперсионный анализ» [18]. Данная работа помогла улучшить понимание принципов работы данного метода, его назначения и пользы.

Важно отметить, что книга имеет практическую направленность, в ней четко, ясно и достаточно кратко излагаются теоретические основы, принципы, практические советы и подтверждения исключительной полезности данного метода.

Остальные источники, так или иначе задействованные при написании работы, отмечены ссылками в тексте. Номер каждой ссылке соответствует источнику из соответствующего раздела работы.

2. Статистические методы для анализа малых выборок

Во втором разделе мы рассмотрим различные статистические методы анализа данных, которые можно применять при работе с малыми выборками. В первом разделе было отмечено, что в данной работе малыми мы считаем выборки, объем которых суммарно (то есть по всем переменным) не превышает 200 значений. Очень малыми выборками считаем те, размер которых не превышает 30 значений. Все методы, которые приведены ниже, позволяют проводить анализ малых выборок. Некоторые имеют ограничения при работе с очень малыми выборками, которые также будут озвучены. Наиболее подходящие методы для очень малых выборок также включены в работу. Методы, которые мы рассмотрим, позволят решать широкий спектр задач, а также позволят получить качественную статистическую информацию, которую можно использовать как при дальнейших исследованиях с помощью более сложных методов, так и при использовании совокупности базовых методов. Задачи и статистические методы их решения более подробно описаны далее.

Рассмотрим задачу статистического оценивания параметров ГС по малой выборке. Она направлена на получение информации о параметрах ГС по малой выборке. Решив эту задачу, мы, с некоторой погрешностью, получим информацию о ГС, не работая с ней напрямую. Перейдем к более подробному определению статистического оценивания.

Пусть СВ, плотность которой зависит от некоторого параметра a, является для нас ГС. Тогда некоторое количество реализаций данной СВ – выборка из ГС. По данной выборке мы хотим узнать неизвестный параметр распределения нашей ГС. Но, так как мы рассматриваем малые выборки, мы можем только оценить этот параметр, а не узнать его с полной уверенностью. Ситуацию оценки параметра ГС по выборке будем называть статистическим оцениванием, а оценки – статистическими оценками, которые бывают двух типов: точечные и интервальные. О методах нахождения точечных оценок (ТО) и интервальных оценок (ИО) неизвестных параметров ГС по малым выборкам поговорим далее.

1. Методы точечного оценивания

Точечное оценивание предполагает получение оценок значений параметров ГС в виде конкретных числовых значений. Работа с ТО на самом деле имеет как

положительные, так и отрицательные стороны. Плюсом является то, что мы получаем конкретное значение, минусом – погрешность ТО для малых выборок несколько больше, чем погрешность ИО. Рассмотрим классическую формулировку точечного оценивания неизвестных параметров ГС. Пусть СВ X все так же зависит от параметра a. Точечная оценка $\overline{a}$, определяемая по элементам выборки, также является СВ из условия случайности получения какой-либо конкретной выборки. Поскольку производится оценивание, тем более по малому набору данных, наша оценка может отличаться от параметра в ГС. Данное отличие выражается математическим ожиданием и дисперсией нашей оценки: $M_{\overline{a}}, \sigma_{\overline{a}}^2$.

Теоретически, если возможно получить все различные выборки из ГС и по ним оценить неизвестный параметр a, то можно построить плотность распределения оценок как плотность распределения СВ $\overline{a}$. На основе построенной плотности можно определить несмещенную оценку неизвестного параметра и погрешность этой оценки. К сожалению, на практике почти никогда не удается реализовать вычисления озвученным выше методом. Получение всех возможных выборок, оценок и построение плотности вероятности оценки требует больших ресурсных и временных затрат, а иногда оно и вовсе невозможно. Чтобы в данной ситуации все-таки произвести оценивание, необходимо использование специальных статистических методов для оценки неизвестных параметров распределения. В данной работе предлагается использовать следующие методы: максимального правдоподобия, моментов, наименьших квадратов. Во время подготовки работы данные методы были протестированы для работы в наших условиях (с выборками объема до 200 объектов) и показали вполне адекватные и пригодные для дальнейших исследований результаты для СВ с нормальным, экспоненциальным и равномерным типами распределения. Более того, метод наименьших квадратов стал основой для применения еще одного метода исследования – регрессионного анализа. Подробное описание методов, примеры использования и сравнение результатов даются далее.

Первым методом ТО неизвестных параметров ГС рассмотрим метод максимального правдоподобия. Данный метод разработан Р. Фишером [12] и используется для точечного оценивания неизвестных параметров распределения СВ по выборке ее реализаций. В основе данного метода лежит функция правдоподобия (ФП). Для использования ММП требуется знать тип распределения СВ, параметры

распределения при этом знать не обязательно. Далее кратко приведем уже известные зависимости для оценки по данному методу параметров нормального, равномерного и экспоненциального распределения. В [7, 12, 14] доказывается, что при использовании ММП в общем случае получаются асимптотически эффективные и несмещенные оценки. Зависимости для нормальной СВ:

$$\bar{M}_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad 2.1.1$$

$$\bar{D}_x = \frac{1}{n}\sum_{i=1}^{n} (x_i - \breve{M}_x)^2 \qquad 2.1.2$$

Нужно уточнить, что в результате вычислений получена смещенная оценка для дисперсии. Исходя из информации в [7, 14, 15, 17], для получения несмещенной оценки дисперсии множитель $\frac{1}{n}$ необходимо заменить на $\frac{1}{n-1}$. Для определения точности оценок используется ковариационная матрица, по которой находятся погрешности наших оценок:

$$\sigma^2_{\bar{M}_x} = \frac{\bar{D}_x}{n} \qquad 2.1.3$$

$$\sigma^2_{\bar{D}_x} = \frac{2\bar{D}_x{}^2}{n} \qquad 2.1.4$$

Дальнейшая задача будет заключаться в использовании данных зависимостей для получения информации по малой выборке (оценка параметров) и проверке качества оценивания. Для понимания отличий в сути и работе рассмотренных методов, а именно ММП и ММ, мы рассмотрим их применение в случае еще двух типов распределения – равномерного и экспоненциального. В случае нормального распределения методы отличаются незначительно.

Для оценки равномерного распределения по ММП известны следующие зависимости:

$$\bar{A} = x_{min} \qquad 2.1.5$$

$$\bar{B} = x_{max} \qquad 2.1.6$$

К сожалению, погрешности оценок для данного распределения вычислить не удается. Это значительный минус.

Пример 1. Имеем малую выборку $x = \{12, 20, 24, 23, 16, 19, 20, 35, 10, 45\}$. По зависимостям получаем ТО границ ГС с равномерным распределением:

$$\bar{A} = 10$$

$$\bar{B} = 45$$

График предполагаемой плотности ГС:

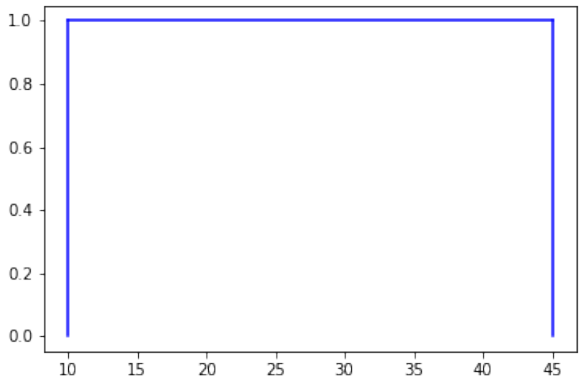

Рисунок 1. Предполагаемое распределение в примере 1

В случае, когда ГС имеет (или предположительно имеет) экспоненциальное распределение, ставится задача точечной оценки параметра интенсивности происходящих событий λ. Данную оценку будем обозначать как $\bar{\lambda}$. Допустим, что проводится N испытаний в течение некоторого временного промежутка T. Тогда время, до наступления события в каждом из испытаний будем обозначать как множество $\{t_1, t_2, \ldots, t_m\}$. Мы также можем определить количество испытаний n, в которых событие не произошло за заданный промежуток времени: $n = N - m$. Сумму наблюдений обозначм как S. Тогда используется зависимость:

$$\bar{\lambda} = \frac{m}{S} \quad 2.1.7$$

Для вычисления погрешности данной оценки получаем формулу:

$$\sigma^2_{\bar{\lambda}} = \frac{\bar{\lambda}^2}{m} \quad 2.1.8$$

Рассмотрим применение полученных зависимостей на примере.

Пример 2. Пусть в результате 10 испытаний с периодом, равным единице, исследуемое событие появилось четыре раза со значениями:

$$t_1 = 0.96, t_2 = 0.8, t_3 = 0.65, t_4 = 0.7$$

Найдем общее время испытаний:

$$S = 0.96 + 0.8 + 0.65 + 0.7 + 6 * 1 = 9.11$$

ТО параметра экспоненциального распределения λ и ее погрешность:

$$\bar{\lambda} = \frac{4}{9.11} = 0.4391$$

$$\sigma_{\bar{\lambda}} = \frac{0.4391^2}{4} = 0.0482$$

График предполагаемой плотности ГС:

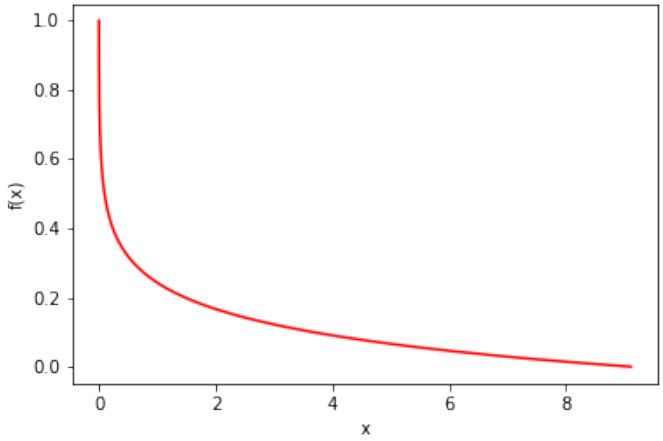

Рисунок 2. Предполагаемое распределение в примере 2

Вторым методом, который активно применяется для точечного оценивания параметров распределения, является метод моментов (ММ). Далее рассмотрим суть и общий вид данного метода, а также его использование для ТО параметров нормального, равномерного и экспоненциального распределения (по аналогии с ММП). Данный метод был создан К. Пирсоном [12] и основывается [12, 13, 15] на получении зависимостей моментов распределения. На практике данный метод чаще всего используется для оценки от одного, до двух (реже для трех) неизвестных параметров, что возвращает нас к работе с МО и дисперсией. Далее приведены известные зависимости для оценки параметров нормального, равномерного и экспоненциального распределения.

$$\bar{M}_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad 2.1.9$$

$$\bar{D}_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{M}_x)^2 \qquad 2.1.10$$

Формулы для ТО МО совпадает с ММП. Формула для ТО дисперсии по ММ приводит к несмещенной оценке, при использовании ММП требуется коррекция знаменателя.

Получается, что принципиальной разницы для оценивания параметров нормального распределения между ММП и ММ нет, за исключением вышеупомянутого нюанса с дисперсией. Поэтому рассматривать какие-либо примеры в данном пункте не будем. Тем не менее, нельзя сказать, что методы взаимозаменяемые или одинаковые. Результаты работы с нормальным распределением не следует распространять на остальные типы. Для поиска отличий рассмотрим ТО равномерного и экспоненциального распределения по ММ и сравним результаты с ММП.

В случае равномерного распределения ГС ставится задача оценки параметров a и b – границ распределения. Для равномерного распределения есть «свои» формулы вычисления МО и дисперсии, которые и выступят в роли моментов:

$$M_x = \frac{b-a}{2} \quad 2.1.11$$

$$D_x = \frac{(b-a)^2}{12} \quad 2.1.12$$

Оценка границ производится по зависимостям:

$$\bar{a} = \bar{M}_x - \sqrt{3\bar{D}_x} \quad 2.1.13$$

$$\bar{b} = \bar{M}_x + \sqrt{3\bar{D}_x} \quad 2.1.14$$

Погрешности оценок границ:

$$\sigma_{\bar{a}} = \sigma_{\bar{b}} = \sigma_{\bar{M}_x} + \frac{3\sigma_{\bar{D}_x}^2}{4\bar{D}_x} \quad 2.1.15$$

Пример 3. Рассмотрим использование полученных по ММ зависимостей в условиях, аналогичных *примеру 1*. Проверим, есть ли отличия в результатах. Используем ту же самую выборку. Оценки границ в данном случае определяем через выборочные МО и дисперсию:

$$\bar{M}_x = 22.4$$

$$\bar{D}_x = 110.93$$

$$\bar{A} = 22.4 - \sqrt{3 * 110.93} = 4.1572$$

$$\bar{B} = 22.4 + \sqrt{3 * 110.93} = 40.6428$$

График предполагаемой плотности ГС:

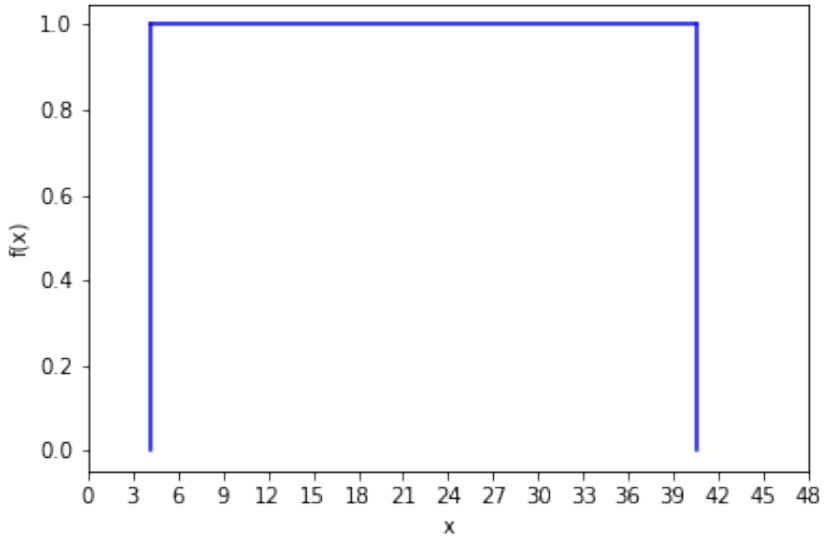

Рисунок 3. Предполагаемая плотность в примере 3

Сравнивая результаты оценки границ по ММП и ММ видим, что оценки по ММП получаются полезнее, потому что они включают в себя все значения из выборки. В случае ММ оценка правой границы не сходится с имеющейся выборкой, в которой имеется большее значение, чем полученная оценка. Поэтому будем использовать оценивание по ММП в дальнейшем.

По аналогии с прошлым пунктом, то есть для сравнения результатов использования ММП и ММ для ТО параметров, рассмотрим задачу точечного оценивания параметров экспоненциального распределения по ММ. У экспоненциального распределения есть единственный параметр λ, характеризующий интенсивность событий. Если выборка извлечена из ГС с экспоненциальным распределением (или с предположительно экспоненциальным распределением), то можно использовать следующие специальные формулы для вычисления МО и дисперсии:

$$M_x = \frac{1}{\lambda} \quad 2.1.16$$

$$D_x = \frac{1}{\lambda^2} \quad 2.1.17$$

Чтобы оценить неизвестный параметр λ по ММ используются следующие зависимости по МО и дисперсии:

$$\bar{\lambda} = \frac{1}{\bar{M}_x}, \sigma_{\bar{\lambda}}^2 = \frac{\sigma_{\bar{M}_x}^2}{\bar{M}_x{}^4} \quad 2.1.18$$

$$\bar{\lambda} = \frac{1}{\sqrt{\bar{D}_x}}, \sigma_{\bar{\lambda}}^2 = \frac{\sigma_{\bar{D}_x}^2}{4\bar{D}_x^{\,3}} \qquad 2.1.19$$

На практике полезно использование обоих вариантов оценивания для выбора наиболее точного. Рассмотрим оценивание неизвестного параметра экспоненциального распределения по ММ на примере. Также сравним полученные результаты с ТО, полученной по ММП.

Пример 4. Рассматриваем ту же выборку, что и в *примере 2*. Найдем выборочные МО и дисперсию и ТО параметра λ двумя способами:

$$\bar{M}_x = 0.911$$

$$\bar{D}_x = 0.1943$$

Оценка и погрешность через МО:

$$\bar{\lambda} = \frac{1}{0.911} = 1.0977$$

$$\sigma_\lambda = 0.0028$$

Оценка и погрешность через дисперсию:

$$\bar{\lambda} = \frac{1}{\sqrt{0.1943}} = 7.4627$$

$$\sigma_{\bar{\lambda}} = 2.573$$

Намного более точную оценку получили через МО, ее и будем использовать для построения графика плотности предполагаемой ГС:

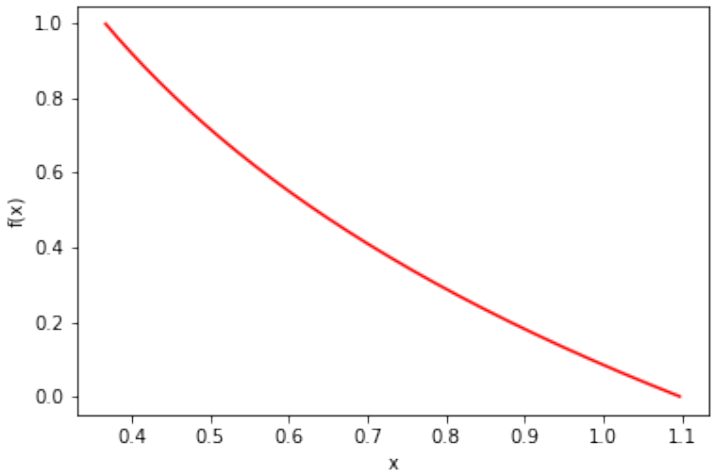

Рисунок 4. Предполагаемое распределение в примере 4

Сравнивая результаты оценки по ММП и ММ видим, что через ММ мы получаем меньшее значение дисперсии оценки, следовательно, она точнее. Далее будем использовать именно этот метод.

Последний метод, который мы рассмотрим для решения задачи точечного оценивания неизвестных параметров распределения ГС по малой выборке – метод наименьших квадратов. Данный метод создан К. Гауссом и А. Лежандром **[8]**. Имеет широкий спектр применения, но нас в данном случае интересует ТО параметров нормального распределения, а также методика, используемая в регрессионном анализе. Плюсом данного метода является то, что он не требует знания типа распределения ГС. Суть метода заключается в том, что производится минимизация квадратов различных отклонений. Но, так как полученные по МНК зависимости в случае нормального распределения не отличаются от уже известных 2.1.9, 2.1.10, просто обозначим их без более подробного рассмотрения

В отличие от ММП, МНК сразу дает несмещенную оценку дисперсии, необходимости в корректировке нет. Погрешность оценки МО аналогична 2.1.3.

В регрессионном анализе используется вариация МНК, при которой лучшая модель выбирается на основе минимизации отклонений значений имеющихся моделей от истинных значений [16]. Более подробно рассмотрим эту методику в соответствующем пункте работы.

2. Методы интервального оценивания

Если до этого мы рассматривали методы, которые позволяют получить оценки неизвестных параметров в виде числовых значений, то в данной главе мы рассмотрим способы оценивания параметров в виде интервалов, то есть интервальное оценивание (ИО). ИО представляет собой построение интервалов, в которых вероятнее всего находятся истинные значения неизвестных параметров в ГС. При работе с малыми выборками ИО считается более эффективным, чем ТО. К тому же на основе ИО можно будет найти (как минимум для нормального распределения) ТО. Методы из прошлого пункта не обладают свойствами для проведения интервального оценивания, поэтому далее в работе будут предложены специальные методы для ИО. Но сначала стоит рассмотреть понятие доверительного интервала. В общем смысле доверительный интервал (ДИ) – такой интервал, в которой с заданной доверительной вероятностью P попадет какое-либо значение

(например, оценка неизвестного параметра). Обычно используется вероятность 95% или 99%. ДИ дают широкие возможности для анализа различных величин, особенно при проверке уровня значимости. Далее рассмотрим статистические методы для интервального оценивания МО и дисперсии, а также произведем оценивание толерантного интервала СВ, который достаточно редко (относительно ранее перечисленных методов анализа) используется в исследованиях, хотя может обеспечить очень полезную информацию.

Рассмотрим задачу интервального оценивания МО нормальной СВ. Доверительный интервал для МО будем строится по распределению Стьюдента [12]. Данный метод используется при нормальном распределении ГС. По выборке необходимо найти точечные оценку и погрешность МО по уже описанным методам. Далее имеет место следующая статистика *t*:

$$t = \frac{\overline{M}_x - M_x}{\sigma_{\overline{M}_x}} \quad 2.2.1$$

Границы ДИ МО оцениваются в соответствии с квантилями распределения Стьюдента и доверительной вероятностью *P*:

$$\overline{M}_{x_L} = \overline{M}_x - q\sigma_{\overline{M}_x} \quad 2.2.2$$

$$\overline{M}_{x_R} = \overline{M}_x + q\sigma_{\overline{M}_x}, \quad 2.2.3$$

где q – квантиль распределения Стьюдента при доверительной вероятности *P* и $k = n - 1$ степенях свободы (*n* – объем выборки), который можно найти в специальных таблицах.

Выше был предложен метод ИО для МО нормальной СВ. Будет логично рассмотреть способы оценки ИО дисперсии нормальной СВ, чтобы иметь полную картину того, как можно построить ДИ для параметров нормального распределения по малой выборке. Аналогично прошлому пункту, рассмотрим малую выборку из нормальной ГС. По данной выборке нам необходимо получить ТО для дисперсии с помощью уже известных методов. Далее наши действия аналогичны тем, что совершались при ИО МО, но необходимо заменить статистику и используемое в методе распределение. В данном случае мы будем использовать распределение «хи-квадрат» и статистику вида:

$$\chi^2 = \frac{n\overline{D}_x}{D_x} \quad 2.2.4$$

В данном методе необходимо задать две доверительные вероятности p для левой и правой границ ДИ соответственно. Используем следующие зависимости:

$$P(D \geq D_L) = p_R \quad 2.2.5$$

$$P(D \leq D_R) = 1 - p_L, \quad 2.2.6$$

где D_R, D_L – правая и левая граница ДИ для дисперсии. Для нахождения левых частей уравнений необходимо воспользоваться таблицей (или проинтегрировать плотность) χ^2 распределения и при числе степеней свободы $k = n - 1$ [15] и соответствующих доверительных вероятностях для границ найти квантили q_L, q_R. Найдя квантили, мы можем построить оценки границ ДИ для дисперсии по следующим формулам:

$$\overline{D_{x_L}} = \frac{n\overline{D_x}}{q_R} \quad 2.2.7$$

$$\overline{D_{x_R}} = \frac{n\overline{D_x}}{q_L} \quad 2.2.8$$

Последней задачей блока методов интервального оценивания предлагаю выбрать построение толерантных интервалов. Откровенно говоря, данная задача на практике встречается реже, чем все, что мы рассмотрели выше. Однако это нисколько не снижает полезности данного анализа. Толерантным интервалом (ТИ) называется множество значений, в которое доля, большая или равная g, реализаций СВ попадет с заданной доверительной вероятностью p (по большей части множество представляет собой отрезок $[a, b]$). Иными словами, если мы найдем ТИ, то мы будем понимать, какие значения в принципе могут быть получены из нашей ГС при различных доверительных вероятностях. В некотором смысле мы получаем своеобразную область допустимых значений, зависящую от вероятности и вычисляемую приближенно.

Для ГС с известной плотностью можно использовать классическое определение вероятности и вычислять оценки границ ТИ по интегралам плотности до каждой из границ:

$$\int_{-\infty}^{a} f(x)dx = 1 - p_L \quad 2.2.9$$

$$\int_{-\infty}^{b} f(x)dx = p_R \quad 2.2.10$$

Для доверительных вероятностей границ действует уже знакомое правило:

$$p_L + p_R = 1 + p \quad 2.2.11$$

Если же плотность распределения ГС неизвестна, то нам необходимо найти ТО МО и СКО. Далее, согласно [15], имеют место следующие зависимости для оценок границ ТИ:

$$a = \bar{M}_x - k\bar{\sigma}_x \quad 2.2.12$$

$$b = \bar{M}_x + k\bar{\sigma}_x, \quad 2.2.13$$

$$\text{где } k = q_g\left(1 + \frac{q_p}{\sqrt{2n}} + \frac{5q_p^2 + 10}{12n}\right) [\mathbf{6}] \quad 2.2.14$$

Мы видим два вида квантилей, один в зависимости от доли g, другой от доверительной вероятности p. В обоих случаях квантили определяются по нормальному распределению и соответствующим таблицам при $p = 0{,}5(1 + g)$ в случае квантили от доли g, а в случае квантили от доверительной вероятности p – по самому значению p.

3. Методы проверки статистических гипотез и комплексные методы

Ранее мы рассмотрели статистические методы точечного и интервального оценивания параметров распределения СВ. Они позволяют получить по малой выборке базовую информацию о ГС, полезную и необходимую для дальнейших исследований уже другими методами. Мы обязательно рассмотрим методы более углубленного исследования чуть позже. На данном этапе предлагаю рассмотреть последний вопрос для получения базовой информации о ГС, а именно задачу проверки статистических гипотез о параметрах и типе распределения. Выше неоднократно упоминалось, что в данной работе рассматриваются СВ с нормальным распределением. Экспоненциальное и равномерное распределение были приведены в качестве примеров отличия принципов работы и результатов оценки по методам точечного оценивания. Поэтому в дальнейшем мы возвращаемся к использованию нормального распределения, а значит нужно привести более точную формулировку задач в рамках данной работы. В следующих пунктах мы рассмотрим последнюю группу методов для получения базовой информации о нормальной ГС по выборке, а именно проверку статистических гипотез о параметрах нормального распределения и проверку гипотез о нормальности распределения (также можно назвать это проверкой распределения на нормальность). Как и выше, мы будем использовать методы, уместные при работе с малыми выборками, которые в нашей работе

ограничиваются объемом в двести объектов. Данные методы приведены и протестированы в различных источниках, что только подтверждает их уместность. Более того, гипотезы могут являться более надежным источником информации, так как сразу показывают вероятность того, что мы не правы. Эта вероятность выражена численно и проста к пониманию, в отличие от дисперсий ТО. Также гипотезы позволяют подбирать наиболее вероятные параметры методом проб и ошибок, а в случае ТО мы имеем лишь единственное значение с не очень понятным уровнем «погрешности» [24, 25].

Предлагаю начать с проверки гипотез МО нормального распределения. Как известно, СВ с нормальным типом распределения имеет два параметра – МО и дисперсия. На практике в большей части исследований данные параметры неизвестны, но имеется выборка, по которой можно проверить некоторые статистические гипотезы о значениях этого параметра. Важно понимать, что мы работаем с гипотезами, которые не дают однозначного ответа в виде значения параметра, а лишь позволяют оценить (в численном вероятностном или процентном формате) уверенность в том, что наша гипотеза верна. Степень уверенности можно регулировать, задавая уровень статистической значимости при проверке гипотез. Данный показатель в чем-то схож с доверительной вероятностью, но используется и интерпретируется несколько иначе. Данный показатель тесно связан с понятием статистических ошибок. Глобально, есть два вида статистических ошибок: ошибки первого и второго рода. Кратко поговорим и принципе статистических гипотез и об этих ошибках.

Исследования методами проверки статистических гипотез заключаются в выдвижении двух гипотез – нулевой и альтернативной. При проверке гипотез нулевой гипотезой является какое-то предположение об исследуемом объекте, которое мы хотим проверить. Альтернативной гипотезой в таком случае будет предположение, в корне обратное содержащемуся в нулевой гипотезе. Например, если при исследовании мы хотим проверить гипотезу о том, что МО ГС в точности равно какому-то значению M_0, то нулевая гипотеза будет иметь вид $H_0: M_x = M_0$. Альтернативной гипотезой в данном случае будет $H_1: M_x \neq M_0$.

В работе с методами проверки статистических гипотез, очевидно, могут возникать ошибки. В основном они регулируются уровнем значимости, а также

корректностью информации, содержащейся в выборке. Поэтому в работе с выборками малого объема необходимо контролировать и проверять их информативность, анализировать различные проблемы, например выбросы, а также использовать наиболее подходящие и точные методы. Так или иначе, при проверке гипотез встречается два типа ошибок:

1. Ошибка первого рода допускается тогда, когда в ходе исследования принимается альтернативная гипотеза, хотя на самом деле верной была нулевая. Данная ошибка наиболее неприятна при прямом типе исследования.
2. Ошибка второго рода допускается в случаях принятия нулевой гипотезы при верности альтернативной. Данная ошибка наиболее неприятна при обратном типе исследования.

В некотором смысле, уровень значимости, который задается при проверке, также означает вероятность совершения ошибки первого рода. После того, как фундаментальные понятия о теме статистических гипотез и их проверке рассмотрены, мы можем перейти к конкретным методам проверки этих гипотез.

К рассмотрению предлагается ГС X, из которой получена выборка. Предполагается, что распределение ГС нормальное. Дальнейшие наши действия зависят от информации о дисперсии ГС. Случай, когда дисперсия известна, на практике встречается достаточно редко, поэтому нет особого смысла рассматривать его. Наиболее часто дисперсия ГС, как и МО, неизвестно. Поэтому для начала нам необходимо найти ТО МО и дисперсии по выборке с использованием уже известных методов, а также погрешности данных оценок. Из ТО дисперсии получим ТО СКО. Для проверки нулевой гипотезы о МО вида $H_0: M_x = M_0$ в **[1]** предлагается использовать z-статистику, частично модифицированную под проверку гипотезы:

$$z = \frac{\bar{M}_x - M_0}{\sigma_{\bar{M}_x}} \qquad 2.3.1$$

Данный показатель при нормальной ГС имеет распределение Стьюдента с $k = n - 1$ степенями свободы. Для того, чтобы сделать вывод относительно выдвинутой гипотезы, необходимо найти критическое значение z_{cr} и сравнить его с полученным значением показателя по выборке. Критическое значение в данном случае будем искать по таблице квантилей распределения Стьюдента, в котором доверительная вероятность определяется как $p = 1 - \alpha$, где α – уровень значимости

при проверке гипотезы, а количество степеней свободы определяется по формуле, данной выше. На основе полученных значений используем решающее правило классического вида:

$$\text{Принимаем } H_0: M_x = M_0, \text{если } |z| \leq z_{cr}$$

$$\text{Отклоняем } H_0: M_x = M_0, \text{если } |z| > z_{cr}$$

Далее рассмотрим один из методов проверки гипотез по двум выборкам, которые зачастую используются при каких-либо сравнениях двух нормальных ГС на схожесть по имеющимся выборкам из них. Подобные исследования позволяют ответить на более глобальные вопросы. Например, о статистических различиях между двумя группами в результате какого-то эксперимента. На рассмотрение и дальнейшее использование мы примем метода проверки гипотез о МО по двум малым выборкам. Здесь и далее будем рассматривать две нормальные СВ X и Y (то есть две разные ГС) и малые выборки из них.

Пусть МО и дисперсии ГС нам неизвестны и требуют точечной оценки по соответствующим методам. Выдвигаем нулевую гипотезу о том, что на самом деле МО наших ГС совпадают, то есть $H_0: M_X = M_Y$ и задаем уровень значимости α. Здесь мы также будем использовать уже знакомую нам статистику *t*, только, опять же, несколько модифицированную под нашу задачу:

$$t = \frac{|\bar{M}_X - \bar{M}_Y|}{\sqrt{\frac{\bar{\sigma}_X^2}{n_1} + \frac{\bar{\sigma}_Y^2}{n_2}}}, \quad 2.3.2$$

где знаменатель отражает погрешность оценки разности МО в числителе. По информации из [14, 15], наша статистика имеет уже привычное распределение Стьюдента с $k = n_1 + n_2 - 2$ степенями свободы. После нахождения критического значения по распределению Стьюдента, формируем классическое решающее правило для нашей задачи:

$$\text{Принимаем } H_0: M_X = M_Y, \text{если } |t| \leq t_{cr}$$

$$\text{Отклоняем } H_0: M_X = M_Y, \text{если } |t| > t_{cr}$$

Далее рассмотрим метод, о котором в работе уже были упоминания – метод проверки статистических гипотез об однородности распределений двух ГС. Примечательно, что нам не нужно достоверно знать типы распределения ГС, поэтому мы выдвигаем гипотезу в общем виде, без подробностей: $H_0: f(X) = f(Y)$.

В [2, 16] утверждается, что лучшим методом проверки является критерий Уилкоксона. Данный критерий основывается на методике ранжирования имеющихся выборок. Также во многих источниках отмечается, что при работе с малыми выборками принципы ранжирования очень актуальны, так как они снижают степень искажения данных. В данном критерии предлагается сформировать одну общую выборку из имеющихся двух, в которой все значения отсортированы по возрастанию и сохранена информация о принадлежности каждого значения к одной из двух исходных выборок. Вместе с этим формируется вектор рангов, значения которого соответствуют номерам элементов общей выборки. Рассмотрим две выборки *x* и *y* из ГС *X* и *Y*. Размер общей выборки равен сумме размеров выборок *x* и *y*. Далее отсортируем общую выборку по возрастанию и заменим ее элементы следующим образом: если значение получено из выборки *x*, тогда заменяем его на *x*, иначе – на *y*. В итоге мы получим нечто похожее на:

$$r = \{1, 2, 3, 4, \dots, n\}, \text{где } n = n_x + n_y \quad 2.3.3$$

$$g_{x,y} = \{x, y, y, x, \dots, y\} \quad 2.3.4$$

Сопоставляя вектор рангов и общую выборку, мы можем определить ранги для элементов каждой из двух начальных выборок. И уже после этих операций мы можем перейти к вычислению показателя Уилкоксона на основе следующих зависимостей:

$$\begin{cases} U_x = \sum_{i=1}^{n_x} r_x \\ U_y = \sum_{j=1}^{n_y} r_y \end{cases} \quad 2.3.5$$

В источниках [2, 15] доказывается, что распределение данных показателей стремится к нормальному при выполнении следующих условий:

$$\begin{cases} n_x + n_y \geq 20 \\ n_x \geq 5, n_y \geq 5 \end{cases} \quad 2.3.6$$

Также приводятся зависимости для вычисления МО и дисперсии данных показателей. И если МО могут отличаться, то дисперсия у U_x и U_y общая, а, следовательно, и равная. Мы можем вычислить МО и дисперсию следующим образом:

$$M_{U_x} = 0{,}5 n_x (n_x + n_y + 1) \quad 2.3.7$$

$$M_{U_y} = 0{,}5n_y(n_x+n_y+1) \quad 2.3.8$$

$$\sigma_U^2 = \frac{1}{12}n_x n_y(n_x+n_y+1) \quad 2.3.9$$

После данных вычислений мы можем построить доверительные интервалы для U_x и U_y. Для заданного уровня значимости α мы находим квантиль q по нормальному распределению и $p = 1 - \frac{\alpha}{2}$ и используем следующие формулы:

$$\begin{cases} U_{x_L} = M_{U_x} - q\sigma_U \\ U_{x_R} = M_{U_x} + q\sigma_U \end{cases} \quad 2.3.10$$

$$\begin{cases} U_{y_L} = M_{U_y} - q\sigma_U \\ U_{y_R} = M_{U_y} + q\sigma_U \end{cases} \quad 2.3.11$$

А далее принцип работы до банального прост. Мы проверяем, попадают ли найденные значения в их построенные ДИ. В данном случае решающее правило меняется и выглядит так:

$$\text{Принимаем } H_0: f(X) = f(Y), \text{если } \begin{cases} U_x \in [U_{x_L}; U_{x_R}] \\ U_y \in [U_{y_L}; U_{y_R}] \end{cases}$$

В любых других случаях у нас будет недостаточно оснований для принятия гипотезы, даже при частичном выполнении условий ее принятия. В таких ситуациях необходимо провести дополнительные исследования, выбрать другой уровень значимости, использовать другие методы или, если возможно, увеличить объемы выборок.

В этом месте нашей работы пришло время рассмотреть методы проверки статистических гипотез о типе распределения ГС. На самом деле, это одна из важнейших задач по той причине, что многие методы для своего использования требуют знания типа или вообще нормальности распределения. Ниже мы рассмотрим два основных и универсальных (можно проверять гипотезы о разных типах распределения) метода проверки гипотез о типе распределения ГС по малым выборкам в достаточно кратком варианте, а их модификации для работы как с малыми, так и с очень малыми выборками рассмотрим более подробно, а также рассмотрим специальный метод для проверки распределения именно нормальность.

Первым методом проверки гипотез о типе распределения рассмотрим использование критерия χ^2. Обращаясь к [13], мы узнаем, что автор приводит данный метод как наиболее универсальный и удобный в случае проверки гипотез о

типе распределения ГС. Действительно, проверка гипотезы по данному критерию производится быстро даже при ручном вычислении без использования компьютера и сложных расчетов. Предлагаю кратко рассмотреть данный метод и удостовериться в словах автора источника. Пусть x – выборка объема n из ГС с неизвестным типом распределения. Мы хотим проверить распределение ГС на нормальность. Иными словами, мы выдвигаем гипотезу $H_0: f(x) = Norm(m, \sigma)$. Стоит отметить, что по данному критерию можно проверять гипотезы и о других типах распределения. Данный метод основан на группировке значений имеющейся выборки. Группами, в данном случае, являются отрезки числовой прямой, которые содержат в себе значения выборки. Но для начала нам необходимо получить точечные оценки параметров распределения по методам оценки параметров нормального распределения из первой главы данного раздела. После получения ТО $\bar{m}_x$и $\bar{\sigma}_x^2$ отсортированные значения нашей выборки мы делим на некоторое количество k отрезков таким образом, чтобы длина отрезков была одинаковой. После этого необходимо определить количество значений n_k, попавших на каждый отрезок. Теперь необходимо определить две вероятности – экспериментальную $p_{k_{\text{эксп}}}$ и теоретическую $p_{k_{\text{т}}}$:

$$p_{k_{\text{эксп}}} = \frac{n_k}{n} \qquad 2.3.12$$

$$p_{k_{\text{т}}} = \int_{a_k}^{b_k} \frac{1}{\bar{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{m}}{\bar{\sigma}}\right)^2}, \qquad 2.3.13$$

где a_k, b_k – границы k-ого интервала, $\bar{\sigma}$ – ТО СКО, вычисленное из ТО дисперсии. В итоге статистика метода вычисляется как:

$$\chi^2 = n \sum_{i=1}^{k} \frac{\left(p_{k_{\text{т}}} - p_{k_{\text{эксп}}}\right)^2}{p_{k_{\text{т}}}} \qquad 2.3.14$$

Далее находится критическое значение, которое сравнивается с вычисленным по формуле выше. К сожалению, метод имеет некоторые недостатки. Деление значений выборки на отрезки приводит к потере общей информации. К тому же, границы, длина и число интервалов сложно выбрать объективным способом. Корректная работа метода доказана только для малых выборок из нормальных ГС, для других видов распределения необходимы достаточно большие выборки. В некоторых

источниках рекомендуется использовать от шести, до двадцати интервалов, на каждом из которых будет более 5-10 значений [16].

Второй метод для проверки гипотез о типе распределения основан на использовании критерия Крамера-Мизеса-Смирнова ω^2, описанном в [14]. Отличительной чертой метода является то, что при его использовании исследуются все значения выборки в совокупности и без разделения на группы с дальнейшей потерей информации. С помощью данного критерия сравниваются предполагаемое распределение ГС в виде ФР ($F(x)$) и эмпирической ФР (ЭФР, $F_э(x)$) по выборке. Для построения ЭФР необходимо отсортировать выборку по возрастанию. Для наглядности и простоты понимания используем немного графического материала:

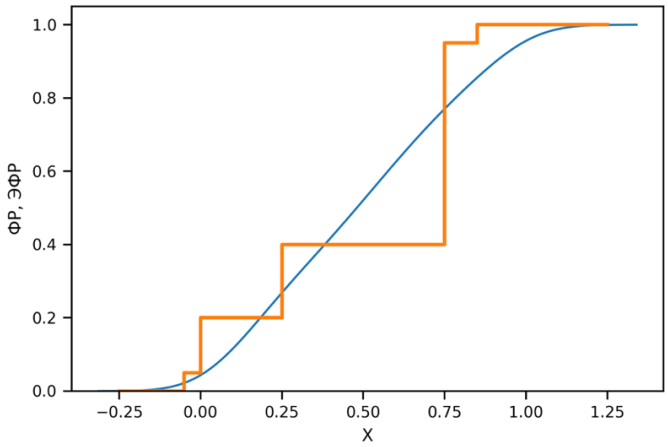

Рисунок 5. Сравнение ЭФР и ФР в критерии Крамера

На данном рисунке изображены примеры графиков ФР для нормального распределения и ЭФР по некоторой выборке. Заметна некоторая схожесть и необходимо каким-то образом проверить отклонения ЭФР от ФР. Для это существует статистика:

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_i) - \frac{2i-1}{2n} \right)^2, \quad 2.3.15$$

где $F(x_i) = 0{,}5 + \Phi\left(\frac{x_i - m}{\sigma}\right)$, Φ – функция Лапласа, вычисляемая для каждого значения выборки.

Данный метод не ограничивает информацию, содержащуюся в выборке, но имеет и минус, точнее рекомендацию к использованию. Часто рекомендуется его

использовать с выборками, объем которых позволяет распределению показателя ω^2 быть стабильным, то есть рекомендуются выборки объема сорок и более значений. Несмотря на удобство вычислений, универсальность и относительную простоту, рассмотренные выше методы проверки статистических гипотез о типе распределения имеют недостатки, которые были озвучены выше. Поэтому исследователи получают новую задачу – разработать более точные и эффективные критерии. Так, в источнике [16], описываются модификации уже рассмотренных методов, основанных на показателях χ^2и ω^2, которые более предпочтительны для малых и очень малых выборок. Данные модификацию могут применяться для проверки гипотез о разных типах распределения.

Модифицированный метод проверки статистических гипотез о типе распределения по показателю χ^2. В данном случае мы рассмотрим сразу несколько распределений. Иными словами, мы выдвигаем сразу несколько гипотез, каждую для предполагаемого типа распределения, и пытаемся выбрать наиболее вероятную. До этого мы находили теоретическую вероятность по зависимости

$$p_{k_т} = \int_{a_k}^{b_k} \frac{1}{\bar{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{m}}{\bar{\sigma}}\right)^2} \quad 2.3.16$$

Данную зависимость в общем виде можно представить как

$$p_{k_т} = f(x \mid \{\bar{a}\}), \quad 2.3.17$$

где $\{\bar{a}\}$ – вектор точечных оценок неизвестных параметров распределения.

Далее предлагается использовать принцип, заложенный в биномиальное распределение [8], формулы для МО и дисперсии данного распределения:

$$M_m = nP \quad 2.3.18$$

$$\sigma_m^2 = nP(1-P) \quad 2.3.19$$

Данные показатели можно интерпретировать как наиболее вероятное число появлений события при заданной вероятности в заданном количестве экспериментов (МО) и погрешность данного числа появлений (дисперсия).

На основе стандартной версии критерия, используя биномиальное распределение, мы можем вычислить показатель вероятного числа значений, которые попадут на каждый отрезок, а также погрешности данного показателя в виде его дисперсии. Данная задача полностью соответствует сути биномиального распределения. У нас есть заданное количество «экспериментов» (объем выборки) и

вероятность появления на каждом отрезке $p_{k_т}$. По данным выше формулам определяем МО и дисперсию для каждого отрезка:

$$M_k = np_{k_т} \quad 2.3.20$$

$$\sigma_k^2 = np_{k_т}(1 - p_{k_т}) \quad 2.3.21$$

Расхождения теоретического и фактического распределения на каждом из отрезков с учетом найденных МО и дисперсии можно определить как:

$$q_k = \frac{(n_k - M_k)}{\sigma_k} \quad 2.3.22$$

Далее вводится удобный показатель S – нормированная и центрированная версия СВ, которая отображает суммарные расхождения теоретического и фактического распределений [16]:

$$S = \sqrt{2K}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{n_k - M_k}{\sigma_k}\right)^2} - 1\right) \quad 2.3.23$$

В [16] говорится об исследовании распределения данного показателя. Оно тестировалось при нескольких типах распределения ГС. Объем выборок начинался с 20 и заканчивался 100, количество отрезков – от 5 до 20. В каждом варианте производилась генерации 1000 выборок.

Для проверки гипотезы необходимо найти критическое значение S_{cr}. Так как доказана нормальность распределения этого показателя, используется функция Лапласа или таблица квантилей стандартного нормального распределения. При этом нужно соблюдать соотношение **[1]**:

$$0{,}5 + \Phi(S_{cr}) = 1 - \alpha, \quad 2.3.24$$

где $\Phi(S_{cr})$ – значение функции Лапласа в точке S_{cr}. Формируется гипотеза, что наша ГС имеется некоторый известный тип распределения g: $H_0: f(x) = g(x)$. Для проверки этой гипотезы задаем уровень значимости α (соответственно появляется $p = 1 - \alpha$ доверительная вероятность) и ищем показатель S по имеющейся выборки и сравниваем с критическим. Решающее правило в данной случае будет иметь вид:

$$\text{Принимаем } H_0: f(x) \sim g(x), \text{если } S \leq S_{cr}$$

$$\text{Отклоняем } H_0: f(x) \sim g(x), \text{если } S > S_{cr}$$

Согласно [16], данный модифицированный метод отличается более высокой эффективностью при работе с малыми выборками, а также отсутствием

необходимости использования таблиц распределения показателя χ^2 или интегрирования его плотности. Но, к сожалению, главный недостаток классического метода не изменился. Мы все равно разбиваем значения выборки на отрезки, не имея четкого метода сделать это объективно. А это влечет за собой искажение информации, которое в одном исследовании может быть не значительно, а в другом – критично. Поэтому при исследованиях, зачастую, не полагаются на какой-то конкретный метод, а используют целые комплексы методов и делают выводы на основе анализа совокупности результатов каждого метода. Исследования таким образом положительно влияют на результаты.

Мы находимся на финишной прямой исследования методов проверки статистических гипотез о типах распределения по малым выборкам. Выше была рассмотрена модификация метода проверки по критерию χ^2, которая, лучше подходит для малых выборок, что доказывается в [16]. Настало время рассмотреть модификацию второго метода проверки статистических гипотез о типах распределения по малым выборкам, а именно модифицированный показатель ω^2 – критерий Крамера-Мизеса-Смирнова. Он по-прежнему обладает плюсом анализа всей информации в выборке без разделения на группы [12, 15]. В модификации степень отклонения ЭФР от ФР (сам показатель ω^2) осредняется и преобразуется для более аккуратного распределения в виде G [16]:

$$G = \left[\frac{1}{12n} + \sum_{i=1}^{n}\left(F(x_i) - \frac{2i-1}{2n}\right)^2\right]^{0,5} \qquad 2.3.25$$

Распределение данного показателя исследовалось при помощи статистического моделирования. Тестировались известные типы распределения с известными параметрами, количество выборок при тестировании было равно 10000, а их размер – от 10 до 100 [16]. В этом же источнике автор говорит, что по результатам тестирования были сделаны следующие выводы:

1. Показатель имеет устойчивое распределение
2. Распределение очень слабо зависит от объема выборки
3. $M_G \in [0{,}372; 0{,}375]$
4. $\sigma_G \in [0{,}152; 0{,}155]$

При тестировании известных распределений с неизвестными параметрами, замененными на ТО, были выявлены некоторые нюансы. При работе с выборками,

объем которых попадает в тестовый отрезок, распределение показателя G не изменилось для распределения Релея, нормального и Максвелла. Небольшие изменения были замечены для равномерного, экспоненциального, Вейбулла и гамма.

Перейдем к способу проверки гипотезы о типе распределения. Важно отметить, что данный метод, как и модифицированный χ^2, позволяет достаточно легко, наглядно, быстро и, главное, точно проверять сразу несколько гипотез о разных типах распределения и выбирать из них наиболее статистически достоверную. Выдвигаем все ту же гипотезу $H_0: f(x) = g(x)$, задаем уровень значимости α и вычисляем доверительную вероятность p. По данной вероятности определяем критическое значение G_{cr} для предполагаемого распределения по специальной таблице (*приложение 3*). Производим ТО параметров распределения, после этого находим показатель G. Решение о принятии гипотезы производится по следующему правилу:

$$\text{Принимаем } H_0: f(x) \sim g(x), \text{ если } G \leq G_{cr}$$

$$\text{Отклоняем } H_0: f(x) \sim g(x), \text{ если } G > G_{cr}$$

Данный метод по-прежнему учитывают всю информацию в выборке и не требует субъективных действий (например, разбиения на интервалы), поэтому должен давать более точную информацию.

И последним рассмотрим специальный метод проверки гипотез. Специальные методы предназначены для проверки гипотез о каком-либо одном известном типе распределения. Для проверки гипотез о нормальном распределении ГС наиболее эффективным [24-26] по совокупности тестирований является метод (критерий) Шапиро-Уилка [3, 5, 24-26]. Критерий Шапиро-Уилка W позволяет проверять только один тип гипотез – о нормальном распределении ГС по полученной выборке. Данный критерий наиболее эффективен при объемах выборки $8 \leq n \leq 50$ [3, 5, 24-26]. В нашей работе мы рады данному ограничению, так как выборки подходящего критерию объема являются объектом наших исследований. Поэтому и критерий можно считать обоснованным и применимым. Для выборок большего объема существует модификация данного метода, которая увеличивает максимальный порог количества значений в выборке до 2000. Данную модификацию рассматривать не будем и сконцентрируемся на версии для малых выборок. Выдвигаем гипотезу

$H_0: f(x) = Norm(M, \sigma)$. Для использования данного критерия необходимо получить ТО МО и дисперсии и упорядочить значения в выборке (по возрастанию). Далее предлагается использование статистики W следующего вида:

$$W = \frac{1}{\bar{\sigma}^2}\left(\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right)^2, \quad 2.3.26$$

где a_{n-i+1} – вспомогательный коэффициент, который определяется по специальной таблице (*приложение 1*). К сожалению, уже при 21 значении в выборке, таблица имеет неприятно большие размеры. Поэтому решалась задача аппроксимации и получения формул для приблизительного расчета удовлетворяющей точности. Полученные формулы для нахождения коэффициентов согласно [3] выглядят так:

$$a_i = a_0\left(z + \frac{1483}{(3-z)^{10,845}} + \frac{71,6 * 10^{-10}}{(1,1-z)^{8,26}}\right), \quad 2.3.27$$

$$a_0 = \frac{0,899}{(n-2,4)^{0,4162}} - 0,02, \quad 2.3.28$$

$$z = \frac{n-2i+1}{n-0,5} \quad 2.3.29$$

Проверка гипотезы, по классике, производится сравнением полученного показателя W и критического значения W_{cr}. Предельное значение также можно определить двумя способами, но перед этим необходимо задать уровень значимости учитывая деталь, которая будет описана далее. Первый способ – табличный (*приложение 2*) – поиск значения по объему выборки и доверительной вероятности. Второй – приближенные формулы.

Проверка гипотезы о нормальности распределения, на этот раз, проводится не так, как в прошлых методах. Во-первых, уровень значимости, который мы задаем, также представляет собой и доверительную вероятность в данном методе. Во-вторых, в отличие от всех прошлых методах, в данном гипотеза принимается, если полученное значение больше критического, а не меньше. То есть имеем решающее правило:

$$\text{Принимаем } H_0: f(x) \sim Norm(M, \sigma), \text{если } W > W_{cr}$$

$$\text{Отклоняем } H_0: f(x) \sim Norm(M, \sigma), \text{если } W \leq W_{cr}$$

В финальной части второго раздела данной работы предлагается рассмотреть еще два статистических метода – дисперсионный и регрессионный анализ. Если

ранее мы рассматривали методы, позволяющие, на самом деле, получить относительно небольшой объем информации, например – оценка параметра, проверка гипотезы о нормальном распределении и так далее. Все эти методы так или иначе являются основой для более комплексных, сложных и информативных методов. Отличие данных методов в том, что с помощью них исследуются более сложные вопросы, например, о наличии зависимости внутри выборки или о различии исследуемых групп. Выше мы несколько раз сталкивались с данными задачами, ведь зависимость, в некотором роде, мы могли определить и по коэффициенту корреляции, а сравнение групп – провести по t-критерию. К сожалению, это самые примитивные способы решения задачи исследования зависимости или различия, они имеют ограничения (например, по t-критерию можно сравнить только две группы) и проблемы интерпретации (например, коэффициент корреляции, на самом деле, не говорит именно о зависимости, он лишь определяет меру взаимосвязи между двумя количественными переменными). Отсюда возникает вопрос, что делать, если исследуется три и более групп на различие или определяется зависимость количественной переменной от качественной. На данный момент методы, которые разбирались в работе, ничем не могут помочь при решении подобных вопросов. И именно для расширения возможностей и списка проблем, которые можно с помощью них решать, используются более продвинутые и сложные методы. В данную работу включены два метода – дисперсионный и регрессионный анализ. Они, в свою очередь, являются комплексами многих статистических методов, часть из которых мы рассмотрели выше. Используя данные методы, мы будем решать более сложные задачи, а именно проверка гипотез о зависимости внутри выборки, как между количественными переменными, так и между количественными и качественными. Дисперсионный и регрессионный анализ применим и к малым выборкам, что означает для нас возможность их использования при анализе выборок установленного в работе объема (до 200 объектов). Конечно, имеются некоторые ограничения для применения каждого метода, но установленный в работе объем выборок в это ограничение почти всегда попадает. Более подробно об ограничениях и использовании методов с малыми выборками будет сказано при описании дисперсионного и регрессионного анализа, к чему мы и перейдем.

Первым мы рассмотрим статистических метод, который называется дисперсионный анализ (ДИ), разработанный в 1918 году Р. Фишером [6, 8, 18]. Во многих англоязычных источниках данный метод носит название ANOVA – Analysis of variance. Это очень эффективный и универсальный статистических метод, который найдет свое применение практически в любом исследовании. Его суть заключается в проверке влияния какого-либо фактора (независимой переменной) на зависимую (признак). Если влияние подтверждается, то можно разработать способы управления данными факторами для получения желаемых значений зависимой переменной. Главной особенностью данного метода является то, что при его использовании можно исследовать как количественные, так и качественные факторы и признаки. Множество других статистических методов возможно использовать с числами, при этом качественные переменные создают проблемы для исследователей. Основным недостатком данного метода являются сложности в автоматизации вычислений, особенно при исследовании влияния более 2 факторов. Это обусловлено тем, что отсутствует четкая структура входных данных, от которых зависят последующие вычисления. Например, фактор может иметь 3 градации, а может 2, 4, 5 и так далее. А когда факторов становится больше, построение универсального алгоритма приводит к большим сложностям. Поэтому во многих исследованиях исследователи не могут полностью отдать основную часть работу программным решениям, логику метода необходимо формировать под конкретную задачу. Мы, в дальнейшем, также будем придерживаться данного способа организации исследований. Стоит отметить, что в любом случае мы не сможем достоверно сказать обо всех факторах, влияющих на изменчивость признака. Но, на самом деле, мы и не ставим себе такой цели. Нашей целью будет проверка некоторой гипотезе о том, что изменчивость вызвана фактором. Определить, какой именно фактор влияет на изменчивость можно на основе статистических методов и сложных вычислений, но для дисперсионного анализа, зачастую, изменчивость предполагают на основе хороших знаний области исследования, прошлых экспериментов, и, что вполне допустимо, здравого смысла и логики. В работе будет использоваться две версии дисперсионного анализа – однофакторная и многофакторная. С помощью дисперсионного анализа мы будем проверять значимость влияния различных факторов на зависимые признаки. Как по-отдельности (в случае однофакторного

ДИ), так и в совокупности (в случае многофакторного). Возникает вопрос, почему анализ именно дисперсионный. Ответ достаточно прост – дисперсия зависимого признака формируется из дисперсий влияющих на нее факторов, дисперсии их взаимодействия и некоторой случайной и не учтенной дисперсии. То есть, мы можем объяснить изменчивость (дисперсию) зависимого признака изменчивостью (дисперсиями) факторов, влияющих на него. Поэтому мы будем работать с оценками изменчивости данных, в том числе и с дисперсией.

Поговорим о применимости дисперсионного анализа к малым выборкам. Были проанализированы многие источники информации о дисперсионном анализе, например [6, 18]. Зачастую, примеры использования ДА или исследований с его помощью приводят на относительно небольших выборках объема от 20 до 500 в зависимости от количества факторов и их уровней. Очевидно, что выборка объема, допустим, 20 объектов, лучше подойдет в случае однофакторного анализа. И дело тут в некоторых базовых ограничениях (и рекомендациях) размера. Каждый фактор вариацией своих значений образует группы значений признака. Почти во всех источниках отмечается, что размер этих групп ограничен снизу – минимум 2 значения в каждой группе. Некоторые авторы дают дополнительные рекомендации по минимальному размеру групп – минимум 10 значений в каждой группе. Именно поэтому однофакторный ДА будет более эффективным для очень малой выборки. Просто по причине того, что размер групп после разбиения будет больше, ведь при использовании многофакторного ДА групп становится больше, а объем выборки не увеличивается. Также при использовании однофакторного ДА существует требование по количеству вариаций фактора, которых должно быть не менее трех. Но, как оказалось на практике, ДА очень часто применяется при работе с выборками объема от 100 до 200 объектов. Главное правильно интерпретировать полученные результаты. Исходя из этого, мы не можем найти причины не использовать данный метод в работе с нашими выборками объема до 200 объектов. Более того, в условиях наших ограниченных размеров, метод показывает достойные и адекватные результаты. Главное – помнить о базовых ограничениях, не ошибаться при выборе уровня значимости и правильно интерпретировать результаты.

Нулевая гипотеза в дисперсионном анализе выдвигается на основе предположения, что влияние фактора или факторов отсутствует. Поэтому для нас,

как для исследователей, появляется задача статистически опровергнуть эту гипотезу. Рассмотрим два вида дисперсионного анализа более подробно.

Однофакторный дисперсионный анализ (one-way ANOVA). Данный статистический метод применяется при проверке гипотезы о влиянии только одного фактора, который может иметь различное число градаций. Для удобства можно формировать «группы» из значений признака, у которых совпадает значение критерия. В результате можно будет говорить о межгрупповых сравнениях, что, на мой взгляд, легче воспринимается и усваивается. Далее будем говорить об исследованиях градаций (уровней) фактора в виде групп. Рассмотрим выборку размера n, которая имеет качественную переменную с $m = 3$ значениями и количественную. По значениям количественной переменной мы формируем три группы. Предполагаем, что существует зависимость и выдвигаем заведомо нежелательную гипотезу $H_0: M_1 = M_2 = M_3$. Данную гипотезу нужно интерпретировать как равенство средних в каждой группе – иными словами, значения фактора не влияют на признак. Далее нам необходимо найти ТО математических ожиданий – общего и групповых:

$$\bar{M} = \bar{M}_1 + \bar{M}_2 + \bar{M}_3 \quad 2.3.30$$

ТО групповых МО находятся по рассмотренным выше методам. После этого можем перейти к одному из основных понятий – общей сумме квадратов SST (sum squares total). Это, своего рода, мера изменчивости данных без учета фактора, что чем-то похоже на дисперсию. Формула для SST следующая:

$$SST = \sum_{i=1}^{n} (x_i - \bar{M})^2 \quad 2.3.31$$

Действительно, внешность зависимости очень напоминает дисперсию. Далее необходимо вычислить количество степеней свободы $df_{SSB} = n - 1$. К слову, если мы поделим SST на df, то получим ТО дисперсии. В свою очередь, общая сумма квадратов состоит из двух других показателей – межгрупповой SSB и внутригрупповой SSW суммы квадратов.

Рассмотрим SSB. Данный показатель характеризует различия или изменчивость в значениях между группами. Для его вычисления понадобится использовать ТО групповых МО. А также определить n_1, n_2, n_3 – количество элементов в каждой группе. Далее используем общее МО и находим SSB:

$$SSB = \sum_{i=1}^{m} n_i(\bar{M}_i - \bar{M})^2 \quad 2.3.32$$

Количество степеней свободы вычисляется как $df = m - 1$.

Далее рассмотрим SSW. Данное значение характеризует общую изменчивость значений внутри каждой группы, в нее также входят случайные и не учтенные расхождения. То есть сравнение будет вестись с групповыми МО. Для условий нашей задачи формулу запишем в следующем виде:

$$SSW = \sum_{i=1}^{n_1}(x_{i1} - \bar{M}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{M}_2)^2 + \sum_{i=1}^{n_3}(x_{i3} - \bar{M}_3)^2, \quad 2.3.33$$

где x_{i1} – i-ый элемент первой группы, со второй и третьей группой аналогично.

Данную формулу можно представить в другом виде, если все группы одинакового размера n. Поделим обе части на $n - 1$ и получим:

$$\frac{SSW}{n-1} = \frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{M}_1)^2}{n-1} + \frac{\sum_{i=1}^{n_2}(x_{i2} - \bar{M}_2)^2}{n-1} + \frac{\sum_{i=1}^{n_3}(x_{i3} - \bar{M}_3)^2}{n-1} \quad 2.3.34$$

В правой части уравнения в явном виде представлены формулы для вычисления групповых дисперсий, преобразуем зависимость итоговому виду:

$$SSW = (n-1)(\bar{\sigma}_1^2 + \bar{\sigma}_2^2 + \bar{\sigma}_3^2) \quad 2.3.35$$

Число степеней свободы для данного показателя вычисляем как $df_{SSW} = n - m$.

В итоге, в общем случае, должно выполняться равенство $SST = SSB + SSW$. Если $SSB > SSW$, то возможны статистические различия между группами, которые нужно подтвердить какой-либо статистикой. Предлагаю использовать F-критерий по нескольким причинам:

1. Успешно работает с исследованием дисперсий, а найденные величины очень на них похожи.
2. Из найденного можем составить осмысленную статистику, которая сильно схожа с F-критерием:

$$F = \frac{SSB/df_{SSB}}{SSW/df_{SSW}} \quad 2.3.36$$

   Важно отметить, что в числителе всегда должно быть большее значение.
3. Данный критерий уместен при малых выборках.

Далее задаем уровень значимости, находим доверительную вероятность и проверяем гипотезу по F-распределению, сравнивая найденное значение с критическим. Решающее правило имеет вид:

$$\text{Принимаем } H_0: M_1 = M_2 = M_3, \text{если } F \leq F_{cr}$$

$$\text{Отклоняем } H_0: M_1 = M_2 = M_3, \text{если } F > F_{cr}$$

Если гипотезу отклонить не удалось, но имеются веские причины предполагать ее истинность, необходимо провести дополнительные исследования другими методами или увеличить объем выборки. Если мы смогли отклонить гипотезу, то это означает, что были получены статистически значимые различия между группами. Однако, если количество исследуемых групп $m \geq 3$, то мы не можем сказать, какие именно группы имеют значимые отличия. Статистически подтвержденные отличия говорят лишь о том, что как минимум две группы отличаются друг от друга. В исследованиях с помощью дисперсионного анализа часто визуализируют результаты. По графику можно предположить направление зависимости признака от фактора. Например, в однофакторном ДА мы могли получить подобные результаты (если фактор имеет два уровня):

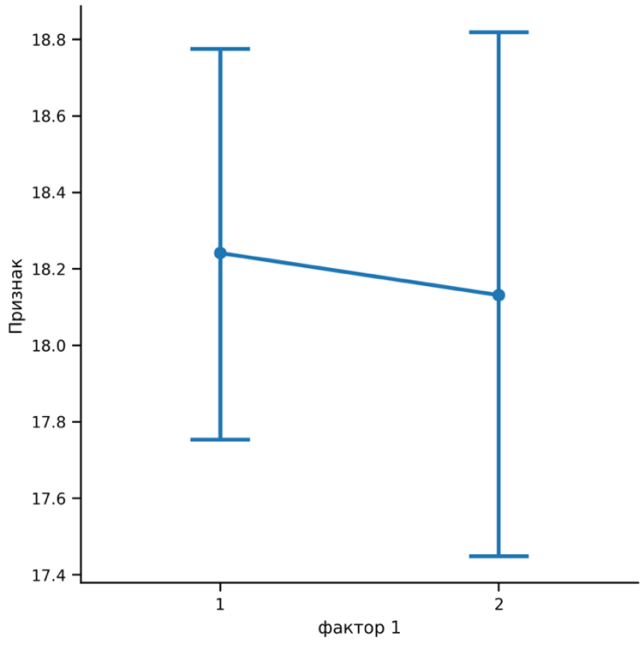

Рисунок 6. Графический пример возможных результатов ОДА

По графику можно заметить, что увеличение (хотя, в случае качественных переменных правильнее будет говорить просто об изменении) значения фактора приводит к снижению значения признака, пусть и не совсем значительному.

Пример использования ОДА для малой выборки приведен в третьем разделе и реализован программным способом.

В итоге можно сказать, что данным методом мы частично подтвердили то, что исследуемый фактор своими вариациями может влиять на признак и, на основе анализа графика результатов, установили направление предполагаемого влияния. Почему подтвердили только частично? Затронем данный вопрос при подведении итогов и после того, как рассмотрим следующий вид дисперсионного анализа – многофакторный, к которому сейчас и перейдем.

Многофакторный дисперсионный анализ позволяет проверять предположения о том, что на зависимый признак имеют влияние сразу несколько факторов. Более того, многофакторный ДА также позволяет проверить гипотезу о том, что один фактор может оказывать влияние на признак в зависимости от значений другого фактора (других факторов). Данную ситуацию будем называть взаимодействием факторов. Так как принцип работы и формулы для вычислений, в сравнении с однофакторным случаем, глобально не меняются, то рассмотрим ситуацию, когда есть основания предполагать, что исследуемый признак зависит от двух факторов, каждый из которых имеет два уровня (значения). Обозначим их цифрами 1 и 2. Также поделим значения признака на 4 группы в зависимости от значений факторов. Удобнее представить данные в виде таблицы:

Таблица 2. Обобщенный пример данных для МДА

Фактор 1	Фактор 2	Группы значений	Средние	Размер групп
1	1	$\{x_{11}\}$	$\bar{M}_{x_{11}}$	n_1
1	2	$\{x_{12}\}$	$\bar{M}_{x_{12}}$	n_2
2	1	$\{x_{21}\}$	$\bar{M}_{x_{21}}$	n_3
2	2	$\{x_{22}\}$	$\bar{M}_{x_{22}}$	n_4

Начнем работу с фактора 1. Необходимо будет вычислить 3 МО:

$$\bar{M}_{a1} = \bar{M}_{x_{11}} + \bar{M}_{x_{12}} \quad 2.3.37$$

$$\bar{M}_{b1} = \bar{M}_{x_{21}} + \bar{M}_{x_{22}} \quad 2.3.38$$

$$\bar{M}_{f1} = \frac{\bar{M}_{a1} + \bar{M}_{b1}}{2} \quad 2.3.39$$

Мы нашли МО для значений, имеющих различные уровни фактора 1, а также общее МО для данного фактора. То есть работали с группировкой по разным значениям фактора 1. На основе этих данных найдем межгрупповую сумму квадратов по фактору 1:

$$SSB_1 = (n_1 + n_2)\left(\bar{M}_{a1} - \bar{M}_{f1}\right) + (n_3 + n_4)(\bar{M}_{b1} - \bar{M}_{f1}) \quad 2.3.40$$

В принципе, наши действия в точности повторяют принцип однофакторного анализа с учетом того, что второй фактор мы «опускаем» и представляем, что у нас всего $m_1 = 2$ группы для сравнения. Число степеней свободы в данном случае равно $df_1 = m - 1 = 1$. Аналогичные действия проводятся и со вторым фактором. В этом случае находим МО:

$$\bar{M}_{a2} = \bar{M}_{x_{11}} + \bar{M}_{x_{21}} \quad 2.3.41$$

$$\bar{M}_{b2} = \bar{M}_{x_{12}} + \bar{M}_{x_{22}} \quad 2.3.42$$

$$\bar{M}_{f2} = \frac{\bar{M}_{a2} + \bar{M}_{b2}}{2} \quad 2.3.43$$

Далее находим межгрупповую сумму квадратов по фактору 2:

$$SSB_2 = (n_1 + n_3)\left(\bar{M}_{a2} - \bar{M}_{f2}\right) + (n_2 + n_4)(\bar{M}_{b2} - \bar{M}_{f2}) \quad 2.3.44$$

Так как фактор 2 тоже имеет только два различных значения, число степеней свободы остается таким же: $df_2 = df_1 = m - 1 = 1$.

Последней суммой квадратов, которую мы рассмотрим, будет межгрупповая, которая отвечает за взаимодействие факторов. Если ранее мы рассматривали группы, которые по очереди формируются значениями только одного фактора, то теперь рассмотрим четыре группы, которые формируются набором уникальных значений первого и второго фактора. Найдем глобальное МО $\bar{M}_{f1f2}$:

$$\bar{M}_{f1f2} = \frac{\bar{M}_{x_{11}} + \bar{M}_{x_{12}} + \bar{M}_{x_{21}} + \bar{M}_{x_{22}}}{m} \quad 2.3.45$$

После этого возможно найти сумму квадратов SSB_{12} для взаимодействия факторов:

$$SSB_{12} = n_1\left(\bar{M}_{x_{11}} - \bar{M}_{f1f2}\right) + n_2\left(\bar{M}_{x_{12}} - \bar{M}_{f1f2}\right) + n_3\left(\bar{M}_{x_{21}} - \bar{M}_{f1f2}\right) + n_4(\bar{M}_{x_{22}} - \bar{M}_{f1f2}) \quad 2.3.46$$

Так как для каждого фактора мы имели по две группы значений, то можно сказать, что было $m_1 = m_2 = 2$ группы для каждого фактора. Исходя из этого, количество степеней свободы для суммы квадратов при взаимодействии найдем как $df_{12} = (m_1 - 1)(m_2 - 1) = 1$. После этого мы находим внутригрупповую сумму

квадратов, которая будет единой для обоих факторов. То есть суммируем квадраты разностей значений и МО для каждой из $m = 4$ начальных групп:

$$SSW = \sum_{i=1}^{n_1}\left(x_{i11} - \bar{M}_{x_{11}}\right) + \sum_{i=1}^{n_2}\left(x_{i12} - \bar{M}_{x_{12}}\right) + \sum_{i=1}^{n_3}\left(x_{i21} - \bar{M}_{x_{21}}\right) + \sum_{i=1}^{n_4}\left(x_{i22} - \bar{M}_{x_{22}}\right) \quad 2.3.47$$

Для данного показателя число степеней свободны равно $df = n_1 + n_2 + n_3 + n_4 - m$. Чтобы проверить набор гипотез о влиянии факторов, найдем значения F-критериев для каждой суммы. Набор гипотез следующий:

1. Значимое влияние первого фактора
2. Значимое влияние второго фактора
3. Значимое влияние при взаимодействии двух факторов

Пользуемся зависимостью для F-критерия, которая уже встречалась в однофакторном ДА, но в данном случае для проверки каждой гипотезы мы меняем значения межгрупповых сумм квадратов:

$$F_i = \frac{SSB_i / df_i}{SSW / df} \quad 2.3.48$$

Для проверки задаем уровень значимости α, находим доверительную вероятность и критическое значение F_{cr} для каждой гипотезы с учетом степеней свободы. Находим либо по таблицам, либо по плотности распределения Фишера.

Для проверки гипотезы имеем решающее правило общего вида:

$$\text{Принимаем гипотезу, если } F_i > F_{cr}$$

$$\text{Отклоняем гипотезу, если } F_i \leq F_{cr}$$

После статистического подтверждения гипотез, очень часто результаты визуализируются в виде облегченной версии графика box-plot. Например, в нашем случае мы могли бы получить такой график:

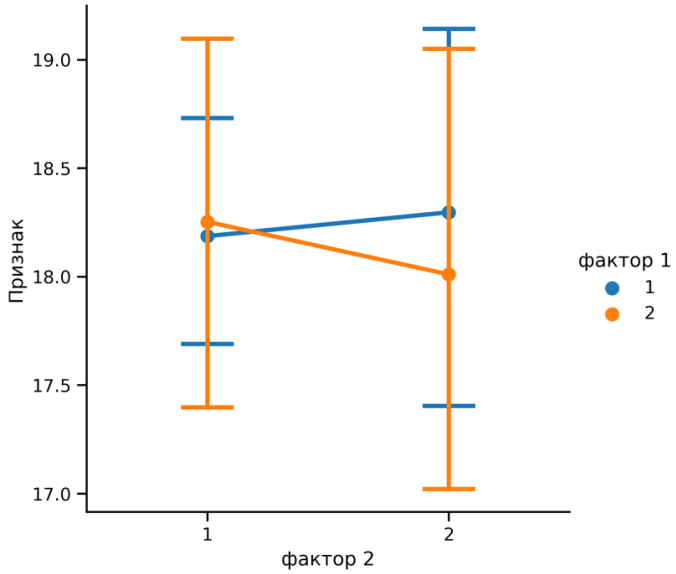

Рисунок 7. Графический пример возможных результатов МДА

В данной ситуации сразу видно как взаимодействие факторов, так и влияние каждого фактора на признак. По графику видно, что при значении 2 фактора 1 изменение значения фактора 2 приводят к понижению значения признака. Для значения 1 фактора 1 ситуация обратная. Как видим, визуальный анализ не стоит списывать со счетов, ведь он может стать хорошим дополнением к уже полученным статистическим выводам.

Подводя итог, дисперсионный анализ является хорошим методом в исследованиях степени зависимости количественной переменной от качественной в малых выборках. Также можно сказать, что данный метод может быть использован для сравнения нескольких групп между собой. Данные группы формируются на основе различных уровней (значений) какого-то фактора (качественной переменной). ДА позволяет предполагать зависимость как от одного фактора, так и от нескольких сразу. Более того, с помощью этого метода можно проверить более комплексные предположения, например, что значения зависимой переменной изменяются под влиянием нескольких взаимодействующих факторов, а не из-за влияния каждого отдельного фактора. Для использования метода необходимо иметь $n_i \geq 2$ наблюдений в каждой группе. Иными словами, данный метод можно применять в исследованиях малых выборок, размер выборки не оказывает колоссального влияния, хотя некоторые источники дают дополнительную рекомендацию для размера $n_i \geq 10$ для каждой группы. Еще одной положительной характеристикой является то, что вычисления в методе относительно простые и

могут быть произведены даже без применения специальных вычислительных средств. К сожалению, из-за того, что заранее неизвестно ни количество факторов, ни количество уровней в факторах, вычисления легче проводить по зависимостям, которые формируются уже после получения данных, а не до этого момента. Выше мы отметили положительные стороны метода, теперь необходимо поговорить и об отрицательных. В работе уже отмечалось, что положительный результат дисперсионного анализа позволяет подтвердить зависимость лишь частично. Это обусловлено тем, что на самом деле положительные результаты дисперсионного анализа, то есть отклонение не желаемой гипотезы и принятие альтернативной или, наоборот, принятие желаемой гипотезы, к большому сожалению, не позволяют уверенно говорить о причинно-следственной связи между исследуемыми объектами. Такой вывод куда более комплексный, сложный, и прийти к нему можно при исследовании выборки совокупностью различных методов, которые дадут статистически значимые различия. Также не нужно забывать о логике, которой можно объяснить или опровергнуть результаты исследования. Ведь может случиться такое, что зависимость будет совсем нелогична или, более того, невозможна, но рассчитанные показатели скажут об обратном. Поэтому при использовании ДА всегда стоит задаваться вопросом о возможности объяснения результатов исследования с обратной стороны. То есть возможна ли такая ситуация, когда различные значения зависимого признака будут влиять на изменение значений фактора, а не наоборот. Но, все-таки, дисперсионный анализ может дать начало дополнительным исследованиям, так как позволяет оперативно и обоснованно предположить реальное наличие зависимости. Еще одним, скорее отрицательным, нюансом является то, что однофакторный дисперсионный анализ требует, как минимум, трех градаций фактора. При двух уровнях результаты будут такие же, как при сравнении средних по t-критерию, вычисление которого займет меньше времени при одинаковом результате. Вообще, требования для использования дисперсионного анализа лучше представить в виде списка:

1. Объем каждой группы $n_i \geq 2$, желательно $n_i \geq 10$.
2. Не менее трех различных значений фактора при однофакторном ДА
3. Получение данных из ГС с непрерывным распределением, дискретное распределение ГС менее предпочтительно, но возможно.

4. Примерное равенство групповых дисперсий без сильных отклонений. Данное требование хоть и существует, но на практике в исследованиях зачастую опускается [19, 22, 23].
5. Независимость наблюдений в группах и групп между собой. То есть градация признака не должна влиять сама на себя.
6. Близкое к нормальному распределение признака. Но, например в [21] отмечается, что это требование весьма условно.

Пример использования МДА для малой выборки приведен в третьем разделе и реализован программным способом.

Последним в списке более комплексных методов и, вместе с этим, последним в работе будет регрессионный анализ (РА) [4, 6, 16]. Этот метод позволяет исследовать зависимость между переменными, выявлять переменную, влияющую на зависимую, если «несколько», а также предсказывать и контролировать значения зависимой переменной [6, 16]. Главной идеей является построение регрессионной модели. Она имеет вид функции зависимой переменной от независимой и может быть линейной и нелинейной. В данной работе мы рассмотрим линейный регрессионный анализ. Это мощный и полезный метод, который позволяет решать большое количество задач и основывается на построении регрессионной модели в линейном виде. Например, с помощью него мы можем проверить зависимость между количественными переменными, что, к слову, невозможно сделать с использованием дисперсионного анализа. Также мы можем проверять не только зависимость, но и определять ее силу. На основе полученных результатов о предполагаемой зависимости мы можем построить модель данной зависимости. А построение модели, в свою очередь, поможет решить еще две важные задачи – предсказание значений и получение путей контроля исследуемого явления с помощью других явлений, от которых оно зависит. Существует несколько разновидностей регрессионного анализа, но в данной работе мы остановимся на двух – одномерной и многомерной линейной регрессии. Это позволит нам работать как с простыми случаями влияния на зависимую переменную одной независимой, так и с более сложными – когда независимую переменную влияют сразу нескольких независимых. Мы также рассмотрим способы выбора наилучшей регрессионной модели с целью получения более точных результатов предсказания или проверки

значений. Как и в случае с дисперсионным анализом, регрессионный анализ также имеет свои требования к данным, о которых будет сказано при описании каждого подвида линейного регрессионного анализа. На данном этапе стоит затронуть тему применимости данного метода к выборкам, объем которых соответствует установленному в работе (до 200 объектов), то есть необходимо проанализировать применение регрессионного анализа к малым выборкам.

Во многих источниках информации о РА вопроса размера исследуемых выборок не касаются. В основных требованиях, которые будут приведены ниже, требования по размеру также отсутствуют. В некоторых источниках даются рекомендации работать с переменными, которые имеют от 20-25 значений. Также присутствуют рекомендации, что количество наблюдений должно в 5-6 раз превышать количество исследуемых переменных. Было принято решение проверить работу с малыми выборками на практике. И здесь подтвердилась основная идея регрессионного анализа – важным обстоятельством является наличие зависимости между переменными в том или ином виде. Проверить это можно, например, по коэффициенту корреляции. Поэтому размер выборки ограничивается тем, что он должен позволять достаточно точно определить наличие связи между переменными. Понятно, что для очень малых выборок, которые суммарно имеют до 30 значений, корректное определение связи будет достаточно сложным и не всегда предсказуемым занятием. Но, когда размер каждой из исследуемых переменных превышал 20-30 значений, результаты были вполне адекватными, как при исследовании связи, так и при использовании линейного регрессионного анализа в целом. Данное обстоятельство подтверждает озвученные выше рекомендации к размеру выборки. Поэтому будем им следовать и применять данный метод к выборкам, в которых исследуемые переменные состоят минимум из 25-30 значений. Не лишним будет отметить, что РА часто применяется на хорошо подготовленных данных. Например, при анализе каких-то показателей по регионам или прочим территориям, когда используется среднее значение для каждого объекта. В таком случае наличие выборки большого размера почти никак не скажется на результатах при условии, что используемые средние значения были качественно вычислены или оценены. При этом в работе с качественными данными стоит все-таки соблюдать вышеупомянутую рекомендацию к минимальному размеру хотя бы из

предположения о том, что меньший объем может не дать адекватно установить связь между переменными.

После озвучивания вопросов размера выборок можно переходить к разбору первой разновидности линейной регрессии – одномерной линейной регрессии (ОЛР). Одномерная линейная регрессия предполагает, что одна независимая переменная X влияет на зависимую переменную Y линейно. Модель регрессии в данном случае выглядит так:

$$Y = b_0 + b_1 X \quad 2.3.49$$

В данном случае это уравнение прямой, которая описывает зависимость между нашими переменными и направление этой зависимости. Основной задачей является определение такой прямой, которая максимально отображает зависимость и ее направление, а также хорошо описывает распределение данных, то есть каждое наше значение по возможности должно быть максимально близко к регрессионной прямой. Основным способом добиться таких результатов является использование метода наименьших квадратов для подбора максимально верных коэффициентов регрессионной прямой. Проще всего данную идею будет рассмотреть с помощью графика:

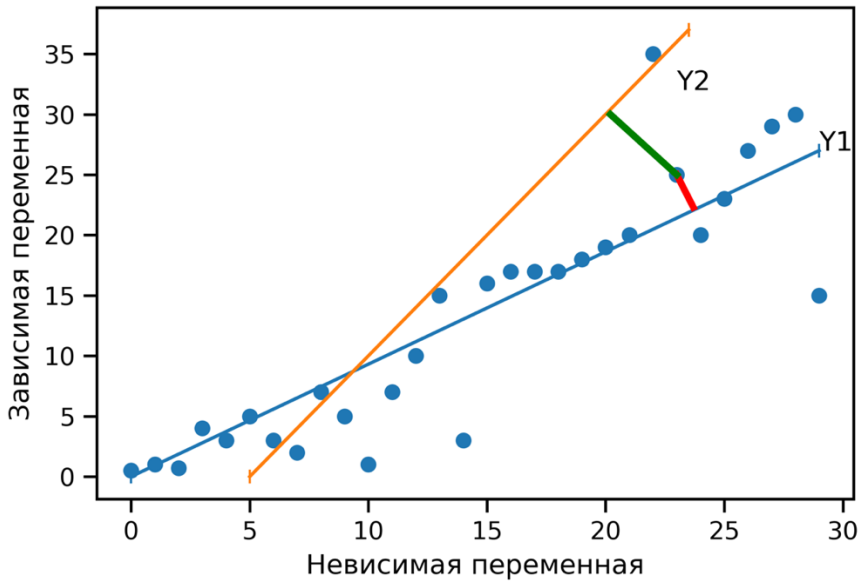

Рисунок 8. Общий принцип линейной регрессии

На данном графике изображены значения зависимой и независимой переменных. Даже по графику мы можем увидеть положительную взаимосвязь. Более того, вид

взаимосвязи приблизительно линейный. В таком случае мы можем построить прямую, которая будет описывать имеющуюся зависимость. И МНК позволяет построить наилучшую подобную прямую. На графике изображены две прямые Y_1 и Y_2. Также из графика очевидно, что прямая Y_1 описывает наши данные лучше. Но как подтвердить это вычислениями? Необходимо найти остатки от разности фактических значений и значений, которые соответствуют фактическим на прямой. Та прямая, у которой сумма квадратов таких остатков будет минимальной, и является наилучшей прямой нашей регрессионной модели. На графике данные остатки изображены красным и зеленым цветом. Объясним использование именно квадрата расстояния. Обозначим остатки как d_1 и d_2, а соответствующие точки на прямых как Y_{1d_1}, Y_{2d_2}. Выбранную точку обозначим как P. Тогда имеем:

$$d_1 = P - Y_{1d_1} \quad 2.3.50$$

$$d_2 = P - Y_{2d_2} \quad 2.3.51$$

В данном случае $d_1 > 0, d_2 < 0$. Для того, чтобы убрать отрицательные значения, которые будут искажать сумму остатков, и используется квадрат подобных разностей. В общем виде правило поиска по МНК можно записать так:

$$\sum_{i=1}^{n} (y_i - \widehat{y_i})^2, \quad 2.3.52$$

где $\widehat{y_i}$ – значение предполагаемой прямой в точке i. В случае линейной зависимости уместна система для поиска коэффициентов:

$$\begin{cases} nb_0 + \sum x = \sum y \\ b_0 \sum x + b_1 \sum x^2 = \sum yx \end{cases} \quad 2.3.53$$

Из этой системы выводятся стандартные зависимости для коэффициентов линейной регрессии:

$$\begin{cases} b_0 = M_y - b_1 M_x \\ b_1 = \dfrac{\sigma_y}{\sigma_x} R_{x,y} \end{cases} \quad 2.3.54$$

Значения МО, СКО и КК можно заменить точечными оценками по методам, ранее описанным в работе.

Рассмотрим более подробно коэффициент b_1 нашей модели. Данный коэффициент отражает наличие взаимодействия исследуемых переменных. В случае отсутствия

корреляции, b_1 будет нулевым, а линия регрессии – параллельна оси X. На основе этого можно сформировать гипотезы. Нулевой будет гипотеза о том, что переменные на самом деле никак не действуют друг на друга. Альтернативной гипотезой будет наличие взаимодействия:

$$H_0: b_1 = 0$$

$$H_1: b_1 \neq 0$$

Важным показателем качества модели ОЛР является коэффициент детерминации (КД) R^2. Он характеризует долю дисперсии зависимой переменной, которую объясняет наша регрессионная модель. Так как мы работаем с одной независимой переменной, то в данном случае можно сказать, что коэффициент детерминации показывает, насколько уверенно изменения зависимой переменной можно объяснить влиянием независимой. Чем больше значение данного коэффициента, тем лучше наша модель описывает и объясняет зависимость между переменными. Также с увеличением значения коэффициента детерминации растет и эффективность предсказания значений по нашей модели. Для вычисления КД используется зависимость вида:

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}, \quad 2.3.55$$

где $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ – сумма квадратов остатков, а $SS_{total} = \sum_{i=1}^{n}(y_i - M_y)^2 = n\sigma_y^2$ – общее значение изменчивости или общая сумма квадратов.

Для возможности использования полученной регрессионной модели значение КД должно быть не менее 0,5. При значениях КД выше 0,8 модель считается хорошей.

И еще одним глобальным показателем качества модели является F-значение, которое в данном случае вычисляется как:

$$F = \frac{R^2}{1 - R^2}\frac{df_2}{df_1}, \quad 2.3.56$$

где $df_1 = 1$ – количество независимых переменных. У нас такая только одна. А $df_2 = n - df_1 - 1$. А далее классический способ проверки. При заданном уровне значимости и найденных значениях числа степеней свободы находим F_{cr}. Если найденное значение критерия $F > F_{cr}$, то наша модель имеет статистическое подтверждение качества и значимости своей работы. Если была получена статистически значимая модель, можно решать задачу предсказания или

регулирования значений ЗП. Для этого просто необходимо подставить нужные значение независимой переменной в модель.

Поговорим о требованиях к данным, которые необходимо соблюдать для качественного использования ОЛР. Список требований ниже:

1. Линейный или почти линейный характер взаимосвязи X и Y.
2. Нормальное или близкое к нормальному распределение остатков.
3. Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной. Иными словами, при визуальном анализе графика должны наблюдаться случайные значения остатков/погрешностей, без их увеличения или уменьшения при увеличении значений независимой переменной.
4. Количественный тип исследуемых переменных.

В конце озвучим два замечания, которые важно осознавать при использовании ОЛР:

1. Установленная и исследуемая зависимость не подтверждает причинно-следственную связь между переменными. То есть даже при получении высокого КК и КД, при построении хорошей модели, мы не можем с полной уверенностью утверждать о неопровержимой зависимости между переменными. Данное утверждение возможно только после исследования сразу несколькими разными методами с получением положительных результатов, а также графического и логического анализа результатов.
2. Результаты нужно осторожно переносить на ГС, желательно с проверкой и другими методами.

Пример использования ОЛР для малой выборки приведен в третьем разделе и реализован программным способом.

И последней мы рассмотрим множественную линейную регрессию (МЛР). Она позволит нам исследовать влияние сразу нескольких независимых переменных на одну зависимую. С помощью нее мы сможем узнать, какие переменные имеют наибольшее влияние, а также получим возможность построить модель с оптимальным количеством переменных, дающую наилучший результат исследования. В основе МЛР лежит предположение о том, что зависимость между зависимой и независимыми переменными можно выразить линейно в виде следующей регрессионной модели:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_m \quad 2.3.57$$

В данном случае каждой независимой переменной соответствует свой коэффициент. Определять эти коэффициенты, называемые регрессионными, все так же будем по МНК. Но, из-за большего количества переменных, процедура нахождения коэффициентов усложняется. Будем работать с данными в матричном виде. Для этого определим вектор X, который состоит из значений зависимой переменной. Определим матрицу значений F, состоящую из векторов-столбцов независимых переменных. Так как при b_0 нет никакой переменной, ее значение примем равным 1. Общий вид данной матрицы:

$$F = (\{1\}\{x_1\}\{x_2\}\ldots\{x_m\}) \quad 2.3.58$$

Вектор неизвестных коэффициентов обозначим как A:

$$A = (b_0\ b_1\ b_2\ \ldots b_m) \quad 2.3.59$$

Используем классическую версию МНК:

$$A = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad 2.3.60$$

В векторно-матричной форме оценку вектора параметров находим по зависимости:

$$\bar{A} = (F^T F)^{-1} F^T X \quad 2.3.61$$

Можно также найти ковариационную матрицу, для понимания связи между независимыми переменными. Поиск производится через дисперсию вектора параметров:

$$\sigma_{\bar{A}}^2 = \frac{1}{n-m} (X - G\bar{A})^T (X - G\bar{A}) \quad 2.3.62$$

$$K_{\bar{A}} = \sigma_{\bar{A}}^2 (G^T - G)^{-1} \quad 2.3.63$$

Для проверки качества модели находим КД. После этого мы можем найти скорректированный КД (СКД), который используется при анализе нескольких переменных. Зависимость для него:

$$R_{adj}^2 = 1 - (1 - R^2) \frac{(n-1)}{(n-m)} \quad 2.3.64$$

Его использование обосновано тем, что стандартный КД будет постоянно расти при добавлении новых независимых переменных в модель. Это имеет негативные последствия, потому что в модели могут присутствовать переменные, «оказывающие» ложное влияние. Поэтому скорректированный КД вводит некий

штраф, который увеличивается с каждой новой добавленной переменной. В итоге, на некотором этапе мы столкнемся с тем, что данный штраф будет больше значимости влияния переменной и СКД начнет уменьшаться. Поэтому появляется задача найти максимальное значение СКД, модель, которая его обеспечит, будет лучшей для нас. В данном методе мы сначала строим модель, включая все переменные, и рассчитываем СКД. Далее поочередно удаляем по одной переменной и рассчитываем коэффициенты и СКД для моделей, состоящих из оставшихся переменных. Повторяем операцию до того момента, когда получим максимально возможное значение СКД. При удалении переменных также может проявиться значимость каких-либо оставшихся, которые ранее, на основе статистических вычислений, мы считали незначимыми.

Еще одним глобальным определением качества модели, как и в случае с ОЛР, является F-критерий. Но для модели множественной ЛР он рассчитывается несколько иначе. А именно, уместна зависимость:

$$F = \frac{SS_{fact}}{SS_{res}} \frac{n-m-1}{m}, \quad 2.3.65$$

где SS_{fact} – факторная сумма квадратов, показывающая отличия значений нашей модели от среднего значения зависимой переменной:

$$SS_{fact} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad 2.3.66$$

В данном случае число степеней свободы находятся так:

$$df_1 = m \quad 2.3.67$$

$$df_2 = n - m - 1 \quad 2.3.68$$

Далее задаем уровень значимости, по таблице F-распределения находим критическое значение F_{cr}. Если найденное значение критерия $F > F_{cr}$, то получено статистическое подтверждение адекватности и значимости построенной регрессионной модели. Если была получена статистически значимая модель, можно решать задачу предсказания или регулирования значений ЗП. Для этого просто необходимо подставить нужные значения независимых переменных в модель.

Множественная линейная регрессия имеет более обширные требования к данным:

1. Линейная зависимость переменных

2. Нормальное распределение остатков
3. Гомоскедастичность.
4. Отсутствие мультиколлинеарности – то есть наши независимые переменные должны быть незначительно связаны между собой. Проверяется по КК.
5. Желательно нормально распределение переменных

Подводя итог, озвучим несколько замечаний:

1. Количество переменных в модели не всегда означает ее качество.
2. Коэффициенты регрессионной модели МЛР показывают то, насколько каждая независимая переменная влияет на зависимую, при учете оставшихся НП.
3. Коэффициенты вычисляются не так, как в одномерном случае. В данном случае вычисления происходят с учетом того, что в модели несколько переменных, каждая из которых может иметь какое-то влияние.
4. Размер выборки не так важен, но строгое правило, что объем каждой выборки должен быть больше числа независимых переменных. Рекомендация – 20-25 значений в каждой, либо в 5-6 раз больше, чем количество НП.

Пример использования МЛР для малой выборки приведен в третьем разделе и реализован программным способом.

3. Программная реализация методов

В третьем разделе данной работы мы перейдем к реализации рассмотренных выше методов на языке Python в среде разработки Jupyter Notebook. При необходимости будут загружаться файлы с данными, которые будем исследовать нашими методами. Во время выполнения программной реализации активно использовалась документация [27-32] по библиотекам языка.

1. Методы точечного оценивания

Начнем с реализации методов точечного оценивания – ММП и ММ. Начнем с ММП. Так как во втором разделе были выведены зависимости для нормального распределения, получившиеся одинаковыми (с учетом коррекции ТО дисперсии по ММП), функция для оценки параметров нормального распределения будет единой. Также эту функцию можно использовать для вычисления выборочного МО и дисперсии у других распределений. Функция возвращает значения в зависимости от переданного параметра версии. Версия определяется на основе целей использования функции. Проверим результат работы, сгенерировав выборку из нормальной ГС с характеристиками $M = 2, \sigma = 1.44, n = 100$. Код функции и ее вывод:

```
def pe_norm(x, version):
    """
    Функция использует входящую выборку в виде списка или массива numpy
    Выбор версии определяет возвращаемые функцией значения
    Т.к. в работе были получены одни и те же зависимости для ТО МО норм
    распределения по ММП, ММ, МНК, то используем одну функцию
    """
    m = (1/len(x)) * sum(x)
    sum_sq = 0
    for i in x:
        sum_sq += (i - m) ** 2
    d_adj = sum_sq / (len(x) - 1)
    sigma_m = d_adj / len(x)
    sigma_d = 2 * ((d_adj) ** 2) / len(x)
    if version == 'both':
        return m, d_adj, sigma_m, sigma_d
    elif version == 'm_full':
        return m, sigma_m
    elif version == 'd_full':
        return d_adj, sigma_d
    elif version == 'm':
```

```
        return m
    elif version == 'd':
        return d_adj
    elif version == 'm_d':
        return m, d_adj
```

```
(2.149878096330244,
 2.1913370255052294,
 0.021913370255052293,
 0.09603915918700212)
```

Рисунок 9. Результаты точечного оценивания параметров нормальной малой выборки и погрешности этих оценок

Получена достаточно точная оценка для данной малой выборки.

Рассмотрим выборку из равномерной ГС с границами 1, 7 и размера 50, которую сгенерируем. Функция:

```
def pe_r_mmp(x):
    """
    Оцениваем границы, на вход выборка в виде массива numpy или списка
    Возвращаем оценки границ распределения
    """
    a = x.min()
    b = x.max()
    return a, b


x = stats.uniform.rvs(1, 7, 50)
pe_r_mmp(x)
```

Вывод:

```
(1.065531869481255, 7.886785816579985)
```

Рисунок 10. Результаты ТО параметров равномерной малой выборки

С небольшой погрешностью получили реальные границы ГС по малой выборке.

Далее рассмотрим оценку параметра экспоненциального распределения по ММ. Проверяем выборку из соответствующего примера второго раздела. Используем функцию:

```
def pe_exp_mm(x, method='m'):
    """
    Оцениваем параметр лямбда, на вход выборка в виде массива numpy
    и метод поиска
    Возвращаем оценки границ распределения
```

```
    """
    if method == 'd':
        pe_lambda = 1 / (x.var()) ** (1 / 2)
        sigma_l = (x.var() ** 2) / (n * 4 * x.var() ** 3)
    pe_lambda = 1 / x.mean()
    sigma_l = (x.var() / len(x)) / (x.mean() ** 4)
    return pe_lambda, sigma_l
```

Ее вывод:

```
(1.0976948408342482, 0.0025391681966473 51)
```

Рисунок 11. Результаты ТО экспоненциальной малой выборки и погрешность оценки

Получены те же самые результаты.

2. Методы интервального оценивания

Следующий пункт – реализация методов интервального оценивания. Первый метод – построение доверительного интервала для МО. Произведем оценку выборки из ГС с параметрами $M = 5\ \sigma = 1.44, n = 50$ по функции:

```
def ie_m_norm(x, p):
    """
    На вход выборка и доверительная вероятность
    Возвращает левую и правую границу ДИ
    """
    m, sigma = pe_norm(x, 'm_full')
    q = stats.t.ppf((1 + p)/2, len(x) - 1)
    m_l = m - q * sigma
    m_r = m + q * sigma
    print(' Доверительный интервал с вероятностью ', p, ' для МО: [',
          round(m_l, 4), ';', round(m_r, 4), ']')
    return m_l, m_r
```

```
Доверительный интервал с вероятностью  0.95  для МО: [ 4.8309 ; 5.0213 ]
```

Рисунок 12. ДИ для МО нормальной малой выборки

Получен достаточно точный ДИ. Функция для ИО дисперсии по той же выборке с общей вероятностью 0.9:

```
def ie_d_norm(x, p_l, p_r):
    """
    На вход выборка и вероятности для левой и правой границы
    Возвращает левую и правую границу ДИ
    """
    d = pe_norm(x, 'd')
```

```
    q_l = stats.chi2.ppf(p_l, len(x) - 1)
    q_r = stats.chi2.ppf(p_r, len(x) - 1)
    d_l = (len(x) * d) / q_r
    d_r = (len(x) * d) / q_l
    print('Доверительный интервал при общей вероятности ', round(p_r - p_l,
2), ' для дисперсии : [', round(d_l, 4), ';', round(d_r, 4), ']')
    return d_l, d_r
```

```
Доверительный интервал при общей вероятности  0.9  для Дисперсии : [ 1.455 ; 2.8447 ]
```

Рисунок 13. ДИ для дисперсии нормальной малой выборки

Получены более заметные расхождения, в сравнении с ДИ для МО.

Последний пункт интервального анализа – построение толерантного интервала. Построим толерантный интервал для выборки из ГС с распределением $Norm(5, 1.44), n = 50$, для доли 0.9 с вероятностью 0.98. Используемая функция и ее вывод:

```
def tolerant(x, p, g):
    """
    Вход - выборка, доверительная вероятность и доля совокупности ГС
    Возвращает границы интервала
    """
    m, std = pe_norm(x, 'm_d')
    std **= 1 / 2
    q = stats.norm.ppf(p)
    k = stats.norm.ppf(0.5 * (1 + g)) * (1 + q / ((2 * len(x)) ** (1 / 2)) +
(5 * (q ** 2) + 10) / (12 * len(x)))
    a = m - k * std
    b = m + k * std
    print('Толерантный интервал доли ', g, ' при доверительной вероятности ',
          p, ' : [', round(a, 4), ';', round(b, 4), ']')
```

```
Толерантный интервал доли  0.9  при доверительной вероятности  0.98  : [ 2.7137 ; 7.5823 ]
```

Рисунок 14. Толерантный интервал для нормальной малой выборки

С вероятностью 0.98 мы можем утверждать, что 90% значений в ГС попадают в полученный интервал.

3. Методы проверки статистических гипотез и комплексные методы

Переходим к самому обширному пункту – проверке различных гипотез. Первая – гипотеза о МО. Проверим гипотезу по все той же нормальной выборке, которую

использовали выше. Уровень значимости зададим равным 0.05. Проверим три гипотезы:

$$H_0: M_x = 5.5$$

$$H_0: M_x = 5$$

$$H_0: M_x = 4.9$$

Предполагаем, что первая и последняя будут опровергнуты, а вторая принята.

Функция для проверки гипотезы:

```
def hypothesis_mo(x, alpha, m0):
    """
    Вход - выборка, уровень значимости и предполагаемое значение МО
    Возвращает логический результат проверки M(x) = m0
    """
    m, sigma = pe_norm(x, 'm_full')
    z = (m - m0) / sigma ** (1 / 2)
    z_cr = stats.t.ppf(1 - alpha, len(x) - 1)
    if abs(z) <= z_cr:
        print('z =', round(z, 4), '≤ z_cr =', round(z_cr, 4) ,'\nПринимаем
гипотезу МО =',
              m0, 'с доверительной вероятностью ', 1 - alpha)
        print('Значение z, при принятой гипотезе, на самом деле может быть ',
              '\neще больше с вероятностью', round((1 - stats.t.cdf(abs(z),
len(x))), 4))
        return True
    else:
        print('Отклоняем гипотезу МО = ', m0)
        return False
```

Она возвращает следующие результаты:

```
Отклоняем гипотезу МО =  5.5
```

Рисунок 15. Проверка первой гипотезы о МО малой выборки

```
z = 0.8345 ≤ z_cr = 1.6766
Принимаем гипотезу МО = 5 с доверительной вероятностью  0.95
Значение z, при принятой гипотезе, на самом деле может быть
еще больше с вероятностью 0.204
```

Рисунок 16. Проверка второй гипотезы о МО малой выборки

```
Отклоняем гипотезу МО =  4.9
```

Рисунок 17. Проверка третьей гипотезы о МО малой выборки

Предположения оказались верны, проверка гипотез работает достаточно точно.

Следующая функция – сравнение МО в двух выборках при помощи гипотез об их равенстве. Сгенерируем две выборки из разных ГС: $x \sim Norm(5,2), n = 50; y \sim Norm(5,3), n = 50$. Проверим гипотезу о равенстве МО с помощью функции:

```
def hypothesis_double_mo(x, y, alpha):
    """
    Вход - две выборки, уровень значимости
    Возвращает логический результат проверки M(x) = M(y)
    """
    m_x, d_x = pe_norm(x, 'm_d')
    m_y, d_y = pe_norm(y, 'm_d')
    t = (abs(m_x - m_y)) / (d_x / len(x) + d_y / len(y)) ** (1 / 2)
    t_cr = stats.t.ppf(1 - alpha, len(x) + len(y) - 2)
    if abs(t) <= t_cr:
        print('t =', round(t, 4), '≤ t_cr =', round(t_cr, 4) ,'\nПринимаем
гипотезу M(X) = M(Y) с доверительной вероятностью ',
              1 - alpha)
        return True
    else:
        print('t =', round(t, 4), '> t_cr =', round(t_cr, 4) ,'\nОтклоняем
гипотезу M(X) = M(Y)')
        return False
```

```
t = 0.0186 ≤ t_cr = 1.6606
Принимаем гипотезу M(X) = M(Y) с доверительной вероятностью  0.95
```

Рисунок 18. Проверка гипотезы о равенстве МО по двум малым выборкам (заведомо верная)

Изменим значение МО второй выборки на 4.9 и выполним еще один тест:

```
t = 1.8259 > t_cr = 1.6606
Отклоняем гипотезу M(X) = M(Y)
```

Рисунок 19. Проверка гипотезы о равенстве МО по двум малым выборкам (заведомо ложная)

Гипотеза отклонена, что соответствует действительности.

Далее займемся проверкой двух распределений на однородность. Проверим работу на трех малых выборках: двух из одинаковых нормальных ГС и одной из равномерной. Сначала сравним нормальные ($Norm(5,3)$) между собой, далее одну из них с равномерной ($R(1,10)$). Уровень значимости установим 0.05. Код функции:

```
def hypothesis_same_distr(x, y, value_name, alpha):
    """
```

```
    Вход - две выборки, имя переменной (должно быть одинаковым в обеих
выборках), уровень значимости
    Возвращает логический результат проверки f(x) = f(y)
    """
    x['own'] = 'x'
    y['own'] = 'y'
    df = pd.concat([x, y], ignore_index=True)
    df = df.sort_values(value_name)
    df['rank'] = [i for i in range(1, len(df.value) + 1)]
    u_x = df[df.own == 'x']['rank'].sum()
    u_y = df[df.own == 'y']['rank'].sum()
    n_x = len(df[df.own == 'x'])
    n_y = len(df[df.own == 'y'])
    m_u_x = 0.5 * n_x * (n_x + n_y + 1)
    m_u_y = 0.5 * n_y * (n_x + n_y + 1)
    sigma_u = (1 / 12) * n_x * n_y * (n_x + n_y + 1)
    q = stats.norm.ppf(1 - alpha / 2)
    u_x_l = m_u_x - q * sigma_u ** (1 / 2)
    u_x_r = m_u_x + q * sigma_u ** (1 / 2)
    u_y_l = m_u_y - q * sigma_u ** (1 / 2)
    u_y_r = m_u_y + q * sigma_u ** (1 / 2)
    if (u_x > u_x_l and u_x < u_x_r and u_y > u_y_l and u_y < u_y_r):
        print('U_x =', u_x, ', Доверительный интервал [', round(u_x_l, 4),
';', round(u_x_r, 4), ']')
        print('U_y =', u_y, ', Доверительный интервал [', round(u_y_l, 4),
';', round(u_y_r, 4), ']')
```

Результат работы:

```
U_x = 2426 , Доверительный интервал [ 2286.4017 ; 2763.5983 ]
U_y = 2624 , Доверительный интервал [ 2286.4017 ; 2763.5983 ]
Принимаем гипотезу f(x) = f(y) с доверительной вероятностью 0.9
U_x = 2279 , Доверительный интервал [ 2286.4017 ; 2763.5983 ]
U_y = 2771 , Доверительный интервал [ 2286.4017 ; 2763.5983 ]
Отклоняем гипотезу f(x) = f(y)
```

Рисунок 20. Проверка гипотезы об однородности распределений по двум малым выборкам

Видим, что метод позволил нам отклонить гипотезу при действительно разных распределениях. При одинаковых распределениях проверка проходит успешно, значения попадают в доверительные интервалы. Результаты получены при достаточно хорошем уровне значимости, $\alpha = 0.05$ – стандарт для многих методов и исследований, не требующих «хирургической» точности.

Многие из прошлых методов требовали знания типа распределения ГС. Далее будут приведены реализации методов проверки гипотез о типе распределения. Использовать будем три – нормальное, экспоненциальное и равномерное. Рассмотрим модификации критериев χ^2 и ω^2. Код для классических версий приведен в приложениях 4 и 5. Функция для модификации критерия χ^2 требует предварительного вызова функции для классического метода. Для его реализации была использована таблица значений интегральной функции Лапласа. Дело в том, что критическое значение определяется как аргумент от этой функции, а не как результат. Модули Python не позволяют выполнить такие вычисления. Поэтому было принято решение использовать таблицу. Код функции:

```
table_laplace = pd.read_csv('laplace.csv')
def chi_sq_adj(x, alpha, n, p_n, p_e, p_r, k):
    types = ['Norm', 'Exp', 'R']
    n_len = [len(i) for i in n]
    m_n, m_e, m_r = [], [], []
    for i in range(k):
        m_n.append(sum(n_len) * p_n[i])
        m_e.append(sum(n_len) * p_e[i])
        m_r.append(sum(n_len) * p_r[i])
    sigma_n, sigma_e, sigma_r = [], [], []
    for i in range(k):
        sigma_n.append((sum(n_len) * p_n[i] * (1 - p_n[i])) ** (1 / 2))
        sigma_e.append((sum(n_len) * p_e[i] * (1 - p_e[i])) ** (1 / 2))
        sigma_r.append((sum(n_len) * p_r[i] * (1 - p_r[i])) ** (1 / 2))
    v_n, v_e, v_r = [], [], []
    for i in range(k):
        v_n.append(((n_len[i] - m_n[i]) / sigma_n[i]) ** 2)
        v_e.append(((n_len[i] - m_e[i]) / sigma_e[i]) ** 2)
        v_r.append(((n_len[i] - m_r[i]) / sigma_r[i]) ** 2)
    v_n, v_e, v_r = (sum(v_n) / k) ** (1 / 2), (sum(v_e) / k) ** (1 / 2),
(sum(v_r) / k) ** (1 / 2)
    s = [((2 * k) ** (1 / 2)) * (v_n - 1), ((2 * k) ** (1 / 2)) * (v_e - 1),
((2 * k) ** (1 / 2)) * (v_r - 1)]
    f = 0.5 - alpha
    s_cr = table_laplace[table_laplace.F < f].max()['x']
    s_plus = 0
    for i in range(3):
        if s[i] <= s_cr:
```

```
            print(types[i], ': S =', round(s[i], 4), '≤ S_cr =', round(s_cr,
4) ,'\nПринимаем гипотезу f(x) ~', types[i], 'с доверительной вероятностью ',
1 - alpha)
        else:
            print('S =', round(s[i], 4), '> S_cr =', round(s_cr, 4)
,'\nОтклоняем гипотезу f(x) ~ ', types[i])
```

Модифицированные критерии проверяем на трех выборках: $Norm(5, 2)$, $Exp(0.09)$, $R(1, 5)$. Размер каждой выборки равен 50 элементам. Начнем проверку с нормального распределения:

```
Passed
S = -0.4012 ≤ S_cr = 1.28
Принимаем гипотезу f(x) ~ Norm с доверительной вероятностью  0.9
S = 6.9198 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  Exp
S = 7.3246 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  R
```

Рисунок 21. Тест нормальной малой выборки

Тип распределения определен корректно. Далее проверим экспоненциальное распределение:

```
Passed
S = 3.1218 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  Norm
S = -0.3259 ≤ S_cr = 1.28
Принимаем гипотезу f(x) ~ Exp с доверительной вероятностью  0.9
S = 3.8319 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  R
```

Рисунок 22. Тест экспоненциальной малой выборки

Критерий также однозначно определил верную гипотезу, работоспособность проверена. Последний проверяемый тип распределения – равномерный:

```
Passed
S = 1.7524 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  Norm
S = 8.9687 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  Exp
S = 1.5536 > S_cr = 1.28
Отклоняем гипотезу f(x) ~  R
```

Рисунок 23. Тест равномерной малой выборки

В случае равномерного распределения появились проблемы. Все три гипотезы были отклонены, поэтому в данной ситуации применять метод нельзя.

Перейдем к методу модифицированного критерия ω^2. Для вычислений нам понадобится еще одна таблица критических значений (*приложение 6*). Внутри основной функции вызываем базовую для данного метода. Код функции:

```
w_adj_table = pd.read_csv('w_adj.csv')
```

```
def kramer_test_adj(x, alpha):
    types = ['Norm', 'Exp', 'R']
    n = len(x)
    w_n, w_e, w_p = kramer_test(x, alpha, 'for_adj')
    w = [(1 / (12 * n) + w_n) ** (1 / 2), (1 / (12 * n) + w_e) ** (1 / 2), (1
/ (12 * n) + w_p) ** (1 / 2)]
    w_cr = w_adj_table[w_adj_table.p == 1 - alpha].values[0, 1]
    for i in range(3):
        if w[i] <= w_cr:
            print('G =', round(w[i], 4), '≤ G_cr =', round(w_cr, 4)
,'\nПринимаем гипотезу f(x) ~', types[i], 'с доверительной вероятностью ', 1
- alpha)
        else:
            print('G =', round(w[i], 4), '> G_cr =', round(w_cr, 4)
,'\nОтклоняем гипотезу f(x) ~', types[i])
```

Проверяем нормальное распределение:

```
G = 0.2442 ≤ G_cr = 0.316
Принимаем гипотезу f(x) ~ Norm с доверительной вероятностью  0.9
G = 1.2389 > G_cr = 0.316
Отклоняем гипотезу f(x) ~ Exp
G = 0.5763 > G_cr = 0.316
Отклоняем гипотезу f(x) ~ R
```

Рисунок 24. Результаты теста нормальной малой выборки

Проверка пройдена, верная гипотеза принята. Экспоненциальный случай:

```
G = 0.5681 > G_cr = 0.316
Отклоняем гипотезу f(x) ~ Norm
G = 0.1742 ≤ G_cr = 0.316
Принимаем гипотезу f(x) ~ Exp с доверительной вероятностью  0.9
G = 0.8028 > G_cr = 0.316
Отклоняем гипотезу f(x) ~ R
```

Рисунок 25. Тест критерия по экспоненциальной малой выборке

Получаем подтверждение корректной работы и в данном случае. Равномерное распределение:

```
G = 0.4087 > G_cr = 0.316
Отклоняем гипотезу f(x) ~ Norm
G = 1.1902 > G_cr = 0.316
Отклоняем гипотезу f(x) ~ Exp
G = 0.2406 ≤ G_cr = 0.316
Принимаем гипотезу f(x) ~ R с доверительной вероятностью  0.9
```

Рисунок 26. Тест критерия по равномерной малой выборке

В отличие от прошлого метода, данный метод позволяет корректно проверить гипотезу о равномерном распределении. Будем считать его наилучшим в данном случае.

Последний метод проверки гипотез – критерий Шапиро-Уилка, с помощью которого проверяют распределение на нормальность. Код функции, определяющей критическое значение, а также код основной функции:

```
def find_critical(x, alpha):
    n = len(x)
    if alpha == 0.1:
        return (-0.0084 * n ** 4 + 1.2513 * n ** 3 - 70.724 * n ** 2 + 1890 *
n + 73840)  / 100000
    elif alpha == 0.05:
        return (-0.0113 * n ** 4 + 1.656 * n ** 3 - 91.88 * n ** 2 + 2408.6 *
n + 67608) / 100000
    elif alpha == 0.01:
        return (-0.0148 * n ** 4 + 2.1875 * n ** 3 - 122.61 * n ** 2 + 3257.3
* n + 55585) / 100000

def shapiro_test(x, alpha):
    x = np.sort(x)
    n = len(x)
    z = []
    a0 = 0.899 / ((n - 2.4) ** 0.4162) - 0.02
    a = []
    for i in range(n):
        z.append((n - 2 * (i + 1) + 1) / (n - 0.5))
        a.append(a0 * (z[i] + 1483 / ((3 - z[i]) ** 10.845) + (71.6 * 10 **
(-10)) / ((1.1 - z[i]) ** 8.26)))
    mean, sigma = pe_norm(x, 'm_d')
    w = []
    for i in range(n):
        w.append(a[n - i - 1]*(x[n-i - 1] - x[i]))
    w = sum(w)
    sigma = [(i - mean) ** 2 for i in x]
    sigma = sum(sigma)
    w **= 2
    w /= sigma
    w_cr = find_critical(x, alpha)
    if w > w_cr:
        print('W =', round(w, 4), '> W_cr =', round(w_cr, 4) ,'\nПринимаем
гипотезу f(x) ~ Norm с доверительной вероятностью ', 1 - alpha)
        return True
    else:
        print('W =', round(w, 4), '≤ W_cr =', round(w_cr, 4) ,'\nОтклоняем
гипотезу f(x) ~ Norm')
```

Сначала проверим его работу на малой нормальной выборке с теми же параметрами, размер которой меньше, чем в прошлых методах (пусть будет 30 элементов). Это вызвано ограничениями по максимальному размеру выборки для данного метода:

```
W = 2.6696 > W_cr = 0.9387
Принимаем гипотезу f(x) ~ Norm с доверительной вероятностью  0.9
```

Рисунок 27. Тест критерия нормальности по малой выборке

Имеется достаточно большое различие экспериментального и критического значения, что для нас является положительным моментом.

В конце раздела рассмотрим реализацию более комплексных методов, которые будем рассматривать на реальных примерах. Полученные результаты дадут

основания делать уже более осмысленные и сложные выводы, чем методы из прошлых пунктов. Первый метод – регрессионный анализ. Рассмотрим однофакторный случай, далее перейдем к многофакторному. Рассмотрим пример клинических испытаний с использованием четырех разных видов терапии. Каждый вид образует группу из 15 пациентов и обозначается как *a, b, c, d* соответственно. Для каждого пациента имеется значение исследуемого показателя. Наша задача – проверить гипотезу о том, что результаты терапии имеют статистически значимые различия, для этого должны статистически значимо отличаться МО ГС, из которых получены выборки для каждой терапии:

$$H_0: M(a) = M(b) = M(c) = M(d)$$

$$H_1: \text{Хотя бы одно среднее отличается}$$

Хотим получить графики попарного сравнения групп, F и p значения. Код функции:

```
def one_way_anova(df, group, value):
    """
    Аргументы: df - выборка, имеющая разделение на группы,
    group - столбец, определяющий группу
    value - столбец зависимых значений

    Производится сравнение нескольких групп между собой.
    H0 - средние в группах не отличаются
    H1 - хотя бы одно среднее значительно отличается

    Вывод: графики, f-value, p-value(т.е. P(>f))
    """
    df_bg = len(df[group].unique()) - 1
    df_wg = len(df[group]) - df_bg - 1



    group_sizes = []
    groups = df[group].unique()
    for gr in groups:
        group_sizes.append(df[df[group] == gr])
    for gr in group_sizes:
        gr.index = range(0, len(gr))
    means = []
    for gr in group_sizes:
        means.append(gr.mean())
    average = sum(means) / len(means)
    ssb = 0
```

```
    for i in range(len(means)):
        ssb += (means[i].value - average) ** 2 * len(group_sizes[i])
    ms_bg = ssb / df_bg
    ssw = 0
    for i in range(len(group_sizes)):
        for j in range(len(group_sizes[i])):
            ssw += (group_sizes[i].value[j] - means[i]) ** 2
    ms_wg = ssw / df_wg
    f_value = ms_bg / ms_wg
    p_value = 1 - stats.f.cdf(f_value, df_bg, df_wg)[0]
    fig, ax = plt.subplots(len(groups) - 1, len(groups) - 1)
    list_df = groups
    print('F =', f_value[0], ', p-value =', p_value)
    for i in range(len(list_df) - 1):
        for j in range(i + 1, len(list_df)):
            df_test1 = df[df.group == list_df[i]]
            df_test2 = df[df.group == list_df[j]]
            ax[i][j - 1].boxplot([df_test1.value, df_test2.value])
            ax[i][j - 1].set_xticklabels([list_df[i],list_df[j]])
    fig.set_size_inches([11, 11])
    fig.savefig('one-way.png')
    return f_value[0], p_value
```

Используем ее для анализа нашего набора данных:

```
F = 6.35639456386063 , p-value = 0.0008731722558327215
```

Рисунок 28. Полученные f-value и p-value

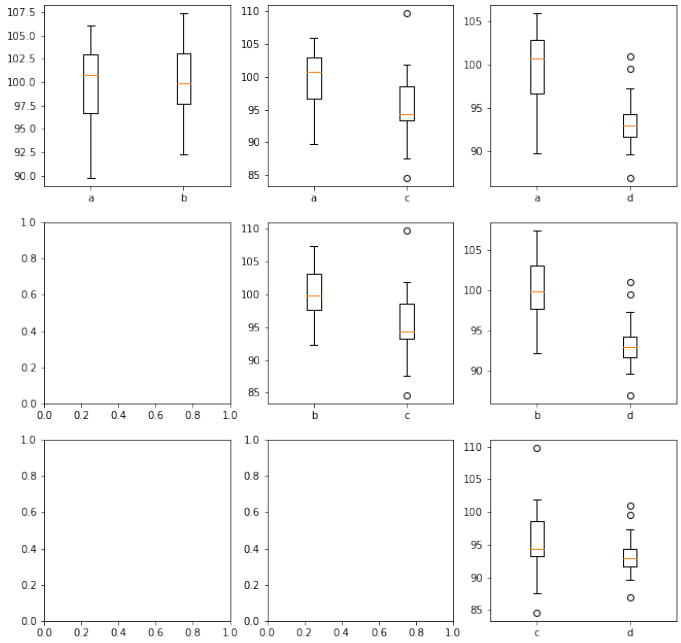

Рисунок 29. Графическое сравнение различных уровней фактора

Однофакторный дисперсионный анализ, в случае работы с тремя и более уровнями фактора, не дает точного ответа на вопрос о том, какие именно уровни (группы) отличаются. Он позволяет только установить наличие статистического отличия. Рассуждать о конкретных отличиях можно, например, по графикам. В нашем случае видим достаточно сильные отличия группы *a* от *d* и *b* от *d*. Дальнейшую проверку предположений можно произвести по t-критерию или другим методам парного сравнения.

Случай многофактороного ДА рассмотрим на примере реализации конкретной задачи, а не функции, как было до этого. В этот раз будем использовать другой набор данных, а именно результаты других клинических испытаний, где молодые и взрослые люди тестировали препарат с различной дозировкой действующего вещества – высокой и низкой. Таким образом, имеется два фактора – возраст и дозировка, и предполагаемое зависимое значение исследуемого показателя. В каждой группе 15 человек. Условия применения соблюдены. Проверяем как воздействие факторов по-отдельности, так и в совокупности.

Код, реализующий данный метод на нашем наборе данных приведен в приложении 6. Это обосновано слишком большим объемом. Результат его работы:

```
F_age = 5.368082124537378 p_age = 0.02394
F_dose = 2.950426806296753 p_dose = 0.09101
F_agg = 8.69154591528831 p_agg 0.00455
```

Рисунок 30. f-value и p-value для каждого фактора и межфакторного взаимодействия

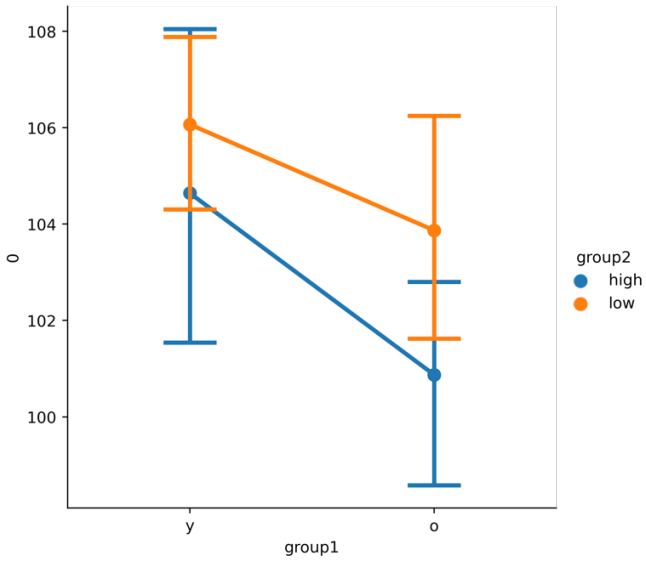

Рисунок 31. Графическое представление результатов работы метода

В данном исследовании мы принимаем различия статистически значимыми, если p-значение не превышает 0.05. Поэтому делаем вывод, что на результаты исследований влияет возраст и взаимодействия фактора возраста и дозировки, причем, согласно полученным значениям, взаимодействие факторов оказывает значительно большее влияние, чем фактор возраста в отдельности. Полученные результаты можно интерпретировать так: исследуемый показатель зависит от возраста, у молодых людей его значение выше. Но более сильная зависимость от совместного влияния возраста и дозировки – у молодых людей с высокой дозировкой показатель почти равен показателю у взрослых с высокой дозировкой.

Последний оставшийся метод – регрессионный анализ, если точнее, то одномерная и многомерная линейная регрессия. Наша функция будет работать в обоих случаях. Данные, с которыми мы будем работать, представляют собой социальную статистику в США. А именно средние значения бедности (как зависимой переменной) и процента городского населения, процент белокожего населения, процент людей с высшим образованием и процент семей, управляемых женщинами (независимые переменные). В одномерном случае рассмотрим зависимость только от процента людей с высшим образованием. В наборе данных 51 штат, включая Вашингтон (округ Колумбия) как отдельный штат. Перейдем к функции. Если быть точнее, их будет две. Одна будет строить модель, а вторая – проверять ее по критериям качества, озвученным во втором разделе работы. Рассмотрим одномерный случай. Проверим зависимость уровня бедности от процента людей с высшим образованием при уровне значимости 0.1. Код:

```
def mult_linregr(df, groups, x):
    x = np.array(df[x])
    intercept = np.array([1 for i in range(len(df))])
    coeff = []
    for i in groups:
        coeff.append(df[i])
    g_t = [intercept]
    for gr in coeff:
        g_t.append(gr)
    g_t = np.array(g_t)
    g = g_t.transpose()
    a = list(map(lambda x: round(x, 3),
                 (np.linalg.matrix_power(g_t.dot(g), -1).dot(g_t)).dot(x)))
    return a, groups
```

```
def test_model(df, x, model, group, alpha):
    x = np.array(df[x])
    model_values = []
    for i in range(len(df)):
        summ = model[0]
        for j in range(1, len(model)):
            summ += model[j] * df[group[j - 1]][i]
        model_values.append(summ)
    ssres = []
    for i in range(len(df)):
        ssres.append((model_values[i]-x[i])**2)
    ssres = sum(ssres)
    sstotal = []
    for i in range(len(df)):
        sstotal.append((x[i] - x.mean())**2)
    sstotal = sum(sstotal)
    r2 = 1 - ssres/sstotal
    r2_adj = 1 - (1 - r2) * (len(df) - 1) / (len(df) - len(group))
    ssfac = []
    for i in range(len(df)):
        ssfac.append((model_values[i] - x.mean())**2)
    ssfac = sum(ssfac)
    f_value = ssfac * (len(df) - len(group) - 1) / (ssres * len(group))
    f_cr = stats.f.ppf(1 - alpha, len(group), len(df) - len(group) - 1)
    final_model = 'y_model = ' + str(model[0]) + ' +'
    for i in range(1, len(model)):
        final_model += ' (' + str(model[i]) + group[i-1] + ') +'
    final_model = final_model[:len(final_model) - 2]
    if (f_value > f_cr) and (r2_adj > 0.5):
        print(final_model)
        print('F =', round(f_value, 4), '> F_cr =', round(f_cr, 4),
              ': passed')
        print('R_2adj =', round(r2_adj, 4), ': passed')
    else:
        print('Not passed')
        print('F =', round(f_value, 4), ', F_cr =', round(f_cr, 4))
        print('R_2adj =', round(r2_adj, 4))
```

Полученные результаты:

```
y_model = 64.781 + (-0.621hs_grad)
F = 61.7647 > F_cr = 2.8108 : passed
R_2adj = 0.5578 : passed
```

Рисунок 32. Оптимальная модель зависимости ЗП и одной НП

Так как модель прошла проверку на качество, мы можем построить график:

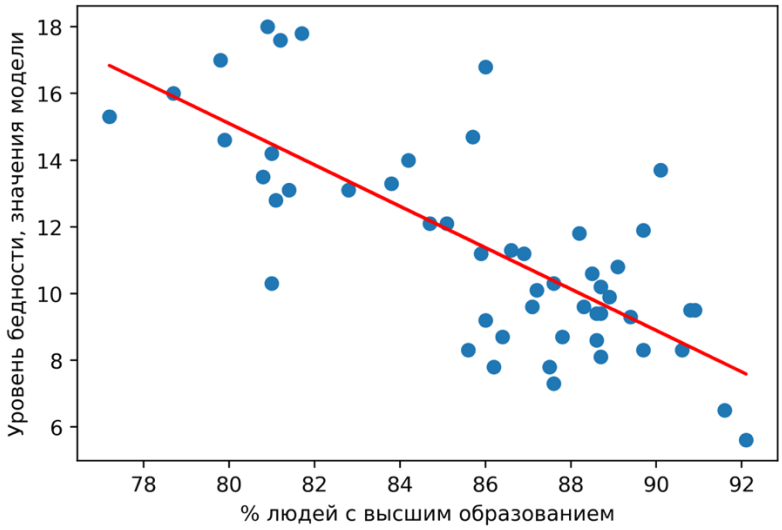

Рисунок 33. Фактические данные и аппроксимирующая модельная прямая

Но более правильным будет рассмотреть регрессионный анализ, включающий все имеющиеся предикторы. Это задача множественной линейной регрессии. В данном случае значения аппроксимируются плоскостью, поэтому график построить не удастся. Но зато мы сможем определить наиболее качественную модель из всех возможных. Но для начала проведем проверку на мультиколлинеарность, то есть сравним значения КК между нашими предикторами:

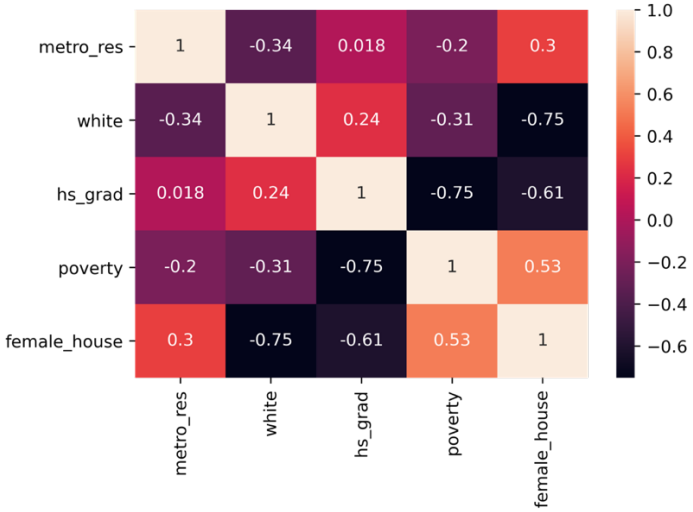

Рисунок 34. Проверка данных на мультиколлинеарность

Наблюдаем достаточно сильную корреляцию между переменной female_house и переменными white и hs_grad. Поэтому female_house – первая переменная-кандидат на проверку. Остальные значения в допустимых пределах. Функцию используем ту же, в качестве аргумента-списка предикторов передаем все переменные:

```
y_model = 66.477 + (-0.056metro_res) + (-0.048white) + (-0.555hs_grad) + (0.051female_house)
F = 20.5787 > F_cr = 2.0712 : passed
R_2adj = 0.6187 : passed
```

Рисунок 35. Первичная модель множественной линейной регрессии

Видим, что дополнительные переменные улучшили качество модели. Теперь будем исключать переменные и проверять, увеличится ли значения коэффициента детерминации. Первой исключим из модели переменную female_house:

```
y_model = 68.722 + (-0.056metro_res) + (-0.053white) + (-0.57hs_grad)
F = 28.0217 > F_cr = 2.2042 : passed
R_2adj = 0.6262 : passed
```

Рисунок 36. Повторный тест первичной модели с удалением переменной female_house

Коэффициент детерминации увеличился. Проведем еще несколько проверок, передавая функции различные списки проверяемых предикторов. В результате дальнейшей проверки увеличения значения КД не получили. Вывод – лучшая модель была получена после удаления переменной female_house. Ее можно использовать для дальнейших исследований и предсказаний.

Заключение

В данной работе рассматривался статистический анализ выборок малого объема. Изучались и сравнивались статистические методы, которые помогают в решении задач в условиях ограниченного количества экспериментальных или исследовательских данных. Вопрос малых выборок затрагивается в достаточно небольшом количестве публикаций и научных трудов, которые содержат достаточно обширную статистическую теорию, но не всегда полноценно охватывают вопросы прикладного характера. В это же время проблема малых выборок часто встречается именно на практике. Также часто невозможно или не целесообразно проводить дополнительные исследования для увеличения объема получаемой информации. Поэтому в таких ситуациях необходимо использовать специальные методы, которые даже по ограниченной выборке дают адекватные и близкие к истине результаты. В данной работе исследуются малые выборки, объем которых не превышает 200 элементов. Из многочисленных статистических методов и критериев были выбраны самые основные и важные, которые позволяют решать обширный список задач. Была произведена попытка сохранения баланса между теоретическим обоснованием и практическими аспектами. Каждый выбранный метод был реализован программным образом и протестирован на различных входных данных. Полученные результаты позволили включить их в работу.

Главной задачей работы был поиск и выбор оптимальных методов для работы с малыми выборками установленного объема. Изучая различные источники, задачу удалось разделить на три подзадачи.

Первым пунктом исследования стали методы, позволяющие проводить точечное оценивание, то есть по малой выборке получать возможные значения какого-либо параметра или какой-либо характеристики ГС, из которой выборка была извлечена. Были рассмотрены три метода точечного оценивания: максимального правдоподобия, моментов и наименьших квадратов. В результате исследования теории мы выяснили, что для нормального распределения все методы дают одинаковые оценки параметров (с учетом коррекции в ММП), в случае экспоненциального распределения теоретически и на практике лучшие результаты показал метод моментов, однако в случае равномерного – метод максимального правдоподобия.

Вторая подзадача заключалась в интервальном оценивании. Она описана кратко на примере построения доверительных интервалов для математического ожидания и дисперсии. Главное здесь было отразить принцип построения ДИ, который повторяется из метода в метод, сменяя только критерии работы и используемые статистики. Стоит еще раз упомянуть, что для малых выборок интервальное оценивание более предпочтительное, чем точечное. Поэтому желательно использовать оба типа оценивания в исследованиях, чтобы была возможность сравнить результаты и сделать какие-либо полезные выводы.

Третья, самая обширная и главная подзадача – проверка гипотез и взаимосвязей в данных. Этой теме посвящена почти половина теоретического материала, а обосновано это исключительной полезностью, гибкостью и универсальностью, которую обеспечивают рассматриваемые методы. Наиболее подробно были рассмотрены методы проверки гипотез о типах распределения и, так называемые, комплексные методы или методы определения взаимосвязей. Случай проверки гипотез о типе распределения объясняется тем, что многие статистические методы требуют знания типа распределения. В то же время, зная тип распределения, можно избежать использования некоторых методов, ограничившись базовыми зависимостями для этого типа и его свойствами. Поэтому в работе присутствует целых пять методов проверки подобных гипотез. Два из них хорошо работают с выборками, объем которых не менее 50 элементов, а их модификации можно успешно применять и с более малыми выборками, также приведен один из мощнейших критериев проверки нормальности распределения по малой выборке. Совместное использование нескольких методов может дать достаточно точный ответ о типе распределения по имеющейся малой выборке. Огромным плюсом данных методов является то, что в каждом из них мы имеем численное выражение вероятности того, что наши предположения верны.

Для исследования зависимостей в данных было предложено два мощных и полезных метода – дисперсионный анализ и линейная регрессия. ДА позволил нам проверять влияние различных градаций качественной и независимой переменной на предполагаемую зависимую количественную. Мы рассмотрели случай изучения влияния как одного, так и совокупности факторов. ЛР позволила определить тип и характер зависимости между количественными переменными. Был приведен как

одномерный, так и множественный случай. Также данный метод позволяет выполнять задачи построения оптимальной модели зависимости и прогнозирования значений. Каждый из этих методов по-своему полезен и может быть качественно применен к выборкам установленного в работе объема.

Заключительная часть работы посвящена реализации выбранных методов на языке Python. В работе использование готовых статистических функций или моделей намеренно сведено к минимуму, оставляя лишь функции и методы для создания визуализаций и генерации выборок, что к нашим задачам имеет лишь косвенное отношение. Был продемонстрирован оригинальный код каждого рассматриваемого метода и результат его выполнения. В целом почти все методы показали достойные результаты в случае работы с нормальным, экспоненциальным и равномерным распределением.

Подводя итог хочется еще раз подчеркнуть, что в работе собрана лишь часть методов, применяемых в работе с малыми выборками. Методы, попавшие в работу, показались автору наиболее полезными, а моментами интересными и необычными. Тем не менее, получилось рассмотреть достаточно обширный список тем, предложить несколько решений в некоторых задачах, реализовать, протестировать и проверить правильность включения методов в работу. Хочется верить, что данная работа, как и множество других, послужит хорошим примером взгляда отдельного человека на проблему статистического анализа малых выборок и, вероятно, внесет какие-либо улучшения в практический аспект данного вопроса.

Список используемых источников

Русскоязычные источники

1. Володин Н. И. Теория вероятностей и ее применения, – Казань, 1967
2. Гаскаров Д. В, Шаповалов В. И. Малая выборка. – М.: Статистика, 1978
3. ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения.
4. Домбровский В. В. Эконометрика. – Томск, НФПК, 2016
5. Казакявичюс К.А. Приближённые формулы для статистической обработки результатов механических испытаний. – Заводская лаборатория, 1988, т. 5, № 12, с. 82-85
6. Ковалев Е. А., Медведев Г. А. Теория вероятностей и математическая статистика для экономистов. – М.: Юрайт, 2016
7. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. – М.: Наука, 1968
8. Лебедев А. В., Фадеева Л. Н. Теория вероятностей и математическая статистика. Изд. 4-е, перераб. и доп. – М., 2018
9. Петров А. А. Проверка гипотезы о нормальности распределений по малым выборкам. ДАН, 1951, т. 76, № 3, с. 355-358
10. Петров А. А. Проверка гипотезы о типе распределения по данным малых выборок. – В кн.: Сборник научных работ кафедры математики МИФИ, вып. 1. М., Атомиздат, 1958, с. 121-136
11. Петров А. А. Проверка статистических гипотез о типе распределения по малым выборкам. – Теория вероятностей и ее применения, 1956, т. 1, № 2, с. 248-270
12. Прохорова Ю. В. Вероятность и математическая статистика: Энциклопедия. – М.: Большая Российская энциклопедия, 2003
13. Пугачев В. С. Теория вероятностей и математическая статистика. – М.: Физика, 2002
14. Рао С. Р. Линейные статистические методы и их применения. – М.: Наука, 1968

15. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. – М.: Наука, 1965
16. Сухорученков Б. И. Анализ малой выборки. Прикладные статистические методы. – М.: Вузовская книга, 2010
17. Уилкс С. Математическая статистика. – М: Наука, 1967
18. Юденков В. А. Дисперсионный анализ. – Минск: Бизнесофсет, 2013

Источники на других языках

19. J. Greene, M. D'Oliveira. Learning to Use Statistical Tests in Psychology. McGraw-Hill International, Berkshire, 2005
20. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. – Giornale dell' Istituto Italiano degli Attuari, 1933, N 4
21. A.K. Kurtz, S.T. Mayo (1979, p.417)
22. J. J. McCall. The Quarterly Journal of Economics, Vol. 84, No. 1 (Feb., 1970), pp. 113-126
23. J. Welkowitz. Introductory Statistics for the Behavioral Sciences, 2006

Интернет-ресурсы

24. https://ru.wikipedia.org
25. https://en.wikipedia.org
26. http://www.machinelearning.ru
27. https://pandas.pydata.org
28. https://www.python.org/doc
29. https://seaborn.pydata.org
30. https://numpy.org
31. https://matplotlib.org/3.5.0/api/_as_gen/matplotlib.pyplot
32. https://scipy.org

Приложения

1. Таблица коэффициентов для критерия Шапиро-Уилка

n	*i*									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	7071									
4	6872	1677								
5	6646	2413								
6	6431	2806	0875							
7	6233	3031	1401							
8	6052	3164	1743	0561						
9	5888	3244	1976	0947						
10	5739	3291	2141	1224	0399					
11	5601	3315	2260	1429	0695					
12	5475	3325	2347	1586	0922	0303				
13	5359	3325	2412	1707	1099	0539				
14	5251	3318	2460	1802	1240	0727	0240			
15	5150	3306	2495	1878	1353	0880	0433			
16	5056	3290	2521	1939	1447	1005	0593	0196		
17	4968	3237	2540	1988	1524	1109	0725	0359		
18	4886	3253	2553	2027	1587	1197	0837	0496	0173	
19	4808	3232	2561	2059	1641	1271	0932	0612	0303	
20	4734	3211	2565	2085	1686	1334	1013	0711	0422	0140
21	4634	3185	2578	2119	1736	1399	1092	0804	0530	0263

2. Таблица зависимостей для поиска критического значения в критерии Шапиро-Уилка

Уровень значимости α	Критическое значение W_{cr}
0,01	$\frac{-0{,}0148n^4 + 2{,}1875n^3 - 122{,}61n^2 + 3257{,}3n + 55585}{100000}$
0,05	$\frac{-0{,}0113n^4 + 1{,}656n^3 - 91{,}88n^2 + 2408{,}6n + 67608}{100000}$
0,1	$\frac{-0{,}0084n^4 + 1{,}2513n^3 - 70{,}724n^2 + 1890n + 73840}{100000}$

3. Таблица критических значений модифицированного показателя ω^2

	p	value
0	0.90	0.316
1	0.95	0.348
2	0.99	0.418

4. Функция для классического критерия χ^2

```
def interval_test(x, k, start):
    n = get_length(x, k)
    if start + n * k > x.max():
        print('Passed')
        return True
def chi_test(x, k, alpha, start, version='def'):
    """
    Вход - выборка, количество интервалов, уровень значимости, стартовая
позиция разбиения на интервалы и версия
    Функция выполняет проверку по критерию хи квадрат о соответствии
выборки нормальному, экспоненциальному и равномерному
    распределению
    """
    types = ['Norm', 'Exp', 'R']
    mean, d = pe_norm(x, 'm_d')
    std = d ** (1 / 2)
    l, err = pe_exp_mm(x, 'm')
    a, b = pe_r_mm(x, 'for_adj')
    n = get_length(x, k)
    if interval_test(x, k, start):
        table = []
        for i in range(k):
            table.append(x[np.where((x > i * n) & (x < (i + 1) * n))])
        p_e = [len(i) / len(x) for i in table]

        p_t_n = [stats.norm(loc=mean, scale=std).cdf(start + n)]
        p_t_e = [1 - math.exp(-l * (start + n))]
        p_t_r = [(start + n - a) / (b - a)]
        for i in range(1, k):
            p_t_n.append(stats.norm(loc=mean, scale=std).cdf((i + 1) * n)
- stats.norm(loc=mean, scale=std).cdf((i * n)))
            p_t_e.append(1 - math.exp(-l * ((i + 1) * n)) - 1 + math.exp(-
l * (i * n)))
            p_t_r.append(((i + 1) * n - a) / (b - a) - ((i * n - a) / (b -
a)))
```

```
        value_n, value_e, value_r = [], [], []
        for i in range(k):
            value_n.append((p_t_n[i] - p_e[i]) ** 2 / p_t_n[i])
            value_e.append((p_t_e[i] - p_e[i]) ** 2 / p_t_e[i])
            value_r.append((p_t_r[i] - p_e[i]) ** 2 / p_t_r[i])
        chi_sq = [len(x) * sum(i) for i in [value_n, value_e, value_r]]
        chi_sq_cr = [stats.chi2.ppf(1 - alpha, k - 3), stats.chi2.ppf(1 -
alpha, k - 2), stats.chi2.ppf(1 - alpha, k - 3)]
        if version == 'for_adj':
            return table, p_t_n, p_t_e, p_t_r, k, alpha
        for i in range(3):
            if chi_sq[i] < chi_sq_cr[i]:
                print('χ^2 =', round(chi_sq[i], 4), '< χ^2_cr =',
round(chi_sq_cr[i], 4) ,'\nПринимаем гипотезу f(x) ~ ', types[i], ' с
доверительной вероятностью ', 1 - alpha)
            else:
                print('χ^2 =', round(chi_sq[i], 4), '≥ χ^2_cr =',
round(chi_sq_cr[i], 4) ,'\nОтклоняем гипотезу f(x) ~ ', types[i])
```

5. Функция для классического критерия ω^2

```
def kramer_test(x, alpha, version='default'):
    '''
    На вход выборка, уровень значимости параметр версии
    В случае использования метода как первого шага модификации
возвращаются результаты проверки на нормальное,
    экспоненциальное и равномерное распределение
    В стандартном случае функция печатает результаты проверки
    '''
    x = np.sort(x)
    n = len(x)
    mean, std = pe_norm(x, 'm_d')
    std **= (1 / 2)
    types = ['Norm', 'Exp', 'R']
    l, err = pe_exp_mm(x, 'm')
    a, b = pe_r_mm(x, 'for_adj')
    wn, we, wr = 1 / (12 * n), 1 / (12 * n), 1 / (12 * n)
    for i in range(n):
        wn += ((stats.norm.cdf((x[i] - mean) / std) - (2 * (i + 1) - 1) /
(2 * n)) ** 2)
        we += (1 - math.exp(-l * x[i]) - (2 * (i + 1) - 1) / (2 * n)) ** 2
        wr += ((x[i] - a) / (b - a) - (2 * (i + 1) - 1) / (2 * n)) ** 2
    p = 1 - alpha
```

```
    w_cr = w_square_table[w_square_table.p == 1 - alpha].values[0][1]
    if version == 'for_adj':
        return wn, we, wr
    w = [wn, we, wr]
    for i in range(3):
        if w[i] <= w_cr:
            print('ω^2 =', round(w[i], 4), '≤ ω^2_cr =', round(w_cr, 4)
,'\nПринимаем гипотезу f(x) ~ ', types[i], 'с доверительной вероятностью
', 1 - alpha)
        else:
            print('ω^2 =', round(w[i], 4), '> ω^2_cr =', round(w_cr, 4)
,'\nОтклоняем гипотезу f(x) ~ ', types[i])
```

6. Код для многофакторного ДА

```
df_test = pd.read_csv('2way_clinic.csv')
groups = ['group1', 'group2']
df_first = len(df_test[groups[0]].unique()) - 1

df_second = len(df_test[groups[1]].unique()) - 1

df_third = 1

first_group = df_test['group1'].unique()
second_group = df_test['group2'].unique()
average1 = []
average2 = []
average3 = []
group_sizes1 = []
group_sizes2 = []
group_sizes3 = []
for val in first_group:
    average1.append(df_test[df_test['group1'] == val].mean())
    group_sizes1.append(len(df_test[df_test['group1'] == val]))
average_global1 = sum(average1) / len(average1)

for val in second_group:
    average2.append(df_test[df_test['group2'] == val].mean())
    group_sizes2.append(len(df_test[df_test['group2'] == val]))
average_global2 = sum(average2) / len(average2)

average3 = []
for gr in first_group:
```

```
    for val in second_group:
        average3.append(df_test[(df_test['group1'] == gr) &
(df_test['group2'] == val)].mean())
        group_sizes3.append(len(df_test[(df_test['group1'] == gr) &
(df_test['group2'] == val)]))
average_global3 = sum(average3) / len(average3)

ssb1 = 0
ssb1 = group_sizes1[0] * (average1[0][0] - average_global1[0]) ** 2
ssb1 += group_sizes1[1] * (average1[1][0] - average_global1[0]) ** 2
ms_bg1 = ssb1 / df_first

ssb2 = 0
ssb2 = group_sizes2[0] * (average2[0][0] - average_global2[0]) ** 2
ssb2 += group_sizes2[1] * (average2[1][0] - average_global2[0]) ** 2
ms_bg2 = ssb2 / df_second

ssb3 = 0
for i in range(4):
    ssb3 += group_sizes3[i] * (average3[i] - average_global3) ** 2
ssb3 = ssb3[0]
ms_bg3 = ssb3 / df_third


def_global = len(df_test) - len(group_sizes1) - len(group_sizes2)

variances = []
group_sizes_global = []
for val in first_group:
    for gr in second_group:
        variances.append(df_test[(df_test['group1'] == val) &
(df_test['group2'] == gr)].var())
        group_sizes_global.append(len(df_test[(df_test['group1'] == val) &
(df_test['group2'] == gr)]))
variances = np.array(variances)

ssw = 0
for i in range(len(group_sizes_global)):
    ssw += (group_sizes_global[i] - 1) * variances[i]
ssw = ssw[0]
ms_wg = ssw / def_global

f_first = ms_bg1 / ms_wg
```

```
f_second = ms_bg2 / ms_wg
f_third = ms_bg3 / ms_wg
p_value1 = round(1 - stats.f.cdf(f_first, df_first, def_global), 5)
p_value2 = round(1 - stats.f.cdf(f_second, df_second, def_global), 5)
p_value3 = round(1 - stats.f.cdf(f_third, df_third, def_global), 5)
sns.catplot(x='group1', y='0', data=df_test, kind='point', hue='group2',
capsize=0.2)
plt.savefig('anova.png', dpi=400, bbox_inches='tight')
plt.show()
```