Владимирский государственный университет

М. А. ГУНДОРОВА С. А. ГРАЧЕВ

ПАКЕТЫ ПРИКЛАДНЫХ СТАТИСТИЧЕСКИХ ПРОГРАММ

Учебное пособие

Владимир 2024

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

М. А. ГУНДОРОВА С. А. ГРАЧЕВ

ПАКЕТЫ ПРИКЛАДНЫХ СТАТИСТИЧЕСКИХ ПРОГРАММ

Учебное пособие

Электронное издание

Владимир 2024

ISBN 978-5-9984-1953-9

УДК 658(075.8)
ББК 65.051.01я

Рецензенты:
Доктор экономических наук, профессор
профессор кафедры бизнес-информатики и экономики
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых
А. М. Губернаторов

Доктор экономических наук, доцент
профессор кафедры менеджмента,
директор Центра стратегического развития Владимирского филиала
Российской академии народного хозяйства и государственной службы
при Президенте Российской Федерации
О. Л. Гойхер

Гундорова, М. А. Пакеты прикладных статистических программ [Электронный ресурс] : учеб. пособие / М. А. Гундорова, С. А. Грачев ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2024. – 420 с. – ISBN 978-5-9984-1953-9. – Электрон. дан. (26,9 Мб). – 1 электрон. опт. диск (CD-ROM). – Систем. требования: Intel от 1,3 ГГц ; Windows XP/7/8/10 ; Adobe Reader ; дисковод CD-ROM. – Загл. с титул. экрана.

Включает полный курс дисциплины «Пакеты прикладных статистических программ». Рассмотрены теоретические и практические основы дисциплины, проанализированы актуальные статистические методы в управлении фирмой.

Предназначено для студентов специалитета, обучающихся по специальности «Экономическая безопасность» всех форм обучения.

Рекомендовано для формирования профессиональных компетенций в соответствии с ФГОС ВО.

Табл. 66. Ил. 251. Библиогр.: 110 назв.

ISBN 978-5-9984-1953-9

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Современная экономика невозможна без применения инструментов математической статистики, процессов моделирования и прогнозирования, корреляционно-регрессионного анализа, анализа трендов и иных методов и моделей. В основе данных процессов лежат закономерности, носящие статистическо-вероятностный характер.

Экономическая система является крайне сложным образованием и включает в себя множество подсистем и элементов, взаимодействие которых происходит по определенным законам. Для снижения уровня неопределённости и повышения эффективности функционирования любого хозяйствующего субъекта необходимо владеть навыками работы со статистической информацией.

Неотъемлемой чертой современного общества является процесс цифровизации, который проникает во все сферы экономики и общества, перестраивая коренным образом многие процессы.

Это наряду с постоянно усложняющейся социально-экономической системой и ростом объёмов данных порождает необходимость применения в рамках исследования и анализа специального программного обеспечения, которое значительно повышает эффективность работы с информацией.

Существует значительное число программных продуктов, позволяющих применять отдельные инструменты математической статистики и статистического анализа при решении прикладных задач, стоящих перед специалистом сегодня.

Пособие содержит описание порядка проведения ряда статистических и экономических расчетов с использованием программных комплексов Statistica и MS Excel. Помимо простого рассмотрения теоретических вопросов анализа данных и описания функциональных возможностей программ, пособие содержит конкретные примеры выполнения прикладных задач.

Издание позволит студентам:

- ознакомиться с общими теоретическими положениями статистического анализа данных;

- рассмотреть основные функциональные возможности программных комплексов Statistica и MS Excel;

- научиться применять программные комплексы при решении задач профессионального характера и в исследовательских вопросах.

Пособие построено таким образом, что позволяет освоить курс «Пакеты прикладного статистического анализа» последовательно, изучив все представленные разделы, но обладая полученными знаниями при освоении дисциплины «Статистика», обучающийся может начать изучение пособия с интересующего его раздела.

ВВОДНАЯ ЧАСТЬ

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ СТАТИСТИЧЕСКИХ МЕТОДОВ

1. Предмет, задачи и методы статистической науки

Слово «статистика» происходит от латинского «status», т.е. состояние. В Средние века данное слово использовалось для обозначения политического состояние государства.

В науку данный термин ввел немецкий ученый Готфрид Ахенвалль. Зарождение статистики как науки произошло только в XVII в., однако элементы статистического учета можно было наблюдать еще в глубокой древности. Из исторических источников известно, что переписи населения в Китае проводились еще за 5 тыс. лет до н. э., древние римляне регулярно проводили статистическое сравнение военного потенциала разных стран, вели учет имущества граждан. В Средние века активно велся учет домашнего имущества и земель [1].

У истоков статистической науки стояли две основные школы (рис. 1.1).

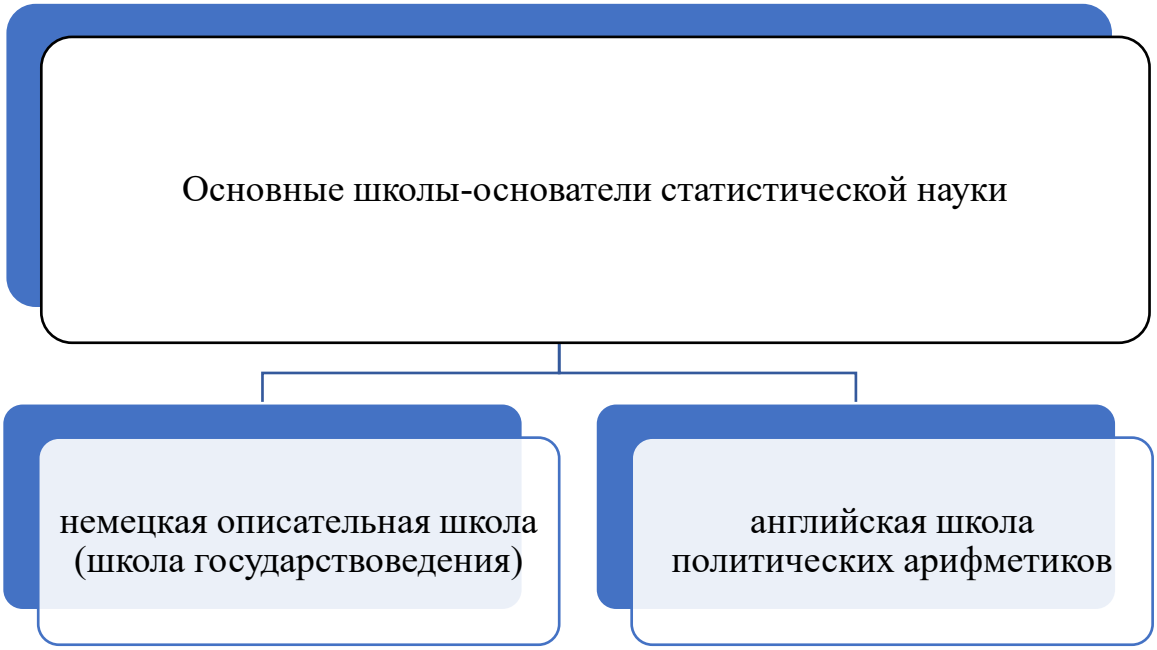


Рис. 1.1. Основные школы-основатели статистической науки

Представители описательной школы считали, что основная задача статистики состоит в описании территории государства, населения, климата, вероисповедания, способов ведения хозяйства и т. п. Также они отмечали, что ведение данного учета важно исключительно в словесной форме.

По сути они продвигали идеи учета исключительно моментной статистики, без учета особенностей развития территорий в те или иные периоды, цифр и динамических характеристик

Основными представителями немецкой школы были Г. Конринг, Г. Ахенвалль, А. Бюшинг и др.

К представителям российской школы государствоведения, разделявшим аналогичные взгляды, можно отнести первооткрывателя табличного метода в статистике И. К. Кириллова, В. Н. Татищева, который занимался проблемой источниковедения, М.В. Ломоносова. Представившего экономико-географическое описание Российского государства и разработал подробную анкету для сбора статистических данных, руководителя первого в стране Статистического комитета К. Ф.

Несмотря на значительные достижения в данной области всех вышеперечисленных ученых, основателем русской государственной статистики принято считать П. П. Семенова-Тян-Шанского. Именно он стал инициатором проведения Всероссийской переписи населения в 1897 г. и отвечал за обработку ее материалов. Семенов-Тян-Шанский является автором множества сборников и справочников по фабрично-заводской статистике.

Основная цель политических арифметиков состояла в изучении общественных массовых явления с применением числовых характеристик.

Именно с открытием данной школы принято выделять новый этап в развитии статистической науки, поскольку от описания явлений и процессов статистика перешла к их измерению и исследованию, к выработке вероятных гипотез будущего развития.

Основной назначение статистической науки политические арифметики видели в изучении массовых общественных явлений. Они осознавали необходимость учета в статистическом исследовании требований закона больших чисел, поскольку закономерность может про-

явиться лишь при достаточно большом объеме анализируемой совокупности. История показала, что последнее слово в статистической науке осталось именно за школой политических арифметиков.

Основателем английской школы принято считать Уильяма Петти. Данный ученый активно интересовался хозяйственными процессами, закономерностями в экономической жизни страны. Именно Петти впервые сделал попытку оценки национального богатства и национального дохода.

Также яркими представителями английской школы политических арифметиков можно назвать Дж. Граунта, который исследовал закономерности воспроизводства населения и построил первую в историю таблицу смертности в своей работе «Естественные и политические наблюдения, перечисленные в прилагаемом оглавлении и сделанные над бюллетенями смертности, по отношению к управлению, религии, торговле, росту, болезням и пр.» , А. Кетле, который возглавлял национальную статистику Бельгии и был основоположником учения о средних величинах. Также Кетле изучал закономерности общественной жизни в области преступности и выявил действие постоянных и случайных причин и впервые ввел термин «средний человек».

Основное преемственное направление английской школы политических арифметиков в статистике – математическое. Оно возникло в XIX веке под влиянием идей Ф. Гальтона, К. Пирсона, У. Госсета (Стьюдента), Р. Фишера и других исследователей.

Существенное влияния на эволюцию статистической методологии оказали труды российских статистиков, представителей так называемой академической статистики, А. А. Чупрова, В. С. Немчинова, С. Г. Струмилина и др.

Развитие статистической науки и расширение области практической статистической работы привели к изменению сущности понятия «статистика». В настоящее время данный термин употребляется в трех значениях:

1) статистика как отрасль практической деятельности, основная цель которой состоит в сборе, обработке и анализе данных о разнообразных явлениях общественной жизни. Полученная в результате статистического исследования информация позволяет решать задачи выявления реально существующих закономерностей, свойственных описываемым процессам и явлениям;

2) статистика — это данные, которые служат для количественной характеристикой общественных явлений или территориального распределения показателя;

3) статистика как это наука. Как и любая другая наука, статистика имеет свой предмет и метод изучения. Предмет статистики заключается в изучении количественной стороны массовых социально-экономических явлений в связи с их качественной стороной, в исследовании количественно выраженных закономерностей общественного развития в конкретных условиях места и времени. Свой предмет статистика изучает при помощи специфического метода. Кратко статистический метод можно охарактеризовать следующим образом: это сбор, обобщение, представление, анализ и интерпретация данных [1].

Так как статистика изучает множество социально-экономических явлений и характерные для них закономерности, то и метод статистики представляет собой целую совокупность приемов, пользуясь которыми статистика исследует свой предмет.

К основным приемам статистической науки относят статистическое наблюдение, метод группировки и обобщения данных с последующим представлением результатов анализа и их интерпретацией

Задачи статистики как науки состоят в следующем:

- ✓ описание структуры экономики;
- ✓ описание тенденций развития экономики в будущем;
- ✓ анализ и прогнозирование различных экономических явлений;
- ✓ выявление факторов развития экономики для принятия управленческих решений.

В России экономико-статистические исследования проводятся научно-исследовательскими институтами, ведомственными статистическими органами и организациями, а также независимыми специалистами, однако преимущественная часть статистической информации формируется в системе официальной государственной статистики. В статистических управлениях первичная статистическая информация последовательно агрегируется с целью получения на уровне Росстата РФ макромодели функционирования экономики страны в виде системы национальных счетов.

Статистические органы преобразуют полученные от респондентов индивидуальные сведения и предоставляют потребителям информацию в полном соответствии с принципом конфиденциальности: только макроданные, относящиеся не менее чем к трем объектам наблюдения. Основные принципы организации работы органов официальной статистики в России (принцип легальности, принципы предметной централизации и региональной децентрализации) соответствуют требованиям Евростата и Департамента статистики ООН. В соответствии с международными стандартами ведения статистики и учета в России к официальной статистике относятся государственные статистические управления и ведомственная статистика (внутренняя и внешняя), т. е. определенные государственные организации, которые выполняют важные статистические работы, связанные с их собственной деятельностью (например, отделы ЗАГС) [1].

Права и обязанности официальной статистики детально урегулированы на федеральном уровне. Наряду с этим существует широкая и разнообразная сфера альтернативной статистики, т. е. частных, неофициальных статистических исследований, организаторы которых не имеют полномочий для проведения обследований с обязанностью предоставления информации широкому кругу лиц.

Основные принципы работы статистических управлений, в том числе в отношении сбора данных о населении представлены на рис. 1.2.

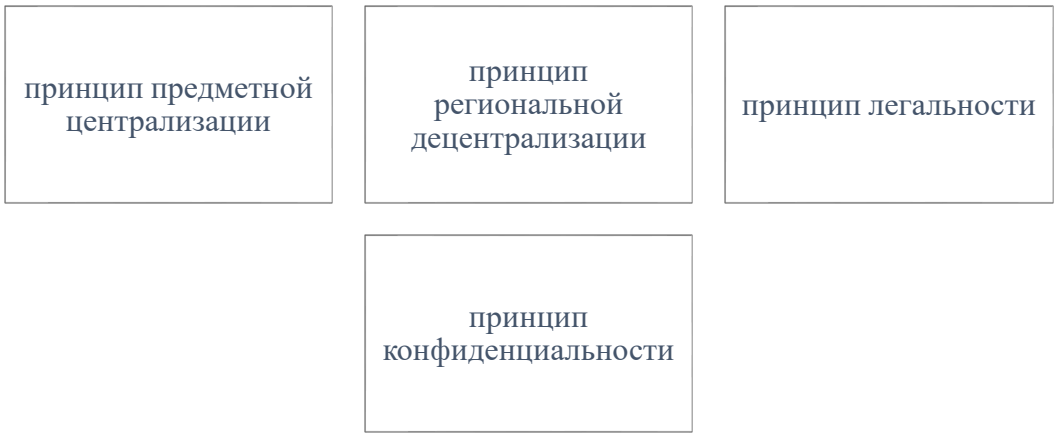


Рис. 1.2. Основные принципы работы статистических управлений

Реализация данных принципов позволяет добиться нейтральной и независимой позиции статистических управлений и тем самым укрепить доверие респондентов и пользователей, без которого статистика не может обойтись.

С мая 2012 г. деятельностью Росстата руководит Правительство РФ. Росстат РФ, его органы в республиках, краях, областях, автономных областях и округах, в городах Москве и Санкт-Петербурге, других городах и районах, а также подведомственные им организации, учреждения и учебные заведения составляют единую систему государственной статистики страны. Формы и методы сбора и обработки статистических данных, методология расчета статистических показателей, установленные Росстатом, являются статистическими стандартами РФ [1].

Основные задачи Росстата РФ представлены на рис. 1.3.

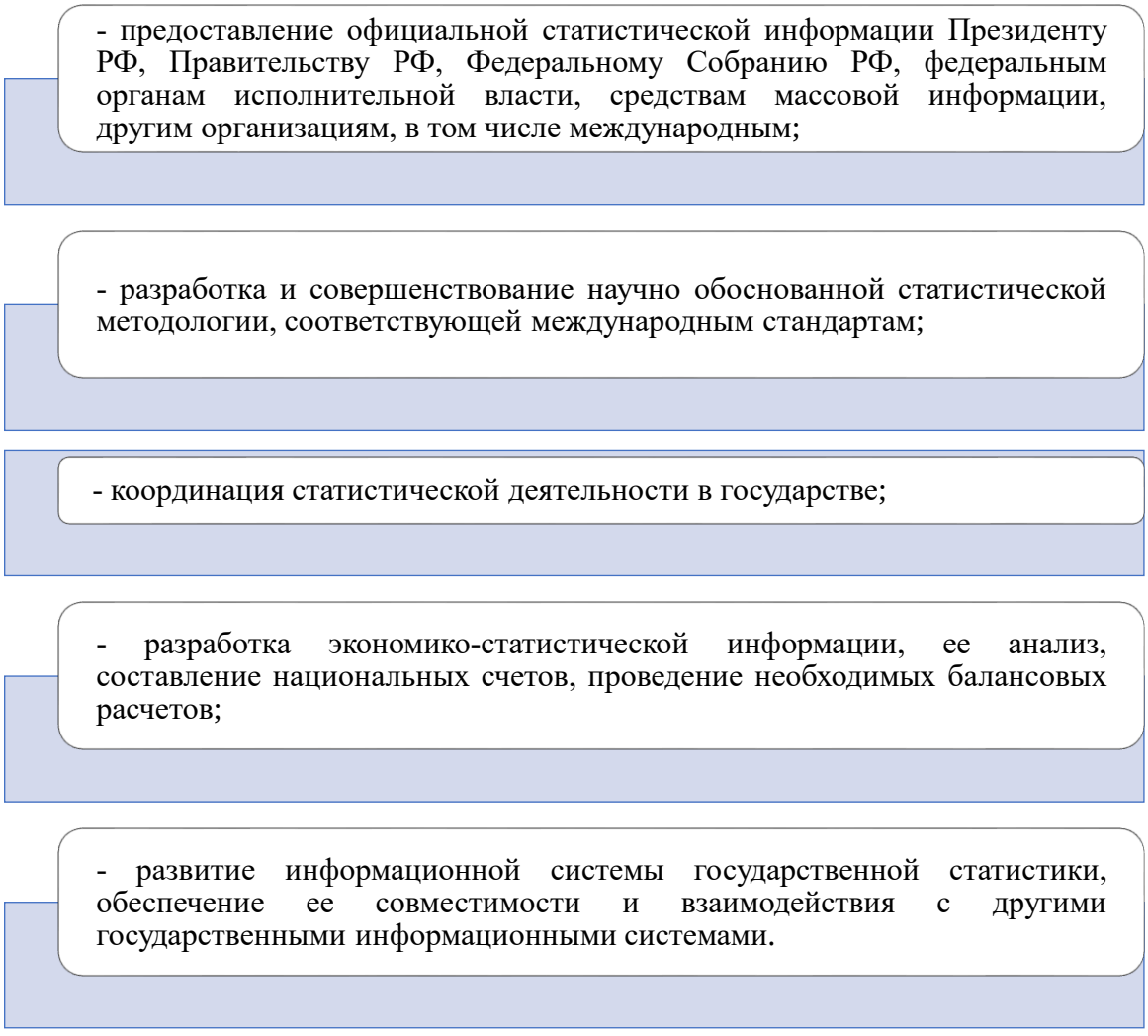


Рис. 1.3. Основные задачи Росстата

Таким образом, в ходе эволюции статистической науки менялись представлении о ее предмете. В данном параграфе были рассмотрены задачи, предмет и методы статистической науки, а также основные принципы организации работы органов официальной статистики в России.

2. Значение статистического наблюдения в управлении фирмой. Классификация статистических наблюдений

Статистическое наблюдение – это научно обоснованная регистрация по единой разработанной программе фактов и их признаков, которые характеризуют явления общественной жизни, и сбор массовых данных.

Любое статистическое наблюдение начинается с планирования и организации исследования. Данный этап включает в себя разработку программы статистического наблюдения, определение критического момента наблюдения, времени и периода наблюдения, определение цели и задач исследования, объекта наблюдения.

Вторым этапом статистического исследования является непосредственно статистическое наблюдение. При его организации важная роль отводится планированию: от качества отобранных статистических данных зависит правильность и достоверность выводов, которые будут использоваться в рамках управляющего воздействия.

Для более четкой организации статистического наблюдения разрабатывают программу наблюдений, которая представляет собой перечень вопросов, по которым собираются сведения, либо перечень признаков и показателей, которые подлежат регистрации [1].

Программа наблюдения оформляется в виде бланка (анкеты, формуляра), в который заносятся первичные сведения. Необходимым дополнением к бланку является инструкция (или указания на самих формулярах), в которой разъясняется смысл представленных вопросов. Состав и содержание вопросов программы наблюдения напрямую зависит от конкретных исследовательских задач и особенностей рассматриваемого явления.

Важным элементом статистического наблюдения является понятие критического момента наблюдения. Данное понятие представляет

собой момент или отрезок времени, по состоянию на который проводится регистрация значений признаков по каждой единице наблюдения.

В зависимости от целей и задач исследования, особенностей структуры совокупности, предмета исследования критическим моментом может быть дата (день, час), неделя, месяц и т. п.

Период наблюдения представляет собой интервал времени, в течение которого осуществляется сбор данных, заполнение бланков программы наблюдения.

Время наблюдения – это время, в течение которого проводится обследование по разработанной программе.

Формулировка цели исследования предполагает постановку научной проблемы, определение свойств и тенденций явления, которые подлежат анализу.

Задачи исследования — совокупность действий, необходимых для достижения цели исследования.

Объект наблюдения — совокупность социально-экономических явлений и процессов, которые подлежат исследованию, или точные границы, в пределах которых будут регистрироваться статистические сведения. Например, при переписи населения необходимо установить, какое именно население подлежит регистрации — наличное, т. е. фактически находящееся в данной местности в момент переписи, или постоянное, т. е. живущее в данной местности постоянно.

Совокупность (статистическая совокупность) представляет собой множество единиц изучаемого явления, объединенных единой качественной основой, но отличающихся друг от друга отдельными признаками. Таковы, например, совокупность домохозяйств, совокупность семей, совокупность предприятий, фирм, объединений и т. п.

Основными свойствами статистической совокупности является однородность, динамичность и независимость единиц. Совокупность называется однородной, если один или несколько изучаемых существенных признаков ее объектов являются общими для всех единиц. Совокупность, в которую входят явления разного типа, считается разнородной. Совокупность может быть однородна в одном отношении и разнородна в другом [1].

В каждом отдельном случае однородность совокупности устанавливается путем проведения качественного анализа, опираясь на сущность изучаемого процесса или явления.

Динамичность совокупности означает, что появление новых элементов совокупности и исчезновение существовавших ранее не отменяет существования совокупности как объекта наблюдения. Например, совокупность студентов высших учебных заведений не исчезает в результате отчисления одних студентов и восстановления других.

Независимость единиц показывает, что значения признаков одних единиц совокупности не могут быть получены как функция значений других ее единиц. Чтобы определить статистическую совокупность, необходимо ответить на два вопроса: какие единицы входят в совокупность и как эти единицы отличаются друг от друга [1].

В статистике выделяют три вида единиц (рис. 1.4).

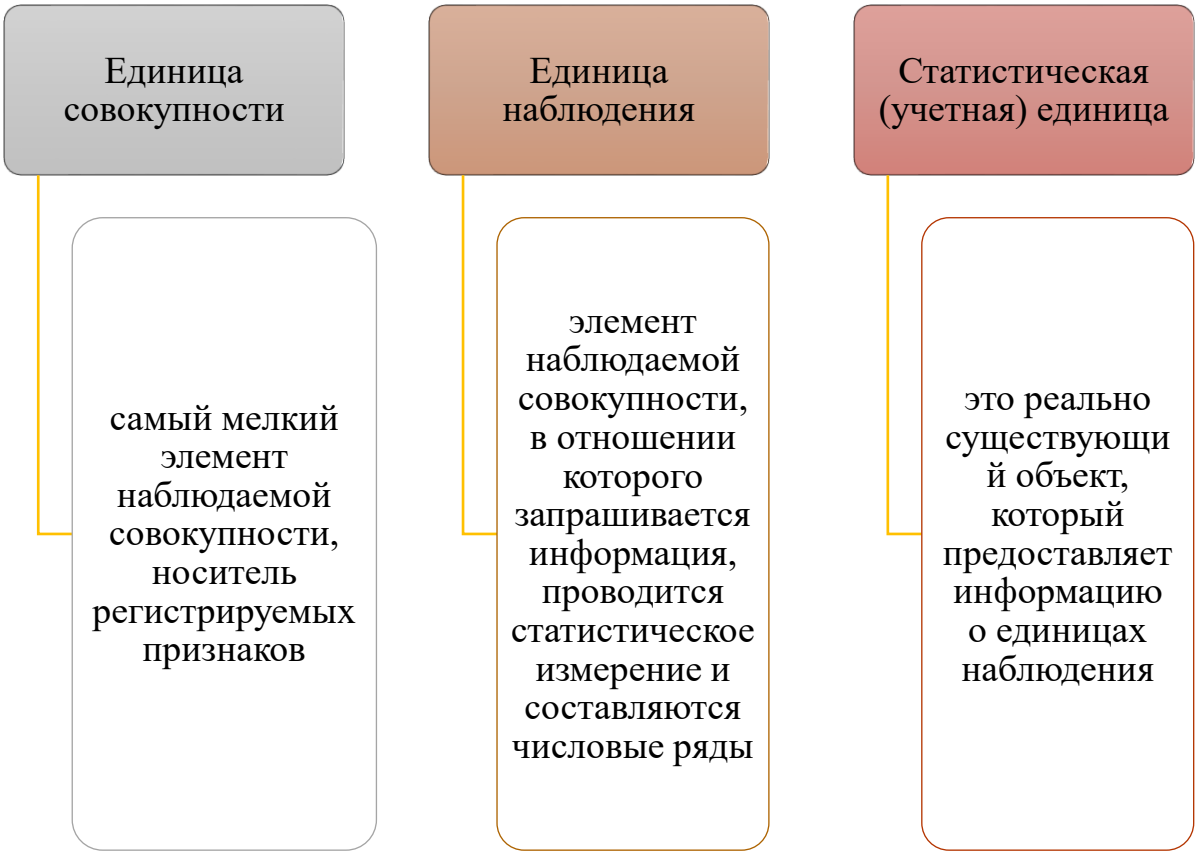


Рис. 1.4. Основные виды единиц в статистике

Стоит отметить, что единица совокупности и единица наблюдения могут совпадать (например, при анализе успеваемости студентов группы, каждый студент является единицей совокупности и единицей наблюдения).

Каждая единица наблюдения является набором значений различных признаков, которые определяют качественные особенности единицы совокупности.

Признаки можно разделить на три группы.

К первой группе относятся признаки, которые присущи всем единицам рассматриваемой статистической совокупности и помогают однозначно определить границы наблюдаемой совокупности. Значения признаков данной группы отвечают на вопросы:

- что изучается?
- когда изучается?
- где изучается?

Таким образом, значения данной группы признаков дают ответ на первый вопрос в определении статистической совокупности — какие единицы входят в данную совокупность.

Вторая группа включает в себя признаки, которые позволяют отличить единицы совокупности друг от друга. Данные признаки являются особенными, индивидуальными и неизменными для каждой единицы совокупности. Данная группа признаков отвечает на второй вопрос в определении статистической совокупности — как единицы совокупности отличаются друг от друга.

Третья группа представляет собой признаки как предмет статистического интереса, т. е. случайным образом варьирующие признаки единиц наблюдения (например, объем произведенных услуг какой-либо конкретной фирмой).

Значения таких признаков могут иногда совпадать у отдельных единиц совокупности, а могут быть различными. Именно эта группа признаков является непосредственным предметом изучения в статистическом исследовании [1].

По характеру отображения свойств единиц изучаемой совокупности признаки делятся на две основные группы (рис. 1.5).

Рис. 1.5. Классификация признаков по характеру отображения свойств единиц изучаемой совокупности

В статистическом исследовании принято различать пять основных шкал измерения признаков (в порядке повышения точности измерения) (рис. 1.6).

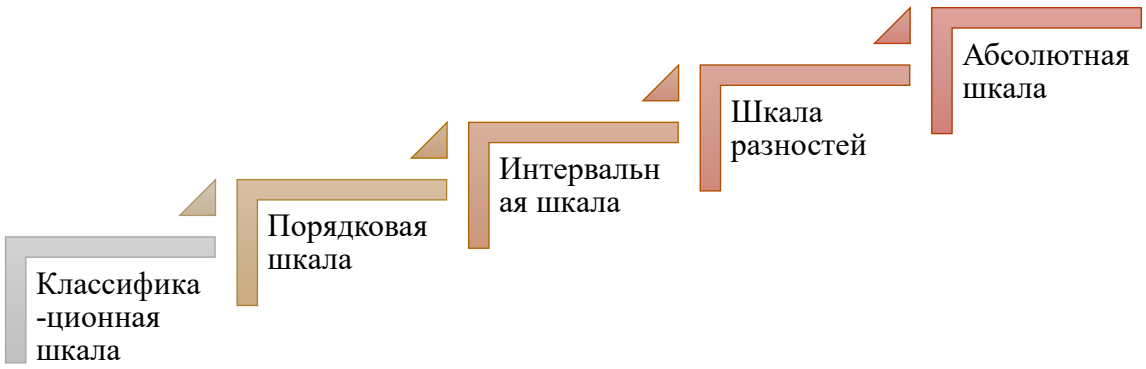


Рис. 1.6. Основные шкалы измерения в статистике

Классификационная шкала представляет собой перечень значений атрибутивного признака (например, телефонный справочник). Эта

шкала не имеет ни нуля (начала отсчета), ни предпочтений, ни единицы измерения (рис. 1.7).

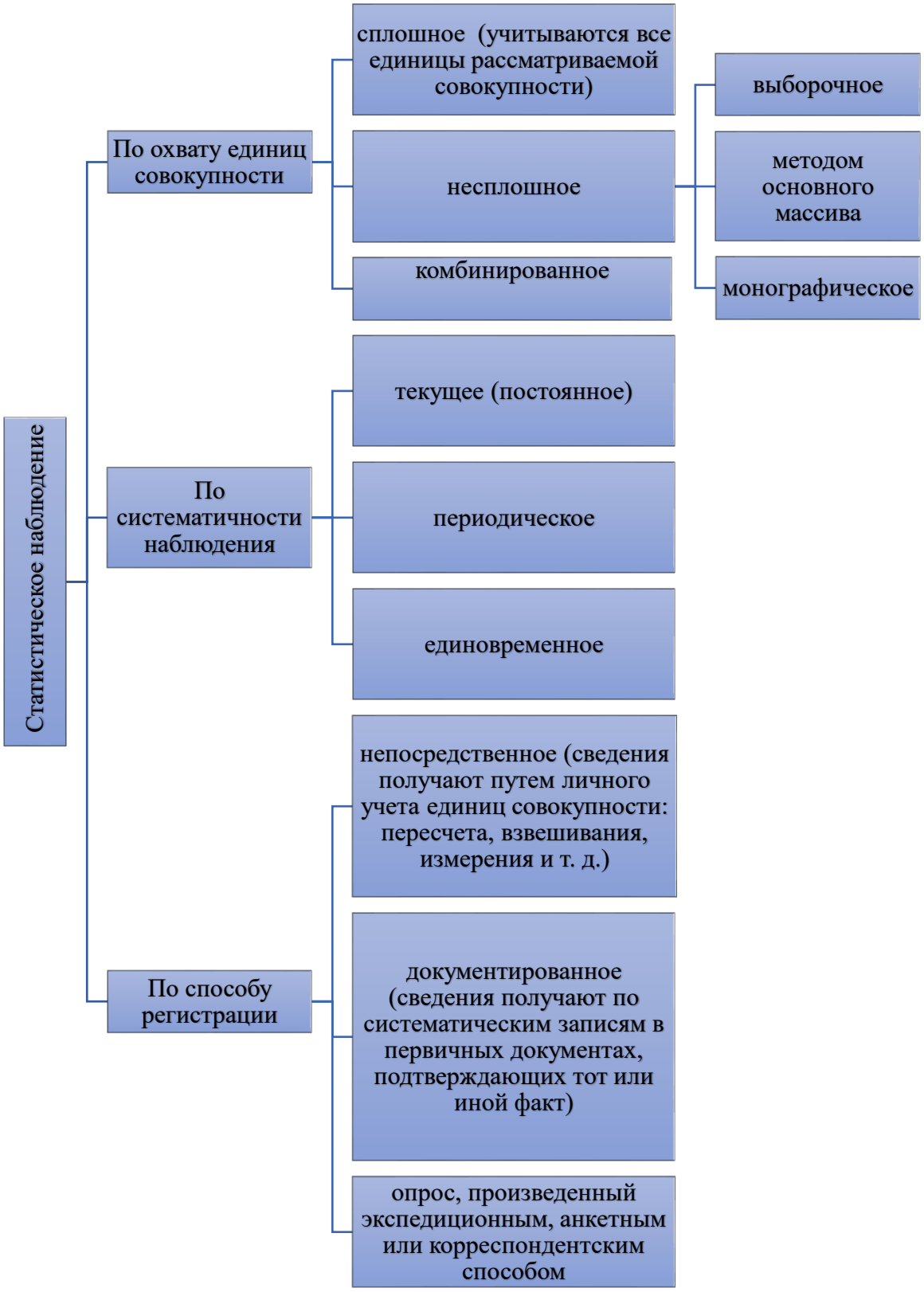


Рис. 1.7. Классификация статистических наблюдений

В порядковой (ранговой) шкале устанавливаются отношения предпочтений между вариантами значений признака. В данной шкале также нет нуля (начала отсчета) и единицы измерения.

Интервальная шкала устанавливает отношения следования между интервалами значений признака. Имеет произвольный нуль и произвольную единицу измерения.

Шкала разностей устанавливает отношения следования между разностями значений признака. Имеет фиксированную единицу измерения и произвольный нуль (например, шкала времени).

В отличие от шкалы разностей, шкала отношений имеет фиксированный нуль, а единица измерения в ней может быть произвольной.

Абсолютная шкала имеет фиксированный нуль и фиксированную единицу измерения показателя.

На этапе статистического наблюдения проводится сбор данных по разработанной программе. Однако не всякий сбор данных можно назвать статистическим наблюдением. О статистическом наблюдении можно говорить только в том случае, если обеспечивается регистрация устанавливаемых фактов в специальных учетных документах и изучаются статистические закономерности, которые проявляются в большом числе единиц некоторой совокупности. Именно поэтому статистическое наблюдение должно быть планомерным, массовым и систематическим.

К статистическому наблюдению предъявляется целый ряд требований, к которым можно отнести:

- ✓ Достоверность данных
- ✓ Полнота данных
- ✓ Точность данных
- ✓ Практическая ценность статистических данных
- ✓ Сопоставимость и единообразие данных [2].

Существуют различные признаки классификации статистических наблюдений, основные из которых приведены на рис. 1.7.

Рассмотрим подробнее разновидности несплошных наблюдений.

Выборочное наблюдение предполагает, что изучается отобранная в случайном порядке часть единиц совокупности с целью характеристики всей совокупности.

Несплошное наблюдение методом основного массива означает, что обследованию подвергается основная часть совокупности, и из генеральной совокупности исключается некоторая часть, о которой заведомо известно, что она не играет большой роли в характеристике всей совокупности [1, 2].

Монографическое исследование предполагает изучение отдельных типичных единиц совокупности

В статистической практике используются три организационные формы наблюдения: отчетность, специальное организованное наблюдение и регистр.

Отчетность — это такая организационная форма, при которой единицы наблюдения предоставляют сведения о своей деятельности в виде формуляров регламентированного образца. Особенность отчетности состоит в том, что она обязательна, документально обоснована и юридически подтверждена подписью руководителя.

Специально организованное наблюдение проводится с целью получения сведений, отсутствующих в отчетности, или для проверки ее данных. Примером специально организованного наблюдения является перепись населения. Кроме этого, органы статистики проводят бюджетные обследования, которые характеризуют структуру потребительских расходов и доходов семей.

Регистр представляет собой систему, постоянно следящую за состоянием единицы наблюдения и оценивающую силу воздействия различных факторов на изучаемые показатели. В практике статистики различают регистры населения и регистры предприятий.

Выделяют следующие способы статистического наблюдения:

- ✓ экспедиционный (специально подготовленные регистраторы путем опроса заполняют формуляры, одновременно контролируя правильность получаемых сведений);
- ✓ саморегистрации (работники статистических органов раздают опросные бланки опрашиваемым лицам, инструктируют их, а затем собирают заполненные формуляры, контролируя полноту и правильность полученных сведений);
- ✓ корреспондентский (статистическими органами организуется специальная сеть корреспондентов из лиц, проживающих на местах, которые проводят наблюдение согласно разработанному

бланку и инструкции и сообщают сведения статистическим органам);

- ✓ анкетный (разработанная анкета рассылается кругу лиц и после заполнения возвращается органам, проводящим наблюдения);
- ✓ явочный (предусматривает предоставление сведений в органы, ведущие наблюдение в явочном порядке).

Перед началом третьего этапа статистического исследования необходимо провести арифметический и логический контроль собранных данных с целью устранения ошибок наблюдения. В статистике ошибкой наблюдения называют расхождение между расчетным и действительным значениями исследуемой величины. В зависимости от причин возникновения различают ошибки регистрации и ошибки репрезентативности. Контрольной проверкой собранных данных статистическое наблюдение завершается [1].

Таким образом, в рамках данного параграфа была рассмотрена сущность статистического наблюдения и определено его значение в процессе управления фирмой. На основании вышеизложенного можно сделать вывод о том, что грамотная организация статистического наблюдения является залогом успешного управления фирмой с использованием статистических методов и приемов.

3. Статистическая сводка и группировка

Одним их из этапов статистического исследования является сводка и группировка данных. Сводка представляет собой упорядочивание и обобщение первичного материала, сводку его в группы и выдачу на этой основе обобщающих характеристик совокупности.

Составными элементами сводки являются: программа сводки; подсчет групповых итогов; оформление конечных результатов сводки в виде таблиц и графиков [1, 2].

Различают простую сводку (подсчет только общих итогов) и статистическую группировку, которая сводится к расчленению совокупности на группы по существенному для единиц совокупности признаку.

Группировка позволяет получить такие результаты, по которым можно выявить состав совокупности, характерные черты и свойства

типичных явлений, обнаружить закономерности и взаимосвязи. Результаты сводки могут быть представлены в виде статистических рядов распределения.

Пусть из совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 – n_2 раз, x_k – n_k раз и $\sum n_i = n$ – объем выборки. Наблюдаемые значения x_i называют вариантами, а последовательность вариант, записанных в возрастающем порядке, - вариационным рядом. Числа наблюдений называют частотами, а их отношения к объему выборки $n_i/n = w_i$ − относительными частотами.

Статистическое распределение выборки представляет собой перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (в качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал) [2].

Статистическим рядом распределения называют упорядоченное распределение единиц совокупности на группы по изучаемому признаку. В зависимости от признака ряды могут быть вариационными (количественными) и атрибутивными (качественными). При построении вариационного ряда с равными интервалами определяют его число групп (n) и величину интервала (h). Число групп можно определить с помощью различных формул. Оптимальное число групп может быть определено по формуле Стерджесса (1.1):

$$n = 1 + 3{,}322 * \lg N, \qquad (1.1)$$

где n – число групп,
N – число единиц совокупности.

В зависимости от исследовательских целей можно использовать равные и неравные интервалы (в последнем случае – равномерно возрастающие или убывающие) открытые и закрытые. Величина равного интервала рассчитывается по формуле (1.2):

$$i = \frac{x_{max} - x_{min}}{n}, \qquad (1.2)$$

где i – длина интервала;

x_{max} – максимальное значение признака в рассматриваемой совокупности;

x_{min} – минимальное значение признака в рассматриваемой совокупности.

При проведении анализа вариационных рядов с неравными интервалами применяется показатель плотности распределения признака. Он может быть рассчитан как частное частоты или частости каждого интервала к его величине.

Для вариационного ряда возможен расчёт накопленных частот. По значению данного параметра можно сформулировать вывод о том, какое число единиц в совокупности имеет значение признака не выше того значения, которое соответствует выбранной величине накопленной частоты.

Группировка является процессом образования групп единиц совокупности однородных в каком-либо отношении, а также имеющих одинаковые или близкие значения группировочного признака.

Группировки бывают простые и комбинационные. Образование простых группировок осуществляется по какому-либо единому признаку, в том время как комбинационная группировка предполагает сочетание двух и более группировочных признаков. Основную задачу метода группировки можно определить, как выявление и выделение основных типов явлений, определение структуры совокупности, изучение взаимосвязи признаков [2].

Классификация видов группировок зависимости от цели и задач исследования приведена на рис. 1.8.

Рис. 1.8. Виды группировок

Если группировка используется для того, чтобы охарактеризовать качественные особенности и различия между типами явлений, принято говорить о типологической группировке.

Данный вид группировки активно используется в различных исследованиях социального и экономического характера.

Если группировка производится с целью выявления состава однородной в качественном отношении совокупности по какому-либо признаку, применяется структурная группировка.

В экономике фирмы примером применения данных видов группировок можно назвать группировку организаций по признаку численности рабочей силы, проценту реализации плана и прочее.

Аналитическая группировка применима в том случае, если основная исследовательская задача состоит в изучении взаимосвязи между процессами, явлениями или признаками. В результате проведения данного вида группировки определяются причинно-следственные связи в рассматриваемых объектах [1, 2].

Таким образом, сводка и группировка данных является важнейшим этапом реализации статистического исследования. В управлении фирмой применение тех или иных методов и приемов опирается на результаты первичной обработки статистических данных. Именно поэтому грамотная реализация данного этапа является залогом успешности проводимых управляющих воздействий.

4. Представление данных

Грамотное представление статистических данных является важнейшим этапом исследования. При этом удобство пользования данными существенно упрощает задачу любого исследования с применением статистических методов или приемов. Статистические данные могут быть представлены в графическом, текстовом и числовом виде.

Наиболее рациональным способом представления итоговых результатов сводки и группировки является формирование статистических таблиц. Основное преимущество использования таблиц состоит в возможности компактного, удобного и рационального представления нужных данных. При этом важное значение уделяется грамотному выделению статистического подлежащего и сказуемого в данной форме. Как правило, строки в таблицах используются для отображения подлежащего, а столбцы – для сказуемого [2].

В зависимости от строения подлежащего принято выделять три вида статистических таблиц (рис. 1.9).

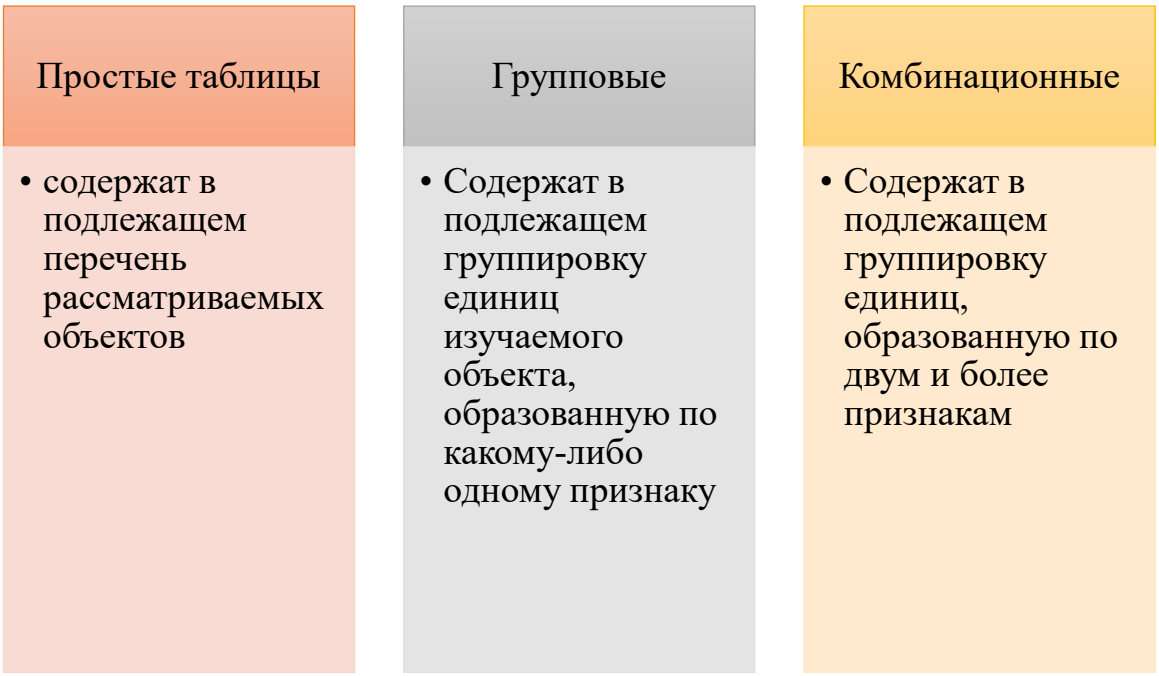


Рис. 1.9. Виды статистических таблиц по типу строения подлежащего

Обобщая все вышесказанное, можно сделать вывод о том, что статистическое подлежащее по своей сути представляет объект конкретного исследования, в сказуемое – систему показателей, которые позволяют охарактеризовать нужные стороны рассматриваемого процесса или явления.

Важно помнить, что наглядности и рациональности представляемых данных можно достичь только в том случае, если четко и грамотно сформулирован предмет и объект статистического исследования. Излишняя детализация представляемых данных в формируемой таблице только усложняет процесс интерпретации результатов сводки и группировки. Если представляется задача группировки большого объема данных, отражающих совокупность различных статистических показателей, на практике зачастую используют разбивку исходных параметров на несколько статистических таблиц, каждая из которых отображает блок рассматриваемых статистических данных по какому-либо обобщенному вопросу [1, 2].

Существует целый ряд практических рекомендаций по составлению и оформлению таблиц, которые позволяют значительно рационализировать представляемые сведения (рис. 1.10).

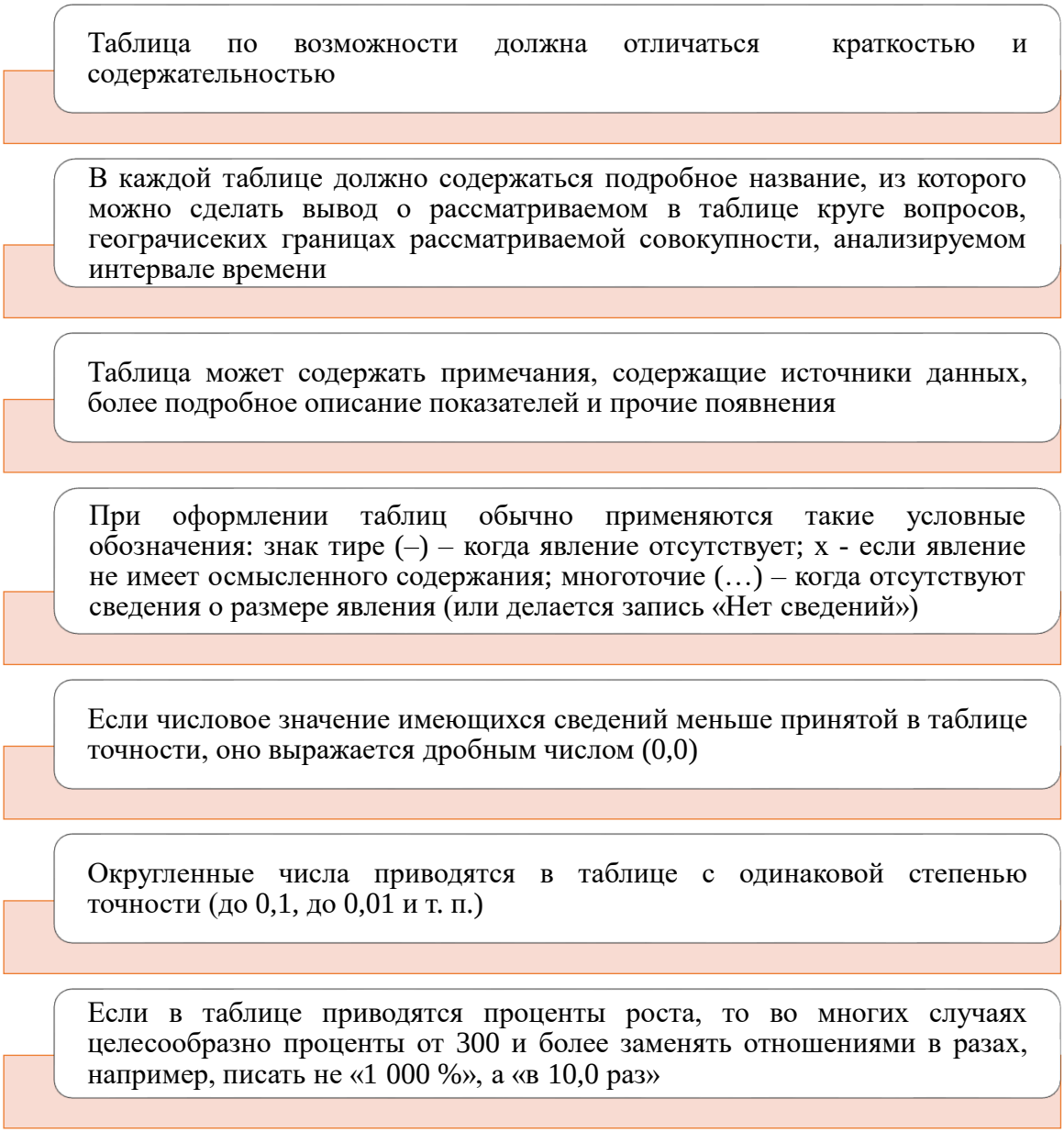


Рис. 1.10. Основные правила составления и оформления статистических таблиц

В таблицах важно уточнять, каковы используемые единицы измерения. Если единицы измерения для различных ячеек таблицы отличаются друг от друга, в верхних или боковых заголовках обязательно необходимо уточнять, в каких единицах измерения приводятся рассматриваемые статистические данные.

Применение графиков в процессе представления анализируемых статистических параметров дает возможность наглядного и вырази-

тельного отображения показателей, характеризующих рассматриваемый процесс или явление. Кроме того, грамотное графическое представление способно существенно облегчить восприятие и продемонстрировать наличие или отсутствие взаимосвязи между параметрами, а также характер данной взаимосвязи. В зависимости от того, что именно демонстрируется на графике (сравнение параметров, динамика их изменения, степень распространения процессов в конкретном подразделении фирмы и т.д.), могут быть использованы различные их виды [2].

Современные автоматизированные компьютерные программы обладают широким спектром возможностей для грамотного отображения на графике рассматриваемых процессов и явлений. Основная задача пользователя при этом сводится к грамотному определению типа представляемых данных и выбору инструментов, которые позволяют исследователю отобразить на графике именно те свойства и тенденции, которые нуждаются в отображении.

По способу построения графиков выделяют три основных вида данных объектов (рис. 1.11).

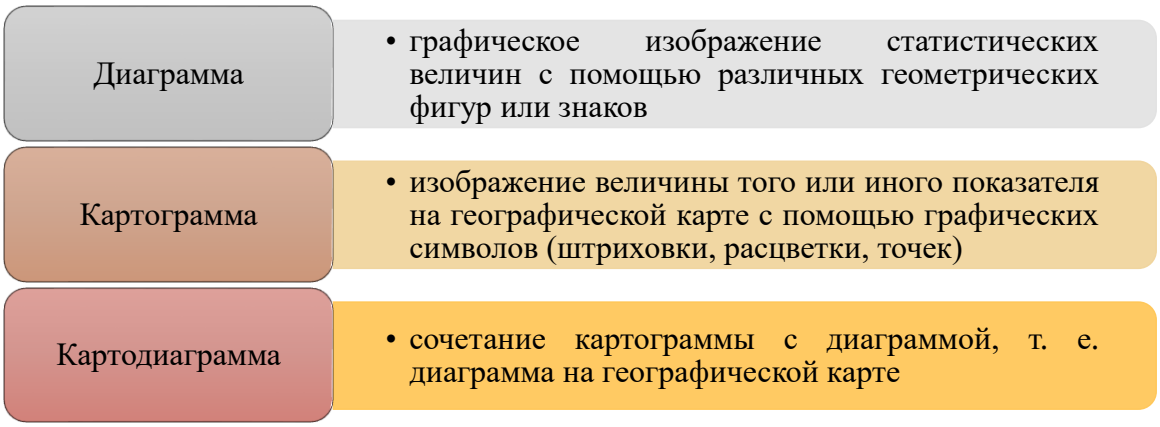


Рис. 1.11. Классификация графиков по способу построения

В зависимости от применяемых графических образов среди диаграмм различают столбиковые, плоскостные, объемные, линейные и др.

Для грамотного графического отображения вариационных рядов применяются линейные и плоскостные диаграммы, построенные в прямоугольной системе координат. Вариационный ряд можно изобразить,

как и любой ряд значений аргумента и функции, используя прямоугольную систему координат и строя точки с координатой (x_1 , f_1); (x_2 , f_2); (…); (x_n , f_n) в виде полигона, гистограммы, кумулятивной кривой (кумуляты), кривой Лоренца [2].

Полигон – графическое изображение дискретного вариационного ряда распределения. Полигон представляет собой замкнутый многоугольник, абсциссами вершин которого являются значения варьирующегося признака, а ординатами — соответствующие им частоты (рис. 1.12).

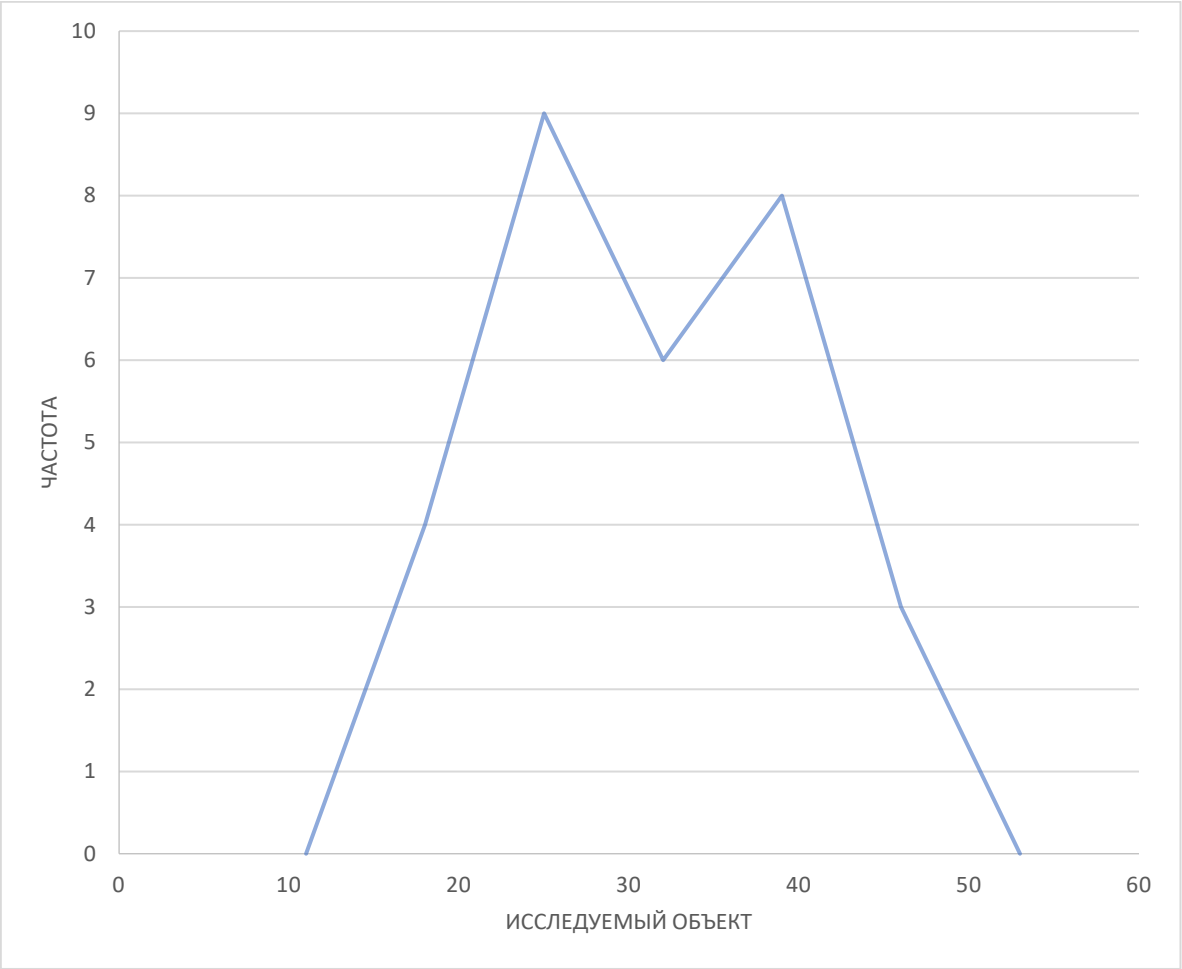


Рис. 1.12. Пример полигона частот

Гистограмма представляет собой графическое изображение интервального вариационного ряда. При ее построении на оси абсцисс откладывают не значения признака, а границы интервалов значений

признака. По оси ординат откладывают частоты (частости) или плотности распределения (все зависит от вида интервального ряда). Если ряд интервальный с равными интервалами, то на оси ординат откладывают частоты (частости), т. е. строят прямоугольники с высотой, равной частоте (частости) заданного интервала. Если ряд интервальный с неравными интервалами, то строят гистограмму плотностей распределения, поскольку в ряду с неравными интервалами именно плотность дает точное представление о количестве единиц в каждом из интервалов. Площадь всей гистограммы, таким образом, численно равна сумме частот или численности единиц в совокупности. Примет гистограммы приведен на рис. 1.13.

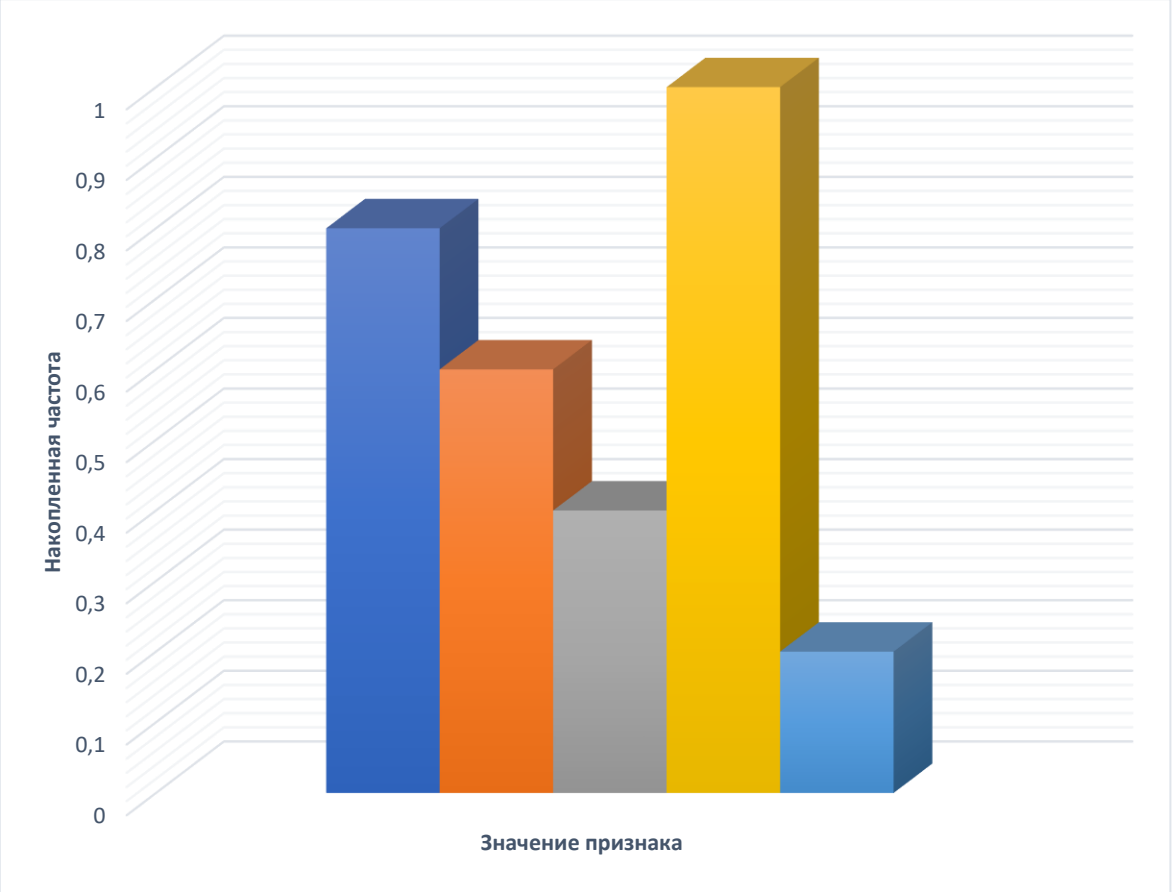


Рис. 1.13. Пример гистограммы

Кумулятивная кривая (кумулята) – кривая, характеризующая динамику накопленной частоты или частости. По оси абсцисс откладывают варианты значений признака (в интервальном ряду – верхние границы интервалов), а на оси ординат – соответствующие накопленные частоты или частости. Полученные точки соединяют отрезками и получают график, который называется кумулятой или кумулятивной кривой (рис. 1.14) [2].

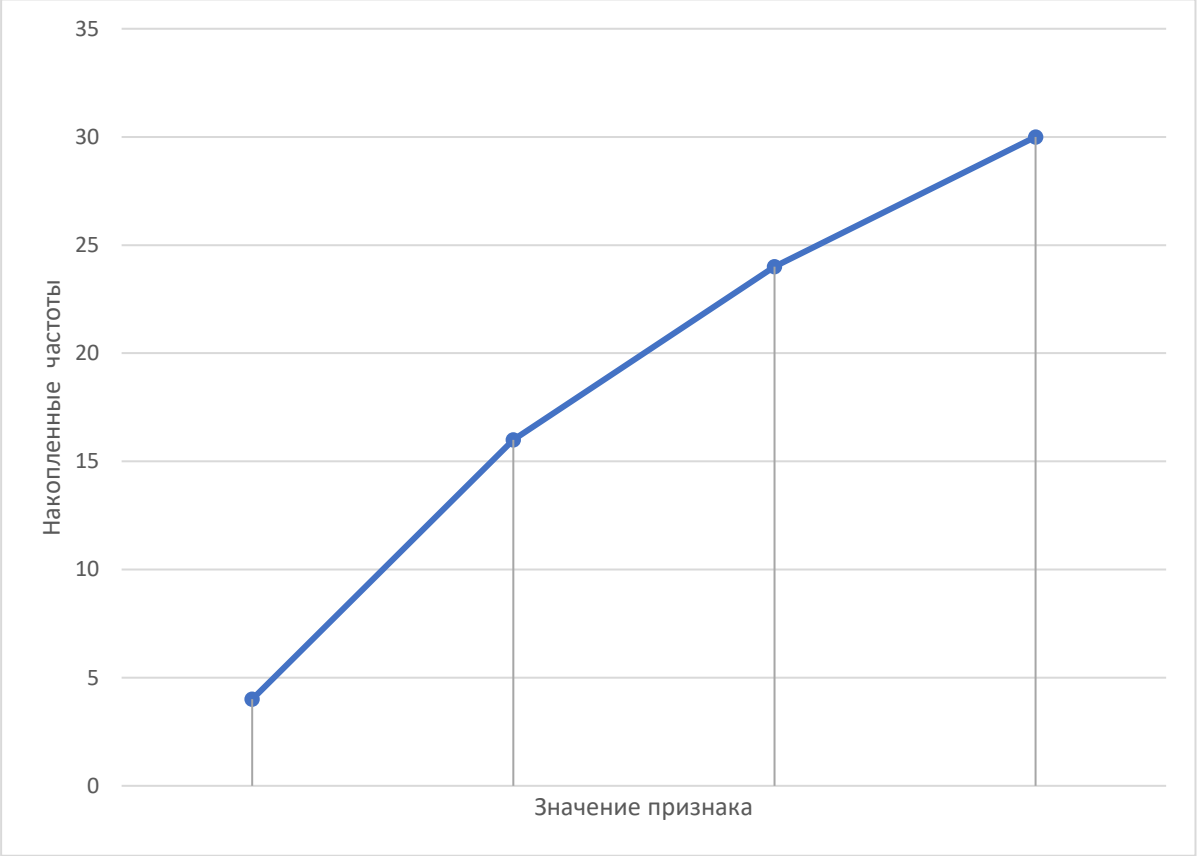


Рис. 1.14. Пример кумуляты

Кривая Лоренца — это график, который используется на практике для характеристики уровня относительной концентрации тех или иных явлений в совокупности. Построение такого графика предполагает указание значений накопленных частностей выделенных групп и значений накопленных долей признака в общем объеме совокупности (в процентах) (рис. 1.15).

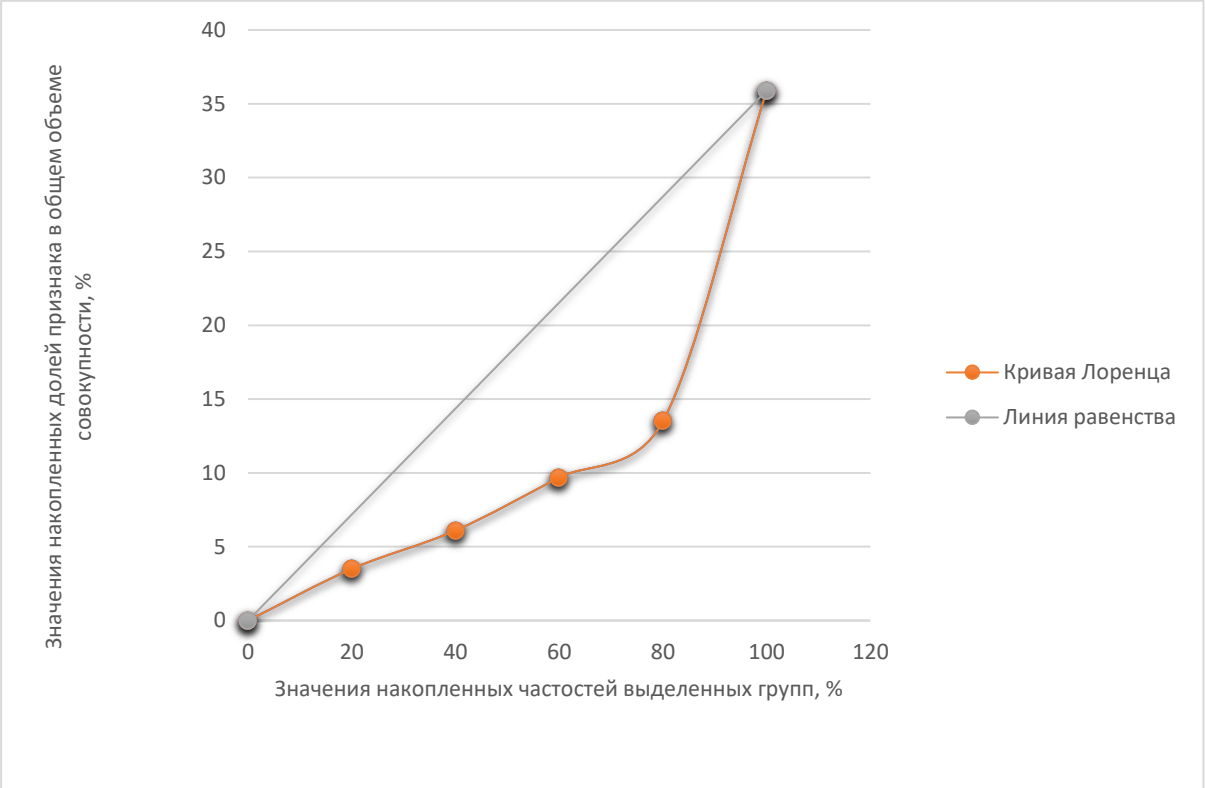


Рис. 1.15. Пример кривой Лоренца

Таким образом, были рассмотрены основные приемы грамотного представления данных при проведении статистического наблюдения. На основании вышеизложенного можно сделать вывод о том, что выбор инструмента представления данных оказывает существенное влияние на процедуру дальнейшего исследования показателей фирмы с использованием статистических методов.

ПРАКТИЧЕСКАЯ ЧАСТЬ

1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Данный раздел содержит описание первичного одномерного статистического анализа данных, который позволяет сформировать представление о наборе данных, а также представлены методы для визуализации анализируемой информации.

Существует достаточно много инструментов, позволяющих упростить работу с исходными данными и сделать их более понятными. Мы рассмотрим некоторые из них.

Средние величины

Средняя величина — обобщающая числовая характеристика изучаемого количественного признака по всем единицам статистической совокупности. В средних погашаются индивидуальные различия единиц совокупности.

Отличительной чертой средних величин является то, что они позволяют проводить сравнение нескольких различных статистических совокупности по какому-то признаку. Применение средних позволяет определить и выявить общие черты и закономерности [14].

Существует значительное количество данных величин, применение которых зависит от решаемых задач и характера анализируемых данных.

Рассмотрим отдельные их них, которые будут позднее представлены в функциях программных комплексов.

Средняя величина (средняя арифметическая простая)

Данный вид средней величины применяется в том случае, когда являются известными значения непосредственно анализируемого признака (x) и число единиц совокупности с данным значением (f).

Следует отметить, что применение является корректным если каждый x встречается только один раз (в этом случае f=1), а также в том случае, когда неизвестно сколько единиц совокупности имеют данное значение признака, т.е. исходные данные пока не упорядочены [103].

Формула для расчета данной средней

$$\bar{x} = \frac{\sum x_i}{n} \qquad (1.1)$$

где $\bar{x}$ – средняя арифметическая простая;

x – значение исследуемого признака;

n- число единиц в анализируемой совокупности.

Пример

Компания, работающая на нескольких рынках, по итогам года получила следующие данные по прибыли в разрезе рынков:

- рынок А – прибыль составила 1 000 тыс. руб.;
- рынок Б – прибыль составила 1 100 тыс. руб.;
- рынок В – прибыль составила 950 тыс. руб.

Необходимо определить среднюю прибыль компании по рынкам.

Решение

Применив формулу (1.1), производим расчет:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1000 + 1100 + 950}{3} = 1016{,}7 \text{ тыс. руб.}$$

Таким образом, становится возможным сделать вывод, что фирма получает в среднем 1016,7 тыс. руб. прибыли.

Средняя гармоническая величина

Данный вид средней величины применяется в том случае, когда при анализе имеются данные не только анализируемого признака x, но и его объем (W). Последний представляет собой произведение значения признака на число единиц данного признака.

Следует отметить, что бывает простая и взвешенная вариации данного показателя [62].

Простая гармоническая средняя рассчитывается в том случае, когда объем признаков одинаков.

Формула для расчета данной средней

$$\bar{x} = \frac{n}{\sum \frac{1}{x_i}} \qquad (1.2)$$

Взвешенная гармоническая рассчитывается в том случае, когда исходные данные содержат значения анализируемого признака и объем признака для каждого значения признака [41].

Формула для расчета данной средней

$$\bar{x} = \frac{\sum W_i}{\sum \frac{W_i}{x_i}} \qquad (1.3)$$

где W_i – объем признака, вес варианты.

Пример

Компания, занимающаяся транспортными услугами, имеет данные, что автомобиль, загрузившись в пункте А товаром в полном объеме, прибыл в пункт Б со скоростью 60 км/ч, обратно будучи незагруженным – со скоростью 75 км/ч. Необходимо определить среднюю скорость, с которой перемещался автомобиль компании.

Решение

Применив формулу (1.2), производим расчет средней сокрости движения автомобиля компании:

$$\bar{x} = \frac{n}{\sum \frac{1}{x_i}} = \frac{2}{{}^{1}/_{60} + {}^{1}/_{75}} = 66{,}7 \frac{\text{км}}{\text{ч}}$$

Таким образом, средняя скорость движения автомобиля компании составила 66,7 км/ч

Пример

По представленным данным о выработке трех проектных команд определить средний размер оплаты труда:

- команда А – при месячном фонде оплаты труда 2 700 тыс. руб. средняя заработная плата составила 90 тыс. руб.;
- команда Б - при месячном фонде оплаты труда 2 400 тыс. руб. средняя заработная плата составила 80 тыс. руб.;
- команда В - при месячном фонде оплаты труда 3 600 тыс. руб. средняя заработная плата составила 80 тыс. руб.

Решение

Применив формулу (1.3), производим расчет

$$\bar{x} = \frac{\sum W_i}{\sum \frac{W_i}{x_i}} = \frac{2700 + 2400 + 3600}{\frac{2700}{90} + \frac{2400}{80} + \frac{3600}{80}} = 82{,}9 \text{ тыс. руб.}$$

Таким образом, средняя заработная плата по трем проектным командам составила 82,9 тыс. руб.

Средняя геометрическая величина

Средняя геометрическая применяется в том случае, если анализируемая величина представлена относительными показателями (коэффициентами) или значения показателя значительно отличаются друг от друга [9].

Данная средняя бывает простой и взвешенной.

Простая средняя геометрическая применяется в том случае, когда расстояния между анализируемыми показателями равные (они имеют равные веса и равнозначны). Также можно отметить, что показатели не имеют равных значений [21].

Для расчета используется формула (1.4):

$$\bar{x} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} \qquad (1.4)$$

Пример

Существуют данные о темпах роста величины валового регионального продукта Владимирской области. Определить среднегодовой темп изменения данного показателя.

Индексы физического объема валового регионального продукта в 2017-2022 гг. (в постоянных ценах; в процентах к предыдущему году) составили: 2017 год - 100,7; 2018 год - 100,5; 2019 год - 106,4; 2020 год - 99,8; 2021 год - 112,9; 2022 год - 93,4.

Решение

Используя формулу (1.4), производим расчет простой геометрической средней:

$$\bar{x} = \sqrt[n]{x_1 \times \ldots \times x_n} = \sqrt[6]{100{,}7 * 100{,}5 * 106{,}4 * 99{,}8 * 112{,}9 * 93{,}4} = 102{,}1$$

Таким образом, средний темп изменения показателя валового регионального продукта Владимирской области за период 2017-2022 годы составил 102,1.

Средняя геометрическая взвешенная – применяется в том случае, когда расстояния между анализируемыми показателями разные (они имеют различные веса и неравнозначны).

Для расчета используется формула (1.5):

$$\bar{x} = \sqrt[\Sigma f_i]{x_1{}^{f_1} \times \ldots \times x_n{}^{f_n}} \quad (1.5)$$

где $f_1, \ldots f_n$ – веса показателей (расстояние между коэффициентами).

Пример

Существуют данные о темпах роста величины валового регионального продукта Владимирской области. Определить среднегодовой темп изменения данного показателя.

Индексы физического объема валового регионального продукта в 2017-2022 гг. (в постоянных ценах; в процентах к предыдущему году) составили: 2017 год - 100,7; 2018 год - 100,7; 2019 год - 106,4; 2020 год - 99,8; 2021 год - 112,9; 2022 год - 93,4.

Решение

Используя формулу (1.5), производим расчет:

$$\bar{x} = \sqrt[\Sigma f_i]{x_1{}^{f_1} \times \ldots \times x_n{}^{f_n}} = \sqrt[6]{100{,}7^2 * 106{,}4^1 * 99{,}8^1 * 112{,}9^1 * 93{,}4^1} = 102{,}1$$

Таким образом, средний темп изменения показателя валового регионального продукта Владимирской области за период 2017-2022 годы составил 102,1.

Мода и медиана

Медиана (M_e) — это величина признака, которая находится в середине ранжированного вариационного ряда, где отдельные значения признака (варианты) расположены в порядке их возрастания или убывания (по рангу) [81].

Как правило, данный показатель применяют в том случае, однородность анализируемой совокупности не подтверждена. Отметим, что в зависимости от вида ряда распределения медиана определяется поразному. Если дискретный ряд представлен нечетным числом членов, то медиана является срединным его членом, разделяя ряд на две равные части [11]. В том же случае, когда число членов дискретного ряда четной, необходимо использовать формулу (1.6).

$$M_e = x_{Me} + h_{Me}\frac{0{,}5\sum f_i - S_{Me-1}}{f_{Me}} \qquad (1.6)$$

где x_{Me} – нижняя граница медианного интервала;

h_{Me} – ширина медианного интервала;

$0{,}5\sum f_i$ – полусумма всех частот анализируемого множества;

S_{Me-1} - накопленная чистота в интервале, предшествующем медианному;

f_{Me} - частота в медианном интервале.

Пример

Необходимо произвести расчет медианы по данным по заработной плате:

Иванов И.И. – 36000 руб.

Боринов М.А. – 50000 руб.

Квадратов А.Е. – 55000 руб.

Птичкина М.А. – 65000 руб.

Москвина Н.Б. – 75000 руб.

Решение

Для расчета медианы, необходимо определиться с методом расчета. В данном случае число членом ряда является нечетным. Соответственно необходимо выбрать срединную величину. Данной величиной является 55000 руб.

Таким образом медианной заработной платой является 55000 руб.

Мода (M_o) – это самые встречающееся значение признака, иными словами, можно сказать, что мода является вариантой, которая соответствует максимальной величине частоты. Аналогично медиане, спо-

соб расчета моды зависит от типа ряда распределения. Если анализируемый ряд данных является дискретным, то показатель моды определяется без вычислений. В данном случае просто выбирается значение анализируемого признака, которое обладает наибольшей (максимальной) частотой [13].

Однако в случае, если анализируемые данные представлены интервальным вариационным рядом, то для определения моды используют формулу (1.7):

$$M_o = x_{Mo} + h_{Mo} \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \qquad (1.7)$$

где x_{Mo} – нижняя граница модельного интервала;
h_{Mo} – величина модельного интервала;
f_{Mo} – частота модального интервала;
f_{Mo-1} – частота интервала, предшествующего модальному;
f_{Mo+1} – частота интервала, следующего за модельным.

Пример

Необходимо определить значение показателя моды для ряда данных по среднедушевому доходу:

Доход до 25000 руб. – 1000 человек;
Доход 25000-30000 руб. – 2000 человек;
Доход 30000-35000 руб. – 2500 человек;
Доход 35000-40000 руб. – 3000 человек;
Доход более 40000 руб. – 1500 человек.

Решение

1. Определяем интервал с наибольшей частотой, в данном случае таковым является доход 35000-40000 руб., т.к. наибольшее количество человек получают именно его – 3000 человек.

2. Используя формулу (1.7), производим расчет:

$$M_o = x_{Mo} + h_{Mo} \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} =$$

$$= 35000 + 5000 \frac{3000 - 2500}{(3000 - 2500) + (3000 - 1500)} = 36250 \text{ руб.}$$

Таким образом, модальная заработная плата составила 36250 руб.

Показатели вариации

Для оценки надежности или типичности среднего значения признака анализируемой совокупности единиц применяют показатели вариации. Они позволяют оценить отклонение фактических или исходных данных от их расчетного (среднего) или теоретического значения.

Размах вариации

Размахом вариации называют разность между максимальным и минимальным значениями анализируемого признака. Следует отметить, что размах вариации, являясь абсолютной величиной, измеряется тех же единицах, что и анализируемая варианта [35].

Для расчета данного показателя применяют формулу (1.8):

$$R_{\text{вар}} = x_{max} - x_{min} \quad (1.8)$$

где $R_{\text{вар}}$ – размах вариации;

x_{max} – максимальное значение варианты;

x_{min} - минимальное значение варианты.

Размах вариации исчисляют с целью оценки типичности определенной ранее средней величины. Росту вероятности нетипичности средней соответствует увеличение значения размаха вероятности.

Среднее квадратическое отклонение

Среднее квадратическое отклонение (σ) – показатель, который оценивает на какое количество единиц с среднем фактические (исходные) значения вариант отключаются от определенного ранее среднего значения. Следует отметить, что показатель среднего квадратического отклонения является абсолютной величиной и, следовательно, измеряется в тех же единицах, что и анализируемые признаки [21].

Для расчета данного показателя используется формула (1.9):

$$\sigma = \sqrt{\sigma^2} \quad (1.9)$$

где σ^2 – дисперсия.

Дисперсия

Дисперсия – это показатель, который характеризует колеблемость значений изучаемой величины относительно исчисленной средней. Данная величина является безразмерной, то есть не имеет единиц измерения.

Следует отметить, что существует несколько способов расчета дисперсии, которые зависят от вида анализируемого признака [3].

Дисперсия количественного признака:

- при расчете используются формулы средней арифметической (простой или взвешенной). В данном случае используются формулы (1.10) - (1.11):

$$\sigma^2 = \frac{\sum(x - \bar{x})^2 f}{\sum f} \quad (1.10)$$

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{n} \quad (1.11)$$

- при расчете используется разность средней квадратов вариант и квадрата средней вариант. В данном случае используются формулы (1.12), (1.13), (1.14).

$$\sigma^2 = \overline{x^2} - \bar{x}^2 \quad (1.12)$$

$$\overline{x^2} = \frac{\sum x^2 f}{\sum f} \quad (1.13)$$

$$\bar{x}^2 = \left(\frac{\sum xf}{\sum f}\right)^2 \quad (1.14)$$

- при расчете применяют способ моментов. Следует отметить, что данный способ применяют только в том случае, когда анализируемый ряд является равноинтервальным, или в том случае, когда дискретный ряд имеет равные шаги (расстояния, лаги) между значениями анализируемого признака [4]. При использовании данного способа используют формулы (1.15) - (1.18).

$$\sigma^2 = h^2(m_2 - {m_1}^2) \quad (1.15)$$

$$m_2 = \frac{\sum (x')^2 f}{\sum f} \quad (1.16)$$

$$m_1 = \frac{\sum x' f}{\sum f} \quad (1.17)$$

$$x' = \frac{x - x_0}{h} \quad (1.18)$$

где m_1 – момент первого порядка;

m_2 – момент второго порядка;

x' - условная варианта;

h - величина интервала.

Пример

Рассчитать дисперсию несколькими способами на основе данных, представленных в таблице

Отдел	**Средняя выработка, ед/чел., x**	**Число сотрудников, чел., f**
А	**110**	**12**
Б	**120**	**10**
В	**130**	**14**
Г	**140**	**8**

Решение

Для расчета первым и вторым способами выполним предварительные расчеты, оформив их в табличной форме:

x	f	$\lvert x - \bar{x} \rvert$	$\lvert x - \bar{x} \rvert f$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$	x^2	$x^2 f$
110	12	14	168	196	2352	12100	145200
120	10	4	40	16	160	14400	144000
130	14	6	84	36	504	16900	236600
140	8	16	128	256	2048	19600	156800
-	44	-	420	-	5064	-	682600

Произведем расчет дисперсии:

- используя формулу (1.10), рассчитываем

$$\sigma^2 = \frac{\sum(x - \bar{x})^2 f}{\sum f} = \frac{5064}{44} = 115$$

- используя формулу (1.12), рассчитываем

$$\sigma^2 = \overline{x^2} - \bar{x}^2 = \frac{682600}{44} - 124{,}09^2 = 115$$

Для расчета третьим способом заполним расчетную таблицу

x	f	$x - x_0$	$x' = \frac{x - x_0}{h}$	$x'f$	$(x')^2$	$(x')^2 f$
110	12	-20	-2	-24	4	48
120	10	-10	-1	-10	1	10
130	14	0	0	0	0	0
140	8	10	1	8	1	8
-	44	-	-	-26	-48	66

x_0 = 130 единиц, h=10 единиц

Производим расчеты:

$$m_2 = \frac{\sum(x')^2 f}{\sum f} = \frac{66}{44} = 1{,}5$$

$$m_1 = \frac{\sum x' f}{\sum f} = \frac{-26}{44} = -0{,}59$$

$$\sigma^2 = h^2(m_2 - {m_1}^2) = 100 \times (1{,}5 - 0{,}59^2) = 115$$

Дисперсия альтернативного признака

Альтернативным признаком является такой признак, который может принимать только два значения. Для расчета дисперсии в данном случае применяется формула (1.19).

$$\sigma^2 = p(1 - p) \qquad (1.19)$$

где p – доля единиц, обладающих признаком.

Коэффициент вариации

Коэффициент вариации показывает отношение среднего квадратического отклонения к исчисленной средней [96]. Существует несколько форм представления данного показателя – в виде коэффициента или в процентах. Коэффициент вариации используют для оценки надежности средней, а также для сравнения изменчивости признака в нескольких анализируемых совокупностях [15].

Для расчета данного показателя используется формула (1.20):

$$V = \frac{|\sigma|}{\bar{x}} \qquad (1.20)$$

где V – коэффициент вариации;

σ – среднее квадратическое отклонение;

$\bar{x}$- среднее значение анализируемого признака.

Как правило, типичная средняя может использоваться для характеристики изучаемой совокупности в том случае, когда коэффициент вариации не превышает 0,33 или 33% [62]. Если данное значение превышено, то следует вывод о случайном характере исчисленной средней и, соответственно, неоднородности исходных данных. В таком случае можно порекомендовать исключение экстремальных значений, а также увеличить размер выборки [7].

Пример

Рассчитать коэффициент вариации на основе данных, представленных в таблице

Отдел	**Средняя выработка, ед/чел., *x***	**Число сотрудников, чел., *f***
А	**110**	**12**
Б	**120**	**10**
В	**130**	**14**
Г	**140**	**8**

Решение

Рассчитаем среднее квадратическое отклонение, выполнив предварительные табличные расчеты

x	f	$\lvert x-\bar{x}\rvert$	$\lvert x-\bar{x}\rvert f$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$	x^2	$x^2 f$
110	12	14	168	196	2352	12100	145200
120	10	4	40	16	160	14400	144000
130	14	6	84	36	504	16900	236600
140	8	16	128	256	2048	19600	156800
-	44	-	420	-	5064	-	682600

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\bar{x})^2 f}{\sum f}} = \sqrt{115} = \pm 10{,}7 \text{ единиц}$$

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{5460}{44} = 124{,}09 \text{ единиц}$$

$$V = \frac{|\sigma|}{\bar{x}} = \frac{10{,}7}{124{,}09} = 0{,}086$$

Таким образом, коэффициент вариации составил 8,6%.

Асимметрия и эксцесс

Коэффициент асимметрии характеризует симметричность в распределении наблюдений [71]. Асимметрию можно измерить с помощью коэффициента асимметрии Пирсона, который рассчитывается по формуле (1.21):

$$A = \frac{\bar{x}-M_o}{\sigma} = \frac{m_3}{\sigma^3} \qquad (1.21)$$

Соответственно возможны две ситуации:

- правосторонняя асимметрия – данная ситуация наблюдается при $\bar{x} > M_o$;

- левосторонняя асимметрия – данная ситуация наблюдается при $\bar{x} < M_o$;

- симметричность относительно средней величины в случае $\bar{x} = M_o$.

Эксцесс – крутизна распределения по сравнению с нормальной – вычисляется с помощью нормированного момента четвертого порядка [75].

Для расчета используется формула для сгруппированных данных (1.22) или (1.23) для не сгруппированных:

$$Ex = \frac{m_4}{\sigma^4} = \frac{\frac{\sum(x_i - \bar{x})^4}{n}}{\sigma^4} - 3 \quad (1.22)$$

$$Ex = \frac{m_4}{\sigma^4} = \frac{\frac{\sum(x_i - \bar{x})^4 f_i}{\sum f_i}}{\sigma^4} - 3 \quad (1.22)$$

Коэффициент эксцесса, как и коэффициент асимметрии, может принимать положительные, отрицательные или нулевые значения.

- нулевой коэффициент эксцесса обозначает такой же эксцесс, как у стандартного нормального распределения (то есть, «нормальный»);

- положительный коэффициент эксцесса обозначает, что распределение имеет более острую вершину (то есть у нас очень много средних значений, но тонкие «хвосты» — мало низких и высоких значений);

- отрицательный коэффициент эксцесса обозначает, что распределение имеет более пологую вершину (то есть у нас меньше средних значений и толстые «хвосты» — много низких и высоких значений) [39].

Пример

Рассчитать коэффициент асимметрии и эксцесс по данным. Распределение компаний по размеру активов характеризуется следующими данными:

Размер активов, млн руб.	До 200	200 - 300	300 - 400	400 - 500	500 - 600	600 и более	Итого
Удельный вес фирм, % к итогу	8	25	52	7	5	3	100

Рассчитаем среднее квадратическое отклонение:

$$\sigma = \sqrt{\frac{1087500}{100}} = 104{,}28$$

$$A = \frac{\bar{x} - M_o}{\sigma} = \frac{m_3}{\sigma^3} = \frac{882750}{104{,}28^3} = 0{,}78$$

$$Ex = \frac{m_4}{\sigma^4} = \frac{\frac{\sum(x_i - \bar{x})^4 f_i}{\sum f_i}}{\sigma^4} - 3 = \frac{521233125}{118265625} - 3 = 1{,}41$$

Таким образом, имеет место правосторонняя асимметрия (A=0,78) и острая вершина (Ex=1,41).

1.1. Описательная статистика в MS Excel

Состав MS Excel включает определенный набор инструментов для анализа данных для решения прикладных задач, в которые входят и статистические направления.

Для включения **Пакета анализа** необходимо:

- перейти **Файл** – **Параметры** и выбрать пункт **Надстройки** рис. 1.1.1);

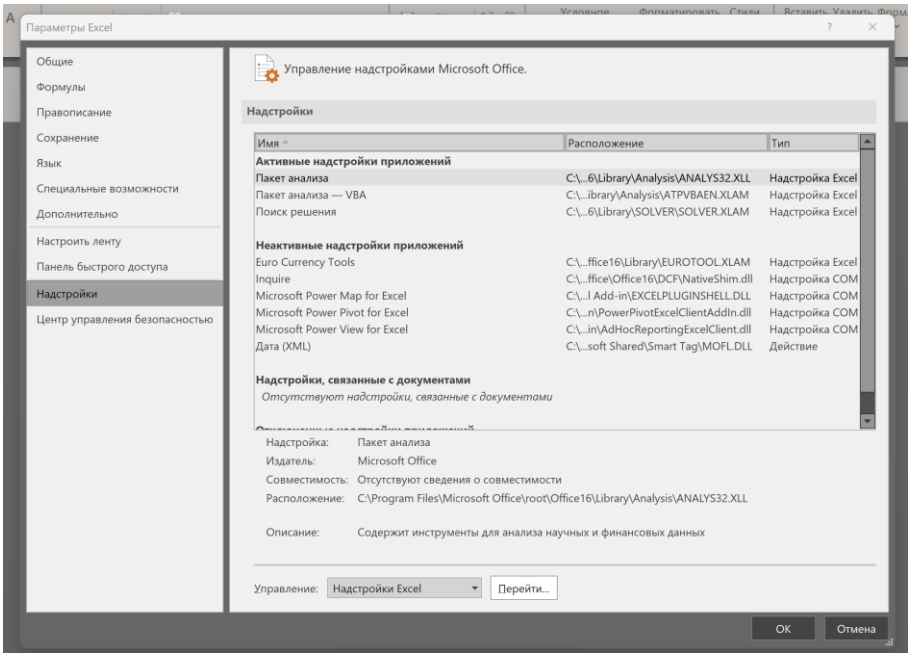


Рис. 1.1.1. Надстройки MS Excel

- далее необходимо в открывшемся меню установить галочки напротив тех пунктов, которые необходимы для проведения анализа (рис. 1.1.2) и нажать **ОК**.

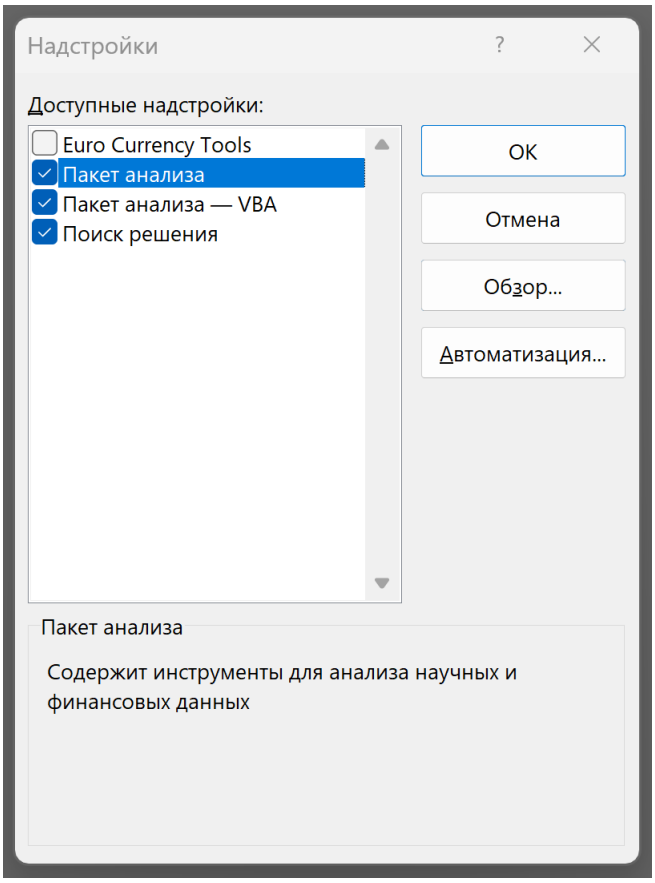


Рис. 1.1.2. Включение Пакета анализа данных

Далее надстройка **Анализ данных** готов к работе. Он находится на панели Данные (рис. 1.1.3).

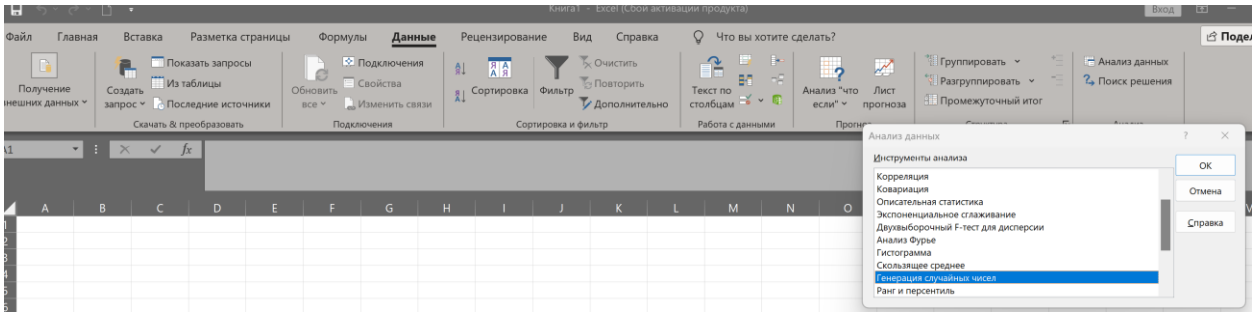


Рис. 1.1.3. Панель Анализа данных

Данный пакет включает в себя ряд функций:

1. Генерация случайных чисел;
2. Выборка;

3. Гистограмма;
4. Описательная статистика;
5. Скользящее среднее;
6. Экспоненциальное сглаживание;
7. Ковариационный анализ;
8. Корреляционный анализ;
9. Двухвыборочный F-тест для дисперсий;
10. Двухвыборочный Z-тест для средних;
11. Парный двухвыборочный /-тест для средних;
12. Двухвыборочный Z-тест с одинаковыми дисперсиями;
13. Двухвыборочный Z-тест с разными дисперсиями;
14. Однофакторный дисперсионный анализ;
15. Двухфакторный дисперсионный анализ с повторениями;
16. Двухфакторный дисперсионный анализ без повторений;
17. Регрессия;
18. Ранг и персентиль;
19. Анализ Фурье.

Инструмент Описательная статистика использует совокупность методов, позволяющих делать научно обоснованные выводы о числовых параметрах распределения генеральной совокупности по случайной выборке из нее. Пусть требуется изучить количественный признак генеральной совокупности.

Тогда с помощью инструмента **Описательная статистика** можно вычислить следующие параметры:

- среднее (статистическая оценка математического ожидания);
- стандартная ошибка (среднего);
- медиана (M_e) — значение признака, приходящееся на середину ранжированной (упорядоченной) совокупности;
- мода (M_o) — значение изучаемого признака, повторяющегося с наибольшей частотой;
- дисперсия выборки;
- стандартное отклонение (среднее квадратическое отклонение);
- эксцесс;
- асимметричность (асимметрия);
- интервал (размах выборки);
- минимальное значение выборки x_{min};
- максимальное значение выборки x_{max};

- сумма всех значений выборки;

- объем выборки n;

- наибольшее значение признака, имеющее разность с порядком x_{max} k единиц;

- наименьшее значение признака, имеющее разность с порядком x_{min} k единиц;

- уровень надежности (предельная ошибка выборки).

Пример

По данным таблицы 1.1.1, необходимо произвести вычисление параметров согласно возможностям инструментов **Описательной статистики**.

Таблица 1.1.1

Исходные данные для анализа

ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек)	
	2022 год
Белгородская область	1514,5
Брянская область	1152,5
Владимирская область	1325,5
Воронежская область	2285,3
Ивановская область	914,7
Калужская область	1070,9
Костромская область	571,9
Курская область	1067,0
Липецкая область	1126,3
Московская область	8591,7
Орловская область	700,3
Рязанская область	1088,9
Смоленская область	873,0
Тамбовская область	966,3
Тверская область	1211,2
Тульская область	1481,5
Ярославская область	1194,6
г. Москва	13104,2

Решение

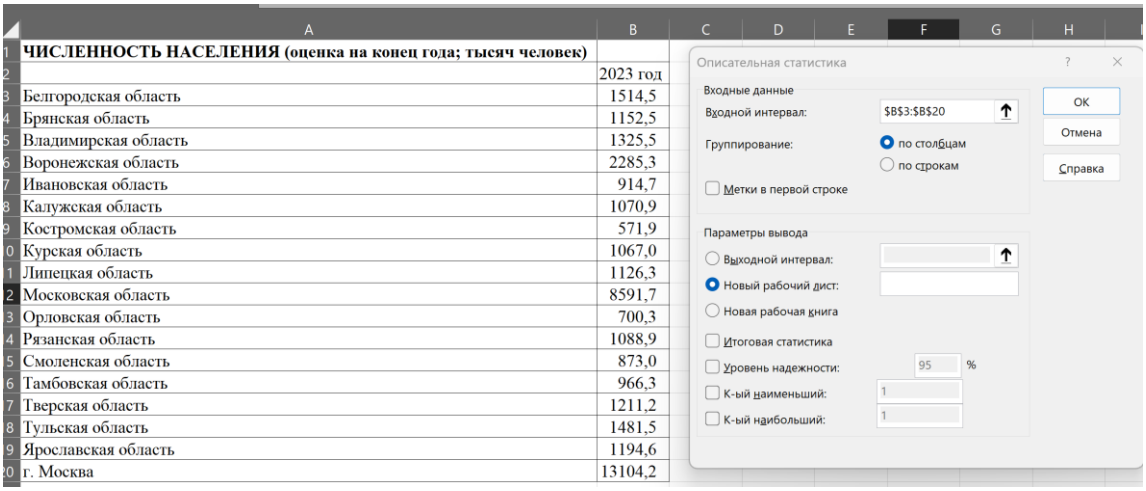
1. Выполнить цепочку действий: **Данные** – **Анализ данных** – **Описательная статистика** – **ОК**;

2. Задать **Входной интервал $B$3:$B$20** (рис. 1.1.4);

3. Группирование – По столбцам;

4. Поставить галочку напротив пунктов **Итоговая статистка**, **Уровень надежности** (со значением 95%), ***К*-ый наименьший** (со значением 1), ***К*-ый наибольший** (со значением 1).

5. Нажать **ОК**.



	A	B
1	**ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек)**	
2		2023 год
3	Белгородская область	1514,5
4	Брянская область	1152,5
5	Владимирская область	1325,5
6	Воронежская область	2285,3
7	Ивановская область	914,7
8	Калужская область	1070,9
9	Костромская область	571,9
10	Курская область	1067,0
11	Липецкая область	1126,3
12	Московская область	8591,7
13	Орловская область	700,3
14	Рязанская область	1088,9
15	Смоленская область	873,0
16	Тамбовская область	966,3
17	Тверская область	1211,2
18	Тульская область	1481,5
19	Ярославская область	1194,6
20	г. Москва	13104,2

Рис. 1.1.4. **Входной интервал** в модуле **Описательная статистика**

На новом листе будет представлено решение в табличной форме (рис. 1.1.5).

Столбец1	
Среднее	2235,572222
Стандартная ошибка	765,59452
Медиана	1139,4
Мода	#Н/Д
Стандартное отклонение	3248,14246
Дисперсия выборки	10550429,44
Эксцесс	8,150473935
Асимметричность	2,924583737
Интервал	12532,3
Минимум	571,9
Максимум	13104,2
Сумма	40240,3
Счет	18
Наибольший(1)	13104,2
Наименьший(1)	571,9
Уровень надежности(95,0%)	1615,263245

Рис. 1.1.5. Пример вывода данных модуля **Описательная статистика**

В таблице представлены все возможные для расчета данные по модулю в соответствии с характером и объемом данных.

Таким образом, можно сделать вывод, что среднее значение численности населения по регионам Центрального федерального округа в 2022 году составило 2235,57 тысяч человек. Стандартная ошибка составила 765,59. Медианное значение по группе регионов 1139,4 тысяч человек. Стандартное отклонение 3248,14, дисперсия выборки 10550429,44. Эксцесс составил 8,15 (распределение имеет более острую вершину (то есть у нас очень много средних значений, но тонкие «хвосты» — мало низких и высоких значений)). Коэффициент асимметрии составил 2,92. Минимальное значение 571,9 тысяч человек (Костромская область), наибольшее значение 13104,2 тысяч человек (г. Москва). Сумма (общая численность по всем субъектам округа) - 40240,3 тысяч человек при числе регионов 18. Уровень надежности (95,0%) составил 1615,26.

1.2. Описательная статистика в Statistica

Statistica — программный пакет для статистического анализа, разработанный компанией StatSoft, реализующий функции анализа данных, управления данными, добычи данных, визуализации данных с привлечением статистических методов.

Инструменты **Описательной статистики** находятся в панели **Анализ** – **Основные статистики и таблицы** (рис. 1.2.1).

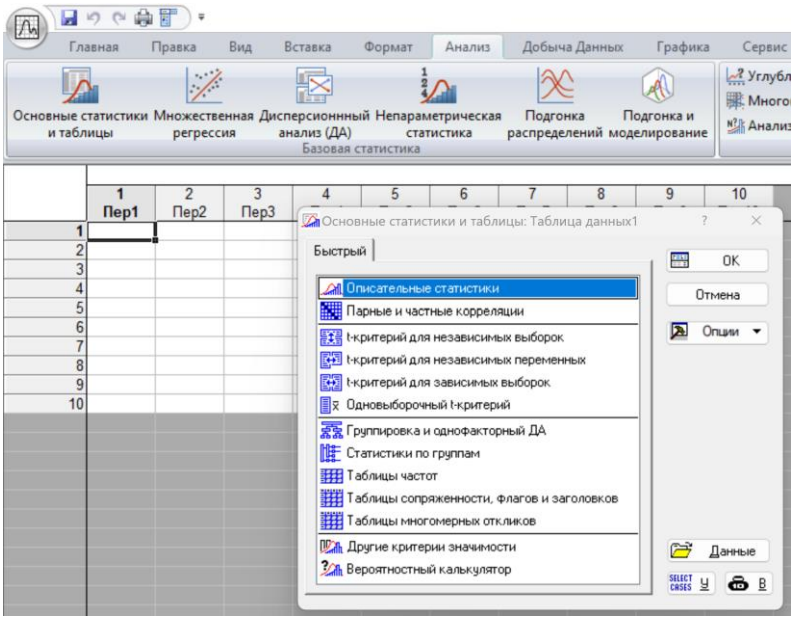


Рис. 1.2.1. Окно **Основные статистики и таблицы** Statistica 10.0

Пример

По данным таблицы 1.2.1, необходимо произвести вычисление параметров согласно возможностям инструментов **Описательной статистики**.

Таблица 1.2.1

Исходные данные для анализа

ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек)	
	2021 год
Белгородская область	1514,5
Брянская область	1152,5
Владимирская область	1325,5
Воронежская область	2285,3
Ивановская область	914,7
Калужская область	1070,9
Костромская область	571,9
Курская область	1067,0
Липецкая область	1126,3
Московская область	8591,7
Орловская область	700,3
Рязанская область	1088,9
Смоленская область	873,0
Тамбовская область	966,3
Тверская область	1211,2
Тульская область	1481,5
Ярославская область	1194,6
г. Москва	13104,2

Решение

1. Выполнить команды **Анализ - Основные статистики и таблицы – Описательные статистики – ОК** (Рис. 1.2.2).

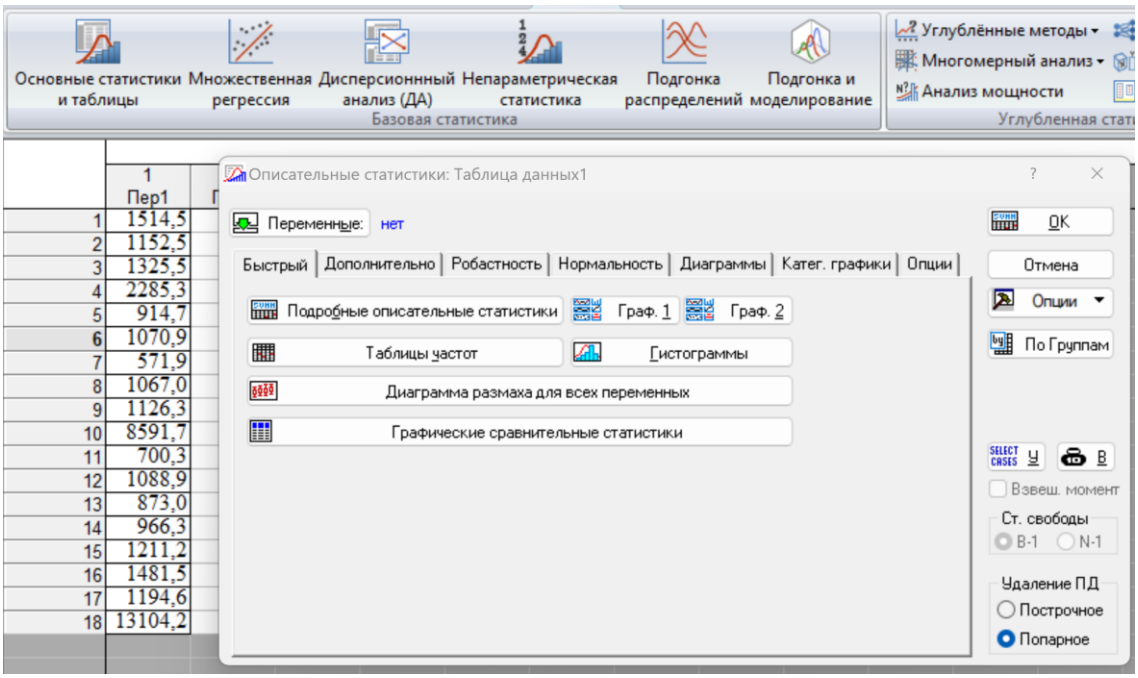


Рис. 1.2.2. Окно **Описательные статистики**

2. Необходимо нажать на **Переменные** и выбрать данные для анализа (рис. 1.2.3). В нашем случае для анализа представлен только один год, поэтому выбираем **Переменную 1** и нажимаем **ОК**.

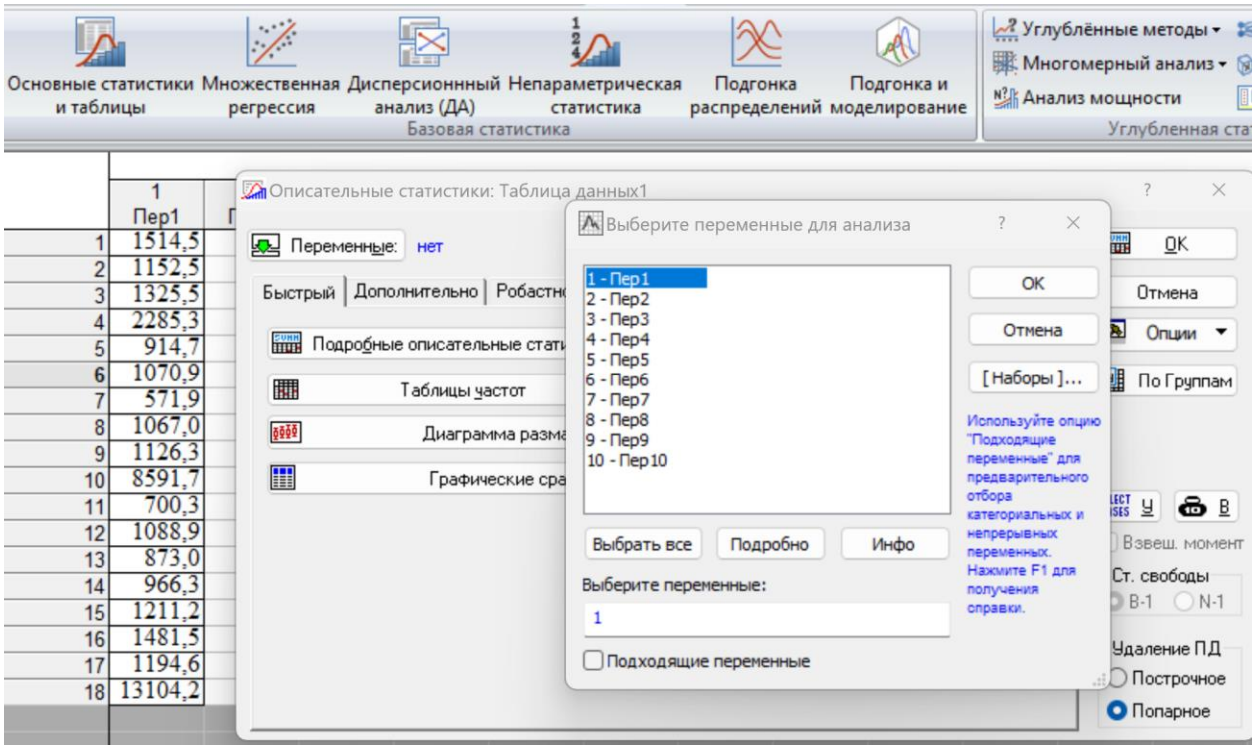


Рис. 1.2.3 Выбор переменных для **Описательной статистики**

Вкладка **Быстрый** предлагает инструменты для экспресс анализа по данным по умолчанию. Среди представленных инструментов:

- подробная описательная статистика – представляется в табличной форме основные данные: число наблюдений, среднее значение анализируемого признака, минимальное значение, максимальное значение, среднее отклонение (рис. 1.2.4);

	Описательные статистики (Таблица данных1)				
Переменная	**N набл.**	Среднее	Минимум	Максим.	Ст.откл.
Пер1	18	2235,572	571,9000	13104,20	3248,142

Рис. 1.2.4. Результаты операции **Быстрый - Подробная описательная статистика**

- таблица частот – в табличной форме представляется группировка данных по интервалам. Происходит определение частот, кумуляты, а также их доли (рис. 1.2.5);

	Таблица частот: Пер1 (Таблица данных1) K-C d=,42117, p<,01 ;Лиллиефорса p<,01					
Группа	**Частота**	Кумул. Частота	Процент допуст.	Кумул. % допуст.	% всех наблюд.	Кумул. % от всех
-2000,00<x<=0,000000	0	0	0,00000	0,0000	0,00000	0,0000
0,000000<x<=2000,000	15	15	83,33333	83,3333	83,33333	83,3333
2000,000<x<=4000,000	1	16	5,55556	88,8889	5,55556	88,8889
4000,000<x<=6000,000	0	16	0,00000	88,8889	0,00000	88,8889
6000,000<x<=8000,000	0	16	0,00000	88,8889	0,00000	88,8889
8000,000<x<=10000,00	1	17	5,55556	94,4444	5,55556	94,4444
10000,00<x<=12000,00	0	17	0,00000	94,4444	0,00000	94,4444
12000,00<x<=14000,00	1	18	5,55556	100,0000	5,55556	100,0000
Пропущ.	0	18	0,00000		0,00000	100,0000

Рис. 1.2.5. Результаты операции **Быстрый – Таблица частот**

- диаграмма размаха всех переменных – происходит вывод диаграммы размаха (рис. 1.2.6);

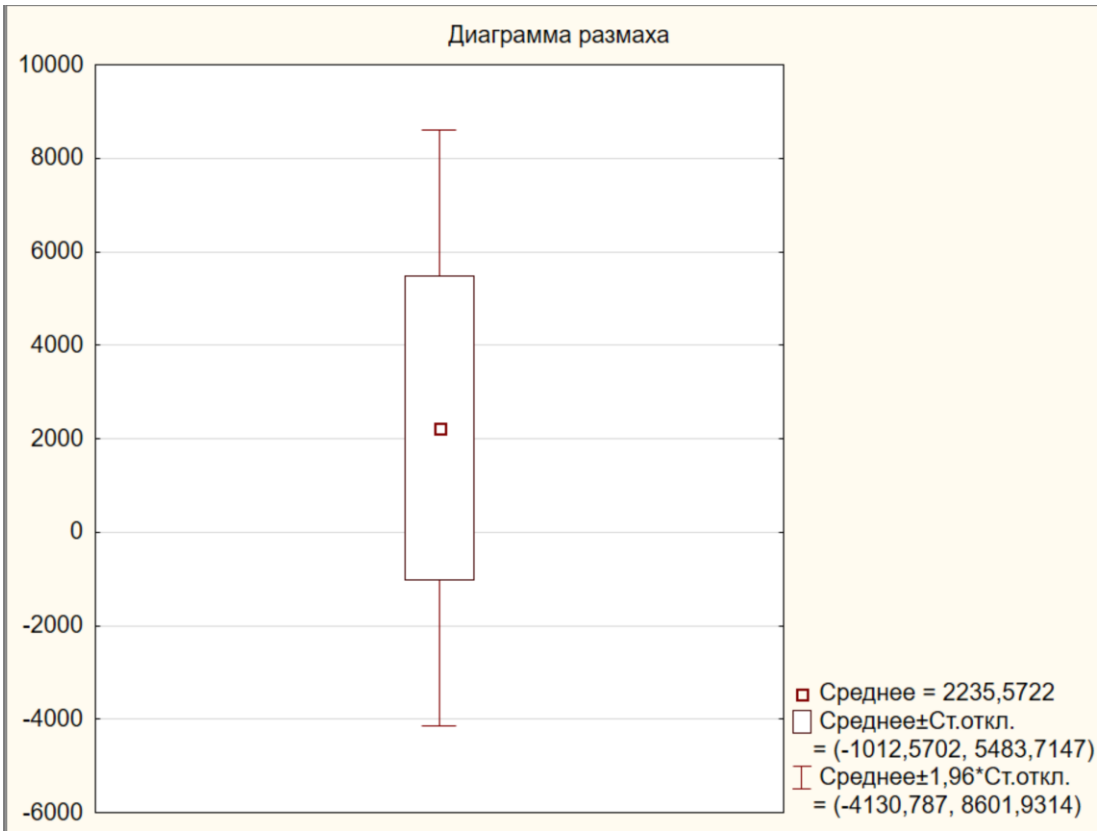


Рис. 1.2.6. Результаты операции **Быстрый – Диаграмма размаха всех переменных**

- графические сравнительные характеристики (рис. 1.2.7) – происходит вывод расчётных данных с графическим их представлением. Происходит расчет числа анализируемых данных, значения средней величины, медианной величины, максимального и минимального значение ряда, дисперсии, среднего отклонения, стандартной ошибки, величины асимметрии, эксцесса, верхней и нижней границ доверительного интервала (95%) стандартного отклонения, верхней и нижней границ доверительного интервала (95%) средней величины.

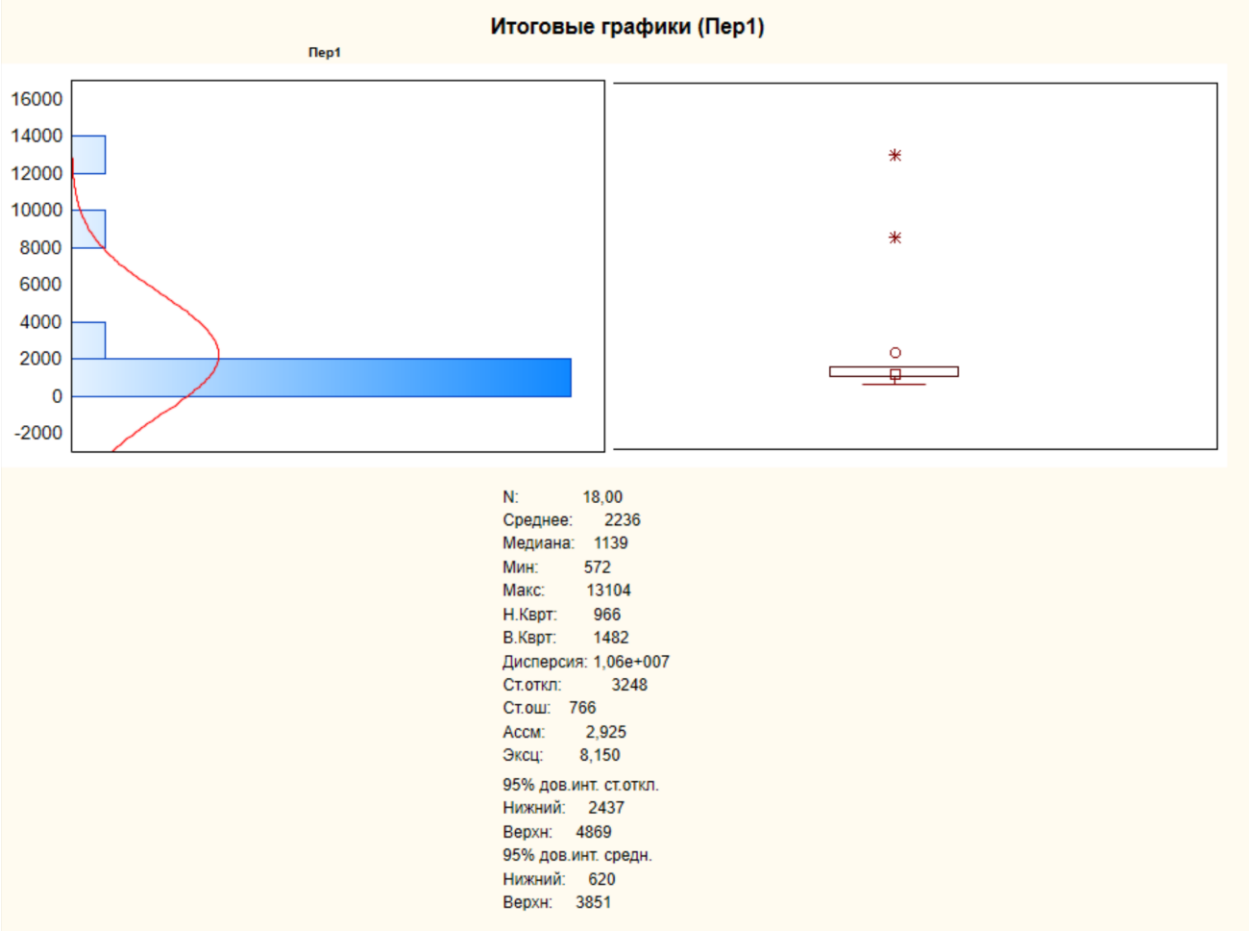


Рис. 1.2.7. Результаты операции **Быстрый – Графические сравнительные характеристики**

- графическое представление данных (2 видов) – происходит построение одного из двух видов графического представления данных.

Вкладка **Дополнительно** позволяет производить расчет дополнительных параметров, добавляя их к набору по умолчанию, или убирая не нужные на данной стадии анализа (рис. 1.2.9).

Также происходит расчет числа наблюдений, минимального и максимального значений, средней величины, среднего отклонения (рис. 1.2.7 – 1.2.8).

Вкладка **Робастность** (рис. 1.2.10) позволяет оценить нечувствительность к различным отклонениям и неоднородностям в выборке, связанным с теми или иными, в общем случае неизвестными, причинами.

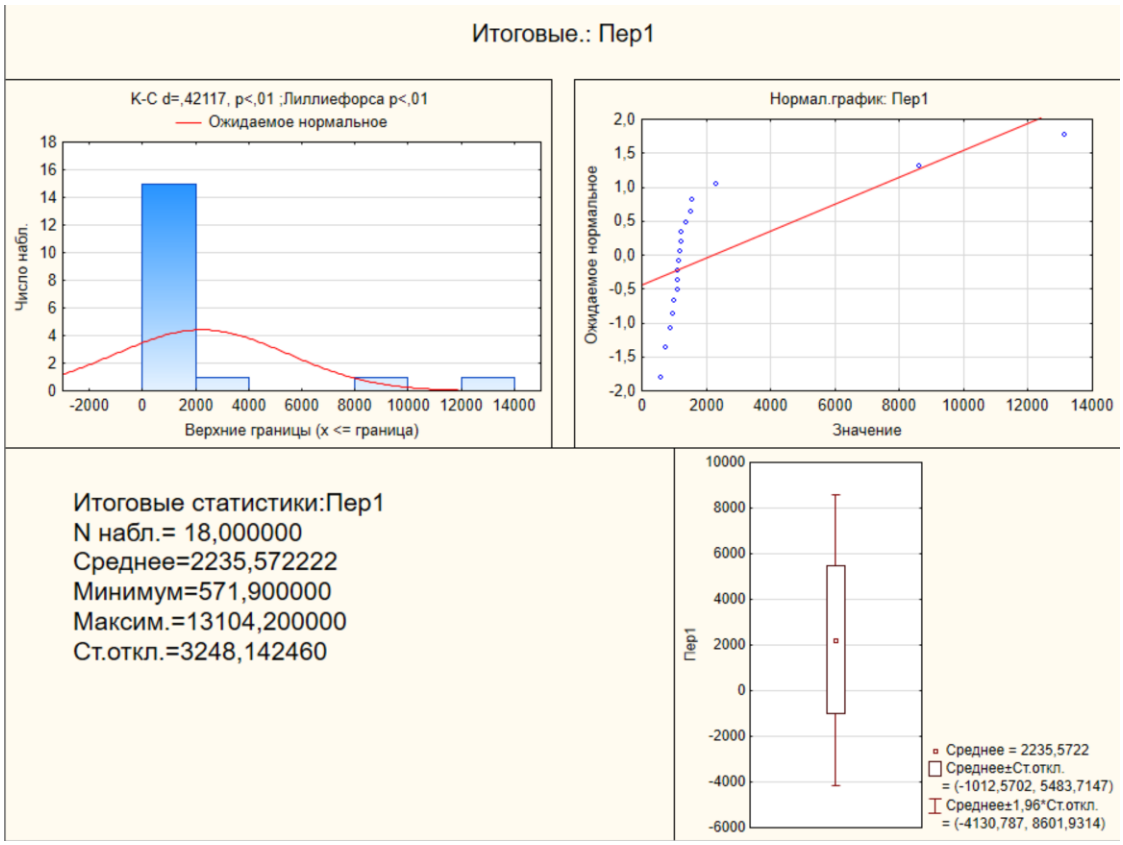


Рис. 1.2.8. Результаты операции **Быстрый – Граф.1**

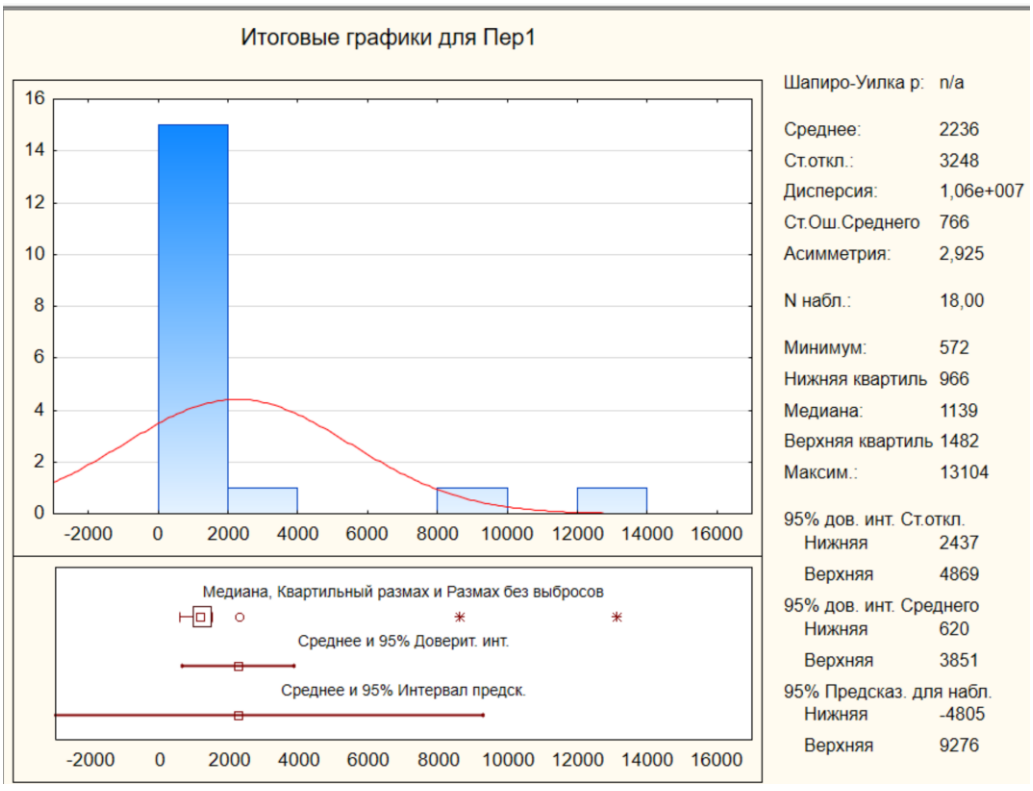


Рис. 1.2.9. Результаты операции **Быстрый – Граф.2**

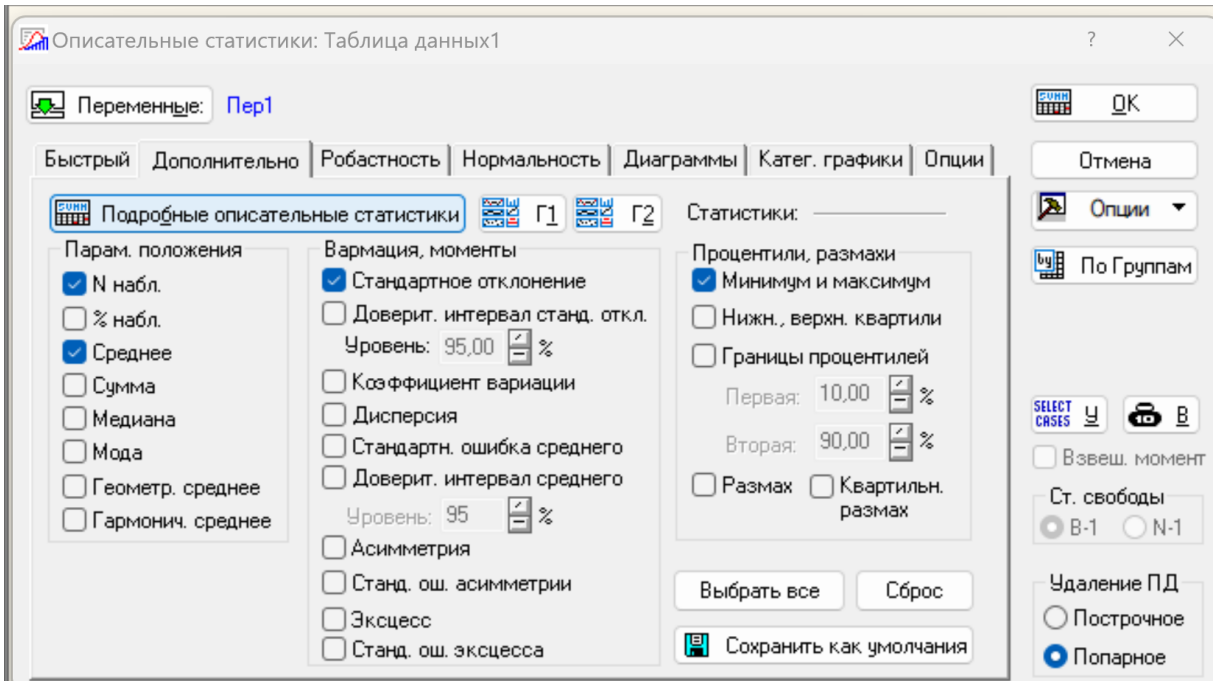


Рис. 1.2.10. Параметры по умолчанию на вкладке **Дополнительно**

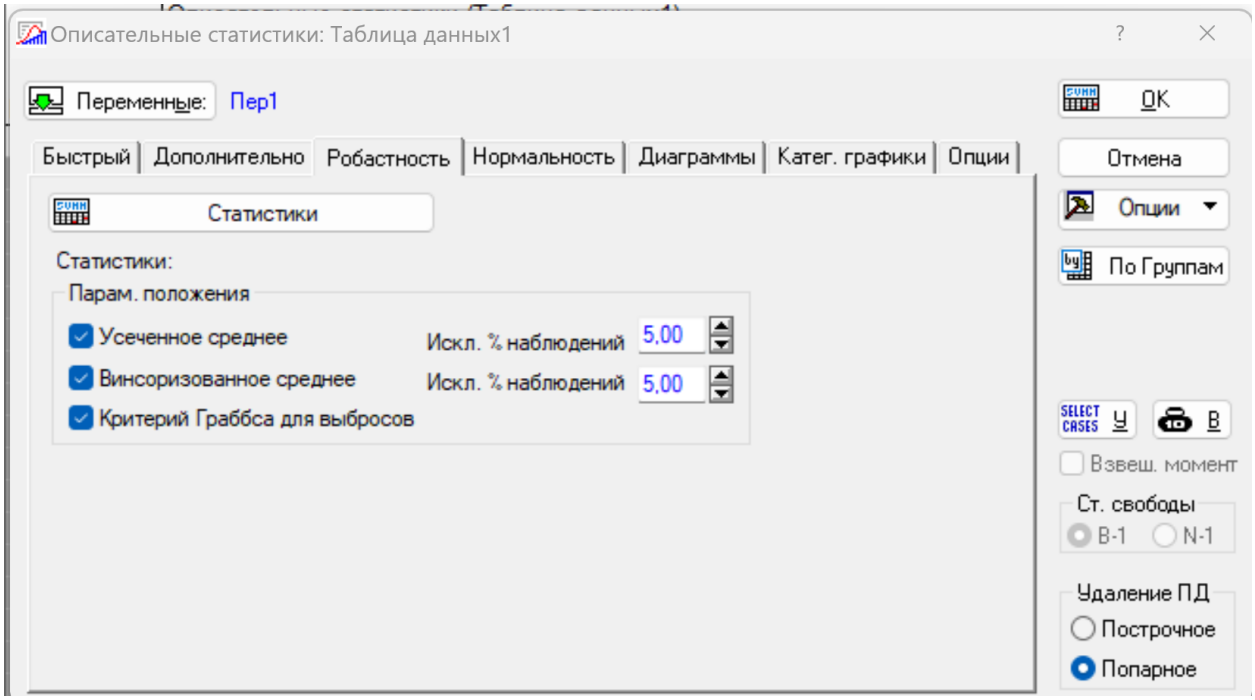


Рис.1.2.11. Параметры на вкладке **Робастность**

Следует отметить, что возможен расчет нескольких параметров, для определения которых необходимо установить галочки напротив желаемых параметров.

Усеченное среднее – один из методов, относящихся к линейным комбинациям порядковых статистик (L-оценки). Для его вычисления усредняются данные вариационного ряда выборки после удаления с обеих сторон определенной доли объектов (она находится в пределах от 5 до 25%)[1]. По умолчанию установлена величина 5%.

Винзоризованное среднее – один из методов, относящихся к линейным комбинациям порядковых статистик (L-оценки). Для его вычисления значения исходную выборку упорядочивают в определенном порядке (например, возрастания), затем с каждой стороны отсекается какой-то процент данных (обычно, берут по 10% или 25% с каждой стороны одинаково), далее убранные специально подобранным образом заменяются на значения из оставшихся чисел, затем вычисляется среднее по всей выборке. По умолчанию установлена величина 5%.

Критерий Грабса — это статистический тест, используемый для обнаружения выбросов в данных. Суть метода заключается в том, чтобы определить, является ли наблюдаемое значение выбросом или нет.

Шаги расчета критерия Грабса:

1. Вычислить среднее значение и стандартное отклонение выборки.

2. Найти максимальное или минимальное значение в выборке.

3. Рассчитать критерий Грабса:

$$G = \frac{x - \bar{x}}{\sigma} \qquad (1.2.1)$$

4. Найти критическое значение $G_{крит}$ из таблицы критических значений (таблица 1.2.2) для критерия Грабса (обычно при уровне значимости 0.05).

[1] Хьюбер П. Робастность в статистике. — М.: Мир, 1984.

Таблица 1.2.2

Критические значения для статистики критерия Граббса

n(L)	Одно наибольшее или одно наименьшее	
	Свыше 1%	Свыше 5%
3	1,155	1,155
4	1,496	1,481
5	1,764	1,715
6	1,973	1,887
7	2,139	2,020
8	2,274	2,126
9	2,387	2,215
10	2,482	2,290
11	2,564	2,355
12	2,636	2,412
13	2,699	2,462
14	2,755	2,507
15	2,806	2,549
16	2,852	2,585
17	2,894	2,620
18	2,932	2,651
19	2,968	2,681
20	3,001	2,709
21	3,031	2,733
22	3,060	2,758
23	3,087	2,781
24	3,112	2,802
25	3,135	2,822
26	3,157	2,841
27	3,178	2,859
28	3,199	2,876
29	3,218	2,893
30	3,236	2,908

5. Если рассчитанное значение G больше критического значения $G_{крит}$, то можно считать соответствующее значение в выборке выбросом.

Таким образом, критерий Грабса помогает определить наличие потенциальных выбросов в данных на основе их удаленности от среднего значения выборки.

Установи все галочки на вкладке Робастность нажимаем ОК и получаем результат в табличной форме (рис. 1.2.12).

Переменная	Описательные статистики (Таблица данных1)								
	N набл.	Среднее	Усеченн. средн. 5,0000%	Винсориз. средн. 5,0000%	Кр. Граббса Статист.	р-значение	Минимум	Максим.	Ст.откл.
Пер1	18	2235,572	1660,263	1992,011	3,346106	0,000291	571,9000	13104,20	3248,142

Рис. 1.2.12. Итоги расчета Параметров на вкладке **Робастность**

Вкладка **Нормальность** (рис. 1.2.13) позволяет устанавливать число выделяемых интервалов (по умолчанию установлено 10), также возможен расчет ожидаемых нормальных частот, критерий Колмогорова-Смирнова, критерий Лиллиефорса, критерий Шапиро-Уилка.

Критерий Колмогорова-Смирнова. Критерий Колмогорова – Смирнова[2] (критерий согласия) предназначен для сопоставления двух распределений:

- эмпирического с теоретическим, например, равномерным или нормальным;
- одного эмпирического распределения с другими эмпирическим распределением.

Критерий позволяет найти точку, в которой сумма накопленных расхождений между двумя распределениями является наибольшей, и оценить достоверность этого расхождения

[2] Гмурман В. Е. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – 479 с

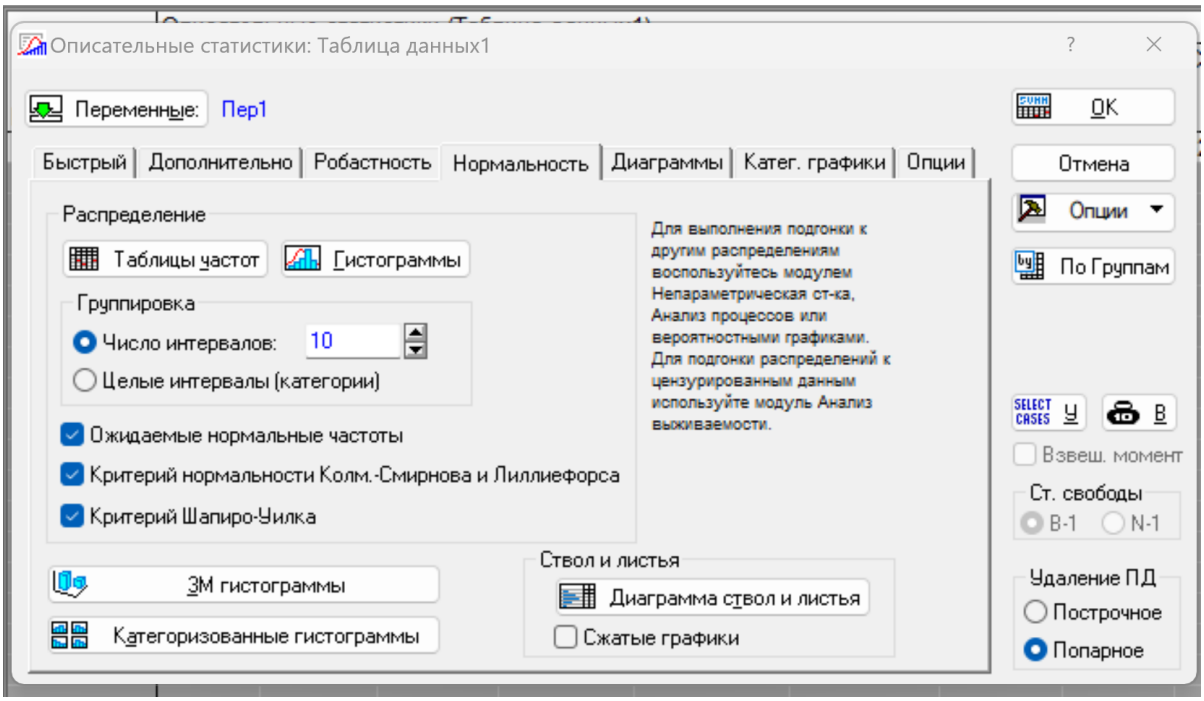


Рис. 1.2.13. Вкладка **Нормальность**

Критерий Лиллиефорса. Используется для проверки нулевой гипотезы о том, что выборка распределена по нормальному закону, когда параметры нормального распределения (математическое ожидание и дисперсия) априори неизвестны. Проверка гипотезы проводится следующим образом:

- оценивается выборочное среднее и дисперсия;

- находится максимальное отклонение между выборочной и теоретической интегральными функциями распределения;

- принимается решение, является ли статистически значимым наблюдаемое отклонение выборочной функции распределения от теоретической.

Критерий Шапиро-Уилка. Критерий Шапиро-Уилка является наиболее эффективным критерием проверки гипотезы о принадлежности выборки к нормальному закону распределения. Следует отметить, что критерий работает одинаково эффективно и при малых, и при больших объемах выборки. Критерий можно применять при объеме выборки $n \geq 3$.

Вкладка **Диаграммы** предоставляет выбор графического представления данных (рис. 1.2.14).

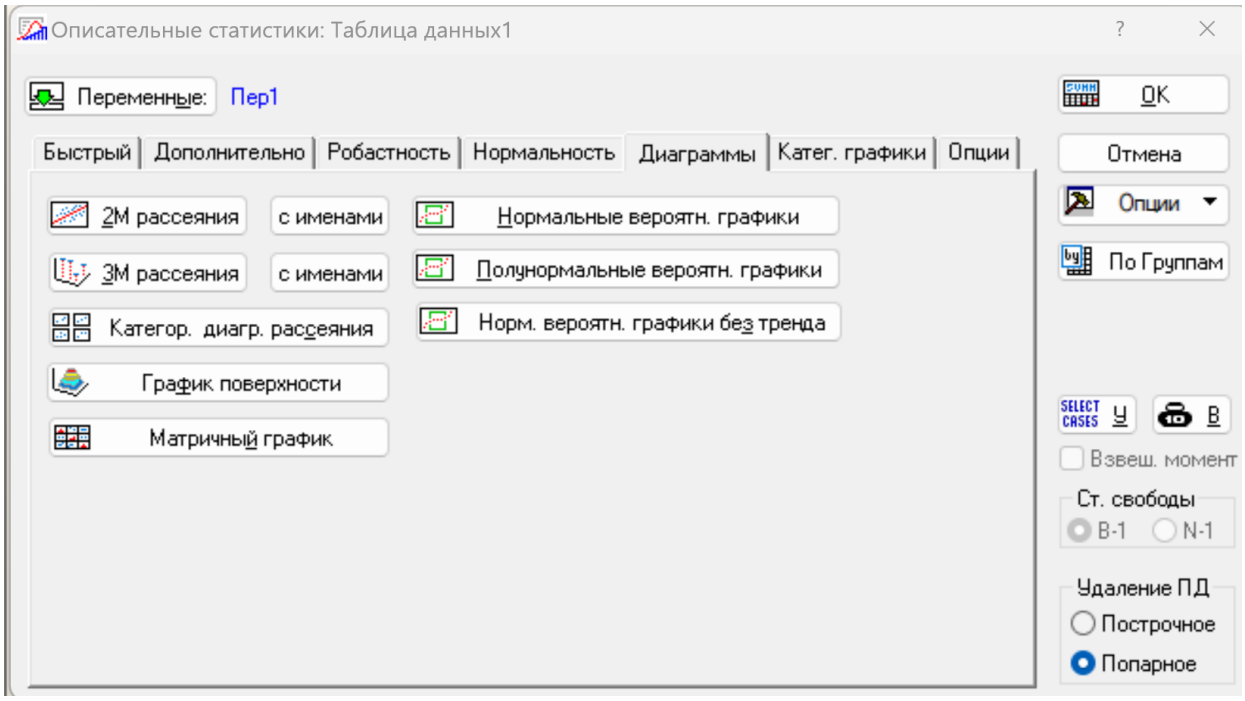


Рис. 1.2.14. Вкладка **Диаграммы**

В зависимости от задач, стоящих перед исследователем и самого набора данных выбираются те или иные графические формы.

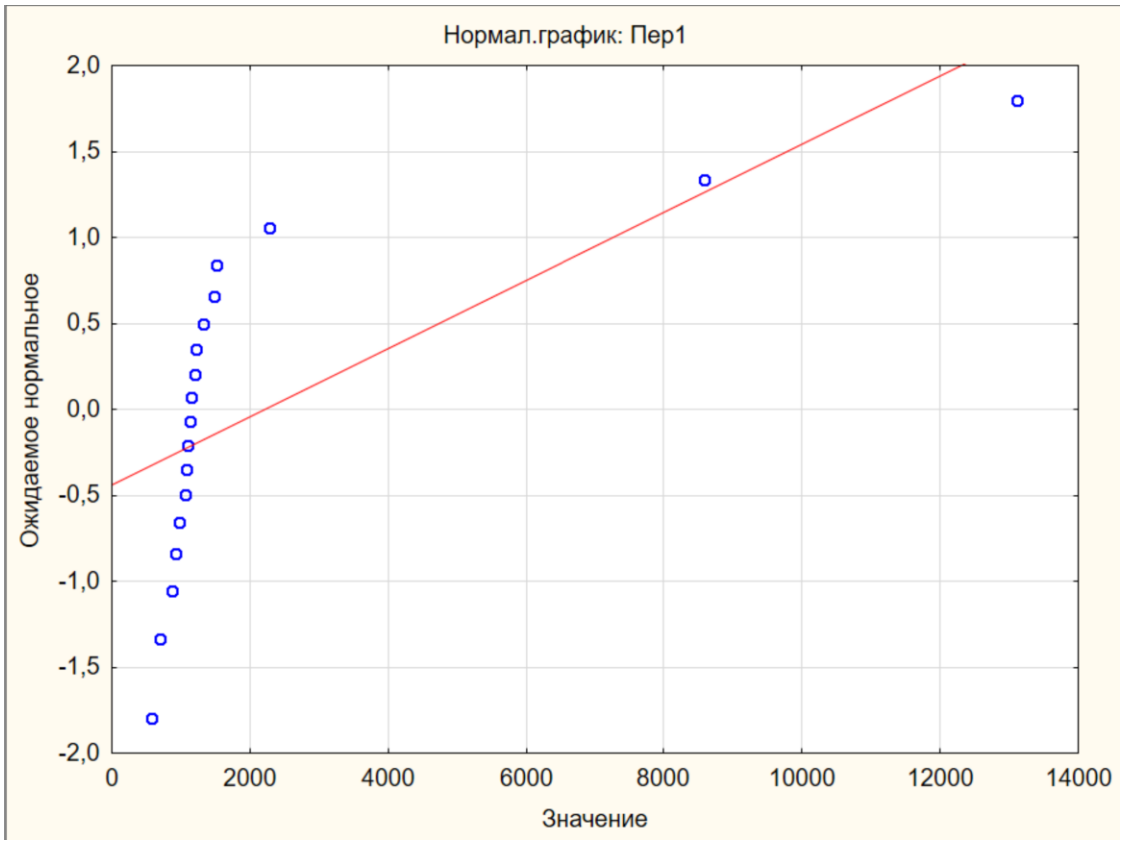


Рис. 1.2.15. Нормальный вероятностный график

Вкладка **Категоризованные графики** предоставляет выбор графических элементов для представления данных (рис. 1.2.16).

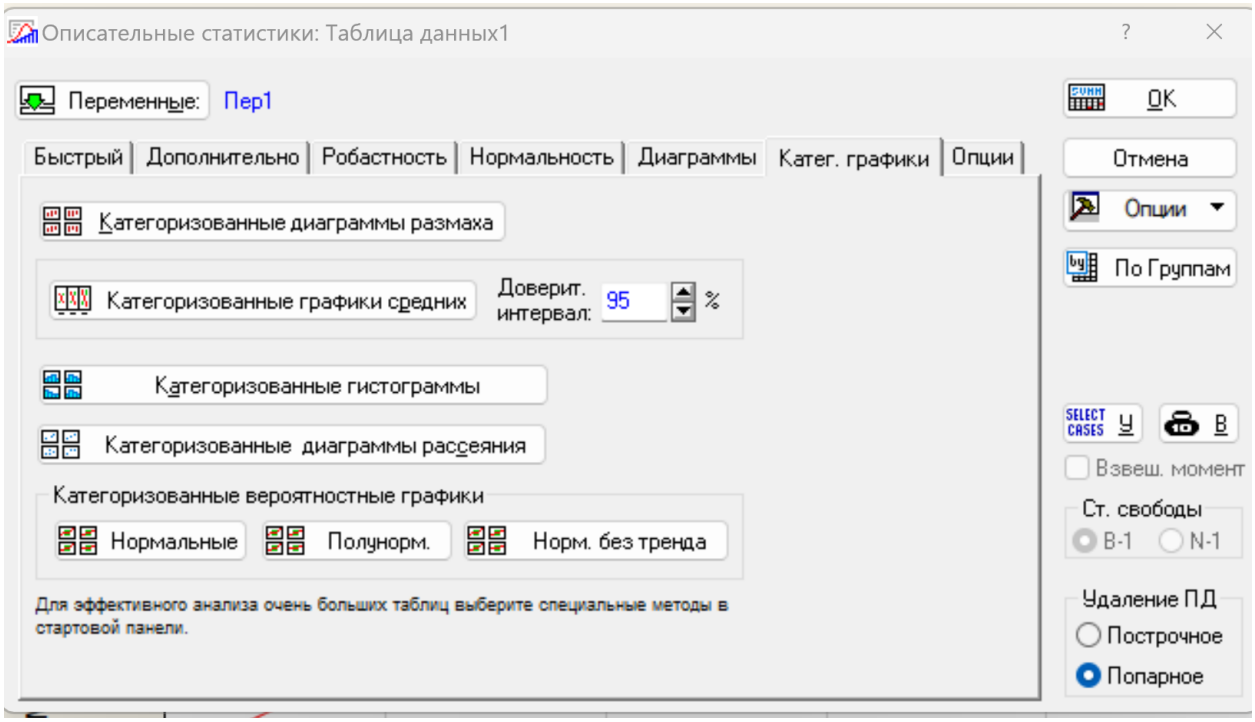


Рис. 1.2.16. **Категоризованные графики**

Таким образом, данный программный продукт предоставляет широкий выбор не только расчетных показателей, но и графического представления анализируемых данных.

Контрольные вопросы по теме

1. Что такое описательная статистика?
2. Какие основные задачи решает описательная статистика?
3. Какие методы используются для описания данных?
4. Что такое меры центральной тенденции?
5. Какие меры центральной тенденции вы знаете?
6. Что такое медиана и как её вычислить?
7. Как определить моду?
8. Какие меры изменчивости данных существуют?
9. Что такое дисперсия и стандартное отклонение?
10. Какие методы используются для визуализации данных?
11. Что такое гистограмма?
12. Как строится "ящик с усами" (box plot)?
13. Что такое квантили и как их интерпретировать?

14. Что такое корреляция и как её измеряют?

15. Что такое коэффициент корреляции Пирсона?

16. Что такое регрессионный анализ и как он используется?

17. Какие методы можно применить для обнаружения выбросов в данных?

18. Что такое нормальное распределение и как его определить?

19. Какие методы используются для проверки гипотез о данных?

20. Что такое доверительный интервал и как он строится?

21. Что такое статистическая значимость и как её определяют?

22. Какие типы данных могут быть описаны с помощью описательной статистики?

23. Какие проблемы могут возникнуть при анализе данных с помощью описательной статистики?

24. Какие плюсы и минусы есть у различных методов описания данных?

25. Какая роль описательной статистики в процессе принятия решений?

26. Как можно интерпретировать результаты описательной статистики?

27. Какие программные инструменты чаще всего используются для анализа данных с помощью описательной статистики?

28. Какие основные шаги следует выполнить при анализе данных с помощью описательной статистики?

29. Как можно определить асимметрию данных?

30. Что такое куртозис и как его можно использовать для анализа данных?

31. Какие методы можно применить для сравнения двух или более наборов данных?

32. Как можно определить выбросы в данных и как с ними работать?

33. Какие типы ошибок могут возникнуть при анализе данных с помощью описательной статистики?

34. В чем разница между параметрическими и непараметрическими методами анализа данных?

35. Как важно учитывать контекст при анализе данных с помощью описательной статистики?

36. Как можно использовать описательную статистику для прогнозирования будущих событий?

37. Какие методы используются для интерпретации результатов анализа данных?

38. Как можно определить, что данные имеют нормальное распределение?

39. Какие методы можно применить для заполнения пропущенных значений в данных перед их анализом?

40. Как можно использовать описательную статистику для выявления тенденций в данных?

41. Как можно определить, что данные имеют выбросы, и как с ними работать при анализе данных?

42. Какие методы можно использовать для сравнения нескольких групп данных между собой?

43. В чем разница между дескриптивной и инферентной статистикой?

44. Как можно использовать описательную статистику для принятия бизнес-решений?

45. Какие методы можно применить для обработки неоднородных данных перед их анализом с помощью описательной статистики?

46. Как можно оценить точность результатов анализа данных с помощью описательной статистики?

47. Как можно использовать описательную статистику для выявления аномалий в данных?

48. Как можно определить, что данные имеют линейную зависимость друг от друга?

49. Как можно использовать описательную статистику для выявления паттернов в данных?

50. Как можно использовать результаты анализа данных с помощью описательной статистики для принятия стратегических решений?

2. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. Понятие и виды корреляционных связей

Изучаемые явления и процессы требуют уточнения взаимодействия с позиций взаимовлияния и зависимостей. Существует достаточно большая классификация видов и типов связей. Однако наиболее общей и при этом крайне важной является разделение связей на две категории: функциональную и корреляционную.

Функциональная связь – это связь между двумя факторами (x и y), при которой каждому значению x соответствует одно или несколько значений y. То есть при изменении независимой величины зависимая меняется строго установленным образом. Как правило, данный тип связи возможно описать точными формулами, при этом следует отметить, что итоговое значение результативного признака зависит только от значений факторного признака или нескольких факторных признаков [107].

Стохастическая связь – это связь между факторами, которая не может быть задана формулой ввиду того, что она является не полной и не точной. Часто такие связи определяют как приближенные [95].

Корреляционная связь – частный вид стохастической связи, при которой при изменении факторного признака результирующая величина изменяется неслучайным образом [111]. Подобные связи становится возможным выявить при значительном количестве наблюдений за явлением.

Как правило, корреляционную связь подразделяют на:

- прямую (положительную) – при росте факторного признака результирующая величина также увеличивается;
- обратная (отрицательную) – с ростом факторного признака результирующий уменьшается.

По типу аналитической формы корреляционную связь делят на:

- линейную – описывается линейным уравнением $y=a+bx$;
- нелинейную – описывается нелинейными уравнениями. Частными случаями являются уравнения, графическим выражением которого является парабола, гипербола, логарифмическая или экспоненциальная функции.

По числу факторов взаимодействия выделяют:

- однофакторные модели – связь наблюдается между одним зависимым и одним независимым факторами;

- многофакторные модели – связь наблюдается между зависимым фактором и несколькими независимыми.

2.2. Общие положения корреляционно-регрессионного анализа

Основными задачами корреляционно-регрессионного анализа являются:

- определение наличия связи между факторным и результирующим показателями;

- определение формы зависимости между факторным и результирующим показателями;

- определение тесноты связи между факторами.

Применять корреляционно-регрессионный анализ становится возможным только при выполнении ряда условий [64]:

- наличие исходных данных за значимые период – в данном случае существует рекомендация, которая заключается, что на каждый факторный признак, участвующий в анализе должно приходиться не менее 6-7 периодов, за которые представлены данные;

- однородность анализируемой совокупности – как правило, данный критерий становится возможным оценить посредством расчета коэффициента вариации. Существуют рекомендации, что инструменты корреляционно-регрессионного анализа целесообразно применять только в том случае, когда значение коэффициента вариации меньше 0,6 (по некоторым источникам 0,8). В противном случае полученные модели будут не отражать существующие закономерности;

- наличие корреляционной связи между анализируемыми показателями;

- наличие нормального распределения единиц совокупности по анализируемым признакам – как правило, оценивают по показателю асимметрии.

2.3. Построение уравнения парной линейной регрессии

Перед построением уравнение регрессии необходимо оценить тесноту связи между анализируемыми признаками. Для этого применяется линейный коэффициент корреляции (2.2.1).

$$r = \frac{\sum xy - \sum x \frac{\sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \times \left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} \quad (2.2.1)$$

где r – линейный коэффициент корреляции;

По значению коэффициента линейной корреляции определяют тесноту связи и направление [23]. Он принимает значения в интервале от +1 до -1. Если значения коэффициента находятся в интервале от -1 до 0, то связь определяется на обратная (отрицательная), если от 0 до +1 – то как прямую (положительную). В случае если значение коэффициента равно 1 делают вывод о том, что анализируемая связь является функциональной. Если рассчитанное значение равно 0, делается вывод об отсутствии связи.

Для трактовки применяют шкалу Чэддока (оценка производится по модулю значения коэффициента линейной корреляции) [74]:

- от 0 до 0,2 – очень слабая связь;
- от 0,2 до 0,3 – слабая связь;
- от 0,3 до 0,5 – умеренная связь;
- от 0,5 до 0,7 – заметная связь;
- от 0,7 до 0,9 – сильная связь;
- свыше 0,9 – очень сильная связь.

Пример

По данным таблицы 2.2.1 произвести расчет коэффициента корреляции и определить тип и тесноту связи.

Таблица 2.2.1

Исходные данные для анализа

Период	Валовый региональный продукт, млн руб.	Численность населения (оценка на конец года; тысяч человек)
2000	42074,5	1507,0
2001	49941,8	1508,1
2002	62404,4	1511,9
2003	76054,5	1513,9
2004	114409,3	1511,7
2005	144987,8	1511,7

2006	178846,1	1514,2
2007	237013,3	1520,1
2008	317656,3	1526,3
2009	304345,3	1531,8
2010	398361,4	1532,4
2011	507839,8	1536,6
2012	545517,2	1541,9
2013	569006,4	1545,5
2014	619677,7	1549,8
2015	693379,4	1552,4
2016	778027,8	1555,5
2017	837306,8	1553,0
2018	911597,9	1550,9
2019	955329,2	1553,0
2020	997330,9	1545,5
2021	1354810,5	1536,5

Решение

Произведем промежуточные расчеты в табличной форме

	y	**x**	**x×y**	**x^2**	**y^2**
2000	42074,5	1507,0	63406271,50	2271049,00	1770263550,25
2001	49941,8	1508,1	75317228,58	2274365,61	2494183387,24
2002	62404,4	1511,9	94349212,36	2285841,61	3894309139,36
2003	76054,5	1513,9	115138907,55	2291893,21	5784286970,25
2004	114409,3	1511,7	172952538,81	2285236,89	13089487926,49
2005	144987,8	1511,7	219178057,26	2285236,89	21021462148,84
2006	178846,1	1514,2	270808764,62	2292801,64	31985927485,21
2007	237013,3	1520,1	360283917,33	2310704,01	56175304376,89
2008	317656,3	1526,3	484838810,69	2329591,69	100905524929,69
2009	304345,3	1531,8	466196130,54	2346411,24	92626061632,09
2010	398361,4	1532,4	610449009,36	2348249,76	158691805009,96
2011	507839,8	1536,6	780346636,68	2361139,56	257901262464,04
2012	545517,2	1541,9	841132970,68	2377455,61	297589015495,84
2013	569006,4	1545,5	879399391,20	2388570,25	323768283240,96
2014	619677,7	1549,8	960376499,46	2401880,04	384000451877,29
2015	693379,4	1552,4	1076402180,56	2409945,76	480774992344,36
2016	778027,8	1555,5	1210222242,90	2419580,25	605327257572,84
2017	837306,8	1553,0	1300337460,40	2411809,00	701082677326,24
2018	911597,9	1550,9	1413797183,11	2405290,81	831010731284,41

2019	955329,2	1553,0	1483626247,60	2411809,00	912653880372,64
2020	997330,9	1545,5	1541374905,95	2388570,25	994668924094,81
2021	1354810,5	1536,5	2081666333,25	2360832,25	1835511490910,25
Сумма	10695918,30	33709,70	16501600900,39	51658264,33	8112727583539,95

Далее используя формулу (2.2.1), производим расчет:

$$r = \frac{\sum xy - \sum x \frac{\sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \times \left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} =$$

$$= \frac{16501600900{,}39 - 33709{,}70 \times \frac{10695918{,}30}{22}}{\sqrt{\left[51658264{,}33 - \frac{(33709{,}70)^2}{22}\right] \times \left[8112727583539{,}95 - \frac{(10695918{,}30)^2}{22}\right]}} =$$

$$= 0{,}83$$

Таким образом, становится возможным сделать вывод, что связь присутствует, является прямой, теснота связи по шкале Чэддока может быть определена как сильная.

Построение уравнения регрессии происходит с исчислением его коэффициентов. Общий вид уравнения парной линейной регрессии имеет вид (2.2.2):

$$y = a + bx + \varepsilon \qquad (2.2.2)$$

Однако на практике построение сводится к поиску коэффициентов a и b. Исчислить их становится возможным применив метод наименьших квадратов [17].

Соответственно коэффициенты уравнения линейной регрессии являются решением системы уравнений (2.2.3):

$$\begin{cases} a + b\bar{x} = \bar{y} \\ a\bar{x} + b\overline{x^2} = \overline{xy} \end{cases} \qquad (2.2.3)$$

Соответственно коэффициенты равны:

$$a = \bar{y} - b\bar{x} = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} \qquad (2.2.4)$$

$$b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{cov(x,y)}{\sigma_x^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad (2.2.5)$$

Пример

По данным, представленным в таблице построить уравнение линейной парной регрессии.

Таблица 2.2.2

Исходные данные для построения

Количество происшествий (шт.)	1	2	3	4	5	6
Нанесенный ущерб (тыс. руб.)	85	140	170	260	310	470

Решение

1. Для расчета коэффициентов a и b уравнения линейной регрессии запишем исходные данные в таблицу 2.2.3:

Таблица 2.2.3

Исходные данные для расчета уравнения линейной регрессии

Номер предприятия	**Количество происшествий (шт.), x**	**Нанесенный ущерб (тыс. руб.), y**	**x^2**	**x*y**
1	1	85	1	85
2	2	140	4	280
3	3	170	9	510
4	4	260	16	1040
5	5	310	25	1550
6	6	470	36	2820
Сумма	21	1435	91	6285

2. Коэффициент b равен:

$$b = \frac{6 \times 6285 - 21 \times 1435}{6 \times 91 - 21^2} = 72{,}14$$

3. Коэффициент a равен:

$$a = \frac{1435 - 72{,}14 \times 21}{6} = -13{,}32$$

4. Следовательно, уравнение линейной регрессии примет вид:

$$y = -13{,}32 + 72{,}14x$$

2.4. Исследование уравнения линейной регрессии

Коэффициент детерминации

Одним из критериев точности описания фактических значений функцией тренда, служит коэффициент детерминации [24]. Коэффициент детерминации показывает, какая доля вариации объясняемой переменной y учтена в модели и обусловлена влиянием на нее факторов, включенных в модель [36]. Для его расчета используется формула (2.4.1):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \qquad (2.4.1)$$

y_i – значения анализируемой переменной;

$\overline{y_i}$ – среднее значение по наблюдаемым данным;

$\widehat{y_i}$ – модельное (теоретическое) значение, построенное по расчетным параметрам.

Коэффициент детерминации может принимать значение в интервале от 0 до 1. Чем ближе значение коэффициента к единице, тем выше доля обусловленное регрессией дисперсии. Соответственно разница единицы и коэффициента детерминации показывает долю дисперсии, вызванной влиянием факторов, которые не были учтены в текущей модели [19].

Чем больше доля объясненной вариации, тем соответственно меньше роль прочих факторов, и, следовательно, линейная модель хорошо аппроксимирует исходные данные и ею можно воспользоваться для прогноза значений результативного признака.

Пример

Определить долю обусловленной регрессией дисперсии по данным представленным в предыдущем примере.

Решение

Было получено уравнение $y = -13{,}32 + 72{,}14x$

Для расчета коэффициента детерминации были выполнены промежуточные расчеты в табличной форме

Номер предприятия	Количество происшествий (шт.), x	Нанесенный ущерб (тыс. руб.), y	$\widehat{y_i}$	$(y_i - \widehat{y_i})^2$	$(y_i - \overline{y_i})^2$
1	1	85	58,82	685,3924	105625
2	2	140	130,96	81,7216	72900
3	3	170	203,1	1095,61	57600
4	4	260	275,24	232,2576	22500
5	5	310	347,38	1397,2644	10000
6	6	470	419,52	2548,2304	3600
7		1435	1501,62	4438,2244	1050625
Сумма	21		419,52	10478,7008	1322850

Производим расчет по формуле (2.4.1):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} = 1 - \frac{10478{,}7008}{1322850} = 0{,}99$$

Таким образом, доля обусловленной регрессией дисперсии составляет 99%.

Средняя ошибка аппроксимации

Фактические значения результативного признака отличаются от теоретических, рассчитанных по уравнению регрессии. Чем меньше это отличие, тем ближе теоретические значения подходят к эмпирическим данным, лучше качество модели [35].

Средняя ошибка аппроксимации ($\bar{A}$) – среднее относительное отклонение расчетных значений от фактических [68].

Для расчета используют формулу (2.4.2):

$$\bar{A} = \frac{1}{n}\sum \left|\frac{y_i - \widehat{y_i}}{y}\right| \times 100\% \qquad (2.4.2)$$

Ошибка аппроксимации в пределах 5-7 % свидетельствует о хорошем подборе модели к исходным данным. Большинство авторов рекомендуют считать модель регрессии адекватной, если средняя относительная ошибка аппроксимации не превышает 12% [11, 23. 31. 72].

Пример

По данным, представленным в таблице, рассчитать среднюю ошибку аппроксимации и сделать вывод о качестве модели.

Таблица 2.2.4

Исходные данные для построения

Количество происшествий (шт.)	1	2	3	4	5	6
Нанесенный ущерб (тыс. руб.)	85	140	170	260	310	470

Решение

Для расчета коэффициента детерминации были выполнены промежуточные расчеты в табличной форме

Номер предприятия	Количество происшествий (шт.), x	Нанесенный ущерб (тыс. руб.), y	$\widehat{y_i}$	$y_i - \widehat{y_i}$	$\frac{y_i - \widehat{y_i}}{y}$
1	1	85	58,82	26,18	0,445087
2	2	140	130,96	9,04	0,069029
3	3	170	203,1	-33,1	-0,16297
4	4	260	275,24	-15,24	-0,05537
5	5	310	347,38	-37,38	-0,10761
6	6	470	419,52	50,48	0,120328
7		1435	1501,62		
Сумма	21		419,52		0,308494

Используя формулу (2.4.2), производим расчет:

$$\bar{A} = \frac{1}{n}\sum \left|\frac{y_i - \widehat{y_i}}{y}\right| \times 100\% = \frac{1}{6} \times 0{,}308494 \times 100\% = 4{,}4\%$$

Таким образом, средняя ошибка аппроксимации составляет 4.4%, что о хорошем подборе модели к исходным данным.

F-критерий Фишера

Данный критерий заключается в проверке гипотезы H_o о статистической незначимости уравнения регрессии. Для этого выполняется сравнение фактического $F_{факт}$ и критического (табличного) $F_{табл.}$ значений F-критерия Фишера [7].

Непосредственному расчету *F*- критерия предшествует анализ дисперсии. Центральное место в нем занимает разложение общей суммы квадратов отклонений переменной *y* от среднего значения y на две части — «объясненную» и «необъясненную». Общая сумма квадратов отклонений индивидуальных значений результативного признака y от среднего значения y вызвана влиянием множества причин. Условно разделим всю совокупность причин на две группы: изучаемый фактор x и прочие факторы. Если фактор не оказывает влияния на результат, то вся дисперсия результативного признака обусловлена воздействием прочих факторов и общая сумма квадратов отклонений совпадет с остаточной. Если же прочие факторы не влияют на результат, то остаточная сумма квадратов равна нулю. В этом случае сумма квадратов отклонений, объясненная регрессией, совпадает с общей суммой квадратов [14].

Поскольку не все точки поля корреляции лежат на линии регрессии, то всегда имеет место их разброс как обусловленный влиянием фактора *x*, т.е. регрессией *y* по *x*, так и вызванный действием прочих причин (необъясненная вариация). Пригодность линии регрессии для прогноза зависит от того, какая часть общей вариации признака *y* приходится на объясненную вариацию. Очевидно, что если сумма квадратов отклонений, обусловленная регрессией, будет больше остаточной суммы квадратов, то уравнение регрессии статистически значимо и фактор *x* оказывает существенное воздействие на результат. Это равносильно тому, что коэффициент детерминации R^2 будет приближаться к единице [19].

Любая сумма квадратов отклонений связана с числом степеней свободы *(df— degrees of freedom)*, т. е. с числом свободы независимого варьирования признака.

Число степеней свободы равно разности между числом независимых наблюдений случайной величины n и числом связей, ограничивающих свободу их изменения. Число степеней свободы связано с числом единиц совокупности *n* и с числом определяемых по ней констант. Число степеней свободы должно показать, сколько независимых отклонений от средней из возможных требуется для образования данной суммы квадратов.

Для общей суммы квадратов необходимо (*n*-1) независимых отклонений, т.к. по совокупности из n единиц после расчета среднего уровня свободно варьируют лишь (*n*-1) отклонение.

Факторная сумма квадратов при парной линейной регрессии зависит только от одной константы – коэффициента регрессии *b*. Поэтому данная сумма имеет одну степень свободы [17].

Разделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим средний квадрат отклонений, или, что тоже самое, дисперсию на одну степень свободы дисперсии [81].

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера, используемого для оценки значимости уравнения регрессии (2.4.3):

$$F_{\text{факт}} = \frac{\sum \frac{(\hat{y}_i - \bar{y}_i)^2}{m}}{\sum \frac{(\hat{y}_i - \bar{y}_i)^2}{n-m-1}} = \frac{R_{xy}^2}{1-R_{xy}^2} \times \frac{n-m-1}{m} \qquad (2.4.3)$$

где *n* – число единиц совокупности;

m – число параметров при переменных.

Для линейной регрессии m = 1.

$F_{табл}$ – максимально возможное значение критерия под влиянием случайных факторов при степенях свободы $k_1 = m$, $k_2 = n - m - 1$ (для линейной регрессии $m = 1$) и уровне значимости α.

Уровень значимости α – вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно величина α принимается равной 0,05 или 0,01.

Если $F_{табл.} < F_{факт}$, то H_0-гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. Если $F_{табл.} > F_{факт}$, то гипотеза H_o не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

Таблица 2.2.5

Таблица значений F-критерия Фишера при уровне значимости $\alpha = 0,05$

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57

40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

Пример

Произвести расчет *F*- критерия по данным, представленным в таблице 2.2.6. Значение α принять 0,95.

Таблица 2.2.6

Исходные данные для анализа

№	*x*	*y*
1	90	41
2	29	49
3	39	56
4	79	64
5	88	72
6	53	65
7	34	63
8	40	87
9	75	77
10	79	62
Сумма	606	636
Среднее	60.6	63.6

Решение

1. Рассчитаем величины дисперсий для *x* и *y*:

$$\sigma_x^2 = 572{,}83$$

$$\sigma_y^2 = 174{,}04$$

2. Определим значение F-критерия Фишера:

$$F_{\text{факт}} = \frac{572{,}83}{174{,}04} = 3{,}29$$

3. Критическое значение F-критерия для степеней свободы равно

$k_1 = k_2 = 10 - 1 = 9$ равно $F_{\text{крит}} = 4{,}03$

Таким образом, $F_{\text{факт}} < F_{\text{крит}}$, так как 3,29 < 4,03, и на уровне значимости 0,05 принимаем нулевую гипотезу, то есть различия в степени однородности показателей статистически незначимы.

Определение достоверности различий по t-критерию Стьюдента

Для оценки статистической значимости коэффициентов регрессии применяется t-критерий Стьюдента, согласно которому выдвигается «нулевая» гипотеза H_0 о статистической незначимости коэффициента уравнения регрессии (т. е. о статистически незначимом отличии величины *a* или *b* от нуля) [44].

Данная гипотеза отвергается при выполнении условия $t_{\text{факт}} > t_{\text{табл}}$, где $t_{\text{табл}}$ определяется по таблицам t-критерия Стьюдента по числу степеней свободы $k_1 = n\text{-}m\text{-}1$ (*m* - число независимых переменных в уравнении регрессии) и заданному уровню значимости α [52, 64].

Стандартные ошибки коэффициента регрессии и параметра *a* определяются по формулам (для парной линейной регрессии) (2.4.4):

$$m_b = \sqrt{\frac{\sum(y - \widehat{y_x})^2 \Big/ (n-2)}{\sum(x - \bar{x})^2}} = \sqrt{\frac{S^2}{\sum(x - \bar{x})^2}} = \frac{S}{\sigma_x \sqrt{n}} \qquad (2.4.4)$$

$$m_a = \sqrt{\frac{\sum(y - \widehat{y_x})^2}{n-2} \times \frac{\sum x^2}{\sum(x - \bar{x})^2}} = \sqrt{S^2 \frac{\sum x^2}{n \sum(x - \bar{x})^2}} \qquad (2.4.5)$$

где S^2 — остаточная дисперсия на одну степень свободы/

Величина ошибки коэффициента корреляции m_r определяется по формуле (2.4.6):

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}} \quad (2.4.6)$$

Оценка значимости коэффициентов регрессии и корреляции проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b} \quad (2.4.7)$$

$$t_a = \frac{a}{m_a} \quad (2.4.8)$$

$$t_r = \frac{r}{m_r} \quad (2.4.9)$$

Далее необходимо провести сравнения фактических и критических (табличных) значений t-статистики. При этом табличные значения (таблица 2.2.7) определяются при определенном уровне значимости α ($\alpha = 0{,}05$ для экономических процессов и явлений) и числе степеней свободы (n - 2) [13]. Если $t_{факт} > t_{крит}$, то H_0 отклоняется, т.е. a, b и r_{xy} не случайно отличаются от нуля и все они статистически значимы. Иначе, гипотеза H_0 о случайной природе показателей принимается.

Таблица 2.2.7

Критические значения *t*-критерия Стьюдента при уровне значимости 0,05; 0,01 и 0,001.

Число степеней свободы k	p=0,05	p=0,01	p=0,001
1	12,70	63,65	636,61
2	4,303	9,925	31,602
3	3,182	5,841	12,923
4	2,776	4,604	8,610
5	2,571	4,032	6,869
6	2,447	3,707	5,959
7	2,365	3,499	5,408
8	2,306	3,355	5,041
9	2,262	3,250	4,781
10	2,228	3,169	4,587
11	2,201	3,106	4,437
12	2,179	3,055	4,318
13	2,160	3,012	4,221
14	2,145	2,977	4,140
15	2,131	2,947	4,073
16	2,120	2,921	4,015
17	2,110	2,898	3,965
18	2,101	2,878	3,922
19	2,093	2,861	3,883

20	2,086	2,845	3,850
21	2,080	2,831	3,819
22	2,074	2,819	3,792
23	2,069	2,807	3,768
24	2,064	2,797	3,745
25	2,060	2,787	3,725
26	2,056	2,779	3,707
27	2,052	2,771	3,690
28	2,049	2,763	3,674
29	2,045	2,756	3,659
30	2,042	2,750	3,646
31	2,040	2,744	3,633
32	2,037	2,738	3,622
33	2,035	2,733	3,611
34	2,032	2,728	3,601
35	2,030	2,724	3,591
36	2,028	2,719	3,582
37	2,026	2,715	3,574
38	2,024	2,712	3,566
39	2,023	2,708	3,558
40	2,021	2,704	3,551
41	2,020	2,701	3,544
42	2,018	2,698	3,538
43	2,017	2,695	3,532

44	2,015	2,692	3,526
45	2,014	2,690	3,520
46	2,013	2,687	3,515
47	2,012	2,685	3,510
48	2,011	2,682	3,505
49	2,010	2,680	3,500
50	2,009	2,678	3,496
51	2,008	2,676	3,492
52	2,007	2,674	3,488
53	2,006	2,672	3,484
54	2,005	2,670	3,480
55	2,004	2,688	3,476
56	2,003	2,667	3,473
57	2,002	2,665	3,470
58	2,002	2,663	3,466
59	2,001	2,662	3,463
60	2,000	2,660	3,460
61	2,000	2,659	3,457
62	1,999	2,657	3,454
63	1,998	2,656	3,452
64	1,998	2,655	3,449
65	1,997	2,654	3,447
66	1,997	2,652	3,444
67	1,996	2,651	3,442

68	1,995	2,650	3,439
69	1,995	2,649	3,437
70	1,994	2,648	3,435
71	1,994	2,647	3,433
72	1,993	2,646	3,431
73	1,993	2,645	3,429
74	1,993	2,644	3,427
75	1,992	2,643	3,425
76	1,992	2,642	3,423
77	1,991	2,641	3,422
78	1,991	2,640	3,420
79	1,990	2,639	3,418
80	1,990	2,639	3,416
90	1,987	2,632	3,402
100	1,984	2,626	3,390
110	1,982	2,621	3,381
120	1,980	2,617	3,373
130	1,978	2,614	3,367
140	1,977	2,611	3,361
150	1,976	2,609	3,357
200	1,972	2,601	3,340
250	1,969	2,596	3,330
300	1,968	2,592	3,323
350	1,967	2,590	3,319

2.5. Построение уравнения линейной парной регрессии в MS Excel

Рассмотрим построение регрессионного уравнения, оценку его параметров, а также анализ тесноты связей на примере.

Пример

Необходимо определить тесноту связей и построить регрессионное уравнения для данных, представленных в таблице 2.5.1.

Таблица 2.5.1

Исходные данные для анализа

Период	Валовый региональный продукт, млн руб., y	Численность населения (оценка на конец года; тысяч человек), x
2000	42074,5	1507,0
2001	49941,8	1508,1
2002	62404,4	1511,9
2003	76054,5	1513,9
2004	114409,3	1511,7
2005	144987,8	1511,7
2006	178846,1	1514,2
2007	237013,3	1520,1
2008	317656,3	1526,3
2009	304345,3	1531,8
2010	398361,4	1532,4
2011	507839,8	1536,6
2012	545517,2	1541,9
2013	569006,4	1545,5
2014	619677,7	1549,8
2015	693379,4	1552,4
2016	778027,8	1555,5
2017	837306,8	1553,0
2018	911597,9	1550,9
2019	955329,2	1553,0
2020	997330,9	1545,5
2021	1354810,5	1536,5

Решение

Для построения корреляционного поля необходимо выполнить команды, выделив ячейки **Вставка** – **Диаграммы** – **Точечная диаграмма** (рис. 2.5.1)

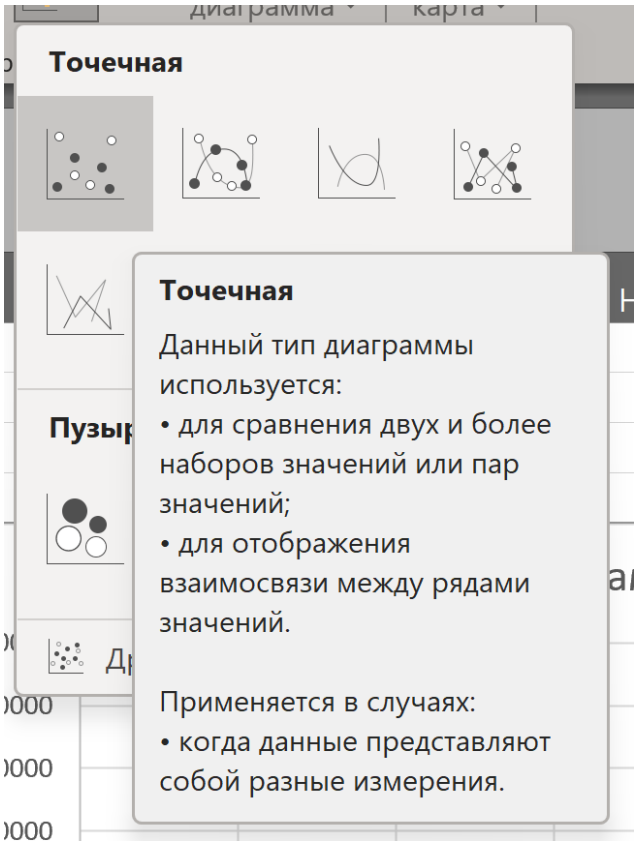


Рис. 2.5.1. Точечная диаграмма

Далее на листе получим точеную диаграмму корреляционного поля (рис. 2.5.2).

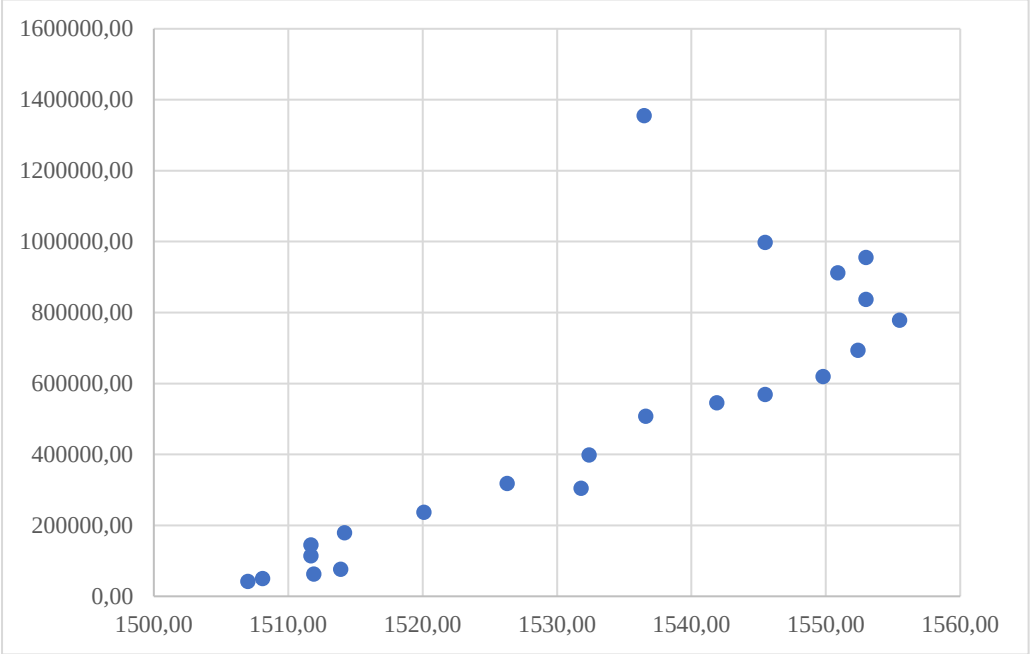


Рис. 2.5.2. Корреляционное поле

Для построения регрессионного уравнения, а аткже оценки его параметров необходимо выполнить команду **Данные – Анализ данных – регрессия** (рис. 2.5.3).

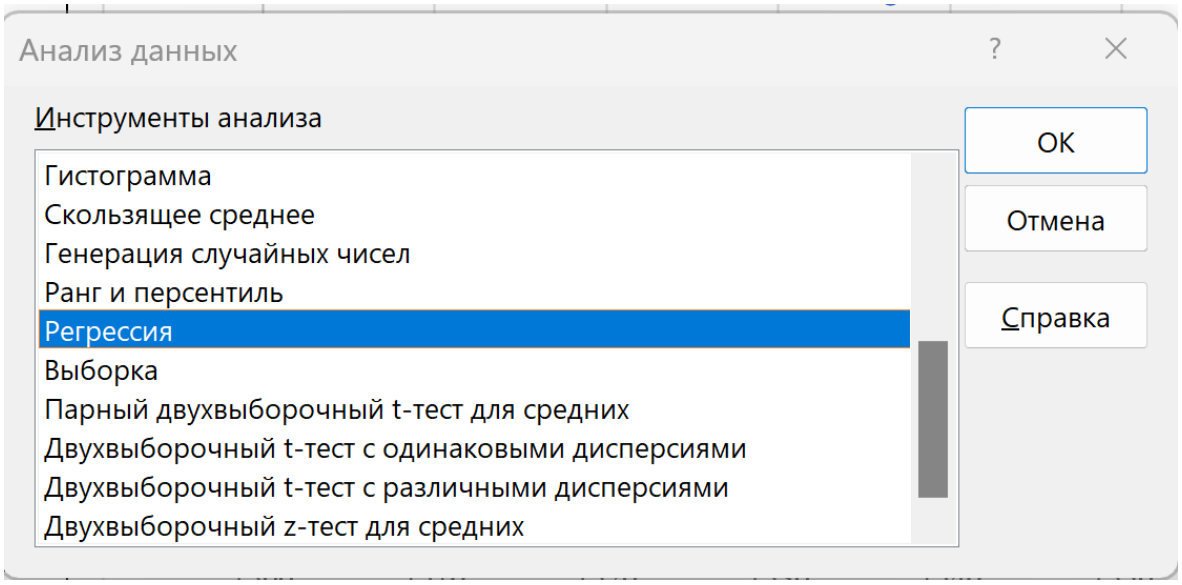


Рис. 2.5.3. Меню **Анализ данных**

Далее необходимо заполнить форму (рис. 2.5.4), указав входной интервал x, интервал y. Отметив параметры вывода, становится возможным поместить результаты анализа или на новый лист или на текущий, указав выходной интервал ячеек. Также можно выгрузить данные в новую книгу. Можно откорректировать уровень надежности, установив галочку напротив соответствующего пункта (по умолчанию задана величина 95%).

Также можно получить расчет остатков и их графическое представление, стандартные остатки, график подбора, а также график нормального распределения. Все перечисленные параметры являются отключенными по умолчанию.

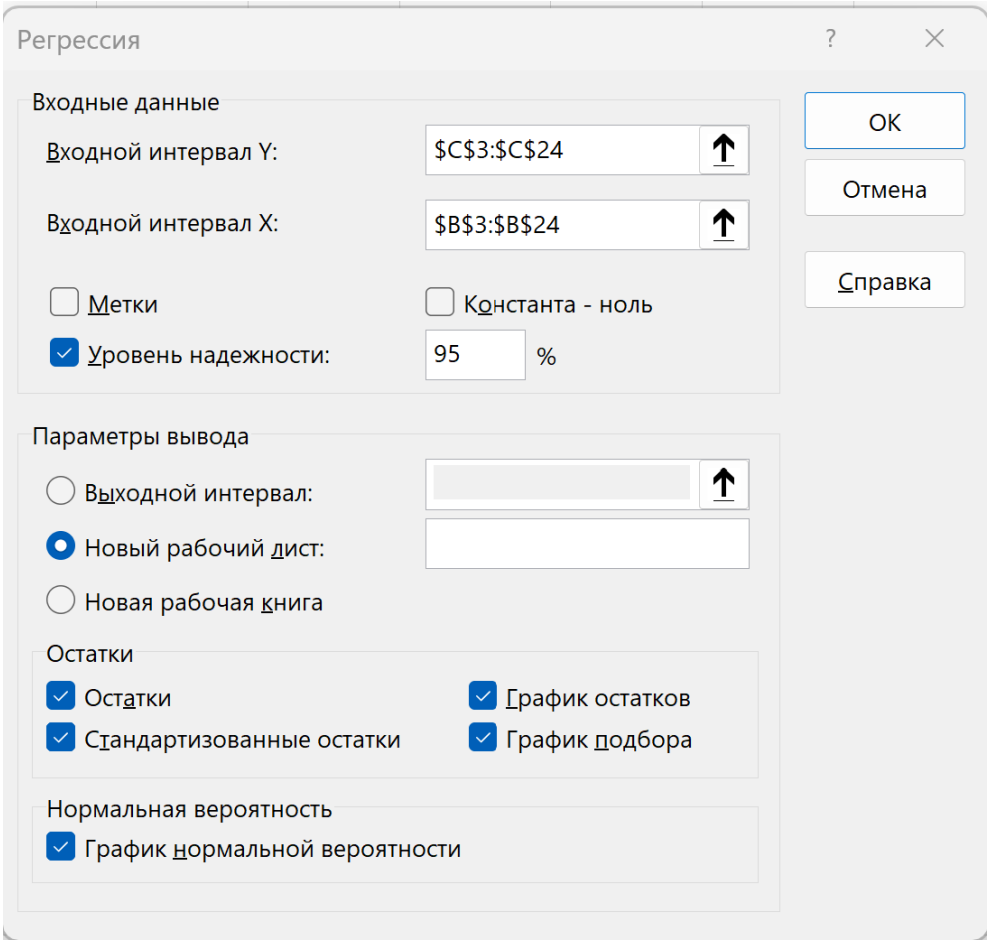


Рис. 2.5.4. Меню **Регрессия**

Отметив нужные позиции и введя необходимые параметры необходимо нажать **ОК**.В результате на новом листе будут выведены итоги регрессионного анализа (рис. 2.5.5).

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множеств	0,833837297							
R-квадрат	0,695284637							
Нормиров	0,680048869							
Стандартн	210655,6294							
Наблюден	22							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	1	2,02509E+12	2,02509E+12	45,63502355	1,43077E-06			
Остаток	20	8,87516E+11	44375794213					
Итого	21	2,91261E+12						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересече	-27050960,05	4076580,492	-6,635698743	1,84088E-06	-35554557,94	-18547362,15	-35554557,94	-18547362,15
Переменн	17971,59391	2660,341916	6,755369979	1,43077E-06	12422,21792	23520,96991	12422,21792	23520,96991

Рис. 2.5.5. Итоги регрессионного анализа

В поле «Регрессионная статистика» представлены:

- Множественный R= 0,833837296714246, что свидетельствует о сильной тесноте связи между показателями;

- R-квадрат= 0,695284637391721, что свидетельствует о том, что порядка 70% изменчивости обусловлено факторами, входящими в регрессионную модель;

- Нормированный R-квадрат= 0,680048869261307. Нормированный R-квадрат (скорректированный коэффициент детерминации) означает, какое влияние корректировка R-квадрата оказала на величину коэффициента детерминации. Недостатком R-квадрата является то, что он увеличивается при добавлении новых объясняющих переменных (хотя это и не обязательно означает улучшение качества регрессионной модели), в то время как нормированный R-квадрат может уменьшаться при введении в модель новых объясняющих переменных, не оказывающих существенное влияние на зависимую переменную. Если нормированный R-квадрат ненамного отличается от коэффициента детерминации, можно сделать вывод о хорошем качестве модели. В данном случае можно сделать вывод о приемлемом качестве регрессионной модели;

- Стандартная ошибка= 210655,62943492;

- Наблюдения=22 – число наблюдений.

Информация о коэффициентах, а также статистика по ним представлена в нижней таблице. Следует помнить, что если p-значение для независимой переменной меньше 0,01, то она является высоко значимой для *y*. Если p-значение от 0,01 до 0,05 – результат значим. Если коэффициент больше 0,05, результат значимым не является.

Таким образом, уравнение имеет вид:

$$y = -27050960{,}06 + 17971{,}59391\text{x}$$

Также на листе представлены график остатков (рис. 2.5.6), график подбора (рис. 2.5.7) и график нормального распределения (рис. 2.5.8).

График остатков - это тип графика, который отображает прогнозируемые значения в сравнении с остаточными значениями для регрессионной модели. Этот тип графика часто используется для оценки того,

подходит ли модель линейной регрессии для данного набора данных, и для проверки гетероскедастичности остатков.

Рис. 2.5.6. График остатков

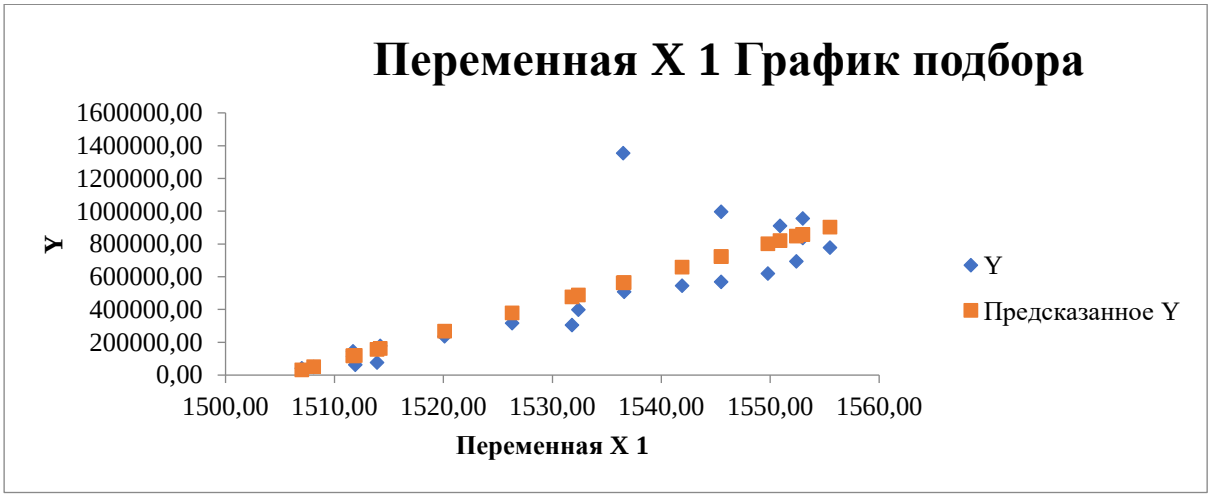


Рис. 2.5.7. График подбора

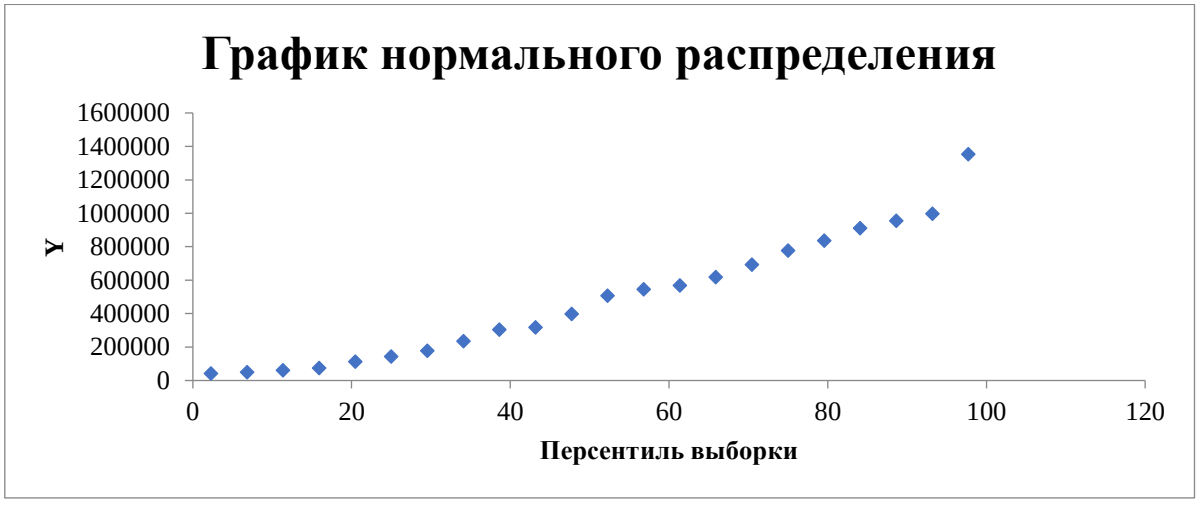


Рис. 2.5.8 График нормального распределения

Также следует отметить, что в MS Excel возможно построить регрессионное уравнение зависимости без расчета дополнительных параметров точности и надежности модели, а также определить корреляционный коэффициент.

Для расчета коэффициента корреляции необходимо в ячейке ввести формулу **=КОРРЕЛ** или перейти **Формулы – Другие формулы – Статистические – КОРРЕЛ** (рис. 2.5.9).

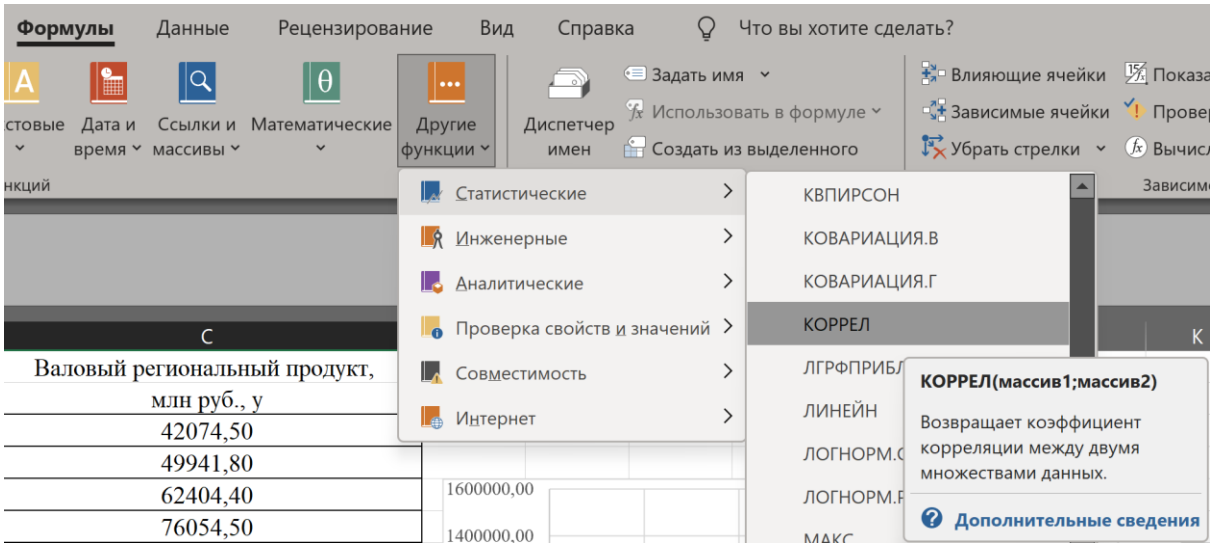


Рис. 2.5.9. Ввод формулы расчета корреляции

Таким образом в ячейке, в которой был осуществлен ввод форму будет рассчитан коэффициент линейной корреляции.

Для построения уравнения регрессии необходимо на графике, отображающем корреляционное поле добавить тренд. Для этого необходимо щелкнув правой кнопкой мышки на любой точке корреляционного поля выбрать пункт **Добавить линию тренда**. Далее в открывшемся меню (рис. 2.5.10) выбрать необходимый тренд. Так как в примере рассматривается линейная регрессии, то будет выбран соответствующий тип тренда. Отметим, что данный пункт является включенным по умолчанию Однако, при необходимости можно установить иные параметры линии тренда:

- экспоненциальная;
- логарифмическая;
- полиномиальная (с возможностью указания степени полинома);
- степенная;
- скользящая среднее (с возможностью указания периода).

В том же меню следует установить галочки напротив пунктов **Показывать уравнение на графике** и **Поместить на диаграмму величину достоверности аппроксимации**.

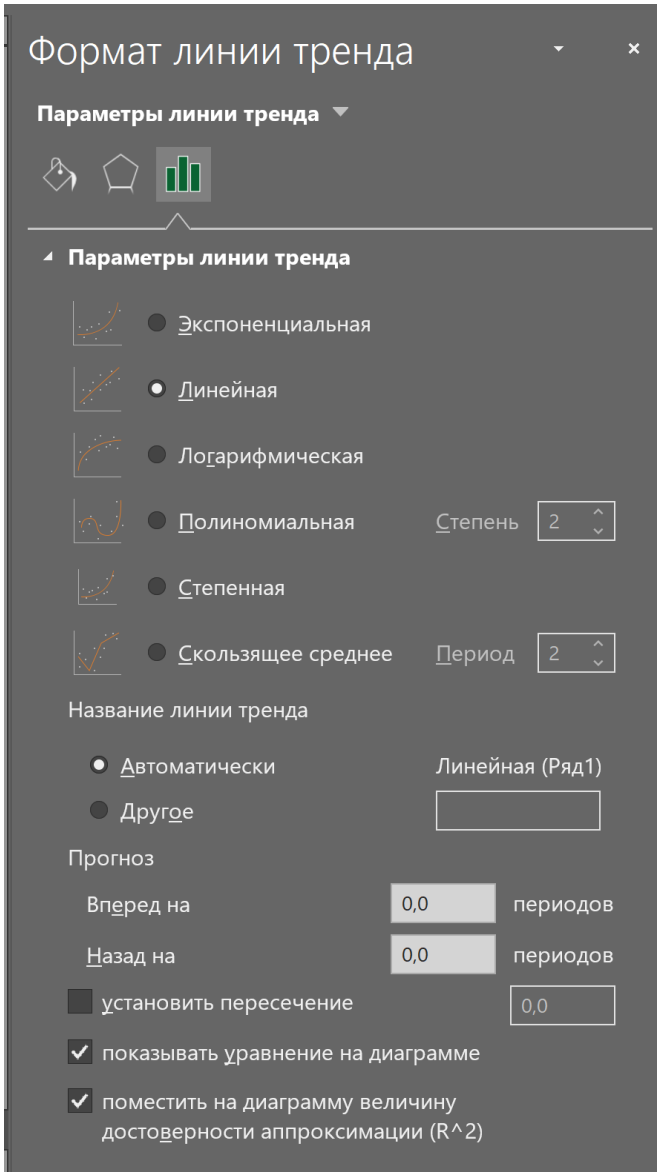


Рис. 2.5.10. Добавление тренда на график корреляционного поля

Для построения уравнения регрессии необходимо на графике, отображающем корреляционное поле добавить тренд. Для этого необходимо щелкнув правой кнопкой мышки на любой точке корреляционного поля выбрать пункт **Добавить линию тренда**. Далее в открывшемся меню (рис. 2.5.9) выбрать необходимый тренд.

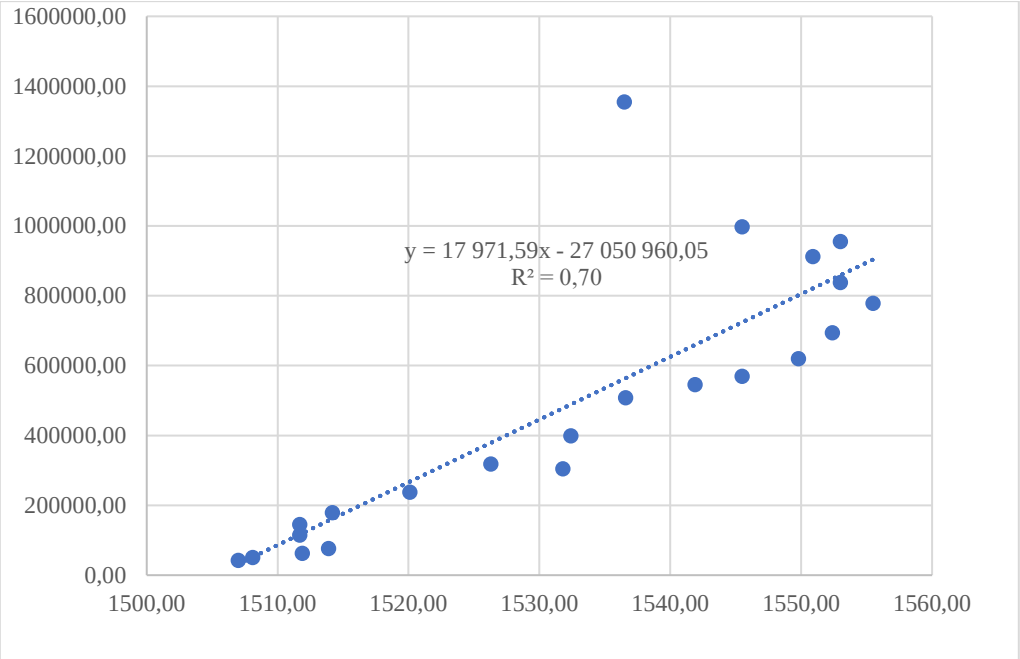


Рис. 2.5.11. График с уравнением регрессии и коэффициентом детерминации

Так как в примере рассматривается линейная регрессии, то будет выбран соответствующий тип тренда. Соответственно на график, отображающем корреляционное поле будут вынесены отмеченные параметры (рис. 2.5.12).

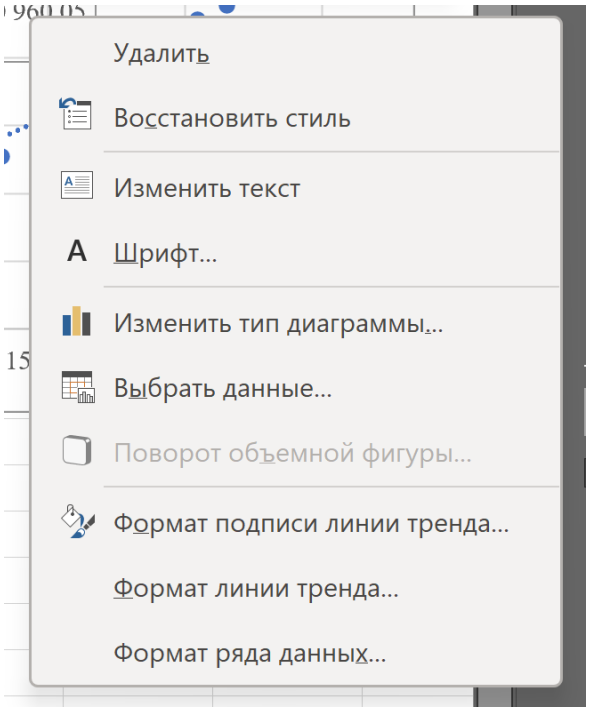


Рис. 2.5.12. Меню выбора формата подписи данных тренда

Как можно отметить, уравнения и коэффициент детерминации полностью соответствуют результатам, полученным по расчету из меню **Анализ данных** с учетом округления. При необходимости можно изменить тип подписи, выбрав иную категорию или установив необходимую точность при округлении. Для этого необходимо щелкнув правой кнопкой мыши по полю уравнения (рис. 2.5.12) выбрать в открывшемся меню необходимые параметры (рис. 2.5.13).

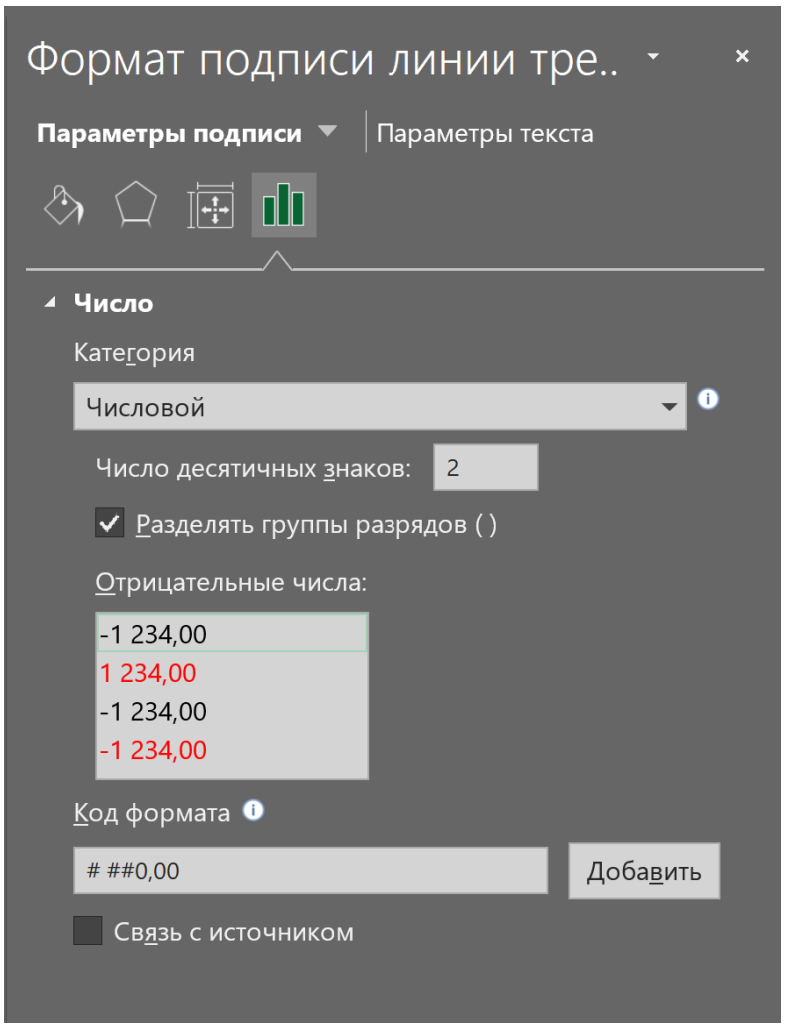


Рис. 2.5.13. Редактирование формата подписи данных тренда

Таким образом были рассмотрены варианты построения регрессионного уравнения в MS Excel.

2.6. Построение уравнения линейной регрессии в Statistica

Рассмотрим построение регрессионного уравнения, оценку его параметров и возможности программного комплекса, а также анализ тесноты связей на примере.

Пример

Необходимо определить тесноту связей и построить регрессионное уравнения для данных, представленных в таблице 2.6.1.

Таблица 2.6.1

Исходные данные для анализа

Период	Валовый региональный продукт, млн руб., y	Численность населения (оценка на конец года; тысяч человек), x
2000	42074,5	1507,0
2001	49941,8	1508,1
2002	62404,4	1511,9
2003	76054,5	1513,9
2004	114409,3	1511,7
2005	144987,8	1511,7
2006	178846,1	1514,2
2007	237013,3	1520,1
2008	317656,3	1526,3
2009	304345,3	1531,8
2010	398361,4	1532,4
2011	507839,8	1536,6
2012	545517,2	1541,9
2013	569006,4	1545,5
2014	619677,7	1549,8
2015	693379,4	1552,4
2016	778027,8	1555,5
2017	837306,8	1553,0
2018	911597,9	1550,9
2019	955329,2	1553,0
2020	997330,9	1545,5
2021	1354810,5	1536,5

Решение

Для расчета коэффициента корреляции необходимо перейти **Основные статистики и таблицы.** В меню выбрать пункт **Парные и частные корреляции - ОК** (рис. 2.6.1).

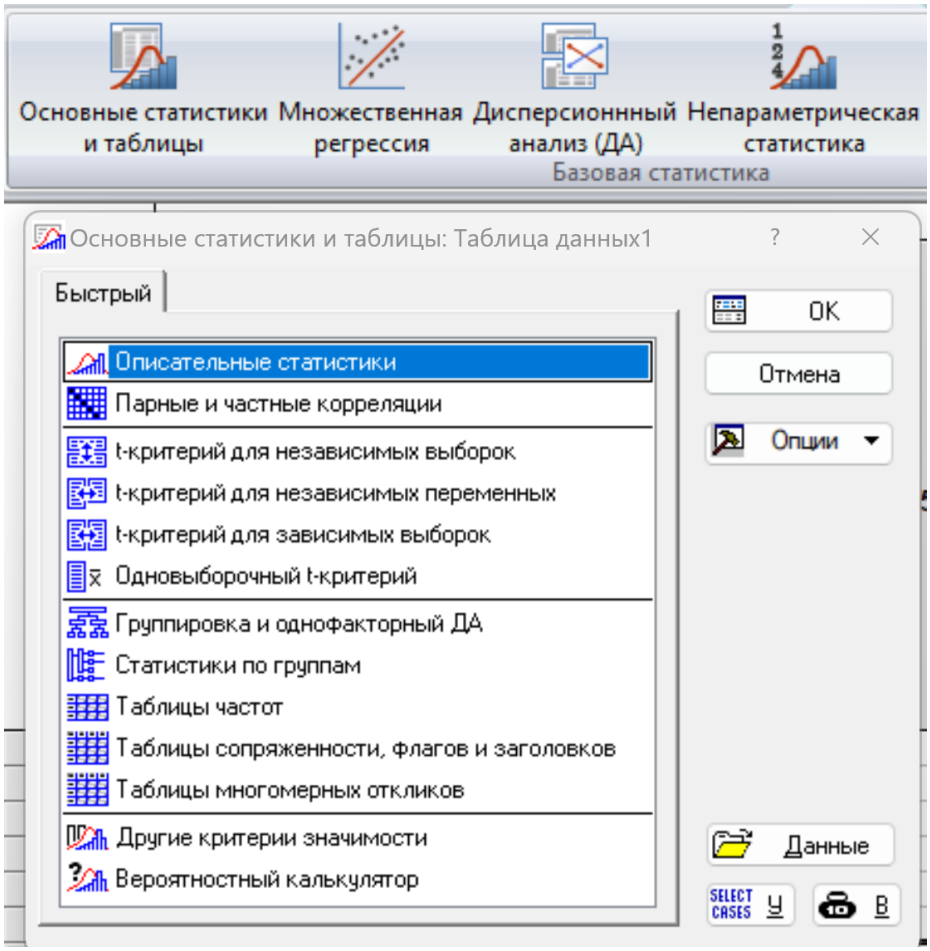


Рис. 2.6.1. Меню расчета тесноты связи

Появится меню выбора параметров расчета (рис. 2.6.2).

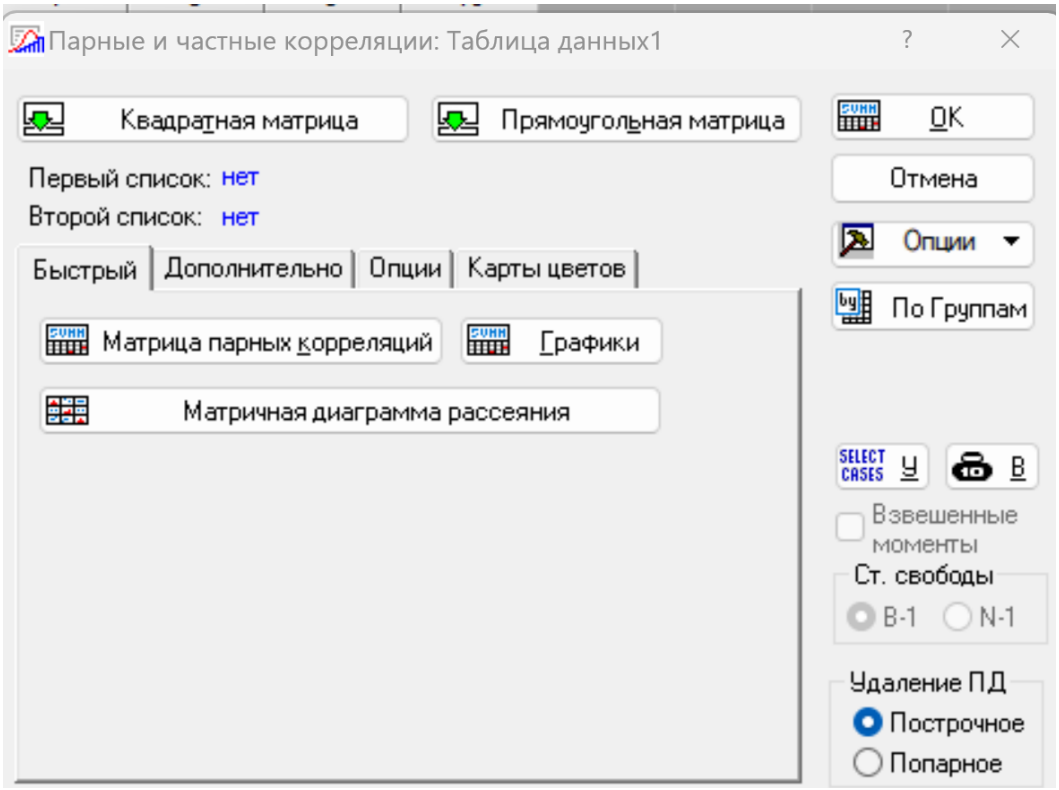


Рис. 2.6.2. Меню параметров расчета тесноты связи

Предоставляется выбор формы представления данных: квадратная или прямоугольная матрица. Квадратная матрица представляет собой расчет корреляционных коэффициентов всех выделенных переменных, которые выделяются единым списком. В прямоугольной матрице можно рассчитать коэффициенты относительно отдельной одного выделенного показателя. В данном случае выделение происходит по двум спискам. В случае расчета коэффициента корреляции между перечнем переменных, в который входят только две переменные безразлично какая матрица будет выбрана.

Далее будет рассмотрен пример с квадратной матрицей. Для ее формирования необходимо выполнить последовательность операция **Квадратная матрица – Выбор переменных для анализа - ОК** (рис. 2.6.3).

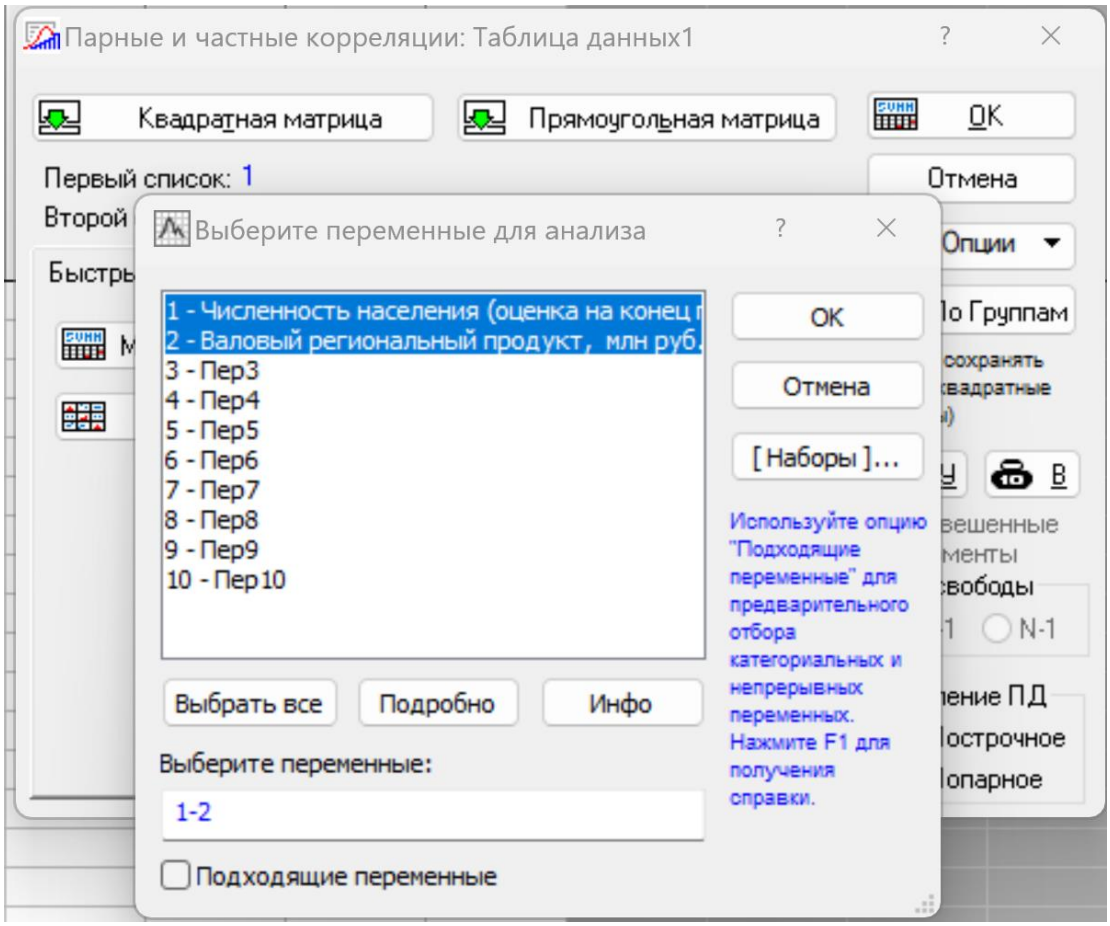


Рис. 2.6.3. Выбор переменных для построения квадратной корреляционной матрицы

Отметим, что в меню есть несколько вкладок, которые позволяют сформировать индивидуальный отчет по представлению результатов корреляционного анализа.

Вкладка **Быстрый**. На данной вкладке представлены наиболее универсальные инструменты.

Матрица парных корреляций (рис. 2.6.4) – происходит построение матрицы корреляционных коэффициентов с представлением дополнительных параметров: число наблюдений (*N*), среднее величина (отдельно по каждому из анализируемых показателей), среднее отклонение (отдельно по каждому из анализируемых показателей).

Переменная	Средние	Ст.откл.	Численность населения (оценка на конец года; тысяч человек), x	Валовый региональный продукт, млн руб., y
	Корреляции (Таблица данных1) Отмеченные корреляции значимы на уровне p <,05000 N=22 (Построчное удаление ПД)			
Численность населения (оценка на конец года; тысяч человек), x	1532,3	17,3	1,000000	0,833837
Валовый региональный продукт, млн руб., y	486178,1	372418,5	0,833837	1,000000

Рис. 2.6.4. Квадратная корреляционная матрица

Отметим, что значимые при p<0,05 корреляции выделяются в матрице красным цветом.

Матричная диаграмма рассеяния – происходит построения корреляционного поля. Необходимо выбрать переменные для построения (рис. 2.6.5), которые будут представлены соответствующими координатами на графике и нажать **ОК**.

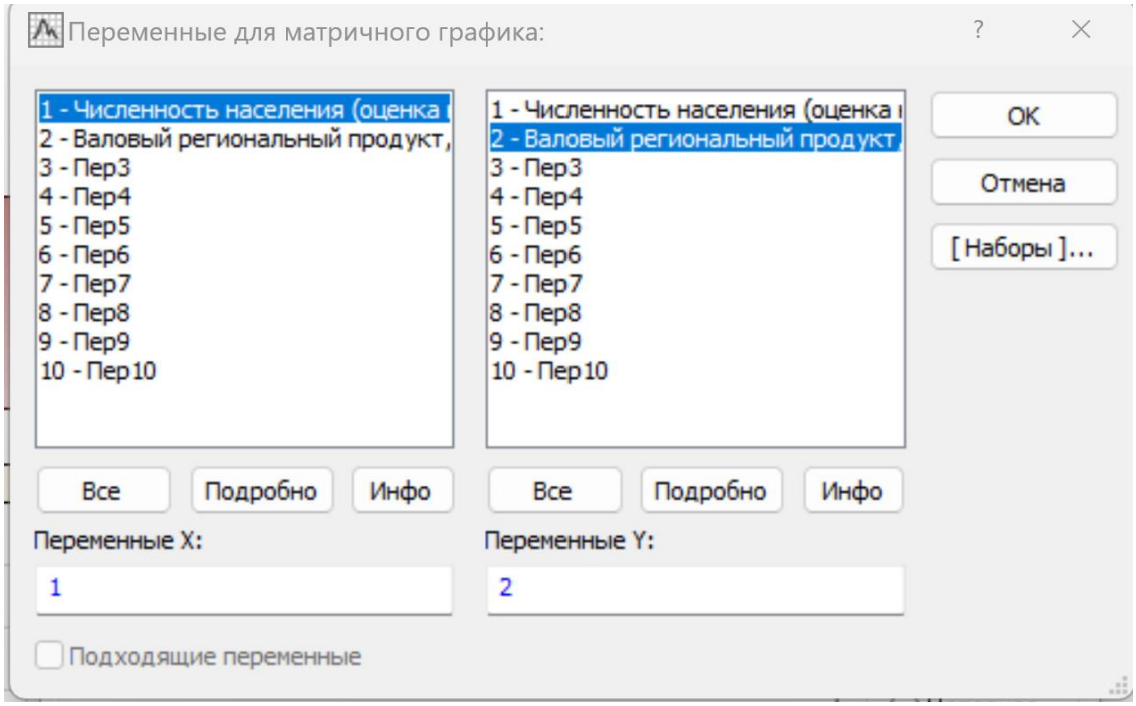


Рис. 2.6.5. Меню построения матричной диаграммы рассеяния

В результате будет представлены графические элементы описания корреляционного взаимодействия переменных (рис. 2.6.6).

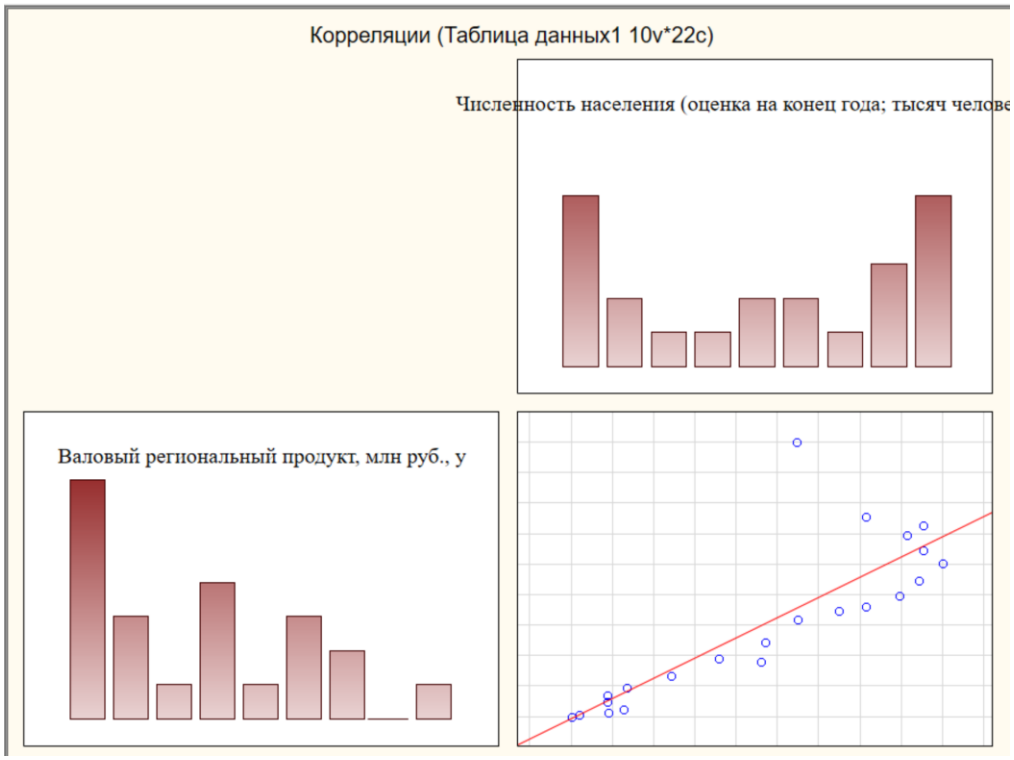


Рис. 2.6.6. **Матричная диаграмма рассеяния**

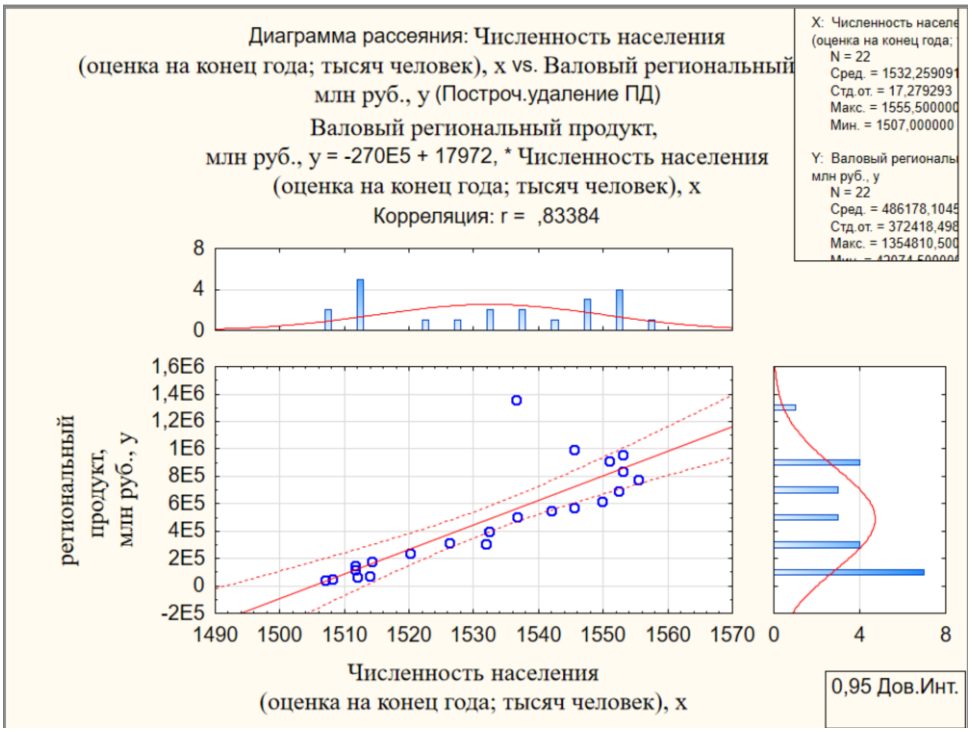


Рис. 2.6.7. Графическое представление корреляционного анализа

Графики – представление результатов корреляционного анализа в графической форме (рис. 2.6.7).

Следует отметить, что происходит построение диаграммы рассеяния, графиков нормальности обеих переменных, а также расчет по каждой из анализируемых переменных: числа наблюдений, средней величины, стандартного отклонения, максимальной и минимальной величин.

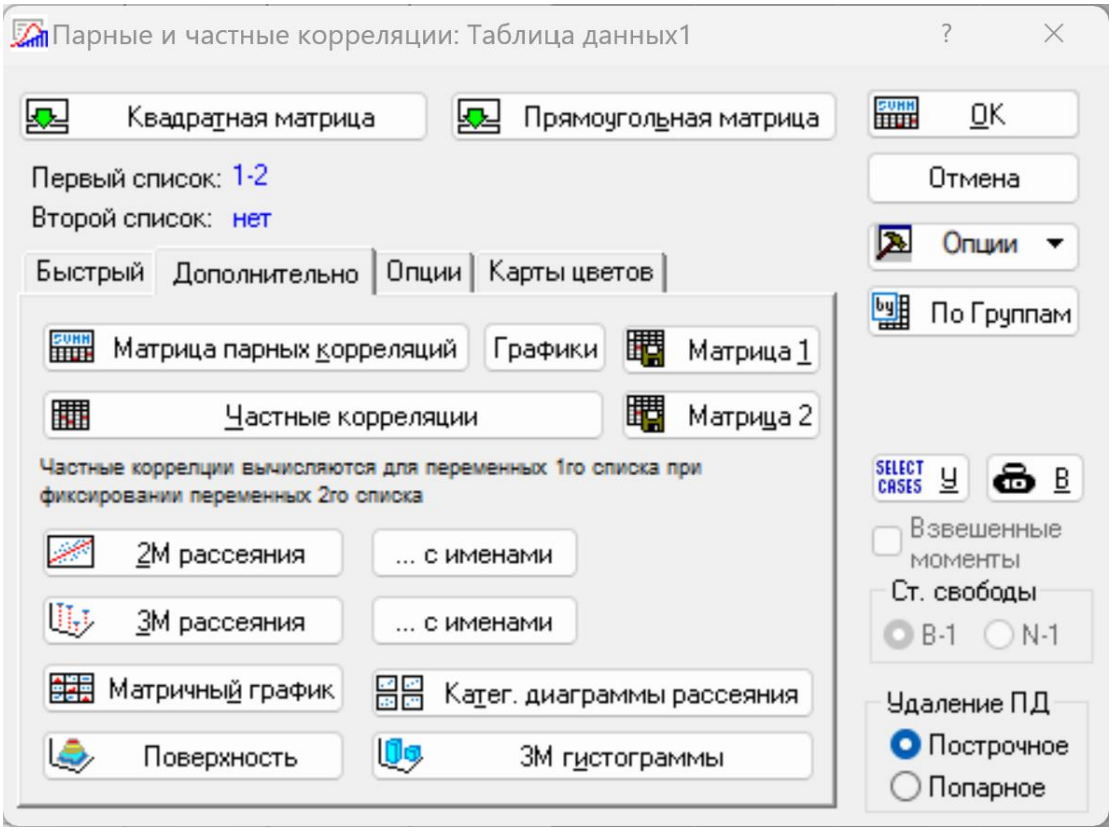


Рис. 2.6.8. Вкладка **Дополнительно**

Вкладка **Дополнительно** (рис. 2.6.8) дает расширенный выбор графического представления корреляционных связей между анализируемыми переменными.

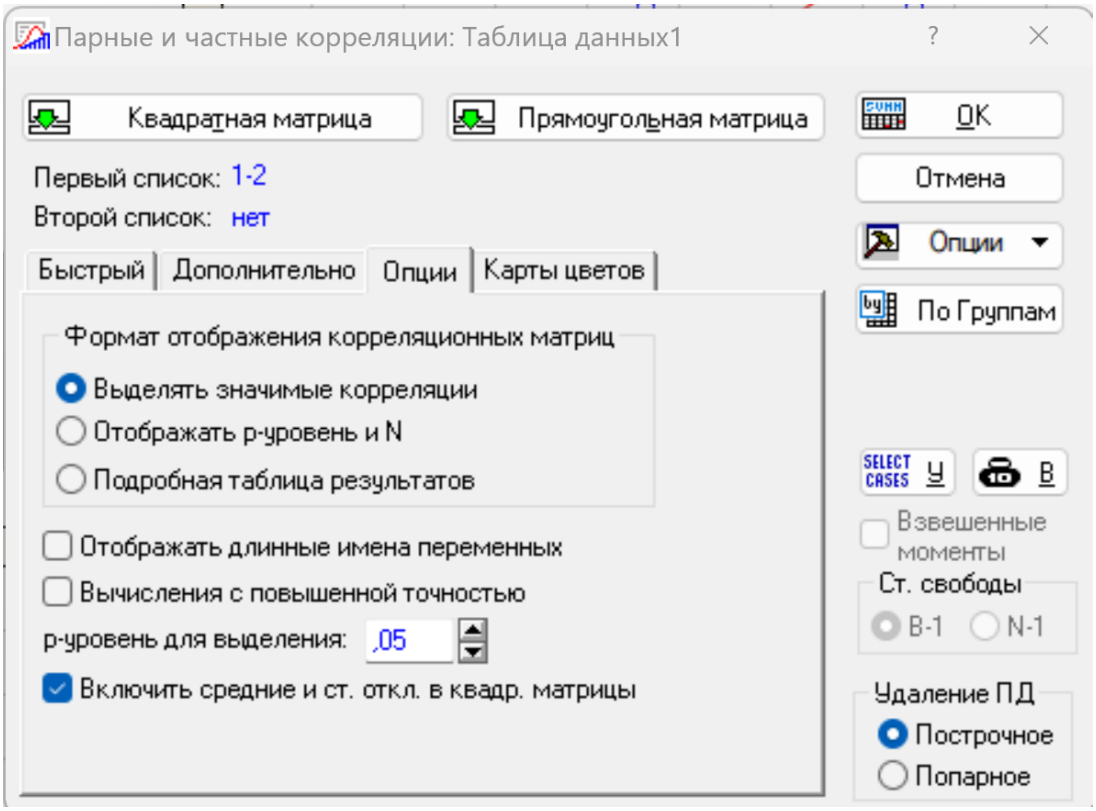


Рис. 2.6.9. Вкладка **Опции**

Вкладка **Опции** (рис. 2.6.9) дает возможность корректировать уровень значимости, используемый при анализе, а также редактировать отображаемые имена переменных. Кроме того при установлении галочки у пункта Подробная таблица результатов, происходит вывод таблицы с итоговыми и промежуточными вычислениями (рис. 2.6.10).

Пер.X и Пер.Y	Среднее	Стд.откл	r(X,Y)	r^2	t	p	N	Св. член завис.Y	Наклон завис.Y	среднее завис.X	**Наклон завис.X**
	Корреляции (Таблица данных1) Отмеченные корреляции значимы на уровне p <,05000 (Построчное удаление ПД)										
ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек), x	1532,3	17,3									
ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек), x	1532,3	17,3	1,000000	1,000000			22	0	1,00	0	1,00
ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек), x	1532,3	17,3									
Валовый региональный продукт, млн руб., y	486178,1	372418,5	0,833837	0,695285	6,755370	0,000001	22	-27050960	17971,59	1513	0,00
Валовый региональный продукт, млн руб., y	486178,1	372418,5									
ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек), x	1532,3	17,3	0,833837	0,695285	6,755370	0,000001	22	1513	0,00	-27050960	17971,59
Валовый региональный продукт, млн руб., y	486178,1	372418,5									
Валовый региональный продукт, млн руб., y	486178,1	372418,5	1,000000	1,000000			22	0	1,00	0	1,00

Рис. 2.6.10. Подробная таблица результатов

Вкладка Карты цветов (рис. 2.6.11) позволяет добавить к стандартной корреляционной матрице цветовую идентификацию для большей наглядности и облегчению трактовки результатов (2.6.12).

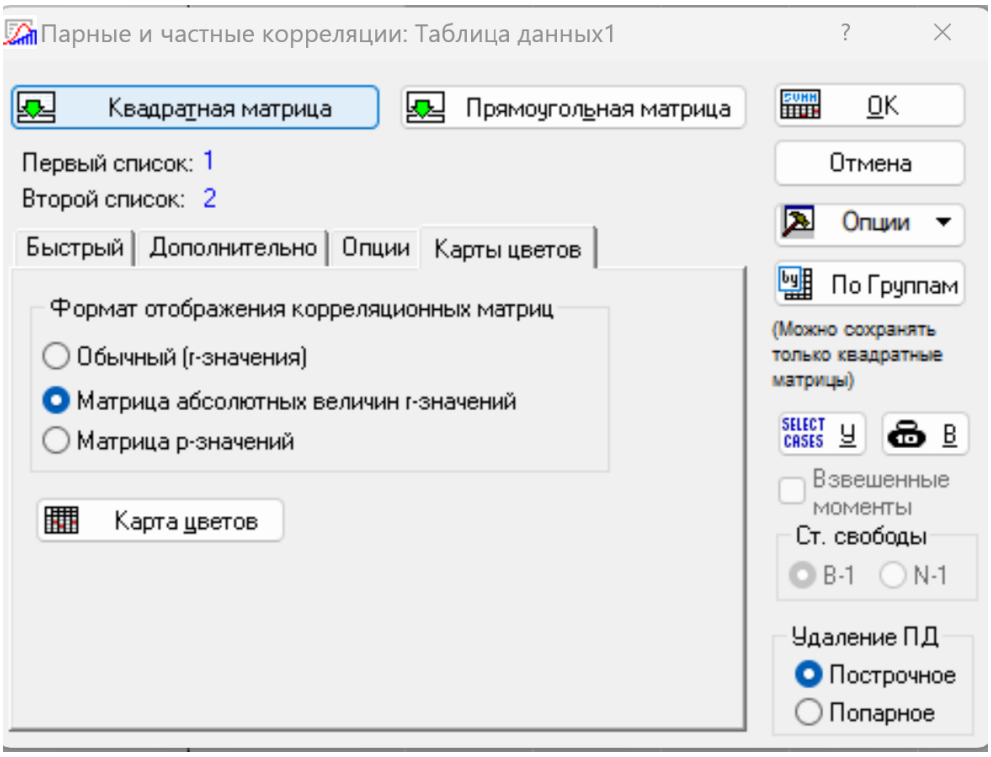


Рис. 2.6.11. Вкладка **Карты цветов**

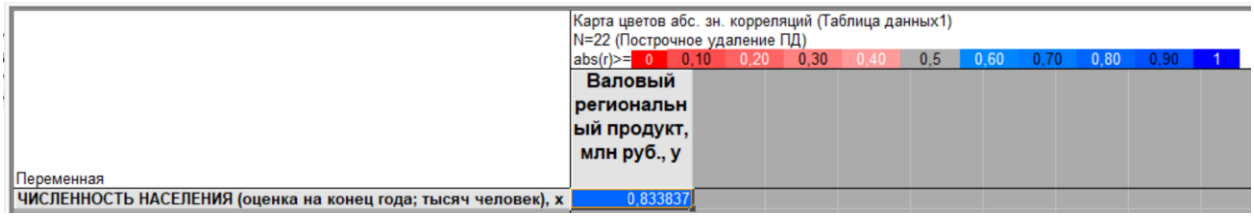


Рис. 2.6.12. Карта цветов абсолютных значений корреляции

Для построения регрессионного уравнения после рассмотрения и подтверждения наличия тесной связи между переменными необходимо перейти во вкладку **Анализ** и выбрать пункт **Множественная регрессия** (рис. 2.6.13).

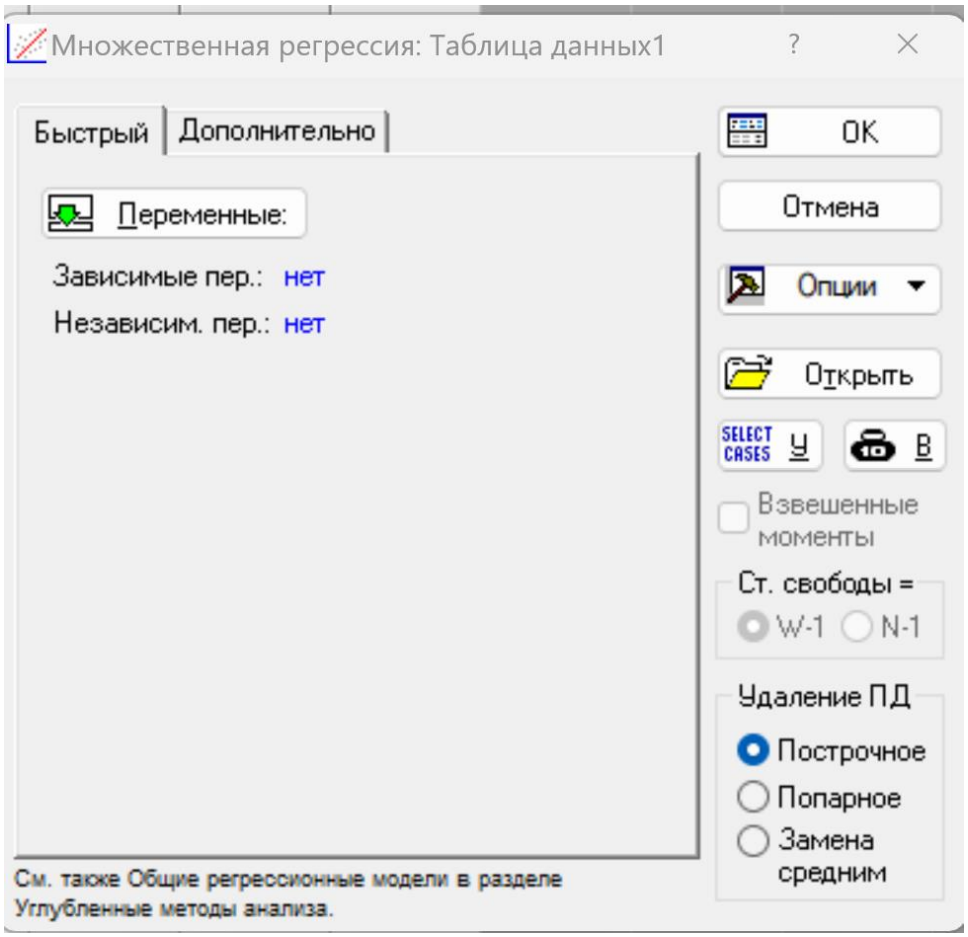


Рис. 2.6.13. Меню **Множественная регрессия**

Чтобы задать зависимую и независимую переменные необходимо нажать кнопку **Переменные** (рис. 2.6.14) и задать их. В рассматриваемом примере зависимой переменной будет валовый региональный продукт, а независимой соответственно численность населения. После выбора необходимо нажать **ОК**.

Следует отметить, что если отсутствует уверенность в характере заданных числовых значений можно установить галочку в поле Подходящие переменные. В данном случае все неподходящие данные не будут учитываться при проведении анализа.

На вкладке **Дополнительно** (рис. 2.6.15) можно задать ряд функций:

- **пошаговая или гребневая регрессия** - для устранения эффекта мультиколлинеарности и нахождения наилучшего регрессионного уравнения в программе Statistica можно воспользоваться методом пошагового регрессионного анализа или гребневой регрессии, который

последовательно проверяет независимые переменные на отсутствие между ними мультиколлинеарности;

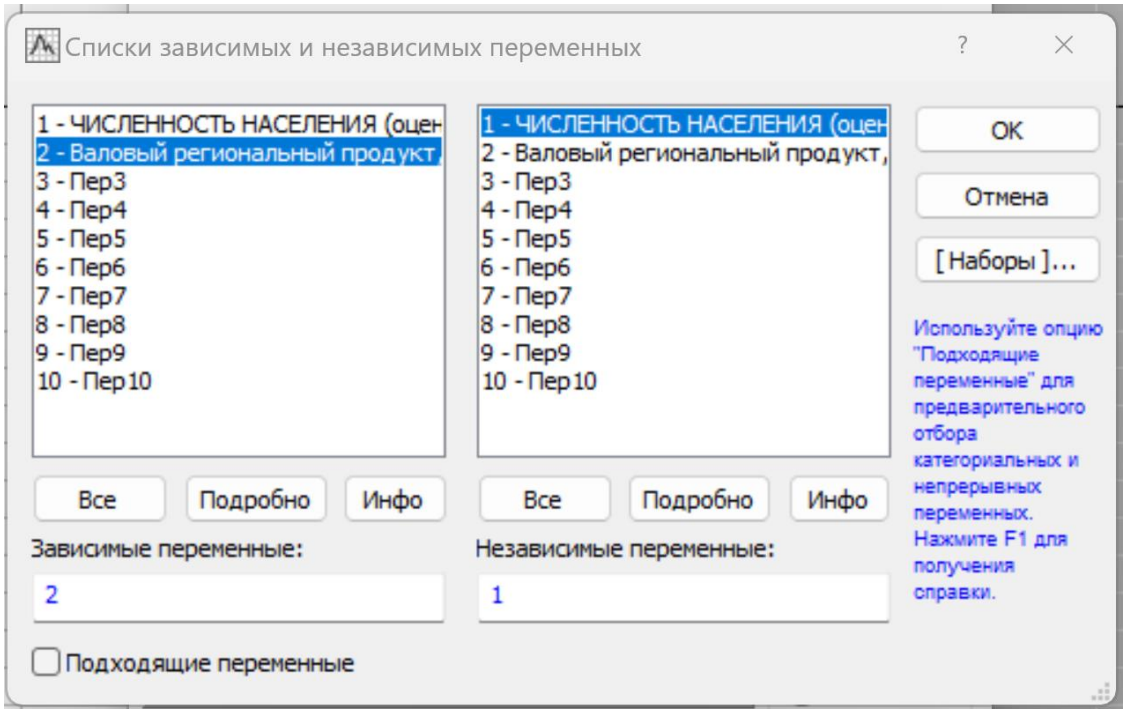


Рис. 2.6.14. Выбор переменных в меню **Множественная регрессия**

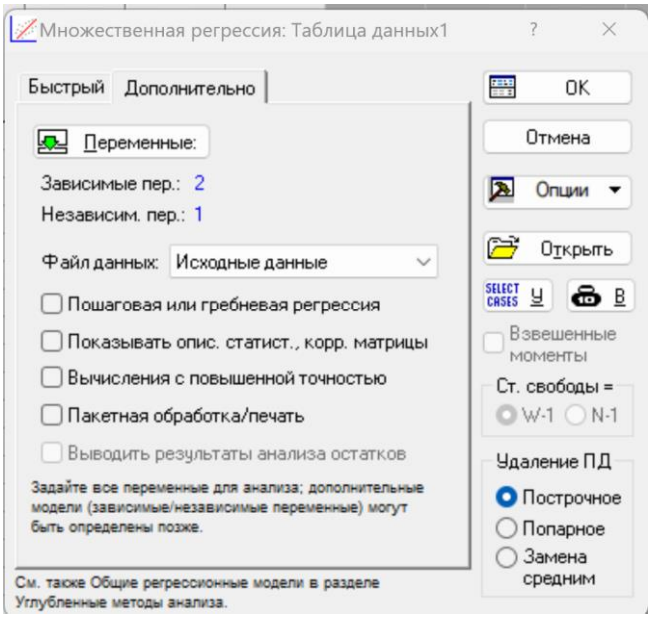


Рис. 2.6.15. Выбор дополнительных параметров регрессионного анализа

- показывать описательную статистику и корреляционные матрицы (рис. 2.6.16) – представляет выбор дополнительно выводимых данных: средние и стандартные отклонения, корреляции, ковариации, диаграмма размаха, матричный график;

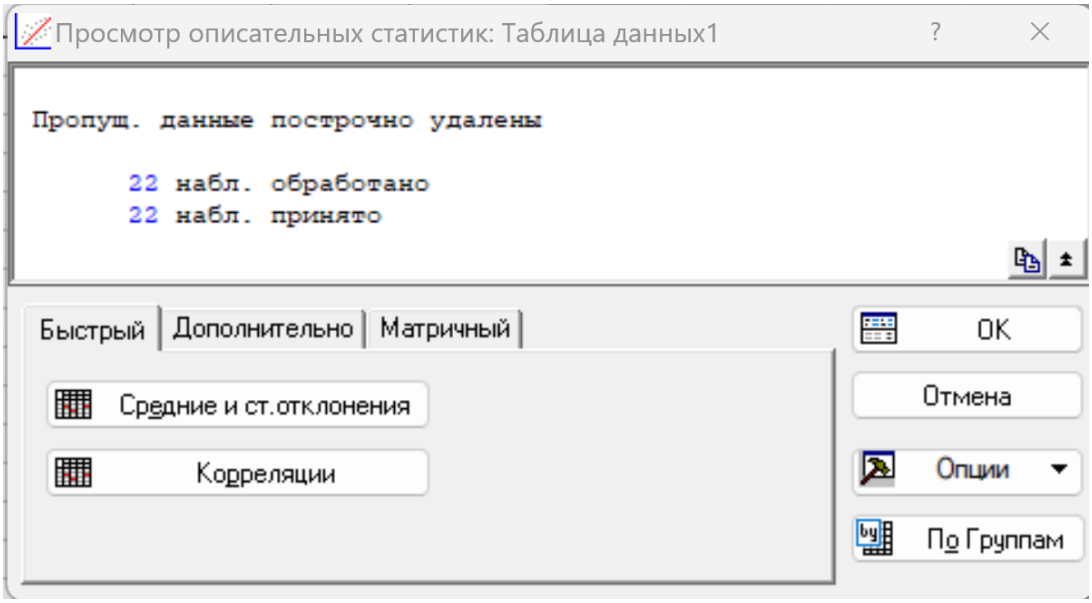


Рис. 2.6.16. Выбор дополнительных параметров регрессионного анализа

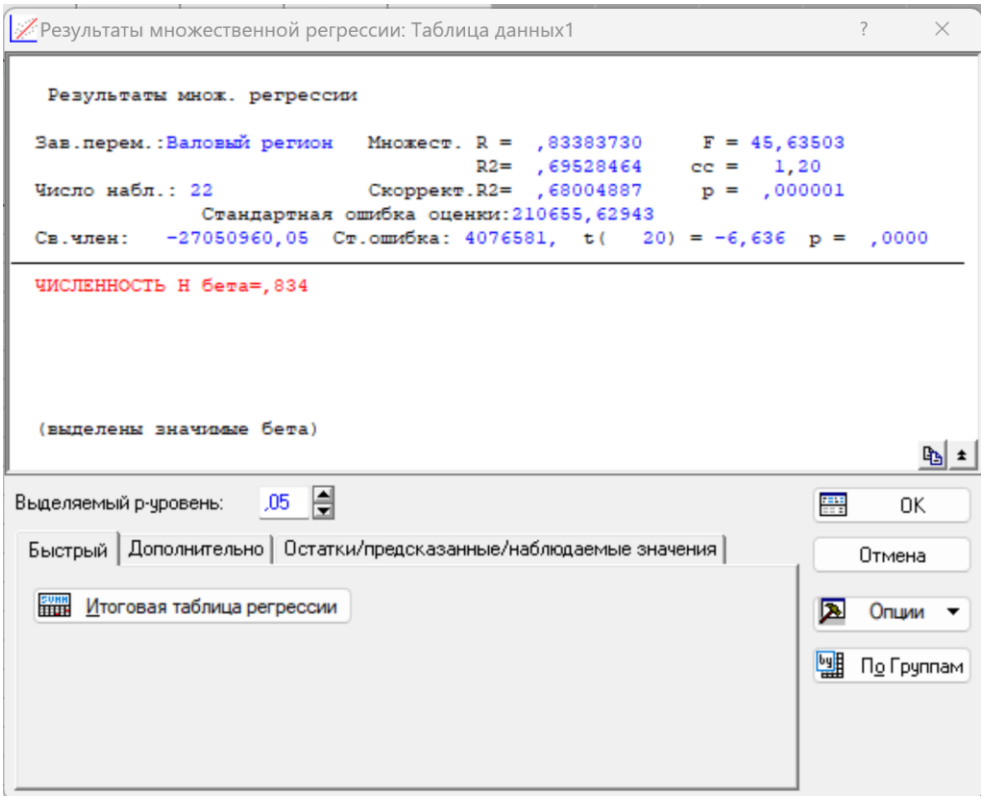


Рис. 2.6.17. Результаты множественной регрессии

- вычисление с повышенной точностью – данную галочку целесообразно устанавливать в том случае, если анализируемые показатели имеют крайне малую относительную дисперсию.

После установки галочек (при необходимости) необходимо нажать **ОК**. Появится вкладка результатов множественной регрессии (рис. 2.6.17).

Вкладка **Быстрый**. При нажатии кнопки «Итоговая таблица регрессии» будут выведены две итоговые формы регрессионного анализа.

Итоговые статистики (рис. 2.6.18) содержит:

- **Множественное R** – показатель множественной корреляции (совпадает с рис считанной ранее в разделе, рассматривающем решение задачи в Excel). В данном случае теснота связи может быть определена как высокая;

Статистика	Итоговые статистик **Значение**
Множест. R	0,833837297
Множест. R2	0,695284637
Скоррект. R2	0,680048869
F(1,20)	45,6350235
p	0,00000143076534
Стд. Ош. Оценки	210655,629

Рис. 2.6.18. Форма **Итоговые статистики**

- **Множественное R^2 –** коэффициент детерминации. В данном случае он равен 0,695, что соответствует 69,5% дисперсии, обусловленной факторами, входящими в модель;

- **Скорректированный R^2 –** скорректированный коэффициент детерминации означает, какое влияние корректировка R-квадрата оказала на величину коэффициента детерминации. Недостатком R-квадрата является то, что он увеличивается при добавлении новых объясняющих переменных (хотя это и не обязательно означает улучшение качества регрессионной модели), в то время как нормированный R-квадрат может уменьшаться при введении в модель новых объясняю-

щих переменных, не оказывающих существенное влияние на зависимую переменную. Если нормированный R-квадрат ненамного отличается от коэффициента детерминации, можно сделать вывод о хорошем качестве модели. В данном случае можно сделать вывод о приемлемом качестве регрессионной модели;

- **F-критерий Фишера** – представляет собой отношение дисперсии (среднего квадрата) вариантов к дисперсии ошибки т.к. $F_{табл.} < F_{факт}$, то H_0-гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность;

- **p-значение** – вероятность значимости – отвергаем нулевую гипотезу об отсутствии различий в целом между средними на уровне 0,000001;

- **стандартная ошибка оценки**.

Итоги регрессии для зависимой переменной (рис. 2.6.19) – в табличной форме представлены основные итоги регрессионного анализа.

Итоги регрессии для зависимой переменной: Валовый региональный продукт, млн руб.
R= ,83383730 R2= ,69528464 Ñêîððåêò. R2= ,68004887
F(1,20)=45,635 p<,00000 Ñòàíä. îøèáêà îöåíêè: 2107E2

N=22	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(20)	p-знач.
Св.член			-27050960	4076580	-6,63570	0,000002
ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (оценка на конец года; тысяч человек), x	0,833837	0,123433	17972	2660	6,75537	0,000001

Рис. 2.6.19. **Итоги регрессии для зависимой переменной**

В шапке таблицы приведены основные сведения: коэффициент корреляции, коэффициент детерминации, значение критерия Фишера, уровень значимости, стандартная ошибка.

В таблице представлены:

- **Бета** - коэффициенты, которые были бы получены, если бы мы заранее стандартизовали все переменные, то есть сделали их среднее равным 0, а стандартное отклонение равным 1. Одно из преимуществ бета-коэффициентов заключается в том, что бета-коэффициенты позволяют сравнить относительные вклады каждой независимой переменной в предсказание зависимой переменной. В случае парной корреляции данный коэффициент равен коэффициенту линейной корреляции;

- **Стандартная ошибка бета**;

- **B-коэффициенты** – коэффициенты регрессии в полученном уравнении;

- **стандартная ошибка B-коэффициентов;**

- t-критерий;

- уровень значимости.

В рассматриваемом примере уравнение будет иметь вид:

$$y=-27050960+17972x$$

Вкладка Дополнительно – предоставляет расширенные возможности представления корреляционно-регрессионного анализа (рис. 2.6.20).

Помимо итоговой таблиц в регрессии имеется возможность проведения **Дисперсионного анализа**.

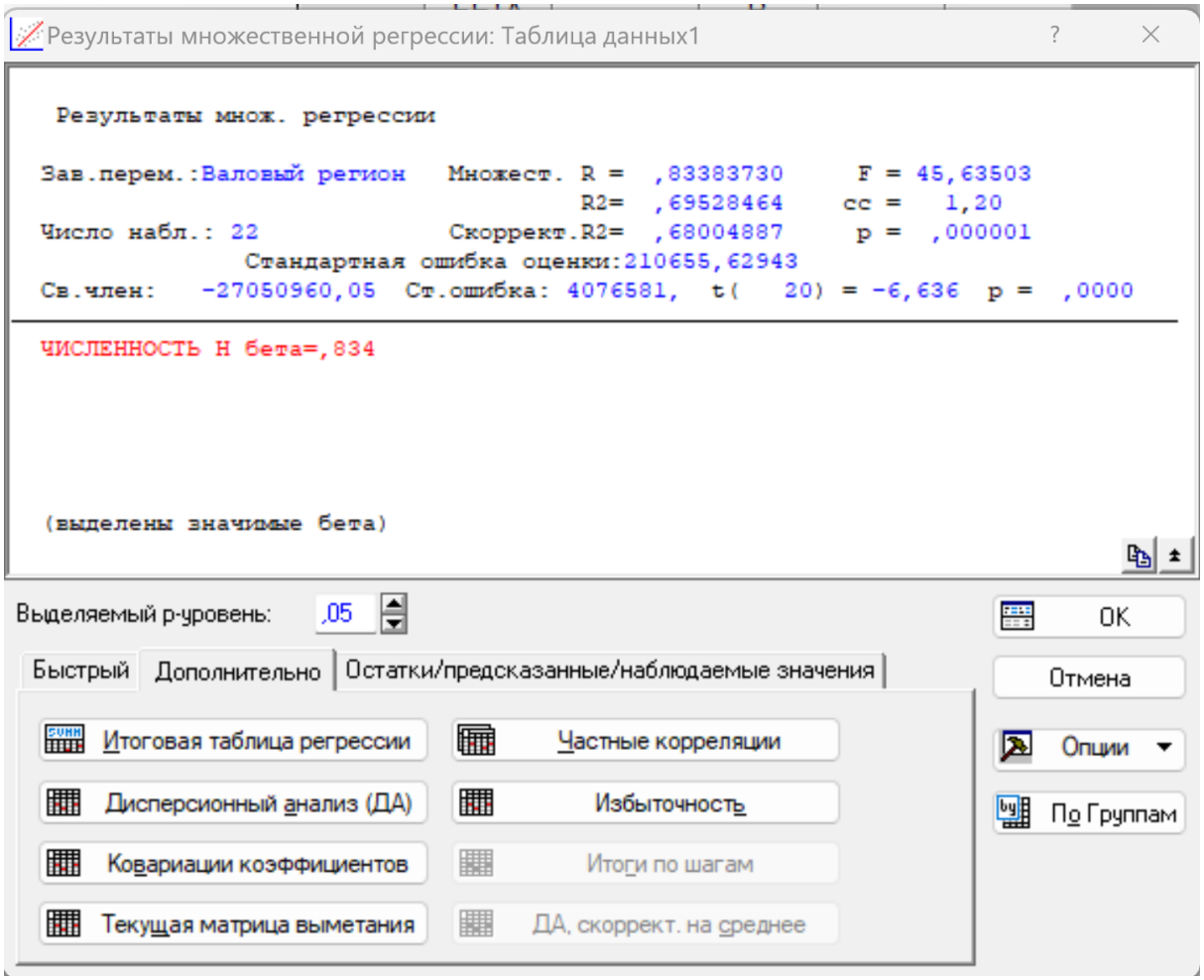


Рис. 2.6.20. **Итоги регрессии для зависимой переменной**

При нажатии кнопки **Дисперсионный анализ** выводятся итоги, которые позволяют более детально ознакомиться с результатами дисперсионного анализа уравнения регрессии (рис. 2.6.21).

Эффект	Дисперсионный анализ; ЗП: Валовый региональный продукт, млн руб. y (Таблица данных1)				
	Сумма квадр.	cc	Средн. квадр.	F	p-знач.
Регресс.	2,025090E+12	1	2,025090E+12	45,63503	0,000001
Остатки	8,875159E+11	20	4,437579E+10		
Итого	2,912606E+12				

Рис. 2.6.21. **Дисперсионный анализ**

В строках таблицы дисперсионного анализа регрессии из общей суммы вариации Итого на регрессию Регрессия приходится 69,5% и 30,5% на остаточную или случайную вариацию Остатки. F - критерий полученного уравнения регрессии значим на 5% уровне. Вероятность нулевой гипотезы p= 0,000001 значительно меньше 0,05, что говорит об общей значимости предлагаемой модели регрессии.

Для построения точечного графика и теоретической линии регрессии необходимо перейти на вкладку **Остатки/предсказанные/наблюдаемые значения (**рис. 2.6.22) и нажать на кнопку **Анализ остатков** (рис. 2.6.23).

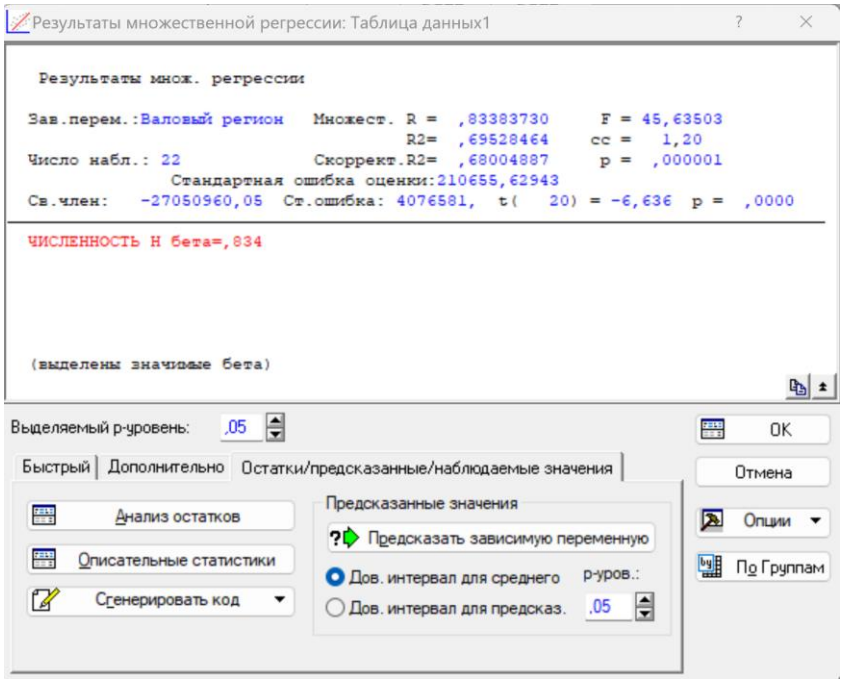


Рис. 2.6.22. Меню **Остатки/предсказанные/наблюдаемые значения**

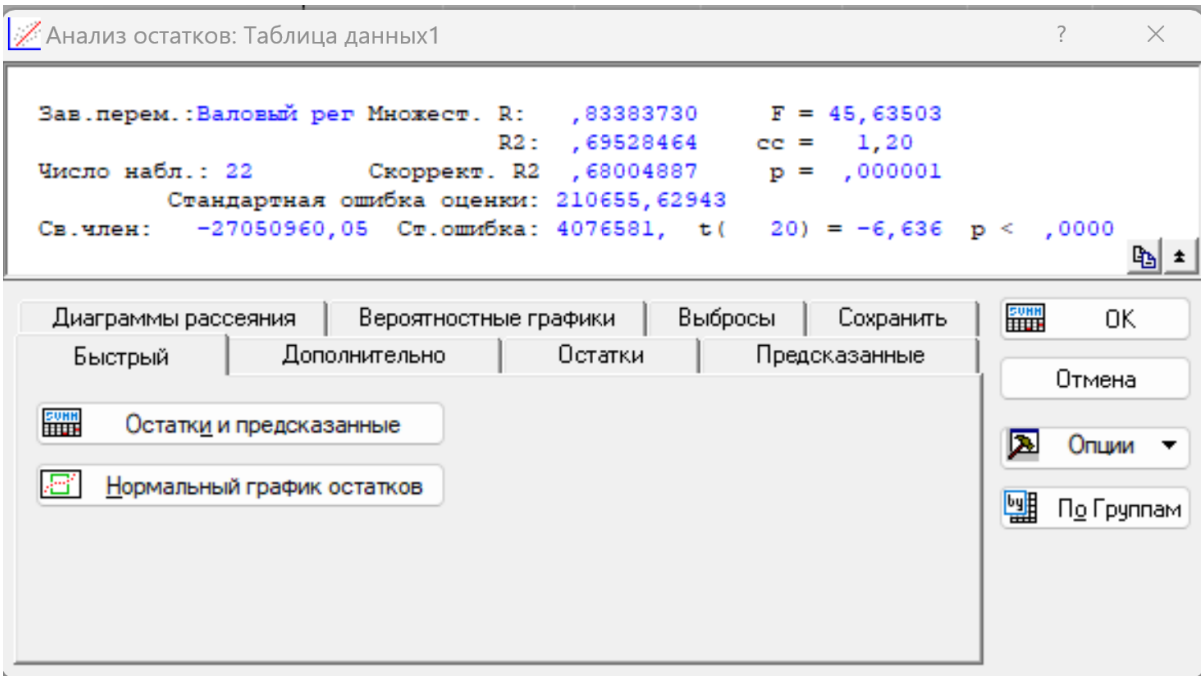


Рис. 2.6.23. Меню **Анализ остатков**

В данном меню представлено восемь вкладок с широким спектром возможностей.

Вкладка **Быстрый**. Содержит две кнопки:

- **Остатки и предсказания** (рис. 2.6.24) – в табличной форме выводятся результаты сравнения фактических (исходных) данных и предсказанных (теоретических, модельных).

Кроме непосредственных наблюдаемых и предсказанных значений, остатков, стандартной ошибки представлены две меры расстояний: расстояние Махаланобиса — мера расстояния между векторами случайных величин, обобщающая понятие евклидова расстояния и расстояние Кука - показывает разницу между вычисленными коэффициентами уравнения регрессии и значениями, которые получились бы при исключении соответствующего наблюдения. В адекватной модели все расстояния Кука должны быть примерно одинаковыми; если это не так, то имеются основания считать, что соответствующее наблюдение (или наблюдения) смещает оценки коэффициентов регрессии;

Набл. No.	Предсказанные значения и остатки (Таблица данных1) Зависимая перемен.: Валовый региональный продукт, млн руб. y								
	Наблюд. Значение	Предсказанные Значение	Остатки	Станд. предск.	Станд. Остатки	Ст.Ош. предск.	Махалан. расст.	Удален. Остатки	Кука расст.
1	42075	32232,0	9843	-1,46181	0,046723	80824,68	2,136897	11542	0,000221
2	49942	52000,7	-2059	-1,39815	-0,009774	78408,55	1,954831	-2390	0,000009
3	62404	120292,8	-57888	-1,17824	-0,274801	70360,64	1,388241	-65157	0,005337
4	76055	156236,0	-80181	-1,06249	-0,380628	66351,86	1,128887	-89013	0,008857
5	114409	116698,5	-2289	-1,18981	-0,010867	70771,03	1,415650	-2580	0,000008
6	144988	116698,5	28289	-1,18981	0,134292	70771,03	1,415650	31888	0,001293
7	178846	161627,5	17219	-1,04513	0,081738	65766,60	1,092295	19078	0,000400
8	237013	267659,9	-30647	-0,70368	-0,145482	55348,28	0,495165	-32919	0,000843
9	317656	379083,8	-61427	-0,34487	-0,291601	47627,79	0,118935	-64737	0,002414
10	304345	477927,5	-173582	-0,02657	-0,824009	44928,54	0,000706	-181854	0,016950
11	398361	488710,5	-90349	0,00815	-0,428895	44913,50	0,000067	-94652	0,004589
12	507840	564191,2	-56351	0,25122	-0,267505	46372,89	0,063112	-59221	0,001915
13	545517	659440,6	-113923	0,55795	-0,540804	51719,51	0,311304	-121231	0,009982
14	569006	724138,4	-155132	0,76629	-0,736425	57078,07	0,587197	-167424	0,023187
15	619678	801416,2	-181739	1,01514	-0,862728	64766,40	1,030510	-200711	0,042906
16	693379	848142,4	-154763	1,16561	-0,734673	69914,81	1,358646	-173921	0,037542
17	778028	903854,3	-125827	1,34502	-0,597309	76419,09	1,809065	-144895	0,031131
18	837307	858925,3	-21619	1,20033	-0,102625	71145,51	1,440800	-24402	0,000765
19	911598	821184,9	90413	1,07880	0,429198	66905,66	1,163810	100557	0,011493
20	955329	858925,3	96404	1,20033	0,457637	71145,51	1,440800	108816	0,015218
21	997331	724138,4	273193	0,76629	1,296868	57078,07	0,587197	294838	0,071909
22	1354811	562394,0	792417	0,24543	3,761668	46307,36	0,060237	832653	0,377490
Минимум	42075	32232,0	-181739	-1,46181	-0,862728	44913,50	0,000067	-200711	0,000008
Максим.	1354811	903854,3	792417	1,34502	3,761668	80824,68	2,136897	832653	0,377490
Среднее	486178	486178,1	-0	-0,00000	-0,000000	62496,61	0,954545	-1170	0,030203
Медиана	453101	525552,3	-43499	0,12679	-0,206493	66059,23	1,110591	-46070	0,007097

Рис. 2.6.24. Таблица данных **Предсказанные значения и остатки**

- **Нормальный график остатков** (рис. 2.6.25) – представление остатков не в табличной как ранее, а в графической форме.

Вкладка Дополнительно (рис. 2.6.26) содержит помимо рассмотренных выше возможностей статистику Дарбина-Уотсона.

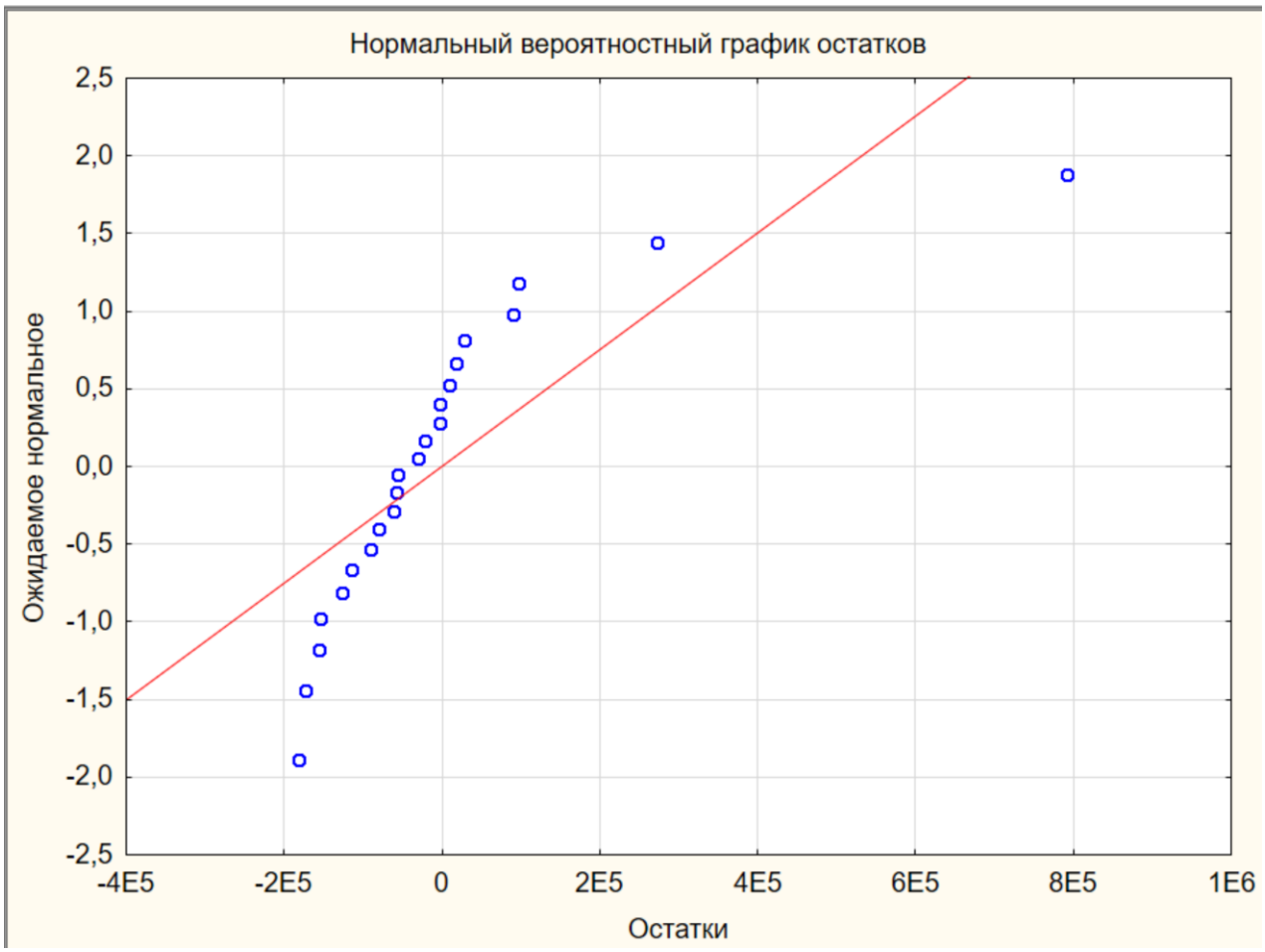


Рис. 2.6.25. Нормальный вероятностный график остатков

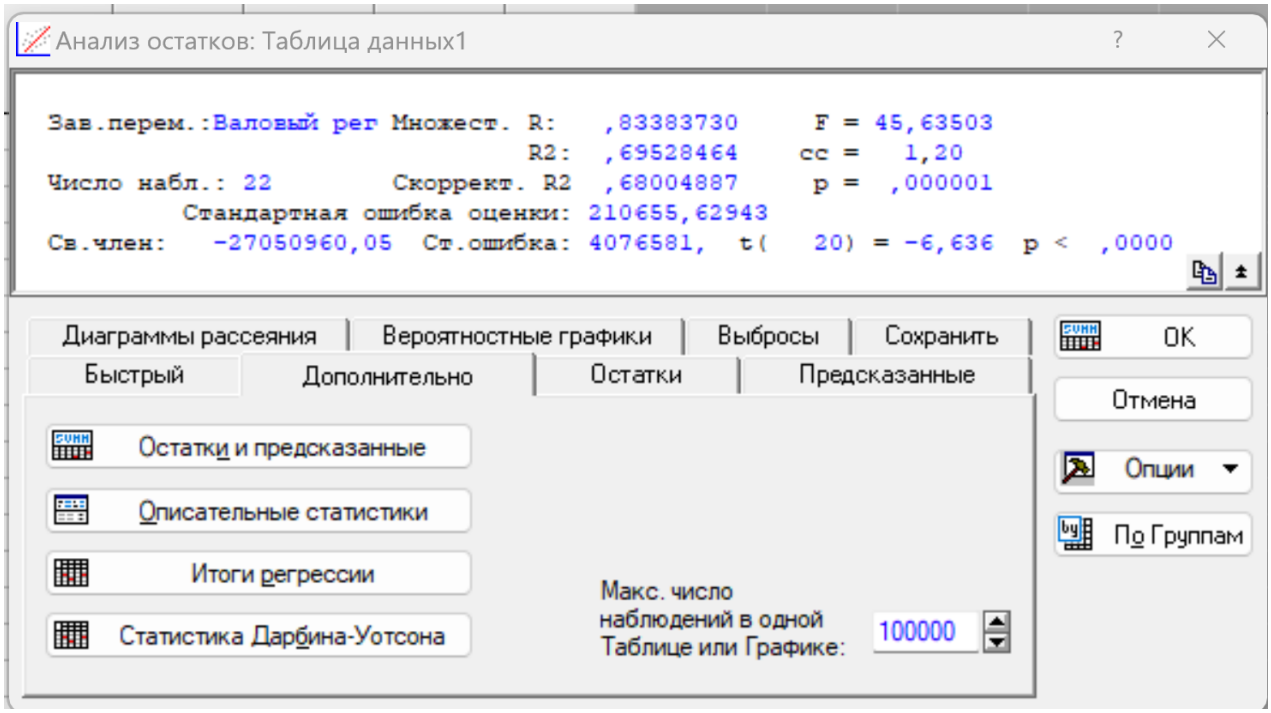


Рис. 2.6.26. Вкладка **Дополнительно**

Статистика Дарбина-Уотсона предназначена для проверки независимости регрессионных остатков. При нажатии соответствующей кнопки выводится табличная форма (рис. 2.6.27).

	Дарбина-Уотсона d (Таблица данных1) и сериальная корреляция остатков	
	Дарбина-Уотсон d	Сериаль Корр.
Оценка	0,412793	1,503610

Рис. 2.6.27. **Статистика Дарбина-Уотсона**

Значение критерия для трактовки критерия представлены в таблице 2.6.2. Существуют полные таблицы, позволяющие трактовать более сложные модели.

Таблица 2.6.2

Распределение критерия Дарбина-Уотсона для положительной автокорреляции (для 5%-го уровня значимости) (фрагмент) (n - объем выборки, m - число объясняющих переменных в уравнении регрессии)

n	***m*=1**		***m* =2**		***m* =3**		***m* =4**		***m* =5**	
	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
6	0,610	1,400								
7	0,7000	1,356	0,467	1,896						
8	0,763	1,332	0,359	1,777	0,368	2,287				
9	0,824	1,320	0,629	1,699	0,435	2,128	0,296	2,388		
10	0,879	1,320	0,697	1,641	0,525	2,016	0,356	2,414	0,243	2,822
11	0,927	1,324	0,658	1,604	0,595	1,928	0,444	2,283	0,316	2,645
12	0,971	1,331	0,812	1,576	0,658	1,864	0,512	2,177	0,379	2,506
13	1,010	1,340	0,861	1,562	0,715	1,816	0,574	2,094	0,445	2,390
14	1,045	1,330	0,905	1,551	0,767	1,779	0,632	2,030	0,505	2,296
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977	0,562	2,220
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935	0,615	2,157
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900	0,664	2,104
18	1,158	1,391	1,046	1,535	0,933	1,696	0,820	1,872	0,710	2,060
19	1,180	1,401	1,074	1,536	0,967	1,685	0,859	1,848	0,752	2,023
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828	0,792	1,991
21	1,221	1,420	1,125	1,538	1,026	1,669	0,927	1,812	0,829	1,964
22	1,239	1,429	1,147	1,541	1,053	1,664	0,958	1,797	0,863	1,940

23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,785	0,895	1,920
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775	0,925	1,902
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767	0,953	1,886
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759	0,979	1,873
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753	1,004	1,861
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747	1,028	1,850
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743	1,050	1,841
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739	1,071	1,833
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735	1,090	1,825
32	1,373	1,502	1,309	1,574	1,244	1,650	1,177	1,732	1,109	1,819
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730	1,217	1,813
34	1,393	1,514	1,333	1,580	1,271	1,652	1,208	1,728	1,144	1,808
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726	1,160	1,803
36	1,411	1,525	1,354	1,587	1,295	1,654	1,236	1,724	1,175	1,799
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723	1,190	1,795
38	1,427	1,535	1,373	1,594	1,318	1,656	1,261	1,722	1,204	1,792
39	1,435	1,540	1,382	1,587	1,328	1,658	1,273	1,722	1,218	1,789
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721	1,230	1,786
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720	1,287	1,776
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721	1,335	1,771
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724	1,374	1,768
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727	1,408	1,767
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731	1,438	1,767
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735	1,464	1,768

Вкладка **Остатки** (рис. 2.6.28) – дает расширенные графические возможности для представления результата анализа

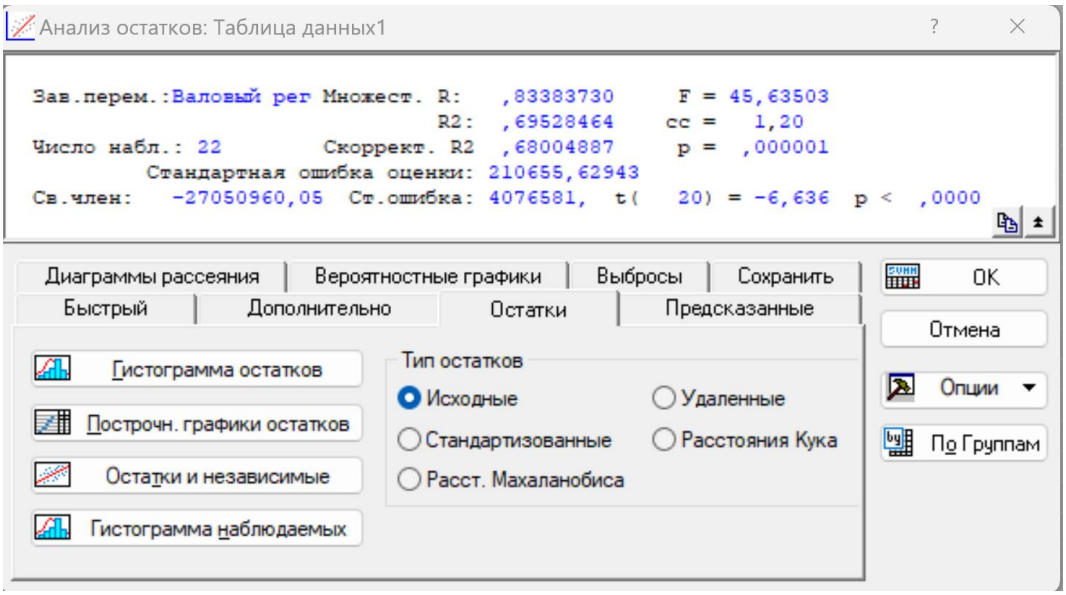


Рис. 2.6.28. Вкладка **Остатки**

Можно построить гистограмму остатков, построчные графики остатков, график остатков и независимых, гистограмму наблюдаемых. В качестве примера ниже представлен график **Остатки и независимые** (рис. 2.6.29). На графике представлены наблюдаемые значения, доверительный интервал, график линейной регрессии. Кроме того, в шапке графика представлены дополнительные параметры.

Удаленные остатки — стандартизованное значение остатка, которое имело бы данное наблюдение, если его значение не учитывать при расчетах регрессионного уравнения. Если удаленный остаток значительно отличается от соответствующего стандартизованного значения остатка, то возможно, что это наблюдение является выбросом, поскольку его исключение изменяет уравнение регрессии.

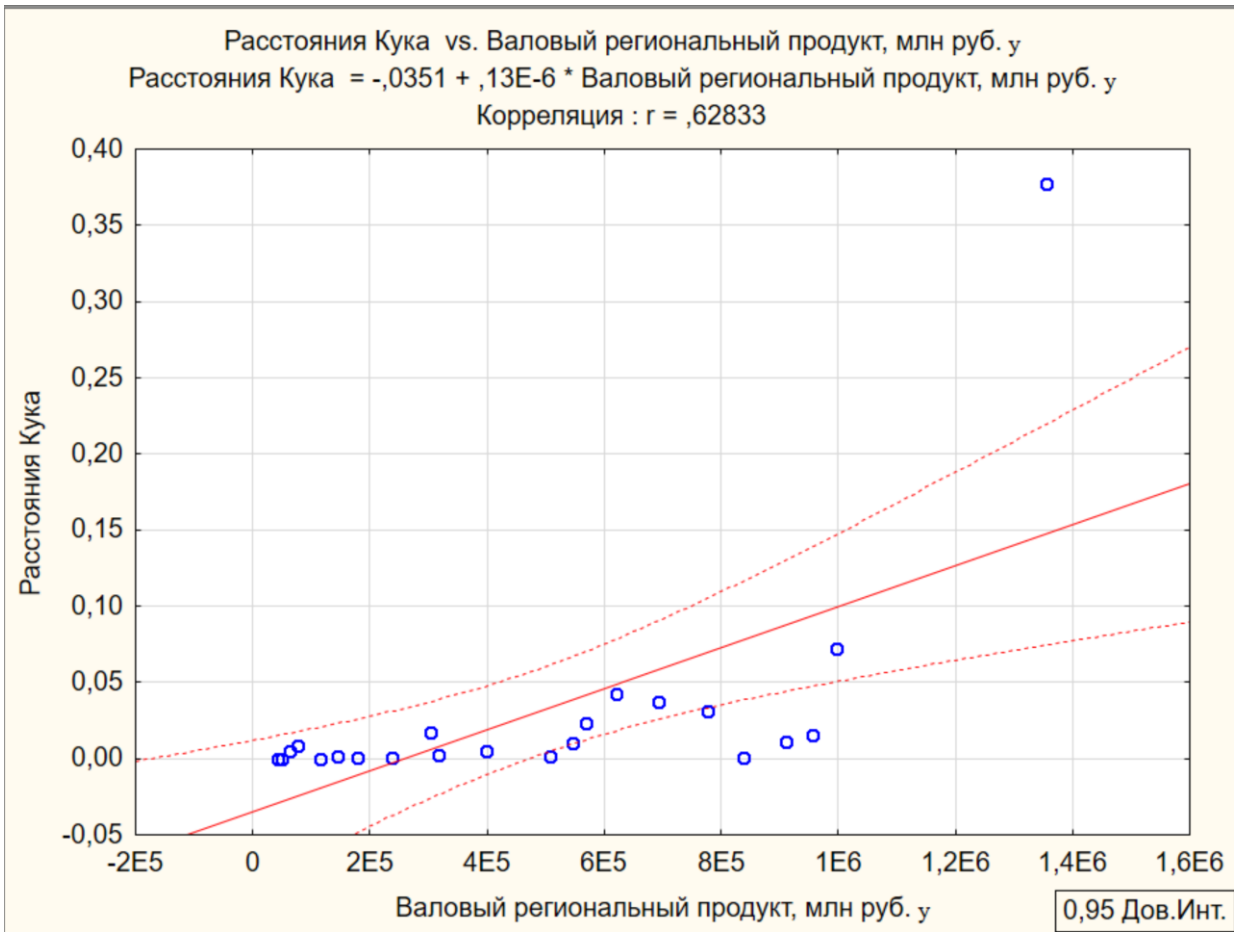


Рис. 2.6.29. График **Остатки и независимые**

На вкладке **Предсказанные** (рис. 2.6.30) представлены инструменты для анализа предсказанных (теоретических, модельных) значений.

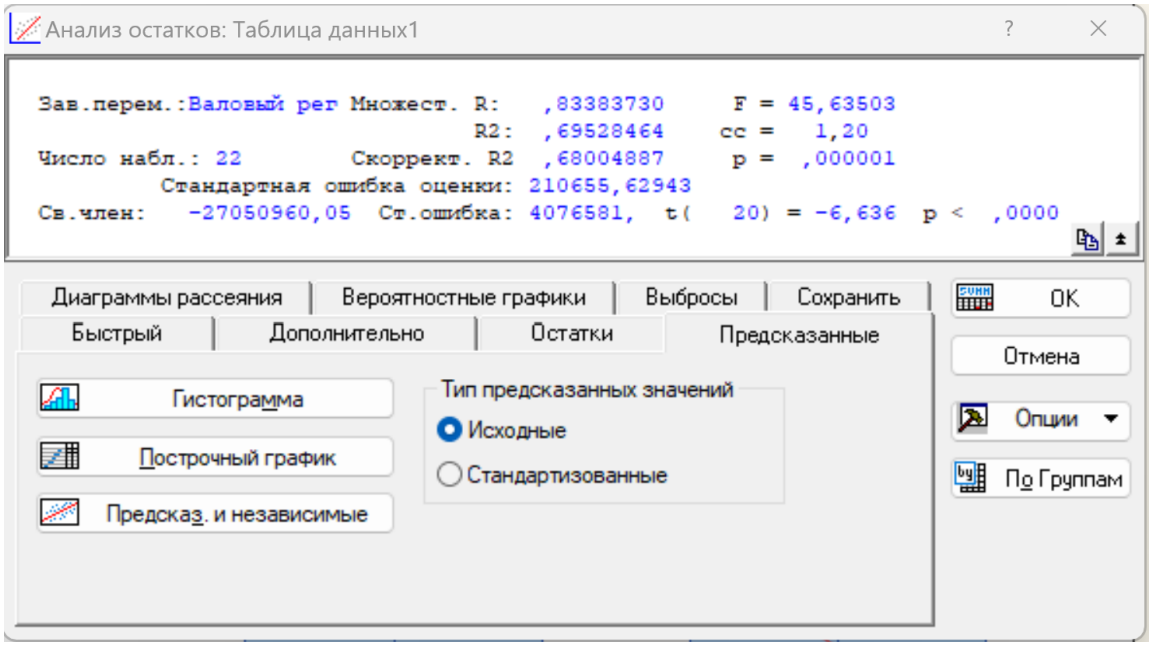


Рис. 2.6.30. Вкладка **Предсказанные**

Помимо рассмотренных выше графических возможностей, на вкладке присутствует возможность построения гистограммы распределения предсказанных значений. Данная функция доступна при нажатии кнопки Гистограмма (рис. 2.6.31).

Рис. 2.6.31. Гистограмма распределения предсказанных значений

На данной гистограмме приведено фактическое распределение предсказанных величин, а также кривая нормального распределения.

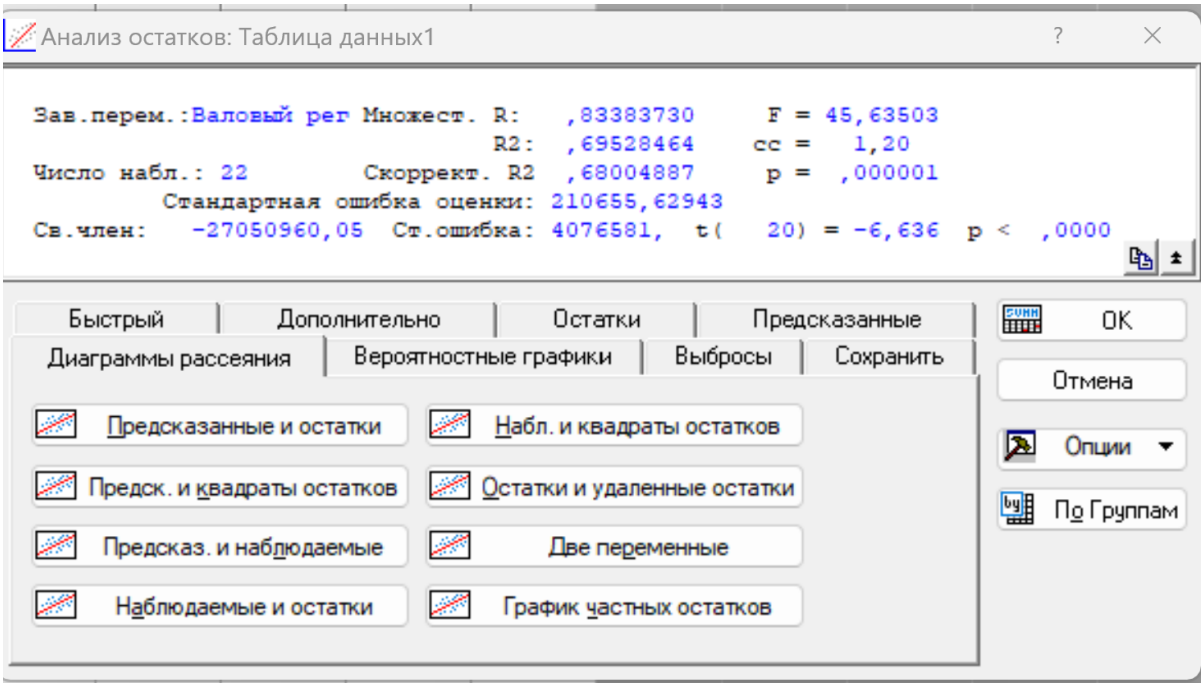


Рис. 2.6.32. **Вкладка Диаграммы**

Вкладка **Диаграммы** (рис. 2.6.32) рассеяния дает возможность построения нескольких типов графиков, облегчающих трактовку полученных результатов анализа остатков.

Помимо рассмотренных выше, тут представлен график, позволяющий на одном поле представить две переменные (соответственно кнопка **Две переменные** – рис. 2.6.33) и график **Наблюдаемые и квадраты остатков** (соответствующая кнопка – рис. 2.6.34). На последнем также представлены границы доверительного интервала и линия регрессии.

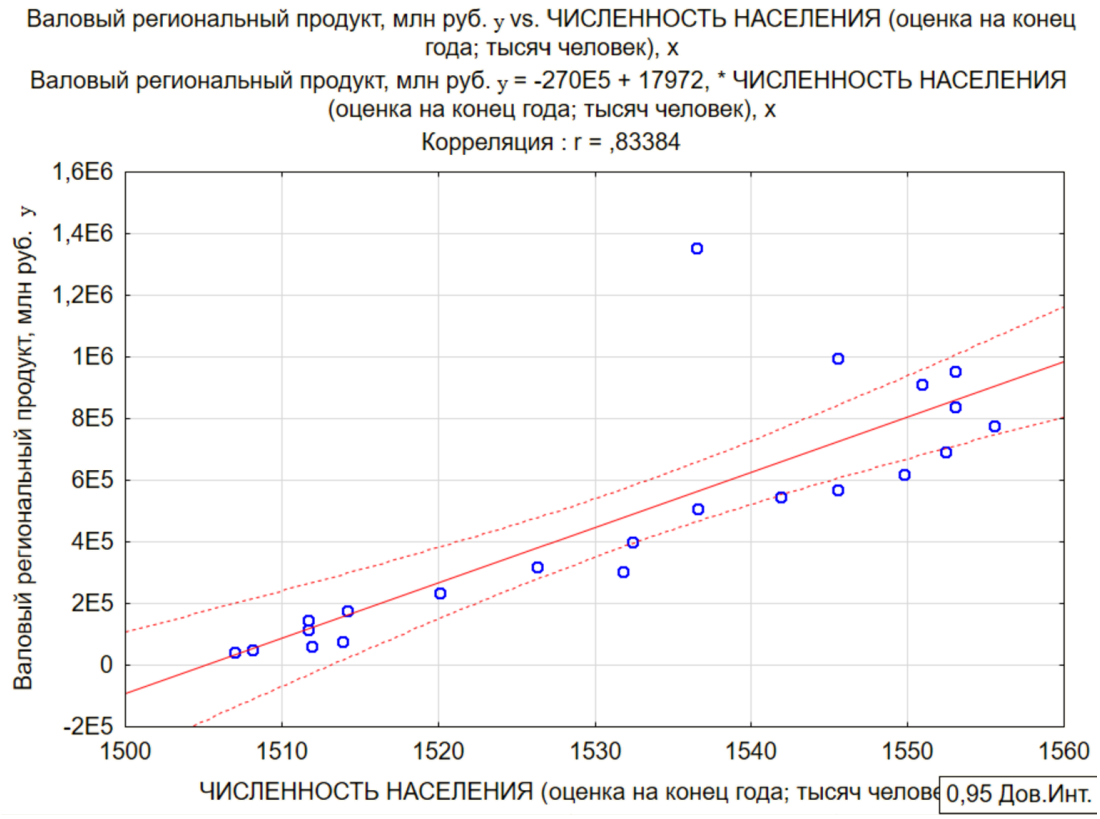


Рис. 2.6.33. График **Две переменные**

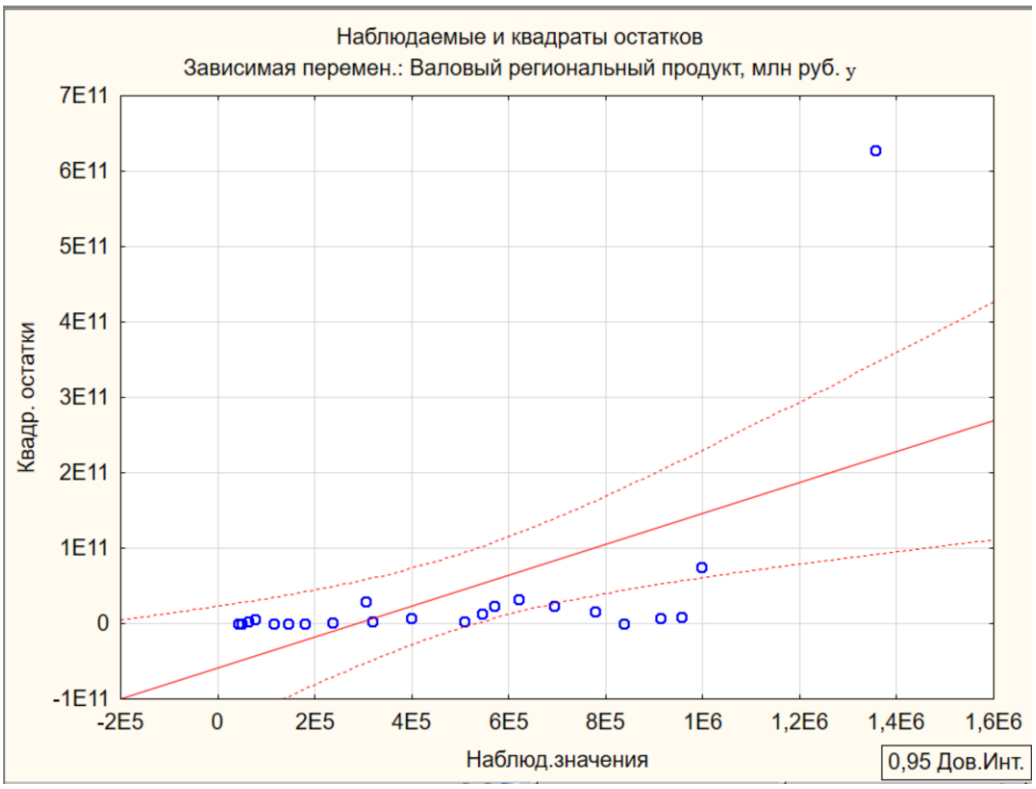


Рис. 2.6.34. График **Наблюдаемые и квадраты остатков**

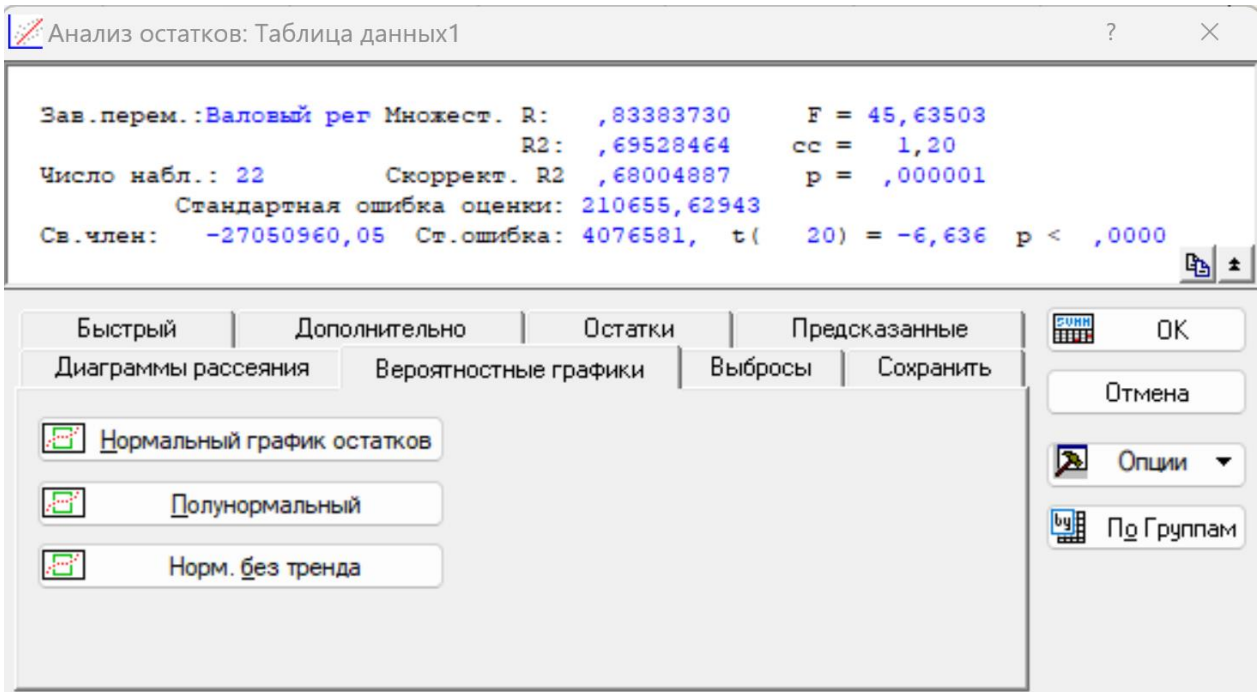


Рис. 2.6.35. Вкладка **Вероятностные графики**

Вкладка **Вероятностные графики** (рис. 2.6.35) предоставляет возможность построения трех типов графиков:

- **нормальный график остатков** (рис. 2.6.36) - если остатки имеют распределение, отличное от нормального, то точки на графике будут отклоняться от прямой. На этом графике также становятся заметны выбросы. Если модель плохо согласуется с наблюдениями, и данные располагаются специальным образом около прямой линии (например, имеют S-образный вид), то, возможно, требуется нелинейное преобразование зависимой переменной, например логарифмирование или извлечение квадратного корня;

- **полунормальный график остатков** (рис. 2.6.37) - данный вероятностный график строится так же, как и нормальный за исключением того, что рассматривается лишь положительная часть распределения и только положительные нормальные значения изображаются на вертикальной оси. Этот график используется в тех случаях, когда исследователя не интересует знак остатка и определяющую роль играют только абсолютные значения;

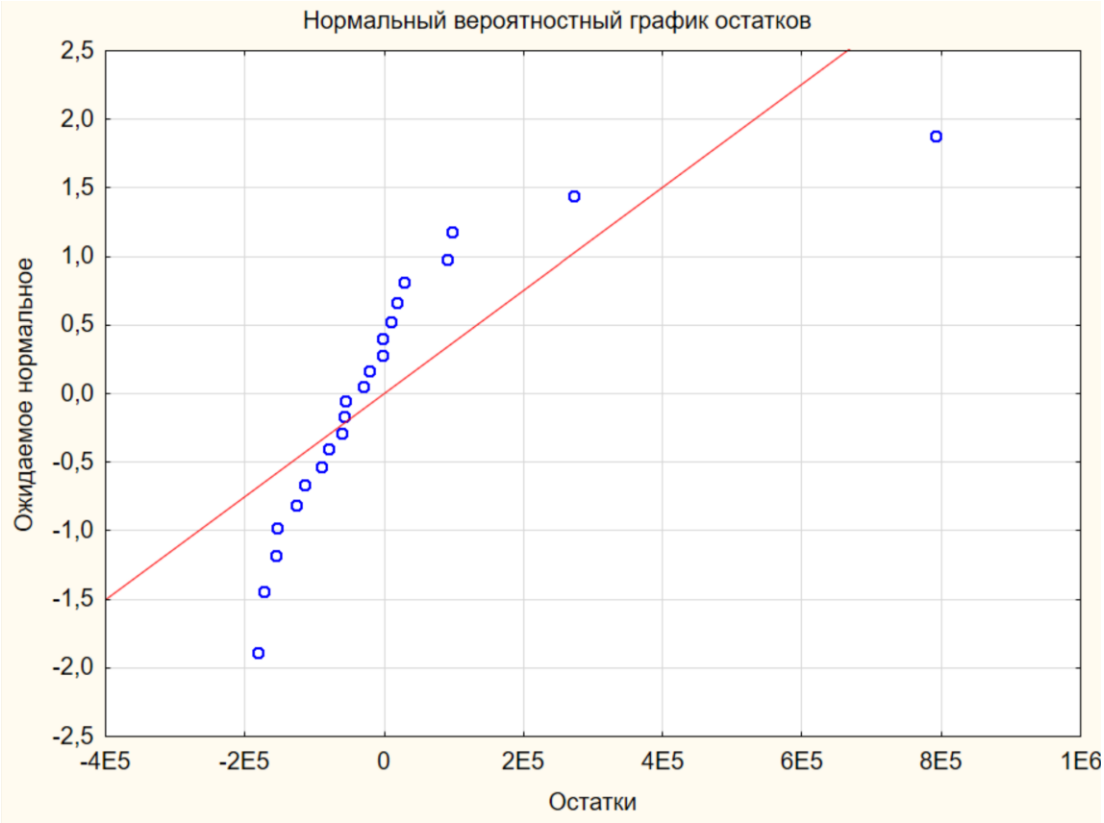


Рис. 2.6.36. **Нормальный график остатков**

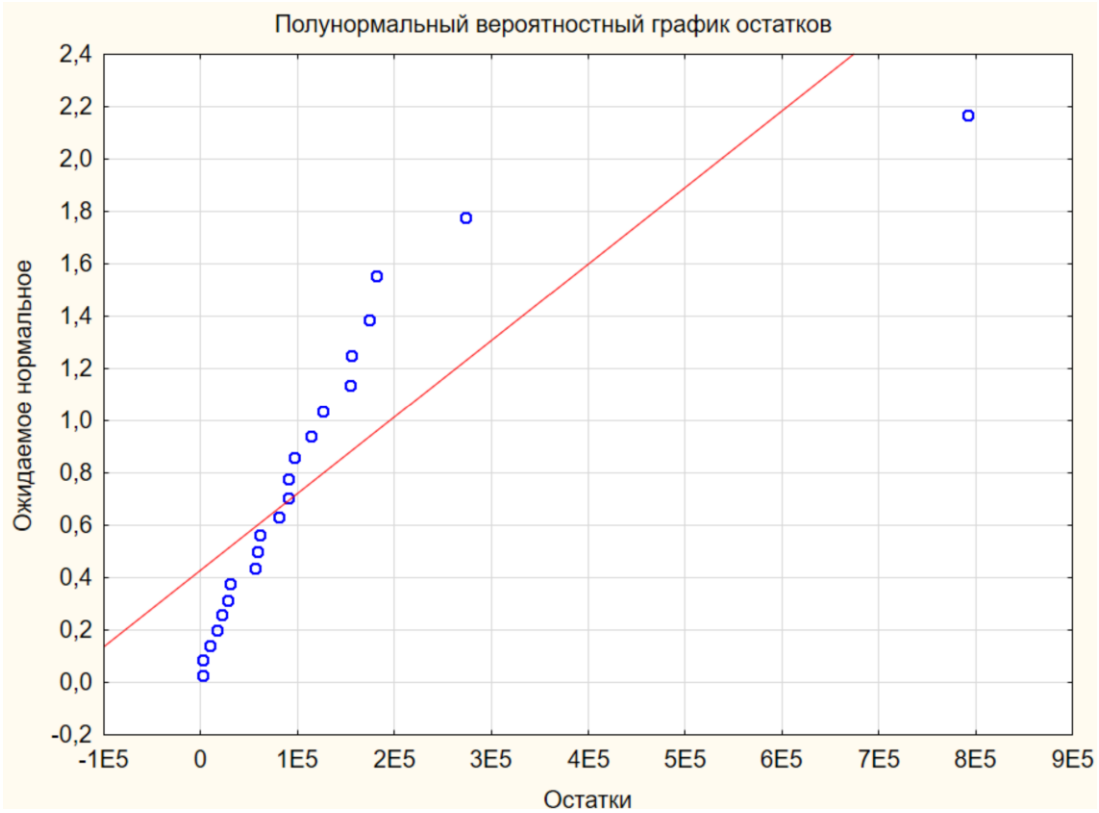


Рис. 2.6.37. **Полунормальный вероятностный график остатков**

- нормальный вероятностный график с исключенным трендом (рис. 2.6.38).

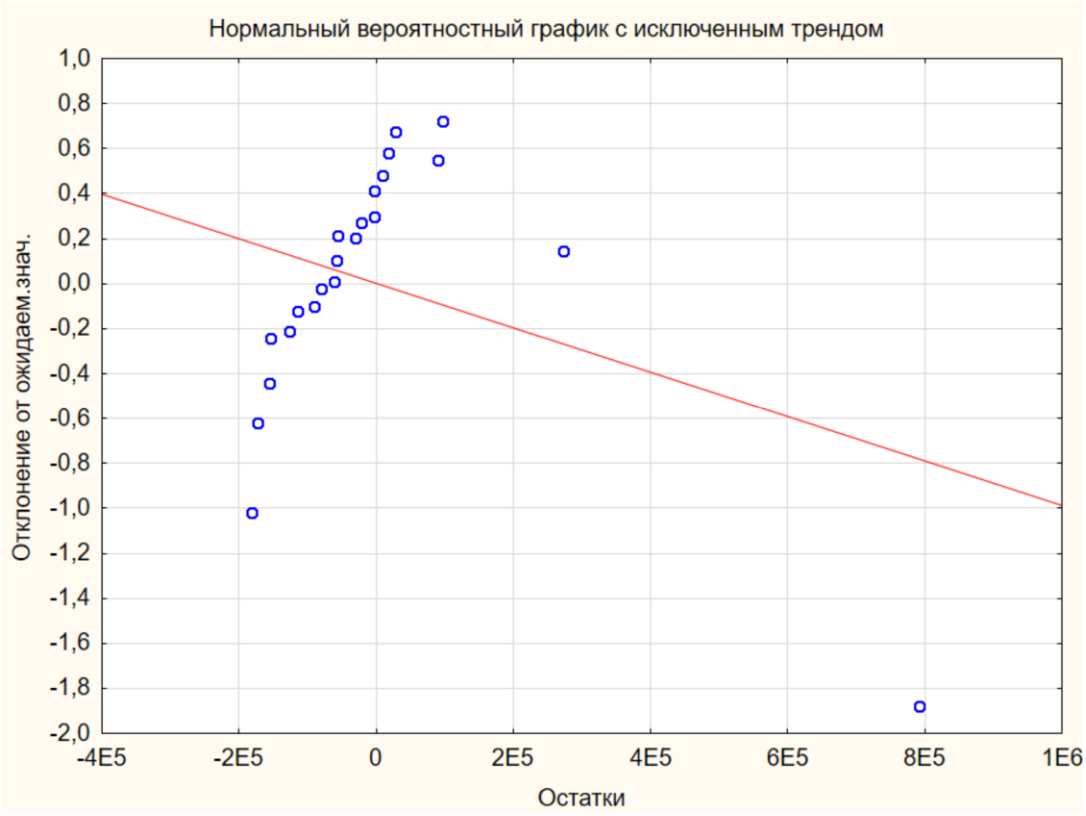


Рис. 2.6.38 **Нормальный вероятностный график с исключенным трендом**

Вкладка **Выбросы** (рис. 2.6.39) – предлагает выбор инструментов для оценки статистических выбросов.

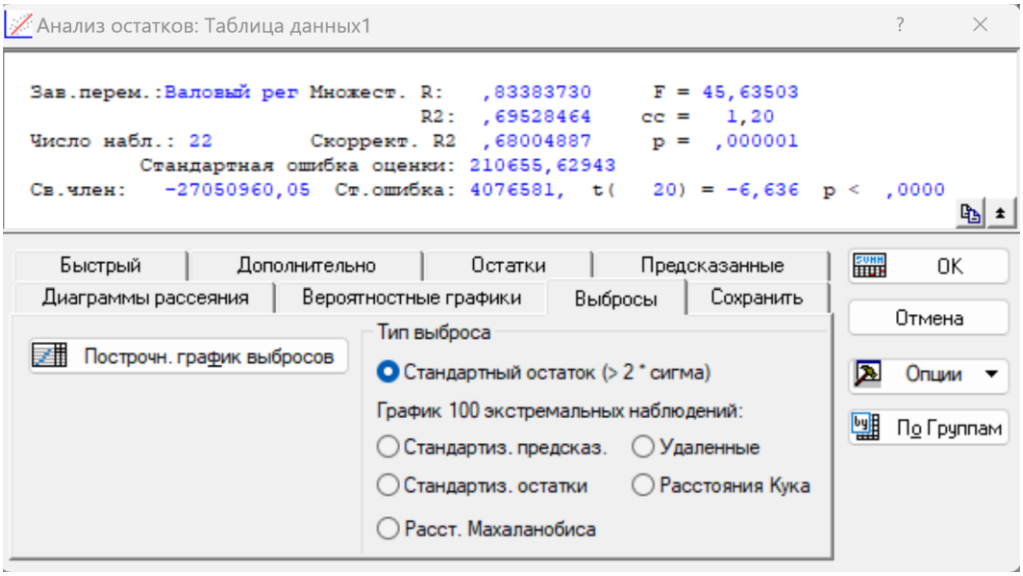


Рис. 2.6.39 Вкладка **Выбросы**

При этом можно выбрать форму представления. По умолчанию задается тип выброса по стандартному остатку. Построчный график выброса представлен на рисунке 2.6.40, на котором приведены помимо прочего и все доступные для расчета меры расстояния.

Станд. остатки (Таблица данных1)
Выбросы

Набл.	Стандартиз. остатки -5. -4. -3. ±2. 3. 4. 5.	Наблюд. Значение	Предсказанные Значение	Остатки	Станд. предск.	Станд. Остатки	Ст.Ош. предск.	Махалан. расст.	Удален. Остатки	Кука расст.
22	. . . | . * . .	1354811	562394,0	792416,5	0,245433	3,761668	46307,36	0,060237	832652,8	0,377490
Минимум	. . . | . * . .	1354811	562394,0	792416,5	0,245433	3,761668	46307,36	0,060237	832652,8	0,377490
Максим.	. . . | . * . .	1354811	562394,0	792416,5	0,245433	3,761668	46307,36	0,060237	832652,8	0,377490
Среднее	. . . | . * . .	1354811	562394,0	792416,5	0,245433	3,761668	46307,36	0,060237	832652,8	0,377490
Медиана	. . . | . * . .	1354811	562394,0	792416,5	0,245433	3,761668	46307,36	0,060237	832652,8	0,377490

Рис. 2.6.40. Построчный график выброса

Таким образом были рассмотрены возможности программного комплекса Statistica для построения парной линейной регрессии на примере.

Контрольные вопросы по теме

1. Что такое корреляционный анализ?
2. Какие виды корреляции существуют?
3. Что такое коэффициент корреляции Пирсона и как его интерпретировать?
4. Как вычисляется коэффициент корреляции Пирсона?
5. Какие значения может принимать коэффициент корреляции Пирсона и что они означают?
6. Что такое коэффициент корреляции Спирмена и в чем его отличие от коэффициента корреляции Пирсона?
7. Как интерпретировать результаты коэффициента корреляции Спирмена?
8. Что такое коэффициент корреляции Кендалла и в чем его преимущества?
9. Какие факторы могут влиять на значение коэффициента корреляции?
10. Как можно определить статистическую значимость коэффициента корреляции?
11. Что такое диаграмма рассеяния и как она используется при корреляционном анализе?
12. Как можно определить наличие линейной зависимости между двумя переменными с помощью диаграммы рассеяния?

13. Какие проблемы могут возникнуть при интерпретации коэффициента корреляции?

14. Какие могут быть причины обнаружения ложной корреляции между переменными?

15. Что такое регрессионный анализ?

16. Какие виды регрессии существуют?

17. Что такое линейная регрессия и как она используется для предсказания значений зависимой переменной?

18. Как строится уравнение линейной регрессии?

19. Как можно интерпретировать коэффициенты уравнения линейной регрессии?

20. Что такое коэффициент детерминации и как его интерпретировать?

21. Как можно определить статистическую значимость уравнения линейной регрессии?

22. Какие предпосылки должны быть выполнены для применения линейной регрессии к данным?

23. Что такое множественная регрессия и в чем её отличие от простой линейной регрессии?

24. Как можно использовать множественную регрессию для анализа влияния нескольких независимых переменных на зависимую переменную?

25. Какие методы используются для проверки предпосылок линейной регрессии?

26. Что такое автокорреляция и как она может повлиять на результаты регрессионного анализа?

27. Что такое мультиколлинеарность и как она может повлиять на результаты множественной регрессии?

28. Как можно использовать результаты регрессионного анализа для прогнозирования будущих значений зависимой переменной?

29. Как можно оценить точность прогнозов, полученных с помощью регрессионного анализа?

30. Какие методы можно использовать для обработки выбросов в данных перед регрессионным анализом?

31. Как можно использовать результаты регрессионного анализа для выявления важных факторов, влияющих на зависимую переменную?

32. Как можно использовать регрессионный анализ для выявления тенденций в данных?

33. Как можно определить, что данные подходят для анализа с помощью корреляционного и регрессионного анализа?

34. Какие программные инструменты чаще всего используются для проведения корреляционно-регрессионного анализа?

35. Как можно использовать результаты корреляционно-регрессионного анализа для принятия бизнес-решений?

36. Как можно использовать результаты корреляционно-регрессионного анализа для планирования маркетинговых стратегий?

37. Как можно использовать результаты корреляционно-регрессионного анализа для оптимизации производственных процессов?

38. Как можно использовать результаты корреляционно-регрессионного анализа для прогнозирования спроса на товары или услуги?

39. Как можно использовать результаты корреляционно-регрессионного анализа для выявления влияния различных факторов на финансовые показатели компании?

40. Как можно использовать результаты корреляционно-регрессионного анализа для оптимизации управленческих процессов в компании?

41. Как можно использовать результаты корреляционно-регрессионного анализа для прогнозирования тенденций на финансовых рынках?

42. Как можно использовать результаты корреляционно-регрессионного анализа для выявления связей между различными социальными явлениями или процессами?

43. Как можно использовать результаты корреляционно-регрессионного анализа для выявления факторов, влияющих на здоровье или качество жизни людей?

44. Как можно использовать результаты корреляционно-регрессионного анализа для оптимизации учебного процесса в образовательных учреждениях?

45. Как можно использовать результаты корреляционно-регрессионного анализа для выявления влияния различных факторов на экологические процессы или явления?

46. Как можно использовать результаты корреляционно-регрессионного анализа для оптимизации процессов в медицине или здравоохранении?

47. Как можно использовать результаты корреляционно-регрессионного анализа для выявления влияния различных факторов на поведение потребителей или клиентов компаний?

48. Как можно использовать результаты корреляционно-регрессионного анализа для выявления связей между различными экономическими явлениями или процессами?

49. Как можно использовать результаты корреляционно-регрессионного анализа для оптимизации процессов в транспортной отрасли или логистике?

50. Как можно использовать результаты корреляционно-регрессионного анализа для выявления влияния различных факторов на политические процессы или явления?

3. ДИСПЕРСИОННЫЙ АНАЛИЗ

3.1. Общее понятие дисперсионного анализа

Дисперсионный анализ представляет собой статистический метод, с помощью которого осуществляется анализ степени влияния факторов на результаты эксперимента, путем исследования значимости различий в средних значениях [21].

Сущность дисперсионного анализа (analysis of variance, ANOVA) заключается в том, чтобы разбить дисперсию измеряемого признака на отдельные элементы, описывающие влияние каждого отдельного фактора, а также их взаимодействия [42].

Последующее сравнение данных элементов позволяет выявить долю вариации исходных данных, обусловленную факторным влиянием и влиянием случайных отклонений.

Соответственно, становится возможной оценка значимости каждого из рассматриваемых факторов, а также их вариантов комбинаций.

Фактором при осуществлении дисперсионного анализа называется переменная, которая, предположительно, может оказывать значимый вклад на формирование конечного результата.

Дисперсионный анализ называется однофакторным, если рассматривается зависимость только от одного фактора, и многофакторным, если анализируется влияние двух или более признаков на результат [2].

Применительно к статистике фирмы могут быть рассмотрены следующие ситуации: предположим, что требуется построение модели объяснения выручки фирм тем, что они расположены в различных городах страны.

В данном случае переменная «месторасположение фирмы» будет выполнять роль анализируемого фактора. Уровнем фактора является конкретное его значение (например, наименование населенного пункта расположения фирмы).

Откликом в дисперсионном анализе принято называть значение измеряемого признака (в рамках рассматриваемого примера откликов выступают конкретные значения выручки фирм).

3.2. Однофакторный дисперсионный анализ

Для однофакторного дисперсионного анализа модель выглядит следующим образом (3.2.1):

$$y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad i = \overline{1, n_j}, \quad j = \overline{1, k} \qquad (3.2.1)$$

где: y_{ij} - i-ое наблюдаемое значение отклика в *j*-ой группе (для *j*-го уровня фактора);

μ - среднее значение отклика по всем уровням фактора (среднее по всей совокупности);

μ_j - среднее значение отклика для *j*-го уровня фактора;

$\alpha_j = \mu_j - \mu$ - дифференциальный эффект среднего, который соответствует *j*-му уровню фактора;

ε_{ij} - независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2.

Очевидно, что при заданных величинах μ_j, величины α_j и μ определяются неоднозначно, поэтому необходимо наложить дополнительное условие, устанавливающие связь между этими величинами. Обычно используют одно из следующих условий (3.2.2 или 3.2.3):

$$\sum_{i=1}^{k} \alpha_i = 0, \qquad (3.2.2)$$

$$\sum_{i=1}^{k} n_i \alpha_i = 0, \ \alpha_k = 0. \qquad (3.2.3)$$

Выражение $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ можно представить в виде (3.2.4):

$$y_{ij} = \mu + (\mu_j - \mu) + (y_{ij} - \mu_j) \qquad (3.2.4)$$

или:

$$y_{ij} - \mu = (\mu_j - \mu) + (y_{ij} - \mu_j). \qquad (3.2.5)$$

Математический смысл данного выражения заключается в следующем: отклонение наблюдаемого значения отклика для j-ой группы складывается из суммы двух слагаемых. Одним из них выступает отклонение отклика от среднего значения *j*-ой группы: $(y_{ij} - \mu_j)$, а вторым отклонение среднего значения *j*-ой группы от среднего значения всей совокупности: $(\mu_j - \mu)$.

Как правило, разложение общей дисперсии для выборочных данных, представляется в виде равенства сумм квадратов соответствующих отклонений:

$$SS_T = SS_B + SS_R \qquad (3.2.6.6)$$

где k – число уровней фактора,

n_j – число наблюдений для *j*-го уровня фактора,

$n = \sum_{j=1}^{k} n_j$ - общее число наблюдений.

Общая сумма квадратов отклонений (3.2.7):

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \qquad (3.2.7)$$

Сумма квадратов отклонений групповых средних от общего среднего (3.2.8):

$$SS_B = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 \qquad (3.2.8)$$

Выражение 3.2.8 также называется эффектом фактора (суммой квадратов эффекта)

Внутригрупповая (остаточная) сумма квадратов отклонений определяется по формуле (3.29):

$$SS_R = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \qquad (3.2.9)$$

Выражение 3.2.9 также называется остаточным эффектом (эффектом ошибок)

Таким образом, в разложении дисперсии на составляющие заключена основная идея дисперсионного анализа: общая вариация переменной, обусловленная влиянием фактора и измеренная суммой SS_T, складывается из двух компонент: SS_B и SS_R, которые используются для описания изменчивости переменной между уровнями фактора (SS_B) и внутри уровней фактора (SS_R).

При осуществлении дисперсионного анализа анализируются не сами значения суммы квадратов отклонений, а средние квадраты. Они получаются путем деления сумм квадратов отклонений на соответствующее число степеней свободы [14, 87, 108].

Число степеней свободы для суммы квадратов случайных величин определяется как общее число линейно независимых слагаемых.

Для полной суммы квадратов $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$ число степеней свободы $\nu_T = n - 1$, так как при ее расчете используются n наблюдений, которые связаны между собой единственным уравнением для общего выборочного среднего всей совокупности.

Для суммы квадратов эффекта фактора $SS_B = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$ число степеней свободы $\nu_B = k - 1$, так как при ее расчете используются k

групповых средних, связанных между собой также одним уравнением для общего выборочного среднего всей совокупности.

Для суммы квадратов ошибок $SS_R = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$ число степеней свободы $\nu_R = n - k$, ибо при его расчете используются n наблюдений, связанных между собой k уравнениями для выборочных средних k групп.

Соответственно выражения для средних квадратов отклонений, которые являются оценками соответствующих дисперсий, имеют вид (3.2.10-3.2.12).

Оценка общей дисперсии:

$$\mathrm{MS_T} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{n_i}}(\mathrm{y_{ij}} - \bar{\mathrm{y}})^2 \qquad (3.2.10)$$

Оценка межгрупповой дисперсии (3.2.11):

$$\mathrm{MS_B} = \frac{1}{\mathrm{k}-1}\sum_{\mathrm{i}=1}^{\mathrm{k}}\mathrm{n_i}(\bar{\mathrm{y}}_\mathrm{i} - \bar{\mathrm{y}})^2 \qquad (3.2.11)$$

Оценка остаточной дисперсии (3.2.12):

$$\mathrm{MS_R} = \frac{1}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{n_i}}(\mathrm{y_{ij}} - \bar{\mathrm{y}}_\mathrm{i})^2 \qquad (3.2.12)$$

При условии истинности нулевой гипотезы $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, статистики MS_B и MS_R, являются несмещенными оценками одной и той же дисперсии σ^2.

В таких условиях сущность проверки нулевой гипотезы состоит в анализе равенства дисперсий на основе F-отношения (3.2.13):

$$\mathrm{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_R}} = \frac{\mathrm{n}-\mathrm{k}}{\mathrm{k}-1}\frac{\mathrm{SS_B}}{\mathrm{SS_R}} \qquad (3.2.13)$$

Если нулевая гипотеза верна, статистика F в случае нормального распределения величин ε_{ij} обладает распределением Фишера с $\nu_1 = k-1$ и $\nu_2 = n-k$ числом степеней свободы.

В случае, если наблюдаемое значение больше или равно критического значения распределения Фишера с уровнем α и с числом степеней свободы $\nu_1 = k-1$ и $\nu_2 = n-k$ (то есть $F_{набл} \geq F_{кр}$), нулевая гипотеза отклоняется и считается, что средние для различных уровней фактора значимо различаются.

Если применяются порядковые данные, непараметрической альтернативой однофакторного дисперсионного анализа являются ранговый дисперсионный анализ Краскела–Уоллиса [18].

Его основой является однофакторный дисперсионный анализ, однако вместо исходных значений переменных анализируются их ранговые характеристики.

Если обозначить через R_{ij} ранг элемента x_{ij}, в общем вариационном ряду значений отклика, то величины $\overline{\mathrm{R}}_{\mathrm{j}} = \frac{1}{\mathrm{n}_{\mathrm{j}}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} \mathrm{R}_{\mathrm{ij}}$ будут определять средние ранги для элементов j-ой группы, а величина $\overline{\mathrm{R}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{R}_{\mathrm{ij}} = \frac{\mathrm{n}+1}{2}$ средний ранг всей совокупности. Таким образом, величина $\sum_{j=1}^{k} n_j (\overline{R}_j - \overline{R})^2$ будет использоваться для характеристики межгруппового разброса рангов.

Если нулевая гипотеза о равенстве средних рангов верна, статистика примет вид (3.2.14):

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j (\overline{R}_j - \overline{R})^2 \qquad (3.2.14)$$

Приближенно она будет равна распределению Хи-квадрат с $k-1$ степенью свободы.

В случае, если соблюдается неравенство $\mathrm{H}_{\text{набл}} \geq \mathrm{H}_{\text{кр}}$, нулевая гипотеза отклоняется и считается, что средние ранги для различных уровней фактора значимо различаются.

Пример

Провести однофакторный дисперсионный анализ по приведенным ниже данным.

В таблице 3.2.1 приведены данные по испытаниям. Из четырёх партий изделий было выбрано по пять штук и проведено испытание на разрушение.

Таблица 3.2.1

Исходные данные для анализа

Партия изделий	Нагрузка разрушения				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

По итогам однофакторного дисперсионного анализа необходимо определить является ли влияние различных партий изделий на нагрузку разрушения существенным при уровне α=0,01.

Решение

1. Рассчитаем средние арифметические величины по каждой из партий:

$$\overline{x_1} = \frac{200 + 140 + 170 + 145 + 165}{5} = 164$$

$$\overline{x_2} = \frac{190 + 150 + 210 + 150 + 150}{5} = 170$$

$$\overline{x_3} = \frac{230 + 190 + 200 + 190 + 200}{5} = 202$$

$$\overline{x_4} = \frac{150 + 170 + 150 + 170 + 180}{5} = 164$$

2. Рассчитаем среднюю арифметическую всей совокупности:

$$\bar{x} = \frac{3500}{20} = 175$$

3. Определим сумму квадратов отклонений между группами SS_1 с $k_1=m-1$ степенями свободы:

$$SS_1 = 4980$$

$$k_1=m-1=4-1=3$$

4. Определим сумму квадратов отклонений внутри группы SS_2 с $k_2=mn-m$ степенями свободы:

$$SS_2 = 7270$$

$$k_1= mn-m =20-4=16$$

5. Полную сумму квадратов SS с $k_1= mn -1$ степенями свободы:

$$SS = 12250$$

$$k_1= mn -1=20-1=19$$

6. Заполним расчетную таблицу

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсий
Межгрупповая	4980	3	1660,0
Внутригрупповая	7270	16	454,4
Полная	1250	19	644,7

7. Рассчитываем F-критерий:

$$F = \frac{4980 \times \frac{1}{3}}{7270 \times \frac{1}{16}} = 3{,}65$$

Сравниваем полученной значение критерия с табличным (5,29). Полученное в результате расчетов значение критерия меньше табличного (критического). Из этого следует вывод, что нулевая гипотеза не может быть отвергнута.

Соответственно делаем вывод по задаче, что различие между деталями в партиях не влияет на величину нагрузки разрушения.

3.3. Однофакторный дисперсионный анализ в MS Excel

Рассмотрим решение задачи проведения однофакторного дисперсионного анализа в MS Excel.

Итак, необходимо провести однофакторный дисперсионный анализ по приведенным ниже данным.

В таблице 3.3.1 приведены данные по испытаниям. Из четырёх партий изделий было выбрано по пять штук и проведено испытание на разрушение.

Таблица 3.3.1

Исходные данные для анализа

Партия изделий	Нагрузка разрушения				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

По итогам однофакторного дисперсионного анализа необходимо определить является ли влияние различных партий изделий на нагрузку разрушения существенным при уровне α=0,01.

Решение

Для начала решения задачи необходимо выполнить послед стельность Данные – Анализ данных – Однофакторный дисперсионный анализ (рис. 3.3.1).

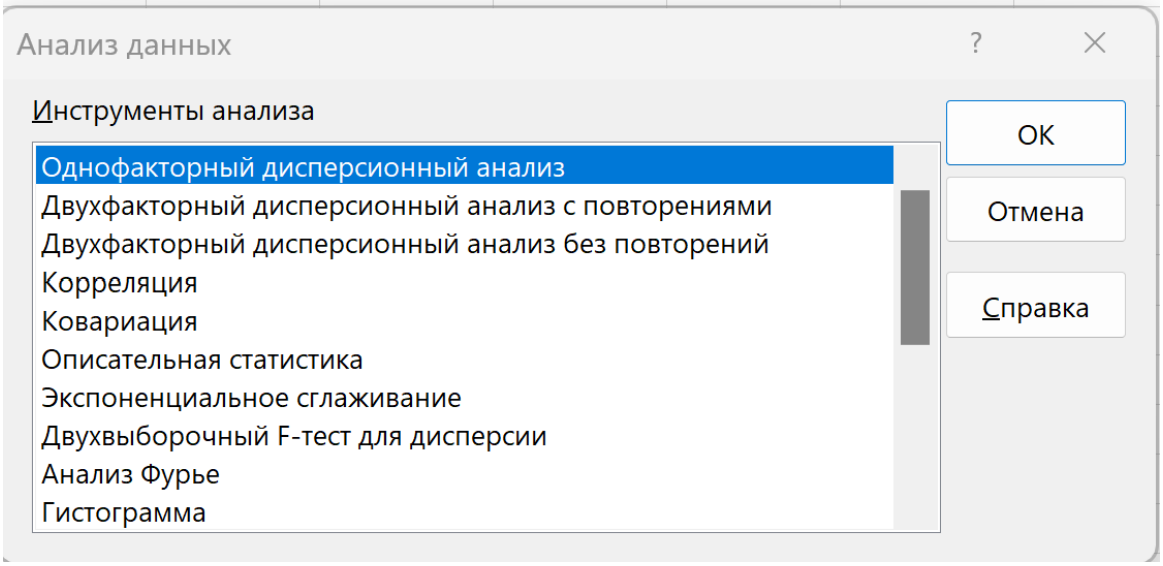


Рис. 3.3.1. Меню однофакторного дисперсионного анализа

После нажатия ОК появится меню дисперсионного анализа (рис. 3.3.2).

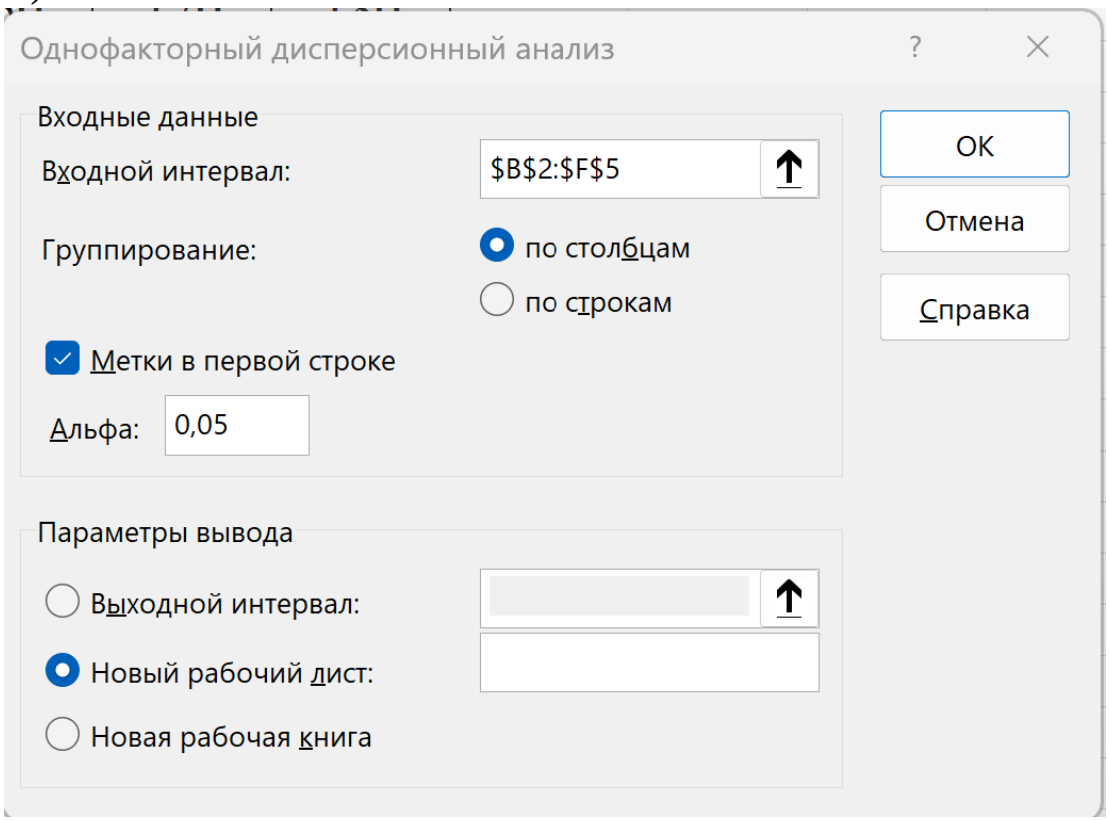


Рис. 3.3.2. Параметры однофакторного дисперсионного анализа

В нем необходимо задать входной интервал данных. В нашем случае это $B$2:$F$5. Отметим галочкой **Метки в первом столбце** (по умолчанию данная функция отключена). При необходимости уровень

альфы можно изменить, введя необходимую величину, которая по умолчанию составляет 0,05. После необходимо нажать **ОК**.

Результаты будут выведены в виде двух таблиц Рис. 3.2.3):

- **Итоги**. В данной таблице представлены промежуточные данные расчетов для каждой группы: число образцов (счет), суммы величин нагрузки разрушения (сумма), среднее арифметическое величин нагрузки разрушения (среднее), дисперсия величин нагрузки разрушения (дисперсия);

- **Дисперсионный анализ**. В данной таблице представлены собственно результаты дисперсионного анализа: компоненты дисперсии (источник вариации), суммы квадратов (*SS*), число степеней свободы (*df*), средний квадрат (*MS*), статистика F-критерия (*F*), вероятность значимости (p-значение), статистика Fкр (F-критическое).

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
200	3	570	190	1600		
140	3	510	170	400		
170	3	560	186,6666667	1033,333333		
145	3	510	170	400		
165	3	530	176,6666667	633,3333333		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	1040	4	260	0,319672131	0,858494084	3,478049691
Внутри групп	8133,333	10	813,3333333			
Итого	9173,333	14				

Рис. 3.3.3. Результаты однофакторного дисперсионного анализа

Полученное в результате расчетов значение критерия меньше табличного (критического). Из этого следует вывод, что нулевая гипотеза не может быть отвергнута.

Соответственно делаем вывод по задаче, что различие между деталями в партиях не влияет на величину нагрузки разрушения.

3.4. Однофакторный дисперсионный анализ в Statistica

Рассмотрим возможности программного комплекса на примере.

В таблице 3.4.1 приведены данные по испытаниям. Из четырёх партий изделий было выбрано по пять штук и проведено испытание на разрушение.

Таблица 3.4.1

Исходные данные для анализа

Партия изделий	Нагрузка разрушения				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

По итогам однофакторного дисперсионного анализа необходимо определить является ли влияние различных партий изделий на нагрузку разрушения существенным при уровне α=0,01.

Решение

Необходимо сформировать данные в рабочей области программы, выделив партии в отдельный столбец (рис. 3.4.1).

	1 Пер1	2 Пер2
1	200	1
2	140	1
3	170	1
4	145	1
5	165	1
6	190	2
7	150	2
8	210	2
9	150	2
10	150	2
11	230	3
12	190	3
13	200	3
14	190	3
15	200	3
16	150	4
17	170	4
18	150	4
19	170	4
20	180	4

Рис. 3.4.1. Исходные данные для анализа

Далее необходимо перейти **Анализ** – **Дисперсионный анализ**. Откроется меню дисперсионного анализа (рис. 3.4.2).

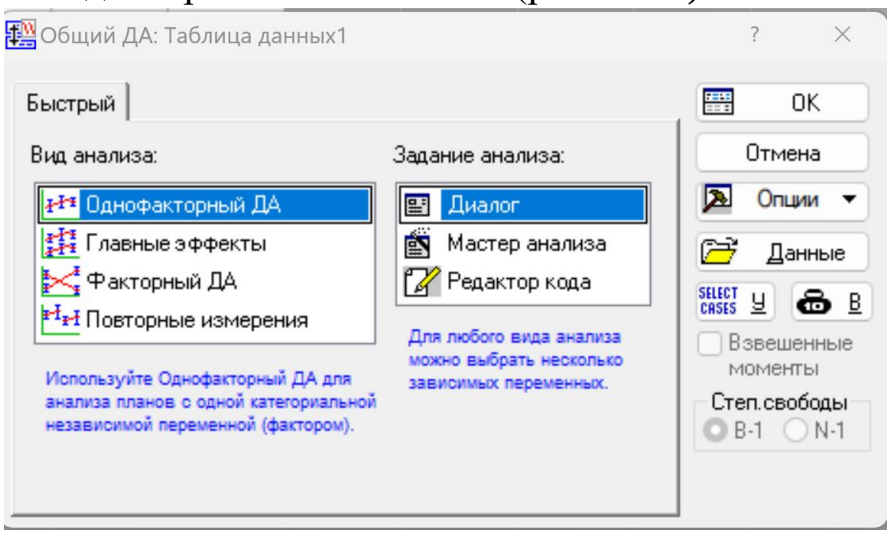


Рис. 3.4.2. Выбор вида дисперсионного анализа

В открывшемся окне необходимо выбрать **Однофакторный дисперсионный анализ** и нажать **ОК**. В результате появится меню итогов анализа (рис. 3.4.3).

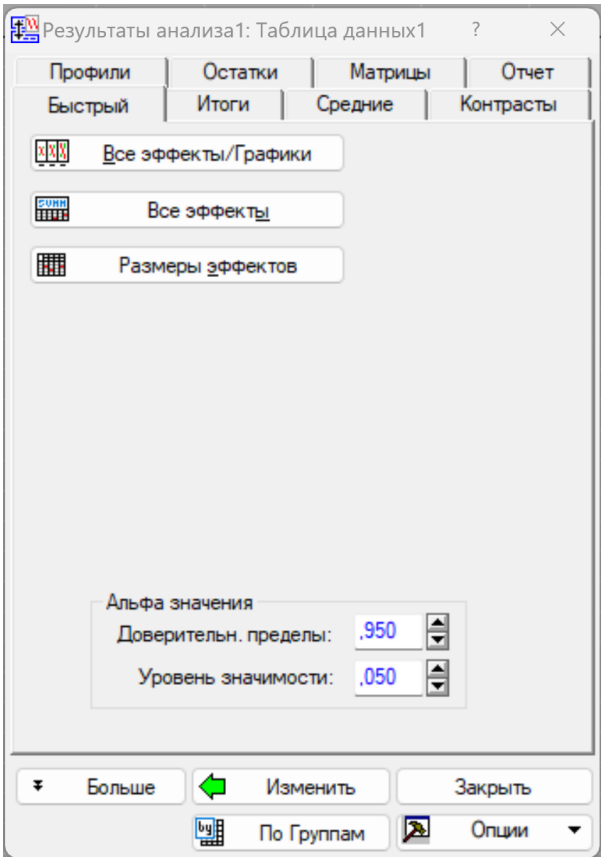


Рис. 3.4.3. Результаты дисперсионного анализа

На вкладке **Быстрый** есть ряд кнопок:

- кнопка **Все эффекты/Графики** (рис. 3.4.3) – выводит форму эффектов и предоставляет выбор ряда параметров: по отображению результатов – в табличной или графической формах, по типу средних – взвешенные, невзвешенные или наименьшие квадраты, по представлению стандартных ошибок – в абсолютном или интервальном виде. При выборе графической формы (рис. 3.4.4) выводится декомпозиция гипотезы. При выборе табличной формы – соответственно таблица с итогами анализа

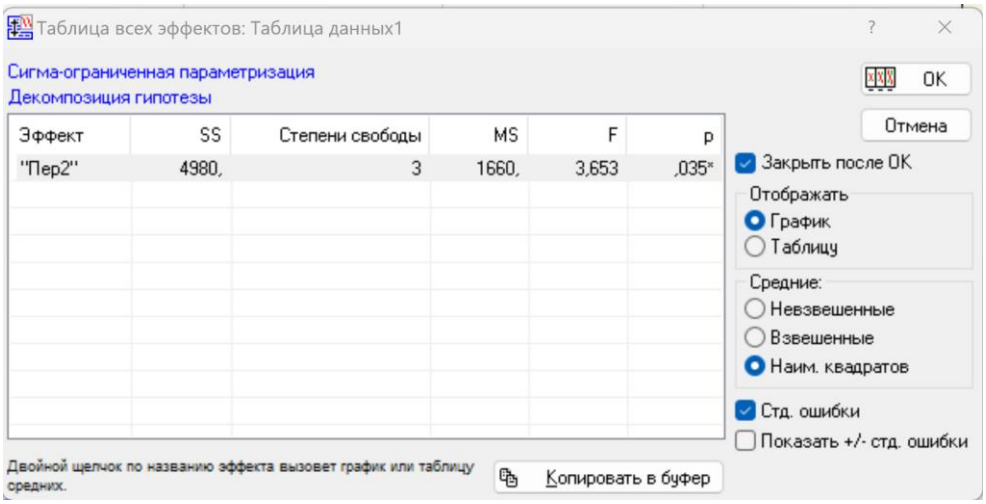


Рис. 3.4.4. Меню **Все эффекты/Графики**

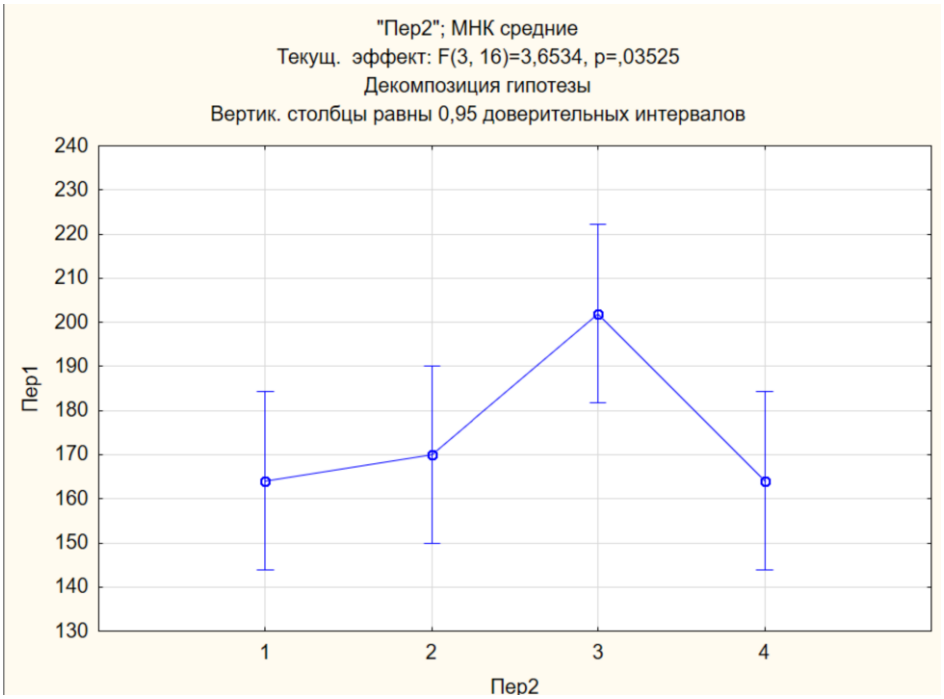


Рис. 3.4.5. **Декомпозиция гипотезы**

N ячейки	"Пер2"; МНК средние (Таблица данных1) Текущ. эффект: F(3, 16)=3,6534, p=,03525 Декомпозиция гипотезы **Пер2**	Пер1 Среднее	Пер1 Стд.ош.	Пер1 -95,00%	Пер1 +95,00%	N
1	1	164,0000	9,532838	143,7913	184,2087	5
2	2	170,0000	9,532838	149,7913	190,2087	5
3	3	202,0000	9,532838	181,7913	222,2087	5
4	4	164,0000	9,532838	143,7913	184,2087	5

Рис. 3.46. Таблица итогов дисперсионного анализа

- кнопка **Все эффекты** (рис. 3.4.7) - в качестве результата в окно выводится таблица. Результаты, приведенные в строке «Пер2», относятся к данным, характеризующим влияние фактора (систематические характеристики), а в строке «Ошибка» – к неучтенным воздействиям (остаточные характеристики). В итоговой таблице использованы следующие обозначения: SS – сумма квадратов, здесь приведены величины вариаций (SS, Season обусловливается межгрупповой изменчивостью, SS, Error – внутригрупповой изменчивостью); Степени свободы – количество степеней свободы; MS – средний квадрат, здесь приведены соответствующие дисперсии; F – наблюдаемое в рассматриваемой задаче значение F- статистики; p – минимальный уровень значимости указанной F-статистики.

Эффект	Одномерный критерий значимости дляПер1 (Таблица данных1) Сигма-ограниченная параметризация Декомпозиция гипотезы **SS**	Степени свободы	MS	F	p
Св. член	612500,0	1	612500,0	1348,006	0,000000
"Пер2"	4980,0	3	1660,0	3,653	0,035248
Ошибка	7270,0	16	454,4		

Рис. 3.4.7. Таблица итогов дисперсионного анализа **Все эффекты**

- Вкладка **Размеры эффектов** (рис. 3.4.8) – в табличной форме выводятся итоги дисперсионного анализа.

Эффект	SS	Степени свободы	MS	F	p	Частичная эта-квадрат.	Нецентрированность.	Наблюдаемая мощность. (альфа=0,05)
Одномерные критерии значимости, размеров эффекта и мощностидляПер1 (Таблица данных1) Сигма-ограниченная параметризация Декомпозиция гипотезы								
Св. член	612500,0	1	612500,0	1348,006	0,000000	0,988270	1348,006	1,000000
"Пер2"	4980,0	3	1660,0	3,653	0,035248	0,406531	10,960	0,688416
Ошибка	7270,0	16	454,4					

Рис. 3.4.8. Таблица итогов дисперсионного анализа **Размеры эффектов**

Остальные вкладки предоставляют расширенные возможности графического и табличного вариантов представления результатов анализа.

3.5. Многофакторный дисперсионный анализ

Если производится анализ двух и более различных факторов на результаты наблюдений, имеет место **многофакторный дисперсионный анализ**.

Примером двухфакторной модели в статистике фирмы может служить следующая модель: например, двухфакторная модель будет применяться при построении объяснения различий в доходах фирм, обусловленных как месторасположением организаций, так и спецификой их вида деятельности [14, 23, 56].

Рассмотрим подробнее влияние на величину X фактора A, имеющего *k* уровней, и фактора B, имеющего m уровней.

Построение двухфакторной модели опирается на выражение (3.5.15):

$$y_{ijl} = \mu_{ij} + \varepsilon_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}, \quad l = \overline{1, n_{ij}}, \quad i = \overline{1, k}, \quad j = \overline{1, m}, \qquad (3.5.15)$$

где: y_{ijl} - *l*-ое наблюдаемое значение отклика для *i*-го уровня фактора A и *j*-го уровня фактора B;

μ - среднее значение отклика по всей совокупности (генеральное среднее);

μ_{ij} - среднее значение отклика для *i*-го уровня фактора A и *j*-го уровня фактора B;

$\alpha_i = \mu_{i*} - \mu$ - главный эффект *i*-го уровня фактора A (μ_{i*} - среднее значение отклика для *i*-го уровня фактора A);

$\beta_j = \mu_{*j} - \mu$ - главный эффект *j*-го уровня фактора B (μ_{*j} - среднее значение отклика для *j*-го уровня фактора B);

$\gamma_{ij} = \mu_{ij} - \mu_{i*} - \mu_{*j} + \mu$ - эффект взаимодействия *i*-го уровня фактора A и *j*-го уровня фактора B;

ε_{ijl} - независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2.

Средние могут определяться через величины μ_{ij}, например, как взвешенные средние:

$$\mu = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}} M(y_{ijl}) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}\mu_{ij};$$

$$\mu_{i*} = \frac{1}{n_{i*}}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}} M(y_{ijl}) = \frac{1}{n_{i*}}\sum_{j=1}^{m} n_{ij}\mu_{ij};$$

$$\mu_{*j} = \frac{1}{n_{*j}}\sum_{i=1}^{k}\sum_{l=1}^{n_{ij}} M(y_{ijl}) = \frac{1}{n_{*j}}\sum_{i=1}^{k} n_{ij}\mu_{ij};$$

$$n = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij},$$

$$n_{i*} = \sum_{j=1}^{m} n_{ij},$$

$$n_{*j} = \sum_{i=1}^{k} n_{ij}.$$

Возможны и другие варианты определения средних величин.

Выражение $y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}$ можно представить в виде (3.5.16):

$$y_{ijl} - \mu = (\mu_{i*} - \mu) + (\mu_{*j} - \mu) + (\mu_{ij} - \mu_{i*} - \mu_{*j} + \mu) + (x_{ijl} - \mu_{ij}) \qquad (3.5.16)$$

Данное соотношение говорит о том, что отклонение наблюдаемого значения отклика складывается из суммы четырех слагаемых: отклонения отклика от среднего значения для *i, j*-го набора уровней факторов А и В $(y_{ijl} - \mu_{ij})$, главных эффектов *i*-го уровня фактора А и *j*-го уровня фактора В и эффекта взаимодействия.

Таким образом, дисперсия отклика может быть представлена в виде суммы четырех дисперсий, одна из которых характеризует внутригрупповую изменчивость для *i, j*-го набора уровней факторов А и В, а остальные соответствующие эффекты [72, 73].

НМК оценки параметров модели двухфакторного дисперсионного анализа достаточно просто получить только в случае пропорциональных частот, то есть при условии (3.5.17):

$$\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \cdot \frac{n_{*j}}{n} \quad \forall\, i = \overline{1,k},\ j = \overline{1,m}. \tag{3.5.17}$$

Данные условия выполняются, например, если количества наблюдений n_{ij} для каждого сочетания уровней факторов совпадают.

Если соблюдается данное условие, эксперимент принято считать сбалансированным [86].

В случае пропорциональных частот исследователь получает понятные МНК-оценки для параметров модели вне зависимости от условий, накладываемых на параметры модели:

$$\hat{\mu} = \bar{y},$$

$$\hat{\alpha}_i = \bar{y}_{i*} - \bar{y},$$

$$\hat{\beta}_j = \bar{y}_{*j} - \bar{y},$$

$$\hat{\gamma}_{ij} = \bar{y}_{ij} - (\bar{y}_{i*} + \bar{y}_{*j}) + \bar{y},$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{l=1}^{n_{ij}} y_{ijl},$$

$$\bar{y}_{i*} = \frac{1}{n_{i*}} \sum_{j=1}^{m} \sum_{l=1}^{n_{ij}} y_{ijl},$$

$$\bar{y}_{*j} = \frac{1}{n_{*j}} \sum_{i=1}^{k} \sum_{l=1}^{n_{ij}} y_{ijl},$$

$$\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} y_{ijl}$$

В случае пропорциональных также справедливо следующее разложение общей суммы квадратов отклонений на составляющие (3.5.18):

$$SS_T = SS_A + SS_B + SS_{AB} + SS_R, \qquad (3.5.18)$$

где:

$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{l=1}^{n_{ij}} (y_{ijl} - \bar{y})^2$ – общая, или полная, сумма квадратов отклонений;

$SS_A = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{l=1}^{n_{ij}} (\bar{y}_{i*} - \bar{y})^2 = \sum_{i=1}^{k} n_{i*} (\bar{y}_{i*} - \bar{y})^2$ – сумма квадратов отклонений средних по уровням фактора А от общего среднего, или сумма квадратов главных эффектов А (можно и так: сумма квадратов, соответствующих эффекту фактора А);

$SS_B = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{l=1}^{n_{ij}} (\bar{y}_{*j} - \bar{y})^2 = \sum_{j=1}^{m} n_{*j} (\bar{y}_{*j} - \bar{y})^2$ – сумма квадратов отклонений средних по уровням фактора В от общего среднего, или сумма квадратов главных эффектов В;

$SS_{AB} = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{l=1}^{n_{ij}} (\bar{y}_{ij} - \bar{y}_{*j} - \bar{y}_{i*} + \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} (\bar{y}_{ij} - \bar{y}_{*j} - \bar{y}_{i*} + \bar{y})^2$ – сумма квадратов взаимодействия эффектов А и В;

$SS_R = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(y_{ijl} - \bar{y}_{ij})^2$ – остаточная сумма квадратов отклонений.

Число степеней свободы для сумм квадратов SS_A и SS_B равно соответственно $\nu_A = k-1$ и $\nu_B = m-1$. Число степеней свободы для суммы квадратов взаимодействия эффектов SS_{AB} равно $\nu_{AB} = (\kappa-1)(m-1)$. Число степеней свободы суммы квадратов остатков SS_R равно $\nu_R = n-km$. Соответственно средние суммы квадратов будут равны:

$$MS_A = \frac{SS_A}{k-1},$$

$$MS_B = \frac{SS_B}{m-1},$$

$$MS_{AB} = \frac{SS_{AB}}{(k-1)(m-1)},$$

$$MS_R = \frac{SS_R}{n-km}.$$

Так как в рамках построения двухфакторной модели рассматриваются различные эффекты от влияния факторов, необходимо осуществлять проверку гипотез значимости различных выявленных эффектов с помощью инструментов статистического анализа.

При условии истинности H_0: «эффект незначим» средний квадрат эффекта является несмещенной оценкой дисперсии σ^2, так же, как и величина MS_R.

Поэтому в качестве статистик критериев проверки гипотез о значимости соответствующих эффектов можно использовать отношения средней суммы квадратов эффектов к средней сумме квадратов остатков [11].

При нормальном распределении остатков данные статистики имеют распределение Фишера с параметрами, определяемыми числами степеней свободы соответствующих сумм, участвующих в отношении.

В том случает, если наблюдаемое значение статистики $F_{набл} \geq F_{кр}$, где $F_{кр}$ - критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы ν_1 и ν_2, то нулевая гипотеза отклоняется и считается, что средние для различных уровней фактора значимо различаются.

В случае непропорциональных частот оценки МНК параметров модели, при заданных $k+m-1$ условиях на параметры модели, могут быть получены численно, однако в этом случае нарушается условие ортогональности сумм квадратов эффектов [68].

Соответственно, независимая одновременная проверка всех гипотез двухфакторного анализа не представляется возможной.

Однако всегда существует возможность проверки значимости любого из выявленных эффектов (либо линейной комбинации эффектов), рассматривая соответствующую однофакторную модель и выделив соответствующую сумму квадратов. Затем можно исключить данную сумму квадратов из общей суммы квадратов и проверить значимость другого эффекта и т.д.

Таким образом, происходит последовательное разложение суммы квадратов на составляющие [103].

Можно использовать и другой подход, выделив на основе однофакторной модели сумму квадратов всех эффектов, за исключением одного, а затем уже исследовать значимость этого эффекта на основе оставшейся суммы квадратов.

В зависимости от того, как формируется разложение суммы квадратов на составляющие, различаются и статистики критериев для проверки гипотез о значимости эффектов в несбалансированной многофакторной модели дисперсионного анализа, при этом общий принцип формирования статистики критерия на основе *F*-отношения сумм квадратов эффекта и остатков остается прежним [77].

В многофакторной несбалансированной модели принято рассматривать 3 основных типа разложения суммы квадратов.

При этом в анализе будут рассматриваться следующие обозначения.

$R(\mu)$ – остаточная сумма квадратов для модели содержащий только параметр μ - общее среднее;

$R(\mu, A)$ – остаточная сумма квадратов для модели учитывающей эффект фактора А, т.е. для модели, содержащей параметр μ и параметры α_i, $i = \overline{1,k}$;

$R(\mu, B)$ – остаточная сумма квадратов для модели учитывающей эффект фактора В, т.е. для модели, содержащей параметр μ и параметры β_j, $j = \overline{1,m}$;

$R(\mu, A, B)$ – остаточная сумма квадратов для модели учитывающей эффекты факторов А и В, без учета взаимодействия факторов;

$R(\mu, A, B, AB)$ – остаточная сумма квадратов для полной модели, учитывающей как эффекты факторов А и В, так и эффект взаимодействия факторов АВ.

Следует отметить, что записи вида: $R(\mu, A)$, $R(\mu, A, B)$, $R(\mu, A, B, AB)$ однозначно характеризуют не только остаточную сумму квадратов, но и саму модель, поэтому в дальнейшем под этой записью будем понимать, как остаточную сумму квадратов, так и модель, в зависимости от контекста.

Под записью $SS(A|B)$, будем понимать сумму квадратов, соответствующих эффекту фактору А, после того как из разложения была удалена сумма квадратов, соответствующая эффекту фактора В.

Тип I называют также последовательной суммой квадратов. Разложение зависит от порядка эффектов в модели. Каждый последующий эффект скорректирован на предыдущие эффекты, эффекты взаимодействия оцениваются после эффектов факторов. Разложение является аддитивным по отношению к общей сумме квадратов.

Пример формирования сумм квадратов на примере двухфакторной модели (в качестве первого фактора выбирается фактор А) приведен ниже [49].

Суммы квадратов I типа, для эффектов модели, где в качестве первого фактора выбран А будут иметь вид:

$$\text{эффект А:} \quad SS(A) = R(\mu) - R(\mu, A);$$

эффект B: $SS(B \mid A) = R(\mu, A) - R(\mu, A, B)$;

эффект AB: $SS(AB \mid A, B) = R(\mu, A, B) - R(\mu, A, B, AB)$.

Все суммы определяются однозначно, вне зависимости от условий, накладываемых на параметры модели. Таким образом, чтобы получить данные суммы потребуется построить 4 различные модели (хотя можно сократить число моделей, используя для построения сумм соответствующие оценочные функции).

В **типе II** суммы квадратов каждого эффекта в модели, корректируются по всем остальным "подходящим" эффектам, то есть, вычисляются после удаления из общей суммы квадратов сумм квадратов "подходящих" эффектов. Под "подходящим" понимается любой эффект, который не содержит исследуемый эффект. Разложение не зависит от порядка эффектов в модели и не является аддитивным по отношению к общей сумме квадратов [98].

Суммы квадратов II типа, для эффектов двухфакторной модели будут иметь вид:

эффект A: $SS(A \mid B) = R(\mu, B) - R(\mu, A, B)$;

эффект B: $SS(B \mid A) = R(\mu, A) - R(\mu, A, B)$;

эффект AB: $SS(AB \mid A, B) = R(\mu, A, B) - R(\mu, A, B, AB)$.

Все суммы определяются однозначно, вне зависимости от условий, накладываемых на параметры модели. Также, как и в предыдущем случае, чтобы получить данные суммы потребуется построить 4 различные модели.

При использовании типа III суммы квадратов каждого эффекта в модели, корректируются по всем остальным эффектам и являются ортогональными к суммам квадратов, содержащим исследуемый эффект. Разложение не зависит от порядка эффектов в модели и не является аддитивным по отношению к общей сумме квадратов.

Суммы квадратов **типа III**, для эффектов двухфакторной модели будут иметь вид:

эффект A: $SS(A \mid B, AB) = R(\mu, B, AB) - R(\mu, A, B, AB)$;

эффект B: $SS(B \mid A, AB) = R(\mu, A, AB) - R(\mu, A, B, AB)$;

эффект AB: $SS(AB \mid A, B) = R(\mu, A, B) - R(\mu, A, B, AB)$.

Суммы квадратов $R(\mu, A, AB)$ и $R(\mu, B, AB)$ зависят от условий, накладываемых на параметры модели.

Для выполнения условий ортогональности, параметры модели должны удовлетворять условиям:

$$\sum_{i=1}^{k} \alpha_i = 0,$$

$$\sum_{j=1}^{m} \beta_j = 0,$$

$$\sum_{i=1}^{k} \gamma_{ij} = 0,\ j = \overline{1, m-1}$$

$$\sum_{j=1}^{m} \gamma_{ij} = 0,\ i = \overline{1, k-1}$$

$$\sum_{i=1}^{k} \sum_{j=1}^{m} \gamma_{ij} = 0,\ i = \overline{1, k},\ j = \overline{1, m}.$$

Как и при использовании двух других типов, получение данных суммы должно опираться на построение 4х различных моделей. Также, как и в предыдущих случаях, чтобы получить данные суммы, потребуется построить 4 различные модели.

3.6. Дисперсионный анализ для повторных наблюдений

На практике часто исследователи сталкиваются с проблемой соответствия различным уровням фактора одних и тех же объектов.

Метод дисперсионного анализа может быть применен для анализа чистой прибыли фирм за различные периоды времени. Однако, сделать вывод о том, что исходные данные не являются взаимокоррелируемыми величинами однозначно нельзя [16].

В такой ситуации анализ эффектов фактора опирается на дисперсионный анализ, предполагающий исключение влияния зависимостей выборок, которые связаны с повторным рассмотрением одних и тех же объектов.

Данный метод - дисперсионный анализ зависимых выборок (повторных наблюдений). С его помощью возможно уменьшение общей дисперсии данных за счет исключения составляющей индивидуальных различий из остаточной дисперсии. Таким образом, повышается мощность критерия, что оказывает положительное влияние на модель [69].

Если имеется k-выборок, объем каждой из которых равен n, соответствующих k различным уровням фактора *W*, влияние которого на значения наблюдаемой переменной рассматривается.

Предполагается, что *j*-ое наблюдаемое значение отклика в каждой группе соответствует одному и тому же объекту наблюдения. В простейшей однофакторной модели дисперсионного анализа повторных измерений исходят из следующей модели порождения данных (3.6.1):

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu + w_i + \tau_j + \varepsilon_{ij} \quad i = \overline{1,k}, \quad j = \overline{1,n}, \qquad (3.6.1)$$

где: y_{ij} - *j*-ое наблюдаемое значение отклика в *i*-ой группе;

μ_{ij} - среднее значение отклика для *i*-го наблюдения в *j*-ой группе;

ε_{ij} - независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2;

μ - среднее значение отклика;

$w_i = \mu_{i*} - \mu$ - эффект *i*-го уровня фактора *W*,

$\mu_{i*} = \frac{1}{n}\sum_{j=1}^{n} \mu_{ij}$ - среднее для *i*-го уровня фактора;

$\tau_j = \mu_{*j} - \mu$ - эффект индивидуальности *j*-го наблюдения,

$\mu_{*j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}$ - среднее для *j*-го объекта наблюдения;

ε_{ij} - независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2.

Уравнение (3.6.1) можно рассматривать как соответствующее классической модели двухфакторного дисперсионного анализа, без компоненты, учитывающей взаимодействие факторов и числом наблюдений для каждой ячейки $n_{ij} = 1$ (что соответствует сбалансированному плану) [35].

Таким образом, его можно записать в виде (3.6.2).

$$y_{ij} - \mu_{*j} = (\mu_{i*} - \mu) + \varepsilon_{ij} = w_i + \varepsilon_{ij}, \quad i = \overline{1,k}, \quad j = \overline{1,n}. \qquad (3.6.2)$$

Для того, чтобы исключить эффект индивидуальности наблюдений и анализировать только влияние фактора *W* на результаты наблюдений, мы должны перейти к рассмотрению величин $y_{ij} - \mu_{*j}$. МНК оценкой параметра μ_{*j} является выборочное среднее $\bar{y}_{*j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij}$. Соответственно, сумма квадратов наблюдений за вычетом суммы квадратов, соответствующих индивидуальным различиям, будет равна $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{*j})^2$.

Если обозначить $z_{ij} = y_{ij} - \bar{y}_{*j}$, то разложение для суммы квадратов $\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^2$ будет иметь вид (3.6.3):

$$\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^2 = \sum_{i=1}^{k} n\bar{z}_i^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij} - \bar{z}_i)^2. \qquad (3.6.3)$$

Число степеней свободы для суммы $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^2$ равно $nk - n$, для суммы $SS_W = \sum_{i=1}^{k} n(\bar{z}_i - \bar{z})^2 = \sum_{i=1}^{k} n\bar{z}_i^2$ равно $k-1$, для суммы $SS_R = \sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij} - \bar{z}_i)^2$ равно $nk - n - k + 1 = (n-1)(k-1)$.

Соответственно, средние суммы квадратов отклонений будут равны (3.6.4 и 3.6.5):

$$MS_W = \frac{SS_W}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k} n\bar{z}_i^2 \qquad (3.6.4)$$

$$MS_R = \frac{SS_R}{(n-1)(k-1)} = \frac{1}{(n-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij} - \bar{z}_i)^2 \quad (3.6.5)$$

F-статистика для проверки гипотезы $H_0 : w_1 = w_2 = \ldots = w_k = 0$ будет иметь вид: $F = \frac{MS_W}{MS_R}$. Если наблюдаемое значение статистики $F_{набл} \geq F_{кр}$, где $F_{кр}$ - критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы $\nu_1 = k-1$ и $\nu_2 = (n-1)(k-1)$, то нулевая гипотеза отклоняется и считается, что средние для различных уровней фактора *W* значимо различаются.

Данный критерий можно получить и как критерий для проверки соответствующей гипотезы в модели двухфакторного дисперсионного анализа.

Для модели, справедливо следующей разложение для суммы квадратов отклонений:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} n(\bar{y}_{i*} - \bar{y})^2 + \sum_{j=1}^{n} k(\bar{y}_{*j} - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{*j} - \bar{y}_{i*} + \bar{y})^2$$

При данном разложении сумма квадратов, которая раньше в двухфакторной модели являлась суммой квадратов взаимодействий, стала остаточной суммой квадратов отклонений, а сумма квадратов,

которая была остаточной суммой в двухфакторной модели, обратилась в ноль.

Сумма квадратов $\sum_{i=1}^{k} n(\bar{y}_{i*} - \bar{y})^2$ соответствует эффекту фактора *W*, а сумма квадратов $\sum_{j=1}^{n} k(\bar{y}_{*j} - \bar{y})^2$ соответствует эффекту индивидуальных различий. Таким образом, имеем (3.6.6 и 3.6.7):

$$SS_W = \sum_{i=1}^{k} n(\bar{y}_{i*} - \bar{y})^2 , \ \nu_W = k - 1 \qquad (3.6.6)$$

$$SS_R = \sum_{i=1}^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{*j} - \bar{y}_{i*} + \bar{y})^2 , \ \nu_W = (n-1)(k-1) \qquad (3.6.7)$$

Таким образом, для расчета статистики можно не осуществлять переход к центрированным по наблюдениям данным, однако, необходимо грамотно определить вид рассчитываемой остаточной суммы квадратов.

Если предположить, что дополнительно все n объектов разбиты на *m* групп по p_j объектов в каждой, $j = \overline{1,m}$ ($\sum_{j=1}^{m} p_j = n$), соответствующих различным уровням фактора *A*, то на результаты наблюдений могут оказывать влияние фактор *W* повторных наблюдений, фактор *A*, воздействующий на слои объектов, и фактор индивидуальных различий каждого объекта. Соответственно, модель порождения данных принимает вид (3.6.8):

$$y_{ijl} = \mu_{ijl} + \varepsilon_{ijl} = \mu + w_i + \alpha_j + \tau_{jl} + \beta_{ij} + \varepsilon_{ijl} \quad i = \overline{1,k}, \quad j = \overline{1,m}, \quad l = \overline{1,p_j} \qquad (3.6.8)$$

где: y_{ijl} - *l*-ое наблюдаемое значение отклика в *i*-ой группе в *j*-ом слое;

$\mu_{ijl} = \mu + w_i + \alpha_j + \tau_{jl} + \beta_{ij}$ - среднее для *l*-го наблюдаемого значения отклика в *i*-ой группе в *j*-ом слое;

μ - среднее значение отклика;

w_i - эффект i-го уровня фактора W;

α_j - эффект j-го уровня фактора A;

τ_{jl} - эффект индивидуальности l-го наблюдения в j-ом слое;

β_{ij} - эффект взаимодействия i-го уровня фактора W и j-го уровня фактора A;

ε_{ijl} - независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2.

Для того, чтобы существовала возможность однозначно определять параметры модели, необходимо на параметры модели наложить дополнительные условия. В качестве этих условий примем:

$$\sum_{i=1}^{k} w_i = 0,$$

$$\sum_{j=1}^{m} \alpha_j = 0,$$

$$\sum_{i=1}^{k} \beta_{ij} = 0,\ j = \overline{1, m-1},$$

$$\sum_{j=1}^{m} \beta_{ij} = 0,\ i = \overline{1, k-1},\ \sum_{i=1}^{k}\sum_{j=1}^{m} \beta_{ij} = 0.$$

Обозначим: $\mu_{*jl} = \frac{1}{k}\sum_{i=1}^{k} \mu_{ijl} = \mu + \alpha_j + \tau_{jl}$ - среднее для l-го объекта наблюдения в j-ом слое. Таким образом, уравнение примет вид (3.6.9):

$$y_{ijl} - \mu_{*jl} = w_i + \beta_{ij} + \varepsilon_{ijl} \quad i = \overline{1, k}, \quad j = \overline{1, n}, \quad l = \overline{1, p_j}. \qquad (3.6.9)$$

Таким образом, на величину $y_{ijl} - \mu_{*jl}$ могут оказывать влияние только главный эффект фактора W и эффект взаимодействия факторов W и A. Соответствующее разложение суммы квадратов в случае равного числа наблюдений p_j для каждого слоя имеет вид (3.6.10):

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{p_j}(y_{ijl}-\bar{y}_{*jl})^2=\sum_{i=1}^{k}n(\bar{y}_{i**}-\bar{y})^2+\sum_{i=1}^{k}\sum_{j=1}^{m}p_j(\bar{y}_{ij*}-\bar{y}_{*j*}-\bar{y}_{i**}+\bar{y})^2+$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{p_j}(y_{ijl}-\bar{y}_{*jl}-(\bar{y}_{ij*}-\bar{y}_{*j*}))^2 \qquad (3.6.10)$$

Сумма квадратов $SS_W=\sum_{i=1}^{k}n(\bar{y}_{i**}-\bar{y})^2$ соответствует эффекту фактора W, сумма квадратов $SS_{WA}=\sum_{i=1}^{k}\sum_{j=1}^{m}p_j(\bar{y}_{ij*}-\bar{y}_{*j*}-\bar{y}_{i**}+\bar{y})^2$ соответствует эффекту взаимодействия факторов W и A, а сумма квадратов $SS_R=\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{p_j}(y_{ijl}-\bar{y}_{*jl}-(\bar{y}_{ij*}-\bar{y}_{*j*}))^2$ является остаточной суммой квадратов. Числа степеней свободы соответствующих сумм квадратов равны:

$$\nu_W=k-1;$$

$$\nu_{WA}=(k-1)(m-1);$$

$$\nu_R=kn-(n-m)-(km-m)-m=(k-1)(n-m)$$

(если заданы средние $\bar{y}_{*j*},\quad j=\overline{1,m}$, то независимо можно определить только $km-m$ величин $\bar{y}_{ij*}$ и $n-m$ величин $\bar{y}_{*jl}$),

$$\nu_W=k-1,$$

$$\nu_{WA}=(k-1)(m-1).$$

Соответственно, средние суммы квадратов отклонений будут равны (3.6.11):

$$MS_W=\frac{SS_W}{k-1},\ MS_{WA}=\frac{SS_{WA}}{(k-1)(m-1)},\ MS_R=\frac{SS_R}{(k-1)(n-m)} \qquad (3.6.11)$$

F-статистика для проверки гипотезы $H_0 : w_1 = w_2 = \ldots = w_k = 0$ будет иметь вид: $F = \frac{MS_W}{MS_R}$. Если наблюдаемое значение статистики $F_{набл} \geq F_{кр}$, где $F_{кр}$ - критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы $\nu_1 = k-1$ и $\nu_2 = m(k-1)(p-1)$, то нулевая гипотеза отклоняется и считается, что эффект фактора *W* значим.

F-статистика для проверки гипотезы $H_0 : \beta_{11} = \beta_{12} = \ldots = \beta_{km} = 0$ соответственно будет иметь вид: $F = \frac{MS_{WA}}{MS_R}$. Если наблюдаемое значение статистики $F_{набл} \geq F_{кр}$, где $F_{кр}$ - критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы $\nu_1 = (k-1)(m-1)$ и $\nu_2 = (k-1)(n-m)$, то нулевая гипотеза отклоняется и считается, что эффект взаимодействия факторов *W* и *A* значим.

На основе разложения невозможно построить критерий для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$. Для этого необходимо использовать разложение суммы квадратов отклонений. То есть, для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ используем разложение, которое соответствует уравнениям (3.6.12):

$$\mu_{*jl} - \mu = \alpha_j + \tau_{jl} \quad i = \overline{1,k}, \quad j = \overline{1,n}, \quad l = \overline{1,p_j}. \qquad (3.6.12)$$

Для модели разложение суммы квадратов имеет вид (3.6.13):

$$k\sum_{j=1}^{m}\sum_{l=1}^{p_j}(\bar{y}_{*jl} - \bar{y})^2 = k\sum_{j=1}^{m} p_j(\bar{y}_{*j*} - \bar{y})^2 + k\sum_{j=1}^{m}\sum_{l=1}^{p_j}(\bar{y}_{*jl} - \bar{y}_{*j*})^2 \qquad (3.6.13)$$

где:

$SS_A = k\sum_{j=1}^{m} p_j(\bar{y}_{*j*} - \bar{y})^2$ - сумма квадратов, соответствующая эффекту фактора *A;*

$SS_{R_1} = k\sum_{i=1}^{k}\sum_{j=1}^{m}(\bar{y}_{*jl} - \bar{y}_{*j*})^2$ - остаточная сумма квадратов отклонений, которая есть не что иное, как сумма квадратов эффекта индивидуальных различий. Число степеней свободы для суммы SS_A: $\nu_A = m-1$, для остаточной суммы: $\nu_{R_1} = n-m$. Таким образом, *F*-отношение для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ будет иметь вид:

$$F = \frac{MS_A}{MS_{R_1}} = \frac{n-m}{m-1}\frac{SS_A}{SS_{R_1}}.$$

Если наблюдаемое значение статистики $F_{набл} \geq F_{кр}$, где $F_{кр}$ - критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы $\nu_1 = m-1$ и $\nu_2 = n-m$, то нулевая гипотеза отклоняется и считается, что эффект фактора *A* значим.

Таким образом, в дисперсионном анализе повторных измерений, по сути, используются две модели.

Первая модель в качестве источника данных использует центрированные по объектам наблюдений данные и исследует различия в повторных наблюдениях и соответствующие эффекты взаимодействия [101, 106].

Вторая модель в качестве источника данных использует средние по объектам наблюдения данные и исследует прочие эффекты. Если число наблюдений в слоях различное, то также как и в многофакторной модели дисперсионного анализа, следует использовать разложения сумм квадратов различных типов. Используемый тип разложения определяется исследователем, исходя из конкретной ситуации. Для большинства приложений дисперсионного анализа рекомендуется использовать III тип разложения [75].

Непараметрическим аналогом дисперсионного анализа повторных измерений является тест Фридмана. В тесте Фридмана предварительно ранжируются наблюдения для каждого из объектов наблюдения. Пусть r_{ij}, $i = \overline{1,k}$, $j = \overline{1,n}$ - ранг *j*-го наблюдения в *i*-ой группе.

Обозначим: $\overline{R}_i = \sum_{j=1}^{n} r_{ij}$ – средний ранг для i-ой группы, $\overline{R} = \frac{k+1}{2}$ - средний ранг всех наблюдений. Тогда выборочная дисперсия для средних рангов групп будет равна (3.6.14):

$$D = \frac{1}{(k-1)} \sum_{i=1}^{k} (\overline{R}_i - \overline{R})^2 \qquad (3.6.14)$$

Максимальное значение дисперсии достигается, когда все наблюдения проранжированы согласованно, в этом случае средние ранги принимают целые значения от 1 до k. Соответствующее наибольшее значение дисперсии при этом равно (3.6.15):

$$D_{\max} = \frac{k(k+1)}{12} \qquad (3.6.15)$$

Отношение $W = \frac{D}{D_{\max}} = \frac{12}{k^3 - k} \sum_{i=1}^{k} (\overline{R}_i - \overline{R})^2$ называется коэффициентом конкордации Кендалла, оно является характеристикой согласованности рангов наблюдений. Значения коэффициента конкордации меняются от 0 до 1, значение 1 соответствует, что все наблюдения проранжированы одинаково, то есть ранги для всех групп максимально различаются. Значение коэффициента конкордации равное нулю соответствует ситуации, когда средние ранги всех групп совпадают. С коэффициентом конкордации связана статистика Фридмана, определяемая по формуле (3.6.16):

$$S = n(k-1)W = \frac{12n}{k(k+1)} \sum_{i=1}^{k} (\overline{R}_i - \overline{R})^2 \qquad (3.6.16)$$

При условии истинности нулевой гипотезы (средние ранги по группам не различаются) статистика S имеет распределение Фридмана. При больших k, n статистика Фридмана приближенно имеет распределение Хи-квадрат с $k-1$ степенью свободы. Если наблюдаемое значение статистики $S_{набл} \geq S_{кр}$, где $S_{кр}$ - критическая точка распреде-

ления Хи-квадрат с числом степеней свободы $k-1$ уровня α (или квантиль уровня $1-\alpha$), то нулевая гипотеза отклоняется и считается, что средние ранги для различных уровней повторных измерений значимо различаются.

Для двух зависимых групп вместо теста Фридмана следует использовать тест Уилкоксона для зависимых выборок (аналог теста Манна-Уитни для независимых выборок).

3.7. Апостериорные множественные сравнения средних

Если в результате дисперсионного анализа получается, что средние значения отклика для разных уровней фактора, различаются, данный результат не является окончательным.

В дальнейшем подразумевается анализ того, для каких уровней фактора средние больше, для каких меньше, а для каких одинаковы. Основная процедура дисперсионного анализа не дает возможности ответить на эти вопросы.

Наиболее простым способом решения данной проблемы является проведение серии попарных сравнений с использованием t-критерия, используя в качестве оценки дисперсии величину MS_R - оценку внутригрупповой дисперсии, полученную в ходе дисперсионного анализа.

Такой подход реализуется в так называемом ***методе наименьшей*** значимой разности (LSD). Статистика критерия LSD для проверки гипотезы равенства средних μ_i и μ_j имеет вид (3.7.1):

$$t = \frac{\overline{\mu}_i - \overline{\mu}_j}{\sqrt{MS_R(1/n_i + 1/n_j)}} \qquad (3.7.1)$$

Если наблюдаемое значение статистики $|t_{набл}| \geq t_{кр}$, где $t_{кр}$ - критическая точка распределения Стьюдента уровня $\alpha/2$ (или квантиль уровня $1-\alpha/2$) с числом степеней свободы $\nu = n-k$, то нулевая гипотеза отклоняется и принимается гипотеза $H_1: \mu_1 \neq \mu_2$.

Существуют разные подходы к решению данной проблемы. Один из них – уменьшить уровень значимости при попарном сравнении так, чтобы вероятность хотя бы одного отклонения нулевой гипотезы равнялось заданному уровню значимости.

Такой подход реализуется в принципе Бонферрони множественных сравнений, в котором при каждом попарном сравнении задается уровень значимости α / C_k^2, где C_k^2 - число сравнений.

Данная величина гарантирует, что вероятность отклонение нулевой гипотезы (при ее истинности) хотя бы в одном из C_k^2 сравнений не превзойдет α. Однако, принцип Бонферрони является чересчур консервативным, он приводит к существенному снижению мощности критерия.

LSD – критерий и критерий Бонферрони занимают полярные позиции в ряду критериев множественных сравнений. Среди остальных критериев множественного сравнения средних можно выделить критерии множественных сравнений Шеффе, Ньюмена-Келса, Тьюки и другие.

В методе множественных сравнений Шеффе для проверки гипотезы равенства средних μ_i и μ_j используется статистика (3.7.2):

$$F = \frac{\left(\overline{\mu}_i - \overline{\mu}_j\right)^2}{(k-1)MS_R(1/n_i + 1/n_j)} \qquad (3.7.2)$$

где MS_R – оценка внутригрупповой (остаточной) дисперсии, полученная в ходе дисперсионного анализа. Если наблюдаемое значение статистики $F_{набл} \geq F_{кр}$, где $F_{кр}$ - критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы $\nu_1 = k-1$ и $\nu_2 = n-k$, то нулевая гипотеза отклоняется и принимается гипотеза $H_1 : \mu_i \neq \mu_j$.

Критерий Шеффе также относится к достаточно консервативным критериям, то есть обладает малой мощностью. Более мощными, соответственно, более чувствительными являются критерии Тьюки, Ньюмена-Келса, Дункана.

В методе множественных сравнений Тьюки (или достоверно значимой разности – HSD) для проверки гипотезы $H_0 : \mu_i = \mu_j$ против альтернативы $H_1 : \mu_i \neq \mu_j$ используется статистика (3.7.3):

$$t_R = \frac{\left|\overline{\mu}_i - \overline{\mu}_j\right|}{\sqrt{MS_R(1/n_i + 1/n_j)/2}} \qquad (3.7.3)$$

Ее значения сравниваются с критическими точками уровня α распределения стьюдентизированного размаха с $\nu_1 = k$ и $\nu_2 = n - k$ степенями свободы.

Если наблюдаемое значение статистики $t_{R\,набл} \geq t_{R\,кр}$, где $t_{R\,кр}$ - критическая точка распределения стьюдентизированного размаха уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы $\nu_1 = k$ и $\nu_2 = n - k$, то нулевая гипотеза отклоняется и принимается гипотеза $H_1 : \mu_i \neq \mu_j$.

Если объемы выборок различаются сильно, то рекомендуется использовать HSD критерий Тьюки для неравных выборок (критерий Spjovoll-Stoline). Статистика критерия в этом случае имеет вид (3.7.4)

$$t_R = \frac{\left|\overline{\mu}_i - \overline{\mu}_j\right|}{\sqrt{MS_R / \min(n_i, n_j)}} \qquad (3.7.4)$$

Критические точки определяются также, как и для критерия HSD Тьюки.

В ***критерии Ньюмана-Келса*** используется та же статистика, что и в критерии Тьюки, однако по другому определяются критические точки. В качестве критических точек критерия Ньюмана-Келса используются критические точки распределения стьюдентизированного размаха с $\nu_1 = r$ и $\nu_2 = n - k$ степенями свободы, где r - число средних расположенных между $\overline{\mu}_i$ и $\overline{\mu}_j$ в вариационном ряду выборочных средних, включая $\overline{\mu}_i$ и $\overline{\mu}_j$. Например, если сравниваются значения $\overline{\mu}_{(i)}$ и $\overline{\mu}_{(i+1)}$ вариационного (упорядоченного) ряда средних, то $r = 2$, если сравниваются значения $\overline{\mu}_{(i)}$ и $\overline{\mu}_{(i+2)}$, то $r = 3$ и так далее.

В пакете STATISTICA используется модифицированный вариант критерия Ньюмана-Келса, в котором в качестве статистики критерия используется величина (3.7.5):

$$t_R = \frac{\left|\overline{\mu}_i - \overline{\mu}_j\right|}{\sqrt{MS_R \frac{1}{k}\sum_{l=1}^{k}\frac{1}{n_l}}} \qquad (3.7.5)$$

Аналогичная статистика используется и в *критерии Дункана*, но в качестве критических точек берутся точки D-распределения Дункана с $\nu_1 = r$ и $\nu_2 = n - k$ степенями свободы, где r - число средних расположенных между $\overline{\mu}_i$ и $\overline{\mu}_j$ в вариационном ряду выборочных средних, включая $\overline{\mu}_i$ и $\overline{\mu}_j$.

Методы множественного сравнения средних можно использовать не только для проверки гипотез о попарном различии средних, а также для проверки гипотез о различии средних для любых выбранных наборов групп. В силу этого, основная гипотеза в данных методах в общем случае имеет вид: $H_0 : \sum_{i=1}^{k} c_i \mu_i = 0$, где $c_i, i = \overline{1,k}$ некоторые заданные константы, удовлетворяющие условию $\sum_{i=1}^{k} c_i = 0$. Например, при $c_3 = c_4 = \ldots = c_k = 0, c_1 = 1, c_2 = -1$, мы будем проверять гипотезу $H_0 : \mu_1 - \mu_2 = 0$ или $\mu_1 = \mu_2$. При $c_1 = 1, c_2 = -1/2, c_3 = -1/2$, $c_4 = c_5 = \ldots = c_k = 0$, будем проверять гипотезу $H_0 : \mu_1 = \frac{1}{2}(\mu_2 + \mu_3)$, то есть, гипотезу однородности первой и совокупности второй и третьей групп и т.д. Линейные комбинации вида: $\alpha(\mu_1 - \mu_2)$, $\alpha(\mu_1 - \frac{1}{2}(\mu_2 + \mu_3))$, то есть величины, пропорциональные разности между средними от средних, называются контрастами.

Критерии LSD, Шеффе, HSD Тьюки легко модифицировать под проверку гипотезы $H_0 : \sum_{i=1}^{k} c_i \mu_i = 0$. Например, статистика LSD критерия для проверки гипотезы $H_0 : \sum_{i=1}^{k} c_i \mu_i = 0$ будет иметь вид (3.7.6):

$$t = \frac{\sum_{i=1}^{k} c_i \overline{\mu}_i}{\sqrt{MS_R \sum_{i=1}^{k} (c_i^2 / n_i)}} \qquad (3.7.6)$$

Критическими точками статистики, по-прежнему, будут являться квантили распределения Стьюдента уровня $1-\alpha/2$ с числом степеней свободы $\nu = n-k$.

3.8. Двухфакторный дисперсионный анализ без повторений в MS Excel

Рассмотрим возможности программного комплекса на примере.

Пример

Необходимо провести дисперсионный анализ без повторений в MS Excel по данным, представленным в таблице (3.8.1).

Таблица 3.8.1

Исходные данные для анализа

	Факторы		
	1	2	3
A_1	1	2	6
A_2	5	6	10

Решение

Для анализа необходимо выполнить ряд операция. Следует перейти **Данные – Анализ данных – Двухфакторный дисперсионный анализ без повторений** (рис. 3.8.1) **– ОК**.

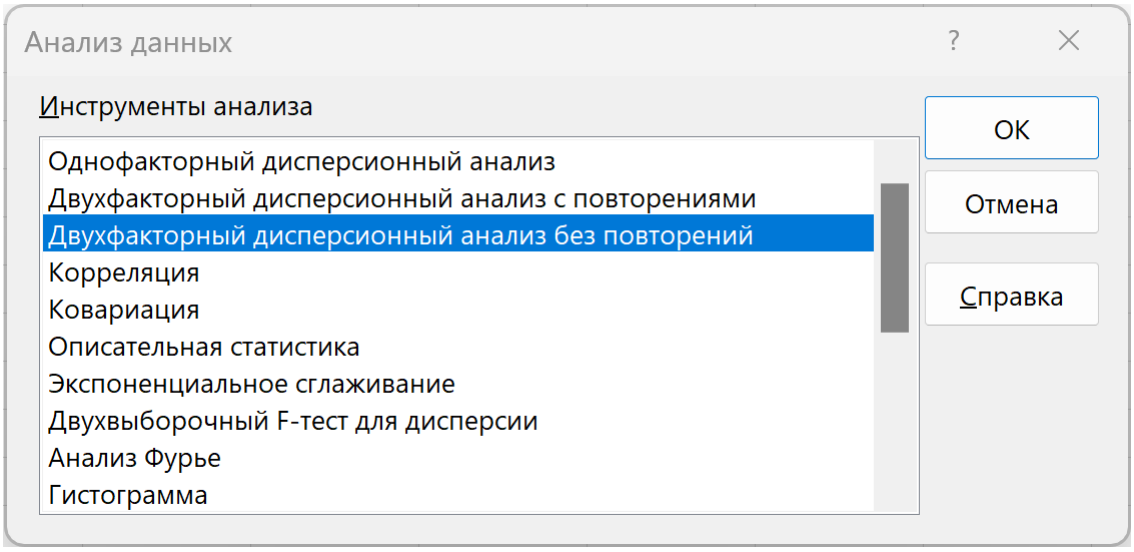


Рис. 3.8.1. Меню **Двухфакторный дисперсионный анализ без повторений**

В открывшемся меню (рис. 3.8.2) необходимо задать входной интервал (в нашем случае это **$B$3:$D$4**).

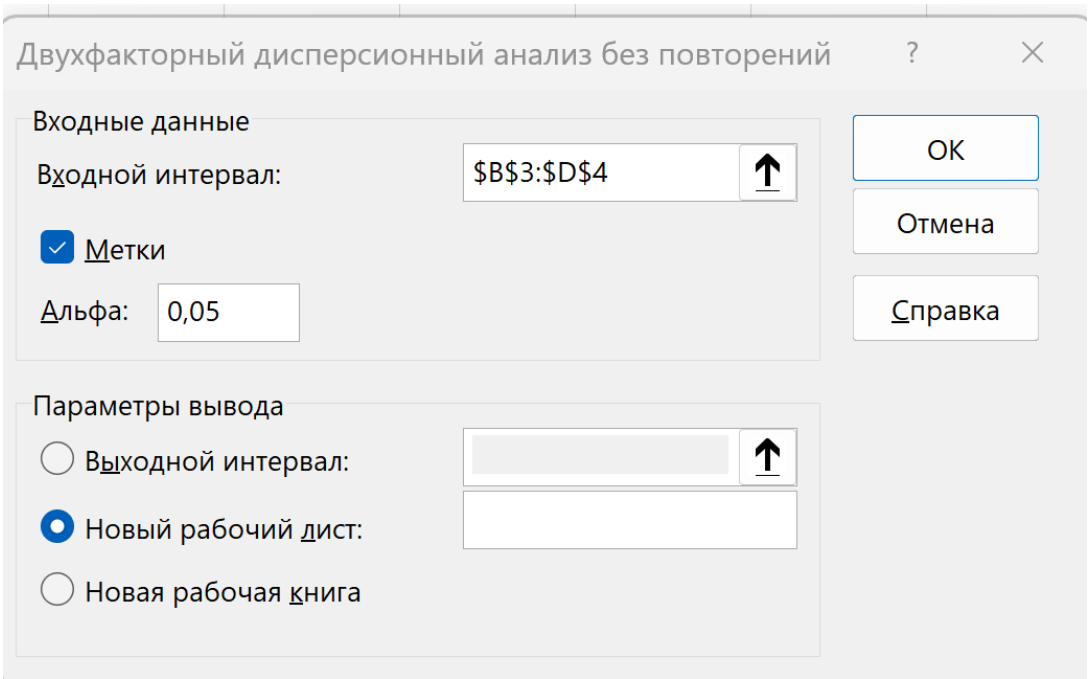


Рис. 3.8.2. Меню ввода параметров двухфакторного дисперсионного анализа без повторений

Для большего удобства следует отметить галочкой пункт **Метки** (галочка не ставится, если входной диапазон не содержит названий строк или столбцов, в том случае подходящие заголовки в выходном

диапазоне будут созданы автоматически)**.** Уровень альфы по умолчанию задан на уровне 0,05, однако при необходимости данный параметр можно скорректировать путем ввода вручную необходимой величины. Также следует поставить галочку напротив желаемого параметра вывода данных.

Выходной диапазон: введите ссылку на ячейку, расположенную в левом верхнем углу выходного диапазона. Размеры выходной области будут рассчитаны автоматически, и соответствующее сообщение появится на экране в том случае, если выходной диапазон занимает место существующих данных или его размеры превышают размеры листа.

Новый лист. Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя.

Новая книга. Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

После установления необходимых параметров и нажатия кнопки **ОК** будет выведена форма результатов проведенного анализа (рис. 3.8.3).

Двухфакторный дисперсионный анализ без повторений						
ИТОГИ	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
	3	6	2	1		
A1	3	9	3	7		
A2	3	21	7	7		
Факторы	3	7	2,333333333	5,333333333		
	3	10	3,333333333	5,333333333		
	3	19	6,333333333	12,33333333		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Строки	42	2	21	21	0,007561437	6,94427191
Столбцы	26	2	13	13	0,017777778	6,94427191
Погрешность	4	4	1			
Итого	72	8				

Рис. 3.8.3. Результаты двухфакторного дисперсионного анализа без повторений

Результаты представлены в виде двух таблиц:

- **Итоги.** В данной таблице представлены промежуточные данные расчетов для каждой строки и столбца: число элементов (**счет**), суммы величин (**сумма**), среднее арифметическое величин (**среднее**), дисперсия величин (**дисперсия**);

- **Дисперсионный анализ**. В данной таблице представлены собственно результаты дисперсионного анализа: компоненты дисперсии (источник вариации), суммы квадратов (*SS*), число степеней свободы (*df*), средний квадрат (*MS*), статистика F_B (*F*), вероятность значимости (*p*-значение), статистика *F* (*F*-критическое).

3.9. Двухфакторный дисперсионный анализ с повторениями в MS Excel

Рассмотрим возможности программного комплекса на примере.

Пример

Ниже в таблице 3.9.1 представлены данные о количестве дефектов на производстве в зависимости от того из какой партии были комплектующие и на каком уровне производилось комплектование оборудования. Необходимо провести двухфакторынй дисперсионный анализ и выявить существует ли влияние уровня комплектации оборудования на количество дефектов.

Таблица 3.9.1

Исходные данные для анализа

Уровень комплектности	Партия комплектующих									
	1					2				
A_1	190	206	170	170	170	190	150	210	150	150
A_2	150	250	220	140	180	230	190	200	190	200
A_3	190	185	135	195	195	150	170	150	170	180

Соответственно, при каждом уровне комплектования (A_j) были исследованы группы по пять образцов для выявления числа дефектов.

Решение

Необходимо отметить, что исходные данные, представленные в таблице 3.9.1 необходимо представить в виде транспонированных показателей (рис. 3.9.1).

	A	B	C	D
1		A1	A2	A3
2	1	190	150	190
3		206	250	185
4		170	220	135
5		170	140	195
6		170	180	195
7	2	190	230	150
8		150	190	170
9		210	200	150
10		150	190	170
11		150	200	180
12				

Рис. 3.9.1. Данные для анализа

Далее необходимо выполнить цепочку действий Данные – Анализ данных – Двухфакторный дисперсионный анализ с повторениями (рис. 3.9.2).

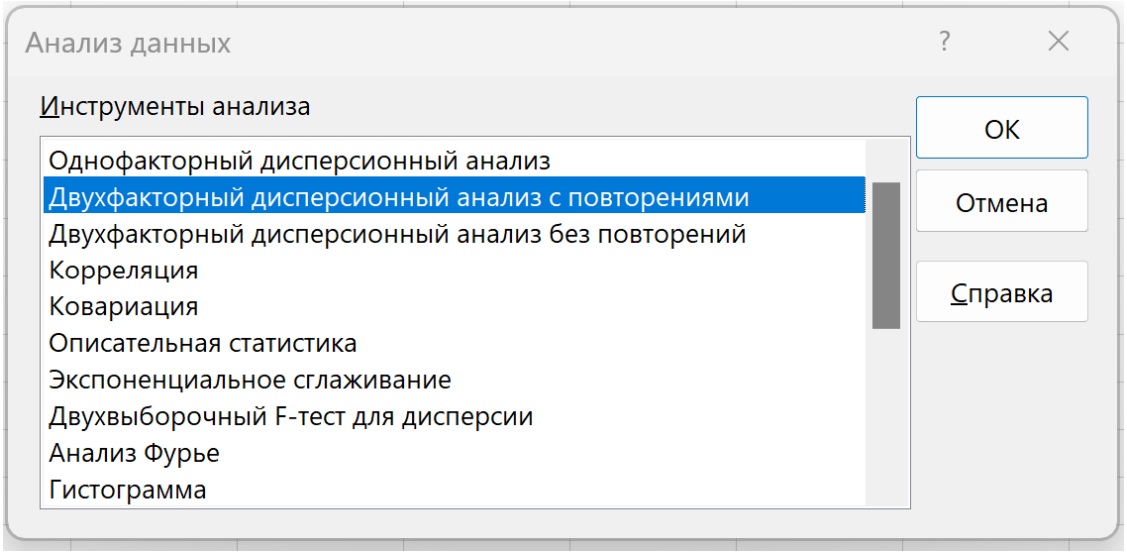


Рис. 3.9.2. Меню выбора **Двухфакторного дисперсионного анализа**

После нажатия кнопки **ОК** появится меню настройки двухфакторного дисперсионного анализа с повторениями (рис. 3.9.3).

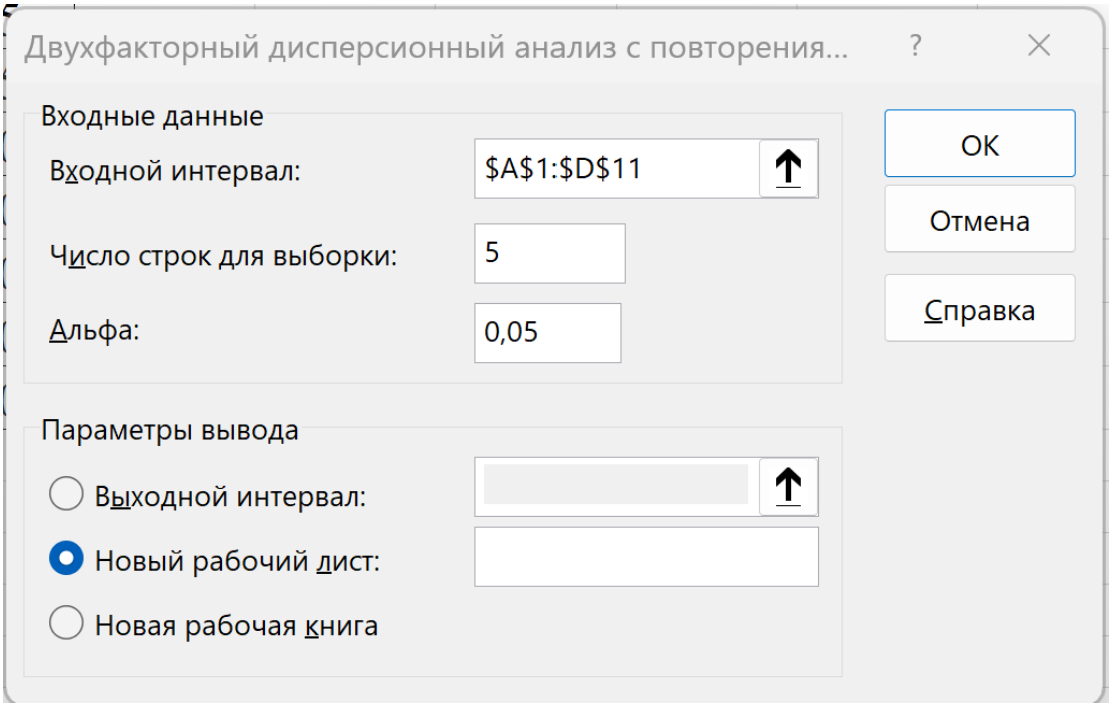


Рис. 3.9.2. Меню настройки **Двухфакторного дисперсионного анализа**

В открывшемся меню необходимо задать входной интервал (в нашем случае это **$A$1:$D$11**).

Число строк для выборки 5. Уровень альфы по умолчанию задан на уровне 0,05, однако при необходимости данный параметр можно

скорректировать путем ввода вручную необходимой величины. Также следует поставить галочку напротив желаемого параметра вывода данных.

Выходной диапазон: введите ссылку на ячейку, расположенную в левом верхнем углу выходного диапазона. Размеры выходной области будут рассчитаны автоматически, и соответствующее сообщение появится на экране в том случае, если выходной диапазон занимает место существующих данных или его размеры превышают размеры листа.

Новый лист. Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя.

Новая книга. Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

После установления необходимых параметров и нажатия кнопки **ОК** будет выведена форма результатов проведенного анализа (рис. 3.9.3).

Результаты представлены в виде двух таблиц:

- **Итоги**. В данной таблице представлены промежуточные данные расчетов для каждой строки и столбца: число элементов (**счет**), суммы величин (**сумма**), среднее арифметическое величин (**среднее**), дисперсия величин (**дисперсия**);

- **Дисперсионный анализ**. В данной таблице представлены собственно результаты дисперсионного анализа: компоненты дисперсии (источник вариации), суммы квадратов (*SS*), число степеней свободы (*df*), средний квадрат (*MS*), статистика *F* (*F*), вероятность значимости *p*-значение), статистика *F*кр (*F* критическое).

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ с повторениями						
2							
3	ИТОГИ	A1	A2	A3	Итого		
4	*1*						
5	Счет	5	5	5	15		
6	Сумма	906	940	900	2746		
7	Среднее	181,2	188	180	183,0666667		
8	Дисперсия	267,2	2170	650	895,352381		
9							
10	*2*						
11	Счет	5	5	5	15		
12	Сумма	850	1010	820	2680		
13	Среднее	170	202	164	178,6666667		
14	Дисперсия	800	270	180	655,2380952		
15							
16	*Итого*						
17	Счет	10	10	10			
18	Сумма	1756	1950	1720			
19	Среднее	175,6	195	172			
20	Дисперсия	509,1555556	1138,888889	440			
21							
22							
23	Дисперсионный анализ						
24	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
25	Выборка	145,2	1	145,2	0,200866919	0,658041916	4,259677273
26	Столбцы	3061,066667	2	1530,533333	2,117310707	0,14228243	3,402826105
27	Взаимодействие	1298,4	2	649,2	0,898090934	0,420600456	3,402826105
28	Внутри	17348,8	24	722,8666667			
29							
30	Итого	21853,46667	29				

Рис. 3.9.3. Результаты **Двухфакторного дисперсионного анализа**

Так как все расчетные значения критерия меньше критических значений, следовательно, влияние фактора (уровень комплектования оборудования) и фактора B (партии комплектующих), а также их взаимодействие на величину количества дефектов незначимо.

3.10. Многофакторный дисперсионный анализ в Statistica

Рассмотрим возможности программного комплекса на примере.

Пример

Сформируем таблицу, в которой в первом столбике будут представлены номер района города, а во втором число квартир, в третьем стоимость квадратного метра жилья (рис. 3.10.1).

	1 Район	2 Количество квартир	3 Цена квад. м
1	1	190	54000
2	1	206	52000
3	1	170	54000
4	1	170	32000
5	1	170	54000
6	2	190	62000
7	2	150	74000
8	2	210	68000
9	2	150	70000
10	2	150	72000
11	3	160	74000
12	3	175	73000
13	3	175	75000
14	3	180	76000
15	3	210	77000

Рис. 3.10.1 Исходные данные для анализа

Поскольку исследуется влияние двух факторов на значение средней цены квадратного метра, то данная задача решается методами многомерного дисперсионного анализа. Можно провести анализ влияния одновременно двух факторов на уровень средней цены без учета взаимодействия факторов. Такой факторный анализ является частным случаем многофакторного дисперсионного анализа и называется дисперсионным анализом главных эффектов. Классический же многомерный анализ, в отличии от анализа главных эффектов предполагает, кроме того, анализ эффектов взаимодействия факторов.

Необходимо перейти **Анализ** – **Дисперсионный анализ** (рис. 3.10.2).

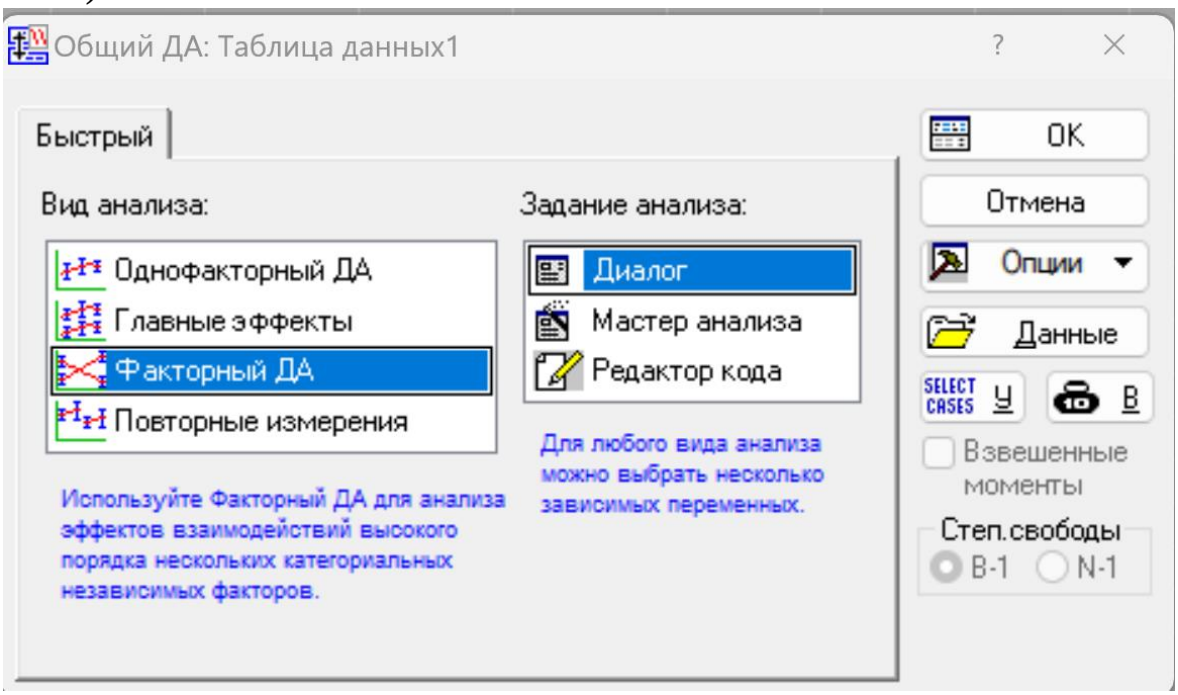


Рис. 3.10.2. Меню выбора параметров дисперсионного анализа

В меню выбрать пункт **Факторный ДА – Диалог**. После нажатия на **OK** попадаем в окно задания условий для дисперсионного анализа. На вкладке **Быстрый** (рис. 3.10.3), нажав на кнопку, попадаем в окно выбора переменных для анализа (рис. 3.10.4).

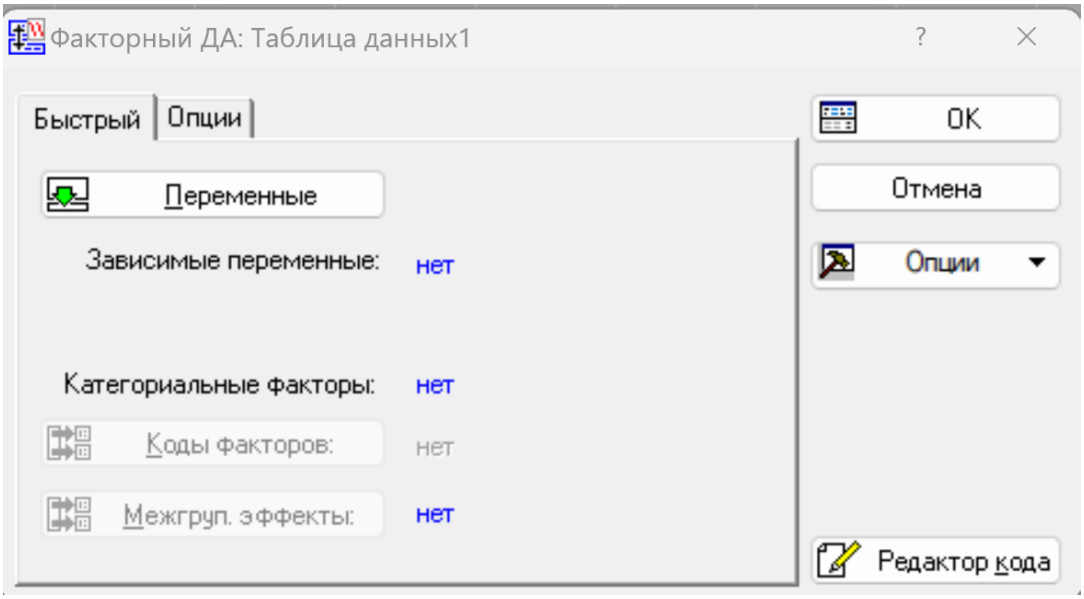


Рис. 3.10.3. Вкладка **Быстрый**

Выбираем в качестве зависимой переменной переменную «Цена квад м», а в качестве категориальных переменных (факторов) - переменные «Район» и «Количество квартир». Можно также выбрать уровни (коды) категориальных переменных, по которым будет проводиться анализ. Если коды не задавать, анализ будет проводиться по всем уровням категориальных переменных.

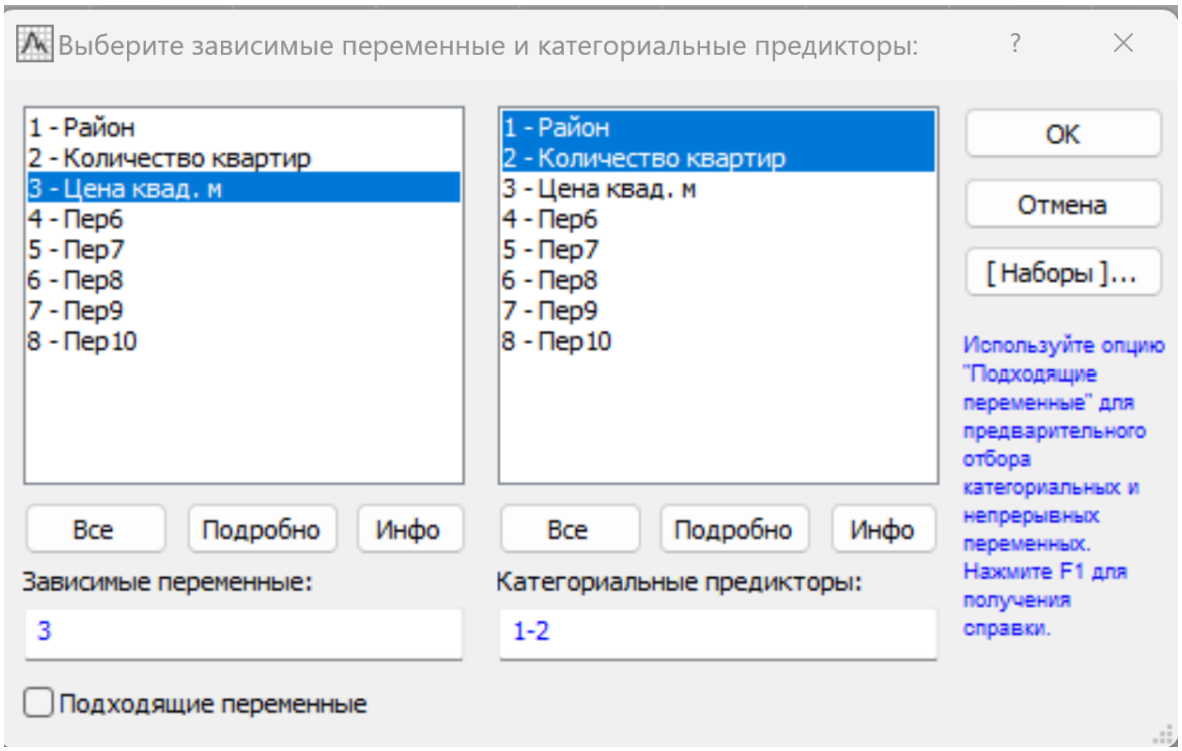


Рис. 3.10.4. Выбор переменных при проведении дисперсионного анализа

На вкладке **Опции** (рис. 3.10.5) можно задать тип суммы квадратов для несбалансированного дисперсионного анализа. По умолчанию стоит тип VI (к данному типу наиболее близок тип III из “классических” тип сумм квадратов).

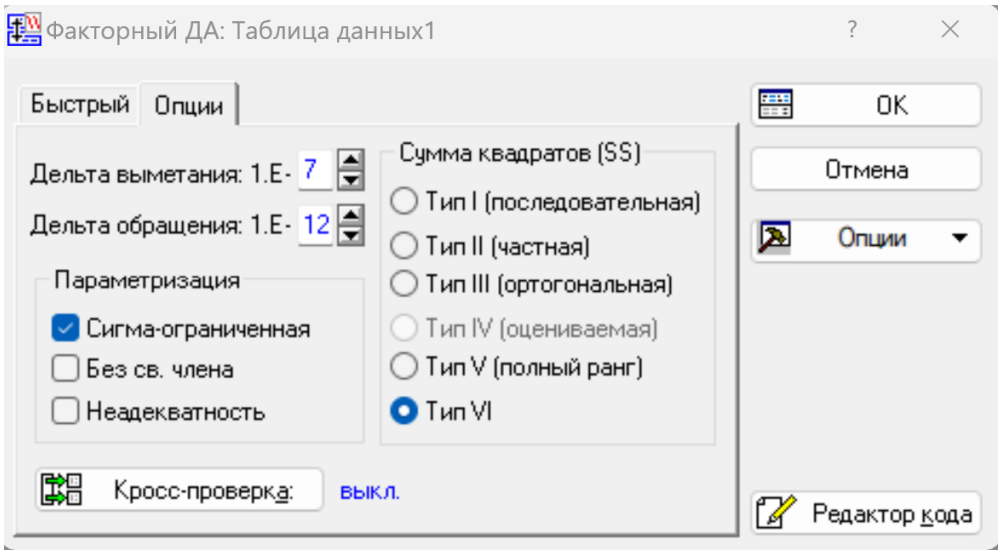


Рис. 3.10.5. Вкладка **Опции** при проведении дисперсионного анализа

Задав переменные и выбрав тип сумм квадратов, после нажатия на кнопку **OK** переходим в окно результатов дисперсионного анализа (рис. 3.10.6) и выбираем вкладку Итоги (рис. 3.10.7).

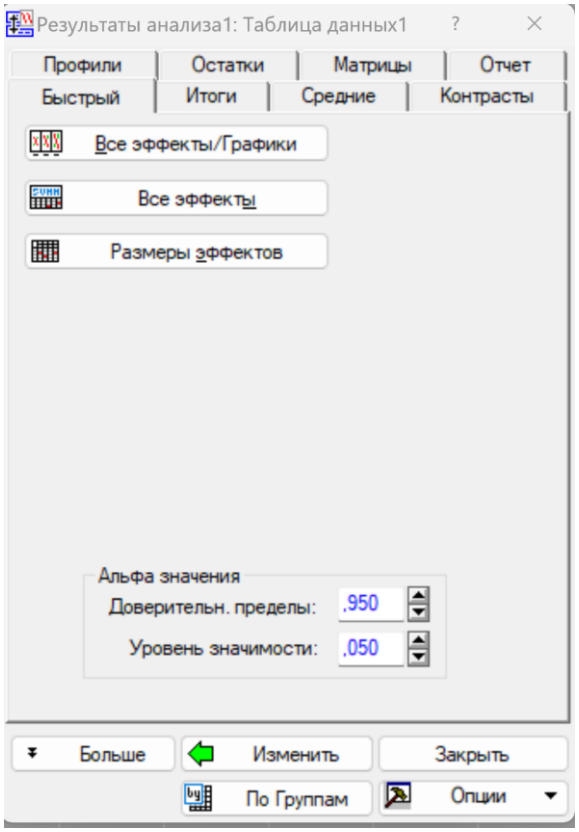


Рис. 3.10.6. Результаты дисперсионного анализа

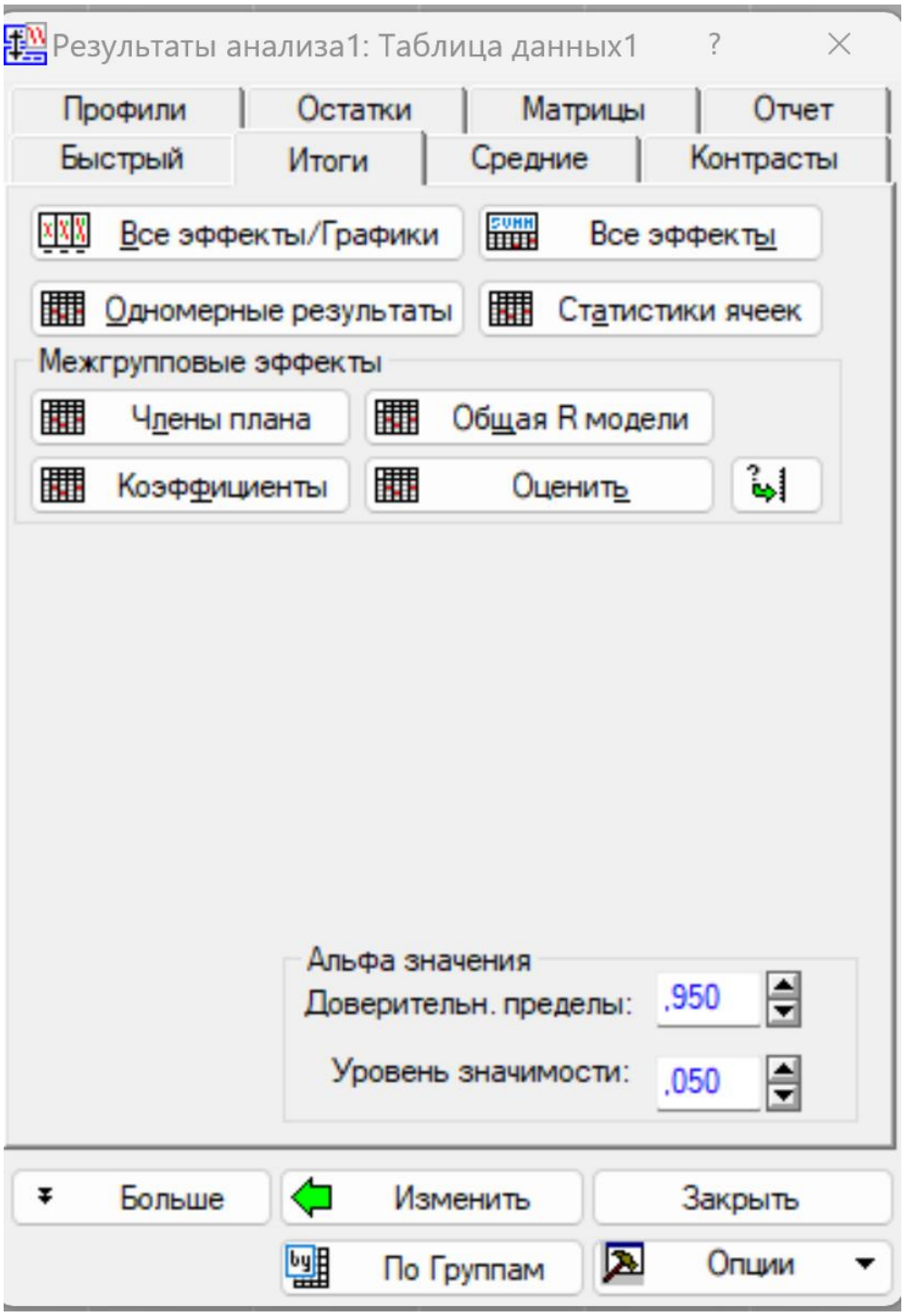


Рис. 3.10.7. Вкладка **Итоги** результатов дисперсионного анализа

Для просмотра описательной статистики на вкладке **Итоги** следует выбрать **Статистика ячеек**. После нажатия кнопки ОК появится таблица данных описательной статистики (рис. 3.10.8).

Эффект	Описательные статистики (Таблица данных1)							
	Уровень Фактор	Уровень Фактор	N	Цена квад. м Среднее	Цена квад. м Ст.откл.	Цена квад. м Стд.ош.	Цена квад. м -95,00%	Цена квад. м +95,00%
Всего			35	66742,86	10253,67	1733,187	63220,6	70265,1
Район	1		7	54571,43	12581,17	4755,233	42935,8	66207,1
Район	2		7	72571,43	2820,00	1065,859	69963,4	75179,5
Район	3		7	68142,86	7840,68	2963,497	60891,4	75394,3
Район	4		7	70857,14	7798,05	2947,384	63645,2	78069,1
Район	5		7	67571,43	8482,47	3206,074	59726,4	75416,4
Количество квартир	130		1	74000,00				
Количество квартир	135		2	76000,00	1414,21	1000,000	63293,8	88706,2
Количество квартир	140		1	62000,00				
Количество квартир	150		8	70500,00	6969,32	2464,027	64673,5	76326,5
Количество квартир	160		1	77000,00				
Количество квартир	170		3	46666,67	12701,71	7333,333	15113,9	78219,5
Количество квартир	175		2	58000,00	5656,85	4000,000	7175,2	108824,8
Количество квартир	180		1	74000,00				
Количество квартир	190		2	61000,00	9899,49	7000,000	-27943,4	149943,4
Количество квартир	195		8	70875,00	7356,97	2601,082	64724,4	77025,6
Количество квартир	206		1	52000,00				
Количество квартир	210		5	67200,00	5019,96	2244,994	60966,9	73433,1
Район*Количество квартир	1	130	1	74000,00				
Район*Количество квартир	1	140	1	62000,00				
Район*Количество квартир	1	170	3	46666,67	12701,71	7333,333	15113,9	78219,5
Район*Количество квартир	1	190	1	54000,00				
Район*Количество квартир	1	206	1	52000,00				
Район*Количество квартир	2	135	1	75000,00				
Район*Количество квартир	2	150	3	72333,33	2081,67	1201,850	67162,2	77504,5
Район*Количество квартир	2	190	1	68000,00				
Район*Количество квартир	2	195	1	76000,00				
Район*Количество квартир	2	210	1	72000,00				
Район*Количество квартир	3	160	1	77000,00				
Район*Количество квартир	3	175	2	58000,00	5656,85	4000,000	7175,2	108824,8
Район*Количество квартир	3	180	1	74000,00				
Район*Количество квартир	3	195	2	71000,00	1414,21	1000,000	58293,8	83706,2
Район*Количество квартир	3	210	1	68000,00				
Район*Количество квартир	4	135	1	77000,00				
Район*Количество квартир	4	150	4	73250,00	2500,00	1250,000	69271,9	77228,1
Район*Количество квартир	4	195	1	54000,00				

Рис. 3.10.8. Описательная статистика дисперсионного анализа

Для просмотра результатов дисперсионного анализа выбираем **Все эффекты**, в результате получаем таблицу, изображенную на рисунке 3.10.9.

Эффект	Одномерный критерий значимости дляЦена квад. м (Таблица данных1) Сигма-ограниченная параметризация Декомпозиция гипотезы				
	SS	Степени свободы	MS	F	p
Св. член	38236928864	1	38236928864	666	0
Район	183737369	4	45934342	1	1
Количество квартир	1072252845	11	97477531	2	0
Ошибка	1090604298	19	57400226		

Рис. 3.10.9. Результаты многофакторного дисперсионного анализа

Первую строку таблицы (эффект **Св. член**) можно проигнорировать (в ней оценивается насколько значимо отличается от нуля значение средней цены по всем квартирам). Во второй и третьих строках

таблицы приводятся эффекты факторов **Район** и **Количество квартир** - суммы квадратов отклонений (*SS*), средние суммы квадратов отклонений (*MS*) с указанием значения статистики Фишера и наблюдаемого уровня значимости *p*. В четвертой строке таблицы приводятся суммы квадратов отклонений (*SS*), средние суммы квадратов отклонений (*MS*) для остатков или внутригруппового разброса.

Для построения графиков средних разных эффектов на вкладке **Итоги** нажимаем на кнопку **Все эффекты/Графики** и в появившемся окне (рис. 3.10.10) выбираем эффект, для которого будут построены графики средних с доверительными интервалами.

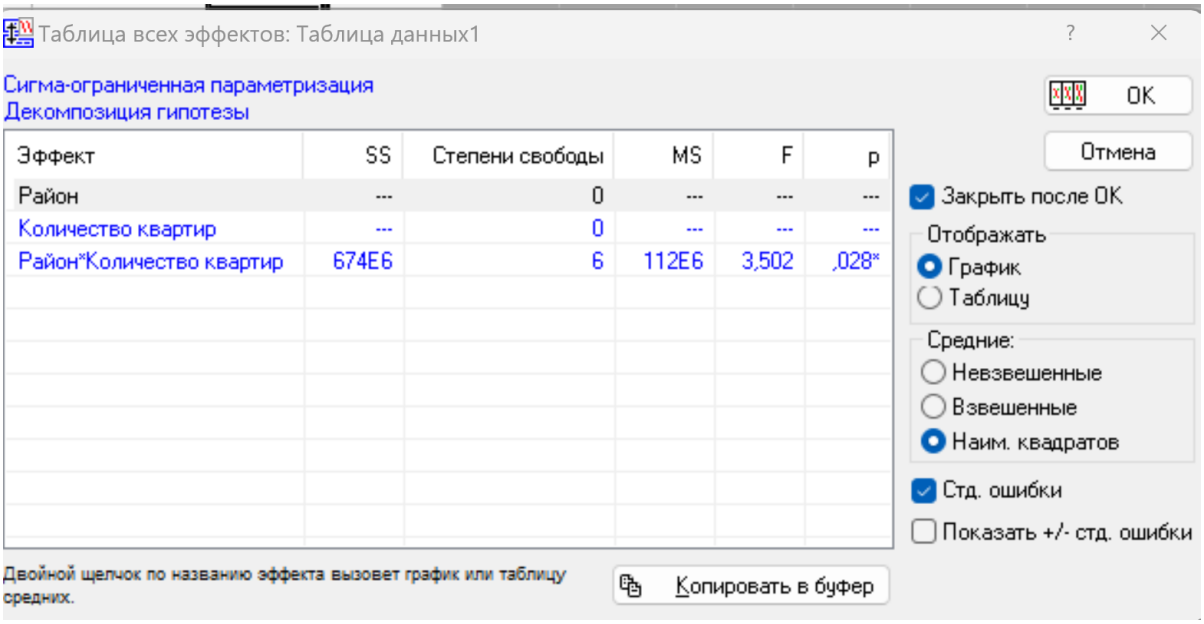


Рис. 3.10.10. Окно для выбора эффектов для построения доверительных интервалов средних

Если для параметра **Отображать** указать значение **Таблицу**, то будет доверительные интервалы для средних, соответствующих разным уровням фактора будут отображены в виде таблицы. По умолчанию строятся графики с 95% доверительными интервалами. Уровень доверия можно задать на вкладке **Итоги** с помощью параметра **Доверительные пределы**.

На рисунке 3.10.11 приведен график средних для эффекта **Район**, на рисунке 3.10.12 - график средних для эффекта **Количество квартир**.

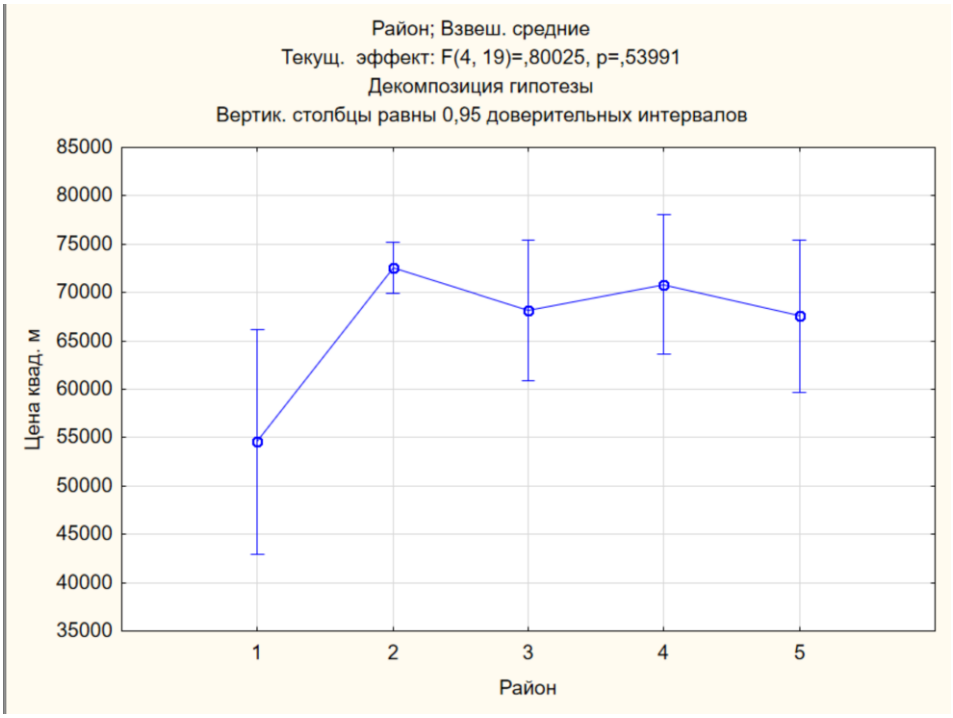


Рис. 3.10.11. График средних для эффекта **Район**

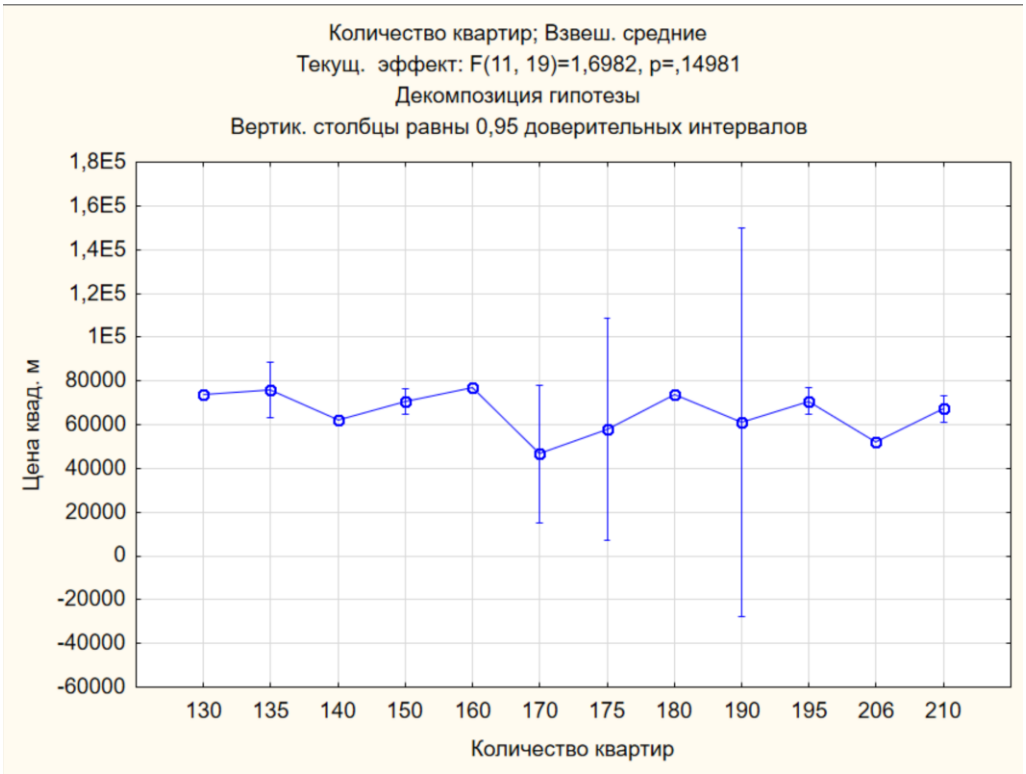


Рис. 3.10.12. График средних для эффекта **Количество квартир**

При построении графиков для эффекта взаимодействия следует указать, какой из факторов будет откладываться по оси OX, а какому будут соответствовать различные графики.

Для выявления значимо различающихся средних для эффектов факторов **Район** и **Количество квартир** используем метод множественных сравнений.

Для этого в модуле результатов дисперсионного анализа, путем нажатия кнопки **Больше**, выбираем расширенный режим, переходим на вкладку апостериорных сравнений средних **Апостер.**, выбираем эффект и выбираем один из методов множественного сравнения (из 8 предлагаемых) (рис. 3.10.13).

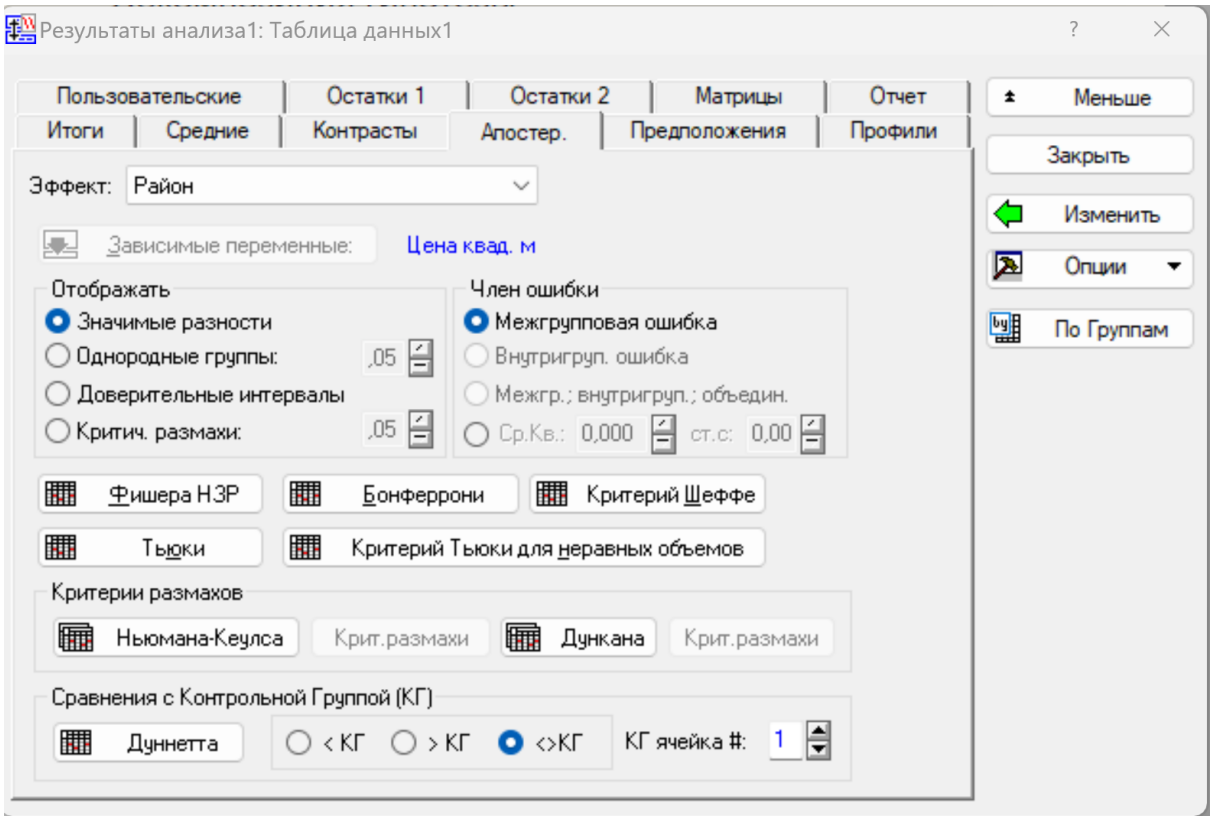


Рис. 3.10.13. Выбор метода множественных сравнений для эффекта **Район**

Для режима отображения (параметр **Отображать**) устанавливаем **Значимые разности**. В результате получим таблицу уровней значимости попарных различий средних для всех комбинаций уровней факторов. На рисунках 3.10.14 – 3.10.16 приведены результаты множественных сравнений средних для эффекта Район по критериям Фишера НЗР, Шеффе и Дункана.

НЗР крит.; перем. Цена квад. м (Таблица данных1)
Вероятности для апостер. критериев
Ошибка: Межгр. MS = 5740E4, cc = 19,000

N ячейки	Район	{1} 54571,	{2} 72571,	{3} 68143,	{4} 70857,	{5} 67571,
1	1		0,000278	0,003355	0,000730	0,004608
2	2	0,000278		0,287821	0,676821	0,232009
3	3	0,003355	0,287821		0,510767	0,889273
4	4	0,000730	0,676821	0,510767		0,427214
5	5	0,004608	0,232009	0,889273	0,427214	

Рис. 3.10.14. Уровни значимости для попарных различий средних для уровней фактора **Район** по критерию Фишера НЗР

Крит. Шеффе; перемен. Цена квад. м (Таблица данных1)
Вероятности для апостер. критериев
Ошибка: Межгр. MS = 5740E4, cc = 19,000

N ячейки	Район	{1} 54571,	{2} 72571,	{3} 68143,	{4} 70857,	{5} 67571,
1	1		0,006692	0,054945	0,015456	0,070760
2	2	0,006692		0,874971	0,995871	0,819343
3	3	0,054945	0,874971		0,976657	0,999946
4	4	0,015456	0,995871	0,976657		0,953742
5	5	0,070760	0,819343	0,999946	0,953742	

Рис. 3.10.15. Уровни значимости для попарных различий средних для уровней фактора **Район** по критерию Шеффе

Крит. Дункана; перем. Цена квад. м (Таблица данных1)
Приближенные вероятности для апостер. критериев
Ошибка: Межгр. MS = 5740E4, cc = 19,000

N ячейки	Район	{1} 54571,	{2} 72571,	{3} 68143,	{4} 70857,	{5} 67571,
1	1		0,000598	0,004549	0,001284	0,004758
2	2	0,000598		0,314120	0,676955	0,271557
3	3	0,004549	0,314120		0,510919	0,889393
4	4	0,001284	0,676955	0,510919		0,452996
5	5	0,004758	0,271557	0,889393	0,452996	

Рис. 3.10.15. Уровни значимости для попарных различий средних для уровней фактора **Район** по критерию Дункана

Как видим, в данном случае, результаты всех тестов практически идентичные.

Для того, чтобы выделить однородные группы, статистически не различающиеся по средней цене, на вкладке **Апостер.** для режима отображения (параметр **Отображать**) устанавливаем значение **Однородные группы**. Выбираем эффект, задаем уровень значимости, например, (чем больше уровень, тем более близкие группы будут выделены) и выбираем критерий множественного сравнения, (например, Щеффе). В результате получаем однородные кластеры групп, расположенные в порядке возрастания средних значений (рис. 3.10.16).

Крит. Шеффе; перемен. Цена квад. м (Таблица данных1)
Однородные группы, alpha = ,05000
Ошибка: Межгр. MS = 5740E4, cc = 19,000

N ячейки	Район	Цена квад. м Среднее	1	2
1	1	54571,43		****
5	5	67571,43	****	****
3	3	68142,86	****	****
4	4	70857,14	****	
2	2	72571,43	****	

Рис. 3.10.16. Однородные кластеры групп в соответствии с выбранным критерием множественного сравнения (Шеффе) и заданным уровнем значимости

Хотя в данном случае условия корректного применения дисперсионного анализа выполнены лишь частично, все же основные результаты дисперсионного анализа можно признать.

Контрольные вопросы по теме

1. Что такое дисперсионный анализ (ANOVA)?
2. Какие основные типы дисперсионного анализа существуют?
3. В чем отличие между однофакторным и многофакторным дисперсионным анализом?
4. Какие предпосылки должны быть выполнены для применения дисперсионного анализа?
5. Как проводится однофакторный дисперсионный анализ?

6. Как интерпретировать результаты однофакторного дисперсионного анализа?

7. Какие значения могут принимать F-статистика и что они означают?

8. Как можно определить статистическую значимость различий между группами с помощью дисперсионного анализа?

9. Как проводится многофакторный дисперсионный анализ?

10. Как интерпретировать результаты многофакторного дисперсионного анализа?

11. Какие значения могут принимать F-статистики в многофакторном дисперсионном анализе и что они означают?

12. Как можно определить влияние каждого фактора на зависимую переменную в многофакторном дисперсионном анализе?

13. Как можно использовать результаты дисперсионного анализа для принятия бизнес-решений?

14. Как можно использовать результаты дисперсионного анализа для оптимизации производственных процессов?

15. Как можно использовать результаты дисперсионного анализа для планирования маркетинговых стратегий?

16. Как можно использовать результаты дисперсионного анализа для выявления влияния различных факторов на финансовые показатели компании?

17. Как можно использовать результаты дисперсионного анализа для оптимизации управленческих процессов в компании?

18. Как можно использовать результаты дисперсионного анализа для прогнозирования тенденций на финансовых рынках?

19. Как можно использовать результаты дисперсионного анализа для выявления связей между различными социальными явлениями или процессами?

20. Как можно использовать результаты дисперсионного анализа для выявления факторов, влияющих на здоровье или качество жизни людей?

21. Как можно использовать результаты дисперсионного анализа для оптимизации учебного процесса в образовательных учреждениях?

22. Как можно использовать результаты дисперсионного анализа для выявления влияния различных факторов на экологические процессы или явления?

23. Как можно использовать результаты дисперсионного анализа для оптимизации процессов в медицине или здравоохранении?

24. Как можно использовать результаты дисперсионного анализа для выявления влияния различных факторов на поведение потребителей или клиентов компаний?

25. Как можно использовать результаты дисперсионного анализа для выявления связей между различными экономическими явлениями или процессами?

26. Как можно использовать результаты дисперсионного анализа для оптимизации процессов в транспортной отрасли или логистике?

27. Как можно использовать результаты дисперсионного анализа для выявления влияния различных факторов на политические процессы или явления?

28. Какие программные инструменты чаще всего используются для проведения дисперсионного анализа?

29. Как можно использовать результаты дисперсионного анализа для оценки эффективности различных методов или стратегий?

30. Как можно использовать результаты дисперсионного анализа для выявления важных факторов, влияющих на зависимую переменную?

31. Как можно использовать результаты дисперсионного анализа для определения оптимальных условий или параметров процессов?

32. Как можно использовать результаты дисперсионного анализа для выявления тенденций или закономерностей в данных?

33. Как можно использовать результаты дисперсионного анализа для прогнозирования будущих значений зависимой переменной?

34. Как можно использовать результаты дисперсионного анализа для сравнения различных групп или образцов данных?

35. Как можно использовать результаты дисперсионного анализа для выявления воздействия различных факторов на изучаемый процесс или явление?

36. Что такое внутригрупповая и межгрупповая изменчивость и как они отражены в результатах дисперсионного анализа?

37. Что такое среднеквадратическое отклонение и как оно используется в дисперсионном анализе?

38. Что такое сумма квадратов отклонений и как она используется при проведении дисперсионного анализа?

39. Что такое общая сумма квадратов отклонений и как она интерпретируется при проведении дисперсионного анализа?

40. Что такое внутригрупповая сумма квадратов отклонений и как она интерпретируется при проведении дисперсионного анализа?

41. Что такое межгрупповая сумма квадратов отклонений и как она интерпретируется при проведении дисперсионного анализа?

42. Что такое степени свободы и как они используются при расчете F-статистики в дисперсионном анализе?

43. Что такое среднеквадратическая ошибка и как она используется при проведении дисперсионного анализа?

44. Что такое F-статистика и как она используется для проверки статистической значимости различий между группами при проведении дисперсионного анализа?

45. Что такое межгрупповая средняя квадратическая ошибка и как она интерпретируется при проведении многофакторного дисперсионного анализа?

46. Что такое внутригрупповая средняя квадратическая ошибка и как она интерпретируется при проведении многофакторного дисперсионного анализа?

47. Что такое межфакторное воздействие и как оно учитывается при проведении многофакторного дисперсионного анализа?

48. Что такое внутрифакторное воздействие и как оно учитывается при проведении многофакторного дисперсионного анализа?

49. Что такое взаимодействие факторов и как оно учитывается при проведении многофакторного дисперсионного анализа?

50. Какие методы используются для проверки предпосылок и условий применения дисперсионного анализа?

4. КЛАСТЕРНЫЙ АНАЛИЗ

4.1. Общие положения кластерного анализа

По своей сути кластерный анализ представляет собой процесс исследования, с помощью которого набор каких-либо объектов (исследуемых параметров) объединяется или группируется в относительно небольшие группы, именуемые кластерами, которые имеют черты внутригруппового сходства и набор четко выраженных отличий с другими кластерами.

Применение для решения задач кластерного анализа пакетов прикладных статистических программ значительно упрощает исследование, поскольку перегруппировка кластеров при изменении, добавлении или исключении какого-либо признака осуществляется без необходимости проводить дополнительные массивные вычисления вручную.

Впервые кластерный анализ как метод исследования нашел свое применение в социологии. Само слово «cluster» происходит от английского определения понятий «скопление» и «гроздь». Родоначальником кластерного исследования можно назвать Роберта Триона, который в 1939 году выпустил труд, систематизирующий представления о предмете, сущности и базовых методах кластеризации. Сущность кластерного анализа, вне зависимости от области его применения, можно определить следующим образом: имеет место некоторое число объектов, которое требуется разбить на подмножества, отличающиеся максимальным сходством внутри исследуемых групп (кластеров) и существенными различиями между самими кластерами. По сути, кластерный анализ представляет собой метод классификации, предполагающий исследование структуры образованных групп.

Основное преимущество кластеризации как метода состоит в том, что выявление групп происходит не по какому-либо одному признаку, а по их совокупности, что позволят проводить всестороннее изучение анализируемых объектов.

Также стоит отметить, что данный математико-статистический метод никак не ограничивает область рассматриваемых объектов, поскольку применим для данных практически любой природы. Данное свойство является наиболее ценным, когда анализируются и исследуются разнообразные признаки. Применение иных эконометрических методов и приемов в таком случае может быть крайне затруднительным.

Применение кластерного анализа позволяет существенно сократить объем данных, сохраняя при этом существенные свойства и признаки выделенных массивов данных. Кроме того, использование данного метода возможно в сочетании с другими качественными методами эконометрики.

Также использование кластеризации активно используется при анализе рядов динамики, позволяя выделять временные интервалы с общими характеристиками.

При проведении кластерного анализа есть возможность применять данный метод циклически, дополняя каждый цикл анализа уточняющей информацией.

Следует отметить, что кластерный анализ имеет также ряд недостатков и ограничений. В частности, качество кластерного анализа во многом обусловлено тщательностью отбора критериев разбиения. Также сведение массива исходных данных к объединениям внутри кластерных групп приводит к некоторым искажениям и отсутствию учета индивидуальных признаков. При кластеризации важно, чтобы соблюдался целый ряд условий.

Выбор масштаба является крайне значимым этапом проведения кластеризации. Как правило, на предварительном этапе исходные данные, характеризующие признак, нормализуются путем вычитания среднего и делением на стандартное отклонение. При такой операции дисперсия равна единице.

Основная задача проведения кластерного анализа состоит в том, чтобы разбить на основании данных, содержащихся в множестве X, множество объектов G на m кластеров таким образом, чтобы каждый из анализируемых объектов G_i принадлежал только одному из определенных подмножеств разбиения. Важно обеспечение сходства между объектами, принадлежащими одному кластеру, и наличие существенных межкластерных различий.

По сути, решение задачи кластерного анализа сводится к разбиению совокупности, которая удовлетворяет условию оптимальности в рамках конкретной исследовательской задачи. По сути критерий оптимальности может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. В роли целевой функции может быть использована внутригрупповая сумма квадратов отклонений.

Важнейшими характеристиками кластера являются его размер, центр, радиус, среднеквадратическое отклонение.

Центр кластера - среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от центра кластера. Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными (по мере сходства могут быть отнесены к нескольким кластерам).

Размер кластера - может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

При возникновении проблемы неопределенности, она может быть решена аналитиком или экспертом.

Как уже было отмечено выше, проблема масштаба является значимой при проведении кластерного исследования. Предположим, что набор данных содержит сведения о двух признаках *x* и *y*. При этом *x* принадлежит диапазону от 100 до 700, а *y* - от 0 до 1. В таком случае корректный расчет расстояний между точками, характеризующими положение объектов, становится невозможным, поскольку переменная, имеющая большие значения, т.е. переменная *x*, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной *y*.

Для устранения данной проблемы используется процедура предварительной стандартизации данных.

Стандартизация (standardization) или нормирование (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Применяются различные способы проведения данной процедуры, например, по формулам (4.1.1-.4.1.4):

$$z = \frac{x - \overline{x}}{\sigma} \qquad (4.1.1)$$

$$z = \frac{x}{\overline{x}} \qquad (4.1.2)$$

$$z = \frac{x}{x_{max}} \qquad (4.1.3)$$

$$z = \frac{x - \bar{x}}{x_{max} - x_{min}} \qquad (4.1.4)$$

где $\bar{x}$ - среднее значение признака;

σ - среднеквадратическое отклонение x;

x_{max} –наибольшее значение признака;

x_{min} –наименьшее значение признака.

Среднеквадратическое отклонение σ определяется, исходя из числа наблюдений, и может быть определено по формуле 4.1.5:

$$\sigma = \sqrt{\frac{1}{n-1}\sum^{n}(x_i - \bar{x})^2} \qquad (4.1.5)$$

Основных видов стандартизации в кластерном анализе представлены в таблице 4.1.1

Таблица 4.1.1

Основные виды стандартизации в кластерном анализе

Стандартизация	Расчет
Z-шкалы (Z-Scores)	Из значений переменных вычитается их среднее, и эти значения делятся на стандартное отклонение.
Разброс от -1 до 1	Линейным преобразованием переменных добиваются разброса значений от -1 до 1
Разброс от 0 до 1	Линейным преобразованием переменных добиваются разброса значений от 0 до 1.

Максимум 1	Значения переменных делятся на их максимум.
Среднее 1	Значения переменных делятся на их среднее
Стандартное отклонение 1	Значения переменных делятся на стандартное отклонение.

Помимо процедуры стандартизации, на практике также применяется вариант решения существующей проблемы путем корректировки параметров с учетом коэффициента важности, который представляет собой весовую характеристику, отражающую значимость конкретной переменной. При определении данного коэффициента может использоваться метод экспертных оценок, предусматривающий проведение экспертного опроса специалистом конкретной предметной области.

Полученные произведения нормированных переменных на соответствующие веса дают возможность адекватной оценки расстояний между точками в многомерном пространстве с учетом неодинакового веса переменных.

Вне зависимости от используемого метода классификации и подходов к определению кластеров, проблема измерения близости объектов является неизбежной. При этом основу данного вопроса составляют два основных положения: неоднозначность выбора способа нормировки и определение расстояния между объектами.

В кластерном анализе для количественной оценки сходства вводится понятие метрики. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние.

Метрикой между анализируемыми объектами в пространстве принято называть такую величину d_{ab}, которая бы удовлетворяла аксиомам (4.1.6-4.1.9):

$$A_1: d_{ab} > 0 \qquad (4.1.6)$$

$$A_2: d_{aa} = 0 \qquad (4.1.7)$$

$$A3: d_{ba} = d_{ab} \qquad (4.1.8)$$

$$A4: d_{ab} + d_{bc} \geq d_{ac} \qquad (4.1.9)$$

В качестве меры близости, характеризующей степень сходства, применяется величина $\mu_{ab,}$ для которой существует конкретный предел, и которая возрастает с усилением степени близости объектов.

Условия меры близости (сходства) объектов становится возможным сформулировать следующим образом:

- μ_{ab} – непрерывна, т. е. малому изменению положения точек в пространстве отвечает малое изменение меры;

- $\mu_{ab} = \mu_{ba}$;

- $0 \leq \mu_{ab} \leq 1$; $\mu_{ab} = 1 \leftrightarrow a = b$ (если a = b).

Для перехода от метрики к расстоянию близости объектов применяют формулу (4.1.10):

$$\mu = \frac{1}{1+d} \qquad (4.1.10)$$

Объединение или метод древовидной кластеризации используется при формировании кластеров несходства или расстояния между объектами. Эти расстояния могут определяться в одномерном или многомерном пространстве. Например, если требуется кластеризовать организации или фирмы, то можно принять во внимание количество работников предприятия, годовой объем выручки, их рентабельность и т.д. Наиболее прямой путь вычисления расстояний между объектами в многомерном пространстве состоит в вычислении евклидовых расстояний. Если имеется двух- или трёхмерное пространство, то данная мера представляет собой реальное геометрическое расстояние между объектами в пространстве.

Данный метод анализа является самым простым с точки зрения вычислений, однако, применение такого алгоритма не дает возможность оценки «реальности» такого расстояния или его производного характера. В зависимости от задач исследования, могут применяться

различные характеристики объектов, что требует подбора адекватных методов оценки.

Евклидово расстояние наиболее часто используется в качестве метрики кластерного анализа и представляет собой простое геометрическое расстояние, определяемое в многомерном пространстве. С точки зрения геометрии применение данного метода целесообразно в том случае, если для объектов характерно шарообразное скопление.

Квадрат евклидова расстояния. Возведение в квадрат стандартного евклидова расстояния позволяет придать больший вес более отдаленным друг от друга объектам.

Обобщенное степенное расстояние является универсальной метрикой и является значимой характеристикой только с математической точки зрения.

Расстояние Чебышева следует применять только в том случае, что необходимо идентифицировать два объекта как различные в том случае, если они отличаются только по определенному критерию.

Манхэттенское расстояние («расстояние городских кварталов», «хэмминговое» расстояние, «сити-блок» расстояние) рассчитывается как среднее разностей по координатам. В большинстве случаев данная мера расстояния приводит к результатам, аналогичным расчетам евклидова расстояния. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат.

Процент несогласия рассчитывается в том случае, если анализируются категориальные данные.

Основные способы определения близости между объектами представлены в таблице 4.1.2.

Таблица 4.1.2

Основные способы определения меры близости при проведении кластерного анализа

Метрика кластерного анализа	Формула для расчета метрики
Линейное расстояние	$d_{Lij} = \sum_{l=1}^{m} \left| x_i^l - x_j^l \right|$

Евклидово расстояние	$d_{Eij} = (\sum_{l=1}^{m}(x_i^l - x_j^l)^2)^{\frac{1}{2}}$
Квадрат евклидова расстояния	$d^2{}_{Eij} = \sum_{l=1}^{m}(x_i^l - x_j^l)^2$
Обобщенное степенное расстояние	$d_{Pij} = (\sum_{l=1}^{m}(x_i^l - x_j^l)^p)^{\frac{1}{p}}$
Расстояние Чебышева	$d_{ij} = \max_{1\le i,j\le l}\lvert x_i - x_j\rvert$
Манхэттенское расстояние	$d_H(x_i, x_j) = \sum_{l=1}^{k}\lvert x_i^l - x_j^l\rvert$

При осуществлении кластерного анализа важно понимать, что современные методы кластеризации могут основываться на работе как с количественными, так и с неколичественными данными. На формальном уровне единицей анализа является поименованная сущность (объект данных), описываемая произвольным набором элементарных свойств (качеств). Другими словами, сущность определяется как подмножество во множестве свойств / качеств. Свойство, в свою очередь, определяет, посредством своей встречаемости, группу сущностей, и, следовательно, может рассматриваться как подмножество во множестве сущностей. На практике набор данных существует как последовательность записей, каждая из которых описывает один объект. Качества могут принадлежать к различным группам. Эти группы могут служить аналогами переменных («полей» - в терминах баз данных), а качества, им принадлежащие, - значениям переменных. Но группы, с одной стороны, могут иметь более одного значения для каждой записи, а с другой стороны, их существование в общем случае необязательно. Более того, группы качеств могут существовать динамически и приобретать различный смысл в процессе анализа.

4.2. Методические основы кластерного анализа

Существуют различные методы кластерного анализа, применяемые на практике (рис. 4.2.1).

Рис. 4.2.1. Методы кластерного анализа

Иерархические и неиерархические методы кластеризации отличаются применяемыми алгоритмами и подходами. Используя различные методы кластерного анализа, у исследователя появляется возможность получения различных результатов при одинаковом наборе статистических данных.

Сущность иерархической кластеризации состоит в том, что меньшие кластеры объединяются в группы больших размеров последовательно, либо происходит обратный процесс разбиения больших кластеров на меньшие группы.

Для иерархических агломеративных методов (Agglomerative Nesting, AGNES) характерно поэтапное объединение исходных элементов в кластеры. При таком подходе происходит поэтапное сокращение общего числа групп. Алгоритм объединения при использовании

данного метода предполагает соединение объектов в кластеры до тех пор, пока все они не будут составлять одну группу.

Принцип работы иерархических дивизимных методов (DIvisive ANAlysis, DIANA) логически противоположен принципам кластеризации с опорой на агломеративные методы. При использовании дивизимных методов процесс формируется от обратного: предполагается, что изначально объекты принадлежат одному кластеру, затем с каждым последующим шагом производит разбиение на меньшие группы.

Применение иерархических методов кластеризации целесообразно в том случае, если исходный объем характеристик для описания кластеров является относительно небольшим. Очевидным преимуществом данных методов является их наглядность, что особенно удобно при необходимости представления результатов исследования в графическом виде.

В результате реализации иерархических алгоритмов становится возможным построение дендрограмм. Само слово в переводе с греческого означает «дерево». С помощью данного инструмента в древовидной форме отражаются результаты кластеризации.

Дендрограммы (древовидные схемы, деревья объединения, деревья иерархической структуры групп) применяются для характеристики отдельных точек и кластеров по отношению друг к другу и демонстрируют в виде особого графика последовательность осуществления объединений в результате кластеризации. Каждый уровень дендрограммы соответствует конкретному шагу поэтопного укрупнения числа кластерных групп.

Дендрограммы могут быть представлены в вертикальном и горизонтальном виде.

При осуществлении кластеризации важно понимать, каким способом определяется расстояние между объектами и каким образом происходит объединение элементов в группу.

Метод одиночной связи («ближайшего» соседа) предполагает наиболее близкое расположение друг по отношению к другу объектов в кластерах по сравнению с соответствующим расстоянием связи (4.2.1):

$$\rho_{min}(K_i, K_j) = \min_{x_i \epsilon K_i, \quad x_j \epsilon K_j} \rho(x_i, x_j) \quad (4.2.1)$$

Метод ближайшего соседа применяется при построении кластеров, которые, как правило, связаны между собой не системными связями, а отдельными элементами, оказавшимися на минимальном расстоянии друг от друга.

Методом, противоположным данному, **является метод полной связи («дальнего» соседа)**. Его реализация предполагает оценку межкластерных расстояний по величине, характеризующей наиболее отдаленное положение всех остальных пар объектов друг от друга (4.2.2):

$$\rho_{min}(K_i, K_j) = \max_{x_i \epsilon K_i, \quad x_j \epsilon K_j} \rho(x_i, x_j) \quad (4.2.2)$$

Метод Варда был открыт в 1963 году. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы дисперсионного анализа.

Каждый шаг реализации кластерного алгоритма предполагает объединять такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. В результате реализации данного метода происходит создание малых групп, а основой реализации способа является объединение близко расположенных кластеров

Метод невзвешенного попарного среднего (метод невзвешенного попарного арифметического среднего) предполагает, что в качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если объекты существенно отличаются друг от друга, в случаях присутствия кластеров «цепочного» типа, а также при предположении неравных размеров кластеров.

Метод взвешенного попарного среднего (метод взвешенного попарного арифметического среднего) отличается от предыдущего рассмотренного метода тем, что численность объектов кластера в нем является весовой характеристикой. В остальном алгоритм реализации

метода аналогичен. Способ взвешенного попарного арифметического среднего следует использовать в том случае, если выдвинута гипотеза о различном размере предполагаемых кластерных групп.

Невзвешенный центроидный метод (метод невзвешенного попарного центроидного усреднения) предполагает использование в качестве расстояния между двумя кластерами расстояние между центрами тяжести соответствующих групп.

Взвешенный центроидный метод (метод взвешенного попарного центроидного усреднения). Этот метод похож на предыдущий, разница состоит лишь в том, что для учета разницы между размерами кластеров (число объектов в них) используются веса. Следует использовать в том случае, если выдвинута гипотеза о существенных различиях в размере предполагаемых кластерных групп.

Помимо иерархических методов, на практике часто используются **итерационные приемы кластеризации**.

Наиболее популярным из них является **метод k-средних**. Впервые полный алгоритм быстрой кластеризации был рассмотрен в работе 1978 года Хартигана и Вонга (Hartigan and Wong). Отличие метода k-средних от иных иерархических методов состоит в необходимости выдвижения гипотезы о числе кластеров до начала проведения анализа.

Метод k-средних имеет свои преимущества и недостатки. К «плюсам» можно отнести:

- относительная простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

Однако существуют и недостатки:

- алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм *k*-медианы;
- алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

Алгоритм реализации метода можно представить следующими этапами:

- 1 этап - выбирается число кластеров *k*;

- 2 этап - из исходного множества данных случайным образом выбираются *k* записей, которые будут служить начальными центрами кластеров;

- 3 этап - для каждой записи исходной выборки определяется ближайший к ней центр кластера. При этом записи, «притянутые» определенным центром, образуют начальные кластеры;

- 4 этап - вычисляются центроиды – центры тяжести кластеров. Каждый центроид – это вектор, элементы которого представляют собой средние значения признаков, вычисленные по всем записям кластера. Затем центр кластера смещается в его центроид.

Суть принципа реализации алгоритма k-средних состоит в следующем: происходит построение кластеров, расстояние между которыми является наибольшим. Выбор числа кластерных групп k может опираться на предыдущие результаты анализа, теоретические положения в конкретной предметной области и т.д.

Процесс итерации прекращается, когда границы кластеров не перестанут изменяться от итерации к итерации, т.е. на каждой итерации в каждом кластере будет оставаться один и тот же набор записей.

После получений результатов кластерного анализа методом *k*-средних необходимо проверять адекватность проведенной процедуры. Суть такой проверки состоит в оценке значимости различий вычисленных кластерных групп. Анализ различий основывается на расчете средних значений групп. Если процедура дала качественные результаты, средние характеристики каждого из выявленных кластеров должны существенно отличаться друг от друга.

4.3. Кластерный анализ в MS Excel

Проведение кластерного анализа возможно в данном программном комплексе, однако, на наш взгляд не является самым простым.

Рассмотрим особенности проведения кластерного анализа на примере.

Пример

Необходимо выполнить классификацию объектов X_1, X_2, X_3, X_4, X_5, X_6, X_7 иерархическим методом. Исходные данные представлены в таблице 4.3.1.

Таблица 4.3.1

Исходные данные для анализа

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	1507,0	1407,9	1558,0	2422,4	1179,0	1060,3	754,0
2	1508,1	1391,4	1539,2	2397,1	1161,9	1049,6	743,9
3	1511,9	1375,0	1520,1	2374,4	1144,5	1038,6	734,5
4	1513,9	1360,2	1509,6	2367,4	1131,0	1031,7	723,1
5	1511,7	1344,1	1497,6	2364,9	1116,7	1027,6	712,0
6	1511,7	1327,7	1486,5	2360,9	1101,8	1023,3	699,8
7	1514,2	1312,7	1475,9	2353,8	1089,8	1020,2	689,8
8	1520,1	1303,3	1466,8	2344,4	1081,1	1017,7	683,4
9	1526,3	1294,3	1457,9	2339,0	1074,3	1015,6	677,8
10	1531,8	1286,5	1449,8	2334,9	1067,8	1015,0	672,9
11	1532,4	1275,3	1441,1	2334,8	1060,1	1009,2	666,4
12	1536,1	1264,4	1431,9	2331,5	1054,0	1008,2	661,8
13	1541,0	1253,6	1421,7	2330,4	1049,0	1005,6	658,9
14	1544,1	1242,6	1413,3	2328,9	1043,1	1004,6	656,4
15	1547,9	1233,0	1405,6	2331,1	1036,9	1010,5	654,4
16	1550,1	1225,8	1397,2	2333,5	1029,8	1009,8	651,5
17	1552,9	1220,5	1389,6	2335,4	1023,2	1014,6	648,2
18	1549,9	1211,0	1378,3	2333,8	1014,6	1012,2	643,3
19	1547,4	1200,2	1365,8	2327,8	1004,2	1009,4	637,2
20	1549,2	1192,5	1358,4	2324,2	997,1	1002,6	633,4
21	1541,3	1182,7	1342,1	2305,6	987,0	1001,0	628,4
22	1531,9	1168,8	1323,7	2287,7	976,9	1012,8	620,8

Решение

Отметим, что данные, представленные в таблице, не являются нормализованными. Соответственно необходимо провести их стандартизацию любым методом. В данном случае данная операция была выполнена путем расчета отношения разницы фактического и среднего значений к величине стандартного отклонения.

Полученные нормализованные данные представлены в таблице 4.3.2.

Таблица 4.3.2

Нормализованные исходные данные

	X1	X2	X3	X4	X5	X6	X7
1	-1,4908	1,8752	1,8519	2,7015	2,0172	2,7735	2,0556
2	-1,4223	1,6405	1,5624	1,8318	1,7153	2,0688	1,7925
3	-1,1858	1,4072	1,2683	1,0514	1,4082	1,3444	1,5476
4	-1,0613	1,1967	1,1067	0,8108	1,1699	0,8900	1,2507
5	-1,1982	0,9677	0,9219	0,7249	0,9175	0,6200	0,9615
6	-1,1982	0,7344	0,7510	0,5874	0,6545	0,3368	0,6438
7	-1,0426	0,5211	0,5877	0,3433	0,4427	0,1326	0,3833
8	-0,6754	0,3873	0,4476	0,0202	0,2892	-0,0320	0,2166
9	-0,2894	0,2593	0,3106	-0,1655	0,1691	-0,1703	0,0707
10	0,0529	0,1484	0,1858	-0,3064	0,0544	-0,2098	-0,0570
11	0,0903	-0,0109	0,0519	-0,3099	-0,0815	-0,5918	-0,2263
12	0,3206	-0,1660	-0,0898	-0,4233	-0,1892	-0,6577	-0,3461
13	0,6256	-0,3196	-0,2469	-0,4611	-0,2774	-0,8289	-0,4216
14	0,8185	-0,4761	-0,3762	-0,5127	-0,3816	-0,8948	-0,4868
15	1,0550	-0,6126	-0,4948	-0,4370	-0,4910	-0,5062	-0,5388
16	1,1920	-0,7150	-0,6241	-0,3545	-0,6163	-0,5523	-0,6144
17	1,3663	-0,7904	-0,7412	-0,2892	-0,7328	-0,2362	-0,7003
18	1,1795	-0,9255	-0,9152	-0,3442	-0,8846	-0,3943	-0,8280
19	1,0239	-1,0792	-1,1076	-0,5505	-1,0682	-0,5787	-0,9869
20	1,1360	-1,1887	-1,2216	-0,6742	-1,1935	-1,0265	-1,0859
21	0,6442	-1,3281	-1,4726	-1,3136	-1,3718	-1,1319	-1,2161
22	0,0591	-1,5258	-1,7559	-1,9290	-1,5501	-0,3547	-1,4141

Далее рассчитаем расстояния между объектами. В данном примере будет исчислено расстояние по формуле Евклидова расстояния. В результате сформирована матрица расстояний (таблица 4.3.3).

Таблица 4.3.3

Матрица расстояний между объектами

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0,0000	8,9100	8,8408	8,5438	8,8567	8,6999	8,8778
X_2	8,9100	0,0000	0,4170	1,8063	0,3413	2,3098	0,6157
X_3	8,8408	0,4170	0,0000	1,8121	0,5016	2,5533	0,9114
X_4	8,5438	1,8063	1,8121	0,0000	1,5833	1,8398	1,4854
X_5	8,8567	0,3413	0,5016	1,5833	0,0000	2,1555	0,4596
X_6	8,6999	2,3098	2,5533	1,8398	2,1555	0,0000	1,8578
X_7	8,8778	0,6157	0,9114	1,4854	0,4596	1,8578	0,0000

Находим пару самых близких объектов по наименьшему расстоянию. В нашем случае это X_2 и X_5. Они имеют самое малое расстояние 0,3413. Объединяем их в кластер А, т.к. число объектов в нем 2, то и весовой коэффициент также 2.

Расстояния от кластера А до всех остальных кластеров (объектов) вычисляются как средние из расстояний от объектов первого кластера до всех остальных. Эти значения заносятся в строку и столбец матрицы расстояний, соответствующие второму объекту из кластера А.

Полученная матрица расстояний представлена в таблице 4.3.4.

Таблица 4.3.4

Матрица расстояний между объектами

	X_1	X_3	X_4	**A**	X_6	X_7
X_1	0	8,8408	8,5438	**8,8834**	8,6999	8,8778
X_3	8,8408	0	1,8121	**0,4593**	2,5533	0,9114
X_4	8,5438	1,8121	0	**1,6948**	1,8398	1,4854
A	**8,8567**	**0,5016**	**1,5833**	**0**	**2,1555**	**0,4596**
X_6	8,6999	2,5533	1,8398	**2,2327**	0	1,8578
X_7	8,8778	0,9114	1,4854	**0,5377**	1,8578	0

Следующий анализ выявил минимальное расстояние между кластером А и X_3. Расстояние равно 0,4593. Таким образом формируем новый кластер Б. Расстояния от кластера Б до всех остальных кластеров

(объектов) вычисляются как средние из расстояний от объектов кластера Б до всех остальных. Эти значения заносятся в строку и столбец матрицы расстояний, соответствующие второму объекту из кластера Б.

Полученная таким образом матрица расстояний представлена в таблице 4.3.5.

Таблица 4.3.5

Матрица расстояний между объектами

	X_1	X_4	**Б**	X_6	X_7
X_1	0	8,8408	**8,8621**	8,6999	8,8778
X_4	8,5438	0	**1,7535**	1,8398	1,4854
Б	**4,6501**	**1,7535**	**0**	**2,393**	**0,7245**
X_6	8,6999	1,8398	**2,393**	0	1,8578
X_7	8,8778	1,4854	**0,7245**	1,8578	0

Следующий анализ выявил минимальное расстояние между кластером Б и X_7. Расстояние равно 0,7245. Таким образом формируем новый кластер В. Расстояния от кластера В до всех остальных кластеров (объектов) вычисляются как средние из расстояний от объектов кластера В до всех остальных. Эти значения заносятся в строку и столбец матрицы расстояний, соответствующие второму объекту из кластера В.

Полученная таким образом матрица расстояний представлена в таблице 4.3.6.

Таблица 4.3.6

Матрица расстояний между объектами

	X_1	X_4	X_6	**В**
X_1	0	8,8408	8,6999	**8,8699**
X_4	8,5438	0	1,8398	**1,6194**
X_6	8,6999	1,8398	0	**2,1254**
В	**8,8699**	**1,6194**	**2,1254**	**0**

Следующий анализ выявил минимальное расстояние между кластером В и X_4. Расстояние равно 1,6194. Таким образом формируем новый кластер Г. Расстояния от кластера Г до всех остальных кластеров

(объектов) вычисляются как средние из расстояний от объектов кластера Г до всех остальных. Эти значения заносятся в строку и столбец матрицы расстояний, соответствующие второму объекту из кластера Г.

Полученная таким образом матрица расстояний представлена в таблице 4.3.7.

Таблица 4.3.7

Матрица расстояний между объектами

	X_1	X_6	**Г**
X_1	0	8,6999	**8,8554**
X_6	8,6999	0	**1,7296**
Г	**8,8554**	**1,7296**	**0**

Следующий анализ выявил минимальное расстояние между кластером Г и X_6. Расстояние равно 1,7296. Таким образом формируем новый кластер Д. Расстояния от кластера Д до всех остальных кластеров (объектов) вычисляются как средние из расстояний от объектов кластера Д до всех остальных. Эти значения заносятся в строку и столбец матрицы расстояний, соответствующие второму объекту из кластера Д.

Полученная таким образом матрица расстояний представлена в таблице 4.3.8.

Таблица 4.3.8

Матрица расстояний между объектами

	X_1	**Д**
X_1	0	**8,7776**
Д	**8,7776**	**0**

Таким образом осталась последняя пара, которая объединяется в кластер Е. Таким образом, схему объединения можно представить следующим образом:

X2 X5
X2 X5 X3
X2 X5 X3 X7
X2 X5 X3 X7 X4
X2 X5 X3 X7 X4 X6
X1 X2 X5 X3 X7 X4 X6

Таким образом задача решена.

4.4. Проведение кластерного анализа в Statistica

Рассмотрим возможности программного комплекса на примерах.

Пример

Необходимо выполнить классификацию объектов X_1, X_2, X_3, X_4, X_5, X_6, X_7 иерархическим методом. Исходные данные представлены в таблице 4.4.1.

Таблица 4.4.1

Исходные данные для анализа

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	1507,0	1407,9	1558,0	2422,4	1179,0	1060,3	754,0
2	1508,1	1391,4	1539,2	2397,1	1161,9	1049,6	743,9
3	1511,9	1375,0	1520,1	2374,4	1144,5	1038,6	734,5
4	1513,9	1360,2	1509,6	2367,4	1131,0	1031,7	723,1
5	1511,7	1344,1	1497,6	2364,9	1116,7	1027,6	712,0
6	1511,7	1327,7	1486,5	2360,9	1101,8	1023,3	699,8
7	1514,2	1312,7	1475,9	2353,8	1089,8	1020,2	689,8
8	1520,1	1303,3	1466,8	2344,4	1081,1	1017,7	683,4
9	1526,3	1294,3	1457,9	2339,0	1074,3	1015,6	677,8
10	1531,8	1286,5	1449,8	2334,9	1067,8	1015,0	672,9
11	1532,4	1275,3	1441,1	2334,8	1060,1	1009,2	666,4
12	1536,1	1264,4	1431,9	2331,5	1054,0	1008,2	661,8
13	1541,0	1253,6	1421,7	2330,4	1049,0	1005,6	658,9
14	1544,1	1242,6	1413,3	2328,9	1043,1	1004,6	656,4
15	1547,9	1233,0	1405,6	2331,1	1036,9	1010,5	654,4
16	1550,1	1225,8	1397,2	2333,5	1029,8	1009,8	651,5
17	1552,9	1220,5	1389,6	2335,4	1023,2	1014,6	648,2
18	1549,9	1211,0	1378,3	2333,8	1014,6	1012,2	643,3
19	1547,4	1200,2	1365,8	2327,8	1004,2	1009,4	637,2
20	1549,2	1192,5	1358,4	2324,2	997,1	1002,6	633,4
21	1541,3	1182,7	1342,1	2305,6	987,0	1001,0	628,4
22	1531,9	1168,8	1323,7	2287,7	976,9	1012,8	620,8

Решение

Отметим, что для стандартизации данных необходимо выполнить **Данные** – **Стандартизировать**. В появившемся окне (рис. 4.4.1) необходимо нажать Переменные и выбрать необходимые (в нашем случае X_1, X_2, X_3, X_4, X_5, X_6, X_7) – рис. 4.4.2.

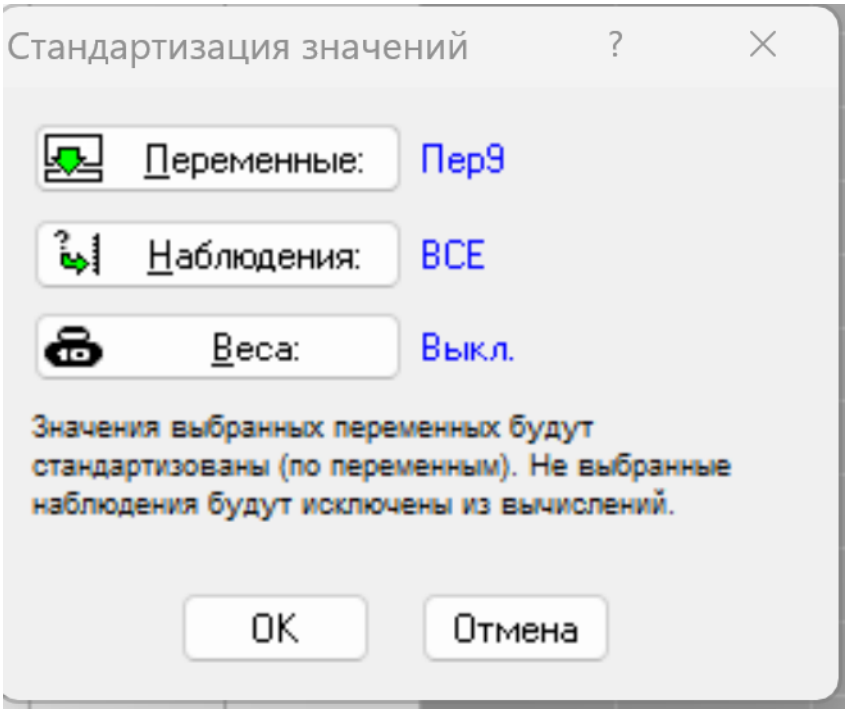


Рис. 4.4.1. Окно команды Стандартизировать

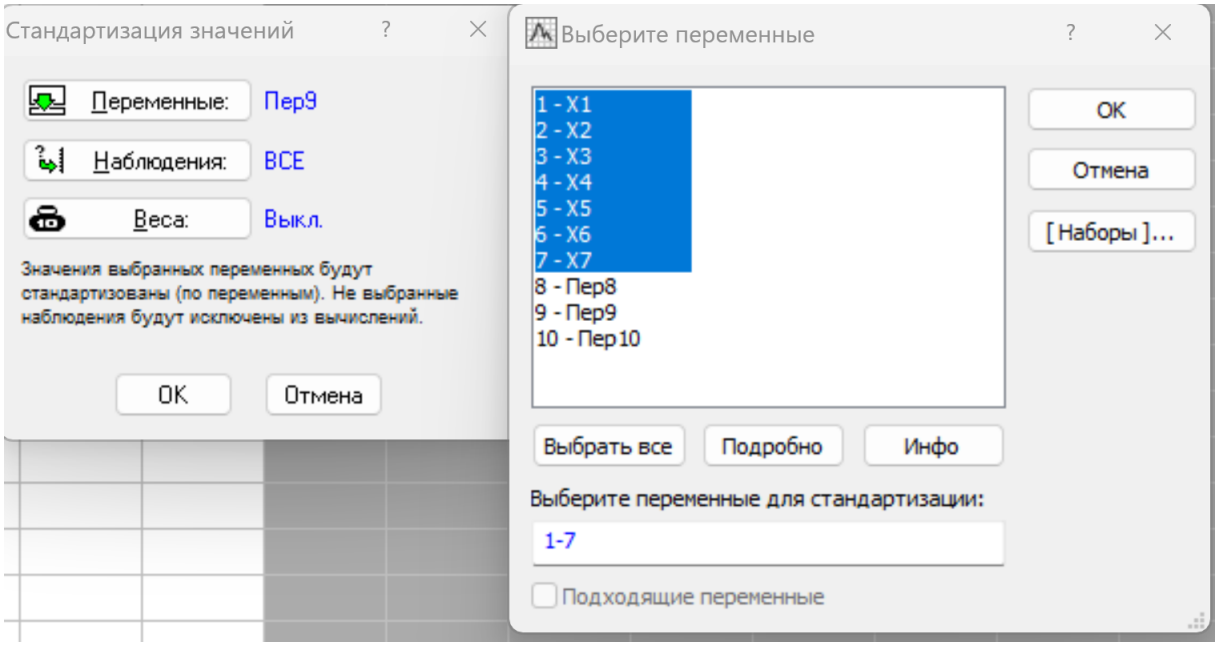


Рис. 4.4.2. Выбор переменных для стандартизации

После нажатия кнопки ОК будет исходные данные будут получена таблица стандартизированных показателей (рис. 4.4.3).

Так по условиям примера необходимо провести классификацию иерархическим методом, то для этого необходимо перейти **Анализ - Многомерный анализ – Кластерный анализ** (рис. 4.4.4).

	1 X1	2 X2	3 X3	4 X4	5 X5	6 X6	7 X7
1	-1,4908	1,8752	1,8519	2,7015	2,0172	2,7735	2,0556
2	-1,4223	1,6405	1,5624	1,8318	1,7153	2,0688	1,7925
3	-1,1858	1,4072	1,2683	1,0514	1,4082	1,3444	1,5476
4	-1,0613	1,1967	1,1067	0,8108	1,1699	0,8900	1,2507
5	-1,1982	0,9677	0,9219	0,7249	0,9175	0,6200	0,9615
6	-1,1982	0,7344	0,7510	0,5874	0,6545	0,3368	0,6438
7	-1,0426	0,5211	0,5877	0,3433	0,4427	0,1326	0,3833
8	-0,6754	0,3873	0,4476	0,0202	0,2892	-0,0320	0,2166
9	-0,2894	0,2593	0,3106	-0,1655	0,1691	-0,1703	0,0707
10	0,0529	0,1484	0,1858	-0,3064	0,0544	-0,2098	-0,0570
11	0,0903	-0,0109	0,0519	-0,3099	-0,0815	-0,5918	-0,2263
12	0,3206	-0,1660	-0,0898	-0,4233	-0,1892	-0,6577	-0,3461
13	0,6256	-0,3196	-0,2469	-0,4611	-0,2774	-0,8289	-0,4216
14	0,8185	-0,4761	-0,3762	-0,5127	-0,3816	-0,8948	-0,4868
15	1,0550	-0,6126	-0,4948	-0,4370	-0,4910	-0,5062	-0,5388
16	1,1920	-0,7150	-0,6241	-0,3545	-0,6163	-0,5523	-0,6144
17	1,3663	-0,7904	-0,7412	-0,2892	-0,7328	-0,2362	-0,7003
18	1,1795	-0,9255	-0,9152	-0,3442	-0,8846	-0,3943	-0,8280
19	1,0239	-1,0792	-1,1076	-0,5505	-1,0682	-0,5787	-0,9869
20	1,1360	-1,1887	-1,2216	-0,6742	-1,1935	-1,0265	-1,0859
21	0,6442	-1,3281	-1,4726	-1,3136	-1,3718	-1,1319	-1,2161
22	0,0591	-1,5258	-1,7559	-1,9290	-1,5501	-0,3547	-1,4141

Рис. 4.4.3. Таблица стандартизированных показателей

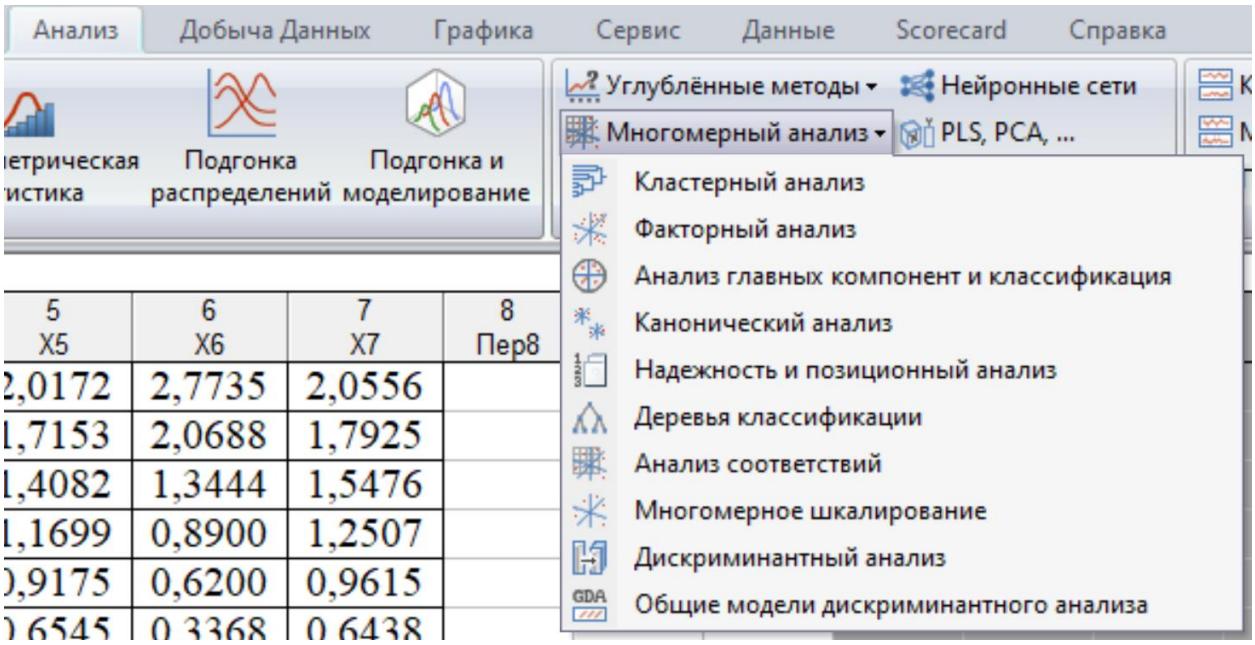


Рис. 4.4.4. Вызов меню кластерного анализа

В появившемся окне необходимо выбрать пункт **Иерархическая классификация** (рис. 4.4.5) и нажать **ОК**.

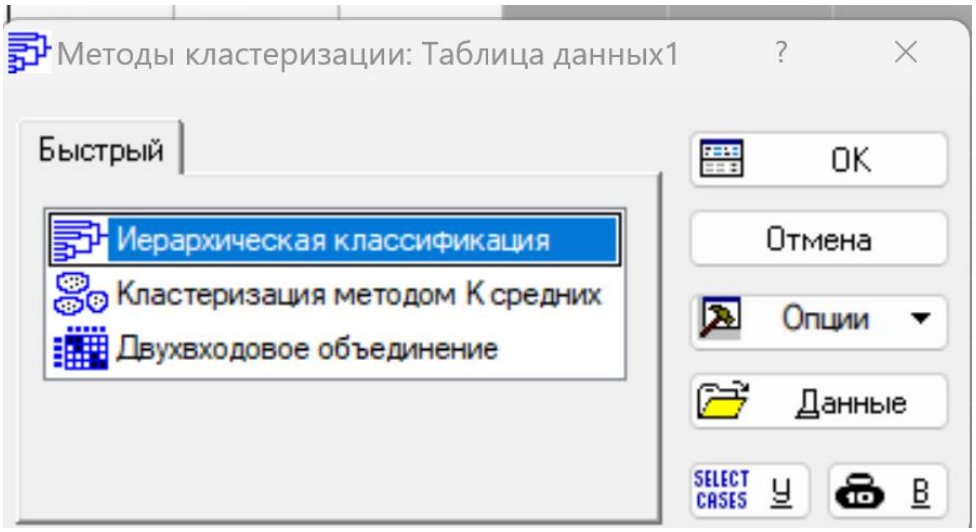


Рис. 4.4.5. Выбор метода кластеризации

В появившемся окне (рис. 4.4.6) необходимо выбрать переменные для проведения анализа.

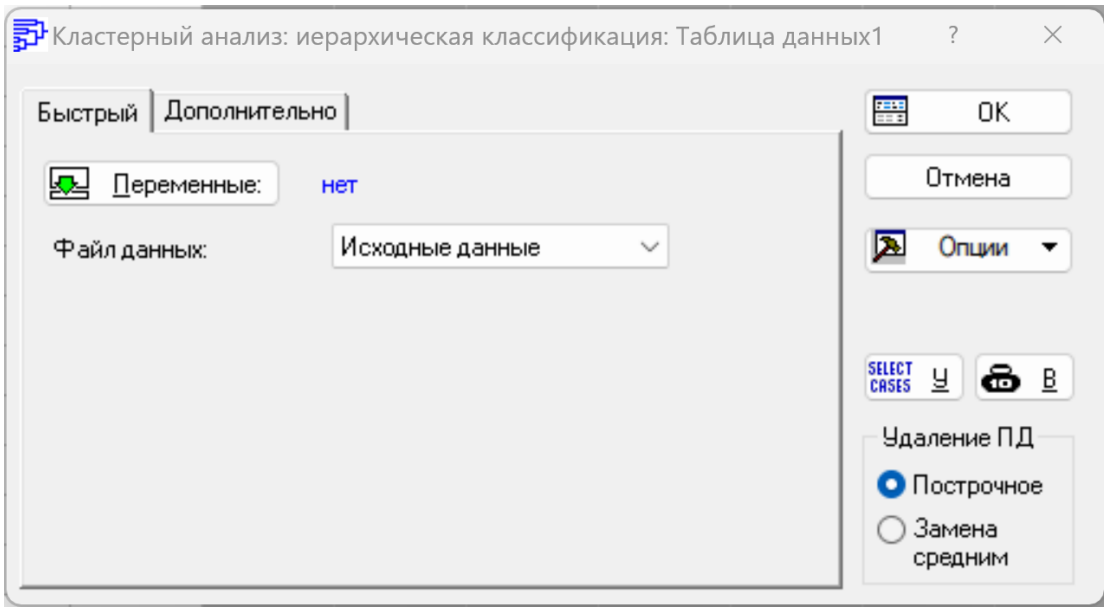


Рис. 4.4.6. Выбор переменных для кластеризации

Для этого на вкладке **Быстрый** необходимо нажать на кнопку **Переменные** и выбрать необходимые. В нашем случае X_1, X_2, X_3, X_4, X_5, X_6, X_7 (рис. 4.4.7).

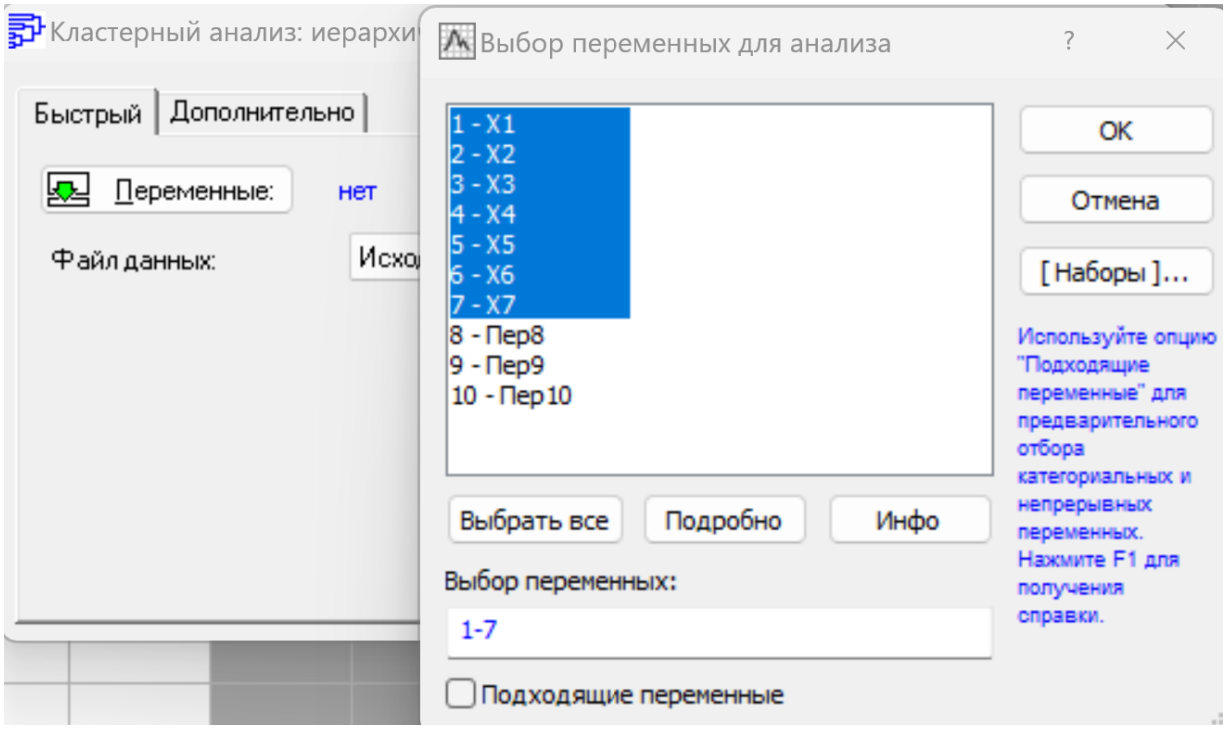


Рис. 4.4.7. Выбор переменных для кластеризации по именам

После нажатия кнопки **ОК** становится возможным перейти на вкладку **Дополнительно** (рис. 4.4.8) и провести настройку проводимого анализа.

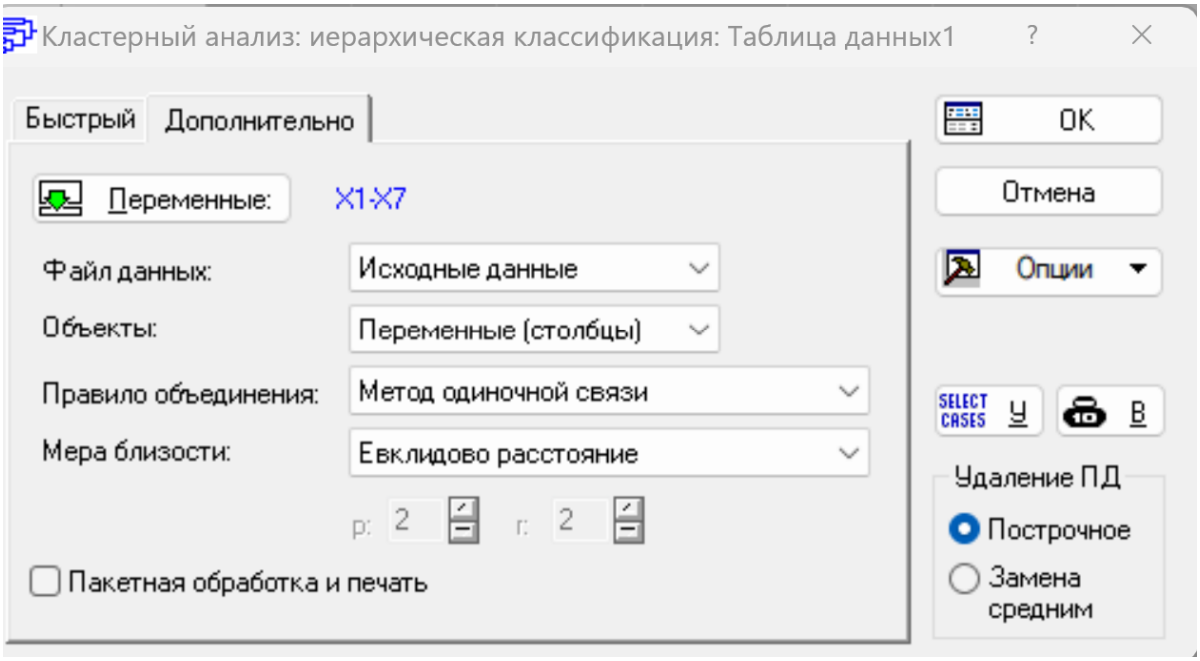


Рис. 4.4.8. Вкладка **Дополнительно**

Среди настаиваемых параметров можно выбрать Правило объединения (рис. 4.4.9). В программном комплексе предоставляется выбор из следующих:

- Метод одиночной связи (по умолчанию);
- Метод полной связи;
- Невзвешенное попарное сравнение;
- Взвешенное попарное сравнение;
- Невзвешенный центроидный метод;
- Взвешенный центроидный метод;
- Метод Варда.

В данном примере выберем **Метод Варда**. Более подробно о данных методах и особенностях их применения представлен материал выше.

Также имеется возможность настройки меры близости (рис. 4.4.10).

Программный комплекс предоставляет на выбор

- Евклидово расстояние (по умолчанию);
- Квадрат евклидова расстояния;

- Расстояние городских кварталов (манхэттенское расстояние);
- Расстояние Чебышева;
- Степенное расстояние;
- Процент несогласия;
- 1-r Пирсона.

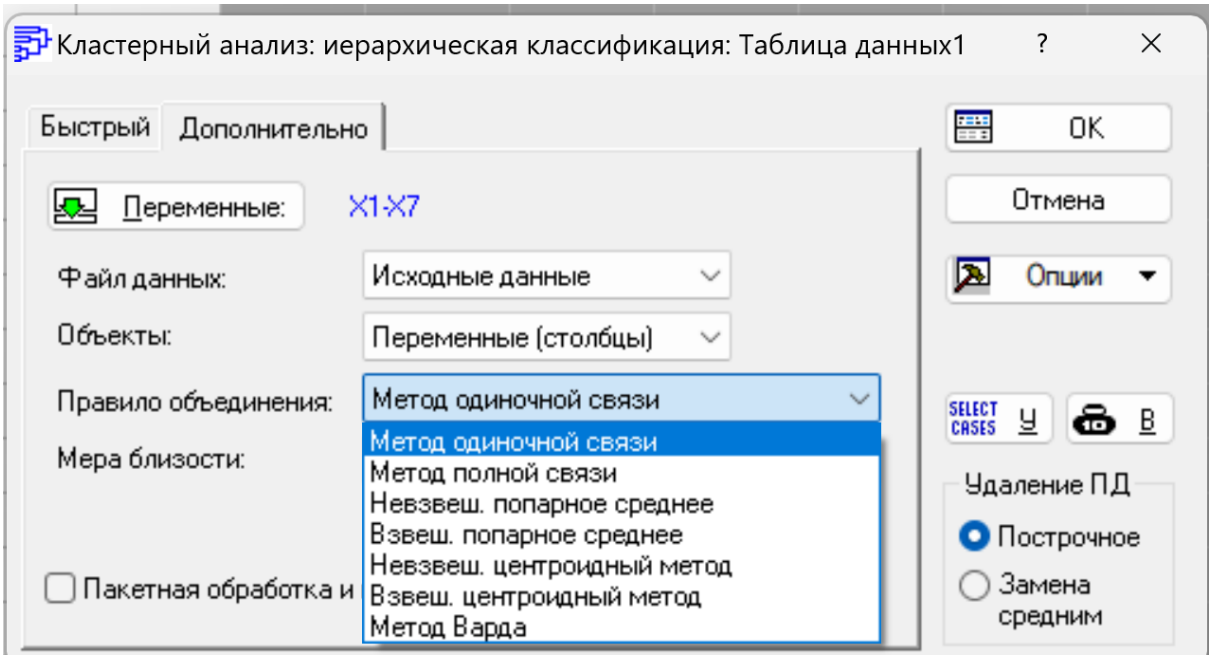


Рис. 4.4.9. Вкладка **Дополнительно.** Выбор правила объединения

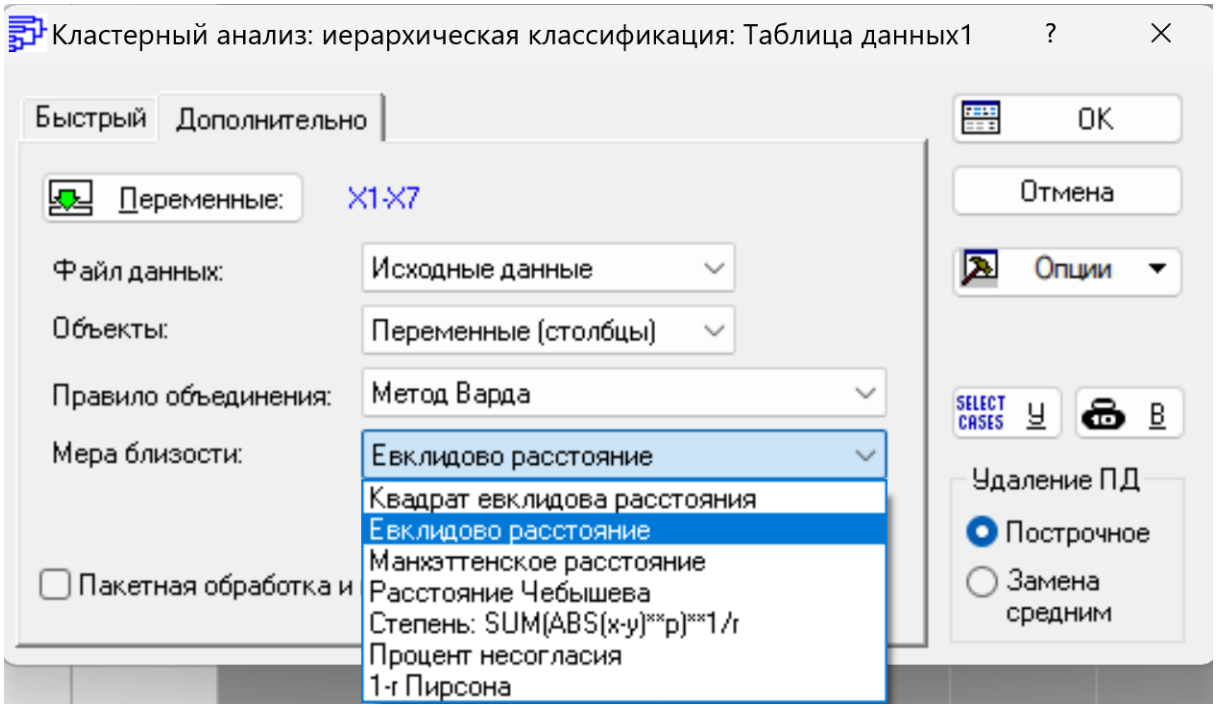


Рис. 4.4.10. Вкладка **Дополнительно.** Выбор меры объединения

Более подробно информация о мерах объединения представлена выше. В данном примере выберем Квадрат евклидова расстояния и нажмем **ОК**. Появится окно результатов иерархической классификации (рис. 4.4.11).

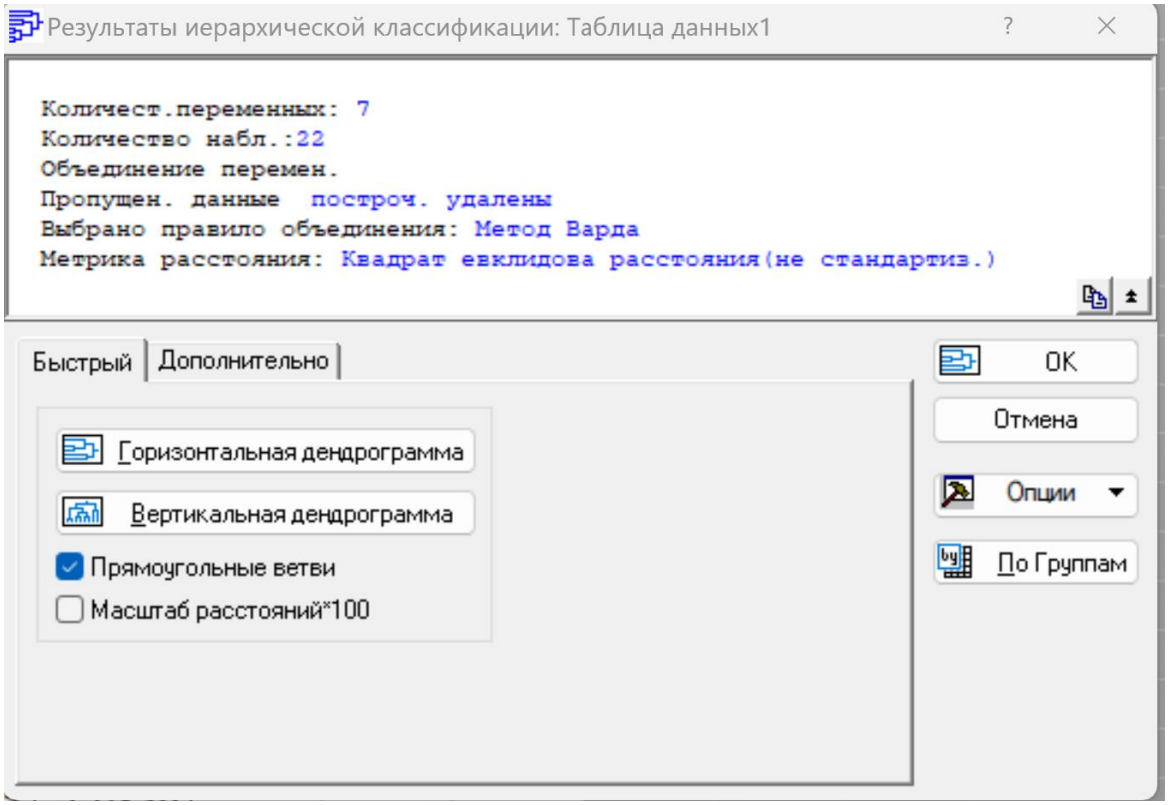


Рис. 4.4.11. Окно результатов иерархической классификации

В верхней части представлены параметры классификации, которые были настроены ранее. Для настройки представления результатов необходимо перейти на вкладку Дополнительно (рис. 4.4.12).

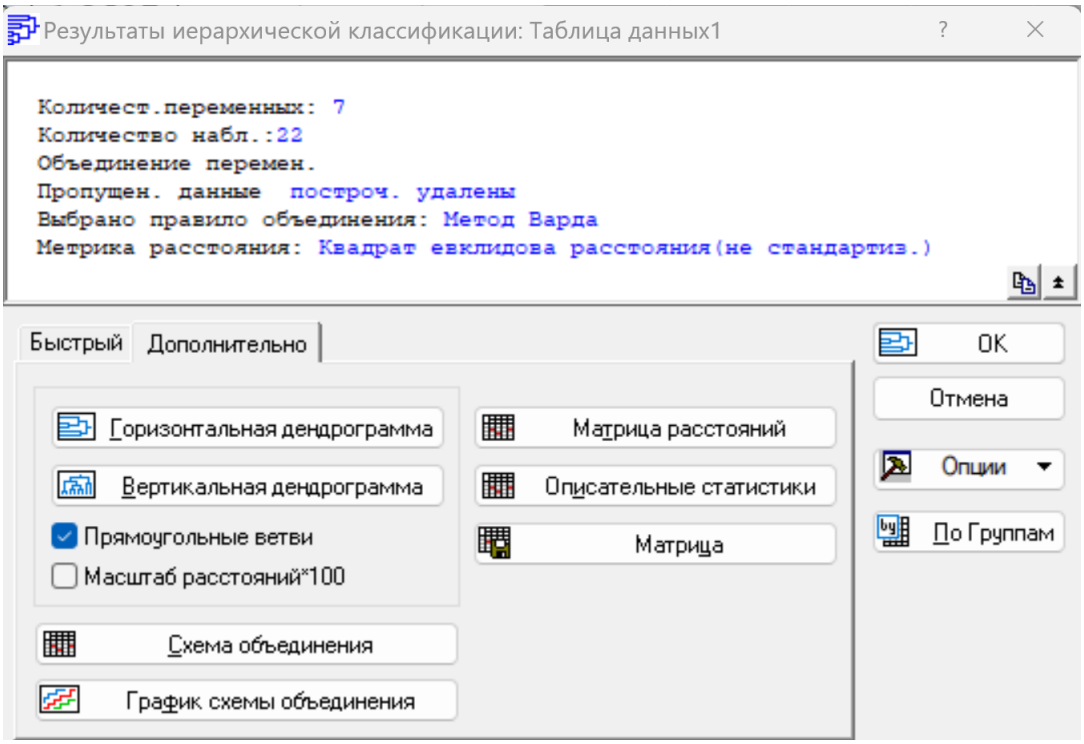


Рис. 4.4.12. Окно результатов иерархической классификации.
Вкладка **Дополнительно**

При нажатии кнопки Горизонтальная дендрограмма будет выведен соответствующий графический элемент (рис. 4.4.13).

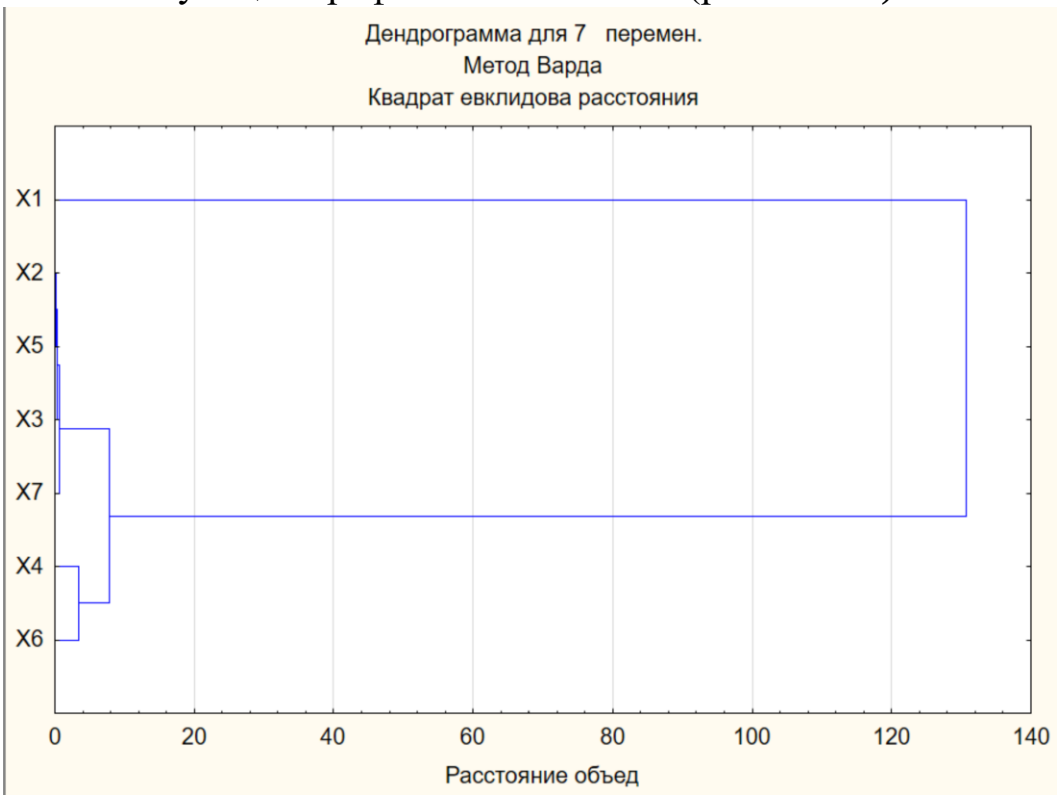


Рис. 4.4.13. Горизонтальная дендрограмма

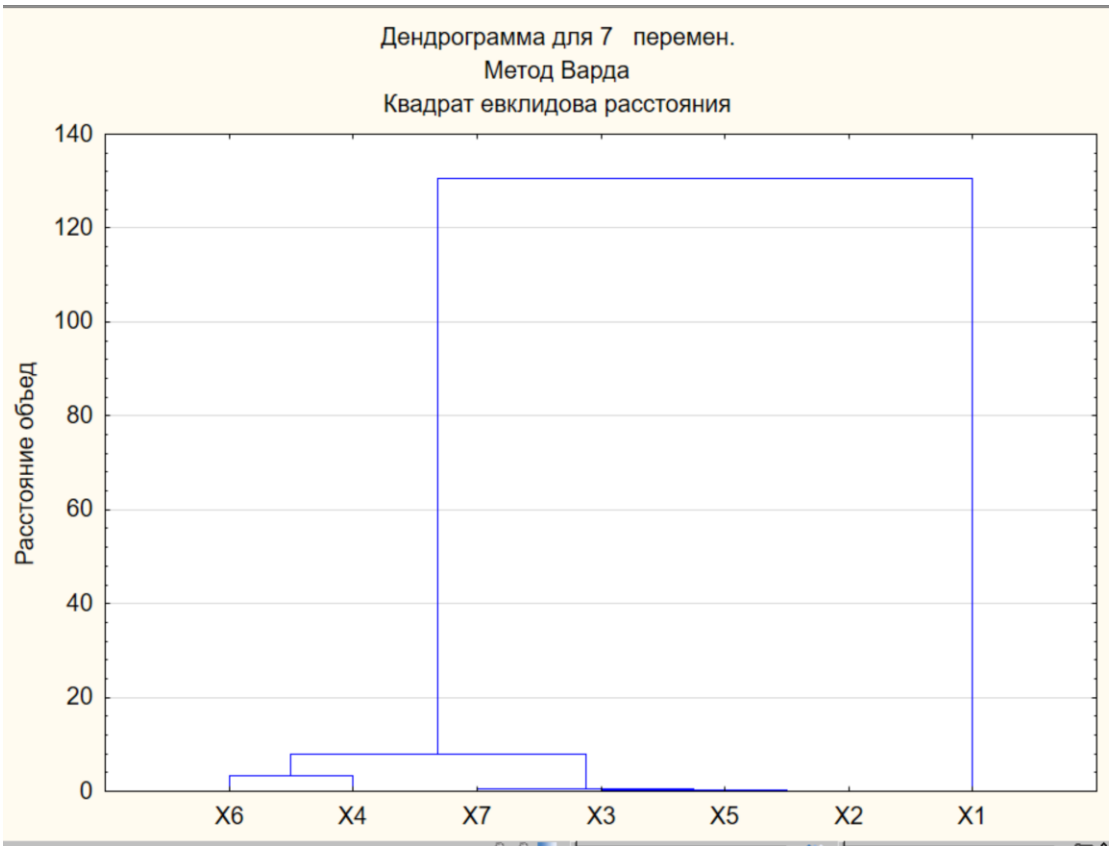


Рис. 4.4.14. Вертикальная дендрограмма

При нажатии кнопки Вертикальная дендрограмма будет выведен соответствующий графический элемент (рис. 4.4.14).

По обеим из них становится возможным определить число кластеров и порядок их объединения.

При нажатии кнопки **Схема объединения** будет выведена табличная форма (рис. 4.4.15).

Схема объединения (Таблица данных1)
Метод Варда
Квадрат евклидова расстояния

расст. объедин.	Объект 1	Объект 2	Объект 3	Объект 4	Объект 5	Объект 6	Объект 7
,1165057	X2	X5					
,2448069	X2	X5	X3				
,6201312	X2	X5	X3	X7			
3,384799	X4	X6					
7,820317	X2	X5	X3	X7	X4	X6	
130,6850	X1	X2	X5	X3	X7	X4	X6

Рис. 4.4.15. **Схема объединения**

В данной форме представлен порядок объединения объектов, а также расстояния объединения.

Кнопка **Марица расстояний** выводит соответствующую табличную форму (рис. 4.4.16).

перемен.	Квадрат евклидова расстояния (Таблица данных1)						
	X1	X2	X3	X4	X5	X6	X7
X1	0,0	79,4	78,2	73,0	78,4	75,7	78,8
X2	79,4	0,0	0,2	3,3	0,1	5,3	0,4
X3	78,2	0,2	0,0	3,3	0,3	6,5	0,8
X4	73,0	3,3	3,3	0,0	2,5	3,4	2,2
X5	78,4	0,1	0,3	2,5	0,0	4,6	0,2
X6	75,7	5,3	6,5	3,4	4,6	0,0	3,5
X7	78,8	0,4	0,8	2,2	0,2	3,5	0,0

Рис. 4.4.16. Марица расстояний

При нажатии кнопки График схемы объединения происходит построение соответствующей графической формы, на которой отмечено объединение по шагам (рис. 4.4.17).

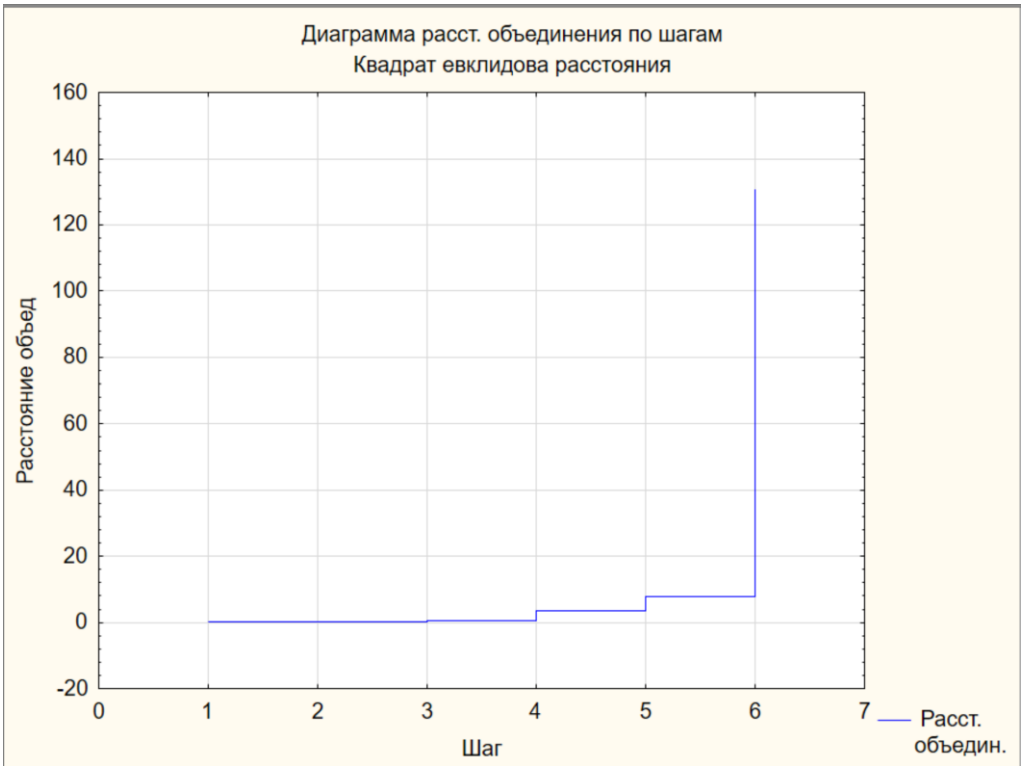


Рис. 4.4.17. График схемы объединения

Таким образом задача классификации иерархическим методом решена.

Пример

Имеются данные о показателях, характеризующих некоторые показатели деятельности 25 объектов.

На основании представленных данных определите кластеры, используя квадрат Евклидова расстояния в качестве меры близости объектов. Сравните результаты иерархической кластеризации и группировки методом k-средних

Решение

Исходные данные представлены в таблицах 4.4.2-4.4.3.

Таблица 4.4.2

Исходные данные для анализа

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	28355,33	29608,58	30104,07	30808,78	32384,35	19422,4	20887,87
2	25400,38	25361,34	26428,4	26611,59	28399,37	18777,76	19248,23
3	23755,73	22875,85	24011,99	23562,54	25383,36	16342,33	16785,77
4	30139,11	29598,57	29356,33	30319,29	32054,02	21243,22	22418,4
5	22582,56	23702,68	24784,76	24527,5	25819,79	15668,65	15891,88
6	27577,55	28620,59	28136,11	29158,13	31425,39	18364,35	19353,33
7	22488,47	23993,97	24769,75	23739,72	25310,29	14609,6	15913,9
8	25839,81	25840,82	26451,43	27302,28	29178,15	17617,6	18561,54
9	27684,66	28483,46	29323,29	30040,01	32511,48	20249,23	21335,31
10	37659,62	40549,51	41327,29	44751,71	47248,2	26416,39	28478,45
11	22862,84	23260,24	24146,12	24919,9	26090,06	16527,51	17261,24
12	24243,22	24598,57	24813,79	25466,44	26912,89	16013	16473,46
13	24787,76	24493,47	25423,4	25913,89	27415,39	17307,29	17198,18
14	25101,08	26195,17	25963,94	26854,83	28182,15	18388,37	19394,38
15	23473,45	23906,88	24101,08	25150,13	27238,21	17147,13	17703,69
16	26312,29	27444,42	27801,77	27235,21	28585,56	18181,16	18657,64
17	27396,37	27846,82	27652,63	27082,06	28686,66	17478,46	17891,87
18	59957,9	59262,2	62594,53	68454,39	74127,05	43841,8	44946,9
19	25759,73	25769,74	26766,74	29179,15	30884,85	18440,42	19308,29
20	33361,33	31558,53	31212,18	33994,96	35391,36	20912,89	20041,02
21	32649,62	32535,5	33194,16	33864,83	35728,69	21838,82	22884,86
22	25627,6	27371,34	26515,49	27008,98	28362,33	15433,42	15905,89
23	25922,9	25928,9	26553,53	27488,46	28933,91	17894,88	19037,02
24	24743,72	27184,16	28026	31372,34	32338,31	18494,48	20253,23
25	36911,88	36151,12	37145,11	41605,56	44281,24	24927,9	25424,4

Таблица 4.4.3

Исходные данные для анализа

	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
1	22095,07	24620,6	26098,07	694072,8	729812,9	786039,4	866845
2	20649,63	22893,87	24494,47	272054,3	281412,1	305564,2	329142,8
3	17830,81	19780,76	20925,91	368857,7	394168,9	413355,6	440983,5
4	23848,83	26556,53	27931,9	806775,6	818100,3	869158,9	944539,2
5	17017	19426,41	20637,62	180698	179074,8	184991,8	198037,6
6	20366,35	23377,35	25085,06	340100,6	372717,4	416382,7	466453,5
7	16676,66	19588,57	20823,8	160740,4	158285,6	167112,2	180467,5
8	19362,34	21587,57	23047,02	337336,4	362756,2	387697,1	428869,7
9	22526,5	24943,92	26884,86	449443,3	484137	506846,8	581084,5
10	30908,88	35234,2	38009,97	3184106	3665962	3783843	4205971
11	18368,35	20237,22	21316,3	208446,1	215571,9	215361,7	230936,9
12	17585,57	19728,71	21214,19	323454,9	334633,4	361292,9	383493,3
13	18056,04	20653,63	22050,03	256963,5	263564,9	291774,9	313169,9
14	20113,09	21785,76	23202,18	317530,9	298037,9	299089,9	331962,8
15	18238,22	20011,99	21631,61	329945,6	361883,7	387912,4	442095,3
16	19997,98	22416,39	23965,94	478015,3	519205,9	557329,6	636769,8
17	19473,45	21335,31	23042,02	443497,2	472816,3	511647,7	561138,5
18	47644,6	54184,13	57164,11	13534384	14251989	15703970	17899398
19	20665,65	23756,73	25762,74	212261,5	231668,9	252087,5	280292,4
20	20922,9	23243,22	24663,64	528931,8	548213,1	576227,8	666401,4
21	24527,5	27689,66	29216,19	628325,8	681162,8	726730,8	820066,2
22	16902,89	19948,93	21041,02	479371,9	477697,6	509276,5	583213
23	20206,19	22856,83	23759,74	350168,4	385884,6	417704,4	461315,8
24	21617,6	24309,29	25860,84	850466,2	917369,1	964767,9	1105540
25	26744,72	30729,7	32542,51	401984,3	432795,2	443052,2	483030,4

На первом шаге решения необходимо выполнить стандартизацию данных. Отметим, что для стандартизации данных необходимо выполнить **Данные** – **Стандартизировать**. В появившемся окне (рис. 4.4.18) необходимо нажать Переменные и выбрать необходимые (в нашем случае X_1 – X_{14}) – рис. 4.4.19.

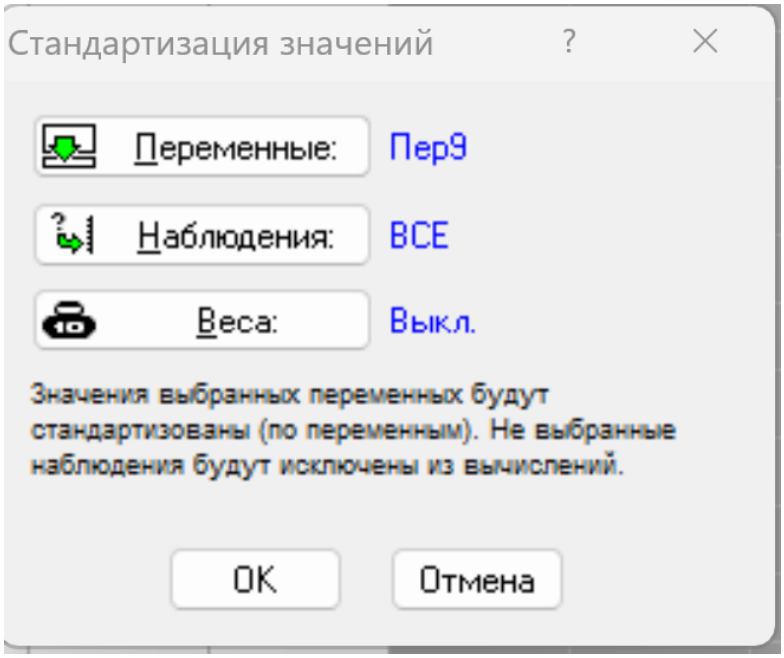


Рис. 4.4.18. Окно команды Стандартизировать

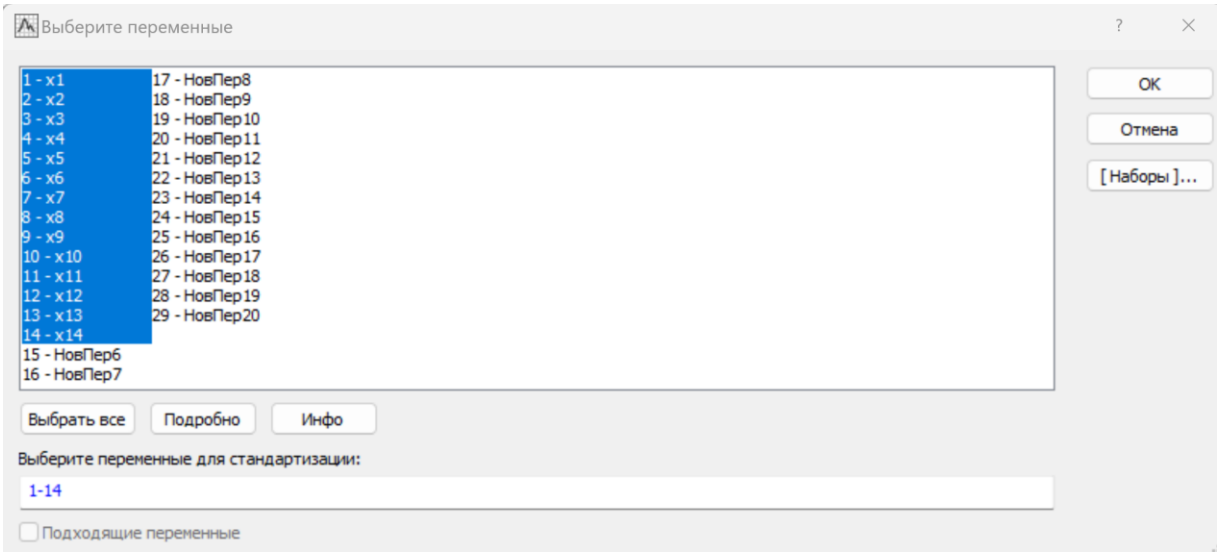


Рис. 4.4.19. Выбор переменных для стандартизации

После нажатия кнопки **ОК** будет исходные данные будут получена таблица стандартизированных показателей (рис. 4.4.20).

	1 x1	2 x2	3 x3	4 x4	5 x5	6 x6	7 x7	8 x8	9 x9	10 x10	11 x11	12 x12	13 x13	14 x14
1	-0,00883	0,095677	0,079671	0,016231	-0,01282	-0,04159	0,073959	0,063505	0,025519	0,010902	-0,13181	-0,13613	-0,13403	-0,13779
2	-0,38998	-0,46647	-0,37796	-0,43115	-0,40398	-0,15367	-0,20316	-0,16529	-0,21305	-0,19921	-0,29008	-0,29519	-0,28901	-0,28999
3	-0,60211	-0,79543	-0,67881	-0,75615	-0,70002	-0,57714	-0,61934	-0,61147	-0,64315	-0,66679	-0,25378	-0,25519	-0,25424	-0,25833
4	0,22125	0,094352	-0,01342	-0,03594	-0,04524	0,275017	0,332636	0,341103	0,292987	0,251183	-0,08954	-0,10481	-0,10722	-0,1158
5	-0,75343	-0,686	-0,5826	-0,65329	-0,65718	-0,69428	-0,77042	-0,74029	-0,69211	-0,70457	-0,32434	-0,33149	-0,3279	-0,3271
6	-0,10915	-0,03509	-0,16535	-0,15971	-0,10695	-0,22556	-0,1854	-0,21013	-0,14625	-0,12183	-0,26456	-0,2628	-0,25327	-0,25112
7	-0,76557	-0,64744	-0,58447	-0,73726	-0,7072	-0,87843	-0,7667	-0,79416	-0,6697	-0,68017	-0,33183	-0,33887	-0,33367	-0,33207
8	-0,3333	-0,403	-0,37509	-0,35753	-0,32753	-0,3554	-0,31922	-0,36905	-0,39352	-0,38887	-0,2656	-0,26634	-0,26252	-0,26176
9	-0,09534	-0,05324	-0,01754	-0,06571	-0,00034	0,102183	0,149582	0,131795	0,070189	0,113992	-0,22355	-0,22328	-0,22409	-0,21868
10	1,191287	1,543762	1,476998	1,502408	1,446195	1,174523	1,356854	1,458621	1,491892	1,571684	0,802059	0,905431	0,832924	0,807358
11	-0,71728	-0,74456	-0,66211	-0,61147	-0,63065	-0,54495	-0,53898	-0,52639	-0,58009	-0,61564	-0,31394	-0,31855	-0,31811	-0,31779
12	-0,53923	-0,56742	-0,57899	-0,55321	-0,54989	-0,63441	-0,67213	-0,65029	-0,65034	-0,62902	-0,2708	-0,27631	-0,27103	-0,27461
13	-0,469	-0,58133	-0,50309	-0,50552	-0,50056	-0,40936	-0,54964	-0,57582	-0,52256	-0,5195	-0,29574	-0,30152	-0,29346	-0,29451
14	-0,42858	-0,3561	-0,43579	-0,40522	-0,4253	-0,22138	-0,17846	-0,25022	-0,36614	-0,36854	-0,27303	-0,28929	-0,2911	-0,28919
15	-0,63852	-0,65897	-0,66772	-0,58692	-0,51795	-0,43721	-0,4642	-0,54699	-0,61121	-0,57433	-0,26837	-0,26665	-0,26245	-0,25802
16	-0,27235	-0,19076	-0,20697	-0,36468	-0,3857	-0,25741	-0,30297	-0,26844	-0,27901	-0,26847	-0,21284	-0,21084	-0,2078	-0,20291
17	-0,13252	-0,1375	-0,22554	-0,381	-0,37578	-0,3796	-0,4324	-0,35147	-0,42838	-0,38952	-0,22578	-0,22729	-0,22254	-0,22432
18	4,06744	4,020476	4,124836	4,028878	4,084579	4,204436	4,14021	4,107677	4,11001	4,081397	4,683842	4,660678	4,677817	4,68332
19	-0,34363	-0,41241	-0,33584	-0,15747	-0,16001	-0,21233	-0,19301	-0,16275	-0,09383	-0,03304	-0,31251	-0,31284	-0,30626	-0,30382
20	0,63687	0,353762	0,217635	0,355846	0,282346	0,21758	-0,06917	-0,12204	-0,16478	-0,17705	-0,19374	-0,20055	-0,20171	-0,19453
21	0,54507	0,483069	0,464397	0,341975	0,315458	0,378579	0,411473	0,448528	0,44954	0,41946	-0,15646	-0,15339	-0,15316	-0,15103
22	-0,36067	-0,20043	-0,36712	-0,38879	-0,40761	-0,73519	-0,76805	-0,75835	-0,61992	-0,65171	-0,21233	-0,22556	-0,2233	-0,21807
23	-0,32258	-0,39135	-0,36238	-0,33768	-0,35151	-0,30719	-0,23886	-0,23548	-0,21816	-0,29548	-0,26078	-0,25813	-0,25284	-0,25258
24	-0,47468	-0,22521	-0,17905	0,076301	-0,01734	-0,20293	-0,0333	-0,01207	-0,01749	-0,02018	-0,07315	-0,06959	-0,07638	-0,07023
25	1,094839	0,961614	0,956303	1,16706	1,154963	0,915706	0,840685	0,799486	0,869551	0,855298	-0,24135	-0,24149	-0,24466	-0,24643

Рис. 4.4.20. Переменные после стандартизации

Так по условиям примера необходимо провести классификацию иерархическим методом, то для этого необходимо перейти **Анализ - Многомерный анализ – Кластерный анализ** и выбрать пункт **Кластеризация методом К-средних** (рис. 4.4.21).

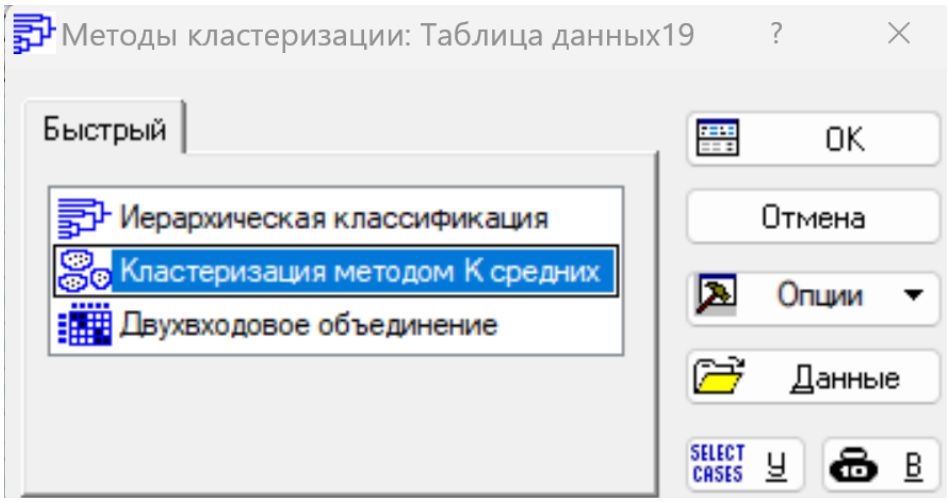


Рис. 4.4.21. Переменные после стандартизации

После нажатия кнопки ОК появится окно настройки анализа. На вкладке Быстрый необходимо в качестве объекта выбрать **Наблюдения** (строки) – это связано с характером таблицы исходных данных (рис. 4.4.22).

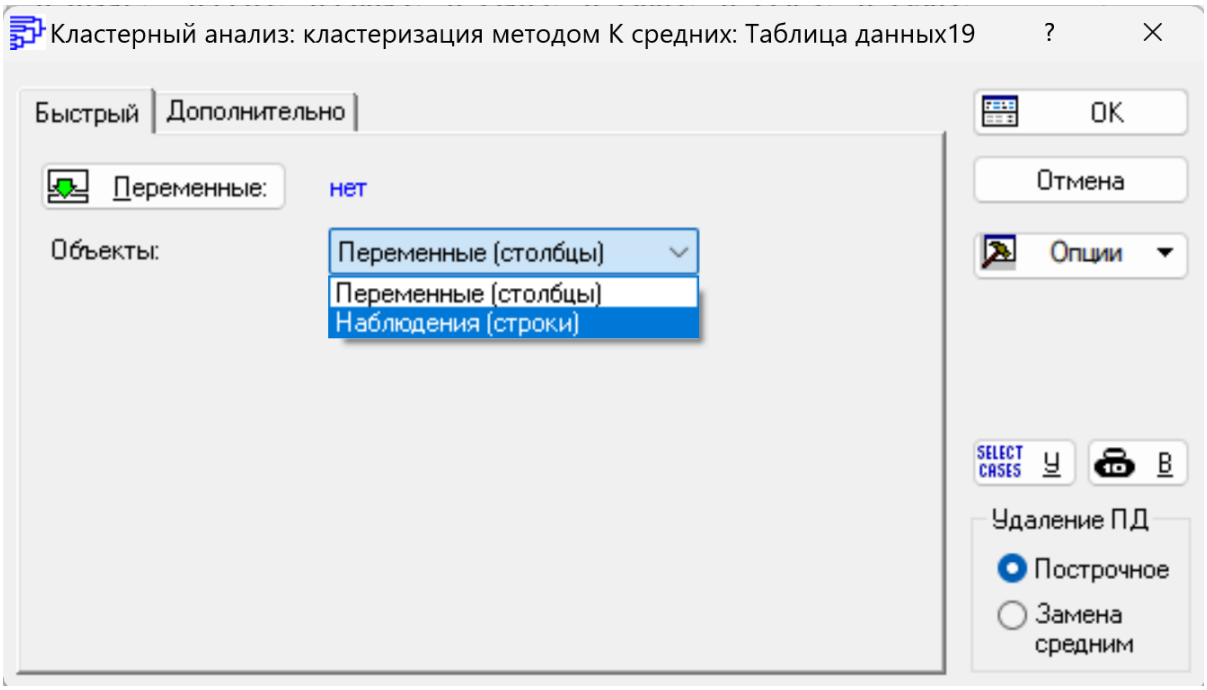


Рис. 4.4.22. Вкладка **Быстрый**

Также необходимо выбрать переменные для анализа – необходимо нажать на кнопку Переменные и выбрать необходимые (рис. 4.4.23).

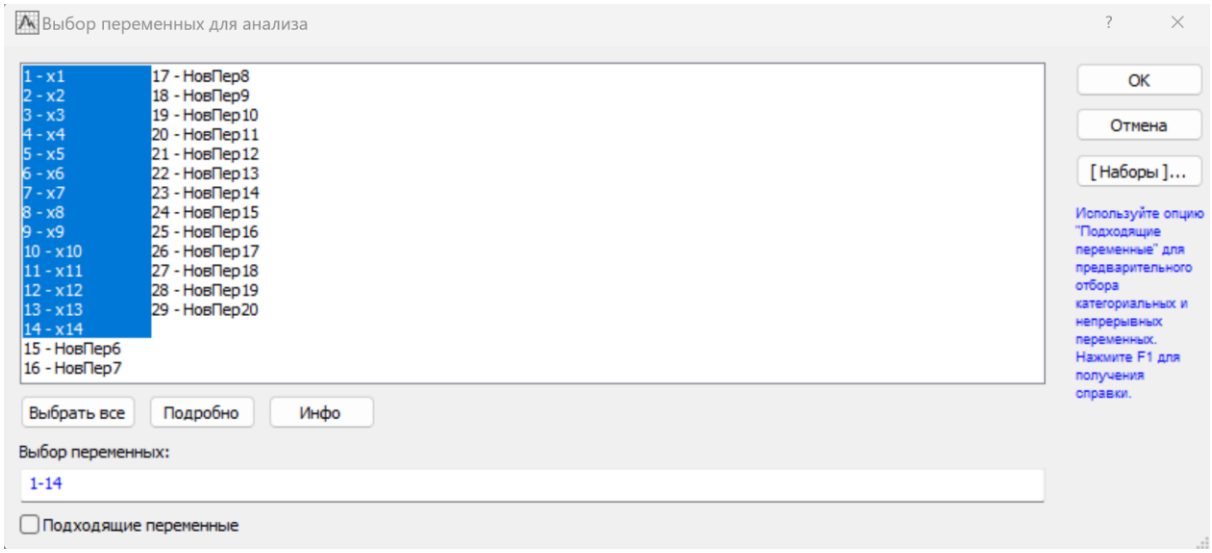


Рис. 4.4.23. Выбор переменных для анализа

На вкладке Дополнительно необходимо настроить анализ (рис. 4.4.24).

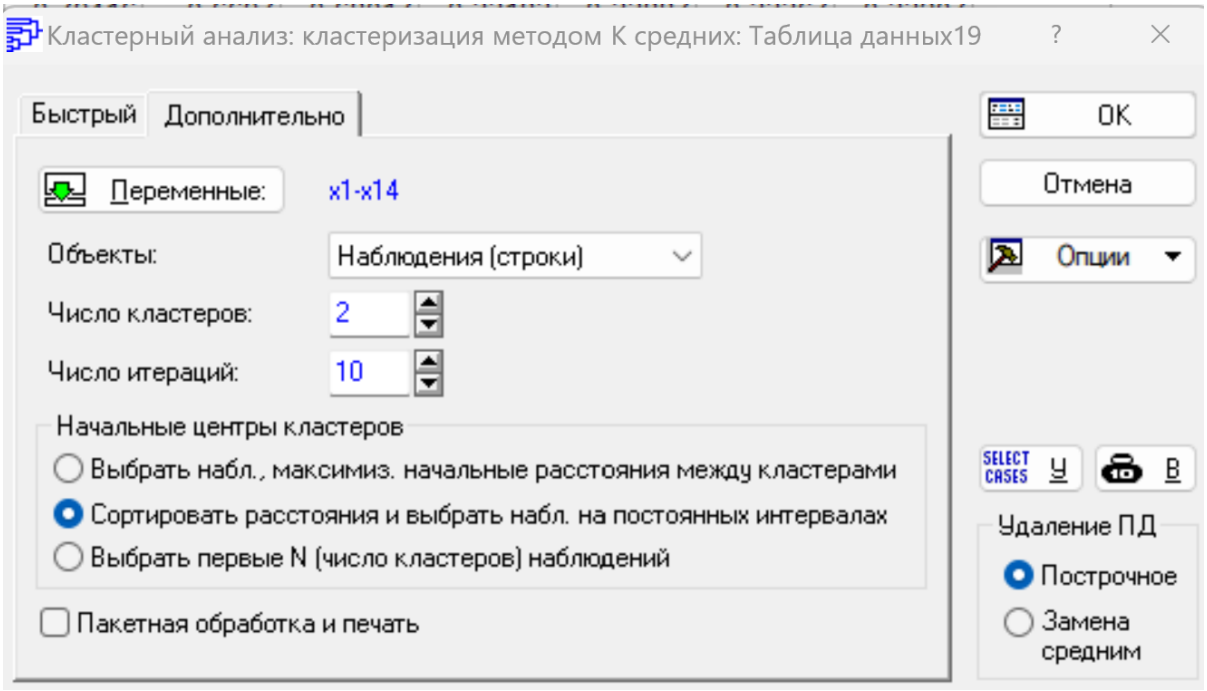


Рис. 4.4.24. Вкладка **Дополнительно**

Оптимальное число групп (интервалов) может быть определено, например, по формуле Стерджесса:

$$K = 1+3.322 \lg n,$$

где n - число единиц совокупности

Для 25 объектов:

$$K = 1+3.322 \lg 25 \approx 6$$

Проверим гипотезу о том, что 25 фирм целесообразно разбить на 6 кластеров методом *k*-средних.

Таким образом, вводим число кластеров, равное 6 и нажимаем кнопку **ОК**. Появляется окно результатов анализа (рис. 4.4.25).

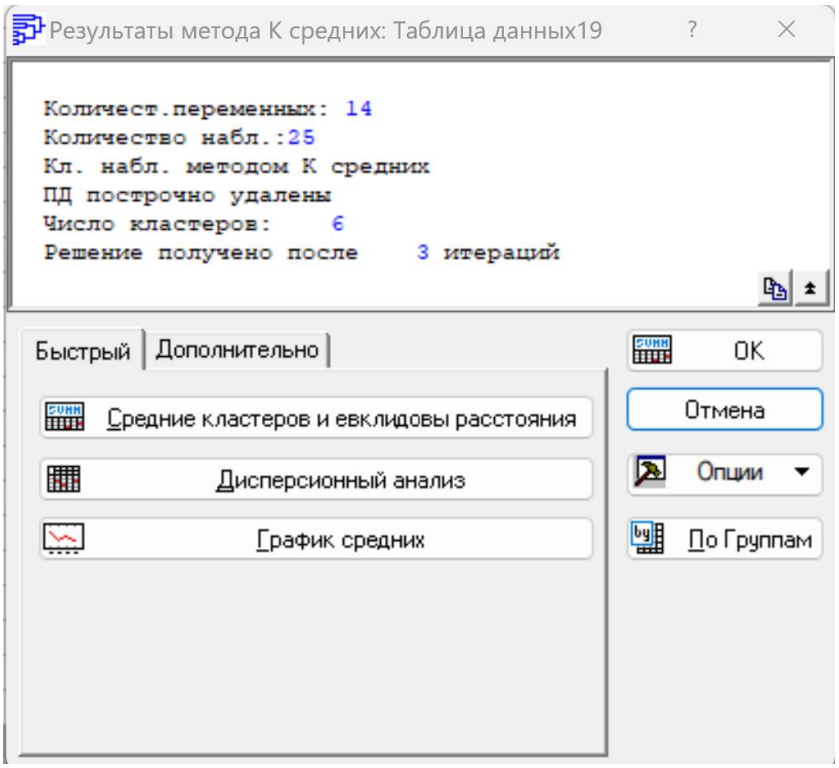


Рис. 4.4.25. Окно результатов анализа

По нажатию на кнопку Средние кластеров и евклидовы расстояния выводится табличная форма разделения 25 объектов на шесть кластеров (рис. 4.4.26).

Кластер Номер	Евклидовы расст. между кластерами (Таблица данных19) Расстояния под диагональю Квадраты расстояний над диагональю					
	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
No. 1	0,000000	11,52278	0,82586	1,60495	2,37857	2,48650
No. 2	3,394523	0,00000	17,74884	20,49991	22,73032	23,05607
No. 3	0,908770	4,21294	0,00000	0,13490	0,41382	0,46697
No. 4	1,266864	4,52768	0,36729	0,00000	0,07810	0,10716
No. 5	1,542260	4,76763	0,64329	0,27947	0,00000	0,01277
No. 6	1,576863	4,80167	0,68335	0,32735	0,11300	0,00000

Рис. 4.4.26. Средние кластеров и евклидовы расстояния

При нажатии на кнопку График средних выводится соответствующий график (рис. 4.4.27).

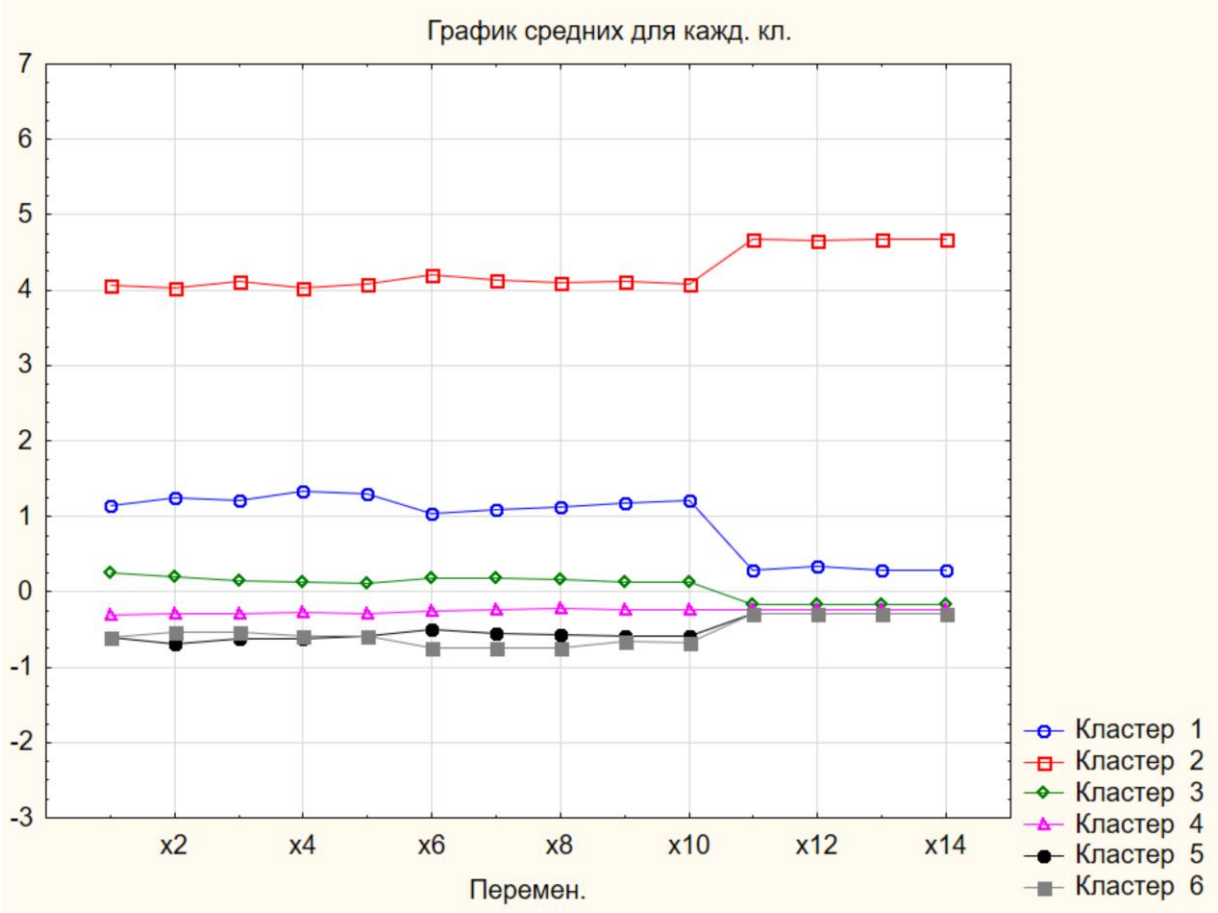


Рис. 4.4.27. График средних

перемен.	Дисперсионный анализ (Таблица данных19)					
	Между SS	cc	Внутри SS	cc	F	значим. p
x1	23,30532	5	0,694675	19	127,4843	0,000000
x2	23,29012	5	0,709881	19	124,6723	0,000000
x3	23,56365	5	0,436352	19	205,2057	0,000000
x4	23,44761	5	0,552387	19	161,3016	0,000000
x5	23,57725	5	0,422747	19	211,9321	0,000000
x6	23,76436	5	0,235638	19	383,2349	0,000000
x7	23,59688	5	0,403122	19	222,4343	0,000000
x8	23,47015	5	0,529848	19	168,3248	0,000000
x9	23,40831	5	0,591688	19	150,3353	0,000000
x10	23,35035	5	0,649651	19	136,5832	0,000000
x11	23,39400	5	0,605995	19	146,6963	0,000000
x12	23,27975	5	0,720249	19	122,8228	0,000000
x13	23,36103	5	0,638969	19	138,9298	0,000000
x14	23,38638	5	0,613615	19	144,8274	0,000000

Рис. 4.4.28. Результаты дисперсионного анализа

Дисперсионный анализ доступен по соответствующей кнопке, при нажатии на которую выводится табличная форма (рис. 4.4.28).

Далее на рисунках 4.4.29.-4.4.34 представлена статистика для каждого из шести выделенных кластеров, которая становится доступной при нажатии кнопки Статистики для каждого кластера на вкладке Дополнительно.

перемен.	Описат.статистики для кластера 1 (Таблица данных19) Кластер содержит 2 набл.		
	Среднее	Стандарт отклон.	Дисперс.
x1	1,143063	0,068199	0,004651
x2	1,252688	0,411641	0,169448
x3	1,216650	0,368187	0,135561
x4	1,334734	0,237127	0,056229
x5	1,300579	0,205932	0,042408
x6	1,045114	0,183011	0,033493
x7	1,098769	0,364987	0,133215
x8	1,129053	0,466079	0,217229
x9	1,180722	0,440061	0,193654
x10	1,213491	0,506561	0,256604
x11	0,280354	0,737803	0,544353
x12	0,331970	0,810996	0,657714
x13	0,294131	0,761969	0,580597
x14	0,280463	0,745141	0,555235

Рис. 4.4.29. Описательные статистики для кластера 1

перемен.	Описат.статистики для кластера 2 (Таблица данных19) Кластер содержит 1 набл.		
	Среднее	Стандарт отклон.	Дисперс.
x1	4,067440	0,00	0,00
x2	4,020476	0,00	0,00
x3	4,124836	0,00	0,00
x4	4,028878	0,00	0,00
x5	4,084579	0,00	0,00
x6	4,204436	0,00	0,00
x7	4,140210	0,00	0,00
x8	4,107677	0,00	0,00
x9	4,110010	0,00	0,00
x10	4,081397	0,00	0,00
x11	4,683842	0,00	0,00
x12	4,660677	0,00	0,00
x13	4,677816	0,00	0,00
x14	4,683320	0,00	0,00

Рис. 4.4.30. Описательные статистики для кластера 2

перемен.	Описат.статистики для кластера 3 (Таблица данных19) Кластер содержит 5 набл.		
	Среднее	Стандарт отклон.	Дисперс.
x1	0,259804	0,325320	0,105833
x2	0,194724	0,217834	0,047451
x3	0,146148	0,201909	0,040767
x4	0,122479	0,208831	0,043610
x5	0,107882	0,175537	0,030813
x6	0,186355	0,161886	0,026207
x7	0,179697	0,194378	0,037783
x8	0,172579	0,226333	0,051227
x9	0,134691	0,239765	0,057487
x10	0,123698	0,227482	0,051748
x11	-0,159020	0,052309	0,002736
x12	-0,163630	0,048068	0,002311
x13	-0,164041	0,048145	0,002318
x14	-0,163565	0,042127	0,001775

Рис. 4.4.30. Описательные статистики для кластера 3

перемен.	Описат.статистики для кластера 4 (Таблица данных19) Кластер содержит 9 набл.		
	Среднее	Стандарт отклон.	Дисперс.
x1	-0,311863	0,123810	0,015329
x2	-0,290876	0,148245	0,021976
x3	-0,295997	0,101332	0,010268
x4	-0,279793	0,166458	0,027708
x5	-0,283786	0,148917	0,022176
x6	-0,257275	0,075069	0,005635
x7	-0,231863	0,111648	0,012465
x8	-0,224990	0,107457	0,011547
x9	-0,239537	0,140247	0,019669
x10	-0,231683	0,146240	0,021386
x11	-0,242036	0,070078	0,004911
x12	-0,243591	0,072679	0,005282
x13	-0,240190	0,069249	0,004795
x14	-0,238436	0,070859	0,005021

Рис. 4.4.31. Описательные статистики для кластера 4

перемен.	Описат.статистики для кластера 5 (Таблица данных19) Кластер содержит 4 набл.		
	Среднее	Стандарт отклон.	Дисперс.
x1	-0,606728	0,103641	0,010742
x2	-0,695073	0,094447	0,008920
x3	-0,627933	0,083519	0,006975
x4	-0,615013	0,104419	0,010903
x5	-0,587298	0,094725	0,008973
x6	-0,492164	0,081414	0,006628
x7	-0,543042	0,063501	0,004032
x8	-0,565168	0,036933	0,001364
x9	-0,589250	0,051379	0,002640
x10	-0,594065	0,062461	0,003901
x11	-0,282955	0,027003	0,000729
x12	-0,285477	0,029569	0,000874
x13	-0,282064	0,029369	0,000863
x14	-0,282162	0,029283	0,000857

Рис. 4.4.32. Описательные статистики для кластера 5

перемен.	Описат.статистики для кластера 6 (Таблица данных19) Кластер содержит 4 набл.		
	Среднее	Стандарт отклон.	Дисперс.
x1	-0,604727	0,193079	0,037279
x2	-0,525324	0,222153	0,049352
x3	-0,528293	0,107474	0,011551
x4	-0,583138	0,149824	0,022447
x5	-0,580469	0,132614	0,017587
x6	-0,735577	0,103839	0,010782
x7	-0,744324	0,048156	0,002319
x8	-0,735775	0,061227	0,003749
x9	-0,658018	0,030601	0,000936
x10	-0,666366	0,032962	0,001086
x11	-0,284826	0,055448	0,003074
x12	-0,293060	0,052953	0,002804
x13	-0,288977	0,052114	0,002716
x14	-0,287963	0,053355	0,002847

Рис. 4.4.32. Описательные статистики для кластера 6

Для определения состава кластеров необходимо нажать кнопку Элементы кластеров и расстояния. После этого будут выведены табличные формы состава кластеров (рисунки 4.4.33-4.4.38).

	Элементы кластера номер 1 (Таблица данных19) и расстояния до центра кластера. Кластер содержит 2 набл.
Наблюд.	**объедин.**
C_10	0,357591
C_25	0,357591

Рис. 4.4.33.. Элементы кластера 1

	Элементы кластера номер 2 (Таблица данных19) и расстояния до центра кластера. Кластер содержит 1 набл.
Наблюд.	**объедин.**
C_18	0,00

Рис. 4.4.34 Элементы кластера 2

	Элементы кластера номер 3 (Таблица данных19) и расстояния до центра кластера. Кластер содержит 5 набл.
Наблюд.	**объедин.**
C_1	0,123921
C_4	0,119621
C_9	0,143877
C_20	0,205680
C_21	0,225338

Рис. 4.4.35. Элементы кластера 3

Наблюд.	Элементы кластера номер 4 (Таблица данных19) и расстояния до центра кластера. Кластер содержит 9 набл. объедин.
C_2	0,087788
C_6	0,117912
C_8	0,090945
C_14	0,092940
C_16	0,060471
C_17	0,123098
C_19	0,097887
C_23	0,046793
C_24	0,196026

Рис. 4.4.36. Элементы кластера 4

Наблюд.	Элементы кластера номер 5 (Таблица данных19) и расстояния до центра кластера. Кластер содержит 4 набл. объедин.
C_3	0,071614
C_11	0,044085
C_13	0,077796
C_15	0,039159

Рис. 4.4.37. Элементы кластера 5

Наблюд.	Элементы кластера номер 6 (Таблица данных19) и расстояния до центра кластера. Кластер содержит 4 набл. объедин.
C_5	0,072111
C_7	0,091280
C_12	0,050461
C_22	0,141546

Рис. 4.4.38. Элементы кластера 6

Таким образом задача классификации решена. Были определены составы всех кластеров и выявлено, что кластер 4 является самым многочисленным и включает 9 объектов.

Контрольные вопросы по теме

1. Что такое кластерный анализ?
2. Какие основные типы кластерного анализа существуют?
3. В чем отличие между иерархическим и непараметрическим кластерным анализом?
4. Какие предпосылки должны быть выполнены для применения кластерного анализа?
5. Как проводится иерархический кластерный анализ?
6. Как проводится непараметрический кластерный анализ?
7. Как интерпретировать результаты кластерного анализа?
8. Как выбрать оптимальное количество кластеров при проведении кластерного анализа?
9. Как можно оценить качество разбиения на кластеры при проведении кластерного анализа?
10. Какие методы и метрики используются для измерения сходства между объектами при проведении кластерного анализа?
11. Как можно использовать результаты кластерного анализа для сегментации рынка?
12. Как можно использовать результаты кластерного анализа для выявления типов поведения потребителей или клиентов компании?
13. Как можно использовать результаты кластерного анализа для определения профилей потребителей или клиентов компании?
14. Как можно использовать результаты кластерного анализа для оптимизации маркетинговых стратегий?
15. Как можно использовать результаты кластерного анализа для улучшения обслуживания клиентов компании?
16. Как можно использовать результаты кластерного анализа для выявления внутренних закономерностей в данных о клиентах или потребителях компании?
17. Как можно использовать результаты кластерного анализа для оптимизации производственных процессов?
18. Как можно использовать результаты кластерного анализа для выявления групп схожих объектов или явлений в данных?
19. Как можно использовать результаты кластерного анализа для прогнозирования тенденций или изменений в данных?
20. Как можно использовать результаты кластерного анализа для выявления внутренних структур или закономерностей в данных?

21. Как можно использовать результаты кластерного анализа для выявления важных факторов, влияющих на зависимую переменную?

22. Как можно использовать результаты кластерного анализа для определения групп риска или возможностей в данных?

23. Как можно использовать результаты кластерного анализа для оптимизации управленческих процессов в компании?

24. Как можно использовать результаты кластерного анализа для выявления внутренних связей между объектами или явлениями?

25. Как можно использовать результаты кластерного анализа для определения оптимальных стратегий действий в различных ситуациях?

26. Как можно использовать результаты кластерного анализа для выявления воздействия различных факторов на изучаемый процесс или явление?

27. Какие программные инструменты чаще всего используются для проведения кластерного анализа?

28. Как можно использовать результаты кластерного анализа для оценки эффективности различных методов или стратегий?

29. Как можно использовать результаты кластерного анализа для выявления тенденций или закономерностей в данных?

30. Как можно использовать результаты кластерного анализа для прогнозирования будущих значений зависимой переменной?

31. Как можно использовать результаты кластерного анализа для сравнения различных групп или образцов данных?

32. Как можно использовать результаты кластерного анализа для определения оптимальных условий или параметров процессов?

33. Что такое метод "k-средних" и как он используется при проведении кластерного анализа?

34. Что такое метод "агломеративная иерархическая кластеризация" и как он используется при проведении кластерного анализа?

35. Что такое метод "DBSCAN" и как он используется при проведении кластерного анализа?

36. Что такое метод "OPTICS" и как он используется при проведении кластерного анализа?

37. Что такое метод "Mean Shift" и как он используется при проведении кластерного анализа?

38. Что такое метод "EM-алгоритм" и как он используется при проведении кластерного анализа?

39. Что такое метод "DBCLASD" и как он используется при проведении кластерного анализа?

40. Что такое метод "BIRCH" и как он используется при проведении кластерного анализа?

41. Что такое метод "Spectral Clustering" и как он используется при проведении кластерного анализа?

42. Что такое метод "OPTICS" и как он используется при проведении кластерного анализа?

43. Что такое метод "Affinity Propagation" и как он используется при проведении кластерного анализа?

44. Что такое метод "Gaussian Mixture Model" и как он используется при проведении кластерного анализа?

45. Что такое метод "Self-Organizing Maps (SOM)" и как он используется при проведении кластерного анализа?

46. Что такое метод "Agglomerative Clustering" и как он используется при проведении кластерного анализа?

47. Что такое метод "Ward's Method" и как он используется при проведении иерархического кластерного анализа?

48. Что такое метод "Complete Linkage" и как он используется при проведении иерархического кластерного анализа?

49. Что такое метод "Single Linkage" и как он используется при проведении иерархического кластерного анализа?

50. Какие методы используются для проверки предпосылок и условий применения кластерного анализа?

5. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

5.1. Основные положения дискриминантного анализа

Задача дискриминантного анализа опирается на следующий набор данных. Допустим, что имеются сведения о n наблюдениях, каждое из которых можно охарактеризоваться по *k* признакам. В таком случае каждое конкретное наблюдение может быть описано с помощью вектора *x*, характеристики которого являются набором случайных величин (5.1.1):

$$X = (X_1, X_2, \ldots X_k)^T \qquad (5.1.1)$$

Суть задачи дискриминации состоит в том, чтобы разбить все множество реализации анализируемой величины на определенное количество областей R_i, затем каждое из новых анализируемых наблюдений относится к какой-либо области опираясь на правило, которое признано в рамках решения конкретной исследовательской задачи решающим. Предполагается, что заранее информация о принадлежности объекта к области недоступна, либо требует значительных затрат ресурсов на ее получение.

Выбор правила осуществления дискриминантного анализа должен базироваться на принципе оптимальности, который представляет собой минимизацию средних потерь от неправильной классификации, исходя из априорных вероятностей p_i извлечения объекта из группы R_i. Решающее правило считается наилучшим в определенном смысле слова, если никакое другое правило не может дать меньшей величины функции потерь.

Значения априорных вероятностей могут быть известны заранее, и определены пользователями заблаговременно (по результатам предварительного анализа), либо заданы в процессе ввода данных модуль. В качестве средних потерь чаще всего принимают вероятность ложной классификации наблюдения.

Построение решающего правила также можно рассматривать как задачу поиска областей R, которые не пересекаются между собой. Дискриминантные функции в этом случае дают определение этих областей путем задания их границ в многомерном пространстве.

В процессе дискриминантного анализа автоматически вычисляются функции классификации, предназначенные для определения той группы, к которой наиболее вероятно принадлежит новый объект. При этом важно равенство количества функций классификации заданной величине кластерных групп.

Считается, что принадлежность наблюдения в определенной группе является объясненной в том случае, если функция классификация максимальна, либо значение апостериорной функции является наибольшим.

Дискриминантный анализ используется для исследования различий заранее заданных групп объектов исследования (фирм, категорий товаров и т.д.). Переменная является группирующей в том случае, если она делит совокупность объектов исследования на конкретные классы

или категории. Дискриминантный анализ используется для того, чтобы оценить межгрупповые развития по определенным признакам. Если признак применяется для выявления различий между группами, они именуются дискриминационными переменными.

Важно грамотное использование шкалирования при анализе: Группирующая переменная должна принадлежать номинальной шкале, а зависимые характеристики быть метрическими. Соблюдение данного условия необходимо для обеспечения высокой точности производимых расчетов. В практической деятельности возможно, чтобы группирующая переменная принадлежала также к порядковой школе, а дискриминационные - к шкале любого типа, однако, надо четко понимать конкретные исследовательские задачи и объекты аналитической практики.

В результате проведения дискриминантного анализа осуществляется построение модели (дискриминантной функции), общий вид которой (5.1.2):

$$D = b_0 + b_i X_1 + b_2 X_2 + ... + b_k X_k \qquad (5.1.2)$$

где D – группирующая (зависимая) переменная;

b_k – коэффициенты дискриминантной функции;

b_o – свободный член (константа);

x_n – дискриминационные (независимые) переменные.

Данная модель позволяет определять принадлежность каждого конкретного объекта к группе, опираясь на базовые характеристики исследования, признанные исследователем как значимые.

Основные цели дискриминантного анализа можно определить как:

- Определение дискриминантных функций или линейных комбинаций независимых переменных, которые наилучшим образом различают (дискриминируют) ка-тегории (группы) зависимой переменной;

- Проверка существования между группами значимых различий с точки зрения независимых переменных;

- Определение предикторов, вносящих наибольший вклад в межгрупповые различия;

- Отнесение случаев к одной из групп (классификация), исходя из значений предикторов;

- Оценка точности классификации данных на группы.

Целесообразность дискриминантного анализа определяется на основе исследовательских задач.

К статистикам, используемым в дискриминантном анализе, относятся следующие коэффициенты и показатели.

Каноническая корреляция используется для измерения степени связи между дискриминантными показателями и группами. Это мера связи между единственной дискриминирующей функцией и набором фиктивных переменных, которые определяют принадлежность к данной группе.

Центроид (средняя точка). Центроид – это средние значения для дискриминантных показателей конкретной группы. Центроидов столько, сколько групп, т.е. один центроид для каждой группы. Средние группы для всех функций – это групповые центроиды.

Классификационная матрица. Иногда ее называют смешанной матрицей, или матрицей предсказания. Классификационная матрица содержит ряд правильно классифицированных и ошибочно классифицированных случаев. Верно классифицированные случаи лежат на диагонали матрицы, поскольку предсказанные и фактические группы одни и те же. Элементы, не лежащие по диагзнали матрицы, представляют случаи, классифицированные ошибочно. Сумма элементов, лежащих на диагонали, деленная на общее количество случаев, дает коэффициент результативности.

Коэффициенты дискриминантной функции. Коэффициенты дискриминантной функции (ненормированные) – это коэффициенты переменных, когда они измерены в первоначальных единицах.

Дискриминантные показатели. Сумма произведений ненормированных коэффициентов дискриминантной функции на значения переменных, добавленная к постоянному члену.

Собственное (характеристическое) значение. Для каждой дискриминантной функции собственное значение – это отношение межгрупповой суммы квадратов к внутри-групповой сумме квадратов. Большие собственные значения указывают на функции более высокого порядка.

F – статистика и ее значимость. Значения F -статистики вычисляют однофакторный дисперсионный анализ, разбивая на группы независимую переменную. Каждый предиктор, в свою очередь, служит в ANOVA метрической зависимой переменной.

Средние группы и групповые стандартные отклонения. Эти показатели вычисляют для каждого предиктора каждой группы.

Объединенная межгрупповая корреляционная матрица. Объединенную межгрупповую корреляционную матрицу вычисляют усреднением отдельных ковариационных матриц для всех групп.

Нормированные коэффициенты дискриминантных функций. Коэффициенты дискриминантных функций используют как множители для нормированных переменных, т.е. переменных с нулевым средним и дисперсией, равной 1.

Структурные коэффициенты корреляции. Также известны как дискри-минантные нагрузки, представляют собой линейные коэффициенты корреляции между предикторами и дискриминантной функцией.

Общая корреляционная матрица. Если при вычислении корреляций наблюдения обрабатывают так, как будто они взяты из одной выборки, то в результате получают общую корреляционную матрицу.

Коэффициент λ Уилкса. Иногда называемый U-статистикой, коэффициент λ Уилкса для каждого предиктора – это отношение внутри групповой суммы квадратов к общей сумме квадратов. Его значение варьирует от 0 до 1. Большое значение λ (около 1) указывает на то, что средние групп не должны различаться. Малые значения λ (около 0) указывают на то, что средние групп различаются.

5.2. Общая процедура дискриминантного анализа

Процедура выполнения дискриминантного анализа состоит из шести основных этапов.

- Определение зависимой и независимой переменной (переменных);
- Выбор метода дискриминантного анализа;
- Определение коэффициентов дискриминантной функции;
- Определение значимости дискриминантной функции;
- Интерпретация полученных результатов;
- Оценка достоверности дискриминантного анализа.

На первом шаге анализа определяется, что является независимой переменной, а что результирующим фактором. Зависимая переменная должна состоять из двух или больше взаимоисключающих и взаимно исчерпывающих категорий. Если зависимая переменная измерена с помощью интервальной или относительной шкалы, то ее следует, в первую очередь, перевести к статусу категориальной. Например, отношение к фирме, измеренное по девятибалльной шкале, можно категоризировать как неблагоприятное (1, 2, 3, 4), нейтральное (5) и благоприятное (6, 7, 8, 9). Также возможно формирование одинаковых групп точками отсечек по результатам графического анализа графика распределения значений зависимой переменной. Предикторы следует выбирать, исходя из теоретической модели или ранее проверенного исследования, или, в случае поискового исследования, из интуиции и опыта исследователя.

Далее выборку делят на две части. Одна из них – анализируемая выборка – используется для вычисления дискриминантной функции. Другая часть – проверочная выборка используется для проверки результатов дискриминантного моделирования.

Если исследователь сталкивается с необходимостью работы с большой по объему выборкой, ее можно разбить на две меньшие равные совокупности, первая из которых будет анализируемой, а вторая проверочной. После анализ повторяется, однако выполняемые выборками функции меняются: анализируемая становится проверочной, а проверочная анализируемой. Таким образом, осуществляется двойная перекрестная проверка результатов моделирования.

Часто распределение количества случаев в анализируемой и проверочной выборки явствует из распределения в общей выборке. Рассмотрим конкретный пример. Предположим, что 50% фирм выборки развиваются стабильно, а в деятельности другой половины организаций часто возникают проблемы под влиянием внешних факторов. Таким образом, ½ анализируемых фирм более подвержены риску, а другая ½ - менее. Или предположим, что четверть организации менее подвержены внешним риска и угрозам. В таком случае выбор проверочной и анализируемой выборки должен руководствоваться теми же соотношениями (25% : 75%).

Для выбора предикторов в дискриминантной функции можно использовать два метода (Для выбора предикторов в дискриминантной

функции можно использовать два метода. Прямой метод – это вычисление дискриминантной функции при одновременном введении всех предикторов. В этом случае учитывается каждая зависимая переменная. При этом ее дискриминирую-щая сила не принимается во внимание. Этот метод больше подходит к ситуации, когда анали-тик, исходя из результатов предыдущего исследования или теоретической модели, хочет, чтобы в основе различения лежали все предикторы.

Альтернативным методом является пошаговый метод. При пошаговом дискриминантном анализе предикторы вво-дят последовательно, исходя из их способности различить (дискриминировать) группы. Этот метод лучше применять в ситуации, когда исследователь хочет отобрать подмножество предикторов для включения их в дискриминатную функцию. На втором этапе осуществляется выбор метода дискриминантного анализа, который описывается числом категорий, имеющихся у зависимой переменной. Выделяют дискриминантный анализ для двух групп и множественный анализ: при осуществлении первого анализируется 2 категории, второй базируется на анализе трех и более категорий.

Главное отличие между ними заключается в том, что при наличии двух групп возможно вывести только одну дискриминантную функцию. Используя множественный дискриминантный анализ, можно вычислить несколько функций.

Рассмотрим случай для двух дискриминантных переменных. Тогда определение коэффициентов дискриминантной функции будет сведено к следующему.

Функция *f(X)* называется канонической дискриминантной функцией, а величины x_1 и x_2 – дискриминантными переменными (5.2.1):

$$f(x) = a_1X_1 + a_2X_2 \qquad (5.2.1)$$

Дискриминантная функция может быть как линейной, так и нелинейной. Выбор вида этой функции зависит от геометрического расположения разделяемых классов в пространстве дискриминантных переменных.

Коэффициенты дискриминантной функции (a_i) определяются таким образом, чтобы $\bar{f}_1(X)$ и $\bar{f}_2(X)$ как можно больше отличались между собой.

Вектор коэффициентов дискриминантной функции *(A)* определяется по формуле (5.2.2):

$$A = S_*^{-1}(\overline{X}_1 - \overline{X}_2) \qquad (5.2.2)$$

Полученные значения коэффициентов подставляют в формулу и для каждого объекта в обоих множествах вычисляют дискриминантные функции f(X*)*, затем находят среднее значение для каждой группы ($\bar{f}_k$).

Таким образом, каждому *i*-му наблюдению, которое первоначально описывалось *m*-переменными, будет соответствовать одно значение дискриминантной функции, и размерность признакового пространства снижается.

Перед тем как приступить непосредственно к процедуре классификации, нужно определить границу, разделяющую два множества. Такой величиной может быть значение функции, равноудаленное от $\bar{f}_1$ и $\bar{f}_2$ (5.2.3):

$$c = \frac{1}{2}(\bar{f}_1 + \bar{f}_2) \qquad (5.2.3)$$

Величина с называется константой дискриминации.

Объекты, расположенные над разделяющей поверхностью $f(x) = a_1x_1 + a_2x_2 + \ldots + a_px_p = c$ находятся ближе к центру множества M_1, следовательно, могут быть отнесены к первой группе, а объекты, расположенные ниже этой поверхности, ближе к центру второго множества, т.е. относятся ко второй группе.

Если граница между группами будет выбрана как сказано выше, то в этом случае суммарная вероятность ошибочной классификации будет минимальной.

Важнейшим этапом дискриминантного исследования является анализ значимости результатов моделирования.

Бессмысленно интерпретировать результаты анализа, если определенные дискриминантные функции не являются статистически значимыми. Поэтому следует выполнить статистическую проверку нулевой гипотезы о равенстве средних всех дискриминантных функций во

всех группах генеральной совокупности. Часто эта проверка базируется на коэффициенте лямбда (λ) Уилкса. Если одновременно проверяют несколько функций, как в случае множественного дискриминантного анализа, то коэффициент λ является суммой одномерных λ для каждой функции. Уровень значимости оценивают, исходя из преобразования λ -статистики в статистику хи-квадрат (исходя из распределения хи-квадрат, которому подчиняется λ -статистика). Если нулевую гипотезу отклоняют, что указывает на значимую дискриминацию, то можно продолжать интерпретировать результаты.

Для интерпретации дискриминантных весов используется процедура, аналогичная множественному регрессионному анализу.

Значение коэффициента для конкретного предиктора зависит от других предикторов, включенных в дискриминантную функцию. Знаки коэффициентов условны, но они указывают, какие значения переменной приводят к большим и маленьким значениям функции и связывают их с конкретными группами.

При наличии мультиколлинеарности между независимыми переменными не существует однозначной меры относительной важности предикторов для дискриминации между группами. Помня об этом предостережении, можно получить некоторое представление об относительной важности переменных, изучив абсолютные значения нормированных коэффициентов дискриминантной функции. Как правило, предикторы с относительно большими нормированными коэффициентами вносят больший вклад в дискриминирующую мощность функции по сравнению с предикторами, имеющими меньшие коэффициенты.

Некоторое представление об относительной важности предикторов можно также получить, изучив структурные коэффициенты корреляции, которые также называют каноническими или дискриминантными нагрузками. Эти линейные коэффициенты корреляции между каждым из предикторов и дискриминантной функцией представляют дисперсию, которую предиктор делит вместе с функцией. Как и нормированные коэффициенты, эти коэффициенты корреляции следует использовать осторожно.

При интерпретации результатов дискриминантного анализа также может помочь разработка характеристической структуры для каждой группы посредством описания каждой группы через групповые средние для предикторов.

Оценка достоверности дискриминантного анализа является заключительным этапом исследования.

Как уже говорилось, данные разбивают случайным образом на две подвыборки. Анализируемую часть выборки используют для вычисления дискриминантной функции, а проверочную – для построения классификационной матрицы.

Дискриминантные веса, определенные анализируемой выборкой, умножают на значения независимых переменных в проверочной выборке, чтобы получить дискриминантные показатели для случаев в этой выборке. Затем случаи распределяют по группам, исходя из дискриминантных показателей и соответствующего правила принятия решения. Например, при дискриминантном анализе двух групп случай может быть отнесен к группе с самым близким по значению центроидом. Затем, сложив элементы, лежащие на диагонали матрицы, и разделив полученную сумму на общее количество случаев, можно определить коэффициент результативности или процент верно классифицированных случаев. Полезно сравнить процент случаев, верно классифицированных с помощью дискриминантного анализа, с процентом случаев, который можно получить случайным образом. Для равных по размеру групп процент случайной классификации равен частному от деления единицы на количество групп. Превысит ли и насколько количество верно классифицированных случаев их случайное количество? Здесь нет общепринятого подхода, хотя некоторые авторы считают, что точность классификации, достигнутая с помощью дискриминантного анализа, должна быть, по крайней мере, на 25% выше, чем точность, которую можно достичь случайным образом.

Большинство программ для выполнения дискриминантного анализа также определяют классификационную матрицу, исходя из анализируемой выборки. Поскольку программы учитывают даже случайные вариации в данных, то полученные результаты всегда точнее, чем классификация данных на основе проверочной выборки.

5.3. Дискриминантный анализ в Statistica

Рассмотрим возможности программного комплекса на примере.

Пример

Существует набор регионов, которые характеризуются набором данных за отдельно взятый период. Есть обучающая выборка.

Необходимо провести классификацию регионов, не вошедших в обучающую выборку.

Исходные данные приведены в таблице 5.3.1

Таблица 5.3.1

Исходные данные для анализа

Регион	x_1	x_2	x_3	x_4	x_5	Группа
Р1	1531,9	41775	165672	735	6,3	
Р2	1168,8	35582	81337	322	4,9	2
Р3	1323,7	39550	103846	357	4,8	1
Р4	2287,7	40830	285892	368	12,2	1
Р5	976,9	32403	44981	337	4,4	
Р6	1012,8	48837	128508	334	4,1	1
Р7	620,8	35967	42743	138	2,4	
Р8	1083,6	40292	193352	376	5,7	1
Р9	1113,7	40188	179400	545	4,7	1
Р10	7768,9	64041	1144660	799	29,1	
Р11	714,1	35754	60612	385	3,3	1
Р12	1085,2	40631	73886	270	5,7	
Р13	909,8	36529	70327	321	4,9	1
Р14	981,0	34438	79397	292	3,9	2
Р15	1230,1	40286	84293	252	5,8	2
Р16	1432,6	44726	182297	405	5,5	
Р17	1227,3	41209	109967	276	6,8	1
Р18	12635,5	112768	4839918	2548	73,5	2
Р19	3153,8	31859	251368	427	12,7	2
Р20	524,1	31362	20435	854	2,3	
Р21	870,5	31712	51063	609	3,9	
Р22	464,2	32846	28083	345	1,9	
Р23	688,1	32999	34554	717	4,7	1
Р24	1516,4	31291	84279	642	4,3	2
Р25	2780,2	37387	254164	275	12,3	2
Р26	4001,6	42848	419337	310	17,4	2
Р27	671,5	35497	35546	223	2,5	1
Р28	770,7	34499	49690	296	4,3	
Р29	3886,4	45800	683305	448	16,5	1

P30	1484,5	39791	117156	253	7,4	1
P31	1198,4	35799	61325	433	6,0	2
P32	2556,8	46267	307824	137	13,3	1
P33	1234,8	36143	75540	114	6,0	2
P34	3144,2	41369	385625	300	15,3	1
P35	1924,6	38357	198131	168	9,2	
P36	1274,1	36031	96202	291	5,6	
P37	3131,7	42771	364151	330	15,2	
P38	2361,0	37408	173054	170	12,3	
P39	1204,0	36126	108493	254	5,2	

Решение

После внесения данных в программу, необходимо перейти **Анализ – Многомерный анализ – Дискриминантный анализ** (рис. 5.3.1).

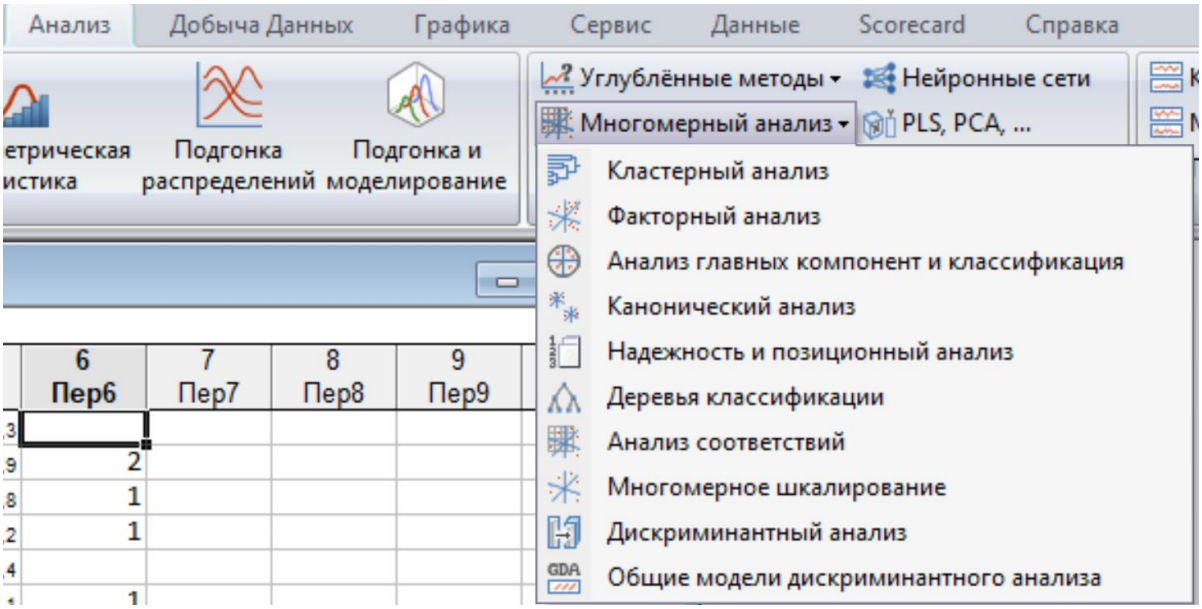


Рис. 5.3.1. Меню выбора дискриминантного анализа

В появившемся окне (рис. 5.3.2) необходимо осуществить выбор переменных по двум типам: группирующая и независимая.

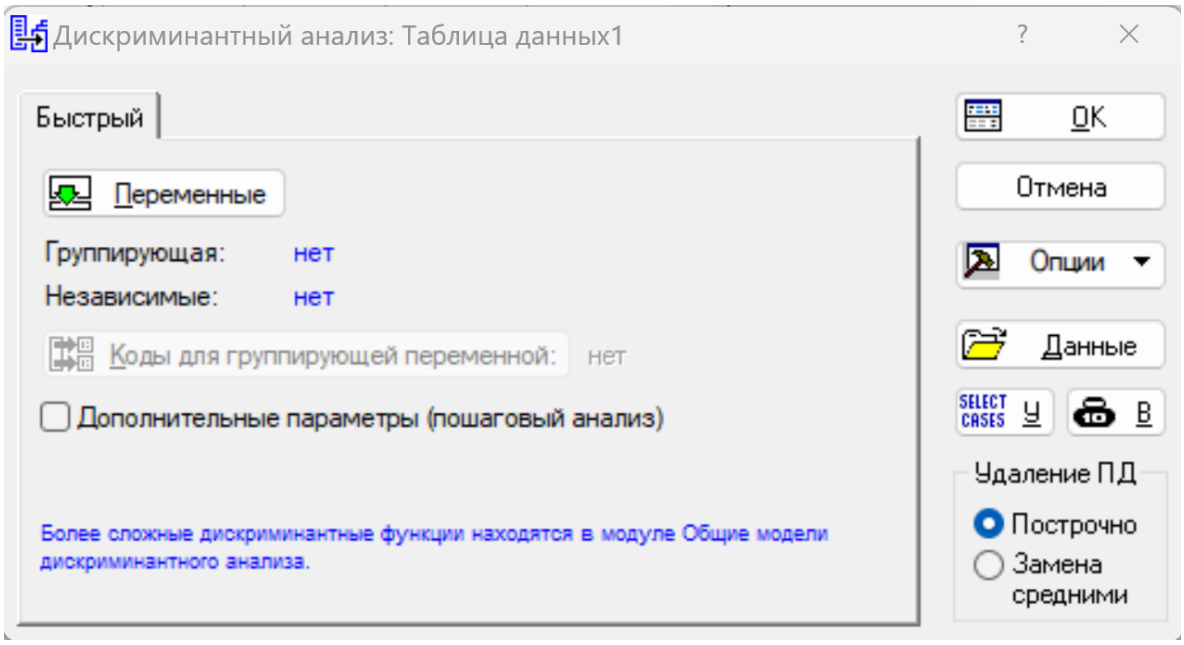


Рис. 5.3.2. Выбор переменных дискриминантного анализа

В случае, рассматриваемом в нашем примере, группирующей переменной будет Пер6, в которую внесены данные обучающей выборки, независимыми соответственно Пер1-Пер5, в которые внесены данные характеристик регионов. Соответственно форма пример вид (рис. 5.3.3).

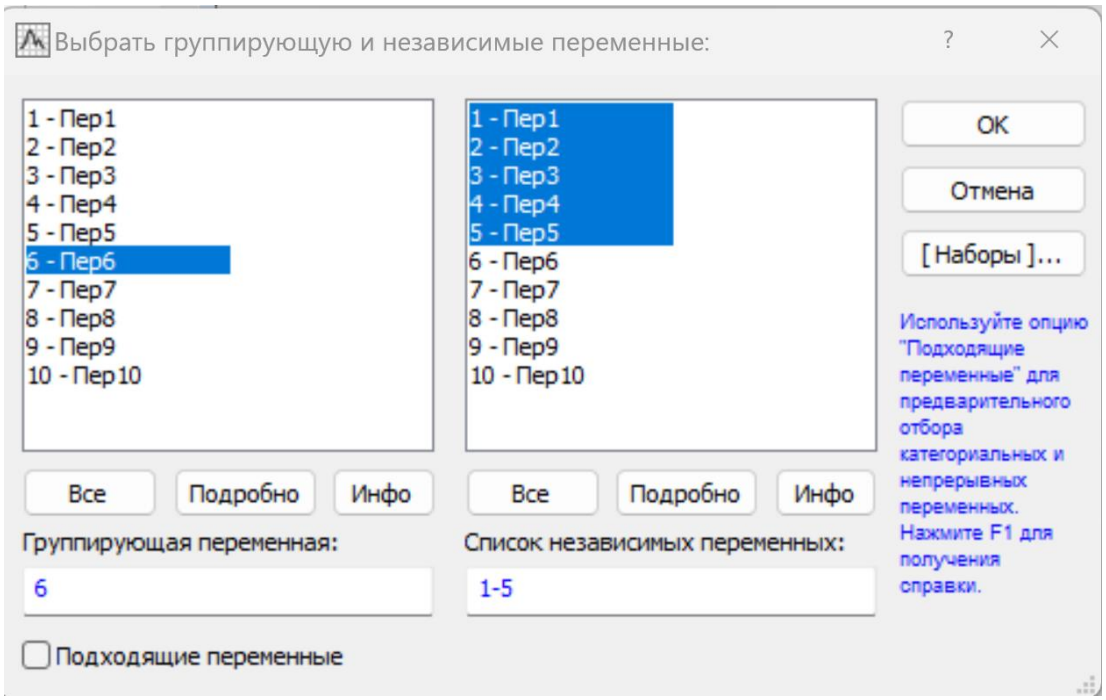


Рис. 5.3.3. Выбор переменных дискриминантного анализа по типам

После нажатия кнопки ОК, появится меню дискриминантного анализа (рис. 5.3.4).

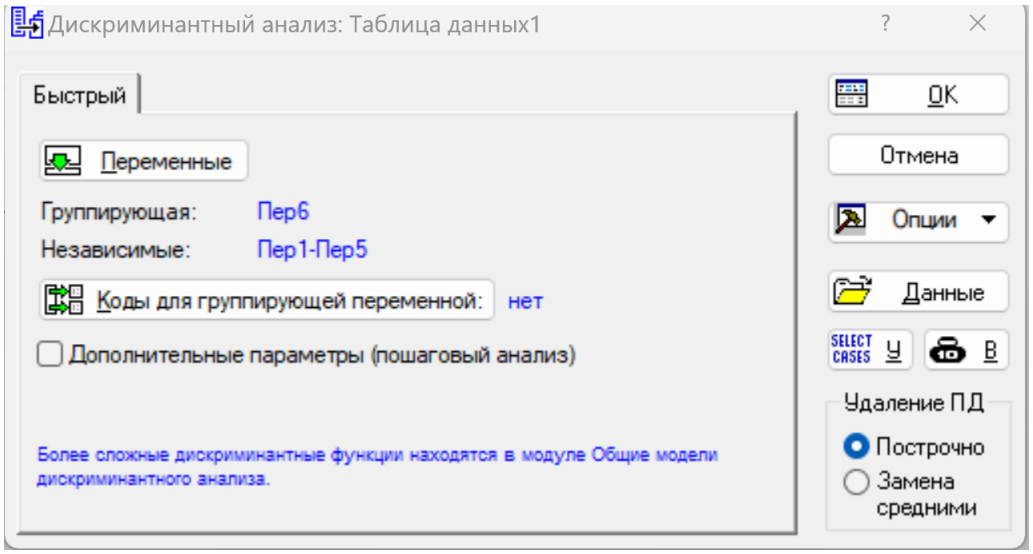


Рис. 5.3.4. Меню дискриминантного анализа

Необходимо ввести коды для группирующей переменной. Для этого необходимо нажать на соответствующую кнопку. В открывшемся окне (рис. 5.3.5).

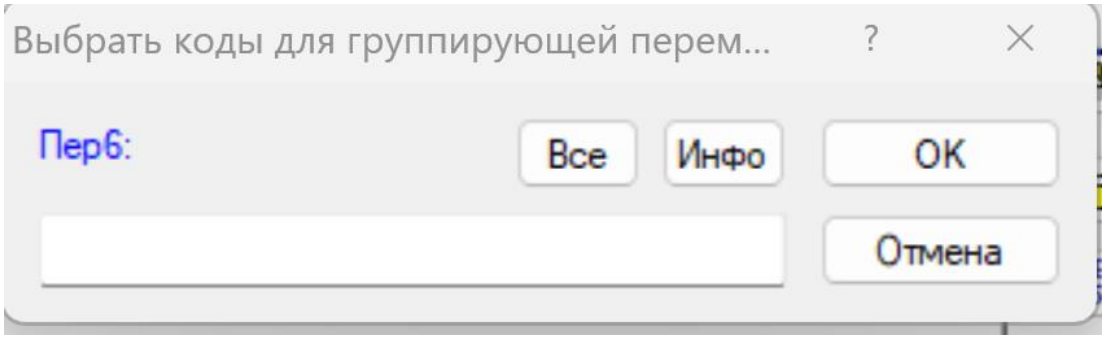


Рис. 5.3.5. Меню ввода кодов для группирующей переменной

В нашем примере, это соответственно 1-2 (именно столько групп выделены в обучающей выборке). Ввод можно осуществить или непосредственно вручную, или нажав на кнопку Все, или нажав на кнопку Инфо и выбрать из появившегося списка (рис. 5.3.6).

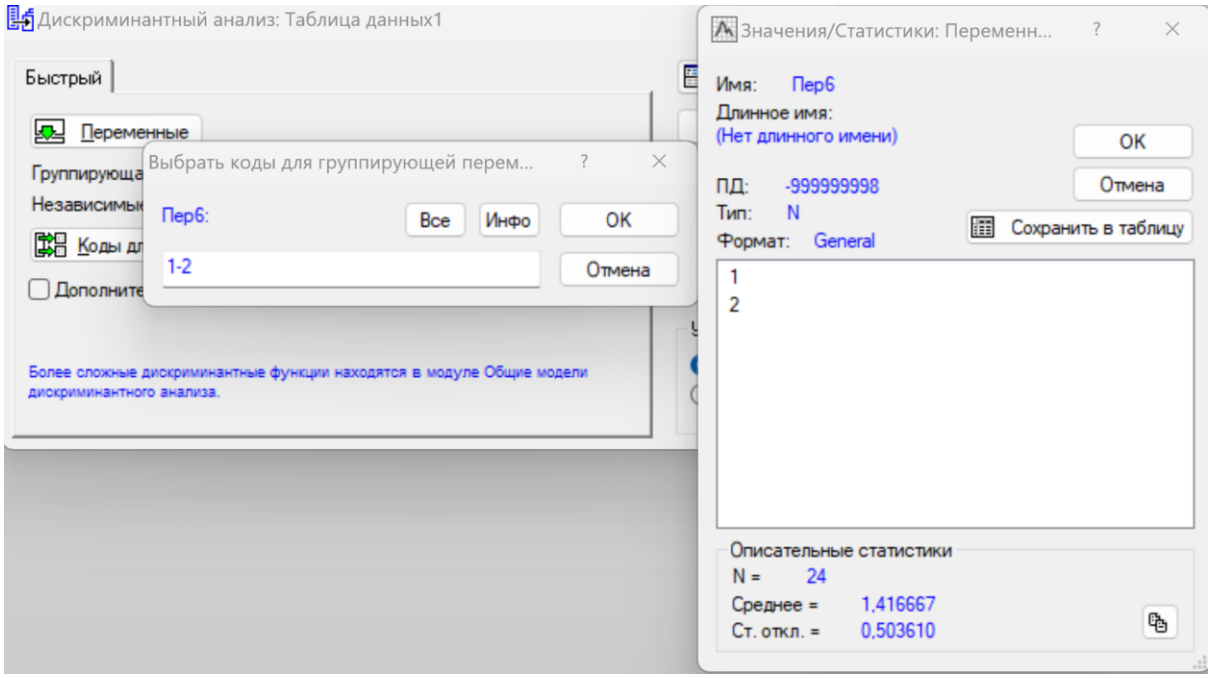


Рис. 5.3.6. Ввод кодов для группирующей переменной

После ввода кодов и нажатия кнопки OK появится форма определения модели (рис. 5.3.7).

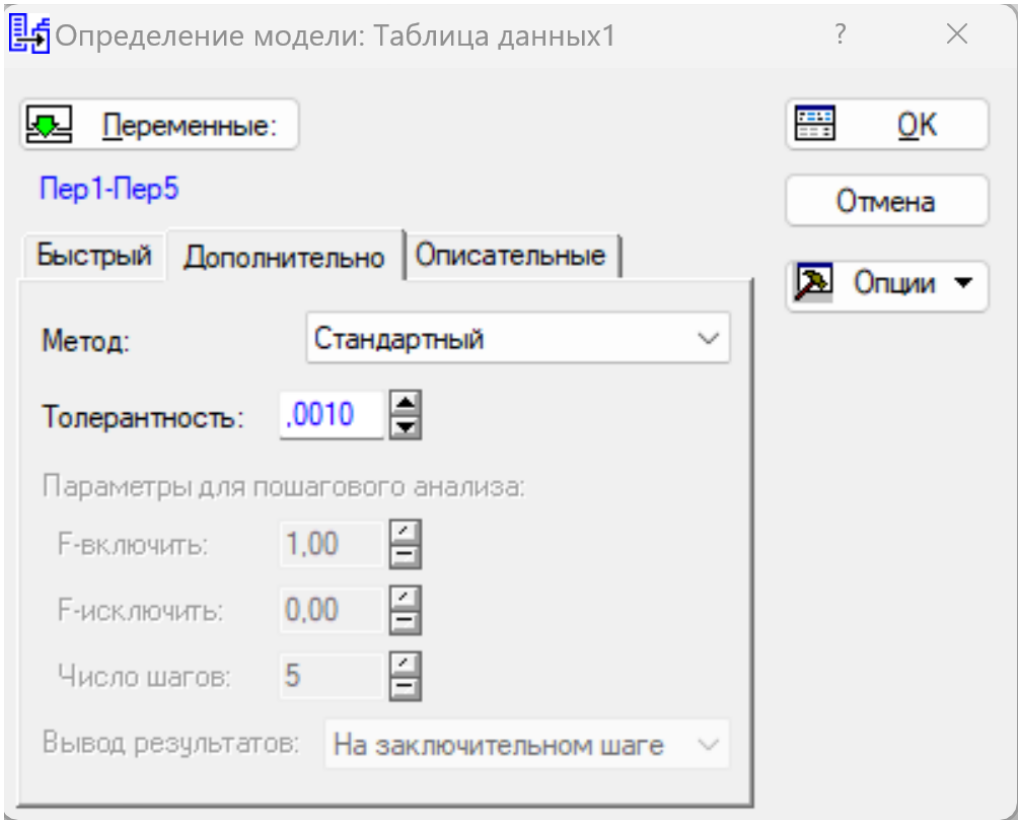


Рис. 5.3.7. Определение модели

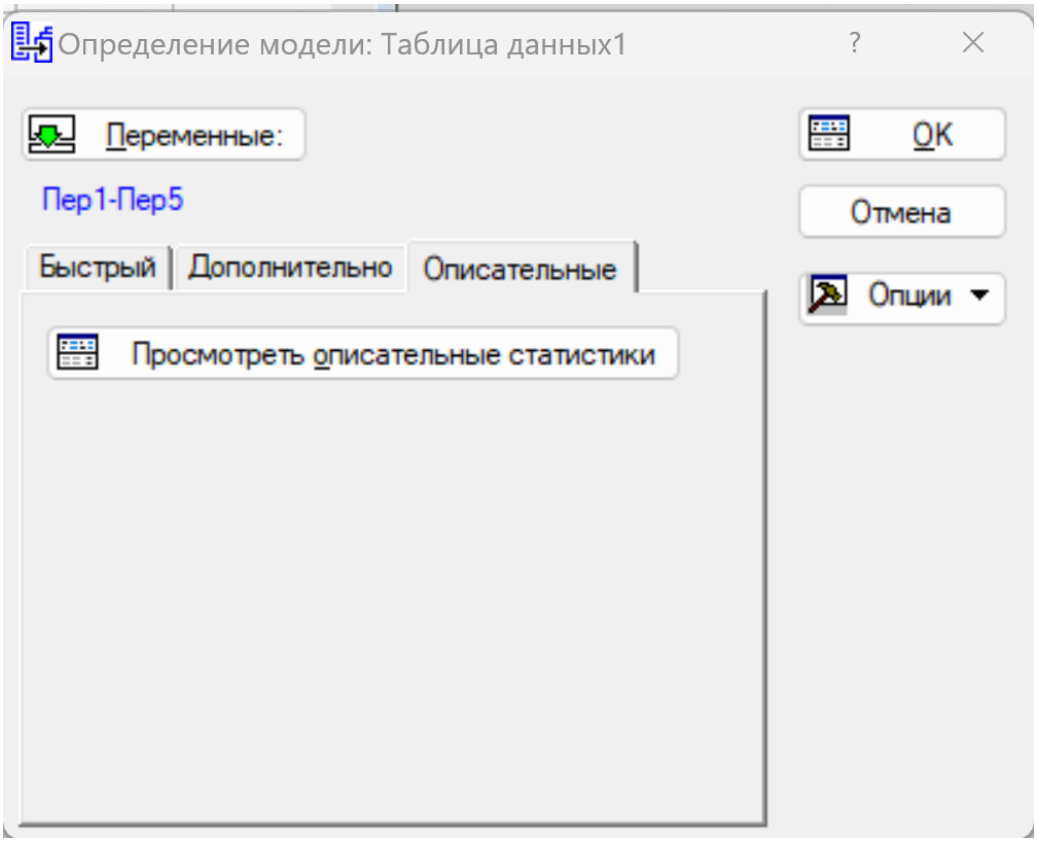


Рис. 5.3.8. Вкладка **Описательные**

Пользователю предоставляется выбор из трех методов:

- стандартный;
- пошаговый с включением;
- пошаговый с исключением.

Для оценки параметров распределения по группам необходимо использовать кнопку **Просмотреть описательные статистики** на вкладке **Описательные** (рис. 5.3.8).

После нажатия на кнопку Просмотреть описательные статистики появится форма (рис. 5.3.9).

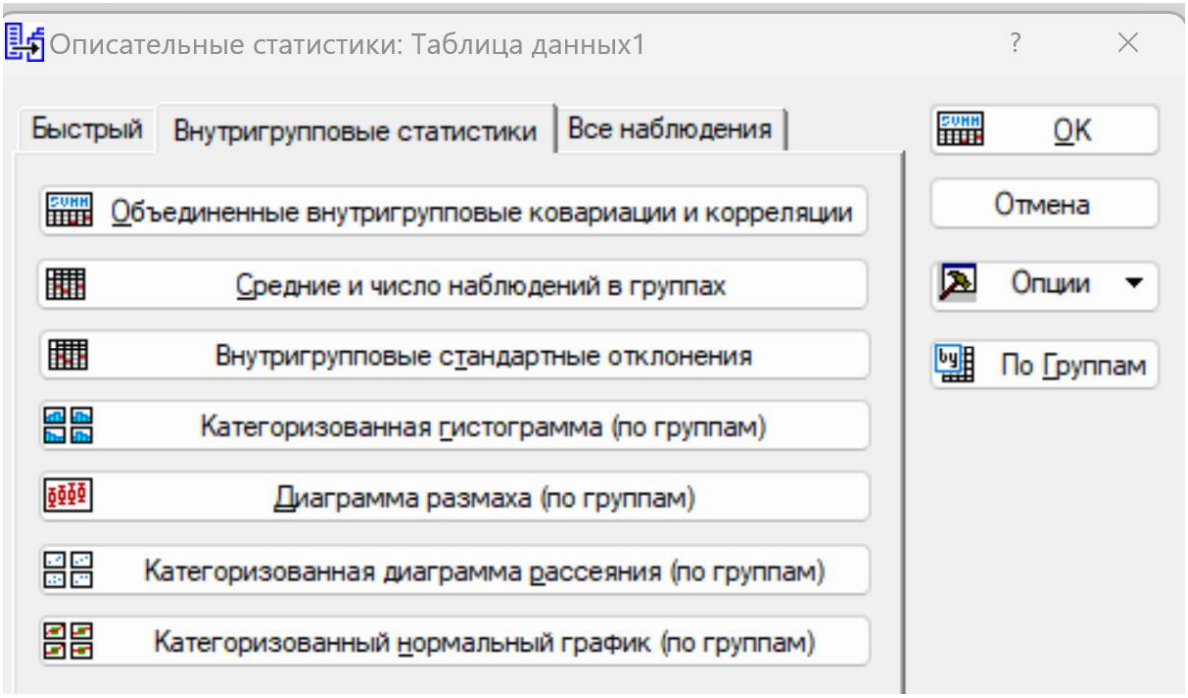


Рис. 5.3.9 Описательные статистики

При нажатии кнопки **Объединенные внутригрупповые ковариации и корреляции** произойдет расчет оценок общих для двух классов ковариационной и корреляционной матриц (рис. 5.3.10).

	Объединенные внутригрупповые корреляции (Таблица данных1)				
Переменная	**Пер1**	Пер2	Пер3	Пер4	Пер5
Пер1	1,00	0,93	0,96	0,85	0,99
Пер2	0,93	1,00	0,98	0,89	0,96
Пер3	0,96	0,98	1,00	0,94	0,98
Пер4	0,85	0,89	0,94	1,00	0,89
Пер5	0,99	0,96	0,98	0,89	1,00

Рис. 5.3.10. **Объединенные внутригрупповые ковариации и корреляции**

При нажатии кнопки **Средние и число наблюдений в группе** появится табличная форма с соответствующими показателями (рис. 5.3.11).

	Средние (Таблица данных1)					
Пер6	**Пер1**	Пер2	Пер3	Пер4	Пер5	N
G_1:1	1578,871	40350,86	192565,3	359,9329	7,57407	14
G_2:2	2990,060	43840,10	623095,9	561,5891	14,67350	10
Все гр.	2166,867	41804,71	371953,0	443,9563	10,53217	24

Рис. 5.3.11. **Средние и число наблюдений в группе**

Нажатие кнопки **Внутригрупповые стандартные отклонения** выведут соответствующую табличную форму (рис. 5.3.12).

	Стандартные отклонения (Таблица данных1)					
Пер6	**Пер1**	Пер2	Пер3	Пер4	Пер5	N
G_1:1	1000,990	4411,90	176880	142,3640	4,68622	14
G_2:2	3544,041	24467,30	1486328	711,6010	21,16508	10
Все гр.	2446,688	15758,92	963926	468,9523	14,15926	24

Рис. 5.3.12. **Внутригрупповые стандартные отклонения**

Остальные кнопки данной формы предназначены для построения различного рода графических форм. Примеры данных графиков представлены на рисунках 5.3.13 – 5.3.16.

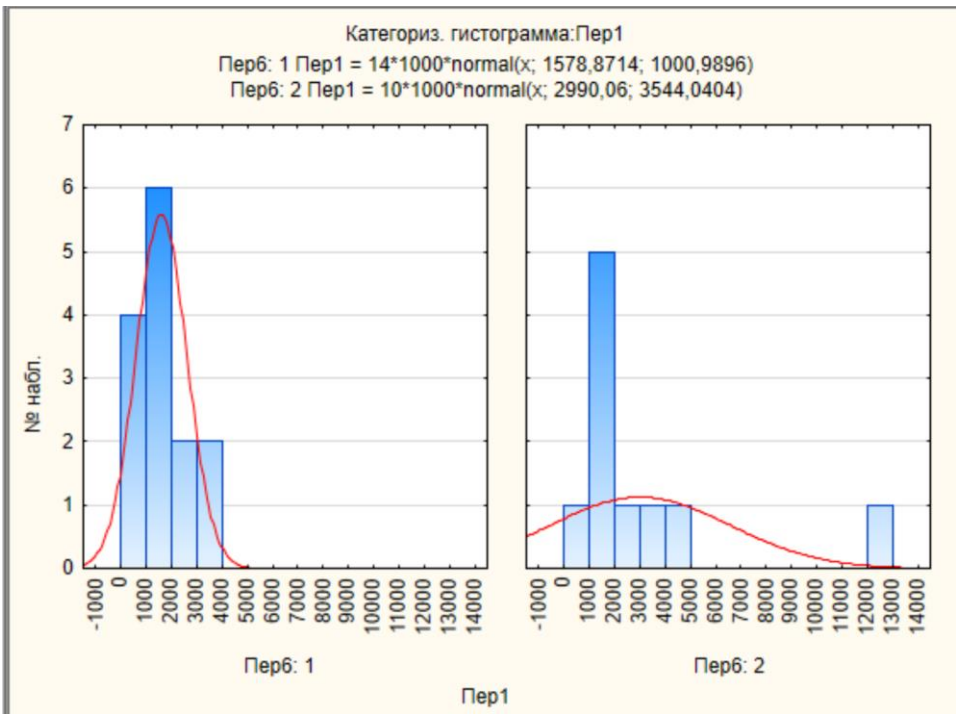


Рис. 5.3.13. Категоризованная гистограмма

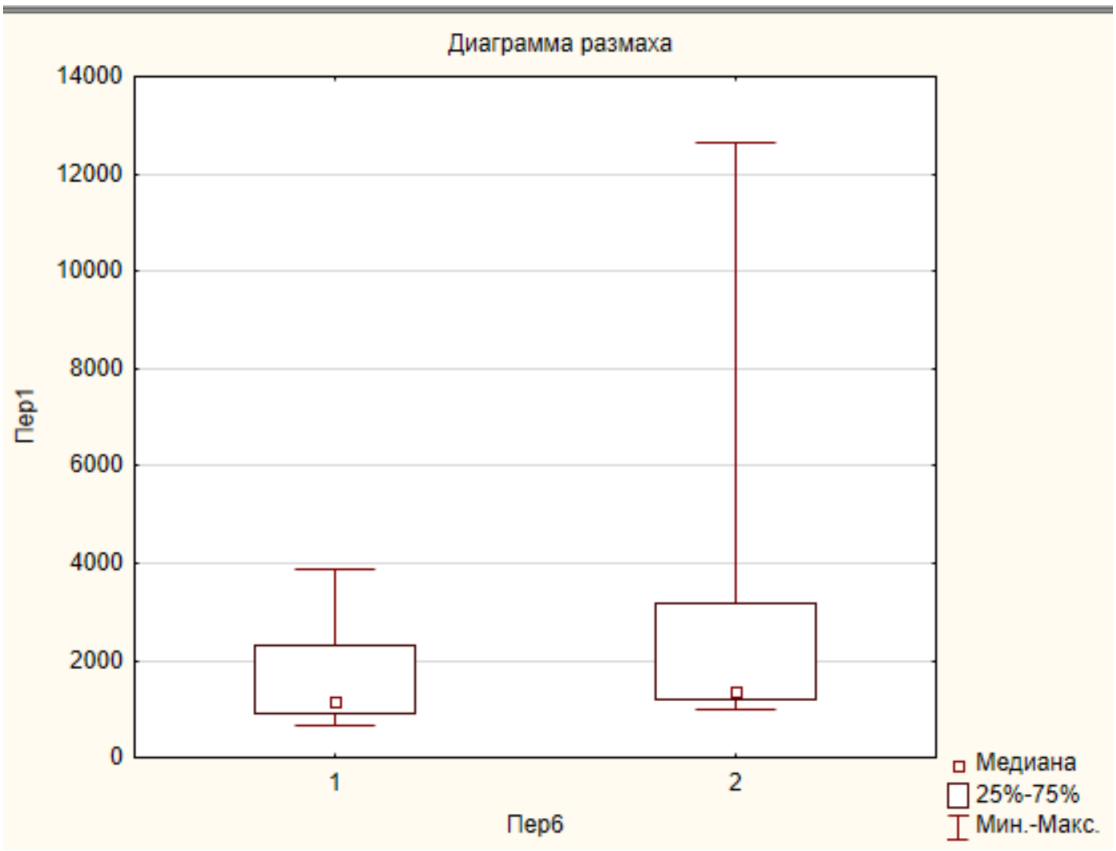


Рис. 5.3.14. Диаграмма размаха

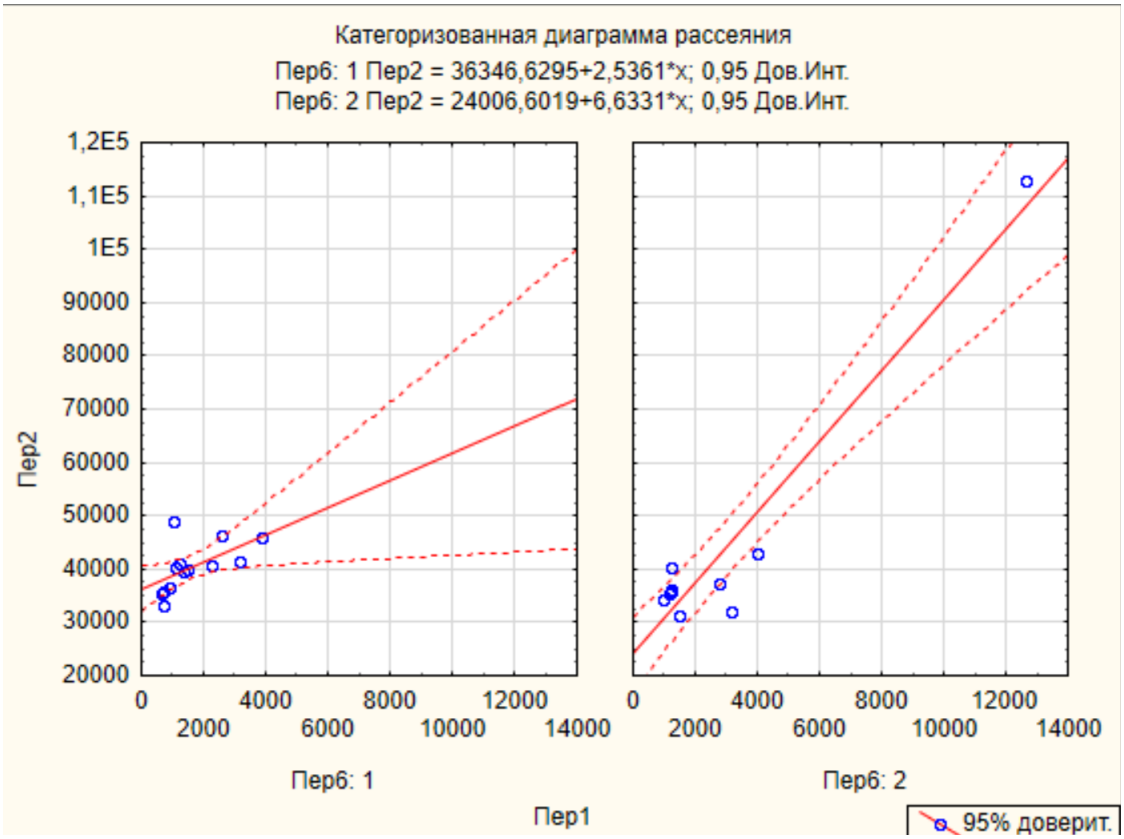


Рис. 5.3.15. Категоризованная диаграмма рассеяния

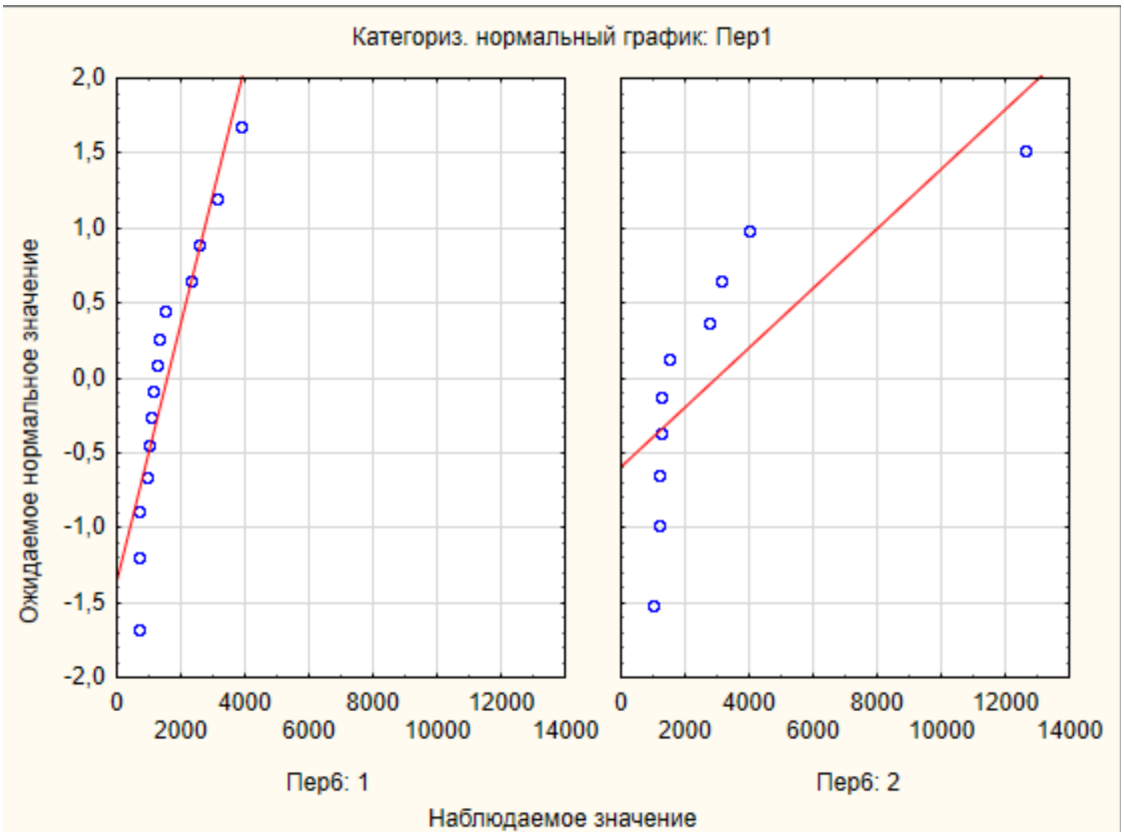


Рис. 5.3.16. Категоризованная нормальный график

После нажатия **ОК** в окне **Определение модели** появится форма **Результаты анализа дискриминантных функций** (рис. 5.3.17).

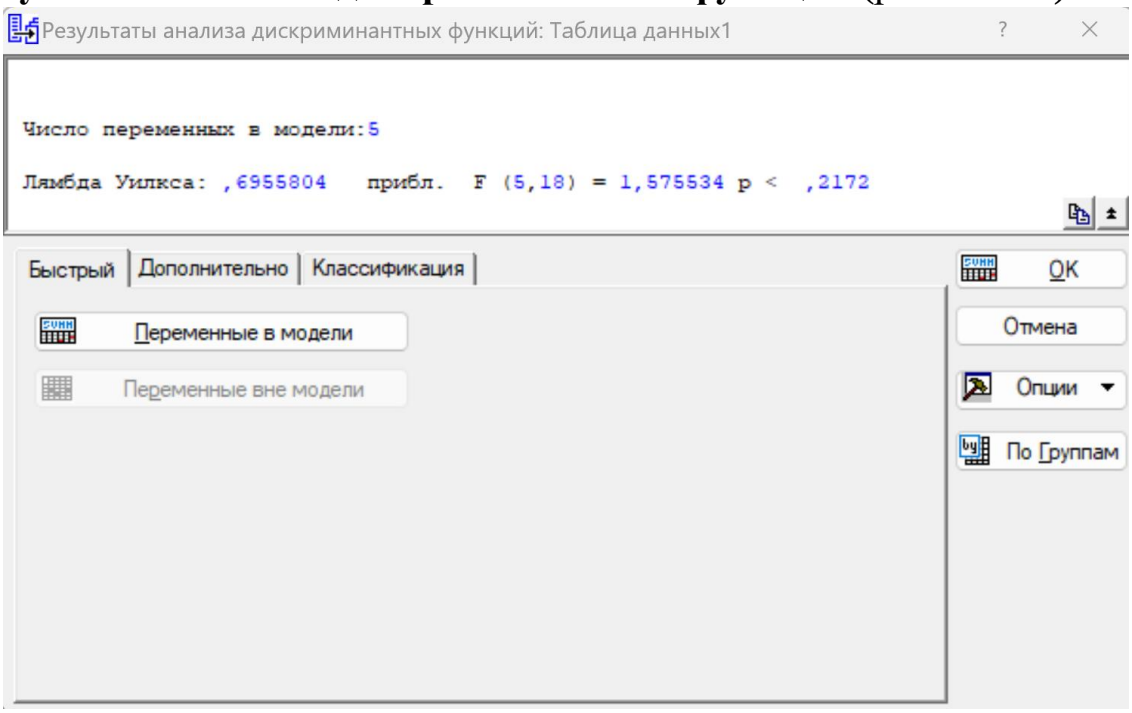


Рис. 5.3.17. Результаты анализа дискриминантных функций

В верхней части представлена лямбда Уилкса и F. После нажатия кнопки **Переменные и модели** выгружается форма итогов с рассчитанными параметрами относительно всех показателей (рис. 5.3.18).

Итоги анализа дискриминантн. функций (Таблица данных1)
Переменных в модели: 5; Группир.: Пер6 (2 гр.)
Лямбда Уилкса: ,69558 прибл. F (5,18)=1,5755 p< ,2172

N=24	Уилкса Лямбда	Частная Лямбда	F-исключ (1,18)	p-уров.	Толер.	1-толер. (R-кв.)
Пер1	0,716253	0,971138	0,534963	0,473943	0,011766	0,988234
Пер2	0,830604	0,837440	3,494086	0,077948	0,030284	0,969716
Пер3	0,749409	0,928172	1,392966	0,253268	0,006282	0,993718
Пер4	0,706869	0,984030	0,292127	0,595486	0,062094	0,937906
Пер5	0,711219	0,978012	0,404692	0,532682	0,005569	0,994431

Рис. 5.3.18. Итоги анализа дискриминантных функций

Вид формы результатов дискриминантного анализа можно указать на вкладке Классификации (рис. 5.3.19).

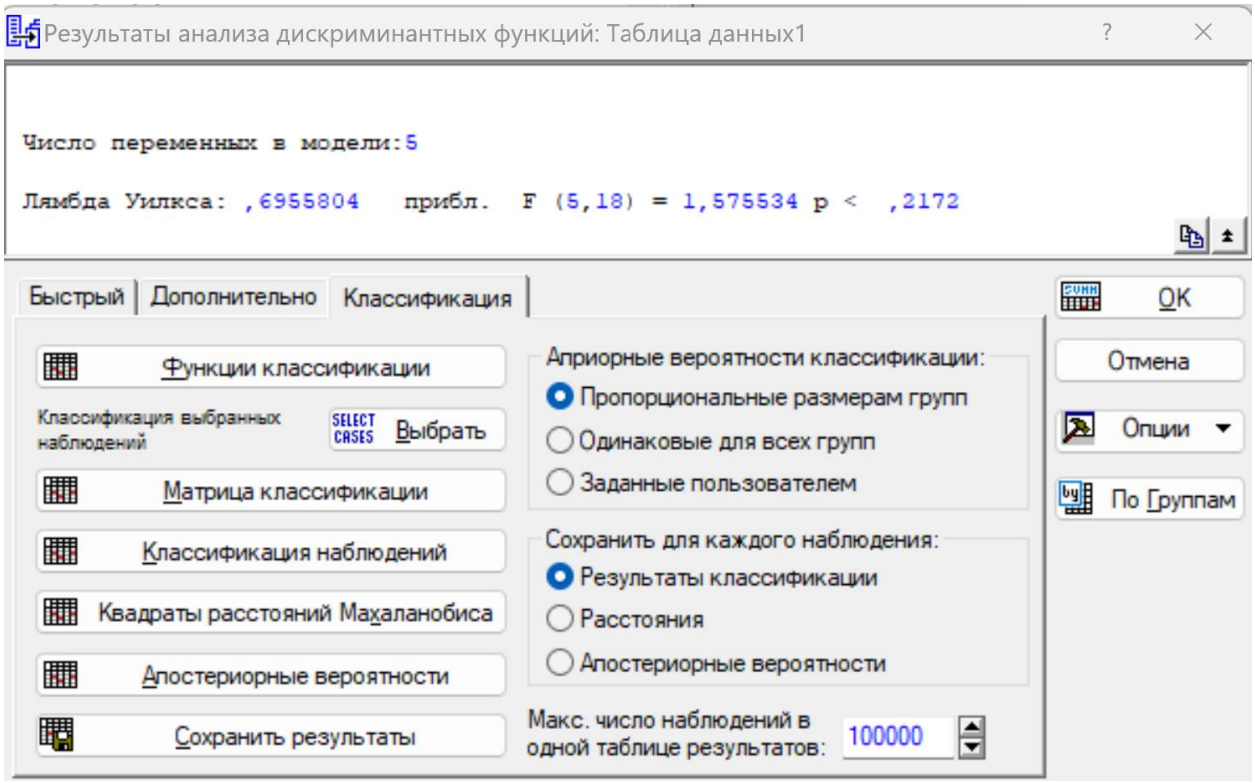


Рис. 5.3.19. Классификации

В меню **Априорные вероятности и классификации** есть выбор из трех вариантов:

- пропорционально объемам групп (по умолчанию);
- одинаковые для всех групп;
- заданные пользователем.

При нажатии кнопки Функции классификации рассчитывается коэффициентов линейных дискриминантных функций Фишера (рис. 5.3.20).

	Функции классификации; группировка: Пер6 (Таблица данных1)	
Переменная	**G_1:1 p=,58333**	G_2:2 p=,41667
Пер1	0,014	0,015
Пер2	0,006	0,005
Пер3	-0,000	-0,000
Пер4	0,092	0,089
Пер5	0,762	0,432
Конст-та	-125,541	-113,682

Рис. 5.3.20. Функции классификации

На основе данных функций происходит классификация объектов обучающих выборок. При нажатии кнопки Матрица классификации будет выведена табличная (матричная) форма результатов данной процедуры (рис. 5.3.21).

	Матрица классификации (Таблица данных1) Строки: наблюдаемые классы Столбцы: предсказанные классы		
Группа	**Процент правиль.**	G_1:1 p=,58333	G_2:2 p=,41667
G_1:1	85,71429	12	2
G_2:2	60,00000	4	6
Всего	75,00000	16	8

Рис. 5.3.21. Матрица классификации

Следует отметить, что если объект, который априори относился в один класс, после проведения процедуры классификации был отнесен к иному классу, то он помечается звездочкой (в данном примере таковые присутствуют).

Подобные результаты доступны при нажатии кнопки Квадраты расстояний Махаланобиса (таблица 5.3.2)

Таблица 5.3.2

Квадраты расстояний Махаланобиса

Квадраты расстояний Махаланобиса до центров (Таблица данных1) Неправильные классификации отмечены *			
	Наблюд. - Класс.	**G_1:1 - p=,58333**	**G_2:2 - p=,41667**
1	---	13,7218	17,9398
*** 2**	G_2:2	1,5674	1,0967
3	G_1:1	1,8210	3,2945
4	G_1:1	2,5930	3,9485
5	---	4,0782	2,4415
6	G_1:1	11,5602	19,2633
7	---	7,8800	7,4093
8	G_1:1	0,7364	2,6696
9	G_1:1	2,3347	4,5673
10	---	112,9087	112,4602
11	G_1:1	1,5820	2,2404
12	---	0,7522	4,3757
13	G_1:1	1,2627	2,3895
14	G_2:2	4,2643	2,8350
*** 15**	G_2:2	0,3388	3,0665
16	---	5,3113	9,3490
17	G_1:1	1,0695	4,7601
18	G_2:2	24,5311	19,6996
19	G_2:2	11,1762	5,0745
20	---	13,6380	14,2495
21	---	5,3384	4,4920
22	---	5,4792	4,3724
23	G_1:1	11,0759	12,9824
24	G_2:2	10,2624	7,0103
25	G_2:2	3,6619	1,3274
26	G_2:2	7,9920	6,4352
27	G_1:1	4,7799	4,4307
28	---	3,3873	3,3887

*** 29**	G_1:1	8,4652	5,0288
30	G_1:1	0,4552	2,6163
*** 31**	G_2:2	1,5373	2,7864
32	G_1:1	4,6605	8,1014
*** 33**	G_2:2	5,0861	4,5803
*** 34**	G_1:1	3,6920	2,9034
35	---	3,3980	2,6393
36	---	1,5580	1,1205
37	---	3,7324	4,3558
38	---	6,2342	6,0249
39	---	2,9861	2,1951

Апостериорные вероятности рассчитываются с нажатием соответствующей кнопки (таблица 5.3.3).

Таблица 5.3.3

Апостериорные вероятности

Апостериорные вероятности (Таблица данных1) Неправильные классификации отмечены *			
	Наблюд. - Класс.	**G_1:1 - p=,58333**	**G_2:2 - p=,41667**
1	---	0,920232	0,079768
*** 2**	G_2:2	0,525257	0,474743
3	G_1:1	0,745210	0,254790
4	G_1:1	0,733843	0,266157
5	---	0,381815	0,618185
6	G_1:1	0,985050	0,014950
7	---	0,525261	0,474739
8	G_1:1	0,786353	0,213647
9	G_1:1	0,810432	0,189568
10	---	0,528029	0,471971
11	G_1:1	0,660533	0,339467
12	---	0,895501	0,104499
13	G_1:1	0,710919	0,289081
14	G_2:2	0,406565	0,593435
*** 15**	G_2:2	0,845580	0,154420
16	---	0,913357	0,086643
17	G_1:1	0,898601	0,101399
18	G_2:2	0,111128	0,888872
19	G_2:2	0,062132	0,937868
20	---	0,655251	0,344749
21	---	0,478333	0,521667

22	---	0,445986	0,554014
23	G_1:1	0,784099	0,215901
24	G_2:2	0,215922	0,784078
25	G_2:2	0,303479	0,696521
26	G_2:2	0,391284	0,608716
27	G_1:1	0,540383	0,459617
28	---	0,583504	0,416496
*** 29**	G_1:1	0,200726	0,799274
30	G_1:1	0,804868	0,195132
*** 31**	G_2:2	0,723322	0,276678
32	G_1:1	0,886643	0,113357
*** 33**	G_2:2	0,520880	0,479120
*** 34**	G_1:1	0,485543	0,514457
35	---	0,489278	0,510722
36	---	0,529386	0,470614
37	---	0,656604	0,343396
38	---	0,557688	0,442312
39	---	0,485240	0,514760

Соответственно классификация регионов произведена, поставленная задача решена.

Контрольные вопросы по теме

1. Что такое дискриминантный анализ?
2. Какие основные цели проведения дискриминантного анализа?
3. В чем отличие между дискриминантным анализом и кластерным анализом?
4. Какие типы дискриминантного анализа существуют?
5. Какие предпосылки должны быть выполнены для применения дискриминантного анализа?
6. Как проводится линейный дискриминантный анализ?
7. Как проводится квадратичный дискриминантный анализ?
8. Как интерпретировать результаты дискриминантного анализа?
9. Как выбрать оптимальные предикторы для дискриминантного анализа?
10. Как оценить качество модели дискриминантного анализа?
11. Какие методы используются для измерения различий между группами при проведении дискриминантного анализа?

12. Как можно использовать результаты дискриминантного анализа для классификации объектов или субъектов?

13. Как можно использовать результаты дискриминантного анализа для прогнозирования принадлежности объектов к определенным группам?

14. Как можно использовать результаты дискриминантного анализа для выявления важных предикторов, влияющих на зависимую переменную?

15. Как можно использовать результаты дискриминантного анализа для выявления групп риска или возможностей в данных?

16. Как можно использовать результаты дискриминантного анализа для определения факторов, объясняющих различия между группами?

17. Как можно использовать результаты дискриминантного анализа для выявления закономерностей в данных и прогнозирования будущих значений переменных?

18. Как можно использовать результаты дискриминантного анализа для определения оптимальных стратегий действий в различных ситуациях?

19. Как можно использовать результаты дискриминантного анализа для определения воздействия различных факторов на изучаемый процесс или явление?

20. Какие программные инструменты чаще всего используются для проведения дискриминантного анализа?

21. Как можно использовать результаты дискриминантного анализа для оценки эффективности различных методов или стратегий?

22. Как можно использовать результаты дискриминантного анализа для выявления тенденций или закономерностей в данных?

23. Как можно использовать результаты дискриминантного анализа для сравнения различных групп или образцов данных?

24. Как можно использовать результаты дискриминантного анализа для определения оптимальных условий или параметров процессов?

25. Что такое линейный дискриминантный анализ и как он используется при проведении дискриминантного анализа?

26. Что такое квадратичный дискриминантный анализ и как он используется при проведении дискриминантного анализа?

27. Что такое непараметрический дискриминантный анализ и как он используется при проведении дискриминантного анализа?

28. Что такое регуляризованный дискриминантный анализ и как он используется при проведении дискриминантного анализа?

29. Что такое метод "Лассо" и как он используется при проведении дискриминантного анализа?

30. Что такое метод "Ridge" и как он используется при проведении дискриминантного анализа?

31. Как можно использовать результаты дискриминантного анализа для прогнозирования классификации новых объектов?

32. Как можно использовать результаты дискриминантного анализа для выявления скрытых паттернов в данных?

33. Как можно использовать результаты дискриминантного анализа для определения групп схожих объектов или явлений в данных?

34. Как можно использовать результаты дискриминантного анализа для выявления внутренних структур или закономерностей в данных?

35. Как можно использовать результаты дискриминантного анализа для выявления внутренних связей между объектами или явлениями?

36. Как можно использовать результаты дискриминантного анализа для определения групп риска или возможностей в данных?

37. Как можно использовать результаты дискриминантного анализа для определения оптимальных стратегий управления в различных ситуациях?

38. Что такое метод "Fisher's Linear Discriminant" и как он используется при проведении дискриминантного анализа?

39. Что такое метод "Quadratic Discriminant Analysis" и как он используется при проведении дискриминантного анализа?

40.

Что такое метод "Regularized Discriminant Analysis" и как он используется при проведении дискриминантного анализа?

41. Что такое метод "Linear Discriminant Analysis" и как он используется при проведении дискриминантного анализа?

42. Что такое метод "Stepwise Discriminant Analysis" и как он используется при проведении дискриминантного анализа?

43. Что такое метод "Linear Discriminant Function" и как он используется при проведении дискриминантного анализа?

44. Что такое метод "Discriminant Function Analysis" и как он используется при проведении дискриминантного анализа?

45. Что такое метод "Discriminant Rule" и как он используется при проведении дискриминантного анализа?

46. Что такое метод "Discriminant Coefficients" и как он используется при проведении дискриминантного анализа?

47. Что такое метод "Discriminant Score" и как он используется при проведении дискриминантного анализа?

48. Что такое метод "Canonical Discriminant Analysis" и как он используется при проведении дискриминантного анализа?

49. Что такое метод "Discriminant Loadings" и как он используется при проведении дискриминантного анализа?

50. Какие методы используются для проверки предпосылок и условий применения дискриминантного анализа?

6. ФАКТОРНЫЙ АНАЛИЗ

6.1. Основные положения факторного анализа

Факторный анализ является процедурой, которая используется для того, чтобы с помощью меньшего количества переменных, именуемых факторами, описать процесс или явление, характеристики которого ранее были представлены большим количеством величин.

Объединение переменных в факторы должно осуществляться на основе подходов:

- Объединение в один фактор только тех переменных, которые сильно коррелируют между собой;
- Переменные, входящие в состав разных факторов, должны слабо коррелировать между собой.

При осуществлении факторного анализа важно понимать, что означает термин «фактор». Под определением данного понятия принято понимать скрытые переменные, именуемые в некоторых источниках «латентные» Факторы не подлежат непосредственному изменению, однако, тесно связаны с измеряемыми параметрами, которые являются проявлением данных факторов.

Сущность факторного анализа состоит в предположении, что имеет место ряд величин, благодаря которым проявляются отношения различного характера между переменными. Также предполагается, что данные величины не известны исследователю.

Для объяснения структуры связей между признаками, подлежащими анализу $x^{(1)}$, $x^{(2)}$, $x^{(3)}$….$x^{(p)}$используется их линейная или иная зависимость от меньшего числа иных факторов, которые непосредственно не измерялись ($f^{(1)}$,…, $f^{(m)}$ ($m < p$)). Данные факторы принято называть общими.

На основании вышеизложенного можно сделать вывод о том, что в широком смысле факторный анализ представляет собой совокупность методов и моделей, основное назначение которых состоит в выявлении, построении и исследовании факторов, определенных на основе их внешних проявлений.

В узком смысле данный термин используется для обозначения методов, которые используются для выявления скрытых, ненаблюдаемых факторов. За счет них объясняются имеющиеся корреляционные матрицы, построенные на основе известных результатов количественных наблюдаемых переменных.

Абсолютное большинство моделей предполагает такое их построение, при котором факторы, определенные по результатам анализа, не коррелируют друг с другом. Также предполагается, что отсутствует возможность однозначного восстановления каждого из значений признака на основе определенных по результатам исследования общих факторов $f(m)$. На каждую переменную также оказывает влияние случайная составляющая *e(j)*. За счет нее и возникает статистическая связь между *x(j)* и *f (m).*

Любой факторный анализ, по сути, сводится к выявлению общих факторов и минимизации степени зависимости признака *x(j)* от собственных случайных компонент *e(j)*. В последующем происходит интерпретация результатов вычисления $f(m)$.

Поскольку факторный анализ является модельной схемой, абсолютное достижение постеленной цели невозможно. Таким образом, происходит приближенное вычисление. Результаты факторного анализа можно считать успешными в том случае, если удалось объяснить как можно большее число исходных переменных минимальным чис-

лом определенных факторов, которые, в зависимости от условий моделирования, могут выполнять роль исходных причин, либо агрегированных теоретических конструкций.

Особенности осуществления факторного анализа должны базироваться на понимании конкретных условий исследования. На практике принято выделять 2 основных уровня:

- Разведочный (эксплораторный) уровень факторного анализа - не известно ни количество факторов, ни структура связи;
- Поверочный (конфирматорный) уровень факторного анализа - осуществляется проверка гипотезы о влиянии факторов.

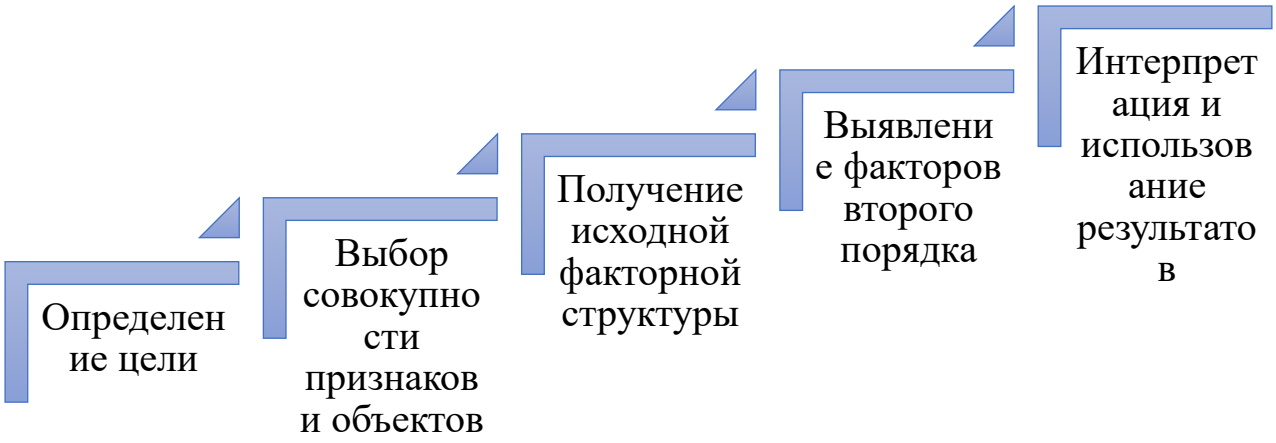


Рис. 6.1.1. Основные этапы факторного анализа

Методы факторного анализа направлены на то, чтобы снизить размерность массива исходных данных, который описывает процесс или явление, на основе сжатия. Достижение данного результата осуществляется, как правило, за счет наличия корреляционных связей в исходном массиве признаков.

Основные этапы осуществления метода представлены на рисунке 6.1.1.

Традиционные представления о факторном анализе при всем многообразии решаемых с его помощью задач сводятся к двум основным положениям:

– Сущность основных процессов заключается в их многообразных, но одновременно с этим простых проявлениях. Они могут быть объяснены при помощи нескольких обобщающих факторов

– Идея постижения сущности наблюдаемых процессов и явлений реализуется с помощью бесконечного приближения.

Основываясь на данных началах, с помощью факторного анализа появляется возможность получения относительно простых моделей, объясняющих взаимосвязи на причинном уровне. Метод факторного анализа активно используется как в теоретических, так и практических исследованиях, в том числе в области экономики.

Для того, чтобы решение задачи факторного анализа стало возможным, необходимо, чтобы соблюдались основные условия применения метода:

- Необходимо установить границы области, наделённой структурой, провести анализ объёма данных, установить уровень различения переменных;

- При выборе переменных важно сохранить возможность их классификации;

- Число переменных p должно соответствовать числу наблюдений n: n >> p.

С помощью факторного анализа становится возможным определение количественных отношений между переменными. Модель можно выразить как при помощи коэффициентов, так и в процентном отношении.

Существуют определенные объекты, по отношению которым возможно проведение процедуры факторного анализа.

К ним можно отнести:

– Вектор отклонений (6.1.1):

$$y = x - \overline{x}, \qquad (6.1.1)$$

причем $E(x - \overline{x}) = E(y = 0)$

В таком случае исходный материал для анализа представлен матрицей ковариаций (6.1.2):

$$K = E(yy') \qquad (6.1.2)$$

– Нормированный вектор (6.1.3):

$$z = S^{-1}(x - \overline{x}) = S^{-1}y \qquad (6.1.3)$$

где S-диагональная матрица стандартных отклонений, E(z = 0).

В таком случае исходный материал представлен ковариационной матрицей R, которая в данном случае равна корреляционной.

– Наблюдаемый вектор X без поправки на среднее.

Матрица наблюдений при традиционном подходе представляется в виде (6.1.4):

$$X = (x_{ik}), \qquad (6.1.4)$$

Где x_{ik} значение k -го признака для i -го объекта, в виде линейных комбинаций значений f_{it} факторов f_t на объектах с невязками e_{ik} (6.1.5):

$$x_{ik} = a_{1k}f_{i1} + ... + a_{tk}f_{it} + e_{ik}, \qquad (6.1.5)$$

причем i = 1….n,

k = 1….p,

t = 1 ….m.

a_{tk} - нагрузки факторов на признак k.

Соотношения факторного анализа формально воспроизводят запись модели множественных регрессий, в которой под $f^{(i)}$ (i = 1,2,…,m) понимаются так называемые объясняющие переменные (факторы-аргументы).

Однако принципиальное отличие модели факторного анализа от регрессионных схем состоит в том, что переменные $f^{(j)}$, выступающие в роли аргументов в моделях регрессии, не являются непосредственно наблюдаемыми в моделях факторного анализа, в то время как в регрессионном анализе значения $f^{(j)}$ измеряются на статистически обследованных объектах.

В матричном виде модель выглядит следующим образом (6.1.6):

$$X = AF + X \qquad (6.1.6)$$

В ходе факторного анализа необходимо оценить минимальное число факторов, определить векторы факторных нагрузок, и вычислить

значения факторов для каждого наблюдаемого объекта. Поскольку число обобщенных факторов предполагается существенно меньше числа исходных признаков, данная задача не имеет однозначного решения. В зависимости от того, какие условия накладываются на обобщенные факторы, существуют различные модели и методы факторного анализа. Можно выделить два основных подхода к построению факторных моделей. В первом случае обобщенные факторы должны выделять большую часть суммарной дисперсии исходных факторов, во втором - обобщенные факторы должны наилучшим образом описывать ковариацию между исходными факторами. Первый метод, по сути, есть рассмотренный ранее метод главных компонент (при условии, что мы строим обобщенные факторы, как линейные комбинации исходных), отличный лишь условиями нормировки, накладываемыми на коэффициенты преобразования. Методы второй группы принято называть каноническими методами факторного анализа.

6.2. Метод главных компонент

Суть метода главных компонент состоит в поиске обобщенных факторов $\mathrm{f}^{(1)}, \mathrm{f}^{(2)}, \ldots, \mathrm{f}^{(\mathrm{m})}$ как линейных комбинаций исходных признаков $\xi_1, \xi_2, \ldots, \xi_{\mathrm{k}}$ (6.2.1):

$$f^{(i)} = \beta^{(i)T}\xi = \beta_1^{(i)}\xi_1 + \beta_2^{(i)}\xi_2 + \ldots + \beta_k^{(i)}\xi_k \qquad (6.2.1)$$

где $\beta^{(i)}$ – векторы неизвестных коэффициентов преобразования, $i = \overline{1,m}$, $(m \le k)$,
$\xi = (\xi_1, \xi_2, \ldots, \xi_k)$ – центрированная многомерная случайная величина с матрицей ковариаций A.

Обобщенные факторы $\mathrm{f}^{(1)}, \mathrm{f}^{(2)}, \ldots, \mathrm{f}^{(\mathrm{m})}$ в рамках реализации метода предполагают некоррелированными случайные величины с нулевым математическим ожиданием и единичной дисперсией. Важно, что выбор линейных комбинаций является процедурой, которая зависит от ряда условий, основное из которых состоит в том, первый из обобщённых факторов объяснял бы наиболее существенную часть дисперсии исходных признаков.

Выбор каждого последующего обобщающего признака базируется на тех же принципах: он должен объяснять максимальную долю оставшейся дисперсии. Кроме того, необходимо, чтобы не наблюдалась межфакторная корреляция.

Результат построения по данной схеме приводит к тому, что в качестве векторов $\beta^{(i)}$, $\mathrm{i}=\overline{1,\mathrm{m}}$ принимаются собственные векторы матрицы ковариаций A, систематизированные в порядке убывания.

Абсолютно линейное преобразование величин ξ_j, $j=\overline{1,k}$ в факторы $f^{(i)}$, $i=\overline{1,k}$ наблюдается, если m=k. В таком случае векторы факторных нагрузок $\alpha^{(\mathrm{i})}$ также должны являться собственными векторами матрицы ковариаций A.

Условный вид нормировок для векторов факторных нагрузок обусловлен тем, что дисперсия, которая объясняется фактором, равна соответствующему собственному значению (6.2.2):

$$\mathrm{D}\left(\sum_{\mathrm{j}=1}^{\mathrm{k}}\alpha_{\mathrm{j}}^{(\mathrm{i})}\mathrm{f}^{(\mathrm{i})}\right)=\sum_{\mathrm{j}=1}^{\mathrm{k}}(\alpha_{\mathrm{j}}^{(\mathrm{i})})^2\mathrm{D}(\mathrm{f}^{(\mathrm{i})})\equiv\lambda_{\mathrm{i}}\quad\Rightarrow\quad|\,\alpha^{(\mathrm{i})}\,|^2=\lambda_{\mathrm{i}}\qquad(6.2.2)$$

Векторы $\beta^{(i)}$ должны удовлетворять условию (6.2.3):

$$\mathrm{D}(\mathrm{f}^{(\mathrm{i})})=\mathrm{D}\left(\sum_{\mathrm{j}=1}^{\mathrm{k}}\beta_{\mathrm{j}}^{(\mathrm{i})}\xi_{\mathrm{j}}\right)=\beta^{(\mathrm{i})^{\mathrm{T}}}\mathrm{A}\beta^{(\mathrm{i})}=1,\ i=\overline{1,m}\qquad(6.2.3)$$

Поскольку данные векторы также являются собственными векторами матрицы ковариаций (6.2.4):

$$|\,\beta^{(\mathrm{i})}\,|^2=1/\lambda_{\mathrm{i}},\quad\alpha^{(\mathrm{i})}=\lambda_{\mathrm{i}}\,\beta^{(\mathrm{i})}\ \text{или}\ \alpha^{(i)}=A\beta^{(i)},\ i=\overline{1,m}\qquad(6.2.4)$$

Кроме того, координаты вектора факторных нагрузок являются ковариациями между фактором и соответствующим исходным параметром (6.2.5):

$$\mathrm{cov}(\xi_{\mathrm{j}},\mathrm{f}^{(\mathrm{s})})=\mathrm{cov}(\xi_{\mathrm{j}},\sum_{\mathrm{i}=1}^{\mathrm{k}}\beta_{\mathrm{i}}^{(\mathrm{s})}\xi_{\mathrm{i}})=\sum_{\mathrm{i}=1}^{\mathrm{k}}\mathrm{cov}(\xi_{\mathrm{j}},\xi_{\mathrm{i}})\beta_{\mathrm{i}}^{(\mathrm{s})}=\alpha_{\mathrm{j}}^{(\mathrm{s})}\qquad(6.2.5)$$

Критерии отбора главных факторов в этом случае аналогичны рассмотренным ранее в методе главных компонент.

6.3. Каноническая модель факторного анализа

Канонической называют такую модель, для которой выполняется ряд условий.

Пусть $A = M(\xi\xi^T)$ - ковариационная матрица исходных признаков, а $\Sigma = M(\varepsilon\varepsilon^T)$ - диагональная матрица ковариаций характерных факторов, тогда справедливо следующее соотношение (6.3.1):

$$A = M((\alpha f + \varepsilon)(\alpha f + \varepsilon)^T) = \alpha M(ff^T)\alpha^T + \Sigma = \alpha\alpha^T + \Sigma \qquad (6.3.1)$$

или (6.3.2)

$$\begin{cases} \operatorname{cov}(\xi_i, \xi_j) = \sum\limits_{s=1}^{m} \alpha_i^{(s)} \alpha_j^{(s)}, \quad i = \overline{1,k}, \quad j = \overline{1,k}, \quad i \neq j \\ D(\xi_i) = \sum\limits_{s=1}^{m} \left(\alpha_i^{(s)}\right)^2 + D(\varepsilon_i), \quad i = \overline{1,k} \end{cases} \qquad (6.3.2)$$

Таким образом, ковариации исходных признаков полностью воспроизводятся матрицей нагрузок, а для воспроизведения их дисперсий нужны также дисперсии характерных факторов. Так как (6.3.3):

$$M(\xi f^T) = M((\alpha f + \varepsilon)f^T) = \alpha M(ff^T) = \alpha, \qquad (6.3.3)$$

то также, как и в методе главных компонент (6.3.4)

$$\operatorname{cov}(\xi_j, f^{(s)}) = \alpha_j^{(s)}, \quad j = \overline{1,k}, \quad s = \overline{1,m} \qquad (6.3.4)$$

Если соблюдаются условия системы уравнений (6.3.2), то модель факторного анализа является канонической. Число уравнений в системе может быть определено по формуле (6.3.5):

$$k(k+1)/2 \qquad (6.3.5)$$

Число неизвестных параметров равно (6.3.6):

$$km + k = k(m + 1). \qquad (6.3.6)$$

В случае, если решение системы существует, оно может рыть рассчитано с точностью до ортогонального преобразования матрицы факторных нагрузок.

Задача факторного анализа обладает единственным решением, если для матрицы факторных нагрузок существуют дополнительные ограничения. Поскольку ортогональное преобразование порядка m однозначно определяется заданием $m(m-1)/2$ элементов, то необходимо, соответственно, обозначить дополнительные $m(m-1)/2$ условий на параметры модели. Обычно полагают, что для элементов матрицы $\alpha = (\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(m)})$ выполнено одно из следующих условий

- матрица $\alpha^T \Sigma^{-1} \alpha$ должна быть диагональной;
- матрица $B^T \alpha$, где B заданная матрица порядков $k \times m$, должна быть нижней треугольной матрицей.

Задача канонического факторного анализа в общем случае не имеет аналитического решения и может быть решена только численно, используя ту или иную итерационную процедуру.

6.4. Параметры факторного моделирования при применении метода максимального правдоподобия

Решение задачи методом максимального правдоподобия предполагает допущение, что обобщенные и характерные факторы имеют нормальное распределение (то есть для исходных признаков характерно многомерное нормальное распределение), а матрица факторных нагрузок удовлетворяет условию (6.4.1):

$$\alpha^T \Sigma^{-1} \alpha = J \qquad (6.4.1)$$

где J представляет собой диагональную матрицу, элементы которой расположены в порядке убывания.

Число дополнительных условий моделирования составляет $m(m-1)/2$. Таким образом, количество условий на параметры теперь

равно $k(k+1)/2+m(m-1)/2$, а число неизвестных параметров по-прежнему равно $k(m+1)$.

Чтобы задача не была неопределенной, требуется, чтобы $k(k+1)/2+m(m-1)/2 \geq k(m+1)$ или $(k-m)^2 \geq (k+m)$.

Так как при $(k-m)^2 = (k+m)$ решение теряет статистическую значимость, то максимальное число параметров для данной модели должно удовлетворять условию (6.4.2):

$$(k-m)^2 > (k+m). \qquad (6.4.2)$$

Метод максимального правдоподобия предполагает, что оценки матрицы факторных нагрузок и дисперсионной матрицы характерных факторов должны удовлетворять системе уравнений (6.4.3):

$$\begin{cases} J\alpha = \Sigma^{-1}(\overline{A}-\Sigma)\alpha \\ \operatorname{diag}\overline{A} = \operatorname{diag}(\alpha\alpha^T + \Sigma) \end{cases} \qquad (6.4.3)$$

Из (6.4.3) следует вывод, что оценки векторов факторных нагрузок $\alpha^{(j)}$ должны быть собственными векторами матрицы $(\overline{A}-\Sigma)\Sigma^{-1}$, а оценки диагональных элементов λ_j матрицы J должны представлять собой соответствующие собственные значения данной матрицы. В качестве собственных векторов берутся m векторов, соответствующих наибольшим собственным значениям λ_j.

Условия нормировки для собственных векторов (6.4.4):

$$\alpha^{(j)T}\Sigma^{-1}\alpha^{(j)} = \lambda_j. \qquad (6.4.4)$$

При заданном числе факторов m решение системы можно осуществить с применением численной итерационной процедуры.

Так как данный метод предполагает оценки правдоподобия, то можно построить критерий отношения правдоподобия для проверки значимости полученной модели (6.4.5):

$$\eta = -\left(n-1-\frac{1}{6}(2k+5)-\frac{2}{3}m\right)\left(\ln\frac{|\overline{A}|}{|\hat{\alpha}\hat{\alpha}^T+\hat{\Sigma}|} - \operatorname{Sp}(\overline{A}(\hat{\alpha}\hat{\alpha}^T+\hat{\Sigma})^{-1}) + k\right) \qquad (6.4.5)$$

При истинности H_0: допустимо представление исходных признаков в виде m-факторной модели, статистика η асимптотически имеет распределение χ^2 с числом степеней свободы и может быть определена по формуле 6.4.6:

$$\nu = \frac{1}{2}((k-m)^2 - (k+m)) \qquad (6.4.6)$$

Если нулевая гипотеза (допустимо представление исходных признаков в виде m- факторной модели) отвергается критерием, то следует перейти к рассмотрению модели с числом обобщенных факторов равным m+1 в том случае, если такое число параметров модели является допустимым.

6.5. Значимость факторных признаков

Зачастую определение факторов анализа не является заключительной целью исследования. Также требуется оценить вес фактора в формировании дисперсии каждого признака, а также определить возможность интерпретации полученных результатов с точки зрения здравого смысла

Пусть имеется m- факторная модель. Пусть $f^{(p)}$ – некоторый фактор, $p \le m$. Общностью фактора $f^{(p)}$ (или вкладом фактора в суммарную дисперсию) признаков называется число $V^{(p)} = \sum_{j=1}^{k} \left[\alpha_j^{(p)}\right]^2$.

Величина $V_0 = \sum_{p=1}^{m} V^{(p)}$ является суммарной общностью факторов $f^{(1)},...,f^{(m)}$. Отношение $V^{(p)}/V_0$ называется долей фактора $f^{(p)}$ в суммарной общности.

Пусть $\alpha_{s_1}^{(p)}, \alpha_{s_2}^{(p)}, \ldots, \alpha_{s_l}^{(p)}$ факторные нагрузки фактора $f^{(p)}$ соответствующие признакам $\xi_{(s_1)},\ldots,\xi_{(s_l)}$. Тогда коэффициент информативности признаков $\xi_{(s_1)},\ldots,\xi_{(s_l)}$ K может быть определен по формуле 6.5.1:

$$K_u = \frac{\sum_{j=1}^{l} \left(\alpha_{s_j}^{(p)}\right)^2}{\left\|\alpha^{(p)}\right\|^2} \qquad (6.5.1)$$

Данное число позволяет определить, какой вклад в признаки $\xi_{(s_1)},\ldots,\xi_{(s_l)}$ вносит в фактор $f^{(p)}$. Если можно выделить какую либо группу признаков, коэффициент информативности которой для фактора $f^{(p)}$ существенно выше коэффициента информативности оставшихся признаков, то первая группа признаков и будет определять содержание фактора.

Предположим, что рассматриваются следующие статистические характеристики деятельности фирмы.

X_1 – уровень выработки в год на одного работника;

X_2 – уровень фондоотдачи

X_3 – размер оборотных производственных средств,

X_4 – размер затрат на выпуск единицы продукции,

X_5 – численность персонала,

X_6 – рентабельность продукции,

X_7 – уровень энерговооруженности труда.

Предполагается, что на предыдущем этапе анализа были определены факторы $f^{(1)}, f^{(2)}$, а также найдены факторные нагрузки $\alpha^{(1)} = (0{,}9; 0{,}8; 0{,}1; 0{,}8; 0{,}3; 0{,}7; 0{,}2)$, $\alpha^{(2)} = (0{,}1; 0{,}4; 0{,}8; 0{,}3; 0{,}7; 0{,}2; 0{,}6)$.

В таком случае коэффициенты нормативности признаков составят:

$$K_{f^{(1)}}(\xi_1, \xi_2, \xi_4, \xi_6) = \frac{0{,}9^2 + 0{,}8^2 + 0{,}8^2 + 0{,}7^2}{0{,}9^2 + 0{,}8^2 + 0{,}1^2 + 0{,}8^2 + 0{,}3^2 + 0{,}7^2 + 0{,}2^2} \approx 0{,}948,$$

$$K_{f^{(2)}}(\xi_3, \xi_5, \xi_7) = \frac{0{,}8^2 + 0{,}7^2 + 0{,}6^2}{0{,}1^2 + 0{,}4^2 + 0{,}8^2 + 0{,}3^2 + 0{,}7^2 + 0{,}2^2 + 0{,}6^2} \approx 0{,}832.$$

Название факторов $f^{(1)}, f^{(2)}$ следует выбирать, опираясь на формирующие их признаки. Фактор $f^{(1)}$ условно можно назвать характеризующим эффективность производства, а фактор $f^{(2)}$ определяющим размеры производства.

Если факторные нагрузки обладают более или менее равномерным распределением, то задача интерпретации фактора усложняется. В этом случае целесообразно прибегать к вращению факторов.

Обобщенные факторы определяются с точностью до ортогонального преобразования. Следовательно, можно реализовывать ортогональное вращение факторного пространства, добиваясь такого расположения осей, при котором факторы допускают наиболее содержательную интерпретацию.

Как правило, используют две основные стратегии вращения. Формулы для их вычисления приведены в таблице 6.5.1.

Таблица 6.5.1

Методы вращения в факторном анализе

Применяемых метод	Сущность подхода в рамках реализаци метода	Формула
Метод варимакс	При вращении максимизируют величину, которая характеризует различие столбцов матрицы факторных нагрузок	$V = \sum_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} \left(\left(\alpha_i^{(j)}\right)^2 - \frac{1}{k} \sum_{s=1}^{k} \left(\alpha_s^{(j)}\right)^2 \right)^2$
Метод квартимакс	При вращении максимизируют величину, которая характеризует различие строк матрицы факторных нагрузок	$Q = \sum_{i=1}^{k} \frac{1}{m} \sum_{j=1}^{m} \left(\left(\alpha_i^{(j)}\right)^2 - \frac{1}{m} \sum_{s=1}^{m} \left(\alpha_i^{(s)}\right)^2 \right)^2$

Учитывая, что метод варимакс основан на упрощении столбцов, а квартимакс на упрощении строк матрицы факторных нагрузок,

можно построить обобщенный метод, основанный на максимизации величины $\alpha Q + \beta V$. В частности, при $\alpha = \beta$ получим критерий, используемый в методе вращения биквартимакс.

Если C представляет собой матрицу поворота, то в новой системе координат матрица факторных нагрузок определится как: $\alpha' = \alpha C$. Обычно в процедуре вращения осуществляют последовательное попарное ортогональное вращение осей. В этом случае полная матрица преобразования будет равна произведению матриц попарных поворотов. Например, для трехфакторной модели при вращении против часовой стрелки будут справедливы следующие формулы (6.5.2 и 6.5.3)

$$C = C_{12} * C_{23} * C_{31} \qquad (6.5.2)$$

$$C = \begin{pmatrix} \cos\varphi_1 & -\sin\varphi_1 & 0 \\ \sin\varphi_1 & \cos\varphi_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi_2 & -\sin\varphi_2 \\ 0 & \sin\varphi_2 & \cos\varphi_2 \end{pmatrix} \cdot \begin{pmatrix} \sin\varphi_3 & 0 & \cos\varphi_3 \\ 0 & 1 & 0 \\ \cos\varphi_3 & 0 & -\sin\varphi_3 \end{pmatrix} \qquad (6.5.3)$$

Для того, чтобы избежать влияния на результаты вращения переменных с большой общностью, осуществляют нормировку факторных нагрузок. Производят данную процедуру путем деления координат векторов факторных нагрузок на соответствующие корни из общностей [107].

После того как найдены оценки параметров факторной модели, как правило, необходимо вычислить оценки значений обобщенных факторов для каждого наблюдения. Эти значения могут быть использованы, например, в задачах классификации или в задачах регрессионного анализа.

Если для факторного анализа использовался метод главных компонент, то задача оценивания значений факторов решается путем выражения главных факторов через линейную комбинацию исходных признаков (6.5.4):

$$f^{(j)} = \beta^{(j)T}\xi = \beta_1^{(j)}\xi_1 + \beta_2^{(j)}\xi_2 + \ldots + \beta_k^{(j)}\xi_k, \quad j = \overline{1,m}, \qquad (6.5.4)$$

где векторы $\beta^{(j)}$, $j=\overline{1,m}$ - собственные векторы матрицы ковариаций A, упорядоченные по убыванию собственных значений λ_j и удовлетворяющие условию $|\beta^{(i)}|^2=1/\lambda_i$.

Пусть X матрица значений исходных признаков размеров $n\times k$, B - матрица, составленная из оценок векторов $\beta^{(j)}$, $j=\overline{1,m}$ размером $k\times m$, F- матрица оценок значений обобщенных факторов размеров $n\times m$, тогда (6.5.5):

$$F = XB \qquad (6.5.5)$$

Если, кроме того, использовалась процедура вращения факторов с матрицей преобразования C, то справедлива формула 6.5.6:

$$F = XBC \qquad (6.5.6)$$

Для оценивания значений факторов канонической модели факторного анализа используют либо метод Бартлетта, либо метод Томпсона.

Согласно методу Бартлетта модель для каждого наблюдения $X^{(i)}, i=\overline{1,n}$ рассматривается как регрессия величины $X^{(i)}$ на факторы $\alpha^{(j)}, j=\overline{1,m}$ с неизвестными коэффициентами $f_i^{(j)}$ (6.5.7):

$$X^{(i)} = f_i^{(1)}\alpha^{(1)} + f_i^{(2)}\alpha^{(2)} + \ldots + f_i^{(m)}\alpha^{(m)} + \varepsilon \qquad (6.5.7)$$

Соответственно для значения $X_j^{(i)}$, которое, по сути, есть значение признака ξ_j в i-ом наблюдении справедливо (6.5.8):

$$X_j^{(i)} = f_i^{(1)}\alpha_j^{(1)} + f_i^{(2)}\alpha_j^{(2)} + \ldots + f_i^{(m)}\alpha_j^{(m)} + \varepsilon_j, \quad j=\overline{1,k}, \qquad (6.5.8)$$

Для оценки факторов следует применять обобщенный метод наименьших квадратов, согласно которому:

$$\hat{f}_i = (\hat{f}_i^{(1)}, \hat{f}_i^{(2)}, \ldots, \hat{f}_i^{(m)})^T = (\alpha^T\Sigma^{-1}\alpha)^{-1}\alpha^T\Sigma^{-1}X^{(i)} \qquad (6.5.9)$$

В методе Томпсона предполагается, что оценки обобщенных факторов есть линейная комбинация исходных признаков, коэффициенты которой определяются из условия минимума среднеквадратичного отклонения оценки от фактора, то есть (6.5.10):

$$\hat{f}^{(j)} = \beta^{(j)T}\xi = \xi^T\beta^{(j)} = \beta_1^{(j)}\xi_1 + \beta_2^{(j)}\xi_2 + \ldots + \beta_k^{(j)}\xi_k, \quad j = \overline{1,m}, \qquad (6.5.10)$$

где вектор коэффициентов удовлетворяет условию (6.5.11):

$$M(f^{(j)} - \hat{f}^{(j)})^2 \to \min \qquad (6.5.11)$$

Таким образом, коэффициенты являются коэффициентами линейной среднеквадратичной регрессии и могут быть выражены через ковариации величин $\xi_1, \xi_2, \ldots, \xi_k$ и $f^{(j)}$ 6.5.12)

$$\beta^{(j)} = A^{-1}(\xi)M(f^{(j)} \cdot \xi) \qquad (6.5.12)$$

здесь: A - матрица ковариаций величин $\xi_1, \xi_2, \ldots, \xi_k$, $M(f^{(j)} \cdot \xi)$ - вектор ковариаций величины $f^{(j)}$ и величин $\xi_1, \xi_2, \ldots, \xi_k$.

Соответственно (6.5.13)

$$\hat{f}^{(j)} = \xi^T A^{-1}(\xi)M(f^{(j)} \cdot \xi) \qquad (6.5.13)$$

Учитывая, что для канонической модели факторного анализа $A(\xi) = \alpha\alpha^T + \Sigma$ и $M(f^{(j)} \cdot \xi) = \alpha^{(j)}$ получим следующую оценку для $f^{(j)}$ (6.5.14):

$$\hat{f}^{(j)} = \xi^T(\alpha\alpha^T + \Sigma)^{-1}\alpha^{(j)}, \quad j = \overline{1,m} \qquad (6.5.14)$$

Соответственно оценка для вектора обобщенных факторов:

$$\hat{f}^T = \xi^T(\alpha\alpha^T + \Sigma)^{-1}\alpha \qquad (6.5.15)$$

Рассмотренные в рамках данного параграфа положения отражают основные положения факторных оценок и могут использоваться для оценки значимости результатов факторного анализа.

6.6. Прогнозирование на основе факторного анализа в MS Excel

Рассмотрим пример проведения факторного анализа в данном программном комплексе.

Пример

Имеются данные по компании, функционирующее на рынке, условно свободном от воздействия иных компаний (монополия отсутствует). Необходимо составить прогноз объема сбыта для данной компании и выявить степень риска при принятии решения.

Исходные данные для проведения анализа представлены в таблице 6.6.1.

Таблица 6.6.1

Исходные данные для анализа

		Факторы влияния		
Период	Объем реализации	F1	F2	F3
1	22	21	11	222
2	33	33	1	455
3	54	44	2	555
4	33	55	66	455
5	21	76	33	566
6	33	98	21	559
7	43	101	32	333
8	44	110	88	455
9	55	121	10	677

Решение

Необходимо определить какие из существующих факторов действительно влияют на изменение объемом продаж. Остальные необходимо исключить.

Оценить данное соответствие становится возможным посредством коэффициента корреляции, который позволит оценить насколько связано (если связано) распределение по времени факторов с объёмом реализации.

Расчет коэффициента корреляции производится с использованием формулы =КОРРЕЛ. Полученные коэффициенты представлены в таблице 6.6.2.

Таблица 6.6.2

Коэффициенты корреляции

	F1	F2	F3
Коэффициент корреляции	0,651996	-0,0574	0,591953

Анализируя данные, представленные в таблице, становится возможным сделать вывод, что второй фактор отсеивается ввиду малой силы связи.

Далее необходимо спрогнозировать значения оставшихся факторов (1 и 3). Это возможно сделать при помощи функции ТЕНДЕНЦИЯ.

Спрогнозируем данные на три периода вперед.

Для этого необходимо продлить число периодов на 3, до 12 соответственно. В первую пустую ячейку фактора 1 ввести функцию ТЕНДЕНЦИЯ и заполнить соответствующие значения в ней (рис. 6.6.1)

Период	Объем реализации	F1	F3
1	22	21	222
2	33	33	455
3	54	54	555
4	33	55	455
5	21	42	566
6	33	98	559
7	43	99	605
8	44	110	455
9	55	121	677
10	=ТЕНДЕНЦИЯ(C17:C25;A17:A25;A26:A28)		
11			
12			

Рис. 6.6.1. Ввод функции ТЕНДЕНЦИЯ для первого фактора

Отметим, что сначала вводится известные значения фактора, далее номера периодов с известными данными, далее номера периодов, на которые строится прогноз.

Аналогичную операцию проделать с первой пустой ячейкой третьего фактора (рис. 6.6.2).

Период	Объем реализации	F1	F3
1	22	21	222
2	33	33	455
3	54	54	555
4	33	55	455
5	21	42	566
6	33	98	559
7	43	99	605
8	44	110	455
9	55	121	677
10	=ТЕНДЕНЦИЯ(D17:D25;A17:A25;A26:A28)		
11			
12			

Рис. 6.6.2. Ввод функции ТЕНДЕНЦИЯ для третьего фактора

Период	Объем реализации	F1	F3
1	22	21	222
2	33	33	455
3	54	54	555
4	33	55	455
5	21	42	566
6	33	98	559
7	43	99	605
8	44	110	455
9	55	121	677
=ТЕНДЕНЦИЯ(B17:B25;A17:A25;A26:A28)			674
11			
12			

Рис. 6.6.3. Ввод функции ТЕНДЕНЦИЯ для объема реализации

Далее аналогичная операция для объема реализации (рис. 6.6.3).

Таким образом получены значения для 10 периода. Путем аналогичных операций заполним 11 и 12 периодов.

В результате будет получена таблица 6.6.3.

Таблица 6.6.3

Таблица данных после ввода функции ТЕНДЕНЦИЯ

Период	Объем реализации	F1	F3
1	22	21	222
2	33	33	455
3	54	54	555
4	33	55	455
5	21	42	566
6	33	98	559
7	43	99	605
8	44	110	455
9	55	121	677
10	49	134	674
11	52	147	708
12	54	159	742

Таким образом задача решена получены все необходимые данные на будущие периоды.

6.7. Факторный анализ в Statistica

Рассмотрим особенности реализации анализа и возможности программного комплекс на примере.

Исходные данные представлены в таблице 6.7.1.

Таблица 6.7.1

Исходные данные для анализа

Номер партии	x1	x2	x3	x4	x5
P1	1531,9	41775	165672	735	6,3
P2	1168,8	35582	81337	322	4,9
P3	1323,7	39550	103846	357	4,8
P4	2287,7	40830	285892	368	12,2
P5	976,9	32403	44981	337	4,4
P6	1012,8	48837	128508	334	4,1
P7	620,8	35967	42743	138	2,4
P8	1083,6	40292	193352	376	5,7
P9	1113,7	40188	179400	545	4,7
P10	7768,9	64041	1144660	799	29,1
P11	714,1	35754	60612	385	3,3
P12	1085,2	40631	73886	270	5,7
P13	909,8	36529	70327	321	4,9
P14	981,0	34438	79397	292	3,9
P15	1230,1	40286	84293	252	5,8
P16	1432,6	44726	182297	405	5,5
P17	1227,3	41209	109967	276	6,8
P18	12635,5	112768	4839918	2548	73,5
P19	3153,8	31859	251368	427	12,7
P20	524,1	31362	20435	854	2,3
P21	870,5	31712	51063	609	3,9
P22	464,2	32846	28083	345	1,9
P23	688,1	32999	34554	717	4,7
P24	1516,4	31291	84279	642	4,3
P25	2780,2	37387	254164	275	12,3
P26	4001,6	42848	419337	310	17,4
P27	671,5	35497	35546	223	2,5
P28	770,7	34499	49690	296	4,3
P29	3886,4	45800	683305	448	16,5
P30	1484,5	39791	117156	253	7,4
P31	1198,4	35799	61325	433	6,0
P32	2556,8	46267	307824	137	13,3
P33	1234,8	36143	75540	114	6,0
P34	3144,2	41369	385625	300	15,3
P35	1924,6	38357	198131	168	9,2
P36	1274,1	36031	96202	291	5,6
P37	3131,7	42771	364151	330	15,2
P38	2361,0	37408	173054	170	12,3
P39	1204,0	36126	108493	254	5,2

Задачей факторного анализа является объединение большого количества показателей, признаков, которыми характеризуется исследуемый объект, в меньшее количество искусственно построенных на их основе факторов, чтобы полученная в итоге система факторов (столь же хорошо описывающая выборочные данные, что и исходная) была наиболее удобна с точки зрения содержательной интерпретации.

Для вызова меню настройки факторного анализа необходимо выбрать раздел программы Анализ – Многомерный анализ – Факторный анализ (рис. 6.7.1).

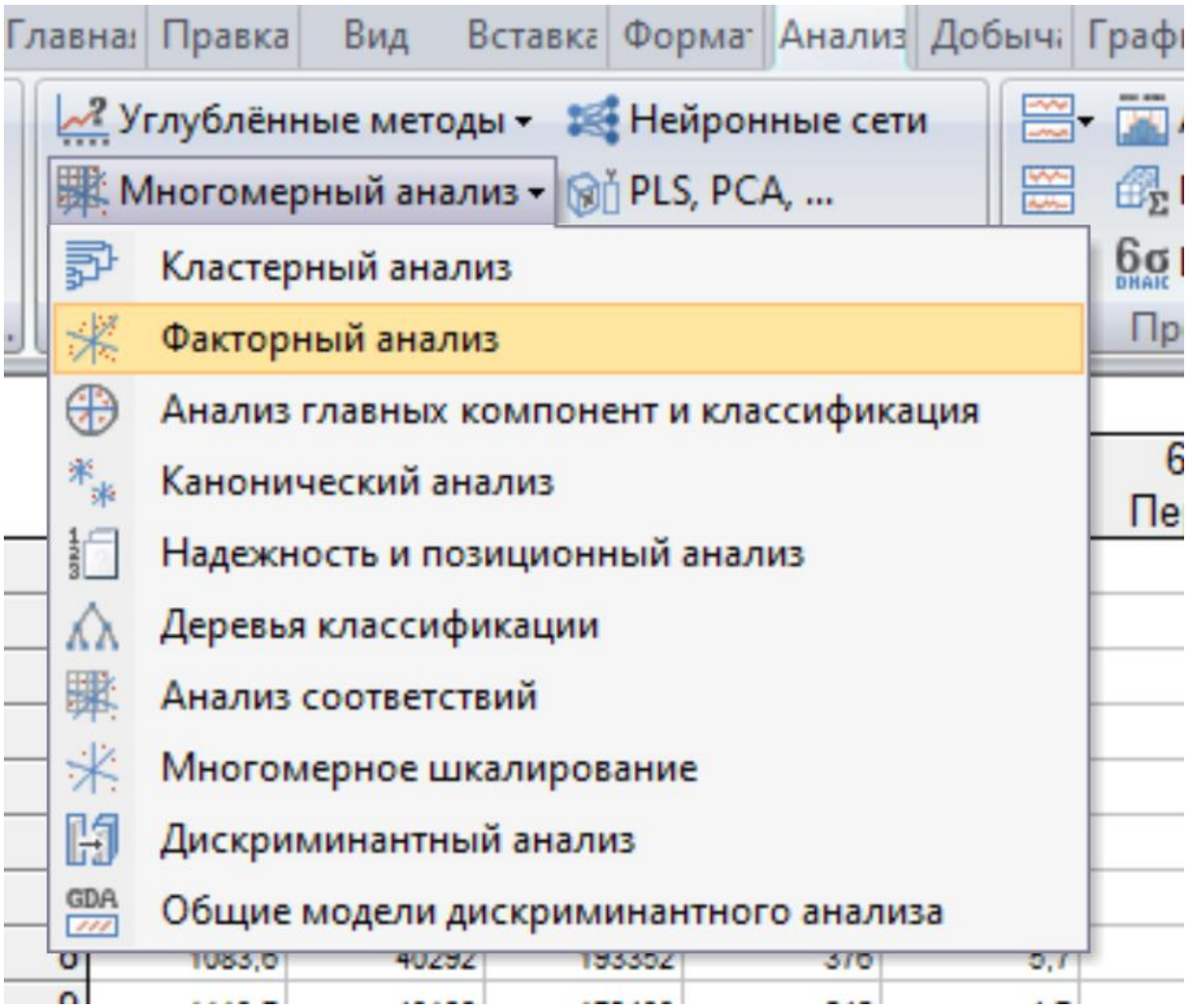


Рис. 6.7.1. Выбор модуля факторного анализа

Появится окно настройки факторного анализа (рис. 6.7.2).

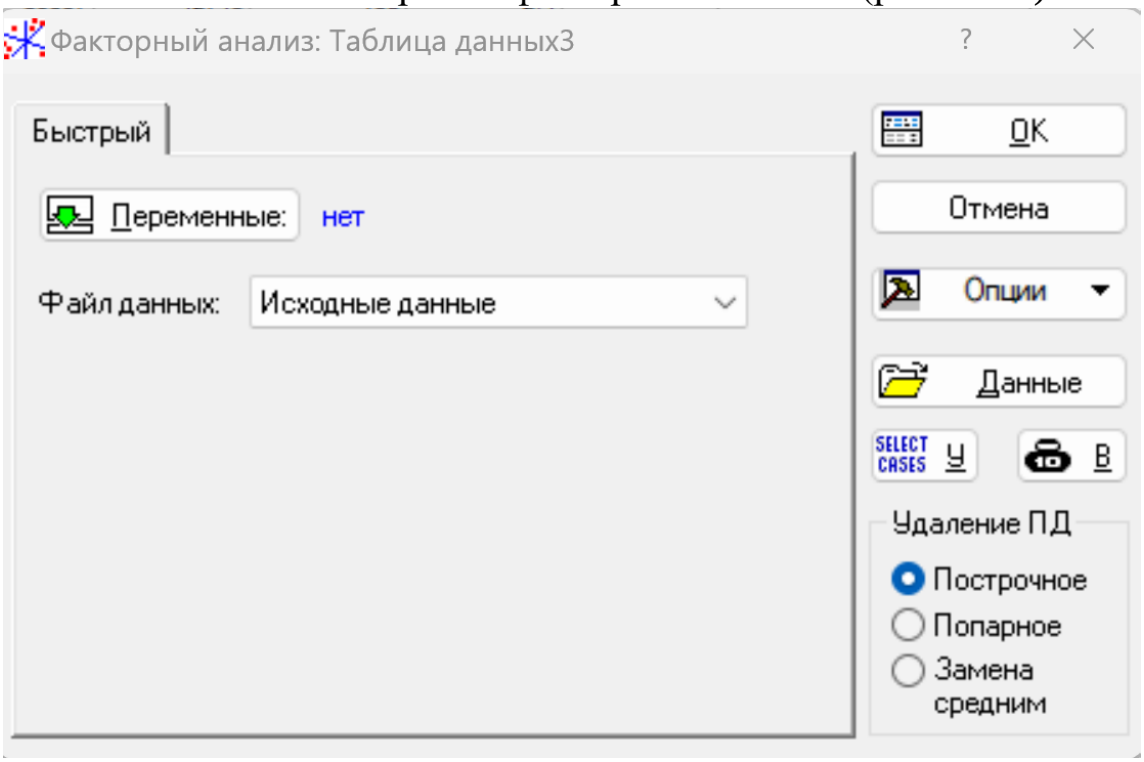


Рис. 6.7.2. Настройка факторного анализа

Перед началом проведения факторного анализа необходимо указать переменные. Для этого необходимо нажать на кнопку Переменные и выбрать необходимые (рис. 6.7.3).

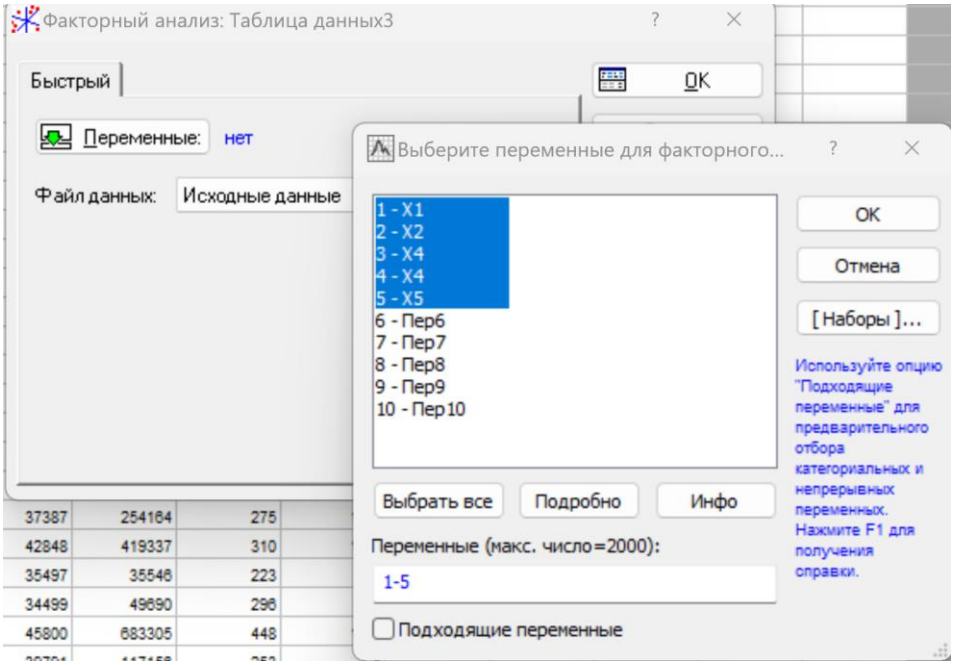


Рис. 6.7.3. Выбор переменных для факторного анализа

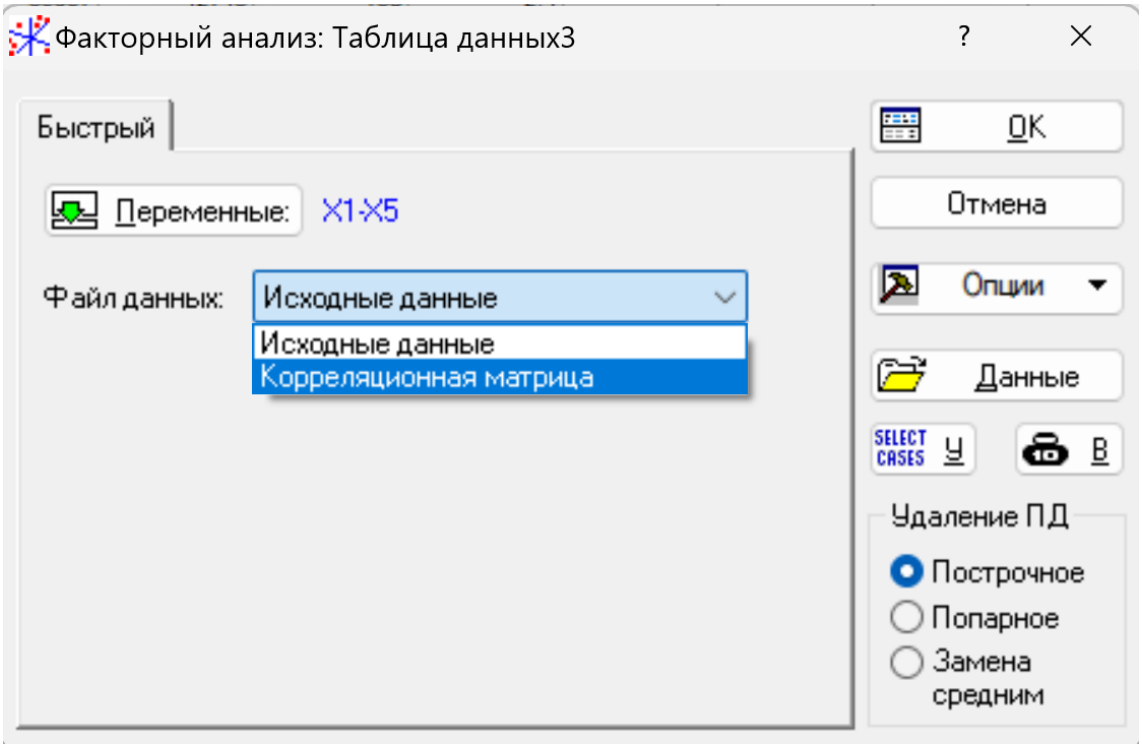


Рис. 6.7.4. Выбор файла данных

Следует отметить, что исходными данными также могут служить корреляционные матрицы. Для использования данного типа исходных данных необходимо в пункте Файл данных выбрать Корреляционная матрица (рис. 6.7.4).

В нашем примере анализ будет производиться по исходным данным.

В правом нижнем углу окна есть раздел Удаление ПД (Замена пропущенных переменных). Путем установки переключателя можно выбрать тот способ обработки пропущенных значений, который наиболее подходит для данного анализа:

- **Построчное** (исключение пропущенных случаев) – в электронной таблице, содержащей данные, игнорируются все строки (случаи), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. В таблице остаются только случаи, в которых нет ни одного пропуска.

- **Попарное** (парное исключение пропущенных значений) – игнорируются пропущенные случаи не для всех переменных, а лишь для выбранной пары. Все случаи, в которых нет пропусков, используются в обработке, например, при поэлементном вычислении корреляцион-

ной матрицы, когда последовательно рассматриваются все пары переменных. Очевидно, в способе Попарное остается больше наблюдений для обработки, чем в способе Построчное. Тонкость, однако, состоит в том, что в способе Попарное оценки различных коэффициентов корреляции строятся по разному числу наблюдений.

- **Замена средним** (подстановка среднего вместо пропущенных значений).

После выбора переменных и нажатия кнопки **ОК** начинается непосредственно анализ (рис. 6.7.5).

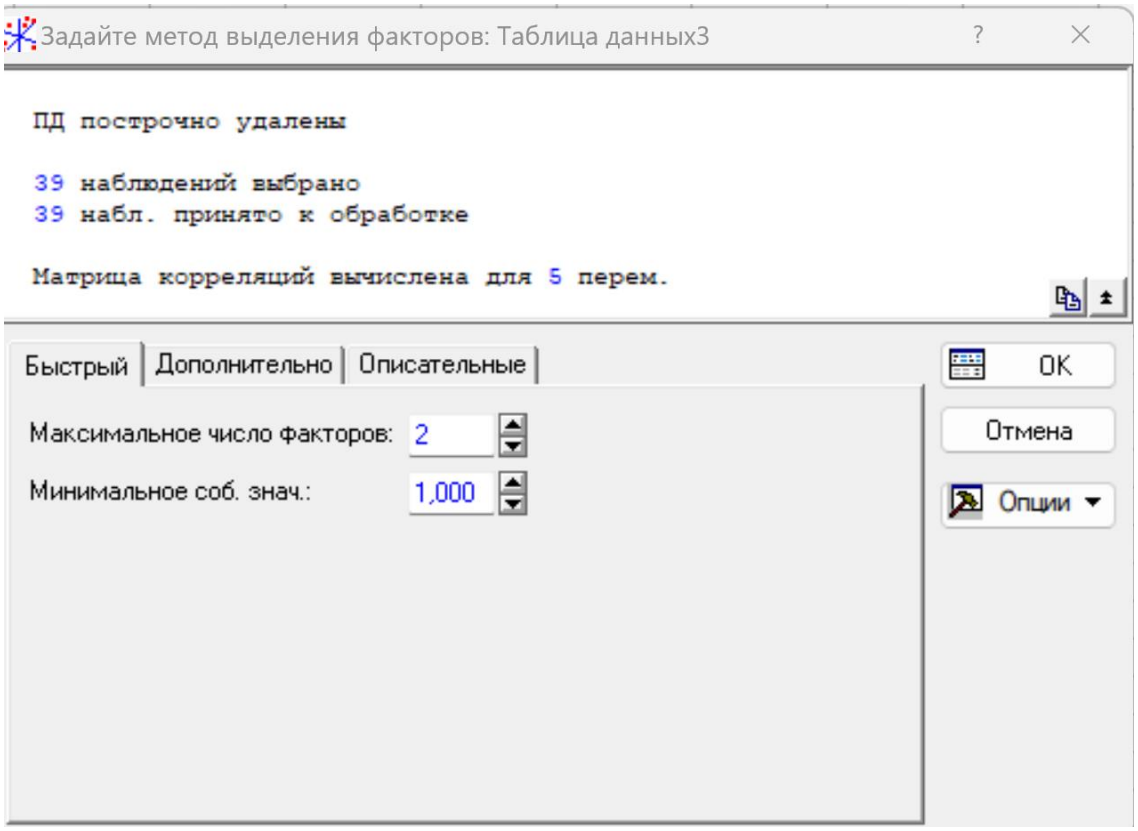


Рис. 6.7.5. Выбор метода выделения факторов

В верхней части формы представлена информация о методе обработки данных, число обработанных случаев, число переменных, для которых исчислена матрица корреляций.

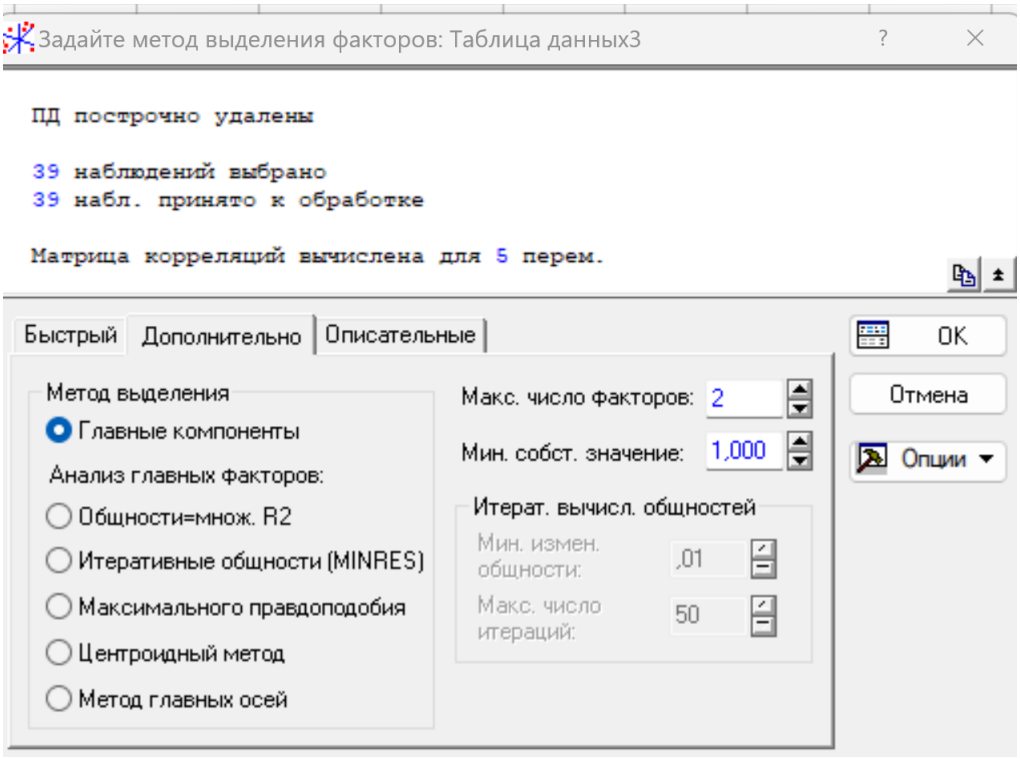


Рис. 6.7.5. Выбор метода выделения факторов

Во вкладке Дополнительно (рис. 6.7.6) представлен выбор настраиваемых параметров.

В части Метод выделения можно выбрать метод обработки.

Главные компоненты - позволяет выделить компоненты, работая с первоначальной матрицей корреляций.

Общности=множ. R2 - на диагонали матрицы корреляций будут находиться оценки квадрата коэффициента множественной корреляции R^2 (соответствующей переменной со всеми другими переменными).

Итеративные общности - метод минимальных остатков - выполняется в два этапа. Сначала оценки квадрата коэффициента множественной корреляции R2 используются для определения общностей, как в предыдущем методе. После первоначального выделения факторов метод корректирует их нагрузки с помощью метода наименьших квадратов с целью минимизировать остаточные суммы квадратов.

Максимального правдоподобия - в этом методе считается заранее известным число факторов (оно устанавливается в поле ввода максимального числа факторов). STATISTICA оценит нагрузки и общности, которые максимизируют вероятность наблюдаемой в таком случае матрицы корреляций. В диалоговом окне результатов анализа доступен

χ-квадрат тест для проверки справедливости принятой гипотезы о числе общих факторов.

Центроидный метод - основан на геометрическом подходе.

Метод главных осей - основан на итеративной процедуре вычисления общностей по текущим собственным значениям и собственным векторам. Итерации продолжаются до тех пор, пока не превышено максимальное число итераций или минимальное изменение в общностях больше, чем это определено в соответствующем поле.

Максимальное число факторов. Заданное в этом поле число определяет, сколько факторов может быть выделено при работе рассмотренных выше методов. Это поле работает вместе с полем Минимальное собственное значение. Часто при заполнении этого поля руководствуются критерием Кайзера, который рекомендует использовать лишь те факторы, для которых собственные значения не меньше 1.

В методе главных компонент по умолчанию предполагается, что дисперсии всех переменных равны 1. Соответственно, общая дисперсия равна числу переменных и максимально возможное число факторов тоже равно числу переменных.

Остальные поля доступны только при выбранном методе **Центроидный метод** или **Метод главных осей**, и определяют необходимые для успешного выполнения последовательных итераций параметры минимального изменения в общностях и максимального числа итераций.

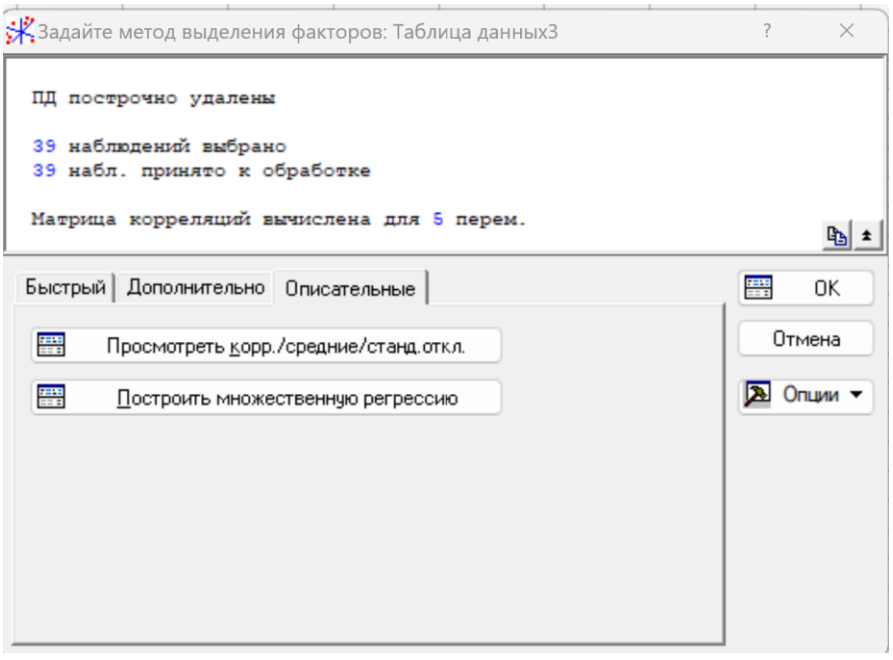


Рис. 6.7.6. Вкладка Описательные

На вкладке **Описательные** (рис. 6.7.6) есть две кнопки:

- **Просмотреть корр./средние/станд.откл** - открывается окно просмотра описательных статистик для анализируемых данных (Рис. 6.7.7), где можно посмотреть средние, стандартные отклонения, корреляции, ковариации, построить различные графики. Здесь можно провести дополнительный анализ текущих данных, проверить соответствие выборочных переменных нормальному закону распределения и существование линейной корреляции между переменными.

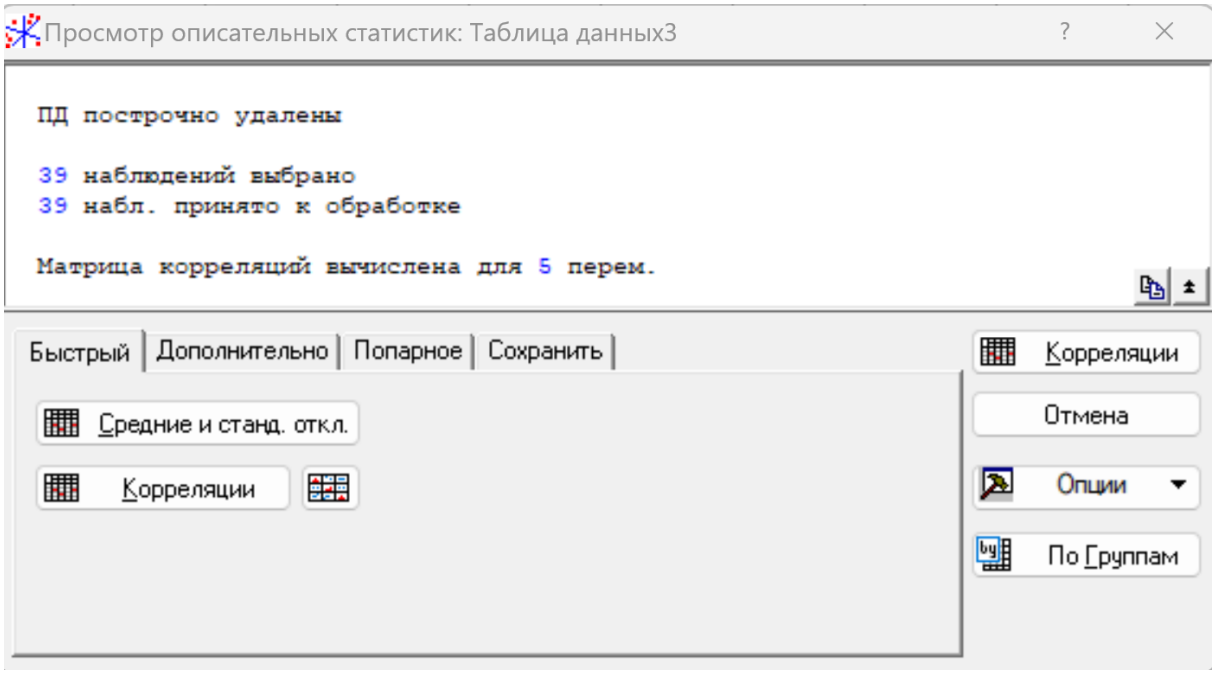


Рис. 6.7.7. Вкладка **Просмотреть корр./средние/станд.откл**

Вкладка Дополнительно (рис. 6.7.8) предоставляет широкий выбор аналитических инструментов.

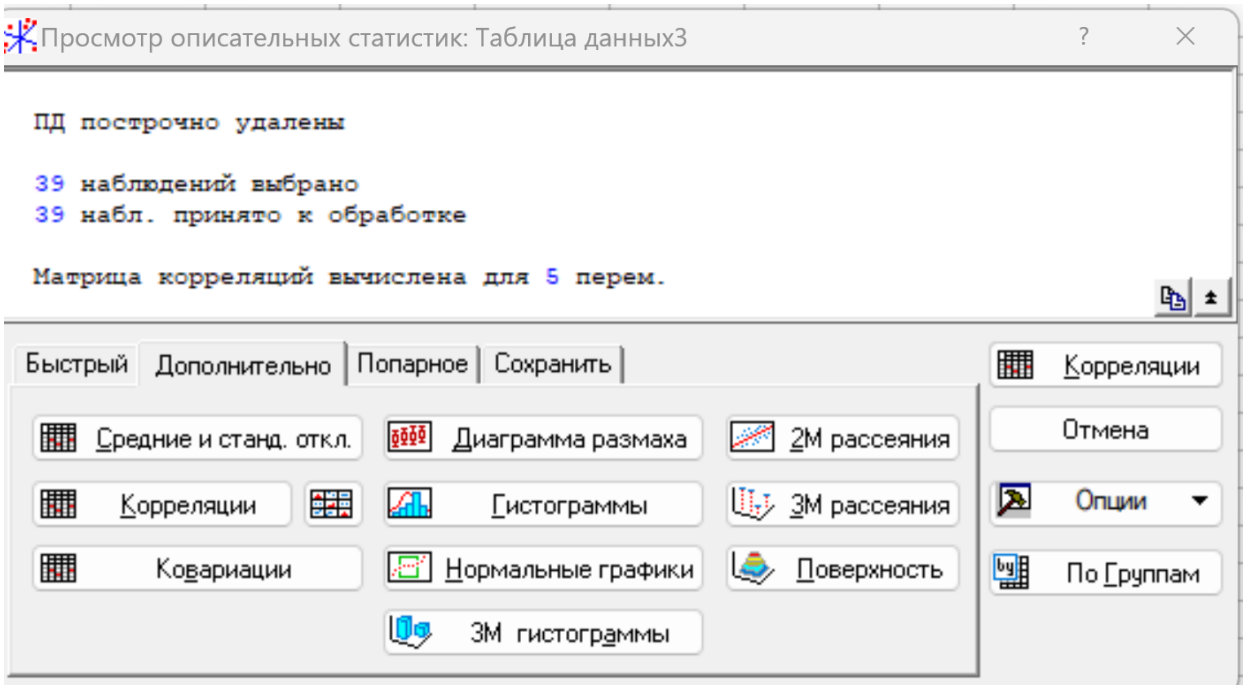


Рис. 6.7.8. Вкладка **Дополнительно**

Кнопка **Средние и стандартные отклонения** выводят соответствующую табличную форму (рис. 6.7.9).

Средние и стандартные отклонения (Таблица данных3)
Построчное удаление ПД
N=39

Переменная	Средние	Стд.Откл
X1	1998,6	2211,6
X2	40614,6	13303,9
X4	299241,4	774732,1
X4	427,0	392,2
X5	9,4	11,9

Рис. 6.7.9. **Средние и стандартные отклонения**

Кнопка Корреляции (доступен выбор переменных для анализа) выводит квадратную матрицу корреляций (рис. 6.7.10).

Переменная	Корреляции (Таблица данных3) Построчное удаление ПД N=39				
	X1	X2	X4	X4	X5
X1	1,00	0,91	0,92	0,76	0,98
X2	0,91	1,00	0,96	0,81	0,94
X4	0,92	0,96	1,00	0,89	0,97
X4	0,76	0,81	0,89	1,00	0,81
X5	0,98	0,94	0,97	0,81	1,00

Рис. 6.7.10. Квадратная матрица корреляций

Кнопка Ковариации выводит соответствующую табличную форму (рис. 6.7.11).

Переменная	Ковариации (Таблица данных3) Построчное удаление ПД N=39				
	X1	X2	X4	X4	X5
X1	4,891079E+06	2,670859E+07	1,575710E+09	655609	25733
X2	2,670859E+07	1,769942E+08	9,869743E+09	4229984	148708
X4	1,575710E+09	9,869743E+09	6,002098E+11	269366590	8937372
X4	6,556086E+05	4,229984E+06	2,693666E+08	153805	3771
X5	2,573309E+04	1,487077E+05	8,937372E+06	3771	142

Рис. 6.7.11. Ковариационная матрица

При нажатии кнопки Диаграмма размаха предоставляется выбор из нескольких параметров для построения необходимого графического элемента (рис. 6.7.12).

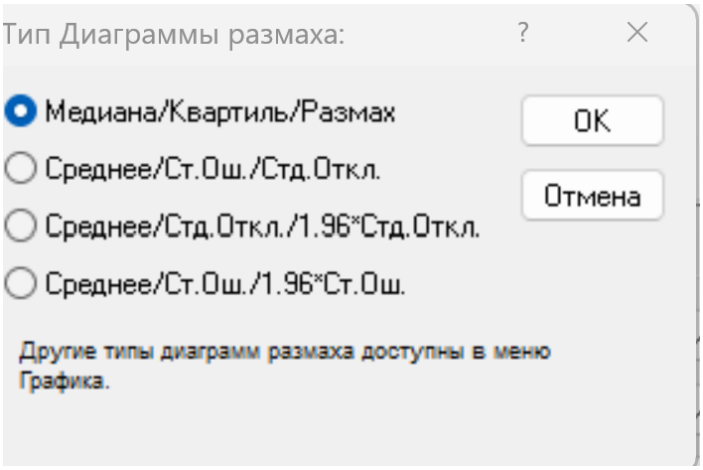


Рис. 6.7.12. Выбор параметров Диаграмм размаха

После выбора необходимых параметров и нажатия кнопки ОК выводится диаграмма размаха (рис. 6.7.13).

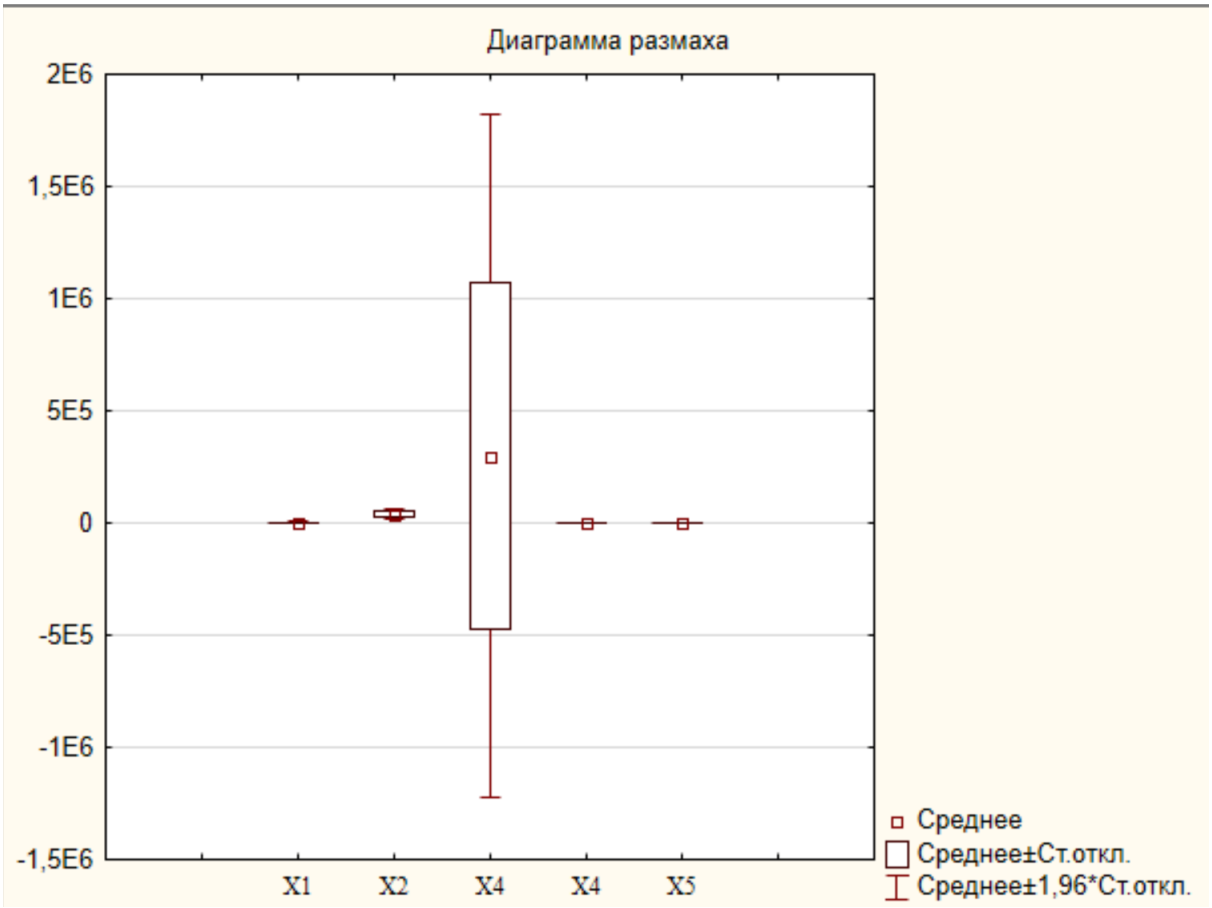


Рис. 6.7.13. Диаграмма размаха

После нажатия кнопки Гистограммы и выбора необходимой переменной происходит построение гистограммы (рис. 6.7.14).

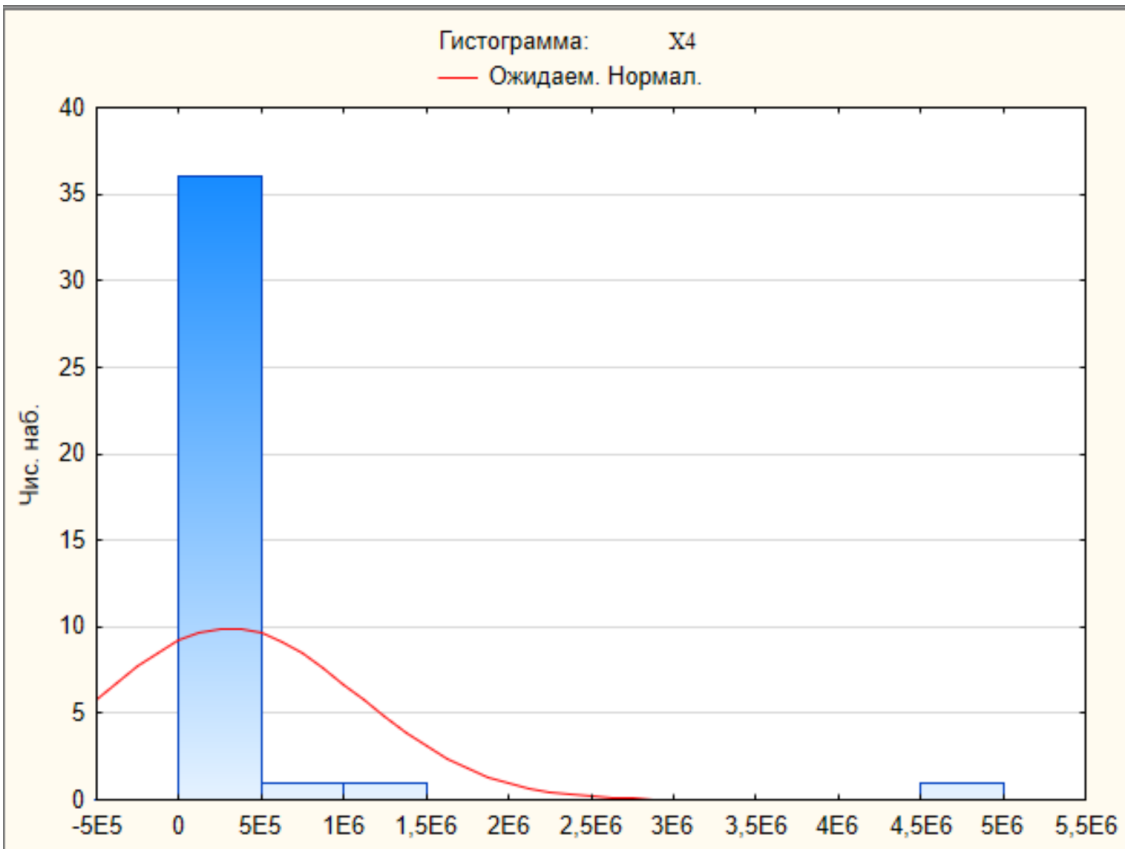


Рис. 6.7.14. Гистограмма

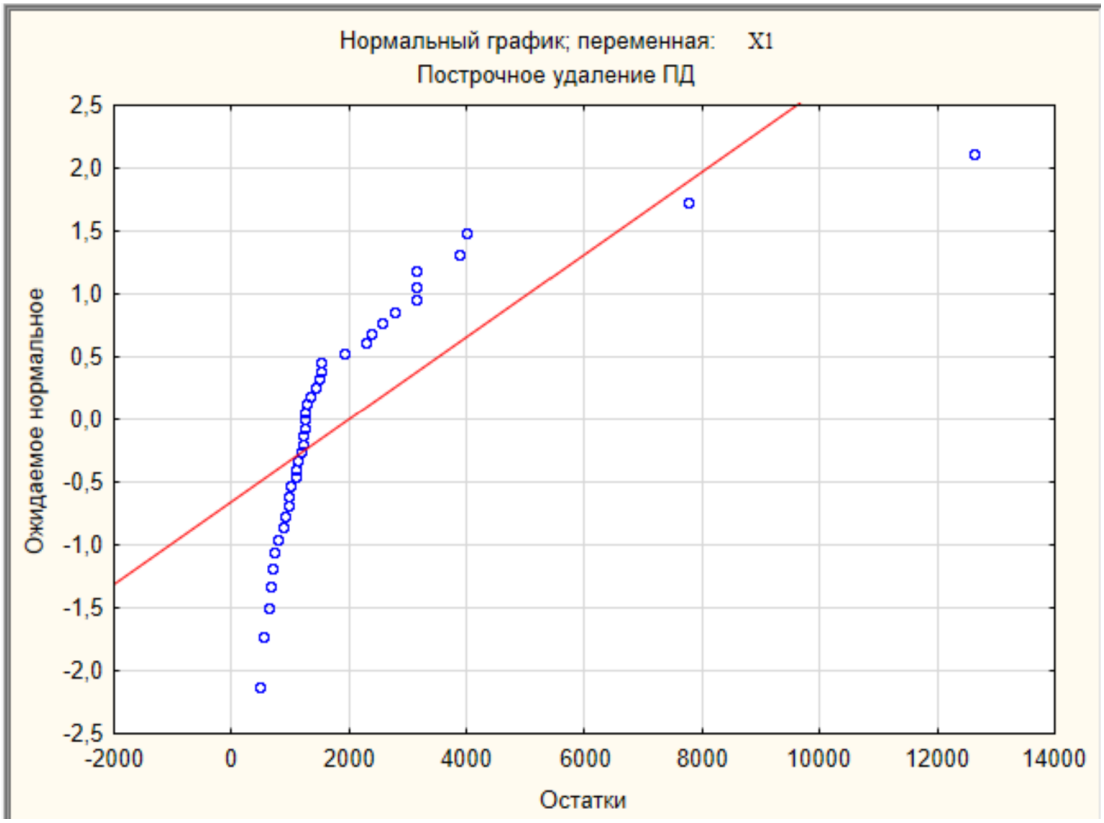


Рис. 6.7.15. Нормальный график переменной

Кнопка Нормальные графики служит для построения соответствующего графика для выбранной переменной (рис. 6.7.15).

Кнопка 3М гистограммы служит для построения трехмерной графической интерпретации данных. Для этого необходимо нажать на кнопку и задать переменный (рис. 6.7.16).

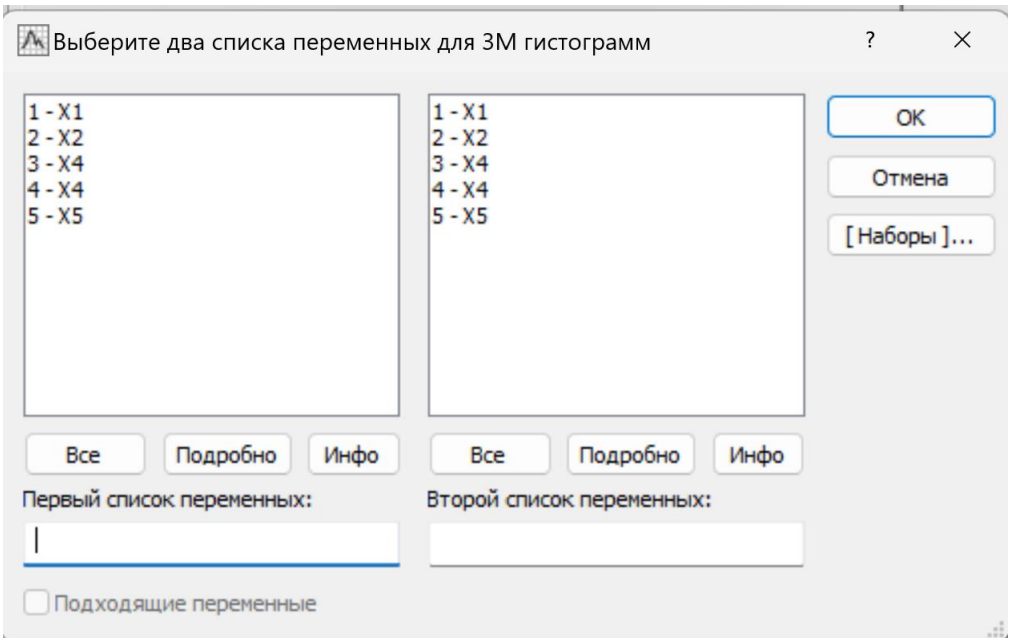


Рис. 6.7.16. Выбор переменных для построения 3М графика

После выбора необходимых переменных и нажатия кнопки ОК будет выведен соответствующий график (рис. 6.7.17).

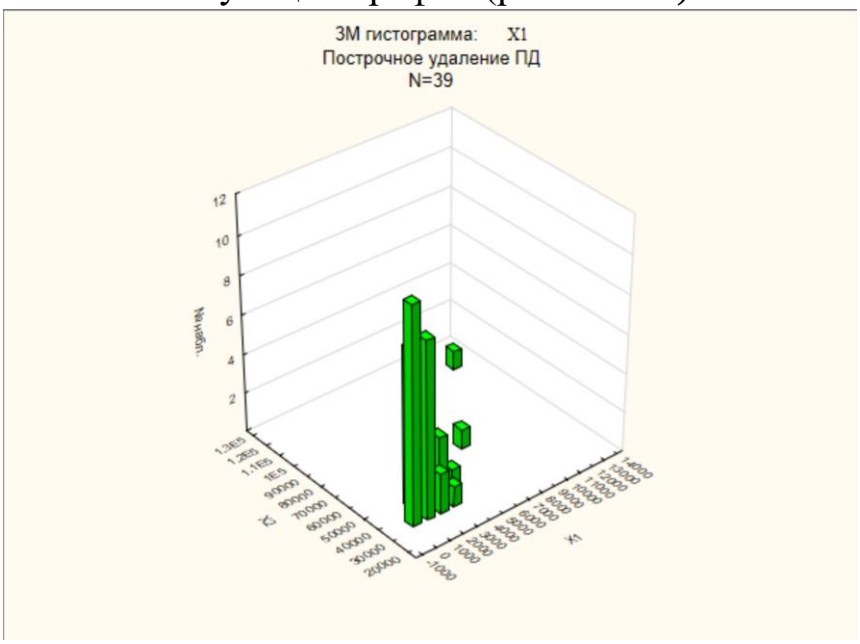


Рис. 6.7.17. Выбор переменных для построения 3М графика

Также возможно построение 2М диаграмм рассеяния (после выбора необходимых переменных) – рис. 6.7.18 и 3М диаграмм рассеяния – рис. 6.7.19.

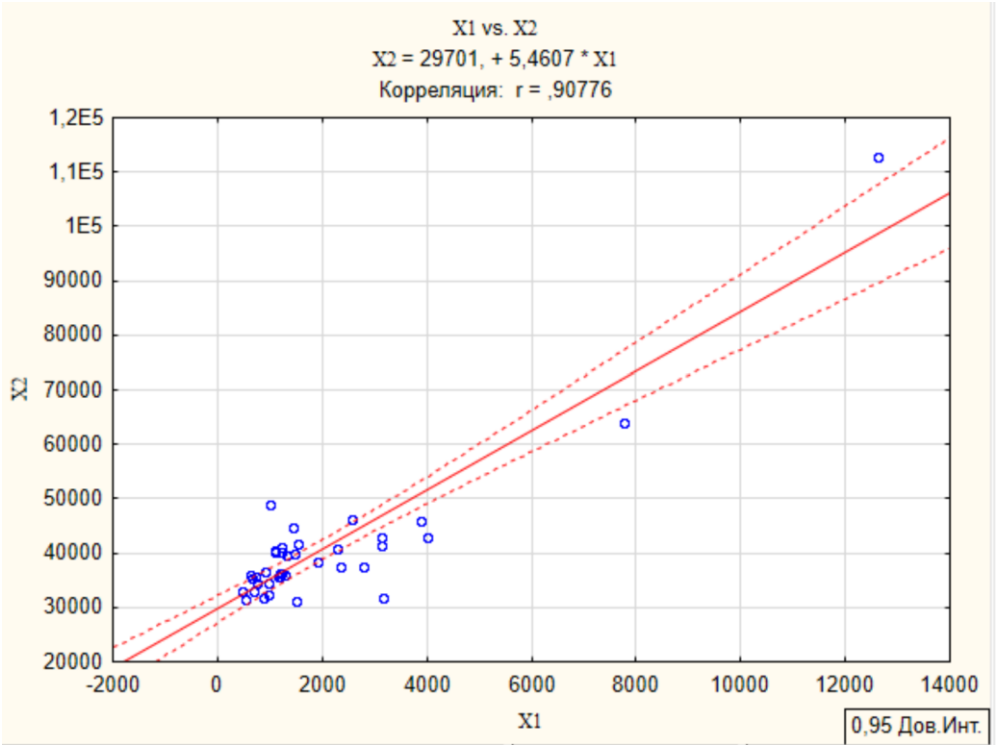


Рис. 6.7.18. 2М диаграмма рассеяния

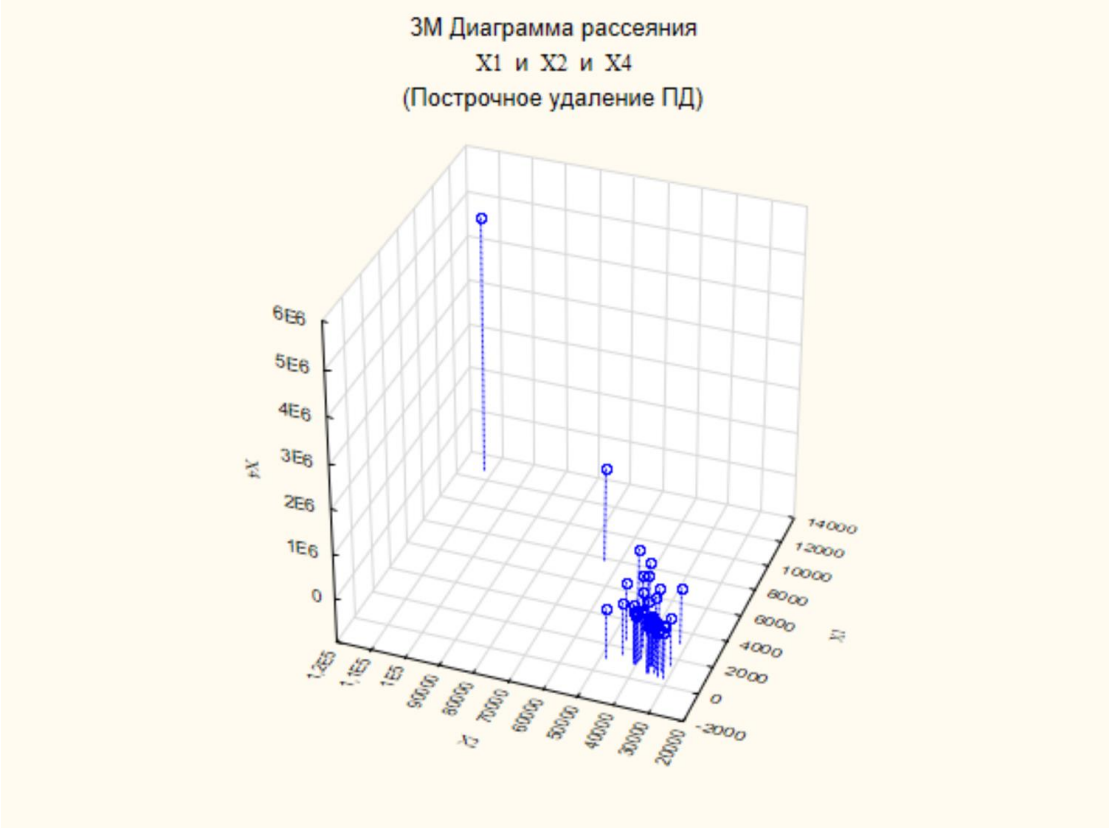


Рис. 6.7.19. 3М диаграмма рассеяния

После ознакомления с описательной статистикой данных перейдем непосредственно к результатам проводимого факторного анализа. После выбора всех необходимых параметров в окне задания метода выделения фактора необходимо нажать ОК. Появится окно Результаты факторного анализа (рис. 6.7.20).

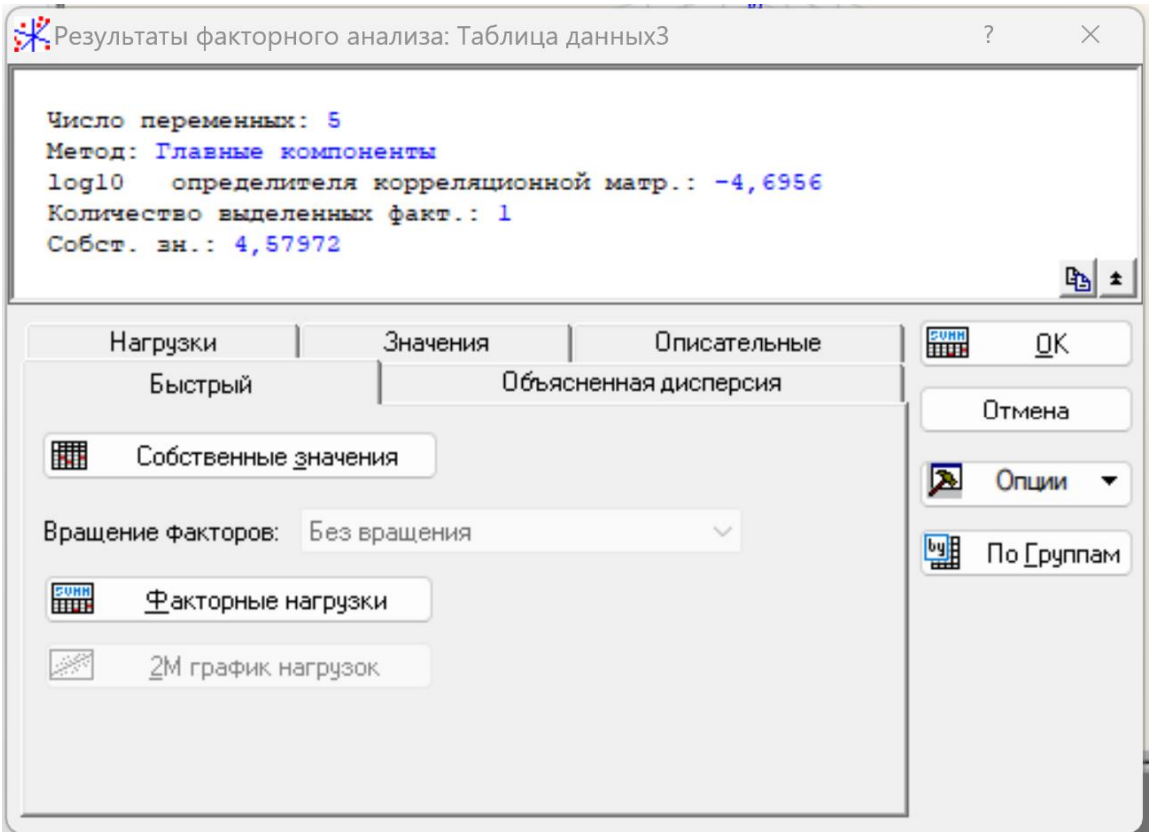


Рис. 6.7.20. Результаты факторного анализа

В нижней части окна находятся подразделы, позволяющие всесторонне просмотреть результаты анализа численно и графически.

График нагрузок, 2D и 3D (двухмерный и трехмерный графики нагрузок) – эти опции построят графики факторных нагрузок в проекции на плоскость любых двух выбранных факторов (рис. 6.7.21) и в проекции в пространство трех выбранных факторов (для чего необходимо наличие как минимум трех выделенных факторов).

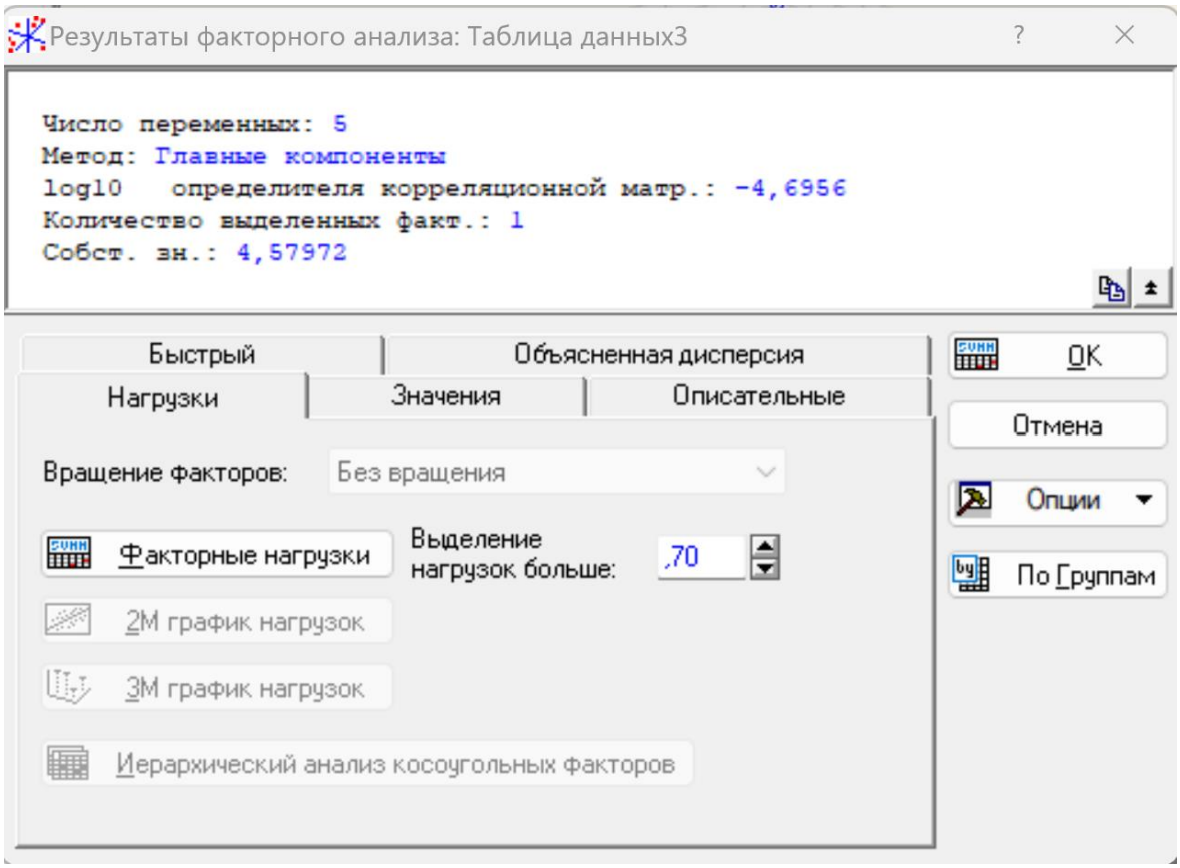


Рис. 6.7.21. Вкладка нагрузки

При нажатии на кнопку Факторные нагрузки открывается таблица с текущими факторными нагрузками (рис. 6.7.22), т.е. вычисленными для данного метода вращения факторов, который указан справа от соответствующей кнопки. В этой таблице факторам соответствуют столбцы, а переменным – строки, и для каждого фактора указывается нагрузка каждой исходной переменной, показывающая относительную величину проекции переменной на факторную координатную ось.

Факторные нагрузки могут интерпретироваться как корреляции между соответствующими переменными и факторами – чем выше нагрузка по модулю, тем больше близость фактора к исходной переменной. То есть они представляют наиболее важную информацию для интерпретации полученных факторов. В сгенерированной таблице для облегчения трактовки будут выделены факторные нагрузки по абсолютной величине больше 0,7.

Перемен.	Фактор.нагрузки (Без вращ.) (Таблица данных3) Выделение: Главные компоненты (Отмечены нагрузки >,700000) **Фактор 1**
X1	-0,955257
X2	-0,966040
X4	-0,989903
X4	-0,887564
X5	-0,983005
Общ.дис.	4,579723
Доля общ	0,915945

Рис. 6.7.22. Факторные нагрузки

В нашем примере выделен один фактор, который описывает порядка 91,6% изменчивости.

Однако не всегда факторный анализ дает такие четкие и легко объяснимые результаты. В таком случае целесообразно прибегнуть к повороту осей, надеясь получить решение, которое можно интерпретировать в предметной области.

Данный вариант доступен, когда выделяется более одного фактора (рис. 6.7.23).

Перемен.	Фактор.нагрузки (Без вращ.) (Таблица данных4) Выделение: Главные компоненты (Отмечены нагрузки >,700000) **Фактор 1**	Фактор 2
Пер1	-0,998417	0,002880
Пер3	-0,998752	0,000383
Пер4	-0,999735	-0,011221
Пер5	-0,997879	-0,003817
Пер6	-0,999427	-0,007085
Пер7	-0,997551	-0,018896
Пер8	-0,998871	-0,015289
Пер9	-0,997442	-0,022747
Пер10	-0,003880	-0,967618
НовПер1	0,082411	-0,963645
Общ.дис.	7,982979	1,866203
Доля общ	0,798298	0,186620

Рис. 6.7.23. Факторные нагрузки

Для выбора варианта вращения необходимо выбрать из выпадающего списка необходимый (рис. 6.7.24).

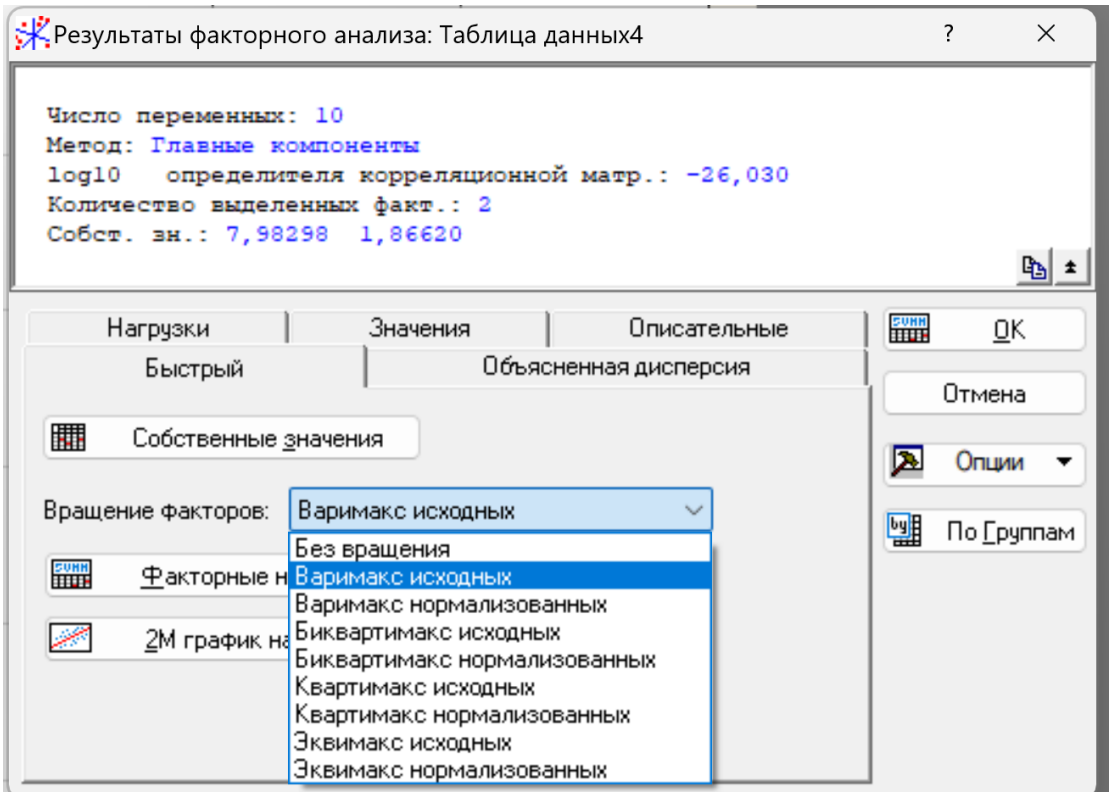


Рис. 6.7.24. Выбор варианта вращения фактора

В раскрывающемся меню можно выбрать различные повороты оси:

– Варимакс – предназначен для максимизации дисперсий квадратов исходных факторных нагрузок по переменным для каждого фактора;

– Биквартимакс – предназначен для одновременной максимизации суммы дисперсий квадратов исходных факторных нагрузок по факторам и максимизации суммы дисперсий квадратов исходных факторных нагрузок по переменным;

– Квартимакс – означает максимизацию дисперсии квадратов факторных нагрузок по факторам;

– Эквимакс – взвешенная смесь вращения по методам варимакс и квартимакс, но в отличие от метода биквартимакс, относительный вес, предназначенный критерию варимакс при вращении, равен числу факторов, деленному на 2.

Возникает вопрос: сколькими же факторами следует ограничиваться на практике? Для этого в программном пакете STATISTICA существует критерий Scree plot (Критерий каменистой осыпи).

Меню выбора данного графика представлено на рисунке 6.7.25.

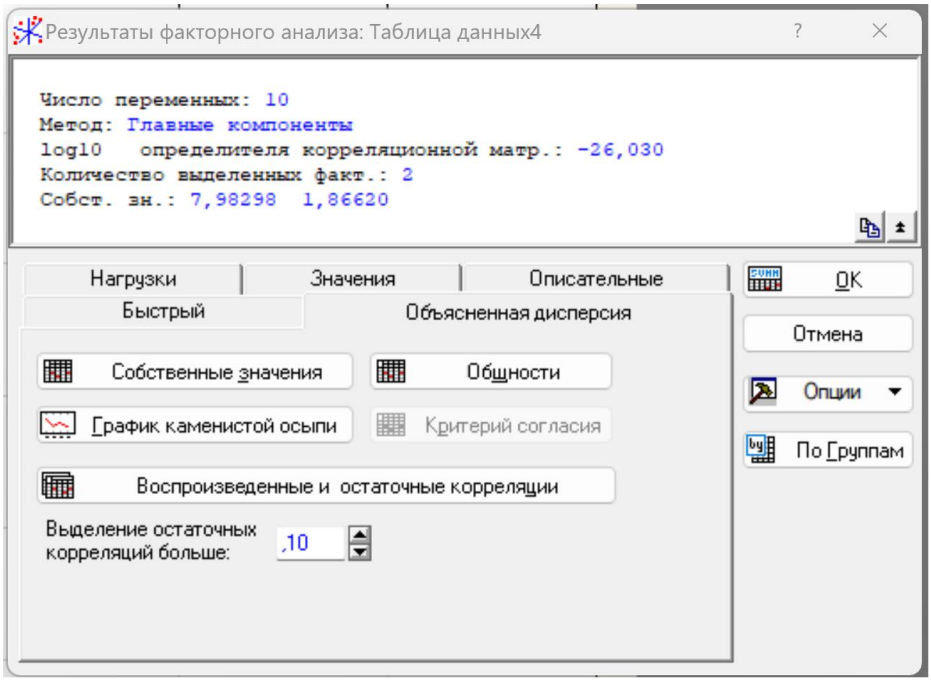


Рис. 6.7.22. Выбор графика каменистой осыпи

График имеет вид – рис. 6.7.23.

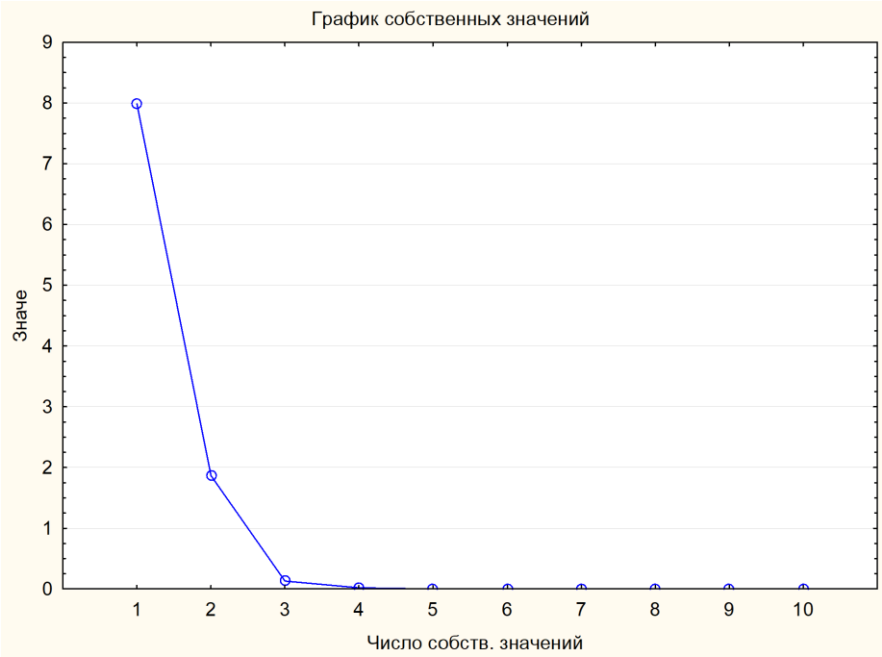


Рис. 6.7.23.График каменистой осыпи

В точках с координатами 1, 2 осыпание замедляется наиболее существенно, следовательно, теоретически можно ограничиваться двумя факторами

Контрольные вопросы по теме

1. Что такое факторный анализ?
2. Какие основные цели проведения факторного анализа?
3. В чем отличие между факторным анализом и компонентным анализом?
4. Какие типы факторного анализа существуют?
5. Какие предпосылки должны быть выполнены для применения факторного анализа?
6. Как проводится обычный факторный анализ?
7. Как проводится метод главных компонент при факторном анализе?
8. Как интерпретировать результаты факторного анализа?
9. Как выбрать оптимальное число факторов при проведении факторного анализа?
10. Как оценить качество модели факторного анализа?
11. Какие методы используются для измерения различий между факторами при проведении факторного анализа?
12. Как можно использовать результаты факторного анализа для уменьшения размерности данных?
13. Как можно использовать результаты факторного анализа для выявления скрытых факторов, влияющих на наблюдаемые переменные?
14. Как можно использовать результаты факторного анализа для создания новых переменных на основе исходных данных?
15. Как можно использовать результаты факторного анализа для выявления внутренних структур в данных?
16. Как можно использовать результаты факторного анализа для определения взаимосвязей между переменными?
17. Как можно использовать результаты факторного анализа для классификации объектов или субъектов?
18. Как можно использовать результаты факторного анализа для прогнозирования значений переменных?

19. Как можно использовать результаты факторного анализа для выявления групп риска или возможностей в данных?

20. Какие программные инструменты чаще всего используются для проведения факторного анализа?

21. Как можно использовать результаты факторного анализа для оценки эффективности различных методов или стратегий?

22. Как можно использовать результаты факторного анализа для выявления тенденций или закономерностей в данных?

23. Как можно использовать результаты факторного анализа для сравнения различных групп или образцов данных?

24. Как можно использовать результаты факторного анализа для определения оптимальных условий или параметров процессов?

25. Что такое общий факторный анализ и как он используется при проведении факторного анализа?

26. Что такое метод главных компонент и как он используется при проведении факторного анализа?

27. Что такое метод максимального правдоподобия и как он используется при проведении факторного анализа?

28. Что такое метод минимальных квадратов и как он используется при проведении факторного анализа?

29. Что такое метод главных осей и как он используется при проведении факторного анализа?

30. Что такое метод кластерного анализа и как он используется при проведении факторного анализа?

31. Как можно использовать результаты факторного анализа для прогнозирования классификации новых объектов?

32. Как можно использовать результаты факторного анализа для выявления скрытых паттернов в данных?

33. Как можно использовать результаты факторного анализа для определения групп схожих объектов или явлений в данных?

34. Как можно использовать результаты факторного анализа для выявления внутренних структур или закономерностей в данных?

35. Как можно использовать результаты факторного анализа для выявления внутренних связей между объектами или явлениями?

36. Как можно использовать результаты факторного анализа для определения групп риска или возможностей в данных?

37. Как можно использовать результаты факторного анализа для определения оптимальных стратегий управления в различных ситуациях?

38. Что такое метод "Principal Axis Factoring" и как он используется при проведении факторного анализа?

39. Что такое метод "Maximum Likelihood Factoring" и как он используется при проведении факторного анализа?

40. Что такое метод "Alpha Factoring" и как он используется при проведении факторного анализа?

41. Что такое метод "Unweighted Least Squares Factoring" и как он используется при проведении факторного анализа?

42. Что такое метод "Weighted Least Squares Factoring" и как он используется при проведении факторного анализа?

43. Что такое метод "Image Factoring" и как он используется пр
и проведении факторного анализа?

44. Что такое метод "Canonical Factor Analysis" и как он используется при проведении факторного анализа?

45. Что такое метод "Common Factor Analysis" и как он используется при проведении факторного анализа?

46. Что такое метод "Generalized Least Squares Factoring" и как он используется при проведении факторного анализа?

47. Что такое метод "Principal Components Analysis" и как он используется при проведении факторного анализа?

48. Что такое метод "Exploratory Factor Analysis" и как он используется при проведении факторного анализа?

49. Что такое метод "Confirmatory Factor Analysis" и как он используется при проведении факторного анализа?

50. Какие методы используются для проверки предпосылок и условий применения факторного анализа?

7. ОСНОВЫ НЕЙРОСЕТЕВОГО ПРОГНОЗИРОВАНИЯ В STATISTICA

7.1. Основы теории нейронных сетей

Нейрон представляет собой единицу обработки информации в нейронной сети. На рисунке 7.1.1 представлена модель нейрона, лежащего в основе искусственных сетей. В этой модели можно выделить три основных элемента:

1. **Набор синапсов (synapse) или связей (connecting link)**, каждый из которых характеризуется своим весом (weight) или силой (strength). В частности, сигнал x_j на входе синапса j, связанного с нейроном k, умножается на вес w_{kj}. Первый индекс синаптического веса w_{kj} относится к рассматриваемому нейрону, а второй – ко входному окончанию синапса, с которым связан данный вес. В отличие от синапсов мозга синаптический вес искусственного нейрона может иметь как положительные, так и отрицательные значения.

2. **Сумматор (adder)** складывает входные сигналы, взвешенные относительно соответствующих синапсов нейрона. Эту операцию называют линейной комбинацией.

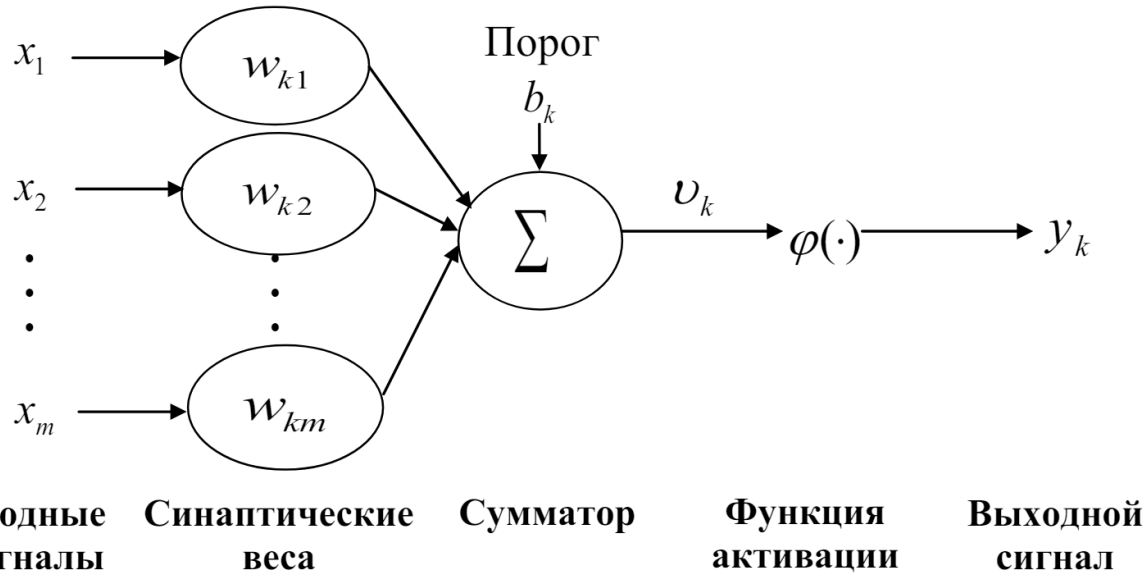


Рис. 7.1.1. Нелинейная модель нейрона

Функция активации (activation function) ограничивает амплитуду выходного сигнала нейрона. Эта функция также называется функцией сжатия (squashing function). Обычно нормализованный диапазон амплитуд выхода нейрона лежит в интервале [0,1] или [1,1].

В модель нейрона, представленную на рисунке 7.1.1, включен пороговый элемент (bias), который обозначен b_k. Эта величина отражает увеличение или уменьшение входного сигнала, подаваемого на функцию активации.

В математическом представлении функционирование нейрона k осуществляется следующим образом (7.1.1) и (7.1.2):

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \qquad (7.1.1)$$

$$y_k = \varphi(u_k + b_k) \qquad (7.1.2)$$

где x_j – входные сигналы;

w_{kj} – синаптические веса нейрона k;

u_k – линейная комбинация входных воздействий (linear combiner output);

b_k – порог;

φ (…) – функция активации (activation function);

y_k – выходной сигнал нейрона.

Использование порога b_k обеспечивает эффект аффинного преобразования (affine transformation) выхода линейного сумматора u_k. В модели, представленной на рисунке 7.1.1, постсинаптический потенциал вычисляется следующим образом (7.1.2):

$$v_k = u_k + b_k \qquad (7.1.2)$$

В зависимости от того, какое значение принимает порог b_k, положительное или отрицательное, индуцированное локальное поле (induced local field) или потенциал активации (activation potential) v_k нейрона k изменяется так, как представлено на рисунке 7.1.2.

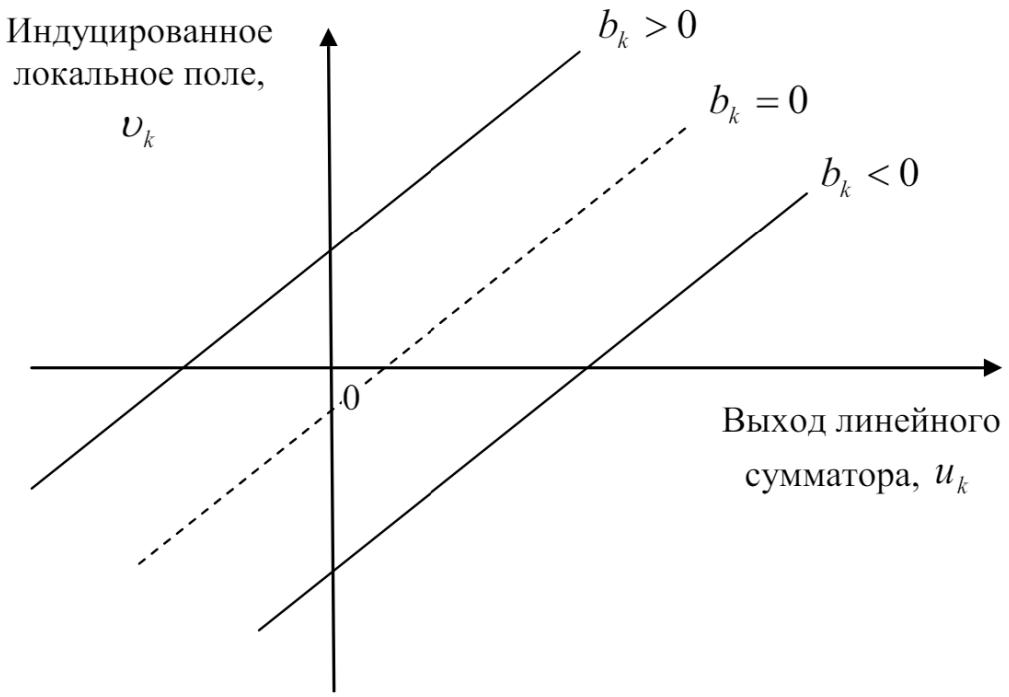


Рис. 7.1.2. Аффинное преобразование, вызванное наличием порога

Поскольку $b_k \neq 0$, то график v_kне проходит через начало координат как график u_k. Порогb_k является внешним параметром искусственного нейрона k. Принимая во внимание выражение (7.1.3), формулы (7.1.1), (7.1.2) можно преобразовать к следующему виду (7.1.4) и (7.1.5).

$$u_k = \sum_{j=0}^{m} w_{kj} x_j \quad (7.1.4)$$

$$u_k = \varphi(v_k) \quad (7.1.5)$$

Таким образом в выражение (7.1.4) добавился новый синапс. Его выходной сигнал равен: $x_0 = +1$, а его вес $w_{k0} = b_k$.

7.2. Архитектура нейросетей

Различают следующие виды сетей по данному классификационному признаку:

1. Однослойные сети прямого распространения. В простейшем случае в сети существует входной слой (input layer) узлов источника, информация от которого передается на выходной слой (output layer) нейронов (вычислительные узлы). Такая сеть называется сетью прямого распространения, которая представлена на рисунке 7.2.1. Нейронная сеть называется **однослойной** (single-layer network), если единственным слоем является слой вычислительных элементов (нейронов).

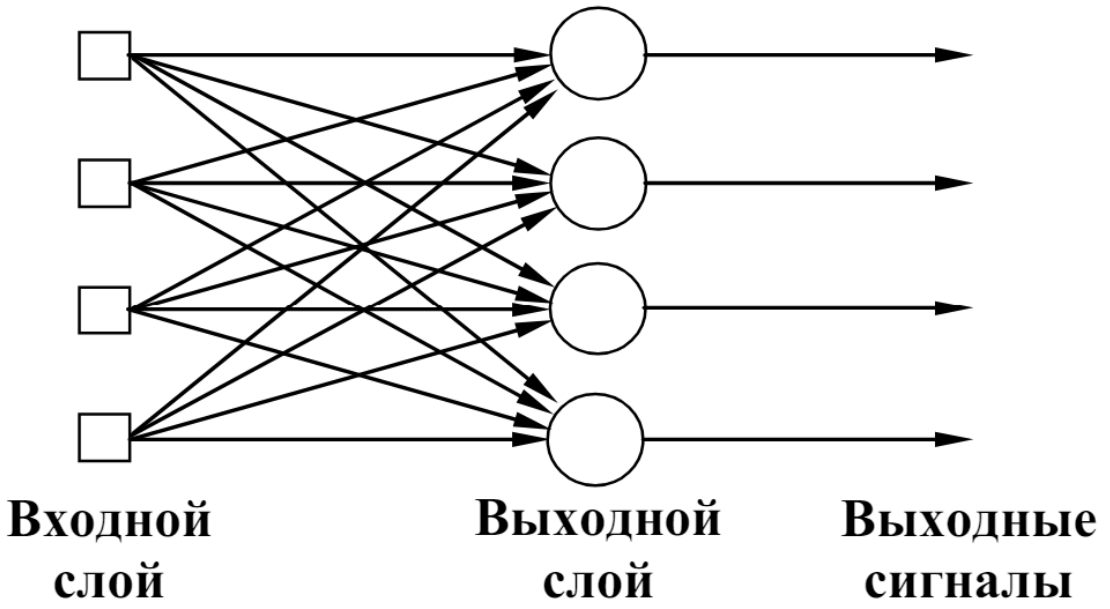


Рис. 7.2.1. Однослойные сети прямого распространения

2. Многослойные сети прямого распространения. Эти сети характеризуются наличием одного или нескольких скрытых слоев (hidden layer), т.е. узлов, которые называются скрытыми нейронами (hidden neuron) или скрытыми элементами (hidden unit). Функция скрытых нейронов заключается в посредничестве между внешним входным сигналом и выходом нейронной сети. Узлы источника входного слоя сети образуют соответствующий входной вектор, который является входным сигналом, поступающим на нейроны 2 слоя (т.е. первого скрытого слоя). Выходные сигналы второго слоя применяются в качестве входов для третьего слоя и так далее. Таким образом, исходящие

сигналы нейронов каждого из слоев сети являются входящими сигналами для следующих за ними слоев.

Сеть, которая представлена на рисунке 7.2.2, называется сетью 4-4-2, потому что имеет 4 входа, 4 скрытых нейрона, 2 выходных нейрона. В общем случае, сеть прямого распространения с m входами, h_1 нейронами первого скрытого слоя, h_2 нейронами второго скрытого слоя и q нейронами выходного слоя называется сетью m - h_1 - h_2 – q.

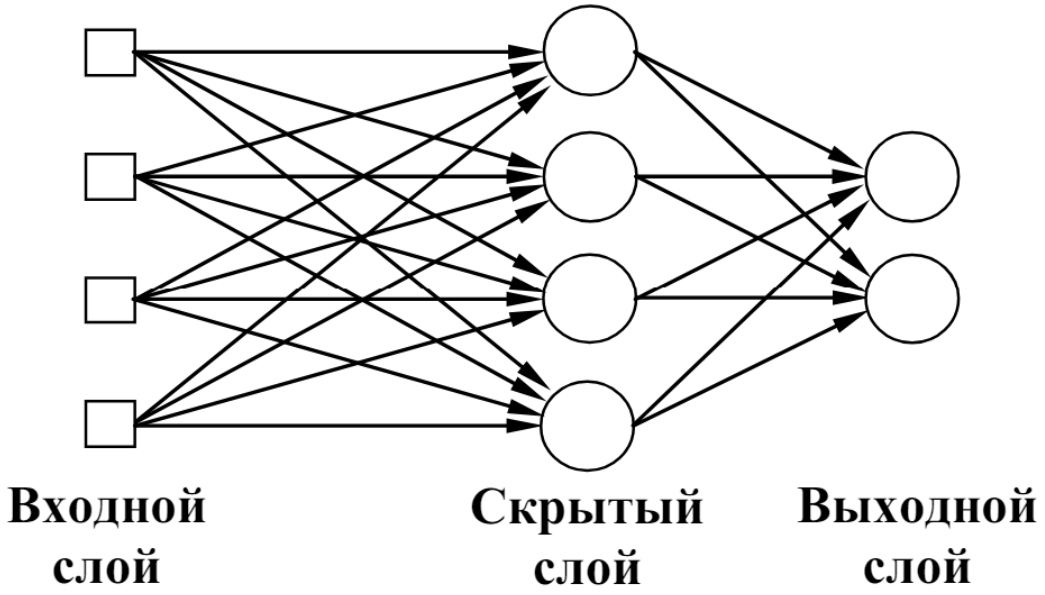


Рис. 7.2.2. Полносвязная сеть прямого распространения

Нейронная сеть, представленная на рисунке 7.2.2, считается полносвязной (fully connected), поскольку все узлы каждого конкретного слоя соединены со всеми узлами смежных слоев. Если некоторые из синаптических связей в сети отсутствуют, то она называется неполносвязной (partially connected).

3. Рекуррентные сети. Рекуррентная нейронная сеть (recurrent network) отличается от сети прямого распространения наличием, по крайней мере, одной обратной связи (feedback loop). Например, рекуррентная сеть может состоять из единственного слоя нейронов, каждый из которых направляет свой выходной сигнал на входы всех остальных нейронов слоя. Архитектура такой сети представлена на рисунках 7.2.3 – 7.2.4.

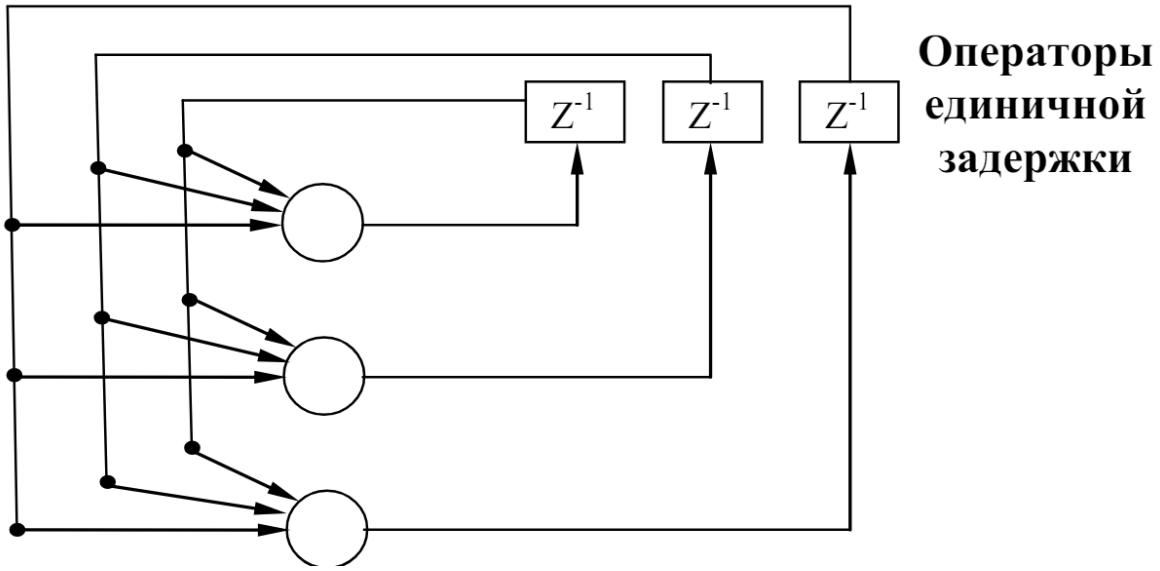


Рис. 7.2.3. Рекуррентная сеть без скрытых нейронов и обратных связей нейронов с самими собой

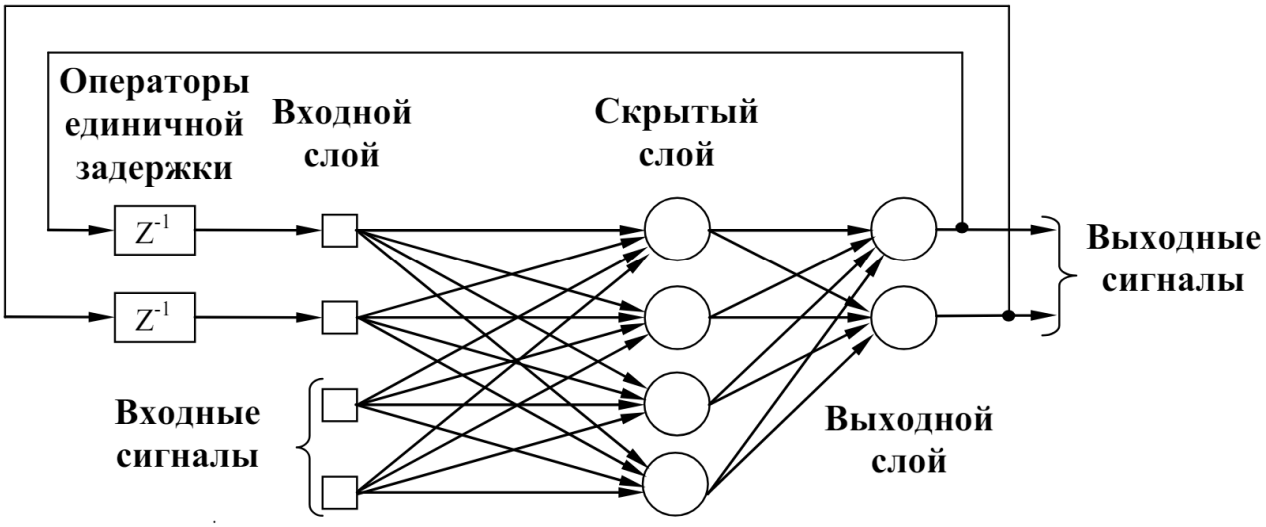


Рис. 7.2.4. Рекуррентная сеть со скрытыми нейронами

Наличие обратных связей в сетях оказывает непосредственное влияние на способность таких сетей к обучению и их производительность. Более того, обратная связь подразумевает исполнение элементов единичной задержки (unit-delay element) (z^{-1}), что приводит к нелинейному динамическому поведению, если в сети содержаться нелинейные нейроны.

7.3. Обучение нейросетей

Можно выделить алгоритмы обучения с учителем и без учителя.

Обучение с учителем

Обучение с учителем предполагает, что для каждого входного вектора существует целевой вектор, представляющий собой требуемый выход. Вместе они называются обучающей парой. Обычно сеть обучается на некотором числе таких обучающих пар. Предъявляется выходной вектор, вычисляется выход сети и сравнивается с соответствующим целевым вектором. Разность (ошибка) с помощью обратной связи подается в сеть, и веса изменяются в соответствии с алгоритмом, стремящимся минимизировать ошибку. Векторы обучающего множества предъявляются последовательно, ошибки вычисляются и веса подстраиваются для каждого вектора до тех пор, пока ошибка по всему обучающему массиву не достигнет приемлемого низкого уровня.

Обучение без учителя

Обучение без учителя осуществляется без вмешательства внешнего учителя, контролирующего процесс обучения. Существует лишь независимый от задачи уровень качества представления, которому должна научиться нейронная сеть, и свободные параметры сети оптимизируются по отношению к этому уровню.

Обучение без учителя является намного более правдоподобной моделью обучения для биологической системы. Развитая Кохоненом Т. и многими другими, она не нуждается в целевом векторе для выходов и, следовательно, не требует сравнения с предопределенными идеальными ответами. Обучающее множество состоит лишь из входных векторов. Обучающий алгоритм подстраивает веса сети так, чтобы получались согласованные выходные векторы, т.е. чтобы предъявление достаточно близких входных векторов давало одинаковые выходы. Процесс обучения, следовательно, выделяет статистические свойства обучающего множества и группирует сходные векторы в классы. Предъявление на вход вектора из данного класса даст определенный выходной вектор, но до обучения невозможно предсказать, какой выход будет производиться данным классом входных векторов. Следовательно, выходы подобной сети должны трансформироваться в некоторую понятную форму, обусловленную процессом обучения. На практике достаточно легко идентифицировать установленную сетью связь между входом и выходом.

7.4. Построение прогноза развития предприятия на среднесрочный период

Для построения среднесрочного прогноза используется программный пакет STATISTICA Automated Neural Networks. Основные статьи операционных расходов представлены в таблице 7.3.1.

Таблица 7.3.1

Основные статьи операционных расходов (в сопоставимых ценах)

№ п/п	Операционные затраты	Ед. изм.	2013	2014	2015	2016
1	Зарплата с начислениями	тыс. руб.	253 670	260 811	255 603	245 696
2	Амортизация	тыс. руб.	116 280	99 611	105 261	76 386
3	Тек. рем. оборудования, зданий	тыс. руб.	48 408	42 639	47 451	42 586
4	Коммунальные платежи	тыс. руб.	11 567	10 842	9 390	9 131
5	ГСМ	тыс. руб.	6 437	7 030	7 768	6 283
6	Связь	тыс. руб.	3 176	2 949	3 642	2 971
7	Страхование, лицензирование	тыс. руб.	17 889	15 669	13 990	7 770
8	Лабораторные затраты	тыс. руб.	659	1 019	806	620
9	Обучение, подписка	тыс. руб.	1 982	2 104	2 579	3 455
10	Командировочные расходы	тыс. руб.	1 410	1 859	1 852	2 445
11	Услуги банка	тыс. руб.	1 656	1 326	933	867
12	Реклама, расходы маркетинга	тыс. руб.	6 086	7 628	7 777	5 569
13	Консульт. услуги	тыс. руб.	35 879	34 553	31 333	28 355
14	Спец. одежд, хоз. расходы	тыс. руб.	4 966	4 542	4 610	4 838
15	Канцтовары	тыс. руб.	793	1 453	1 412	1 979

16	Расходы охране	тыс. руб.	13 337	12 758	11 614	11 079
17	Мед. осмотр, медикаменты	тыс. руб.	1 998	2 865	2 021	1 573
18	Представительские расходы	тыс. руб.	1 103	982	988	547
19	ТБ, охрана труда, пож. служба	тыс. руб.	4 082	2 775	3 447	2 873
20	Налог на землю	тыс. руб.	905	87	505	610
21	Налог на транспорт	тыс. руб.	467	327	213	282

На основе структуры данных строятся нейронные сети для наиболее важных показателей:

1) Зарплата с начислениями.
2) Амортизация.
3) Текущий ремонт оборудования, зданий.
4) Коммунальные платежи.
5) Обучение, подписка.
6) Реклама, расходы маркетинга.
7) Налог на землю.
8) Налог на транспорт.
9) Чистая прибыль.

1) Зарплата с начислениями. На данном показателе разбирается весь процесс построения и обучения нейронной сети, в дальнейшем некоторые шаги опускаются для удобства восприятия информации.

На первом шаге загружаются исходные данные в программу для преобразования в удобный формат.

На втором шаге открывается стартовая панель для работы с нейронными сетями (рис. 7.3.2) и выбираются временные ряды (регрессия).

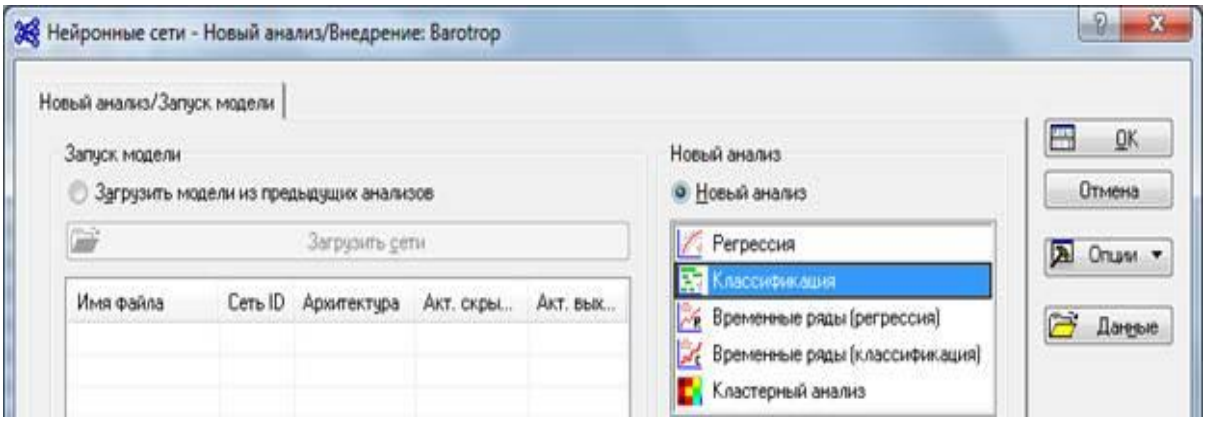


Рис. 7.3.2. Стартовая панель

На третьем шаге открывается окно выбора данных (рис. 7.3.3), выбираются переменные и стратегия автоматизированная нейронная сеть (АНС).

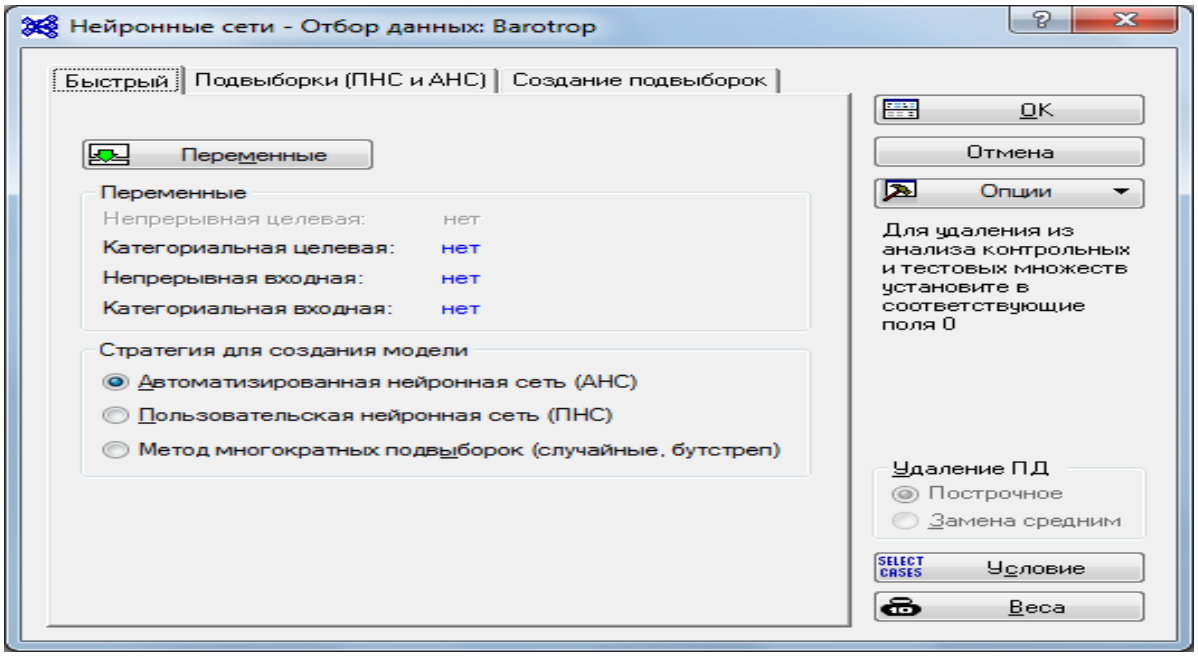


Рис. 7.3.3. Отбор данных

На четвертом шаге в окне “автоматизированные сети” (рис. 7.3.4) выбирается тип сети, количество скрытых нейронов и сетей для обучения.

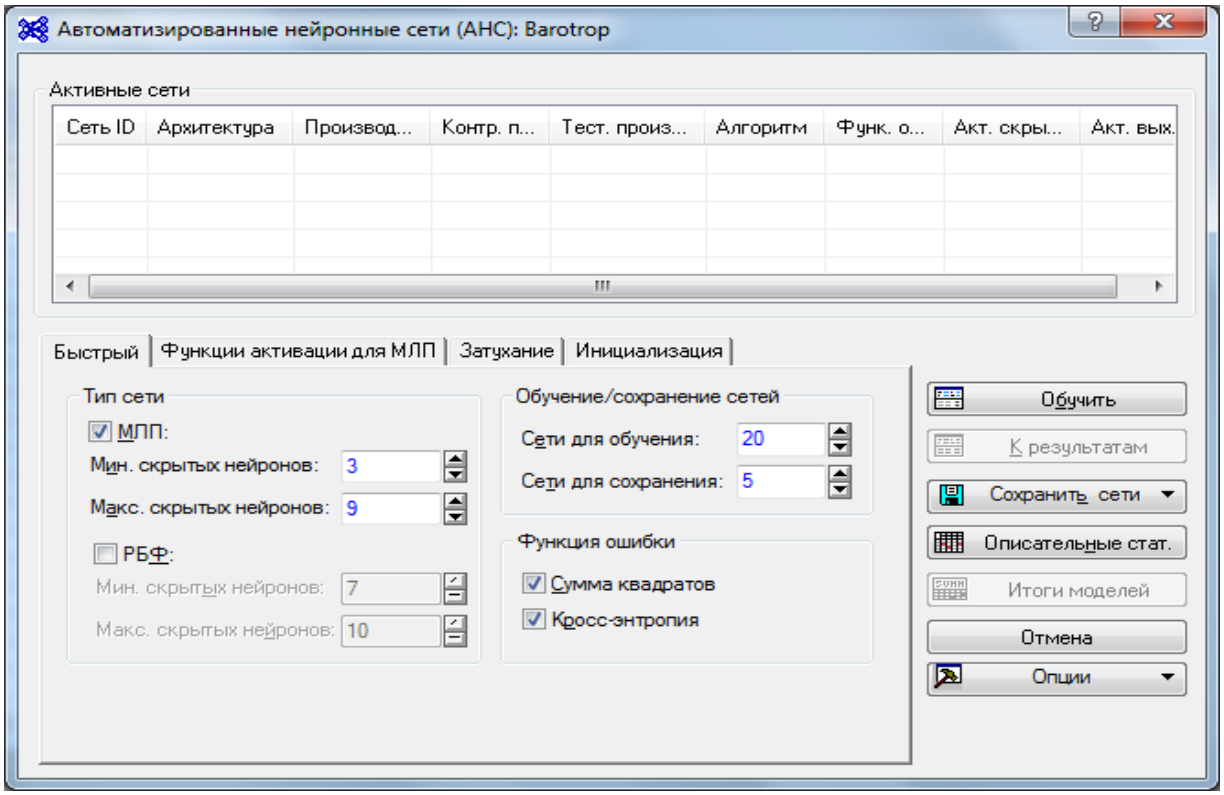


Рис. 7.3.4. Автоматизированные нейронные сети

На пятом шаге анализируются полученные сети и строится их графическое отображение (рис. 7.3.5 и рис. 7.3.6).

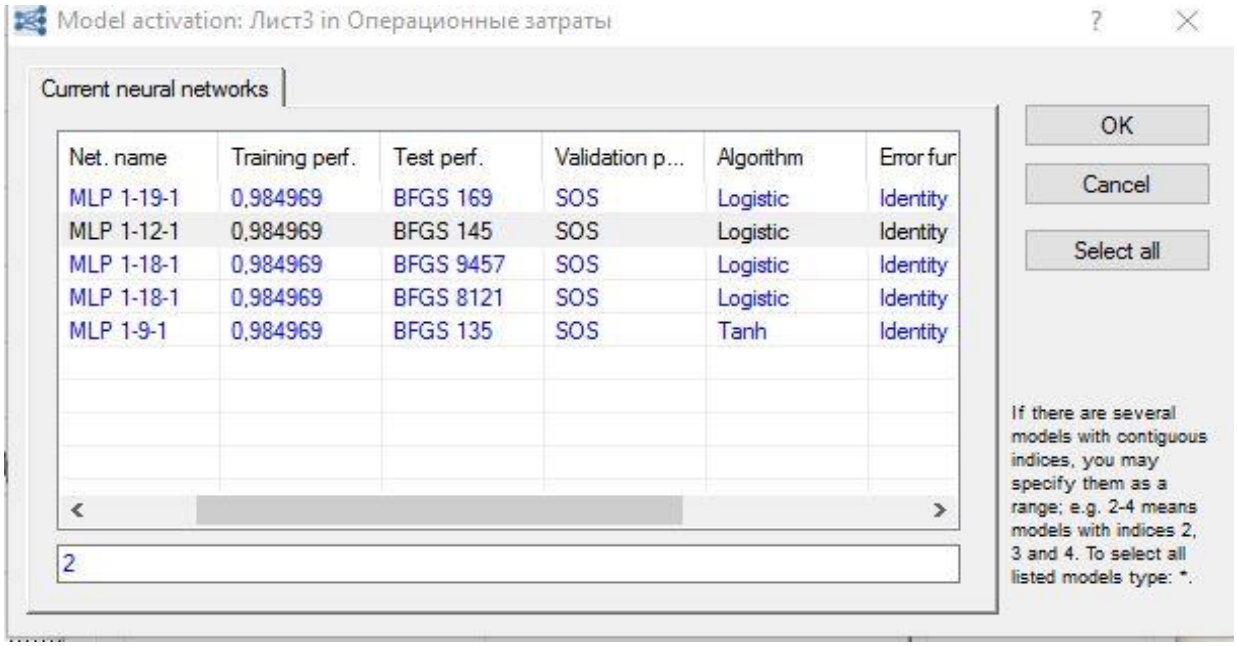


Рис. 7.3.5. Полученные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

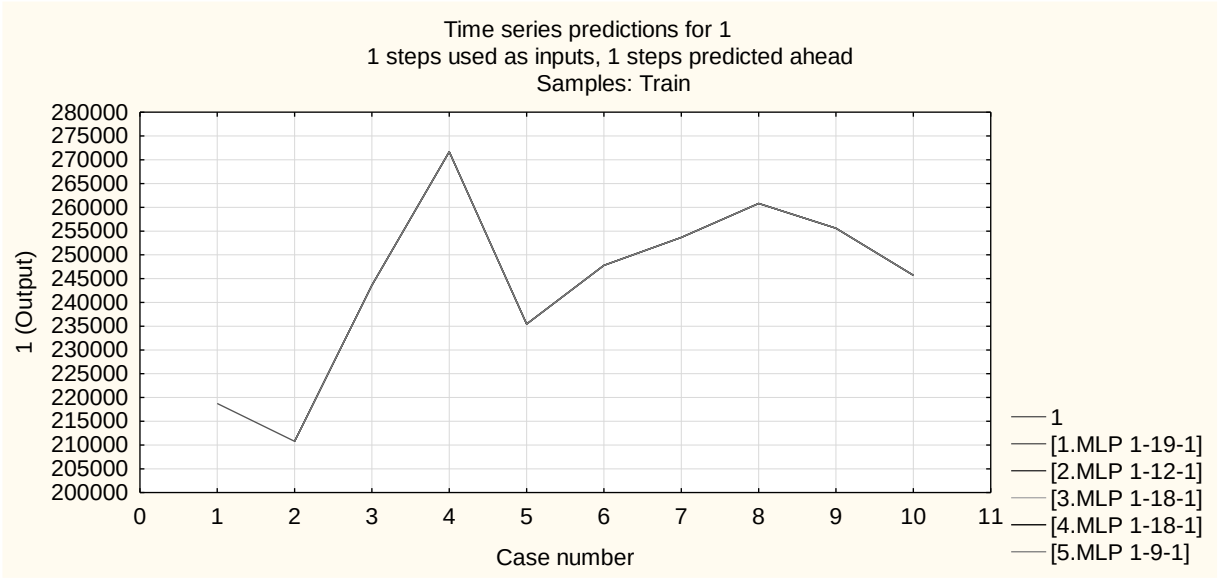


Рис. 7.3.6. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что построенные нейронные сети хорошо улавливают модель поведения показателя.

На шестом шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.7 и рис. 7.3.8).

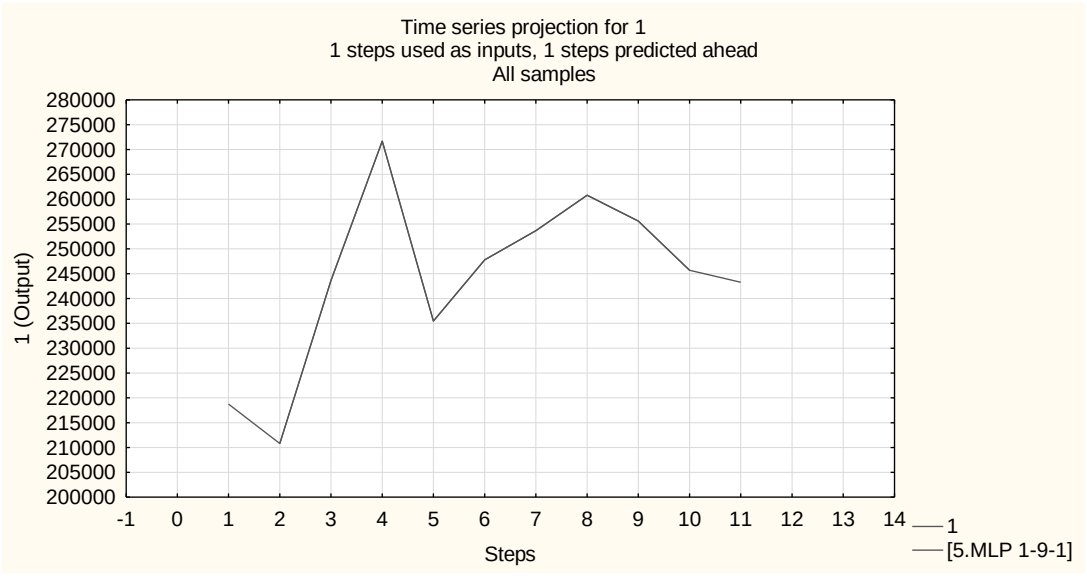


Рис. 7.3.7. Наиболее точная и адекватная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, девять срытых нейронов и один выходной нейрон.

	Time series projection for 1 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
Case name	1 Target	1(Output) MLP 1-9-1
2007	218741,0	
2008	210797,0	210797,0
2009	243611,0	243611,0
2010	271702,0	271702,0
2011	235458,0	235458,0
2012	247795,0	247795,0
2013	253670,0	253670,0
2014	260811,0	260811,0
2015	255603,0	255603,0
2016	245696,0	245696,0
2017		243277,3

Рис. 7.3.8. Прогноз на 2017 г., тыс. руб.

На седьмом шаге анализируются полученные данные (Рис.). В процессе моделирования получается следующий результат в 2017 г. - затраты на оплату труда и начисления уменьшатся на 0,98% или на 2 418,7 тыс. руб. по сравнению с 2016 г.

2) Амортизация.

На этом шаге анализируются полученные сети и строится их графическое отображение (рис. 7.3.9, рис. 7.3.10).

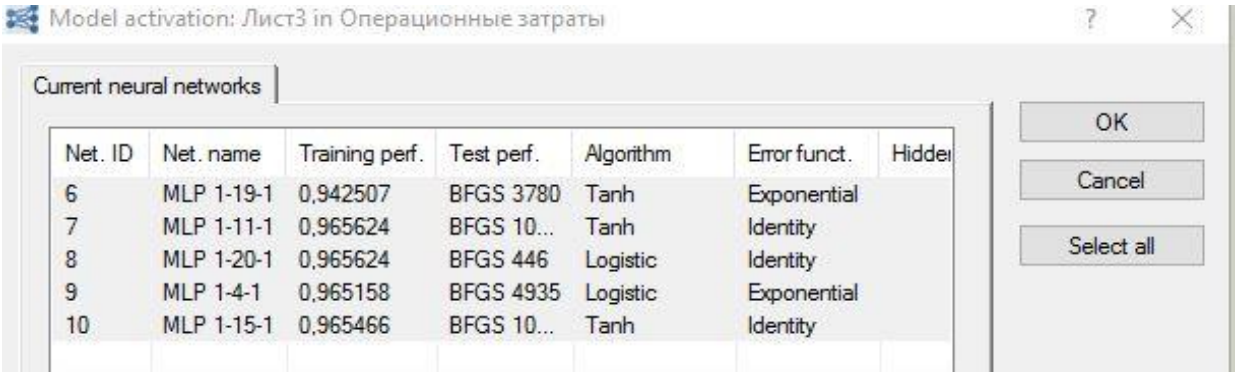


Рис. 7.3.9. Полученные нейронные сети

По данным рисунка отмечается, что все построенные сети кроме 6 обладают хорошей производительностью и высокой точностью.

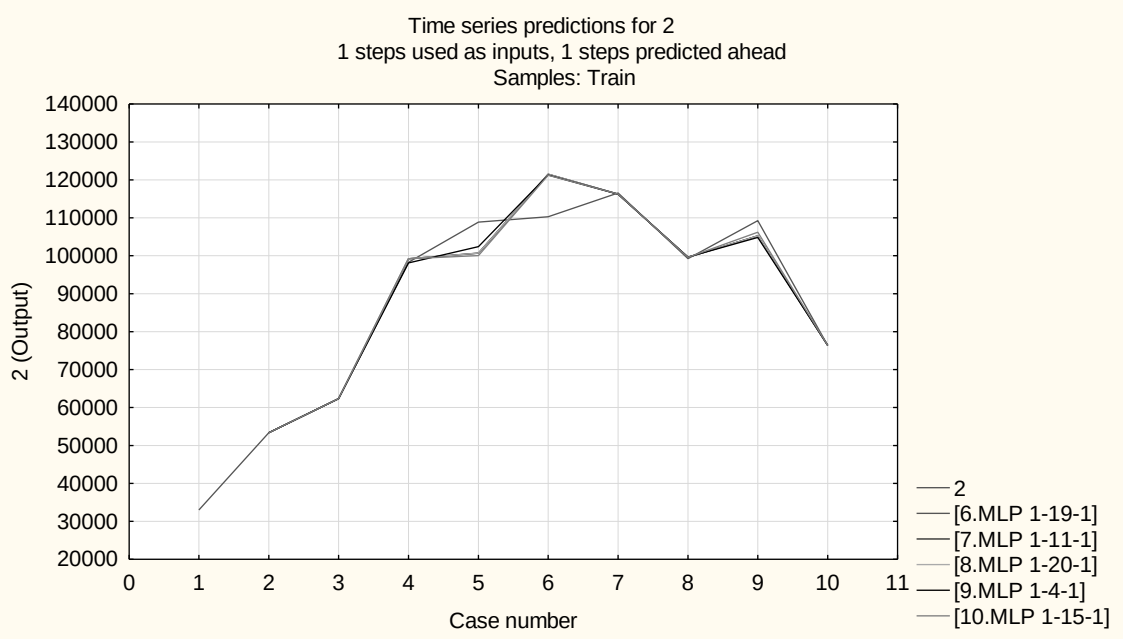


Рис. 7.3.10. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети кроме 6 хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.11 и рис. 7.3.12).

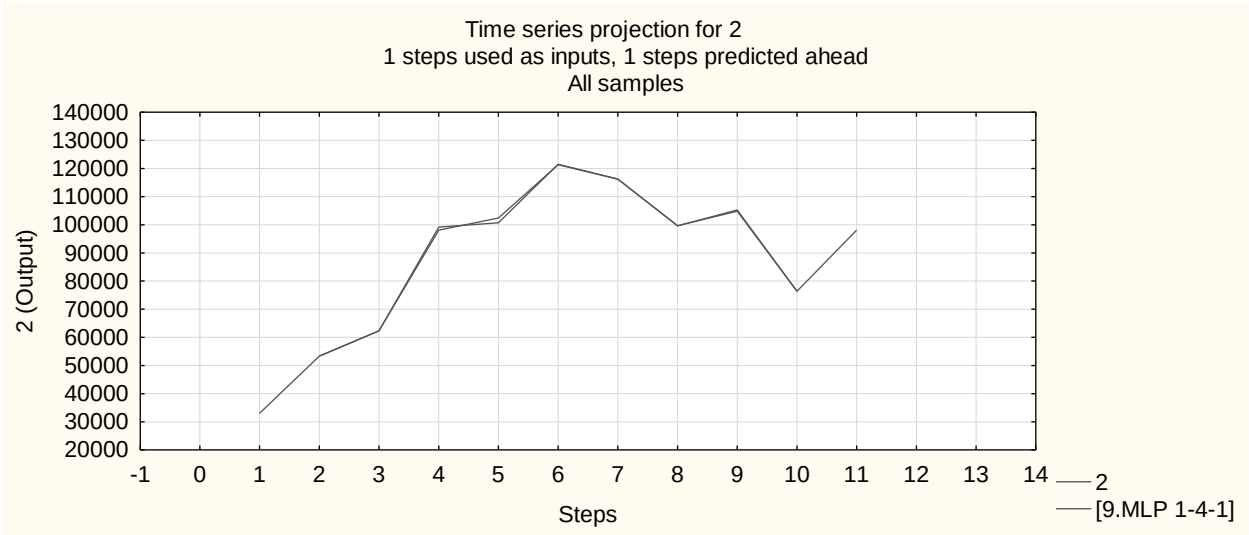


Рис. 7.3.11. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, четыре срытых нейронов и один выходной нейрон.

	Time series projection for 2 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
Case name	2 Target	2(Output) MLP 1-4-1
2007	32998,0	
2008	53375,0	53375,3
2009	62328,0	62331,5
2010	99192,0	98107,9
2011	100712,0	102409,5
2012	121499,0	121307,4
2013	116280,0	116247,4
2014	99611,0	99652,1
2015	105261,0	104839,3
2016	76386,0	76386,0
2017		98107,9

Рис. 7.3.12. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат в 2017 г. - расходы на амортизацию увеличатся на 28,44% и составят 98107 тыс. руб. по сравнению с 2016 г. в связи с покупкой нового оборудования.

3) Текущий ремонт оборудования, зданий.

На этом шаге анализируются полученные сети и строится их графическое отображение (рис. 7.3.13, рис. 7.3.14).

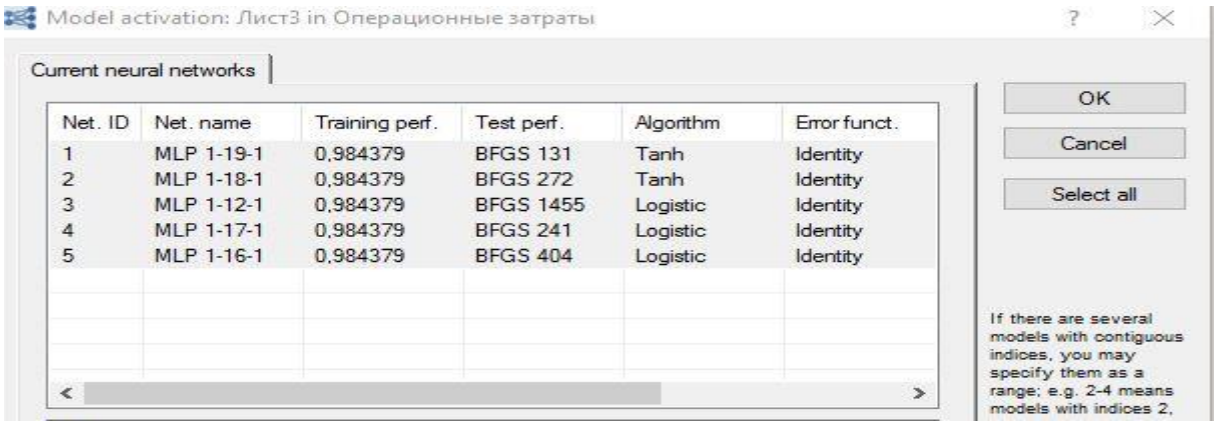


Рис. 7.3.13. Построенные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

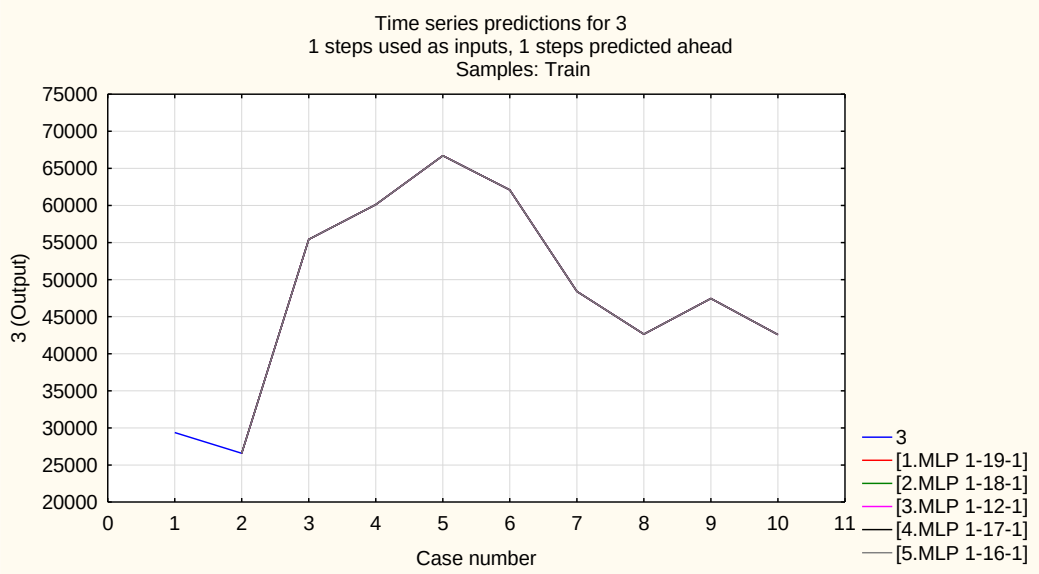


Рис. 7.3.14. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.15 и рис. 7.3.16).

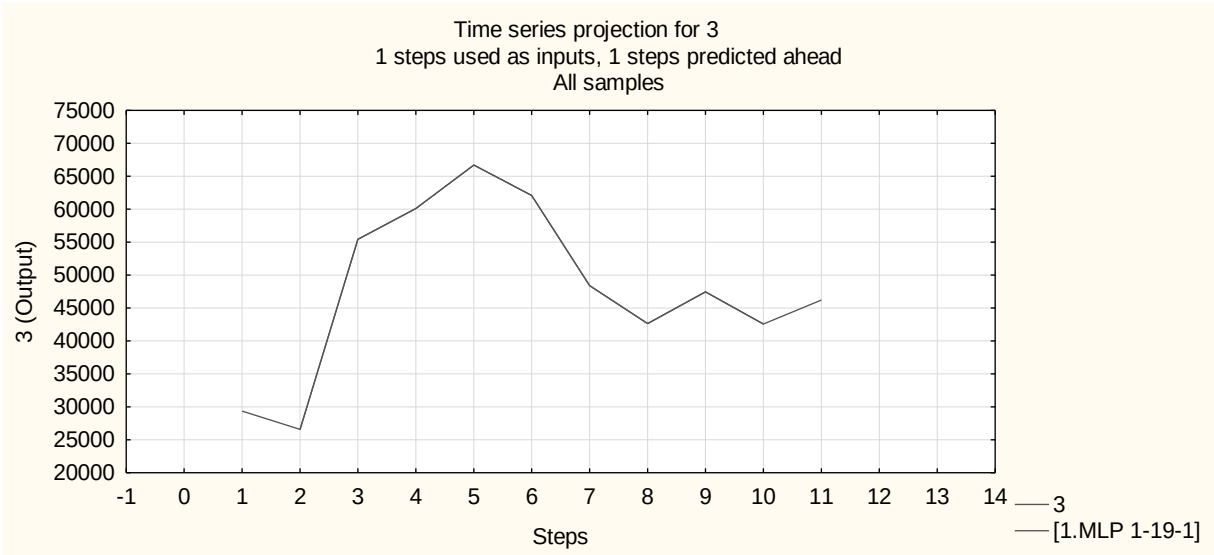


Рис. 7.3.15. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, девятнадцать срытых нейронов и один выходной нейрон.

	Time series projection for 3 (Лист3 in Операционные затрать 1 steps used as inputs, 1 steps predicted ahead All samples	
Case name	3 Target	3(Output) MLP 1-19-1
2007	29372,00	
2008	26582,00	26582,00
2009	55428,00	55428,00
2010	60116,00	60116,00
2011	66703,00	66703,00
2012	62092,00	62092,00
2013	48408,00	48408,00
2014	42639,00	42639,00
2015	47451,00	47451,00
2016	42586,00	42586,00
2017		46228,46

Рис. 7.3.16. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат - в 2017 г. расходы на текущий ремонт увеличатся на 8,55% и составят 46228,46 тыс. руб. по сравнению с 2016 г. в связи с запланированным ремонтом зданий.

4) Коммунальные платежи.

На этом шаге анализируются полученные сети и строится их графическое отображение (рис. 7.3.17, рис. 7.3.18).

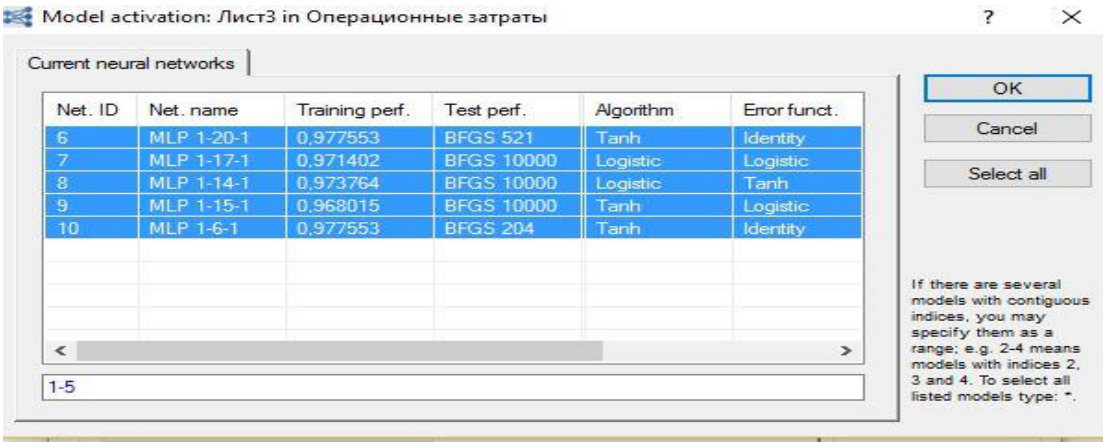


Рис. 7.3.17. Построенные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

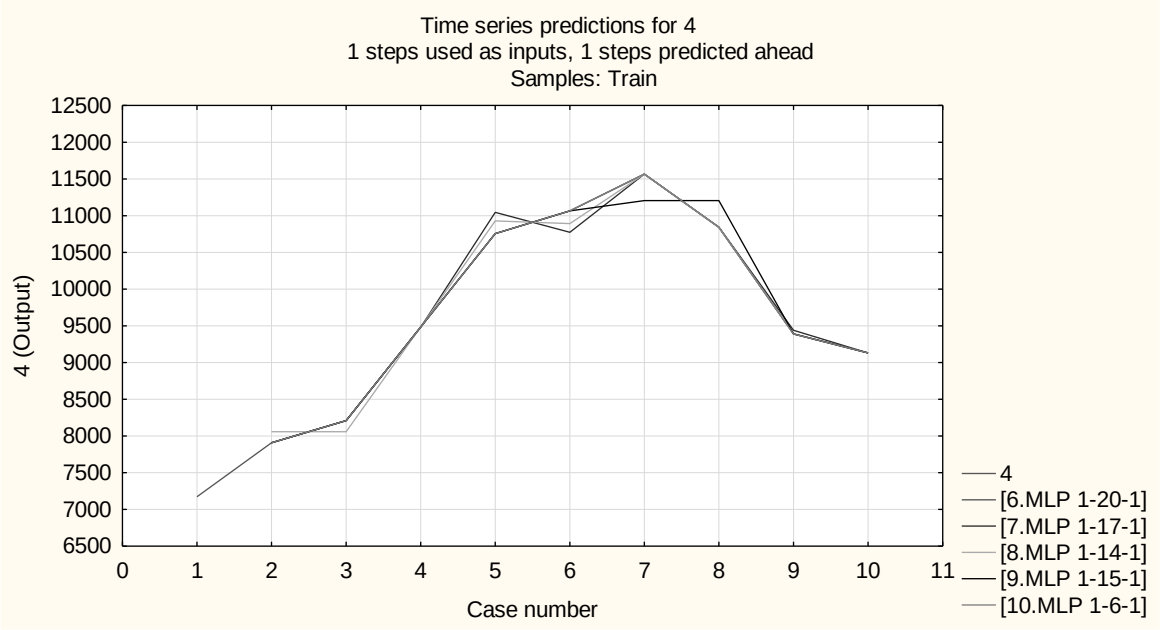


Рис. 7.3.18. Графическое отображение нейронные сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 г. (рис. 7.3.19, рис. 7.3.20).

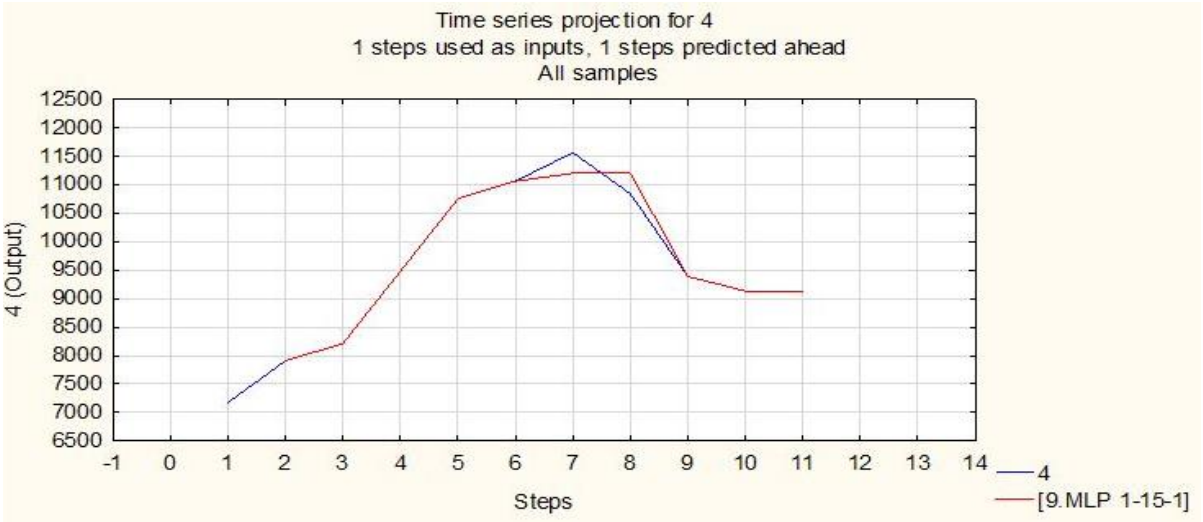


Рис. 7.3.19. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, пятнадцать срытых нейронов и один выходной нейрон.

Case name	Time series projection for 4 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples 4 Target	4(Output) MLP 1-15-1
2007	7173,00	
2008	7908,00	7908,00
2009	8210,00	8210,03
2010	9483,00	9482,94
2011	10756,00	10755,97
2012	11065,00	11064,91
2013	11567,00	11204,44
2014	10842,00	11204,73
2015	9390,00	9390,10
2016	9131,00	9130,99
2017		9130,96

Рис. 7.3.20. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат - в 2017 г. расходы на коммунальные платежи вырастут незначительно и составят 9130,96 тыс. руб. по сравнению с 2016 г.

5) Обучение, подписка.

На этом шаге анализируются полученные сети и строятся их графическое отображение (рис. 7.3.21, рис. 7.3.22).

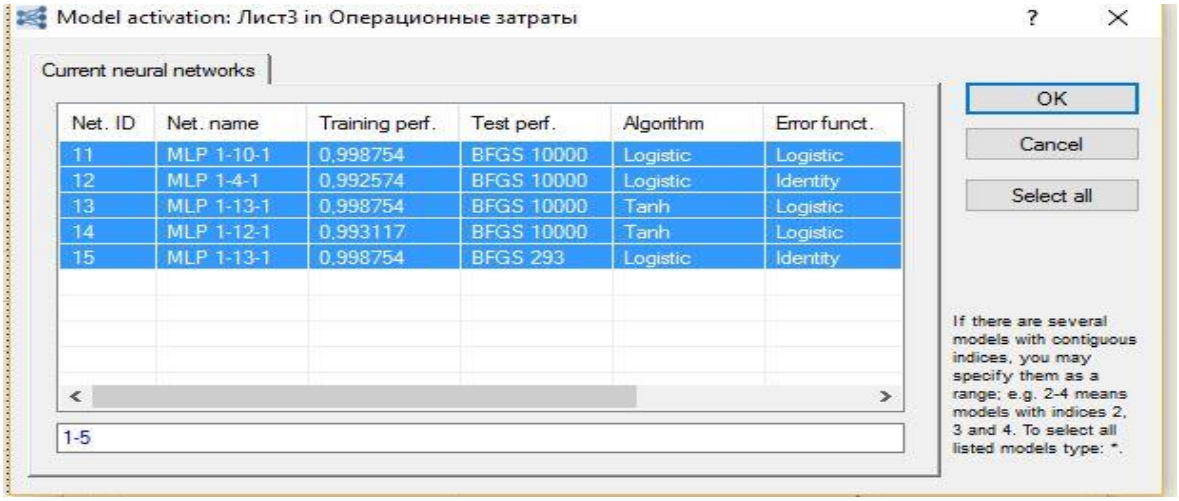


Рис. 7.3.21. Построенные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

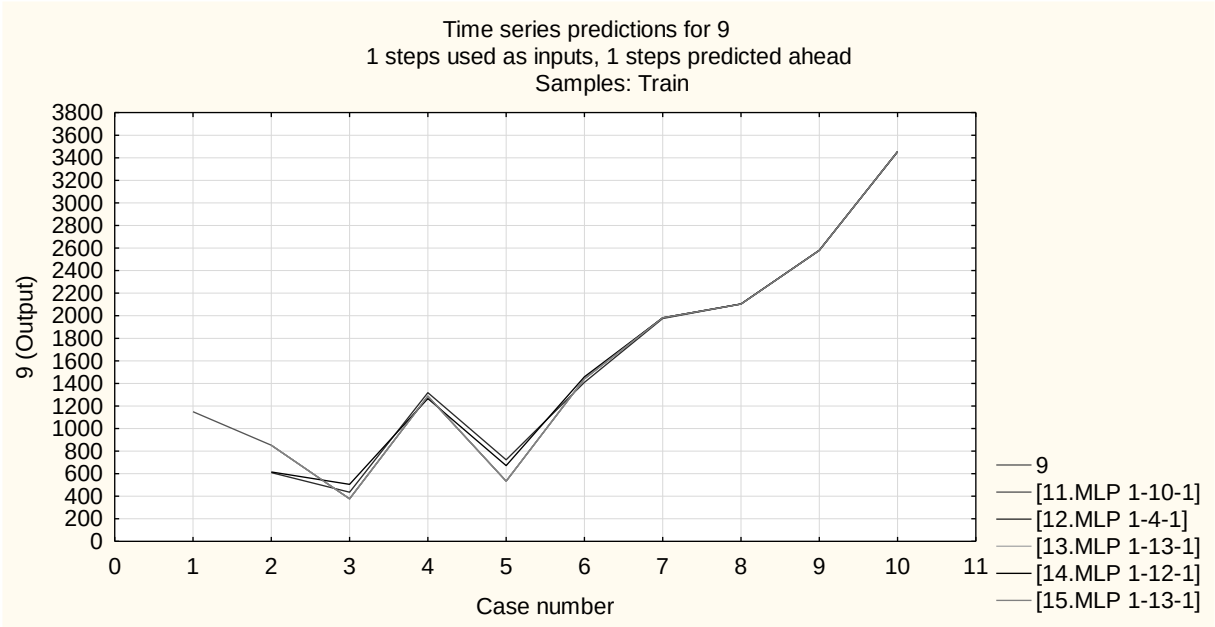


Рис. 7.3.22. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.23, рис. 7.3.24).

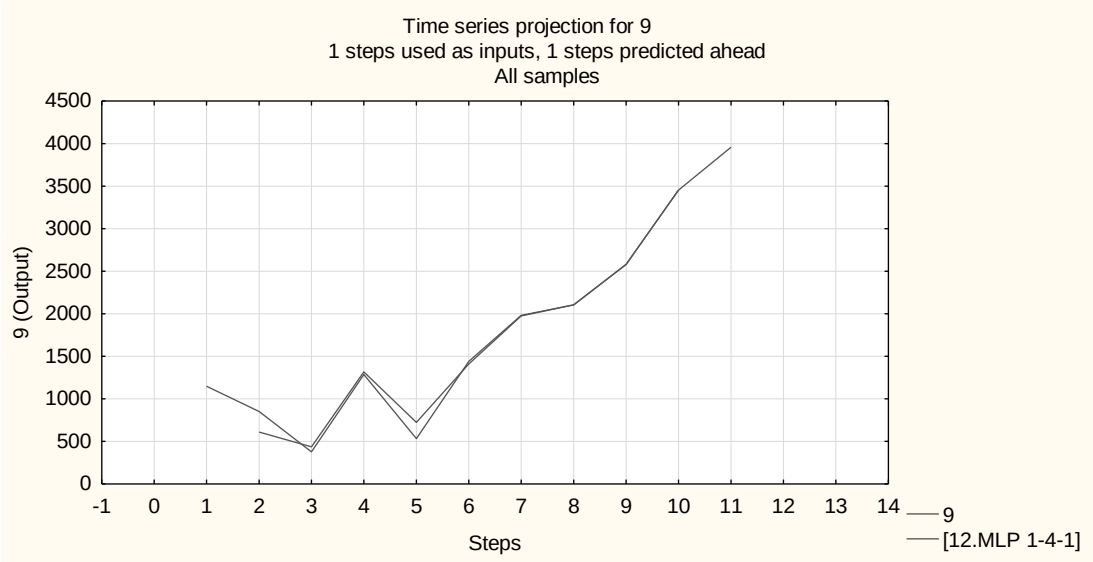


Рис. 7.3.23. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, четыре срытых нейронов и один выходной нейрон.

Case name	Time series projection for 9 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
	9 Target	9(Output) MLP 1-4-1
2007	1149,000	
2008	852,000	609,907
2009	377,000	436,235
2010	1286,000	1317,954
2011	534,000	723,287
2012	1441,000	1408,968
2013	1982,000	1975,506
2014	2104,000	2103,964
2015	2579,000	2581,722
2016	3455,000	3452,203
2017		3957,155

Рис. 7.3.24. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получили следующий результат - в 2017 г. расходы на обучение увеличатся на 14,63% и составят 3957,16 тыс. руб. по сравнению с 2016 г. в связи с переобучением сотрудников.

6) Реклама, расходы маркетинга

На этом шаге анализируются полученные сети и строятся их графическое отображение (рис. 7.3.25, рис. 7.3.26).

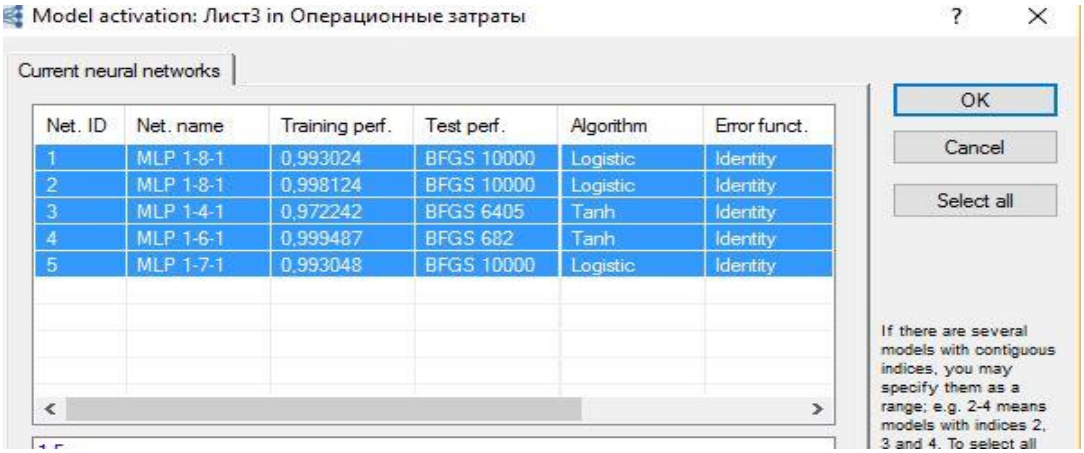


Рис. 7.3.25. Построенные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

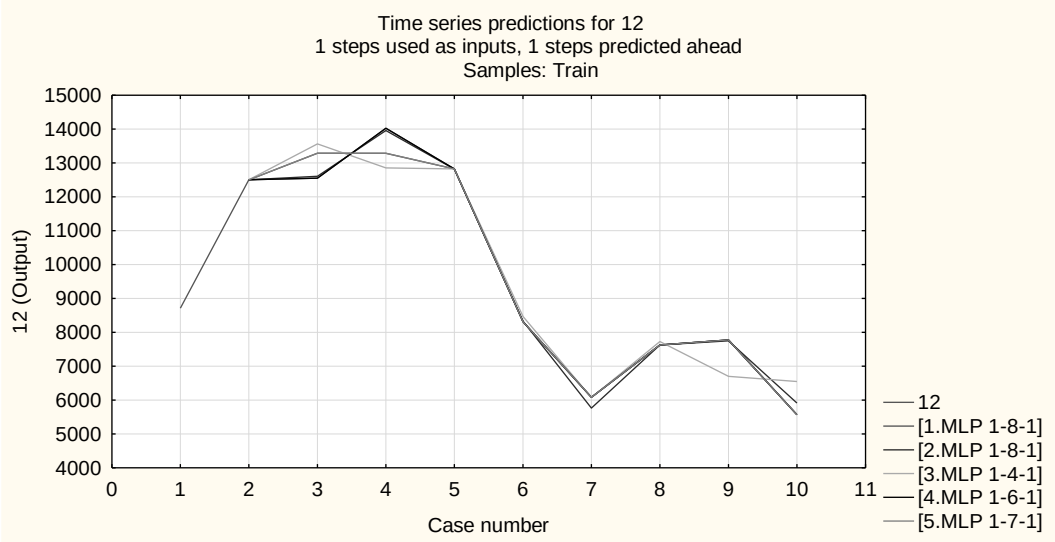


Рис. 7.3.26. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.27, рис. 7.3.28).

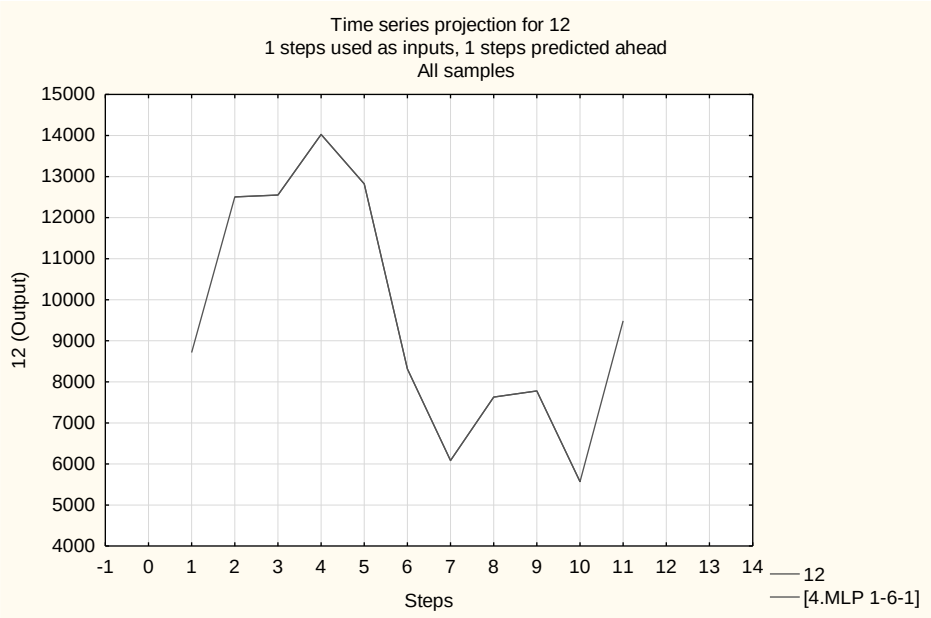


Рис. 7.3.27. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, шесть срытых нейронов и один выходной нейрон.

	Time series projection for 12 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
Case name	12 Target	12(Output) MLP 1-6-1
2007	8716,00	
2008	12504,00	12504,00
2009	12552,00	12552,00
2010	14025,00	14025,00
2011	12824,00	12824,00
2012	8316,00	8316,00
2013	6086,00	6086,00
2014	7628,00	7628,00
2015	7778,00	7778,00
2016	5569,00	5569,00
2017		9485,71

Рис. 7.3.28. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат - в 2017 г. затраты на расходы маркетинга и рекламу вырастут на 70,33% и составят 9485,71 тыс. руб. по сравнению с 2016 г. в связи с выходом на новые рынки (от Калининграда до Владивостока будет продаваться баранка владимирская и «челночок» владимирский)

7) Налог на землю.

На этом шаге анализируются полученные сети и строятся их графическое отображение (рис. 7.3.29, рис. 7.3.30).

Рис. 7.3.29. Построенные нейронные сети

По данным рисунка отмечается, что только девятнадцатая сеть обладает хорошей производительностью и высокой точностью.

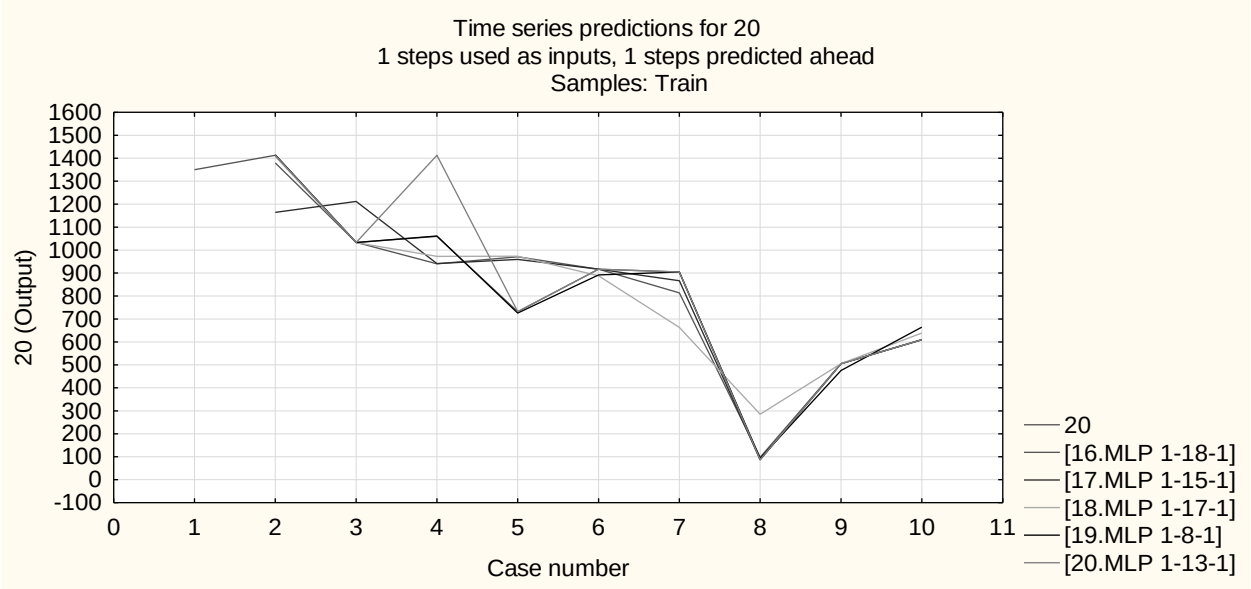


Рис. 7.3.30. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что только девятнадцатая нейронная сеть хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.31, рис. 7.3.32).

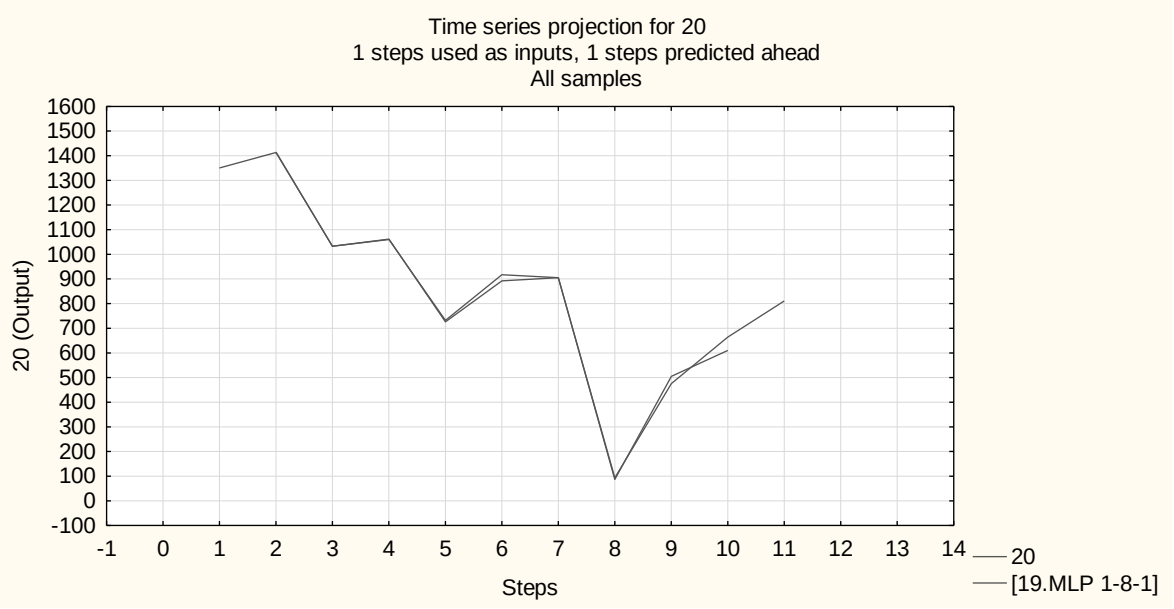


Рис. 7.3.31. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, восемь срытых нейронов и один выходной нейрон.

Case name	Time series projection for 20 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
	20 Target	20(Output) MLP 1-8-1
2007	1350,000	
2008	1413,000	1414,951
2009	1033,000	1032,948
2010	1060,000	1061,343
2011	732,000	725,976
2012	917,000	891,995
2013	905,000	904,983
2014	87,000	93,283
2015	505,000	475,826
2016	610,000	664,453
2017		811,136

Рис. 7.3.32. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат - в 2017 г. расходы на налог на землю вырастут на 22,08% и составят 811,136 тыс. руб. по сравнению с 2016 г. в связи с покупкой земель для строительства комплекса по переработке зерна и хранению муки.

8) Налог на транспорт.

На этом шаге анализируются полученные сети и строятся их графическое отображение (рис. 7.3.33, рис. 7.3.34).

Рис. 7.3.33. Построенные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

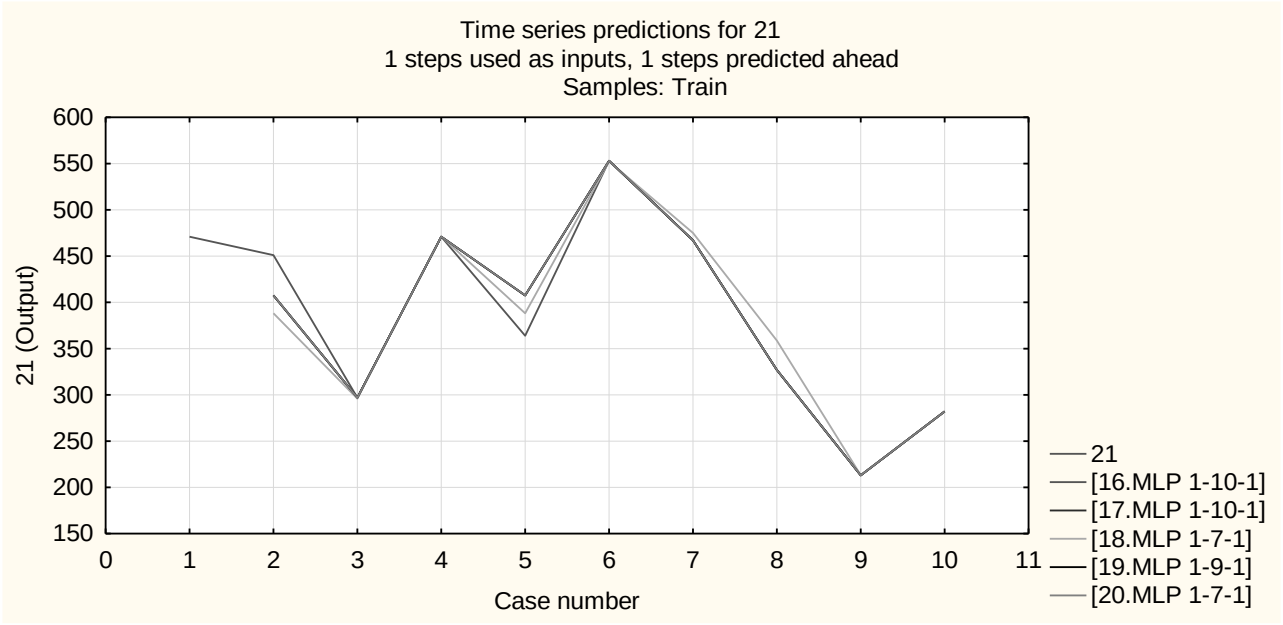


Рис. 7.3.34. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.35, рис. 7.3.36).

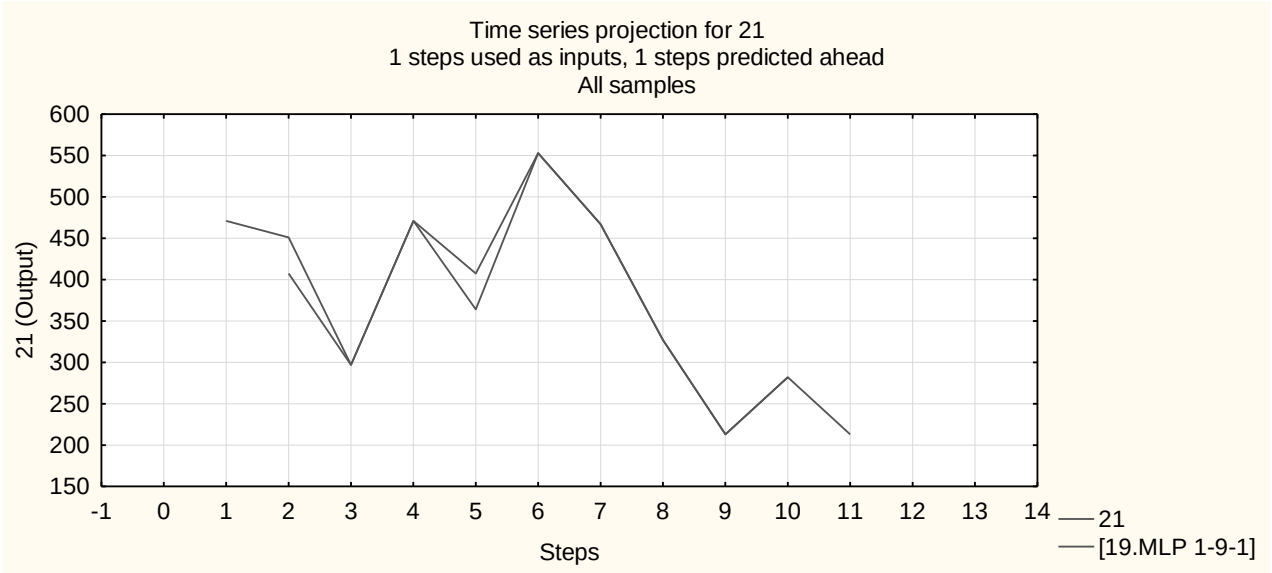


Рис. 7.3.35. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, девять срытых нейронов и один выходной нейрон.

	Time series projection for 21 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
Case name	21 Target	21(Output) MLP 1-9-1
2007	471,0000	
2008	451,0000	407,4879
2009	297,0000	296,9871
2010	471,0000	470,9999
2011	364,0000	407,4879
2012	553,0000	552,9952
2013	467,0000	466,9976
2014	327,0000	327,0480
2015	213,0000	213,0000
2016	282,0000	282,0000
2017		213,0000

Рис. 7.3.36. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат - в 2017 г. расходы на налог на транспорт снизятся на 24,46% и составят 213 тыс. руб. по сравнению с 2016 г. в связи с продажей автомобильного транспорта.

9) Чистая прибыль.

На этом шаге анализируются полученные сети и строятся их графическое отображение (рис. 7.3.37, рис. 7.3.38).

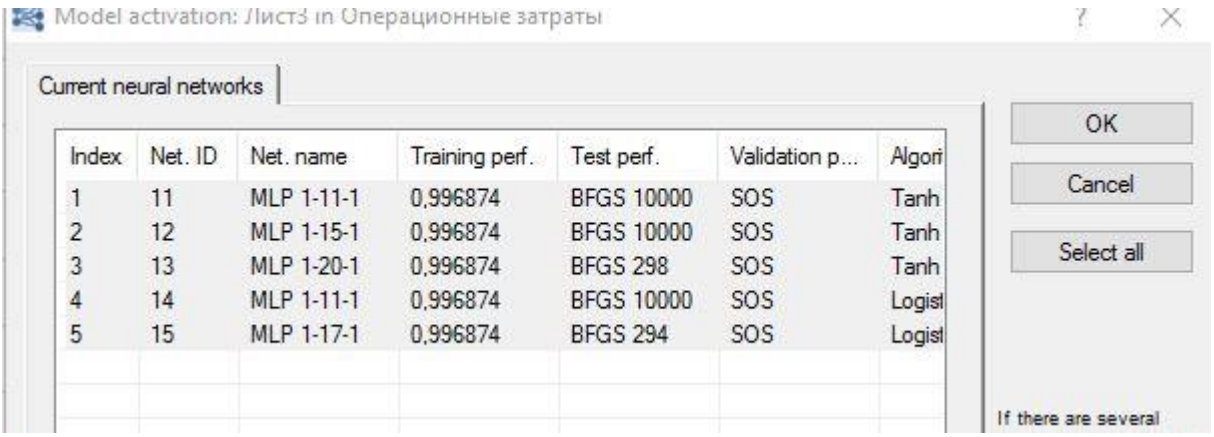


Рис. 7.3.37. Построенные нейронные сети

По данным рисунка отмечается, что все построенные сети обладают хорошей производительностью и высокой точностью.

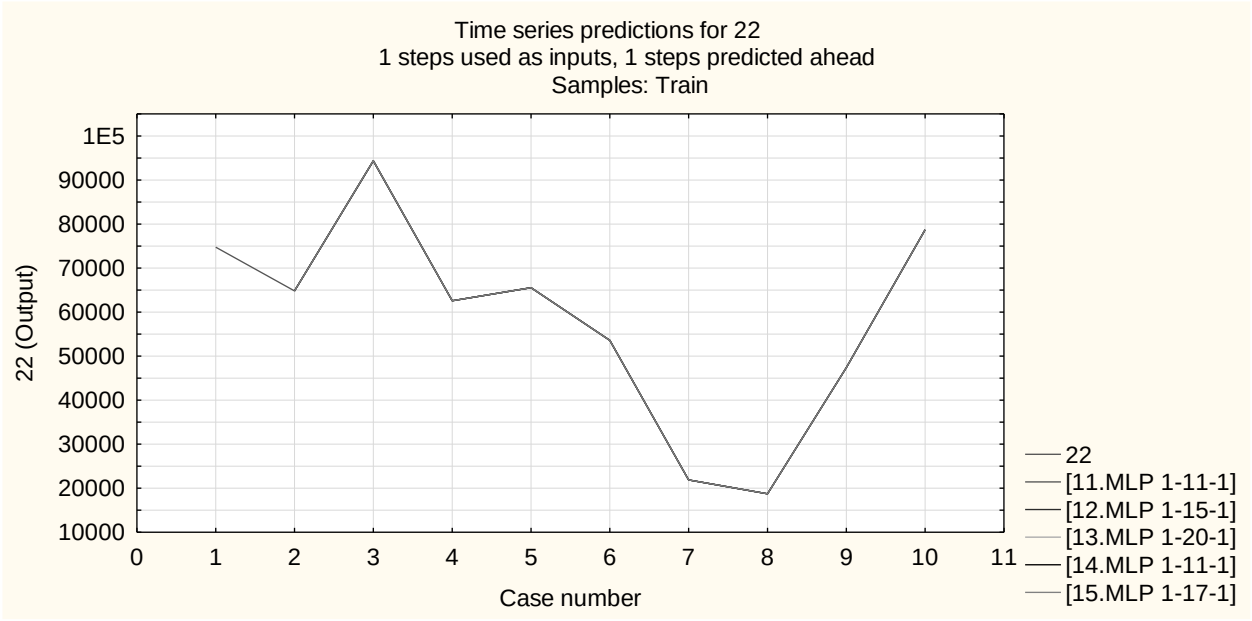


Рис. 7.3.38. Графическое отображение нейронных сетей, тыс. руб.

По данным графика делается вывод, что все построенные нейронные сети хорошо улавливают модель поведения показателя.

На следующем шаге выбирается наиболее точная и адекватная сеть и строится прогноз на 2017 год (рис. 7.3.39, рис. 7.3.40).

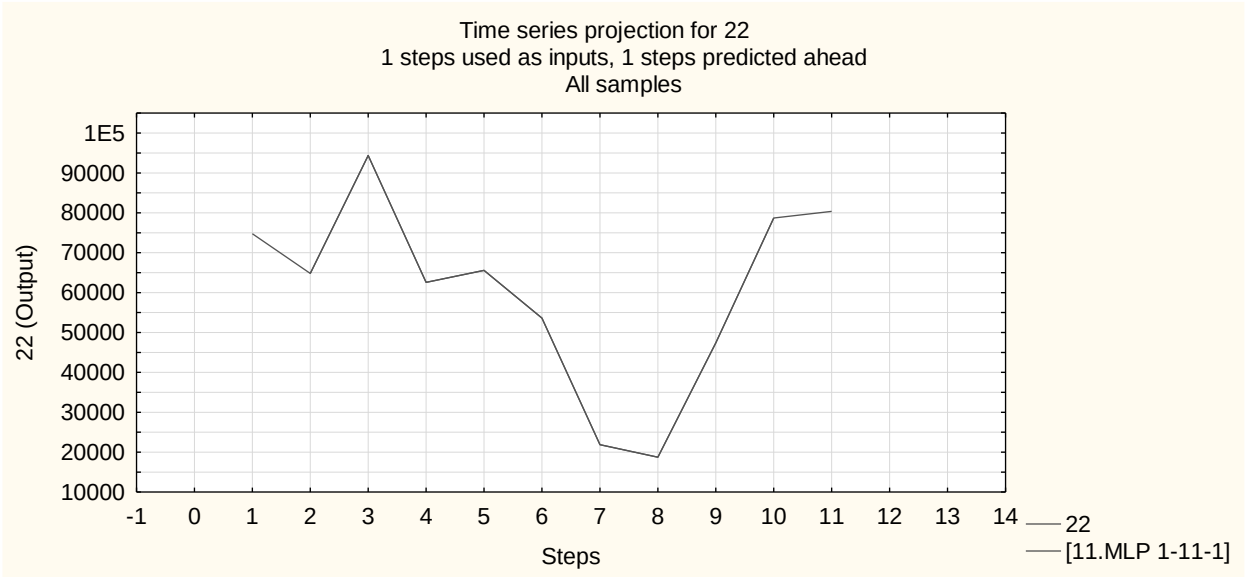


Рис. 7.3.39. Отобранная нейронная сеть, тыс. руб.

На данном рисунке отображается поведение показателя в прогнозном периоде, построенная сеть использует один входной нейрон, девять срытых нейронов и один выходной нейрон.

	Time series projection for 22 (Лист3 in Операционные затраты) 1 steps used as inputs, 1 steps predicted ahead All samples	
Case name	22 Target	22(Output) MLP 1-11-1
2007	74759,00	
2008	64807,00	64806,99
2009	94368,00	94366,50
2010	62559,00	62558,99
2011	65555,00	65555,00
2012	53592,00	53592,00
2013	21879,00	21879,00
2014	18742,00	18742,00
2015	47388,00	47388,00
2016	78704,00	78704,00
2017		80351,05

Рис. 7.3.40. Прогноз на 2017 г., тыс. руб.

В процессе моделирования получается следующий результат - в 2017 г. чистая прибыль предприятия вырастет на 2,09 % и составит 80351,05 тыс. руб. по сравнению с 2016 г. Незначительный рост связывается с увеличением расходов на маркетинг и амортизацию оборудования

Контрольные вопросы по теме

1. Что такое нейронные сети и как они используются в программном комплексе Statistica?
2. Какие типы нейронных сетей поддерживает Statistica?
3. Как создать нейронную сеть в Statistica?
4. Какие методы обучения нейронных сетей доступны в Statistica?
5. Как провести предварительную обработку данных для обучения нейронной сети в Statistica?
6. Как настраивать параметры нейронной сети в Statistica?
7. Как провести кросс-валидацию нейронной сети в Statistica?
8. Как оценить производительность нейронной сети в Statistica?
9. Как провести тестирование обученной нейронной сети в Statistica?
10. Как провести анализ результатов работы нейронной сети в Statistica?
11. Какие методы регуляризации доступны при обучении нейронных сетей в Statistica?

12. Какие функции активации можно использовать при создании нейронных сетей в Statistica?

13. Какие методы оптимизации доступны для обучения нейронных сетей в Statistica?

14. Как проводится выбор архитектуры нейронной сети в Statistica?

15. Какие методы подбора гиперпараметров используются при обучении нейронных сетей в Statistica?

16. Как провести визуализацию архитектуры нейронной сети в Statistica?

17. Как провести анализ ошибок работы нейронной сети в Statistica?

18. Как использовать предобученные модели нейронных сетей в Statistica?

19. Как провести анализ важности признаков для работы нейронной сети в Statistica?

20. Какие методы объединения нейронных сетей доступны в Statistica?

21. Как провести анализ стабильности работы нейронной сети в Statistica?

22. Как использовать рекуррентные нейронные сети в программном комплексе Statistica?

23. Как использовать сверточные нейронные сети в программном комплексе Statistica?

24. Как использовать генеративно-состязательные сети (GAN) в программном комплексе Statistica?

25. Как использовать автоэнкодеры (autoencoders) в программном комплексе Statistica?

26. Как использовать улучшенные методы обучения, такие как обучение со случайным отсевом (dropout) или батч-нормализация (batch normalization) в программном комплексе Statistica?

27. Как проводится выбор функции потерь для обучения нейронной сети в Statistica?

28. Как проводится выбор метрик для оценки производительности нейронной сети в Statistica?

29. Как проводится выбор оптимального размера батча при обучении нейронной сети в Statistica?

30. Как проводится выбор оптимального шага обучения (learning rate) при обучении нейронной сети в Statistica?

31. Какие методы подготовки данных для нейросетей предусмотрены в программном комплексе Statistica?

32. Как можно использовать техники аугментации данных при обучении нейронных сетей в Statistica?

33. Как провести анализ распределения классов при работе с несбалансированными данными при использовании нейросетей в Statistica?

34. Как реализованы возможности работы с текстовыми данными при использовании нейросетей в программном комплексе Statistica?

35. Как реализованы возможности работы с изображениями при использовании нейросетей в программном комплексе Statistica?

36. Как реализованы возможности работы с временными рядами при использовании нейросетей в программном комплексе Statistica?

37. Как провести анализ переобучения (overfitting) и недообучения (underfitting) при работе с нейросетями в программном комплексе Statistica?

38. Как провести анализ градиентов и их влияния на обучение нейросети в программном комплексе Statistica?

39. Как провести анализ работы различных слоев и их влияния на производительность нейросети в программном комплексе Statistica?

40. Как провести анализ работы различных функций активации и их влияния на производительность нейросети в программном комплексе Statistica?

41. Как провести анализ работы различных методов оптимизации и их влияния на производительность нейросети в программном комплексе Statistica?

42. Как провести анализ работы различных методов регуляризации и их влияния на производительность нейросети в программном комплексе Statistica?

43. Как провести анализ работы различных методов подбора гиперпараметров и их влияния на производительность нейросети в программном комплексе Statistica?

44. Как провести анализ работы различных методов объединения слоев и их влияния на производительность нейросети в программном комплексе Statistica?

45. Как провести анализ работы различных методов рекуррентных связей и их влияния на производительность нейросети в программном комплексе Statistica?

46. Как провести анализ работы различных методов сверточных операций и их влияния на производительность нейросети в программном комплексе Statistica?

47. Как провести анализ работы различных методов генеративно-состязательных сетей (GAN) и их применимость для конкретных задач в программном комплексе Statistica?

48. Как провести анализ работы различных методов автоэнкодеров (autoencoders) и их применимость для конкретных задач в программном комплексе Statistica?

49. Как провести анализ работы различных методов улучшенных методов обучения, таких как обучение со случайным отсевом (dropout) или батч-нормализация (batch normalization), и их применимость для конкретных задач в программном комплексе Statistica?

50. Как провести анализ работы различных методов подготовки данных, таких как техники аугментации данных, при работе с текстом, изображениями или временными рядами, и их применимость для конкретных задач в программном комплексе Statistica?

8. ЗАДАНИЯ ДЛЯ ПРОВЕДЕНИЯ ПРАКТИЧЕСКИХ ЗАНЯТИЙ ПО КУРСУ

Занятие 1. Описательная статистика

На основе данных по вариантам (таблица 8.1) необходимо выполнить расчет **в MS Excel и Statistica:**

- среднее (статистическая оценка математического ожидания);

- стандартная ошибка (среднего);

- медиана (M_e) — значение признака, приходящееся на середину ранжированной (упорядоченной) совокупности;

- мода (M_o) — значение изучаемого признака, повторяющегося с наибольшей частотой;

- дисперсия выборки;

- стандартное отклонение (среднее квадратическое отклонение);

- эксцесс;

- асимметричность (асимметрия);

- интервал (размах выборки);

- минимальное значение выборки x_{min};

- максимальное значение выборки x_{max};

- сумма всех значений выборки;

- объем выборки n;

- наибольшее значение признака, имеющее разность с порядком x_{max} k единиц;

- наименьшее значение признака, имеющее разность с порядком x_{min} k единиц;

- уровень надежности (предельная ошибка выборки).

Сделать вывод.

Таблица 8.1

Исходные данные для анализа

	Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
Регион 1	72770	71547	72357	72273	72985
Регион 2	19531	19375	19597	19612	19843
Регион 3	741	748	720	718	734
Регион 4	665	652	636	649	656
Регион 5	798	785	808	793	807
Регион 6	1171	1106	1140	1110	1097
Регион 7	592	572	567	565	574
Регион 8	561	542	521	539	550
Регион 9	380	374	371	362	367
Регион 10	636	609	616	582	592
Регион 11	601	598	607	604	592
Регион 12	3612	3505	3505	3618	3705

Регион 13	438	429	443	429	429
Регион 14	605	587	619	589	588
Регион 15	548	538	539	524	528
Регион 16	574	527	552	547	522
Регион 17	750	735	739	718	715
Регион 18	837	828	803	804	824
Регион 19	746	723	713	722	698
Регион 20	5277	5517	5699	5738	5864

Занятие 2. Корреляционно-регрессионный анализ

На основе данных по вариантам, представленных в таблицах 8.2 и 8.3:

- произвести расчет матрицы корреляций и сделать вывод о наличии/отсутствии связи и ее силе;
- построить график корреляционного поля;
- провести построение линейного парного регрессионного уравнения;
- оценить качество полученной модели;
- построить графическую интерпретацию регрессионной модели.

Таблица 8.2

Исходные данные для анализа, y_i

Период	Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
1	42074,5	24650,5	33017,7	49523,9	16900,0
2	49941,8	30110,3	42075,4	60014,6	22175,9
3	62404,4	37374,1	50359,9	83001,1	26981,3
4	76054,5	43700,3	61818,6	100143,3	33214,6
5	114409,3	51003,4	74207,0	117197,6	40159,4
6	144987,8	66692,3	86926,8	133586,6	44415,4
7	178846,1	82100,4	112841,7	166176,5	55090,0
8	237013,3	102706,2	146663,0	222811,9	74752,0
9	317656,3	125834,4	175395,7	287072,1	86980,3
10	304345,3	126477,4	185824,6	301729,1	87061,9
11	398361,4	147024,0	224759,2	346568,2	109884,5
12	507839,8	174211,8	261222,6	474973,9	128905,4
13	545517,2	207397,5	286018,6	563965,4	136115,0
14	569006,4	219502,8	306641,4	611720,4	158228,7

15	619677,7	242722,4	328064,2	717667,2	151876,8
16	693379,4	271782,5	368489,2	805969,6	180517,5
17	778027,8	316489,4	431549,8	827928,6	205818,6
18	837306,8	341177,8	449849,2	873429,4	212464,7
19	911597,9	367157,1	480027,8	951292,3	232493,6
20	955329,2	399113,8	535493,4	1001790,3	254968,9
21	999081,6	412335,5	554204,3	1063999,2	271653,7

Таблица 8.3

Исходные данные для анализа, x_i

Период	Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
1	741	665	798	1171	592
2	748	652	785	1106	572
3	720	636	808	1140	567
4	718	649	793	1110	565
5	734	656	807	1097	574
6	715	650	811	1154	572
7	756	661	797	1174	556
8	732	659	797	1147	564
9	761	654	786	1158	567
10	791	637	764	1168	545
11	779	650	762	1152	556
12	767	644	773	1172	553
13	787	644	776	1164	557
14	810	638	770	1154	547
15	814	635	759	1161	538
16	806	624	760	1162	548
17	822	625	737	1165	536
18	824	613	733	1179	544
19	825	610	722	1185	525
20	826	595	721	1182	517
21	834	583	710	1172	514

Занятие 3. Дисперсионный анализ

На основе данных по вариантам, представленных в таблицах 8.4 и 8.5, 8.6 провести многофакторный анализ, сделать выводы.

Таблица 8.4

Исходные данные для анализа, варианты 1-2

Округ	Вариант 1		Вариант 2	
	Число объектов	Характеристика	Число объектов	Характеристика
1	1836	8,6	2451	8,6
	2073	10,4	2053	10,4
	1668	8,9	1398	8,9
	1614	6,9	1201	6,9
	2317	6,6	1883	6,6
	1696	5,0	1188	5,0
2	1393	8,1	2524	8,1
	2116	6,0	1967	6,0
	2499	7,2	2049	7,2
	1050	7,0	1814	7,0
	2069	10,3	1400	10,3
	1245	6,3	1600	6,3
3	1295	53,2	2335	53,2
	1451	14,6	1338	14,6
	1131	12,0	2061	12,0
	1163	10,4	1044	10,4
	2184	35,0	1733	35,0
	1126	6,3	1675	6,0
4	2271	53,2	1284	9,2
	1135	14,6	1881	6,6
	2453	7,2	1765	6,3
	2286	6,0	1700	6,3
	1710	9,2	2294	53,2
	992	6,6	2047	14,6
5	1235	6,3	1825	7,2
	1473	53,2	2253	6,0
	2021	14,6	1886	9,2
	1987	8,7	2386	6,6
	1874	11,7	1148	6,3
	1882	8,1	2195	53,2

Таблица 8.5

Исходные данные для анализа, варианты 4-5

Округ	Вариант 3		Вариант 4	
	Число объектов	Характеристика	Число объектов	Характеристика

1	2011	6,0	2008	6,3
	2192	9,2	1708	53,2
	1044	6,6	2100	14,6
	1996	6,3	2531	7,2
	1607	6,3	2093	6,0
	1044	53,2	1797	9,2
2	1063	14,6	1396	6,6
	2118	7,2	2250	6,3
	1053	6,0	1055	53,2
	1046	9,2	2252	14,6
	1037	6,6	2393	8,7
	1395	6,3	2176	11,7
3	2493	53,2	1768	5,0
	1544	14,6	1385	8,1
	1689	8,7	2028	6,0
	1213	11,7	1271	7,2
	1227	8,1	2144	7,0
	1022	6,0	2126	10,3
4	1885	7,2	1159	6,3
	2206	7,0	1244	6,3
	1718	10,3	1949	53,2
	1215	6,3	1086	14,6
	2330	53,2	1634	7,2
	2091	14,6	2313	6,0
5	1663	12,0	1863	9,2
	2522	10,4	2090	6,6
	1224	35,0	1363	6,3
	2091	6,3	1705	53,2
	1160	53,2	2244	14,6
	1973	14,6	2355	8,7

Таблица 8.6

Исходные данные для анализа, вариант 5

	Вариант 5	
Округ	Число объектов	Характеристика
1	2095	6,0

	2147	9,2
	1497	6,6
	1627	6,3
	2497	53,2
	2507	6,9
2	1661	6,6
	1359	5,0
	2374	8,1
	1539	6,0
	1721	7,2
	2130	7,0
3	2133	10,3
	1803	6,3
	2027	53,2
	2559	14,6
	2024	12,0
	1921	10,4
4	1788	6,3
	1741	53,2
	1999	14,6
	1960	7,2
	2190	6,0
	1948	9,2
5	2440	6,6
	2014	6,3
	1842	53,2
	1699	14,6
	1754	8,7
	1574	11,7

Занятие 4. Кластерный анализ

На основе данных, представленных ниже в таблицах провести кластерный анализ иерархическим методом, выполнить построение дендрограмм, сделать вывод.

Таблица 8.7

Исходные данные для анализа, вариант 1

	x_1	x_2	x_3	x_4	x_5	x_6
1	4,8	40 272	27,7	229	347854	425954
2	3,5	54 727	31,8	203	175219	228500
3	4,2	35 612	21,5	259	932	984
4	3,4	31 608	23,0	168	320	1539
5	3,9	28 489	24,9	199	1892	2975
6	3,8	35 100	26,8	273	6145	6410
7	4,5	28 680	24,1	180	423	944
8	4,0	35 028	21,6	234	3023	5339
9	4,4	28 560	21,6	199	100	141
10	4,0	32 715	19,7	185	1044	1077
11	4,2	35 124	23,8	228	350	187
12	3,4	53 793	33,4	235	40667	42671
13	4,7	29 846	24,9	197	289	901
14	4,0	30 495	26,8	227	984	1215
15	5,0	30 731	25,1	213	501	348
16	3,9	30 241	28,3	159	544	1044
17	3,9	30 528	22,3	182	1815	2980
18	3,8	32 131	26,9	117	3174	3976
19	5,9	33 124	30,2	217	2561	3010
20	2,6	88 831	36,3	179	110455	152759
21	3,9	44 531	27,7	261	46573	60224
22	6,6	35 173	28,6	218	542	563
23	7,0	40 877	26,7	255	44285,4	1025
24	9,0	37 061	28,0	184	48160,2	744
25	10,6	33 246	27,5	230	52035	801
26	7,1	29 431	27,1	225	55909,8	11
27	6,6	25 615	26,7	279	59784,6	733
28	4,7	21 800	25,8	231	63659,4	288
29	5,0	17 985	21,4	230	67534,2	941
30	3,7	14 169	25,0	277	71409	2409

Таблица 8.8

Исходные данные для анализа, вариант 2

	x_1	x_2	x_3	x_4	x_5	x_6
1	50,0	691	849	268	700	445
2	38,0	448	470	739	382	114

3	73,0	529	497	742	713	715
4	55,0	649	480	176	584	518
5	13,0	601	346	717	319	306
6	29,0	790	855	219	329	402
7	68,0	377	363	371	689	285
8	64,0	571	484	415	618	402
9	17,0	486	652	373	839	510
10	77,0	489	435	307	730	736
11	101,0	361	317	520	737	471
12	60,0	239	297	575	548	744
13	76,0	552	733	688	191	577
14	13,0	631	567	170	422	496
15	52,0	802	559	812	524	722
16	81,0	301	792	664	670	274
17	30,0	338	566	451	325	187
18	62,0	626	542	435	584	334
19	51,0	763	407	765	819	473
20	19,0	193	826	456	205	230
21	26,0	671	199	408	750	135
22	37,0	217	480	618	669	403
23	45,0	311	462	603	546	422
24	50,0	340	792	742	650	788
25	64,0	345	702	209	818	282
26	50,0	539	452	158	415	371
27	33,0	311	488	296	141	797
28	93,0	277	450	315	217	466
29	31,0	435	624	490	726	316
30	32,0	742	148	514	546	655

Таблица 8.9

Исходные данные для анализа, вариант 3

	x_1	x_2	x_3	x_4	x_5	x_6
1	85,0	176	449	310	415	351
2	70,0	160	527	388	412	393

3	93,0	583	377	377	401	313
4	96,0	723	404	586	533	209
5	71,0	810	837	123	575	554
6	64,0	674	555	217	355	261
7	95,0	713	368	844	459	523
8	11,0	316	327	135	780	240
9	26,0	243	848	745	110	291
10	100,0	512	262	556	374	257
11	32,0	567	144	666	579	811
12	54,0	751	140	216	553	394
13	32,0	725	121	639	406	175
14	70,0	605	637	160	848	625
15	95,0	326	320	371	748	128
16	95,0	467	316	311	499	817
17	13,0	368	339	557	483	775
18	43,0	217	344	506	460	411
19	63,0	775	719	686	332	210
20	18,0	714	727	155	383	798
21	39,0	506	556	450	346	475
22	11,0	344	811	660	644	151
23	47,0	240	842	511	490	601
24	76,0	691	156	157	819	349
25	43,0	256	601	466	699	783
26	81,0	646	434	713	710	559
27	10,0	260	505	528	818	184
28	11,0	848	403	615	351	536
29	59,0	783	181	199	426	514
30	69,0	799	648	625	477	279

Таблица 8.10

Исходные данные для анализа, вариант 4

	x_1	x_2	x_3	x_4	x_5	x_6
1	39,0	726	491	315	816	695
2	70,0	242	609	449	260	448

3	45,0	377	108	223	107	417
4	78,0	837	784	854	260	360
5	37,0	826	348	786	677	692
6	14,0	701	447	534	635	325
7	84,0	440	473	609	604	444
8	36,0	213	191	247	209	388
9	56,0	589	504	253	763	545
10	44,0	569	208	561	269	589
11	98,0	224	268	601	312	453
12	27,0	852	493	761	776	849
13	91,0	474	711	316	501	352
14	11,0	707	371	188	650	251
15	40,0	517	150	783	674	550
16	59,0	384	560	185	241	668
17	100,0	227	488	706	840	201
18	33,0	779	323	547	351	347
19	58,0	533	201	179	797	538
20	19,0	433	373	252	332	755
21	62,0	254	732	398	358	753
22	61,0	155	710	750	300	294
23	64,0	770	114	674	243	244
24	47,0	574	614	615	407	164
25	89,0	224	283	753	511	564
26	85,0	350	772	518	470	283
27	22,0	213	563	166	574	837
28	100,0	687	332	587	143	267
29	98,0	177	594	627	763	415
30	53,0	544	468	658	562	180

Таблица 8.11

Исходные данные для анализа, вариант 5

	x_1	x_2	x_3	x_4	x_5	x_6
1	62,0	565	603	638	320	368
2	101,0	549	543	396	248	251

3	98,0	394	174	767	390	143
4	70,0	173	689	540	645	401
5	85,0	685	404	268	792	714
6	61,0	194	309	515	238	197
7	91,0	311	736	689	239	291
8	21,0	246	748	312	227	658
9	42,0	778	477	524	768	615
10	90,0	448	762	528	539	812
11	87,0	224	345	193	295	822
12	61,0	612	812	307	333	799
13	58,0	555	671	624	666	790
14	98,0	270	254	663	285	760
15	28,0	811	581	197	756	747
16	42,0	499	658	673	489	158
17	75,0	340	151	559	228	236
18	13,0	845	536	813	466	647
19	95,0	329	657	210	233	200
20	79,0	693	749	714	800	283
21	93,0	804	445	224	450	539
22	74,0	457	351	593	264	123
23	12,0	385	725	135	624	817
24	62,0	406	254	516	208	736
25	84,0	427	421	435	554	694
26	58,0	599	841	164	728	480
27	19,0	179	402	541	769	617
28	37,0	564	482	466	499	151
29	73,0	224	633	555	685	581
30	95,0	559	144	334	552	817

Занятие 5. Дискриминантный анализ

Существует набор регионов, которые характеризуются набором данных за отдельно взятый период. Есть обучающая выборка.

Необходимо провести классификацию регионов, не вошедших в обучающую выборку.

Таблица 8.12

Исходные данные для анализа, вариант 1

	x_1	x_2	x_3	x_4	x_5	x_6	Класс
1	71,0	188	469	228	570	244	
2	101,0	400	758	765	540	632	
3	68,0	412	481	165	447	821	2
4	89,0	581	842	743	824	576	2
5	81,0	211	595	487	265	452	1
6	89,0	383	675	512	116	739	1
7	63,0	643	587	636	796	780	
8	48,0	509	759	368	753	480	
9	11,0	671	630	405	826	574	
10	100,0	776	567	280	165	284	3
11	61,0	616	716	258	659	586	3
12	96,0	708	772	207	560	744	2
13	42,0	478	134	514	245	796	2
14	91,0	360	798	841	302	594	1
15	56,0	850	667	689	640	370	1
16	39,0	414	251	566	257	211	
17	47,0	464	225	497	118	764	
18	35,0	554	376	389	328	310	
19	19,0	495	331	219	547	841	3
20	102,0	529	583	581	699	204	3
21	25,0	487	665	848	559	558	3
22	62,0	522	473	428	233	418	
23	75,0	166	145	264	775	299	
24	48,0	568	264	737	400	134	2
25	43,0	729	266	464	209	222	1
26	83,0	692	545	617	584	207	1
27	32,0	746	771	684	688	538	
28	99,0	397	504	538	853	355	2
29	11,0	679	259	105	549	551	2
30	14,0	574	688	644	737	164	2

Таблица 8.13

Исходные данные для анализа, вариант 2

	x_1	x_2	x_3	x_4	x_5	x_6	Класс
1	23,0	375	394	314	519	855	
2	14,0	450	603	537	372	793	

3	84,0	797	204	188	580	220	2
4	19,0	450	293	166	776	676	2
5	37,0	321	430	192	586	771	1
6	82,0	479	506	108	608	665	1
7	69,0	809	809	463	544	223	
8	61,0	118	383	226	690	795	
9	10,0	180	742	768	855	666	
10	86,0	160	191	437	795	539	3
11	20,0	471	272	493	657	527	3
12	57,0	516	567	658	290	679	1
13	13,0	215	508	651	521	425	2
14	39,0	830	842	615	676	162	1
15	93,0	209	699	660	268	527	1
16	26,0	481	749	516	506	604	
17	72,0	737	471	379	726	508	
18	17,0	267	813	835	506	836	
19	67,0	145	188	630	395	110	3
20	100,0	232	258	110	531	599	3
21	11,0	578	546	725	794	288	3
22	22,0	730	124	731	409	405	
23	47,0	165	607	110	769	471	
24	83,0	555	248	335	596	799	2
25	67,0	826	829	819	641	486	1
26	13,0	609	536	120	687	336	1
27	85,0	781	235	148	231	465	
28	92,0	277	815	205	357	213	2
29	36,0	392	265	702	187	732	2
30	69,0	343	663	186	137	316	2

Таблица 8.14

Исходные данные для анализа, вариант 3

	x_1	x_2	x_3	x_4	x_5	x_6	Класс
1	103,0	270	520	477	448	150	
2	46,0	201	186	336	772	542	

3	44,0	371	479	718	551	399	2
4	75,0	301	716	825	825	177	2
5	72,0	173	627	519	364	149	1
6	18,0	636	391	283	817	619	1
7	50,0	336	635	158	388	143	
8	18,0	134	171	807	319	302	
9	57,0	135	277	589	358	431	
10	62,0	476	305	386	447	740	2
11	80,0	408	703	707	770	599	1
12	96,0	729	519	216	379	150	1
13	61,0	262	125	805	210	312	2
14	94,0	693	168	256	744	293	1
15	83,0	640	836	248	391	226	1
16	103,0	520	842	136	341	396	
17	52,0	780	557	304	463	545	
18	81,0	556	286	318	421	290	
19	65,0	742	233	359	425	118	2
20	39,0	835	162	677	522	139	2
21	85,0	730	635	157	542	827	1
22	15,0	515	514	613	357	140	
23	81,0	326	851	437	500	242	
24	99,0	463	277	376	478	522	2
25	96,0	156	469	475	687	670	1
26	24,0	321	804	420	693	813	1
27	51,0	824	343	112	443	720	
28	57,0	241	734	751	510	822	2
29	82,0	853	747	667	740	756	2
30	20,0	306	459	697	537	526	2

Таблица 8.15

Исходные данные для анализа, вариант 4

	x_1	x_2	x_3	x_4	x_5	x_6	Класс
1	91,0	848	311	698	792	492	
2	21,0	853	356	195	704	809	

3	11,0	222	851	709	216	228	2
4	103,0	677	667	828	198	551	1
5	57,0	815	726	828	456	124	1
6	102,0	590	386	456	349	634	1
7	87,0	507	118	447	601	469	
8	33,0	826	327	693	338	330	
9	54,0	125	365	204	404	223	
10	16,0	545	482	133	144	772	2
11	78,0	794	645	214	401	709	1
12	87,0	431	220	403	368	179	1
13	69,0	850	293	295	125	791	2
14	15,0	196	231	176	401	620	1
15	58,0	477	575	289	804	344	1
16	45,0	141	804	700	584	175	
17	28,0	824	606	366	206	487	
18	31,0	830	644	338	564	562	
19	60,0	461	219	433	705	538	2
20	88,0	619	155	481	349	643	2
21	85,0	576	125	293	266	355	1
22	79,0	829	398	633	635	152	
23	73,0	227	109	292	108	534	
24	19,0	563	343	649	166	407	2
25	68,0	791	410	519	710	179	1
26	85,0	839	169	272	250	558	1
27	43,0	475	277	260	584	506	
28	47,0	783	240	622	451	579	2
29	16,0	793	143	219	194	617	2
30	38,0	674	771	302	210	721	2

Таблица 8.16

Исходные данные для анализа, вариант 5

	x_1	x_2	x_3	x_4	x_5	x_6	Класс
1	73,0	372	592	177	622	216	
2	13,0	596	561	742	306	281	

3	24,0	205	515	467	419	808	2
4	57,0	461	147	395	356	585	1
5	19,0	377	410	382	552	423	2
6	20,0	243	746	524	712	833	1
7	94,0	335	394	609	718	121	
8	99,0	733	773	386	343	517	
9	61,0	789	783	513	737	601	
10	38,0	408	672	552	314	207	2
11	66,0	843	300	201	156	577	1
12	81,0	332	273	311	718	136	1
13	64,0	407	844	676	342	337	2
14	56,0	374	502	591	382	358	1
15	12,0	522	584	300	654	756	1
16	87,0	393	404	492	170	397	
17	58,0	391	693	638	835	615	
18	90,0	356	643	117	577	470	
19	65,0	292	286	638	204	589	2
20	30,0	661	573	321	521	768	2
21	33,0	758	579	646	391	247	1
22	46,0	790	392	201	320	516	
23	41,0	354	728	662	113	474	
24	33,0	621	175	855	270	179	2
25	20,0	497	592	505	816	276	1
26	24,0	136	613	520	136	767	1
27	36,0	535	388	834	629	706	
28	90,0	423	680	275	727	505	2
29	25,0	519	397	146	701	719	2
30	41,0	693	279	669	528	515	2

Занятие 6. Факторный анализ

Провести факторный анализ на основе представленных ниже таблиц по вариантам.

Таблица 8.17

Исходные данные для анализа, вариант 1

	x_1	x_2	x_3	x_4	x_5	x_6
1	83,0	394	570	668	194	203
2	13,0	804	243	340	806	553
3	89,0	637	348	535	822	574
4	10,0	778	677	349	835	323
5	86,0	395	826	322	443	745
6	24,0	762	181	797	736	129
7	90,0	488	324	808	156	220
8	92,0	324	621	846	742	815
9	16,0	474	431	840	187	410
10	41,0	458	368	702	281	621
11	23,0	327	805	119	475	177
12	42,0	202	343	785	146	109
13	46,0	592	182	417	612	658
14	23,0	590	376	211	112	468
15	36,0	758	202	605	394	795
16	44,0	656	462	445	289	853
17	35,0	579	769	304	836	698
18	95,0	516	668	249	631	357
19	75,0	414	849	549	501	193
20	23,0	434	516	589	353	642
21	75,0	833	367	767	757	405
22	44,0	297	739	633	655	118
23	44,0	665	235	364	704	244
24	44,0	803	390	430	228	640
25	35,0	745	596	297	341	114
26	43,0	535	322	547	220	817
27	91,0	346	372	223	732	146
28	32,0	230	683	528	742	204
29	79,0	482	159	664	792	847
30	35,0	588	427	491	709	284
31	84,0	810	202	162	335	842
32	22,0	203	390	621	474	755
33	96,0	106	131	443	817	801
34	26,0	609	359	351	650	536
35	75,0	341	661	412	372	673
36	49,0	726	678	597	571	714
37	88,0	299	269	692	204	323

38	93,0	311	703	487	769	163
39	46,0	840	235	253	766	666
40	86,0	142	716	773	354	356
41	32,0	361	315	451	466	799
42	68,0	379	438	384	735	446
43	33,0	201	687	503	412	326
44	62,0	172	623	542	389	521
45	71,0	484	558	340	507	533
46	82,0	796	338	262	668	538
47	47,0	734	418	114	427	171
48	91,0	517	261	309	215	585
49	36,0	496	133	406	835	438
50	15,0	399	587	320	838	690

Таблица 8.18

Исходные данные для анализа, вариант 2

	x_1	x_2	x_3	x_4	x_5	x_6
1	25,0	641	457	223	337	143
2	34,0	584	439	628	261	515
3	47,0	386	591	158	365	723
4	73,0	536	546	786	358	216
5	53,0	289	813	778	817	685
6	100,0	263	206	165	607	820
7	21,0	670	842	113	258	616
8	44,0	417	473	776	422	286
9	24,0	832	404	628	337	115
10	30,0	522	114	170	117	190
11	56,0	696	487	315	128	625
12	96,0	105	662	767	310	794
13	39,0	664	638	338	204	142
14	81,0	642	818	660	562	326
15	75,0	256	794	391	205	812
16	33,0	378	479	685	496	230
17	28,0	311	662	366	711	595
18	98,0	315	605	688	213	814
19	100,0	608	823	602	257	706
20	71,0	448	499	471	762	422
21	55,0	198	113	637	707	803

22	84,0	303	683	332	159	159
23	97,0	666	437	161	813	121
24	27,0	596	836	307	175	126
25	95,0	255	576	713	454	236
26	65,0	544	391	377	134	141
27	94,0	159	529	363	296	144
28	41,0	308	377	634	784	640
29	66,0	506	118	416	844	283
30	27,0	494	708	784	359	614
31	33,0	289	829	684	761	645
32	16,0	237	344	430	133	427
33	91,0	483	826	182	247	362
34	95,0	511	364	686	680	205
35	30,0	546	614	200	663	124
36	33,0	544	598	253	245	694
37	76,0	786	634	766	578	591
38	53,0	589	398	297	365	296
39	24,0	785	397	189	647	348
40	34,0	203	668	597	798	673
41	60,0	123	301	272	817	148
42	42,0	416	730	405	628	251
43	32,0	179	293	328	614	568
44	93,0	189	649	387	487	642
45	103,0	555	411	765	818	255
46	75,0	537	440	300	286	588
47	62,0	626	378	687	782	252
48	41,0	651	215	573	745	680
49	33,0	736	620	777	262	346
50	84,0	161	709	301	239	105

Таблица 8.19

Исходные данные для анализа, вариант 3

	x_1	x_2	x_3	x_4	x_5	x_6
1	56,0	498	175	663	241	534
2	101,0	192	599	205	108	411

3	71,0	803	791	531	423	493
4	103,0	105	310	498	181	227
5	12,0	784	614	617	483	317
6	38,0	833	475	509	378	715
7	12,0	668	736	317	763	503
8	81,0	173	257	521	198	415
9	83,0	522	503	538	408	272
10	74,0	724	367	434	246	223
11	101,0	837	682	563	716	504
12	53,0	258	438	265	615	727
13	86,0	120	329	721	602	566
14	64,0	352	787	615	794	477
15	61,0	388	288	165	344	512
16	26,0	500	384	734	526	233
17	41,0	546	545	154	180	522
18	52,0	732	617	805	363	494
19	36,0	620	398	734	417	495
20	73,0	177	802	391	381	686
21	20,0	134	138	617	284	333
22	57,0	478	159	792	530	440
23	87,0	412	490	698	836	593
24	73,0	281	161	144	733	455
25	70,0	571	692	805	603	110
26	97,0	758	176	770	795	656
27	58,0	438	384	109	229	166
28	53,0	122	385	156	651	478
29	33,0	365	200	406	621	473
30	99,0	168	334	548	257	847
31	100,0	364	247	701	611	444
32	17,0	405	516	794	737	639
33	15,0	531	330	812	732	848
34	45,0	144	158	435	329	156
35	36,0	510	806	551	725	198
36	56,0	179	457	244	487	621
37	27,0	657	389	255	541	371
38	23,0	539	269	142	769	583
39	12,0	123	576	145	544	813
40	51,0	416	846	247	253	298

41	23,0	179	715	640	195	682
42	89,0	743	788	501	163	686
43	72,0	385	782	641	734	400
44	97,0	825	434	686	488	241
45	57,0	144	554	788	760	131
46	24,0	477	687	777	758	152
47	21,0	745	109	658	454	452
48	87,0	446	794	636	204	467
49	33,0	363	267	111	427	526
50	66,0	480	489	255	497	772

Таблица 8.20

Исходные данные для анализа, вариант 4

	x_1	x_2	x_3	x_4	x_5	x_6
1	78,0	315	182	750	536	159
2	99,0	321	294	654	835	790
3	66,0	761	528	484	327	821
4	75,0	121	480	161	224	414
5	67,0	195	357	787	832	364
6	58,0	767	703	783	772	611
7	95,0	190	779	469	409	531
8	33,0	613	239	381	841	649
9	55,0	162	406	392	390	378
10	16,0	363	584	501	396	380
11	32,0	173	351	221	726	661
12	81,0	735	142	470	377	215
13	85,0	352	485	116	802	364
14	95,0	780	176	558	123	407
15	94,0	821	515	592	712	352
16	100,0	787	821	342	361	489
17	32,0	540	842	166	791	258
18	83,0	738	430	680	541	119
19	17,0	178	272	279	711	181
20	93,0	491	848	574	695	328
21	77,0	690	129	616	315	686
22	80,0	716	803	193	111	450
23	61,0	287	782	676	676	660
24	37,0	205	384	683	805	673

25	34,0	846	687	529	478	205
26	68,0	134	134	584	263	539
27	40,0	144	195	578	491	771
28	75,0	276	376	265	112	408
29	12,0	378	227	431	630	831
30	10,0	317	518	213	658	318
31	81,0	216	395	378	679	560
32	72,0	385	169	712	776	564
33	51,0	632	712	216	726	576
34	55,0	297	489	462	654	401
35	59,0	602	185	320	470	763
36	43,0	574	403	769	767	469
37	88,0	623	353	156	469	164
38	85,0	578	213	848	175	182
39	33,0	490	632	483	638	384
40	70,0	207	205	771	619	215
41	12,0	112	843	788	642	723
42	81,0	342	383	849	480	708
43	33,0	505	406	180	415	211
44	69,0	530	530	221	167	656
45	47,0	119	759	558	284	364
46	77,0	564	566	535	804	441
47	102,0	382	758	800	826	360
48	101,0	432	548	569	786	743
49	74,0	177	312	774	168	459
50	56,0	560	622	834	160	580

Таблица 8.21

Исходные данные для анализа, вариант 5

	x_1	x_2	x_3	x_4	x_5	x_6
1	39,0	425	208	586	119	359
2	18,0	693	765	638	815	528

3	100,0	412	127	336	646	516
4	15,0	617	668	472	844	852
5	98,0	405	250	325	124	470
6	55,0	524	673	241	630	478
7	35,0	625	742	802	408	301
8	90,0	854	174	849	790	383
9	90,0	180	644	725	497	168
10	91,0	538	222	211	332	445
11	35,0	408	683	111	370	364
12	61,0	711	483	536	537	467
13	83,0	399	487	791	156	742
14	92,0	373	845	409	546	290
15	90,0	288	128	416	589	343
16	21,0	137	722	369	633	167
17	61,0	334	814	440	789	429
18	35,0	636	149	441	841	397
19	90,0	371	838	477	320	429
20	47,0	817	360	366	814	698
21	51,0	496	295	635	470	140
22	88,0	693	429	478	384	571
23	42,0	547	703	416	157	697
24	37,0	825	597	307	202	218
25	10,0	550	465	597	692	205
26	21,0	633	315	620	151	794
27	74,0	466	336	591	379	189
28	62,0	685	654	472	811	383
29	26,0	674	487	835	506	123
30	69,0	172	543	786	473	172
31	96,0	686	240	754	393	432
32	48,0	310	241	150	567	454
33	75,0	331	280	194	498	421
34	11,0	525	620	110	121	637
35	80,0	141	518	373	476	743
36	92,0	795	402	395	343	293
37	70,0	113	285	329	793	686
38	89,0	458	479	662	624	300
39	47,0	731	656	123	503	594
40	79,0	843	599	336	787	246

41	19,0	759	651	351	523	405
42	63,0	606	529	699	750	715
43	57,0	482	533	353	230	256
44	99,0	836	148	266	398	457
45	53,0	408	261	336	781	284
46	73,0	647	351	783	351	195
47	71,0	850	152	829	683	530
48	47,0	134	473	275	485	423
49	17,0	458	152	574	408	144
50	93,0	301	265	248	498	145

Занятие 7. Нейросети

Спрогнозировать значение показателя на перспективу, используя данные, приведенные в таблицах по вариантам.

Таблица 8.22

Исходные данные для анализа, вариант 1

Месяц	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	116	122	144	105	129	177	122	139	144	147	130	155	107
2	156	113	120	141	121	119	152	176	112	144	116	106	103
3	112	163	137	134	134	102	160	108	163	174	140	118	143
4	100	127	122	146	170	109	147	169	102	100	103	116	124
5	144	165	165	130	141	135	139	115	134	177	111	117	153
6	111	107	117	155	161	113	123	113	170	126	180	176	145
7	147	128	135	137	146	101	146	167	146	172	135	102	137
8	108	117	138	167	153	145	131	157	114	153	130	129	134
9	151	100	176	166	147	147	107	180	171	124	170	178	115
10	121	143	167	143	170	100	155	166	147	148	178	116	169
11	124	131	155	125	166	132	111	105	126	138	121	134	100
12	155	133	160	109	161	172	120	163	106	149	153	165	161

Таблица 8.23

Исходные данные для анализа, вариант 2

Месяц	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	127	168	167	138	140	122	177	100	114	137	101	167	124
2	141	179	115	168	113	114	158	115	168	153	113	109	137
3	130	161	146	164	124	147	124	127	178	134	126	151	123
4	177	179	155	152	158	102	127	102	142	115	146	148	105
5	141	172	171	137	152	159	149	112	166	100	126	160	171
6	174	177	116	142	109	139	155	102	170	173	177	108	103
7	166	151	142	149	177	162	123	168	177	116	146	156	131
8	103	136	137	178	163	108	166	121	137	145	107	159	101
9	148	136	140	108	107	155	148	156	118	117	116	150	168
10	152	110	140	151	179	149	177	114	112	111	114	118	167
11	107	153	162	164	162	149	177	112	122	130	155	117	122
12	148	164	163	102	120	117	132	147	180	168	169	169	107

Таблица 8.24

Исходные данные для анализа, вариант 3

Месяц	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	153	170	125	104	127	162	174	155	100	152	158	146	132
2	126	135	179	125	166	115	112	167	111	147	113	136	177
3	128	160	170	159	105	115	141	163	147	163	102	178	156
4	177	120	172	143	117	130	156	150	140	154	153	107	109
5	143	115	148	100	104	174	145	139	130	129	177	116	168
6	140	144	162	129	156	102	163	152	110	102	175	151	127
7	127	168	153	105	109	149	153	163	117	106	157	112	110
8	140	112	110	149	128	168	137	153	174	102	154	177	124
9	135	169	145	160	165	161	112	155	116	123	127	178	130
10	174	162	147	148	121	135	151	147	120	172	180	117	121
11	135	131	129	122	138	180	151	107	169	173	103	106	142
12	101	168	178	118	109	124	163	166	159	100	115	161	174

Таблица 8.25

Исходные данные для анализа, вариант 4

Месяц	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	104	129	105	142	138	175	111	142	116	109	142	140	117
2	113	169	105	160	120	155	101	135	174	144	157	176	118
3	133	111	106	117	164	152	122	146	125	147	120	156	105
4	157	131	120	160	162	164	117	108	134	141	114	171	150
5	150	165	116	118	124	150	136	107	140	174	173	104	104
6	152	159	111	169	158	174	136	116	164	130	127	148	109
7	106	134	150	146	148	119	179	121	148	117	105	103	137
8	168	134	125	141	107	124	134	133	146	150	136	113	127
9	135	178	116	127	169	151	100	105	115	169	148	144	148
10	127	103	153	148	143	130	107	118	130	123	169	157	175
11	170	157	170	163	129	167	122	102	138	160	166	138	118
12	171	178	110	179	139	159	117	138	101	157	102	129	127

Таблица 8.26

Исходные данные для анализа, вариант 5

Месяц	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	127	172	169	106	161	117	120	139	160	121	125	102	148
2	171	132	151	140	125	125	123	179	117	119	140	162	116
3	178	174	110	155	139	103	156	147	173	153	102	162	159
4	156	174	138	158	166	123	176	131	103	111	170	119	103
5	141	134	101	102	176	156	168	171	110	135	132	151	148
6	119	152	139	124	126	113	103	129	165	123	110	128	137
7	175	159	107	102	102	154	130	111	139	126	169	128	165
8	138	146	173	161	171	163	115	168	173	121	118	110	160
9	112	161	134	171	132	176	143	166	175	116	152	100	160
10	173	147	103	160	156	117	171	109	168	154	140	106	135
11	116	125	180	180	159	127	115	100	173	131	151	123	148
12	164	169	169	159	101	166	121	118	132	165	166	111	176

ПРИМЕРНЫЙ СОСТАВ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ

1. Перечень компетенций и планируемые результаты обучения по дисциплине

Код формируемых компетенций	Уровень освоения компетенции	Планируемые результаты обучения по дисциплине характеризующие этапы формирования компетенций (показатели освоения компетенции)
1	*2*	*3*
ОПК -1 - способность применять математический инструментарий для решения экономических задач	*Частичное освоение компетенции*	Знать: - способность применять математический инструментарий для решения экономических задач; Уметь: применять математический инструментарий для решения экономических задач; Владеть: - способностью применять математический инструментарий для решения экономических задач.
ПК-1 - способность подготавливать исходные данные, необходимые для расчета экономических и социально-экономических показателей, характеризующих деятельность хозяйствующих субъектов	*Частичное освоение компетенции*	Знать: методы подготовки исходных данных, необходимых для расчета экономических и социально-экономических показателей, характеризующих деятельность хозяйствующих субъектов; Уметь: подготавливать исходные данные, необходимые для расчета экономических и социально-экономических показателей, характеризующих деятельность хозяйствующих субъектов; Владеть: способностью подготавливать исходные данные, необходимые для расчета экономических и социально-экономических показателей, характеризующих деятельность хозяйствующих субъектов.
ПК-2 - способностью обосновывать выбор методик расчета экономических показателей	*Частичное освоение компетенции*	Знать: методики расчета отдельных экономических показателей предприятия; Уметь: обосновывать выбор методик расчета экономических показателей; Владеть: методиками расчета отдельных экономических показателей предприятия.

2. Оценочные материалы для проведения текущего контроля успеваемости по дисциплине

Задания к рейтинг-контролю №1

1. Дайте определение понятий «метод» и «модель». Приведите примеры использования моделей в процессе применения метода.
2. Дайте определение понятий «прогноз», «прогнозирование», «методы прогнозирования». Что является предметом процесса прогнозирования. Назови те основные группы методов прогнозирования.

3. Раскройте понятия «экстраполяция» и «интерполяция». Назовите методы прогнозирования, в которых используются экстраполяционные приемы.
4. По данным, представленным в таблице, изучается зависимость объема валового национального продукта X от переменных X_1 - потребление и X_2 - инвестиции (все данные в млрд долл.).

Исходные данные

Параметр	1	2	3	4	5	6	7	8	9	10
Y	8	9.5	11	12	13	14	15	16.5	17	18
X_1	1.65	1.8	2.0	2.1	2.2	2.4	2.65	2,85	3.2	3.55
X_2	14	16	18	20	23	23.5	25	26.5	28,5	30.5

а) Для заданного набора данных постройте линейную модель множественной регрессии. Оцените точность и адекватность построенного уравнения регрессии. Дайте экономическую интерпретацию параметров модели. Определите коэффициент эластичности.
б) Проверьте полученную модель на наличие автокорреляции остатков с помощью теста Дарбина - Уотсона.
в) Получите прогнозные значения валового национального продукта при сохранении средних темпов роста потребления и инвестиций.
г) Используя статистические данные, проведите аналогичные расчеты для РФ за период 1997-2007 гг.

2. По статистическим данным изучается зависимость оборота розничной торговли Y от ряда факторов. В таблице представлены следующие данные: Y - оборот розничной торговли, млрд руб.; X_1 - денежные доходы населения, млрд руб.; X_2 - денежные расходы на покупку товаров и услуг, млрд руб.; X_3 - численность безработных, тыс. человек.

Исходные данные

Год	Y	X_1	X_2	X_3
1	5100,3	7100	5175	3888,6
2	51 200	91 095	64170	6684,3
3	2352	3983,9	3009,4	7059,1
4	3070	6831	5001,8	6287,9
5	3765	8900,5	6147,3	6154,7
6	4529	10 976,3	7670,7	5683,3

7	5642	13 522,5	9615,3	5775,2
8	7038	13 862	9923	5208,3
9	7465	14 675,3	10 781,3	5222,5
10	8793	15 325,7	11 562,8	2156,7

а) Для заданного набора данных постройте линейную модель множественной регрессии. Оцените точность и адекватность построенного уравнения регрессии.
б) Дайте экономическую интерпретацию параметров модели. Рассчитайте коэффициенты эластичности. Определить стандартизированные коэффициенты регрессионной модели.
в) Получите прогнозные значения результативного показателя в зависимости от средних темпов прироста факторных показателей.
г) Проведите аналогичный расчет на материалах одного из регионов РФ.

Задания к рейтинг-контролю №2

1. Рассчитайте точечный и интервальный прогнозы по следующим значениям временного ряда.

Временной ряд

Год	1	2	3	4	5	6	7	8
Y	16.7	15.3	20,2	17.1	15.3	14.4	13,5	12,1

2. Рассчитайте поданным таблица параметры уравнения тренда и показатели адекватности функции реальным условиям. Является ли исследуемая тенденция устойчивой?

Временной ряд спроса на продукт за один год

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
Спрос, усл. ед.	90	111	99	89	87	84	104	102	95	114	103	113

3. Проверьте гипотезу о наличии автокорреляции в остатках методом Дарбина - Уотсона для аддитивной модели временного ряда, представленного в таблице

Временной рядспроса на продукт за два года

Месяц	1	2	3	4	5	6	7	8	9	10	11	12

Спрос, тыс. шт.	46	56	54	43	57	56	67	62	50	56	47	56
Месяц Показатель	13	14	15	16	17	18	19	20	21	22	23	24
Спрос, тыс. шт.	54	42	64	60	70	66	57	55	52	62	70	72

4. Имеются условные данные об объеме потребления электроэнергии (y_t) жителями региона за 16 кварталов. Проверьте ряд на стационарность. Постройте автокорреляционную функцию и сделайте вывод о наличии сезонных колебаний. Постройте авторегрессионную модель временного ряда. Сделайте прогноз на два квартала вперед.

Объем потребления электроэнергии жителями региона

Год	1	2	3	4	5	6	7	8
W	5,8	4,5	5,1	9.1	7	5	6	10,1
Год	9	10	11	12	13	14	15	16
W	7,9	5,5	6.3	10.8	9	6.5	7	11,1

5. Используя данные таблицы, рассчитайте прогноз инвестиций в основной капитал предприятий региона на 2010 г. методом экспоненциального сглаживания и методом гармонических весов.

Объем инвестиций в основной капитал предприятий региона

Год	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Объем инвестиций. млн руб.	93.6	177.0	303.0	512.0	683.0	736.0	712,0	781,0	988.0	1220	1381

6. Фирма рассматривает инвестиционный проект по производству продукта «Б». В процессе предварительного анализа экспертами были выявлены три ключевых параметра проекта и определены возможные границы их изменений Прочие параметры проекта считаются постоянными величинами. Проведите процедуру имитационного моделирования и определите величину ожидаемого показателя чистой приведенной стоимости проекта. Проведите вероятностный анализ полученных данных.

Вероятностные сценарии реализации проекта (условных единиц)

Сценарий Показатель	Наихудший (P = 0.3)	Наилучший (P = 0.2)	Вероятный (P = 0.5)
Объем выпуска, Q	240	550	380
Цена за штуку, P	42	60	50
Переменные затраты, V	39	35	37

Неизменяемые параметры проекта по производству продукта «Б» (условныхединиц)

Показатель	Наиболее вероятное значение	Показатели	Наиболее вероятное значение
Постоянные затраты	720	Норма дисконта	18
Амортизация	200	Срок проекта, п (лет)	5
Ставка налога на прибыль	20	Начальные инвестиции	3800

Задания к рейтинг-контролю №3

1. Проводится экспертная оценка по прогнозированию валового сбора зерновых культур в Курской области. В таблице приведены результаты третьего тура опроса по методу Дельфи. На основе статистической характеристики ответов экспертов сделайте вывод о степени согласованности мнений и возможности завершения экспертизы. В случае положительного результата приведите прогнозное значение валового сбора зерновых культур.

Ранжированный ряд экспертных оценок

Эксперт	1	2	3	4	5	6	7	8	9	10	11
Валовой сбор зерна, тыс. т	900	950	1000	1100	1200	1200	1300	1400	1500	1900	2000

2. Определите прогнозное значение производства продукции на восьмой год методом Дельфи и методом экстраполяции тренда по данным таблицы

Динамика производства продукции

Год	1	2	3	4	5	6	7
Производство продукции, млн руб.	61,5	61,3	62,4	65,5	64,8	64,3	64,7

Осуществите синтез результатов, полученных двумя методами прогнозирования. Вес прогноза по методу Дельфи $p_1 = 0,7$, а вес прогноза по методу экстраполяции тренда $p_2 = 0,3$.

3. Изучите хозяйственную деятельность предприятия вашего региона. Составьте морфологическую таблицу проблем предприятия в условиях мирового финансового кризиса. В строках таблицы можно указать функциональную принадлежность проблем (например, проблемы маркетинговой, производственной, сбытовой, инвестиционной и иной деятельности). По каждой проблеме предложите одно или несколько решений, которые необходимо записать в морфологическую таблицу решений. Далее проанализируйте решения с точки зрения известных, вновь полученных и интересных решений.

Рейтинг-контроль 1	Развернутый ответ на вопрос, тест	10 баллов
Рейтинг-контроль 2	Развернутый ответ на вопрос, тест	10 баллов
Рейтинг-контроль 3	Развернутый ответ на вопрос, тест	15 баллов
Выполнение семестрового плана самостоятельной работы	Выполнение заданий для самостоятельной работы	15 баллов
Посещение занятий студентом		5 баллов
Дополнительные баллы (бонусы)	Выполнение практических заданий, подготовка докладов	5 баллов
Всего по дисциплине		60 баллов

Критерии оценки ответов на вопросы рейтинг-контроля

(шкалы – 10/15 баллов)

Баллы рейтинго-вой оценки	Критерии оценки
9-10/13-15 баллов	Ответ на вопрос представлен комплексно и полно, представлены необходимые определения и понятия, все положения теоретического и практического характера обоснованы, приведены ссылки на текущие стандарты и положения (при необходимости)
7-8/10-12 балла	Ответ на вопрос представлен комплексно и полно, представлены необходимые определения и понятия, все положения теоретического и практического характера недостаточно обоснованы, не приведены ссылки на текущие стандарты и положения (при необходимости)
5-6/7-9 балла	Ответ на вопрос представлен некомплексно и неполно, не представлены необходимые определения и понятия, все положения теоретического и практического характера недостаточно обоснованы, не приведены ссылки на текущие стандарты и положения (при необходимости)
3-4/4-6 балла	Ответ на вопрос представлен частично и неполно, не представлены необходимые определения и понятия, все положения теоретического и практического характера недостаточно обоснованы, не приведены ссылки на текущие стандарты и положения (при необходимости)
1-2/1-3 балла	Ответ отсутствует

Критерии оценки посещения занятий

Баллы рейтинговой оценки	Критерии оценки
5	Студент посетил все занятия
4	Студент по уважительной причине пропустил 4 часа аудиторных занятий
3	Студент по уважительной причине пропустил 6 часов аудиторных занятий
1-2	Студент по уважительной причине пропустил 10 часов аудиторных занятий

САМОСТОЯТЕЛЬНАЯ РАБОТА ОБУЧАЮЩЕГОСЯ

Тема 1. Генезис экономического прогнозирования и планирования:

1. Изучение рекомендованной литературы к практическому занятию.
2. Подготовка доклада/презентации к практическому занятию

Тема 2. Методология прогнозирования и планирования экономических объектов:

1. Изучение рекомендованной литературы к практическому занятию.
2. Подготовка доклада/презентации к практическому занятию

Тема 3. Регрессионные модели прогнозирования:

1. Изучение рекомендованной литературы к практическому занятию.
2. Подготовка доклада/презентации к практическому занятию

Тема 4. Методы прогнозирования экономической динамики:

1. Изучение рекомендованной литературы к практическому занятию.
2. Подготовка доклада/презентации к практическому занятию

Тема 5. Имитационное моделирование в прогнозирование:

1. Изучение рекомендованной литературы к практическому занятию.
2. Подготовка доклада/презентации к практическому занятию

Тема 6. Интуитивные и эвристические методы прогнозирования:

- методы экспертных оценок в прогнозировании;

1. Изучение рекомендованной литературы к практическому занятию.

Подготовка доклада/презентации к практическому занятию

Критерии оценки доклада

Оценка	Критерии оценивания
«отлично»	- Тема раскрыта исчерпывающе, автор продемонстрировал глубокие знания. - Цель сформулирована, четко обоснована, дан подробный план ее достижения. - Доклад отличается творческим подходом, собственным оригинальным отношением автора к теме. - Печатный вариант доклада полностью соответствует предъявляемым требованиям. Отличается четкой структурой и грамотным оформлением. - Качественно оформлена презентация и автору удалось вызвать интерес аудитории и уложиться в регламент.
«хорошо»	- Тема доклада раскрыта, автор показал хорошее знание тематики исследования. - Цель сформулирована, обоснована, дан схематичный план ее достижения. - Работа над докладом была самостоятельная, демонстрирующая серьезную заинтересованность автора, была предпринята попытка представить личный взгляд, применены элементы творчества. - Печатный вариант доклада не полностью соответствует предъявляемым требованиям. Предприняты попытки оформить работу, придать ей соответствующую структуру. - В наличии презентация и автору удалось вызвать интерес аудитории, но он вышел за рамки регламента.
«удовлетворительно»	- Тема доклада раскрыта фрагментарно. - Цель сформулирована, но план ее достижения отсутствует. - Автор проявил незначительный интерес к теме доклада, но не продемонстрировал самостоятельности в работе над докладом, не использовал возможности творческого подхода. - Печатный вариант доклада не соответствует предъявляемым требованиям. Отсутствуют порядок и четкая структура работы. Есть ошибки в оформлении. - Материал изложен с учетом регламента, однако автору не удалось заинтересовать аудиторию.

«неудовлетворительно»	- Тема доклада не раскрыта и не исследована. - Цель не сформулирована. - Доклад шаблонный, показывающий формальное отношение автора. - Доклад в печатном варианте отсутствует. - Презентация не проведена.

3. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

Перечень вопросов к экзамену

1. Прогноз: понятие, классификация.
2. Прогнозирование: понятие, генезис понятия.
3. Индикативный план.
4. Метод прогнозирования.
5. Фактографические методы прогнозирования.
6. Экспертные методы прогнозирования.
7. Комбинированные методы прогнозирования.
8. Этапы прогнозирования и моделирования.
9. Традиционные модели планирования.
10. Модель программно-целевого планирования.
11. Модель интегрированного планирования.
12. Метод Паттерн.
13. Основные этапы системного планирования.
14. Модель Тевеса.
15. Модель Калдора.
16. Понятие регрессии.
17. Линейная регрессия.
18. Метод наименьших квадратов.
19. Коэффициент детерминации.
20. Нелинейные регрессионные модели.
21. Элементы временного ряда.
22. Коэффициент корреляции рангов Спирмена.
23. Модели стационарных временных рядов.
24. Модели нестационарных временных рядов.
25. Адаптивные методы прогнозирования.
26. Метод экспоненциального сглаживания.
27. Метод гармонических весов.
28. Метод экспертных оценок.

29. Метод Дельфи.
30. Метод «мозгового штурма».
31. Эвристические методы прогнозирования.
32. Метод интервью.
33. Метод составления сценариев.
34. Морфологический анализ.

Критерии оценивания:

В экзаменационный билет включен два теоретических вопроса, соответствующий содержанию формируемой компетенции и задача.

Экзамен проводится в устной форме. На ответ студенту отводится 20 минут. За ответ на теоретические вопросы студент может получить максимально 40 баллов.

Перевод баллов в оценку:

Если общая сумма баллов, набранных обучающимся в течении работы за семестр, а также в ходе проведения итогового испытания по дисциплине, превысила 91 балл, это соответствует оценке «отлично».

Если общая сумма баллов, набранных обучающимся в течении работы за семестр, а также в ходе проведения итогового испытания по дисциплине, превысила превышает 74, но меньше или равна 90 баллов, это соответствует оценке «хорошо».

Если общая сумма баллов, набранных обучающимся в течении работы за семестр, а также в ходе проведения итогового испытания по дисциплине, превысила превышает 61, но меньше или равна 73 баллам, это соответствует оценке «удовлетворительно».

Если общая сумма баллов, набранных обучающимся в течении работы за семестр, а также в ходе проведения итогового испытания по дисциплине, превысила не превышает 60 баллов, это соответствует оценке «неудовлетворительно».

Критерии оценки компетенций по осваиваемой дисциплине при проведении промежуточной аттестации

Баллы*	Оценка	Требования к знаниям
91-100	*«отлично» / «зачтено»*	Оценка «отлично» выставляется студенту, если он глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически стройно его излагает, умеет тесно увязывать теорию с практикой, свободно

Баллы*	Оценка	Требования к знаниям
		справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при видоизменении заданий, использует в ответе материал монографической литературы, правильно обосновывает принятое решение. Учебные достижения в семестровый период и результаты текущего контроля демонстрируют высокую степень овладения программным материалом.
74-90	*«хорошо» / «зачтено»*	Оценка «хорошо» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, владеет необходимыми навыками и приемами их выполнения. Учебные достижения в семестровый период и результаты текущего контроля демонстрируют хорошую степень овладения программным материалом.
61-73	*«удовлетворительно» / «зачтено»*	Оценка «удовлетворительно» выставляется студенту, если он знает только основной материал, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, нарушения логической последовательности в изложении программного материала, испытывает затруднения при выполнении практических работ. Учебные достижения в семестровый период и результаты текущего контроля

Баллы*	Оценка	Требования к знаниям
		демонстрируют достаточную степень овладения программным материалом.
0-60	*«неудовлетворительно» / «не зачтено»*	Оценка «неудовлетворительно» выставляется студенту, который не знает значительной части программного материала, допускает существенные ошибки, неуверенно, с большими затруднениями выполняет практические работы. Оценка «неудовлетворительно» ставится студентам, которые не могут продолжить обучение без дополнительных занятий по соответствующей дисциплине. Учебные достижения в семестровый период и результаты текущего контроля демонстрируют не высокую степень овладения программным материалом.

1. ИТОГОВЫЕ ТЕСТОВЫЕ ЗАДАНИЯ ПО ДИСЦИПЛИНЕ

№ п/п	Тестовые задания	Код контролируемой компетенции
1.	Выявление назревающих проблем путем логического продолжения в будущее тенденций, закономерности которых в прошлом и настоящем достаточно хорошо известны – это … a) Генетический прогноз; b) Телеологический прогноз; c) Смешанный прогноз.	ОПК-1

2.	Модели структурно-балансового типа, где наряду с разбиением агрегата на составляющие его элементы также рассматривается связь между этими элементами: a) Структурные модели; b) Трендовые модели; c) Смешанные модели.	ОПК-1
3.	**Формализованное описание исследуемого экономического процесса или объекта в виде математических зависимостей и отношений, т.е. формул – это …** a) Экономико-математическая модель; b) Видение; c) Формализация.	ПК-2
4.	**Назовите два любых принципа прогнозирования.**	ПК-2
5.	**Частный случай стохастической связи, при этом среднее изменение факторного признака приводит к изменению среднего значения результативного признака – это …** a) Функциональная связь; b) Корреляционная связь; c) Множественная связь.	ПК-2
6.	**Уравнение регрессии всегда дополняется *показателями тесноты связи.* При использовании линейной регрессии в качестве такого показателя выступает …** a) Линейный коэффициент корреляции; **b)** Коэффициент вариации; c) Коэффициент детерминации.	ПК-2
7.	Сформулируйте понятие тренда.	ПК-1
8.	Сформулируйте понятие тенденции.	ПК-1

9.	Как определяется медиана ряда динамики с нечетным числом числа членов ряда при выявлении тенденции методом, основанным на медиане выборки	ПК-1
10.	При использовании методики построения регрессионной модели при расчете коэффициента линейной корреляции, какие он может принимать значения?	ПК-2
11.	Какими методами в парной регрессии может проводится выбор вида математической функции?	ПК-2
12.	Охарактеризуйте зависимость коэффициента регрессии и коэффициента корреляции.	ОПК-1
13.	Сформулируйте условие наличия корреляционной связи.	ОПК-1
14.	Рассчитайте долю дисперсии, приходящейся на факторы, не обусловленные регрессией, если в результате расчета был получен коэффициент детерминации 0,86.	ПК-1
15.	Охарактеризуйте зависимость доли объясненной вариации и качества линейной модели	ПК-1
16.	Назовите интервал допустимых значений для показателя ошибки аппроксимации как показателя качества регрессионной модели	ПК-1
17.	Сформулируйте понятие имитационной модель.	ПК-2
18.	Сформулируйте два любых основных принципа имитационного моделирования, используемых при подготовке данных.	ПК-2
19.	Дайте понятие экзогенной переменной с точки зрения имитационного моделирования.	ПК-2

20.	Дайте понятие эндогенной переменной с точки зрения имитационного моделирования	ПК-2

КЛЮЧИ К ТЕСТУ

№ задания	Ответ
1.	a
2.	a
3.	a
4.	Выделяют следующие принципы прогнозирования и планирования: альтернативности, своевременности, системности, комплексности, непрерывности, адекватности и обоснованности, целенаправленности и приоритетности, социальной ориентации, оптимальности, сбалансированности и пропорциональности, сочетания отраслевого и регионального аспектов прогнозирования и планирования, информативности.
5.	b
6.	a
7.	Тренд отражает тенденцию, под которой в статистической литературе понимают основное направление развития, выраженное более или менее гладкой траекторией
8.	***Тенденция*** - это некоторая закономерность, присущая анализируемому объекту, которая прослеживается с момента начала анализа, на всем протяжении и в подводимых на его основании итогах, и, исходя из наблюдаемого, наиболее вероятно будет иметь в дальнейшем прогнозируемое развитие и предугадываемые вариации.
9.	Медиана ряда динамики с нечетным числом числа членов ряда при выявлении тенденции методом, основанным на медиане выборки определяется как серединное значение ранжированного ряда
10.	Коэффициент линейной корреляции может принимать значения от -1 до 1
11.	В парной регрессии выбор вида математической функции может быть осуществлен тремя методами: графическим; аналитическим, т. е. исходя из теории изучаемой взаимосвязи; экспериментальным.
12.	Если коэффициент регрессии $b > 0$, то коэффициент корреляции $0 < R < 1$, и, наоборот, при $b<0$, $-1 < R < 0$.
13.	Если изменение факторного признака или комбинации признаков приводит к вариации значений результативного, то говорят о наличии стохастической или корреляционной (частного случая) связи.
14.	0,14
15.	Чем больше доля объясненной вариации, тем соответственно меньше роль прочих факторов, и, следовательно, линейная модель хорошо аппроксимирует исходные данные и ею можно воспользоваться для прогноза значений результативного признака.

16.	Ошибка аппроксимации в пределах 5-7 % свидетельствует о хорошем подборе модели к исходным данным. Большинство авторов рекомендуют считать модель регрессии адекватной, если средняя относительная ошибка аппроксимации не превышает 12%.
17.	Имитационная модель представляет собой развернутую схему с детально описанной структурой и поведением изучаемого объекта.
18.	Основные принципы имитационного моделирования, используемые при подготовке данных, включают: принцип информационной достаточности; принцип целесообразности; принцип осуществимости, принцип множественности моделей, принцип агрегирования, принцип параметризации.
19.	Экзогенные (независимые) переменные, называются также входными — переменные, которые порождаются вне системы или являются результатом воздействия внешних причин.
20.	Эндогенные (зависимые) – это переменные, которые возникают в системе или в результате воздействия внутренних причин, часто являются выходными.

Критерии оценки итогового теста

(1 вопрос – 1 балл)

Оценка в баллах	**Оценка за итоговый тест**
18-20	«Отлично»
15-17	«Хорошо»
12-14	«Удовлетворительно»
Менее 12 баллов	«Неудовлетворительно»

Методические рекомендации по выполнению курсовых работ (проектов)

1. ОБЩИЕ ПОЛОЖЕНИЯ

Курсовая работа (проект) представляет собой вид учебной и научно- исследовательской работы студента, является индивидуальным, завершенным трудом, отражающим знания, навыки и умения студента, полученные в ходе освоения дисциплины.

Тема курсовой работы (проекта) не может носить описательного характера, в ее формулировку должна быть заложена исследовательская проблема. Курсовая работа (проект) подготавливает студента к выполнению более сложной задачи – написанию выпускной квалификационной или дипломной работ.

Рациональнее темы курсовых работ (проектов), выполняемых студентами за весь период обучения, подбирать таким образом, чтобы они вместе с выпускной работой составляли единую систему последовательно усложняемых и взаимосвязанных работ. При защите работы студент учится не только правильно излагать свои мысли, но и аргументированно отстаивать, защищать выдвигаемые выводы и решения. Формулировка темы должна быть по возможности краткой и соответствовать содержанию работы.

2. ЦЕЛИ И ЗАДАЧИ ВЫПОЛНЕНИЯ КУРСОВЫХ РАБОТ (ПРОЕКТОВ)

Основной целью выполнения курсовой работы (проекта) является развитие мышления, творческих способностей студента, привитие навыков самостоятельной работы, связанной с поиском, систематизацией и обобщением научной и учебной литературы, углублённым изучением определенного вопроса, темы, раздела учебной дисциплины, формирование умений анализировать и критически оценивать исследуемый научный и практический материал, овладение методами современных научных исследований.

Курсовая работа (проект) представляет собой:

- изложение результатов исследования с учетом вопросов теории и практики в проделах выбранной темы;

- авторский труд, самостоятельное творчество студента, формирование его личной позиции и практического подхода к выбранной теме;

- отражение умения студентом логично, аргументировано, ясно, последовательно и кратко излагать свои мысли.

Основные отличия курсовой работы (проекта) от контрольной работы:

- курсовая работа требует более глубокого анализа проблемы, поэтому её минимальный требуемый объем значительно больше,

- обязательно включает практический раздел, направленный на отработку факто- логического материала, в курсовой работе должно найти отражение взаимосвязи теоретических положений с практикой;

- контроль за ходом написания курсовой работы осуществляется кафедрой.

Научно-консультационную и методическую помощь студенту оказывает научный руководитель. Работа над избранной темой требует от студента знаний основ методологии исследования, творческого мышления, прилежания и профессионализма.

Задачами выполнения курсовых работ (проектов) являются:

- систематизация, закрепление, углубление и расширение приобретенных студентом знаний, умений, навыков по учебным дисциплинам профессиональной подготовки;
- овладение методами научных исследований;
- формирование навыков решения творческих задач в ходе научного исследования, художественного творчества или проектирования по определенной теме;
- овладение современными методами поиска, обработки и использования информации.
- подготовка к написанию дипломной работы (материалы курсовых работ могут входить в дипломную работу).

При выполнении курсовых работ (проектов) студент должен продемонстрировать способности:

- выдвинуть научную (рабочую) гипотезу;
- собрать и обработать информацию по теме;
- изучить и критически проанализировать полученные материалы;
- систематизировать и обобщить имеющуюся информацию;

– самостоятельно решить поставленные творческие задачи;
– логически обосновать и сформулировать выводы, предложения и рекомендации.

Особенности курсовых работ (проектов) в зависимости от курса обучения проявляются в постепенном усложнении объектов и методов исследования (проектирования).

Количество курсовых работ (проектов), наименование дисциплин, по которым они предусматриваются, определяется учебным планом. Общее число курсовых работ (проектов) по дисциплинам учебного плана не может превышать 5-6 на весь период обучения, если иное не предусматривается государственным образовательным стандартом и примерным учебным планом по соответствующей специальности (направлению). Курсовая работа (проект) рассматривается как вид учебной работы по дисциплине и выполняется в пределах часов, отводимых на ее изучение. Курсовые работы (проекты) рассматриваются как форма отчетности.

Полные названия курсовых работ (проектов) вносятся в экзаменационные ведомости, зачетные книжки студентов и в приложения к дипломам.

Согласно номенклатуре дел курсовые работы (проекты) учитываются и хранятся на кафедре в течение пяти лет. По истечении указанного срока все курсовые работы (проекты), не представляющие учебно-методической ценности, списываются по акту и уничтожаются.

3. СТРУКТУРА И ЭТАПЫ ВЫПОЛНЕНИЯ КР (КП)

Тематику курсовых работ (проектов) разрабатывает кафедра в учебном году, предшествующем выполнению курсовой работы (проекта).

Выбор и утверждение темы курсовой работы (проекта) происходит в следующем порядке:

- тематика курсовых работ (проектов) сообщается студентам;
- студент может выбрать тему курсовой работы (проекта) из числа тем, предложенных кафедрой;
- студент может также самостоятельно предложить тему курсовой работы (проекта) с обоснованием ее целесообразности;

- тематика курсовых работ (проектов) на предстоящий учебный год утверждается на заседании кафедры, о чем в протоколе заседания делается соответствующая запись.

Студент выполняет курсовую работу (проект) по утвержденной теме под руководством преподавателя, являющегося его научным руководителем.

Темы курсовых работ (проектов) утверждаются на заседании кафедры и подтверждаются соответствующими заявлениями студентов о выборе темы.

Руководителем курсовой работы (проекта) по дисциплине учебного плана является, как правило, лектор, ведущий данную дисциплину, преподаватель, ведущий практические занятия. Руководителем курсовой работы (проекта) по специальным дисциплинам, дисциплинам специализации может быть назначен приглашенный специалист, выполняющий учебную нагрузку на условиях почасовой оплаты на условиях почасовой оплаты.

Научный руководитель составляет задание на курсовую работу (проект), осуществляет ее текущее руководство. Текущее руководство курсовой работой (проектом) включает систематические консультации с целью оказания организационной и научно-методической помощи студенту, контроль за осуществлением выполнения работы в соответствии с планом – графиком, проверку содержания и оформления завершенной работы.

Завершенная курсовая работа (проект) передается студентом на кафедру за неделю до защиты для ее анализа.

Написание работы - процесс, включающий в себя ряд взаимосвязанных этапов:

1. Выбор темы. Рекомендованная тематика курсовых работ содержится в рабочих программах дисциплин, по которым формой промежуточной аттестации является курсовая работа (проект). При выборе темы курсовой работы (проекта) можно рекомендовать студенту четко определить круг своих интересов и выполнять весь комплекс курсовых работ (в рамках соответствующих учебных дисциплин) по одной проблематике. Это позволит существенно повысить качество выполняемых курсовых работ (проектов) и даст возможность студенту лучше подготовиться к выполнению выпускной квалификационной работы.

2. Разработка структуры и оформление содержания. Структура работы должна быть согласована с научным руководителем.

3. Сбор, анализ и обобщение материалов исследования, написание текста работы:

- сбор материалов, необходимых для выполнения курсовой работы (проекта), посредством использования литературных источников, нормативных актов, директивных документов и документации предприятия (организации) по рассматриваемой в работе проблематике;
- систематизация и обработка собранного материала по каждому из разрабатываемых в курсовой работе (проекту) вопросу или проблеме. На базе систематизированного материала формируются основные направления анализа. Одновременно выясняется необходимость сбора дополнительной информации по отдельному вопросу или вопросам;
- сбор дополнительной информации и разработка аналитической части курсовой работы (проекта). На этом этапе выявляются негативные моменты и недостатки функционирования объекта исследования;
- разработка и обоснование предложений по основным направлениям деятельности объекта исследования. На основе разработанных предложений и рекомендаций формулируются соответствующие выводы

4. Оформление работы и её представление для проверки.

5. Защита курсовой работы. Работа предоставляется на кафедру (руководителю) заранее, не позднее, чем за 10 дней до защиты.

Методологической основой курсовой работы (проекта) являются законодательные акты Российской Федерации по экономике, в целом, и по изучаемой дисциплине, в частности, программные документы и решения правительства РФ по хозяйственным вопросам.

По выбранной теме курсовой работы (проекта) рекомендуется использовать данные Росстата, материалы Института исследования товародвижения и конъюнктуры оптового рынка (ИТКОР), учебную специальную литературу, монографии, брошюры, статьи. Целесообразно изучить зарубежный опыт применительно к рассматриваемой теме. Важным условием успешного раскрытия избранной темы является

ознакомление с материалами, опубликованными в периодических изданиях и др.

Желательно, чтобы курсовой проект выполнялся на материалах предприятия или организации по месту работы студентов заочной формы обучения или по месту прохождения производственной практики студентов очной формы обучения. В качестве основы написания курсовой работы (проекта) могут быть использованы материалы, собранные для курсовых работ по смежным дисциплинам, изученным ранее, а также материалы, собранные в ходе учебной и производственной практик

4. ФОРМЫ И ПОРЯДОК АТТЕСТАЦИИ КУРСОВЫХ РАБОТ (ПРОЕКТОВ)

Формами аттестации студента по результатам выполнения курсовой работы являются зачёт (зачтено/не зачтено), а по результатам курсового проекта дифференцированный зачёт ("отлично" - "хорошо" - "удовлетворительно" - "неудовлетворительно"). Форма аттестации по курсовым работам (проектам) по дисциплинам учебного плана вносится в рабочий учебный план специальности (направления) и утверждается Ученым советом института.

Аттестация всех курсовых работ (проектов) должна быть проведена до начала экзаменационной сессии, в сроки, указанные рабочим учебным планом специальности (направления).

Аттестация по курсовым работам (проектам) производится в виде ее защиты перед группой и научным руководителем работы (проекта).

Решение об оценке курсовой работы (проекта) принимается преподавателем по результатам трех рейтингов, проводимых в течение семестра, для которых деканатом выдается отдельная ведомость, аналогичная ведомости текущего рейтинг-контроля, а также по итогам анализа предъявленной курсовой работы (проекта), доклада студента и его ответов на вопросы. Оценка по курсовой работе (проекту) вносится в экзаменационную ведомость, зачетную книжку студента научным руководителем.

Студент, по неуважительной причине не предоставивший в установленный срок или не защитивший курсовую работу (проект), считается имеющим академическую задолженность. Научный руководитель курсовой работы (проекта) проставляет в экзаменационную ведомость

неудовлетворительную оценку. В случае наличия уважительных причин, подтвержденных документально, распоряжением по институту (факультету) студенту устанавливаются индивидуальный порядок и сроки выполнения и защиты курсовой работы (проекта). Курсовая работа, оцененная неудовлетворительно перерабатывается студентом и возвращается на проверку тому же преподавателю.

Критериями оценки курсовой работы являются*:*

- актуальность и степень разработанности темы;
- творческий подход и самостоятельность в анализе, обобщениях и выводах;
- полнота охвата первоисточников и исследовательской литературы;
- уровень овладения методикой исследования;
- научная обоснованность и аргументированность обобщений, выводов и рекомендаций;
- научный стиль изложения;
- соблюдение всех требований к оформлению курсовой работы и сроков ее исполнения.

5. ТРЕБОВАНИЯ К СОДЕРЖАНИЮ КР (КП)

Курсовая работа (проект) имеет ряд структурных элементов: введение, теоретическая часть, практическая часть, заключение.

Разработка введения. Во-первых, во введении следует обосновать актуальность избранной темы курсовой работы (проекта), раскрыть ее теоретическую и практическую значимость, сформулировать цели и задачи работы.

Во-вторых, во введении, а также в той части работы, где рассматривается теоретический аспект данной проблемы, автор должен дать, хотя бы кратко, обзор литературы, изданной по этой теме.

Введение должно подготовить читателя к восприятию основного текста работы. Оно состоит из обязательных элементов, которые необходимо правильно сформулировать. В первом предложении называется тема курсовой работы.

Актуальность исследования (почему это следует изучать?). Актуальность исследования рассматривается с позиций социальной и практической значимости. В данном пункте необходимо раскрыть суть ис-

следуемой проблемы и показать степень ее проработанности в различных трудах. Здесь же можно перечислить источники информации, используемые для исследования (Информационная база исследования может быть вынесена в первую главу).

Цель исследования (какой результат будет получен?). Цель должна заключаться в решении исследуемой проблемы путем ее анализа и практической реализации. Цель всегда направлена на объект.

Объект исследования (что будет исследоваться?). Объект предполагает работу с понятиями. В данном пункте дается определение экономическому явлению, на которое направлена исследовательская деятельность. Объектом может быть личность, среда, процесс, структура, хозяйственная деятельность предприятия (организации).

Предмет исследования (как, через что будет идти поиск?). Здесь необходимо дать определение планируемым к исследованию конкретным свойствам объекта или способам изучения экономического явления. Предмет исследования направлен на практическую деятельность и отражается через результаты этих действий.

Задачи исследования (как идти к результату?), пути достижения цели. Задачи соотносятся с гипотезой. Определяются они исходя из целей работы. Формулировки задач необходимо делать как можно более тщательно, поскольку описание их решения должно составить содержание глав и параграфов работы. Как правило, формулируются 3-4 задачи.

Примерный перечень рекомендуемых задач:

1. «На основе теоретического анализа литературы разработать...» (ключевые понятия, основные концепции).
2. «Определить... » (выделить основные условия, факторы, причины, влияющие на объект исследования).
3. «Раскрыть... » (выделить основные условия, факторы, причины, влияющие на предмет исследования).
4. «Разработать... » (средства, условия, формы, программы).
5. «Апробировать…» (что разработали) и дать рекомендации...

Методы исследования (как исследовали?): дается краткое перечисление методов исследования через запятую без обоснования.

Структура работы – это завершающая часть введения (что в итоге в работе/проекте представлено).

В завершающей части в назывном порядке перечисляются структурные части работы (проекта), например: «Структура работы соответствует логике исследования и включает в себя введение, теоретическую часть, практическую часть, заключение, список литературы, 5 приложений».

Здесь допустимо дать развернутую структуру курсовой работы (проекта) и кратко изложить содержание глав. (Чаще содержание глав курсовой работы излагается в заключении).

Таким образом, введение должно подготовить к восприятию основного текста работы.

Краткие комментарии по формулированию элементов введения представлены в таблице 1.

Таблица 1 – Комментарии по формулированию элементов введения

Элемент введения	Комментарий к формулировке
Актуальность темы	Почему это следует изучать? Раскрыть суть исследуемой проблемы и показать степень ее проработанности.
Цель исследования	Какой результат будет получен? Должна заключаться в решении исследуемой проблемы путем ее анализа и практической реализации.
Объект исследования	Что будет исследоваться? Дать определение явлению или проблеме, на которое направлена исследовательская деятельность.
Предмет исследования	Как и через что будет идти поиск? Дать определение планируемым к исследованию конкретным свойствам объекта или способам изучения явления или проблемы.
Задачи работы	Как идти к результату? Определяются исходя из целей работы и в развитие поставленных целей. Формулировки задач необходимо делать как можно более тщательно, поскольку описание их решения должно составить содержание глав и параграфов работы. Рекомендуется сформулировать 3 – 4 задачи.
Методы исследования	Как изучали? Краткое перечисление методов через запятую без обоснования.
Элемент введения	Комментарий к формулировке

Структура работы (завершающая часть введения)	Что в итоге в работе/проекте представлено. Краткое изложение перечня и/или содержания глав работы/проекта.

Разработка основной части курсовой работы/проекта. Основная часть обычно состоит из двух-трех разделов: в первом содержатся теоретические основы темы; дается история вопроса, уровень разработанности вопроса темы в теории и практике посредством сравнительного анализа литературы.

В теоретической части рекомендуется излагать наиболее общие положения, касающиеся данной темы, а не вторгаться во все проблемы в глобальном масштабе. Теоретическая часть предполагает анализ объекта исследования и должна содержать ключевые понятия, историю вопроса, уровень разработанности проблемы в теории и практике. Излагая содержание публикаций других авторов, необходимо обязательно давать ссылки на них с указанием номеров страниц этих информационных источников.

Вторым разделом является практическая часть, которая должна носить сугубо прикладной характер. В ней необходимо описать конкретный объект исследования, привести результаты практических расчетов и направления их использования, а также сформулировать направления совершенствования, либо вынести их в отдельных – третий раздел курсовой работы (проекта).

Важно глубоко изучить наиболее существенные с точки зрения задач курсовой работы (проекта) стороны и особенности.

Разработка заключения. По окончании исследования подводятся итоги по теме. Заключение носит форму синтеза полученных в работе результатов. Его основное назначение - резюмировать содержание работы, подвести итоги проведенного исследования. В заключении излагаются полученные выводы и их соотношение с целью исследования, конкретными задачами, гипотезой, сформулированными во введении.

Проведенное исследование должно подтвердить или опровергнуть гипотезу исследования. В случае опровержения гипотезы даются рекомендации по возможному совершенствованию деятельности в свете исследуемой проблемы.

Составление списка литературы. В список источников и литературы включаются источники, изученные Вами в процессе подготовки

работы, в т.ч. те, на которые Вы ссылаетесь в тексте курсовой работы/проекта.

Список используемой литературы должен содержать не менее 20 источников (не менее 10 книг и 10-15 материалов периодической печати), с которыми работал автор курсовой работы (проекта).

Список используемой литературы включает в себя:

- нормативные правовые акты;
- научную литературу и материалы периодической печати;
- практические материалы.

6. ТРЕБОВАНИЯ К ОФОРМЛЕНИЮ КР (КП)

Курсовые работы (проекты) следует оформлять в печатном виде с использованием компьютера и принтера распечатывать на одной стороне листа белой бумаги формата А4. Рукописное оформление работы не допускается (разрешается вписывать черными чернилами отдельные слова, формулы, условные знаки, а также выполнять отдельные иллюстрации).

Вне зависимости от способа выполнения работы качество напечатанного текста и оформления иллюстраций, таблиц, распечаток с ЭВМ должно удовлетворять требованию их четкого воспроизведения. При выполнении отчета необходимо соблюдать равномерную плотность, контрастность и четкость изображения по всему отчету. В отчете должны быть четкие, не расплывшиеся линии, буквы, цифры и знаки.

Расположение текста должно обеспечивать соблюдение следующих полей:

- левое поле - не менее 30 мм;
- правое поле - не менее 10 мм;
- верхнее поле - не менее 20 мм;
- нижнее поле - не менее 20 мм.

Все страницы курсовой работы (проекта), включая приложения, должны быть пронумерованы арабскими цифрами сквозной нумерацией по всему тексту. Первой страницей является титульный лист, на котором номер страницы не проставляется. Порядковый номер страницы помещается в нижнем правом углу колонтитула.

Структура выпускной квалификационной работы состоит из следующих элементов:

1. *Титульный лист,* образец которого представлен в приложении А
2. *Пояснительная записка*:
   - Содержание (см. Приложение Б) - включает в себя перечень частей ВКР с указанием страниц, соответствующих началу каждой части работы;
   - Введение - раскрывает актуальность выбранной темы исследования, степень разработанности темы, цели, задачи, объект, предмет, гипотезу и методы исследования, структуру работы;
   - Основная часть - состоит из нескольких глав, содержащих параграфы;
   - Заключение - подводятся основные итоги работы, обобщаются полученные результаты, освещаются рекомендации по конкретному использованию результатов выпускной квалификационной работы и направления дальнейших исследований;
   - Список использованных источников - он включает литературу, используемую при подготовке текста: цитируемую, упоминаемую, а также имеющую непосредственное отношение к исследуемой теме. Полнота списка зависит от тщательности сбора публикаций. Правильно составленный и грамотно оформленный список свидетельствует о том, насколько автор знаком с литературой по теме исследования. Важным компонентом является работа автора с литературой последних трех-пяти лет, как показатель ориентированности автора в современном состоянии научной изученности темы исследования. Библиографический список должен включать не менее 20 источников.
   - Приложения (если таковые имеются).

Оформление заголовков и основного текста.

Текст работы следует разделять на разделы, подразделы и пункты. Разделы и подразделы должны иметь заголовки. Наименования структурных элементов отчета "СОДЕРЖАНИЕ", "ВВЕДЕНИЕ», "ЗАКЛЮЧЕНИЕ", "СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ", "ПРИЛОЖЕНИЕ" служат заголовками структурных элементов работы (проекта). Заголовки структурных элементов (введение, заключение, главы и т.п.) следует располагать в середине строки без точки в

конце и печатать прописными (заглавными) буквами, не подчеркивая, полужирный шрифт не применяется.

Разделы основной части пояснительной записки работы (проекта) должны иметь порядковые номера в пределах всего документа, обозначенные арабскими цифрами без точки и записанные с абзацного отступа. Подразделы должны иметь нумерацию в пределах каждого раздела. В конце номера подраздела точка не ставится.

Каждый раздел следует начинать с нового листа (страницы). Расстояние между заголовками раздела или подраздела приблизительно 1.5-2 см. Расстояние между заголовками раздела и текстом должно быть равно 1,5-2 см.

Расположение текста должно обеспечивать следующих полей:

- левое поле – не менее 30 мм;
- правое поле – не менее 10 мм;
- верхнее поле – не менее 20 мм;
- нижнее поле – не менее 20 мм.

Все страницы курсовой работы (проекта), включая приложения, должны быть пронумерованы арабскими цифрами, шрифт Times New Roman, 12 пт. Порядковый номер страницы помещается в нижнем правом углу колонтитула.

Оформление заголовков раздела (ВВЕДЕНИЕ, ГЛАВА и т.д.):

- междустрочный интервал - 1,5;
- шрифт Times New Roman;
- **написание - прописные (заглавные) буквы;**
- **полужирный шрифт не применяется**;
- размер шрифта 14 пт;
- **режим выравнивания - по центру**;
- отступ в начале абзаца - 15 мм.

Оформление заголовков подраздела и подпункта (1.1, 1.2 и т.д.):

- междустрочный интервал - 1,5;
- шрифт Times New Roman;
- **написание - первая заглавная, остальные строчные буквы;**
- **полужирный шрифт не применяется;**
- размер шрифта 14 пт;
- **режим выравнивания - слева;**

– отступ в начале абзаца - 15 мм.

Оформление основного текста работы (проекта):

– междустрочный интервал - 1,5;
– шрифт Times New Roman;
– **полужирный шрифт не применяется**;
– размер шрифта 14 пт (для таблиц допускается 12 пт);
– **режим выравнивания - по ширине**;
– отступ в начале абзаца - 15 мм.

Разрешается использовать компьютерные возможности акцентирования внимания на определенных терминах, формулах, теоремах, применяя шрифты разной гарнитуры.

Числовые значения величин в тексте следует указывать с необходимой степенью точности, при этом в ряду величин осуществляется выравнивание числа знаков после запятой. Округление числовых значений величин до первого, второго, третьего и т.д. десятичного знака для величин одного наименования должны быть одинаковыми. Например: 1,50; 1,75; 2,00.

Оформление списков.

Внутри пунктов или подпунктов раздела могут быть приведены перечисления, которые записываются с абзацного отступа. **Перед каждой позицией перечисления следует ставить дефис**, а при необходимости ссылки в тексте ВКР на один из элементов перечисления вместо дефиса ставятся строчные буквы в порядке русского алфавита, начиная с буквы а (за исключением букв ё, з, й, о, ч, ъ, ы, ь). Для дальнейшей детализации перечислений необходимо использовать арабские цифры, после которых ставится скобка, а запись производится с абзацного отступа.

Примеры приведены на рисунках 1 и 2.

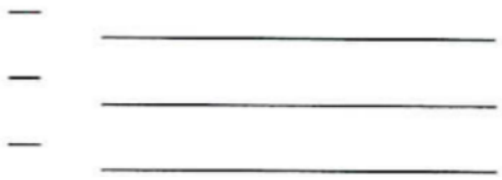


Рис. 1 – Пример оформления списка

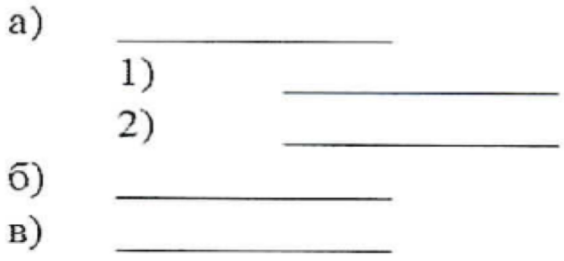


Рис. 2 – Пример оформления списка при необходимости дальнейшей ссылки на один из его элементов

Оформление формул.

Уравнения и формулы следует выделять из текста в отдельную строку. Выше и ниже каждой формулы или уравнения должно быть оставлено не менее одной свободной строки. Пояснения символов и числовых коэффициентов, входящих в формулу, если они не пояснены ранее в тексте, должны быть приведены непосредственно под формулой. Пояснения каждого символа следует давать с новой строки в той последовательности, в которой символы приведены в формуле. Первая строка пояснения должна начинаться со слова "где" без двоеточия после него.

Формулы должны нумероваться сквозной нумерацией арабскими цифрами, которые записывают на уровне формулы в крайнем положении справа в круглых скобках. Одну формулу обозначают - (1).

Ссылки в тексте на порядковые номера формул дают в скобках, например "... в формуле (1)".

Пример:

Плотность каждого образца ρ, кг/м^3, вычисляют по формуле

$$\rho = \frac{m}{V}, \qquad (1)$$

где m - масса образца, кг;

V - объем образца, м^3.

Оформление таблиц.

Таблицу следует располагать непосредственно после текста, в котором она упоминается впервые. При ссылке следует писать слово "таблица" с указанием ее номера.

Все таблицы должны иметь название и порядковую нумерацию. Таблицы нумеруются арабскими цифрами сквозной нумерацией в пределах всей работы (за исключением таблиц приложений). Номер таблицы следует проставлять в левом верхнем углу над таблицей после слова Таблица, без знака №, например, "Таблица 1". В приложениях таблицы обозначают отдельной нумерацией арабскими цифрами с добавлением перед цифрой обозначения приложения, например "Таблица В.1", если она приведена в приложении В.

Название таблицы должно отражать ее содержание, быть точным кратким. **Наименование таблицы следует помещать над таблицей слева, без абзацного отступа в одну строку с ее номером через тире**.

Таблицы выравниваются по центру страницы и оформляются в соответствии с рисунком 3. Выше и ниже каждой таблицы должно быть оставлено не менее одной свободной строки.

Рис. 3 – Оформление таблиц

В каждой таблице следует указывать единицы измерения показателей и период времени к которому относятся данные. если единица измерения в таблице является общей для всех числовых данных то ее приводят в заголовке таблицы после ее названия.

Таблицу с большим числом строк допускается переносить на другой лист (страницу). При переносе части таблицы на другой лист (страницу) слово "Таблица", ее номер и наименование указывают один раз слева над первой частью таблицы и указывают номер таблицы (рис. 4).

Расчет влияния каждого фактора представлен в таблице 1.

Таблица 1 - Расчет влияния факторов на деятельность производственного предприятия

Фактор	Расчет
1. Выручка от реализации	$\Delta П_{п_в} = (В_1 - В_0)$ х $R1_0$ / 100 (6) где $\Delta П_{п_в}$ - изменение суммы прибыли от продаж за счет изменения

Продолжение таблицы 1

	объемов выручки; $В_1$ и $В_0$ - соответственно выручка от продажи в отчетном (1) и базисном (0) периодах; $R1_0$ - рентабельность продаж в базисном периоде
2. Себестоимость реализованной продукции	$\Delta П_{п_с} = В_1 \times (УС_1 - УС_0) / 100$ (7) где $УС_1$ и $УС_0$ – соответственно уровни себестоимости в отчетном и базисном периодах.
3. Коммерческие расходы	$\Delta П_{п_{кр}} = В_1 \times (УКР_1 - УКР_0)/100$ (8) где $УКР_1$ и $УКР_0$ – соответственно уровни коммерческих расходов в отчетном и базисном периодах.
4. Управленческие расходы	$\Delta П_{п_{ур}} = В_1 \times (УУР_1 - УУР_0)/100$ (9) где $УУР_1$ и $УУР_0$ – соответственно уровни управленческих расходов в отчетном и базисном периодах.

Рис. 4 – Оформление при делении таблиц

Оформление иллюстраций и графической части.

Весь графический материал (схемы, диаграммы, фотографии, чертежи и т.п.), расположенный по тексту работы (не включая приложения), следует нумеровать арабскими цифрами сквозной нумерацией. Если рис. один, то он обозначается "Рис. 1". Графики, схемы, диаграммы, располагаются в работе непосредственно после текста, имеющего на них ссылку, или на следующей странице. Поясняющие данные помещают под иллюстрацией, а **ниже по центру печатают слово "Рис.", его номер, а через знак "–" и его наименование**. Иллюстрации каждого приложения обозначают отдельной нумерацией арабскими цифрами с добавлением перед цифрой обозначения приложения. Например, "Рис. А.3 – Детали прибора".

Пример оформления иллюстраций представлен на рисунке 5.

Структура продаж товаров основных товарных групп магазина представлена на рисунке 1.

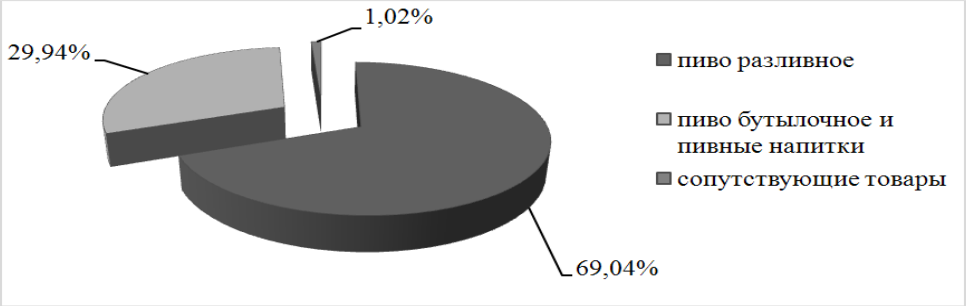


Рисунок 1 – Структура продаж магазина «Золотая кружка» (ИП Медведев К.Д.) по данным на 2015г., %

Рис. 5 – Пример оформления иллюстраций и графической части

При ссылках на иллюстрации следует писать "... в соответствии с рисунком 2".

Выше и ниже каждого рисунка должно быть оставлено не менее одной свободной строки.

Оформление приложений.

Материал, дополняющий текст документа, допускается помещать в приложениях. Приложения могут быть, например, графический материал, таблицы большого формата, расчеты, описания аппаратуры и приборов, описания алгоритмов и программ задач, решаемых на ЭВМ и т.д. Приложения располагают в порядке появления ссылок на них в тексте документа. В тексте документа на все приложения должны быть даны ссылки.

Каждое приложение следует начинать с новой страницы с указанием наверху посередине страницы слова "ПРИЛОЖЕНИЕ" и его обозначения.

Приложения обозначают заглавными буквами русского алфавита, начиная с А. Буквы Ё, З, Й, О, Ч, Ь, Ы, Ъ для обозначения приложений НЕ используются.

Приложение должно иметь заголовок, который записывают симметрично относительно текста (выравнивание по тексту) с прописной (заглавной) буквы с новой строки.

Пример оформления приложения представлен на рисунке 6.

ПРИЛОЖЕНИЕ А

Основные группы показателей экономической эффективности хозяйственной деятельности предприятия

Таблица А.1 - Основные группы показателей экономической эффективности хозяйственной деятельности предприятия

Показатели	Характеристика	Способ расчета
I. Производительность труда		
1. Выработка	Отражает количество продукции, произведенной в единицу рабочего времени или приходящееся на одного среднесписочного работника в месяц, квартал, год	Отношение количества произведенной продукции к затратам рабочего времени на производство этой продукции
2. Трудоемкость	Величина, обратная выработке, характеризует затраты труда на производство единицы продукции	Отношение затрат труда к объему продукции

Рис. 6 – Пример оформления приложения

Оформление библиографического списка используемой литературы.

Список используемой литературы содержит перечень источников, используемых обучающимся при работе над темой работы (проекта). Список используемой литературы нумеруется арабскими цифрами, после которых ставится скобка, запись производится с абзацного отступа. Сведения об источниках следует располагать в порядке появления ссылок на источники в тексте работы (проекта).

При написании работы обучающийся обязан давать ссылку на источник, библиографическое описание которого должно приводиться в списке используемых источников. Порядковый номер ссылки в тексте работы заключают в квадратные скобки.

Для каждого учебника, книги обязательно должен быть указан уникальный номер книжного издания ISBN, для периодических изданий – ISSN, для электронных ресурсов – ссылка (URL) и дата обращения.

7. ПРОЦЕДУРА ЗАЩИТЫ КР (КП)

Курсовая работа (проект), выполненная с соблюдением рекомендуемых требований, оценивается и допускается к защите. Защита должна производиться до начала экзамена или зачета по дисциплине.

Процедура защиты курсовой работы/проекта включает в себя:

- выступление студента по теме и результатам работы (5-8 мин),
- ответы на вопросы аудитории и научного руководителя работы.

Окончательная оценка за курсовую работу (проект) выставляется преподавателем после защиты.

Результаты защиты курсового проекта оцениваются по четырехбалльной системе: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно», а при выполнении курсовой работы по двухбалльной системе: «зачтено» или «незачтено».

Положительная оценка по той дисциплине, по которой предусматривается курсовая работа (проект), выставляется только при условии успешной сдачи курсовой работы (проекта) на оценку не ниже «удовлетворительно».

К защите курсовой работы (проекта) предъявляются следующие требования:

1. Глубокая теоретическая проработка исследуемых проблем на основе анализа экономической литературы.
2. Умелая систематизация цифровых данных в виде таблиц и графиков с необходимым анализом, обобщением и выявлением тенденций развития исследуемых явлений и процессов.
3. Критический подход к изучаемым фактическим материалам с целью поиска направлений совершенствования деятельности.
4. Аргументированность выводов, обоснованность предложений и рекомендаций.
5. Логически последовательное и самостоятельное изложение материала.
6. Оформление материала в соответствии с установленными требованиями.

Для выступления на защите необходимо заранее подготовить и согласовать с руководителем тезисы доклада и иллюстрационный материал в виде презентации.

Рекомендуемые структура, объем и время доклада приведены в таблице 2.

Таблица 2 – Структура, объем и время доклада

№	Структура доклада	Объем	Время
1.	Представление темы работы.	До 1,5 страниц	До 2 минут
2.	Актуальность темы.		
3.	Цель работы.		
4.	Постановка задачи, результаты ее решения и сделанные выводы (по каждой из задач, которые были поставлены для достижения цели курсовой работы/ проекта).	До 6 страниц	До 7 минут
5.	Перспективы и направления дальнейшего исследования данной темы.	До 0,5 страницы	До 1 минуты

При составлении тезисов необходимо учитывать ориентировочное время доклада на защите, которое составляет 8-10 минут. Доклад целесообразно строить не путем изложения содержания работы по главам, а по задачам, то есть, раскрывая логику получения значимых результатов. В докладе обязательно должно присутствовать обращение к иллюстративному материалу, который будет использоваться в ходе защиты работы. Объем доклада должен составлять 7-8 страниц текста в формате Word, размер шрифта 14, полуторный интервал [110].

Темы примерных курсовых работ по курсу

1. Прогнозирование экономического роста страны с использованием статистических моделей;
2. Анализ влияния инфляции на экономические показатели с помощью статистического моделирования;
3. Оценка влияния безработицы на экономическую активность с применением статистических методов;

4. Сравнительный анализ различных моделей прогнозирования цен на товары и услуги;

5. Исследование факторов, влияющих на валютный курс с использованием статистических моделей;

6. Применение временных рядов для прогнозирования динамики фондового рынка;

7. Моделирование зависимости между уровнем доходов населения и объемом потребления;

8. Анализ влияния изменений налоговой политики на экономическую среду с помощью статистических методов;

9. Прогнозирование спроса на определенный вид продукции с использованием регрессионного анализа;

10. Оценка эффективности маркетинговых кампаний с помощью статистических моделей;

11. Исследование взаимосвязи между инвестициями и экономическим ростом с применением статистических методов;

12. Прогнозирование цен на нефть с использованием временных рядов;

13. Анализ воздействия климатических изменений на экономику с помощью статистического моделирования;

14. Исследование факторов, влияющих на развитие малого и среднего бизнеса с применением статистических методов;

15. Сравнительный анализ различных методов прогнозирования инфляции;

16. Моделирование влияния глобальных экономических кризисов на отдельные отрасли экономики;

17. Прогнозирование изменений в потребительском спросе с использованием статистических моделей;

18. Анализ эффективности монетарной политики с помощью статистического моделирования;

19. Исследование влияния демографических факторов на развитие рынка труда с применением статистических методов;

20. Прогнозирование изменений в инвестиционной активности компаний с использованием регрессионного анализа;

21. Анализ взаимосвязи между образованием и экономическим развитием с применением статистических методов;

22. Моделирование влияния глобализации на экономику страны с использованием временных рядов;

23. Прогнозирование изменений в уровне безработицы с использованием статистических моделей;

24. Анализ влияния политических решений на экономическую конъюнктуру с помощью статистического моделирования;

25. Исследование долгосрочных тенденций развития финансовых рынков с применением статистических методов;

26. Сравнительный анализ различных подходов к прогнозированию роста ВВП;

27. Моделирование воздействия технологических инноваций на экономику с использованием статистических моделей;

28. Прогнозирование изменений в индексе потребительских цен с использованием временных рядов;

29. Анализ влияния торговой политики на экономическую активность с применением статистических методов

30. Исследование факторов, определяющих конкурентоспособность страны, с использованием статистических моделей;

31. Прогнозирование изменений в объеме экспорта и импорта с помощью статистического моделирования;

32. Анализ воздействия кризисов на финансовые рынки с применением статистических методов;

33. Исследование взаимосвязи между уровнем инвестиций и инновационной активностью компаний с использованием статистических моделей;

34. Сравнительный анализ различных моделей прогнозирования уровня безработицы;

35. Моделирование влияния роста населения на социально-экономическое развитие страны с использованием временных рядов;

36. Прогнозирование изменений в финансовой устойчивости банковской системы с использованием статистических моделей;

37. Анализ влияния цен на энергоносители на экономику страны с помощью статистического моделирования;

38. Исследование факторов, определяющих инвестиционный климат в стране, с применением статистических методов;

39. Прогнозирование изменений в потребительском поведении с использованием регрессионного анализа;

40. Анализ эффективности государственных программ поддержки экономики с помощью статистических методов;

41. Исследование взаимосвязи между уровнем образования населения и экономическим развитием страны с применением статистических моделей;

42. Сравнительный анализ различных подходов к прогнозированию инвестиционной активности компаний;

43. Моделирование воздействия миграционных процессов на экономику страны с использованием временных рядов;

44. Прогнозирование изменений в индексе промышленного производства с использованием статистических моделей;

45. Анализ влияния изменений демографической ситуации на рынок недвижимости с помощью статистического моделирования;

46. Исследование факторов, влияющих на развитие цифровой экономики, с применением статистических методов;

47. Прогнозирование изменений в уровне инфляции с использованием регрессионного анализа;

48. Анализ эффективности инвестиционных проектов с помощью статистических методов;

49. Исследование взаимосвязи между уровнем коррупции и экономическим развитием страны с применением статистических моделей;

50. Сравнительный анализ различных методов прогнозирования рентабельности предприятий;

51. Моделирование воздействия изменений курса иностранной валюты на экспортно-импортные операции с использованием временных рядов;

52. Прогнозирование изменений в индексе производственной активности с использованием статистических моделей;

53. Анализ влияния технологических инноваций на конкурентоспособность предприятий с помощью статистического моделирования;

54. Исследование факторов, определяющих эластичность спроса, с применением статистических методов;

55. Прогнозирование изменений в объеме инвестиций в основной капитал с использованием регрессионного анализа;

56. Анализ эффективности мер государственной поддержки малого и среднего бизнеса с помощью статистических методов;

57. Исследование взаимосвязи между уровнем заработной платы и уровнем жизни населения с применением статистических моделей;

58. Сравнительный анализ различных подходов к прогнозированию спроса на товары и услуги;

59. Моделирование воздействия изменений законодательства на бизнес-среду с использованием временных рядов;

60. Прогнозирование изменений в индексе потребительской удовлетворенности с использованием статистических моделей;

61. Анализ влияния глобальных экологических проблем на экономику страны с помощью статистического моделирования;

62. Исследование факторов, определяющих эффективность инновационной политики государства, с применением статистических методов

63. Сравнительный анализ различных методов прогнозирования изменений в индексе деловой активности;

64. Моделирование воздействия изменений тарифного регулирования на отраслевую конкуренцию с использованием временных рядов;

65. Прогнозирование изменений в объеме занятости населения с использованием статистических моделей;

66. Анализ влияния социальной политики государства на уровень бедности и социальной напряженности с помощью статистического моделирования;

67. Исследование связи между уровнем инноваций и конкурентоспособностью предприятий с применением статистических методов;

68. Сравнительный анализ различных подходов к прогнозированию изменений в объеме производства товаров и услуг;

69. Моделирование воздействия изменений таможенного регулирования на объем международной торговли с использованием временных рядов;

70. Прогнозирование изменений в индексе индустриализации страны с использованием статистических моделей;

71. Анализ воздействия цифровизации экономики на развитие малого и среднего бизнеса с помощью статистического моделирования;

72. Исследование факторов, определяющих эффективность программ поддержки занятости, с применением статистических методов;

73. Сравнительный анализ различных методов прогнозирования изменений в индексе инноваций;

74. Моделирование воздействия изменений трудового законодательства на динамику заработной платы с использованием временных рядов;

75. Прогнозирование изменений в объеме инвестиций в человеческий капитал с использованием статистических моделей.

ЗАКЛЮЧЕНИЕ

В настоящее время в управлении фирмой активно применяются различные статистические методы. Возрастает актуальность повышения качества прогнозных исследований на микроуровне. Все это обусловливает необходимость углубленного изучения и разработки основных проблем применения статистических методов, с которыми можно столкнуться в процессе управления организацией.

В процессе систематизированного научно обоснованного прогнозирования развития социально-экономических процессов происходило развитие методологии прогнозирования как совокупности методов, приемов и способов мышления, позволяющих на основе анализа ретроспективных данных, экзогенных и эндогенных связей объекта прогнозирования, а также их измерений в рамках рассматриваемого явления или процесса вывести суждения определенной достоверности относительно его будущего развития.

Исследование различных классификационных схем методов прогнозирования позволяет выделить в качестве основных классов фактографические, экспертные и комбинированные методы, специализация которых обусловлена спецификой целей и задач, количеством и качеством исходной информации, периодом упреждения прогноза.

Прогнозы должны предшествовать планам, содержать оценку хода, последствий выполнения (или невыполнения) планов, охватывать все, что не поддается планированию, решению. Прогноз и план различаются способами оперирования информацией о будущем.

В первой главе пособия были рассмотрены теоретические аспекты применения статистических методов в управлении фирмой, в последующих девяти главах – даны практические аспекты моделирования и прогнозирования на микроуровне.

Приведенные в пособии положения могут быть использованы на лекционных и практических занятиях по дисциплине «Статистические методы в управлении фирмой».

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Шорохова, И.С. Статистические методы анализа : [учеб. пособие] / И. С. Шорохова, Н. В. Кисляк, О. С. Мариев; М-во образования и науки Рос. Федерации, Урал. федер. ун-т. – Екатеринбург : Изд-во Урал. ун-та, 2015. – 300 с. – ISBN 978-5-7996-1633-5

2. Курмангалиева А.К. Статистические методы анализа : [учеб. пособие]. – Костанай, КГУ им. А. Байтурсынова, 2018 – 112 с. – ISBN 978-601-7955-08-3

3. Басовский, Л. Е. Прогнозирование и планирование в условиях рынка : учеб. пособие / Л. Е. Басовский. – М. : ИНФРА-М, 2023. – 260 с. – (Высшее образование: Бакалавриат). – ISBN 978-5-16-004198-8.

4. Герасимов, А. Н. Социально-экономическое прогнозирование : учеб. пособие / А. Н. Герасимов, Е. И. Громов, Ю. С. Скрипниченко. – М. : СтГАУ : Агрус, 2017. – 144 с. – ISBN 978-5-9596-1294-8.

5. Почекутова, Е. Н. Прогнозирование и планирование : учеб.-метод. пособие / Е. Н. Почекутова, А. П. Феденко. – Красноярск : СФУ, 2016. – 126 с. – ISBN 978-5-7638-3439-0.

6. Кулешова, Е. С. Макроэкономическое планирование и прогнозирование : учеб. пособие / Е. С. Кулешова. – 2-е изд., доп. – Томск : Эль-Контент, 2015. – 178 с. – ISBN 978-5-4332-0252-8.

7. Петросов, А. А. Стратегическое планирование и прогнозирование / Петросов А.А. – М. :МГГУ, 2001. – 464 с.: ISBN 5-7418-0145-5.

8. Прогнозирование и планирование в условиях рынка [Электронный ресурс] : учеб. пособие / Т. Н. Бабич [и др.]. – М. : ИНФРА-М, 2020. – 336 с. – (Высшее образование: Бакалавриат). – URL: www.dx.doi.org/10.12737/2517. – ISBN 978-5-16-004577-1 (дата обращения: 15.03.2023).

9. Судакова, А. Е. Бюджетное планирование и прогнозирование : учеб. пособие / А. Е. Судакова, Г. А. Агарков, А. А. Тарасьев. - 2-е изд., стер. – М. : ФЛИНТА : Изд-во Урал. ун-та, 2022.– 308 с. - ISBN 978-5-9765-5024-7 (ФЛИНТА), ISBN 978-5-7996-2922-9 (Изд-во Урал. ун-та).

10. Основы экономического прогнозирования [Электронный ресурс] : учебное пособие / Е.В. Смирнова, Е.В. Чмышенко, И.Ю. Цыганова; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2019. – 145 с. ISBN 978-5-7410-2425-6

11. Социально-экономическое прогнозирование: учебное пособие / Ю. А. Антохина, А. М. Колесникова, С. Н. Медведева ; М-во образования и науки Российской Федерации, Федеральное гос. бюджетное образовательное учреждение высш. образования Санкт-Петербургский гос. ун-т аэрокосмического приборостроения. - Санкт-Петербург : ГУАП, 2016. – 177 с. : табл.; 21 см.; ISBN 978-5-8088-1103-4.

12. Виноградская, Н. А. Управление производством : методы экономического прогнозирования и планирования : практикум / Н. А. Виноградская, Е. Н. Елисеева, О. О. Скрябин. – М. : Изд. Дом МИСиС, 2013. - 96 с. - ISBN 978-5-87623-687-6

13. Орехов, А. М. Методы экономических исследований : учебное пособие / А.М. Орехов. — 2-е изд. — Москва : ИНФРА-М, 2023. — 344 с. — (Высшее образование: Бакалавриат). - ISBN 978-5-16-005748-4

14. Бабешко, Л. О. Эконометрика и эконометрическое моделирование : учебник / Л.О. Бабешко, М.Г. Бич, И.В. Орлова. — 2-е изд., испр. и доп. – Москва : ИНФРА-М, 2023. – 387 с. : ил. – (Высшее образование: Бакалавриат). – DOI 10.12737/1141216. – ISBN 978-5-16-016417-5

15. Лычкина, Н. Н. Имитационное моделирование экономических процессов : учебное пособие / Н.Н. Лычкина. – Москва : ИНФРА-М, 2022. – 254 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/724. – ISBN 978-5-16-017094-7

16. Бакланова, И. И. Теория вероятности : учебно-методическое пособие / И. И. Бакланова, Е. В. Матвеева, Л. А. Медведков. - Йошкар-Ола : Поволжский государственный технологический университет, 2017. - 64 с. - ISBN 978-5-8158-1801-9

17. Бочаров, П. П. Теория вероятностей. Математическая статистика [Электронный ресурс] / П. П. Бочаров, А. В. Печинкин. – 2-е изд. - Москва : ФИЗМАТЛИТ, 2005. - 296 с. - ISBN 5-9221-0633-3

18. Коган, Е. А. Теория вероятностей и математическая статистика : учебник / Е.А. Коган, А.А. Юрченко. – М. : ИНФРА-М, 2020. – 250 с. – (Среднее профессиональное образование). – ISBN 978-5-16-015649-1

19. Бакланова, И. И. Теория вероятности : учебно-методическое пособие / И. И. Бакланова, Е. В. Матвеева, Л. А. Медведков. – Йошкар-Ола: Поволжский государственный технологический университет, 2017. – 64 с. – ISBN 978-5-8158-1801-9

20. Теория вероятностей и математическая статистика : учебное пособие / Л.Г. Бирюкова, Г.И. Бобрик, Р.В. Сагитов [и др.] ; под ред. В.И. Матвеева. – 2-е изд., испр. и доп. – М.: ИНФРА-М, 2020. – 289 с. – (Среднее профессиональное образование). – ISBN 978-5-16-015712-2

21. Мхитарян, В. С. Теория вероятностей и математическая статистика [Электронный ресурс] : учеб. пособие / В. С. Мхитарян, Е. В. Астафьева, Ю. Н. Миронкина, Л. И. Трошин; под ред. В. С. Мхитаряна. – 2-е изд., перераб. и доп. – М. : Московский финансово-промышленный университет «Синергия», 2013. – (Университетская серия). – ISBN 978-5-4257-0106-0

22. Лычкина, Н. Н. Имитационное моделирование экономических процессов : учебное пособие / Н.Н. Лычкина. – Москва : ИНФРА-М, 2022. – 254 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/724. – ISBN 978-5-16-017094-7

23. Булыгина, О. В. Имитационное моделирование в экономике и управлении : учебник / О.В. Булыгина, А.А. Емельянов, Н.З. Емельянова ; под ред. д-ра экон. наук, проф. А.А. Емельянова. – Москва : ИНФРА-М, 2021. – 592 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/textbook_5b5ab5571bd995.05564317. – ISBN 978-5-16-014523-5

24. Гулай, Т.А. Теория вероятностей и математическая статистика [Электронный ресурс] : учебное пособие / Т.А. Гулай, А.Ф. Долгополова, Д.Б. Литвин, С.В. Мелешко. – 2-е изд., доп. – Ставрополь: АГРУС, 2013. – 260 с

25. Осипов, Г. В. Моделирование социальных явлений и процессов с применением математических методов : учебное пособие / Г. В. Осипов, В. А. Лисичкин ; под общ. ред. В. А. Садовничего. – Москва : Норма : ИНФРА-М, 2022. – 192 с. : ил. – (Социальные науки и математика). – ISBN 978-5-91768-533-5

26. Криволапов, С. Я. Теория вероятностей в примерах и задачах на языке R : учебник / С.Я. Криволапов. – М.: ИНФРА-М, 2023. – 412 с. + Доп. материалы [Электронный ресурс]. – (Высшее образование). – DOI 10.12737/1898404. – ISBN 978-5-16-017941-4

27. Ларионова, И. А. Статистика. Анализ временных рядов: учебное пособие / И. А. Ларионова. – М.: ИД МИСиС, 2001. – 73 с.

28. Воейко, О. А. Анализ временных рядов и прогнозирование : практикум / О. А. Воейко. – М.; Берлин : Директ-Медиа, 2019. – 175 с. – ISBN 978-5-4499-0178-1

29. Ярушкина, Н. Г. Интеллектуальный анализ временных рядов : учебное пособие / Н.Г. Ярушкина, Т.В. Афанасьева, И.Г. Перфильева. – М. : ИД «ФОРУМ»: ИНФРА-М, 2018. – 160 с. – (Высшее образование). – ISBN 978-5-8199-0496-1

30. Карпенко, Н. В. Эконометрика. Анализ и прогнозирование временного ряда : учебное пособие / Н. В. Карпенко. – М.: РУТ (МИИТ), 2018. – 132 с.

31. Замедлина, Е. А. Статистика: учебное пособие / Е.А. Замедлина – М.: РИОР : ИНФРА-М, 2019. – 160 с. – (Среднее профессиональное образование). – ISBN 978-5-369-01303-8

32. Шумак, О. А. Статистика: Учебное пособие / О.А. Шумак, А.В. Гераськин. – М.: ИЦ РИОР: НИЦ Инфра-М, 2019. – 311 с.: ил.; – (Высшее образование: Бакалавриат). – ISBN 978-5-369-01048-8

33. Статистика : учебник / В.В. Глинский, В.Г. Ионин, Л.К. Серга [и др.] ; под ред. В.Г. Ионина. – 4-е изд., перераб. и доп. – М.: ИНФРА-М, 2023. – 355 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/25127. – ISBN 978-5-16-012070-6

34. Сергеева, И. И. Статистика: учебник / И.И. Сергеева, Т.А. Чекулина, С.А. Тимофеева. – 2-е изд., испр. и доп. — Москва : ФОРУМ : ИНФРА-М, 2021. – 304 с. – (Среднее профессиональное образование). – ISBN 978-5-8199-0888-4

35. Годин, А. М. Статистика : учебник для бакалавров / А. М. Годин. - 12-е изд., стер. – Москва : Издательско-торговая корпорация «Дашков и К°», 2020. – 410 с. – ISBN 978-5-394-03485-5

36. Тимофеева, И. Ю. Статистика. Часть 1. Общая теория статистики: Учебное пособие / Тимофеева И.Ю., Лаврова Е.В., Полякова О.Е. – М.:НИЦ ИНФРА-М, 2018. – 104 с. (Высшее образование)ISBN 978-5-16-107041-3

37. Сидоренко, М. Г. Статистика : учебное пособие / М.Г. Сидоренко. – М. : ФОРУМ : ИНФРА-М, 2022. – 160 с. – (Высшее образование). – ISBN 978-5-91134-160-2

38. Иода, Е. В. Статистика: Учебное пособие / Иода Е.В. - М.:Вузовский учебник, НИЦ ИНФРА-М, 2018. – 303 с. ISBN 978-5-9558-0144-5

39. Ивченко, Ю. С. Статистика: Учебное пособие / Ю.С. Ивченко. - Москва : ИЦ РИОР: ИНФРА-М, 2018. – 375 с.: – (Высшее образование). – ISBN 978-5-369-00636-8

40. Непомнящая, Н. В. Статистика: общая теория статистики, экономическая статистика. Практикум/НепомнящаяН.В., ГригорьеваЕ.Г. – Краснояр.: СФУ, 2015. – 376 с.: ISBN 978-5-7638-3185-6

41. Гужова, О. А. Статистика в управлении социально-экономическими процессами : учебное пособие / О.А. Гужова, Ю.А. Токарев. – М. : ИНФРА-М, 2020. – 172 с. – (Высшее образование: Бакалавриат). – ISBN 978-5-16-012151-2

42. Мелкумов, Я. С. Социально-экономическая статистика : учебное пособие / Я.С. Мелкумов. – 2-е изд., перераб. и доп. – М. : ИНФРА-М, 2023. – 186 с. – (Высшее образование: Бакалавриат). - ISBN 978-5-16-005424-7

43. Экономическая статистика. Практикум: учебное пособие / Ю.Н. Иванов, Г.Л. Громыко, А.Н. Воробьев [и др.] ; под ред. д-ра экон. наук, проф. Ю.Н. Иванова. – М. : ИНФРА-М, 2022. – 176 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/23950. - ISBN 978-5-16-012772-9

44. Батракова, Л. Г. Социально-экономическая статистика : учебник / Л. Г. Батракова. – М.: Логос, 2020. – 480 с. – ISBN 978-5-98704-657-9

45. Гусева, Е. Н. Теория вероятностей и математическая статистика [Электронный ресурс] : Уч. пособ. / Е. Н. Гусева. - 5-е изд., стереотип. – М.: Флинта, 2011. – 220 с. – ISBN 978-5-9765-1192-7.

46. Непомнящая, Н. В. Статистика: общая теория статистики, экономическая статистика. Практикум/НепомнящаяН.В., ГригорьеваЕ.Г. – Краснояр.: СФУ, 2015. – 376 с.: ISBN 978-5-7638-3185-6

47. Теория вероятностей и математическая статистика : учебное пособие / Л.Г. Бирюкова, Г.И. Бобрик, Р.В. Сагитов [и др.] ; под ред. В.И. Матвеева. – 2-е изд., испр. и доп. – М. : ИНФРА-М, 2021. – 289 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/18865. - ISBN 978-5-16-011793-5

48. Теоретическая метрология : методические указания к практическим занятиям / сост. И. А. Петюль, В. Д. Борозна. – Витебск : УО «ВГТУ», 2017. – 66 с.

49. Белько, И. В. Теория вероятностей, математическая статистика, математическое программирование : учебное пособие / И.В. Белько, И.М. Морозова, Е.А. Криштапович. – М.: ИНФРА-М, 2022. – 299 с. : ил. – (Высшее образование: Бакалавриат). – ISBN 978-5-16-011748-5

50. Эрастов, В. Е. Метрология, стандартизация и сертификация : учебное пособие / В.Е. Эрастов. – 2-е изд., перераб. и доп. — Москва : ИНФРА-М, 2022. – 196 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/23696. – ISBN 978-5-16-012324-0

51. Белокопытов, В. И. Организация, планирование и обработка результатов эксперимента : учебное пособие / В. И. Белокопытов. – Красноярск : Сиб. федер. ун-т, 2020. – 132 с. – ISBN 978-5-7638-4297-5

52. Шпаков, П. С. Математическая обработка результатов измерений / П. С. Шпаков, Ю. Л. Юнаков. – Красноярск : СФУ, 2014. - 410 с. – ISBN 978-5-7638-3077-4.

53. Гальянов, А. В. Математическая обработка результатов измерений в горном деле : учебное пособие / А. В. Гальянов. – 2-е изд., испр. и доп. – М.; Вологда : Инфра-Инженерия, 2022. – 292 с. - ISBN 978-5-9729-0815-8

54. Организация производства и управление предприятием : учебник / под ред. О.Г. Туровца. – 3-е изд. – М.: ИНФРА-М, 2022. – 506 с. – (Среднее профессиональное образование). – ISBN 978-5-16-015612-5

55. Пискажова, Т. В. Математическое моделирование объектов и систем управления : учебное пособие / Т. В. Пискажова, Т. В. Донцова, Г. Б. Даныкина. – Красноярск : Сиб. федер. ун-т, 2020. – 230 с. – ISBN 978-5-7638-4184-8

56. С.С. Бондарчук, И.С. Бондарчук. Б81 Статобработка экспериментальных данных в MS Excel: учебное пособие. – Томск: Издательство Томского государственного педагогического университета, 2018. – 433

57. Жуков, В. И. Методология математического моделирования управления социальными процессами : монография / В.И. Жуков, Г.С. Жукова. – М.: ИНФРА-М, 2019. – 207 с. – (Научная мысль). – ISBN 978-5-16-108296-6

58. Двойцова, И. Н. Основы математического моделирования социально-экономических процессов : учебное пособие / И. Н. Двойцова.

– Железногорск : ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России, 2022. – 112 с.

59. Инструментальные средства математического моделирования: учебное пособие / Золотарев А.А., Бычков А.А., Золотарева Л.И. – Ростов н/Д: Издательство ЮФУ, 2011. – 90 с. ISBN 978-5-9275-0887-7

60. Колпаков, В. Ф. Экономико-математическое и эконометрическое моделирование: компьютерный практикум : учеб. пособие / В.Ф. Колпаков. – М. : ИНФРА-М, 2018. — 396 с. – (Высшее образование: Бакалавриат). – www.dx.doi.org/10.12737/24417. - ISBN 978-5-16-010967-1

61. Гусева, Е. Н. Экономическо-математическое моделирование [Электронный ресурс] : Уч. пособ. / Е. Н. Гусева. – 2-е изд., стереотип. - Москва : Флинта : МПСИ, 2011. – 216 с. – ISBN 978-5-89349-976-6 (Флинта), ISBN 978-5-9770-0256-1 (МПСИ)

62. Кундышева, Е. С. Экономико-математическое моделирование : учебник / Е. С. Кундышева ; под ред. Б. А. Суслакова. – 4-е изд. - Москва : Дашков и К°, 2012. – 424 с. - ISBN 978-5-394-01716-2

63. Введение в математическое моделирование : учебное пособие / В. Н. Ашихмин, М. Б. Гитман, И. Э. Келлер [и др.] ; под. ред. П. В. Трусова. – Москва : Логос, 2020. – 440 с. – ISBN 978-5-98704-637-1

64. Орлова, И. В. Экономико-математическое моделирование: практическое пособие по решению задач / И.В. Орлова, М.Г. Бич. – 3-е изд., испр. и доп. – Москва : Вузовский учебник : ИНФРА-М, 2023. – 190 с. – ISBN 978-5-9558-0527-6

65. Федосеев, В.В. Математическое моделирование в экономике и социологии труда. Методы, модели, задачи: учеб. пособие для студентов вузов, обучающихся по специальностям 080104 «Экономика труда», 080116 «Математические методы в экономике» / В.В. Федосеев. – М.: ЮНИТИ-ДАНА, 2017. – 167 с. – ISBN 978-5-238-01114-8

66. Математическое моделирование и проектирование : учеб. пособие / А.С. Коломейченко, И.Н. Кравченко, А.Н. Ставцев, А.А. Полухин ; под ред. А.С. Коломейченко. – М.: ИНФРА-М, 2018. – 181 с. – (Высшее образование: Магистратура). – www.dx.doi.org/10.12737/textbook_59688803c3cb35.15568286. – ISBN 978-5-16-012890-0

67. Гаврилов, Л. П. Информационные технологии в коммерции : учебное пособие / Л.П. Гаврилов. – 2-е изд., перераб. и доп. – Москва :

ИНФРА-М, 2022. – 369 с. + Доп. материалы [Электронный ресурс]. – (Высшее образование: Бакалавриат). – DOI 10.12737/1085795. – ISBN 978-5-16-016187-7

68. Математическое моделирование и проектирование : учебное пособие / А.С. Коломейченко, И.Н. Кравченко, А.Н. Ставцев, А.А. Полухин ; под ред. А.С. Коломейченко. – М.: ИНФРА-М, 2021. – 181 с. – (Среднее профессиональное образование). – ISBN 978-5-16-015651-4

69. Тарасик, В. П. Математическое моделирование технических систем : учебник / В.П. Тарасик. – Минск : Новое знание ; Москва : ИНФРА-М, 2022. – 592 с. – (Высшее образование: Бакалавриат). – ISBN 978-5-16-011996-0

70. Назарова, Ю. Н. Математическое моделирование в экономике : практикум : специальность : 38.05.01 «Экономическая безопасность». Специализация : «Судебная экономическая экспертиза» / Ю. Н. Назарова. – Волгоград : ФГБОУ ВО Волгоградский ГАУ, 2019. – 68 с.

71. Введение в математическое моделирование : учебное пособие / В. Н. Ашихмин, М. Б. Гитман, И. Э. Келлер [и др.] ; под. ред. П. В. Трусова. – Москва : Логос, 2020. – 440 с. – ISBN 978-5-98704-637-1

72. Математическое моделирование и проектирование : учебное пособие / А.С. Коломейченко, И.Н. Кравченко, А.Н. Ставцев, А.А. Полухин ; под ред. А.С. Коломейченко. – Москва : ИНФРА-М, 2021. – 181 с. – (Среднее профессиональное образование). – ISBN 978-5-16-015651-4

73. Колпаков, В. Ф. Экономико-математическое и эконометрическое моделирование: компьютерный практикум : учеб. пособие / В.Ф. Колпаков. – М.: ИНФРА-М, 2018. – 396 с. – (Высшее образование: Бакалавриат). – www.dx.doi.org/10.12737/24417. – ISBN 978-5-16-0109671

74. Белько, И. В. Теория вероятностей, математическая статистика, математическое программирование : учебное пособие / И.В. Белько, И.М. Морозова, Е.А. Криштапович. – М.: ИНФРА-М, 2022. – 299 с. : ил. – (Высшее образование: Бакалавриат). – ISBN 978-5-16-011748-5

75. Михалева, М. Ю. Математическое моделирование и количественные методы исследований в менеджменте : учеб. пособие / М.Ю. Михалева, И.В. Орлова. – М.: Вузовский учебник : ИНФРА-М, 2018. – 296 с. – (Высшее образование: Магистратура). —

www.dx.doi.org/10.12737/textbook_5b03f73021f562.03199866. - ISBN 978-5-9558-0607-5

76. Орлова, И. В. Экономико-математическое моделирование: практическое пособие по решению задач / И.В. Орлова, М.Г. Бич. – 3-е изд., испр. и доп. – М.: Вузовский учебник : ИНФРА-М, 2023. – 190 с. - ISBN 978-5-9558-0527-6

77. Анализ и диагностика финансово-хозяйственной деятельности предприятия : учебник / под ред. А.П. Гарнова. — Москва : ИНФРА-М, 2023. – 366 с. + Доп. материалы [Электронный ресурс]. — (Высшее образование: Бакалавриат). – DOI 10.12737/8240. - ISBN 978-5-16-009995-8

78. Погорелова, М. Я. Экономический анализ: теория и практика: Учебное пособие / Погорелова М.Я. – М.:ИЦ РИОР, НИЦ ИНФРА-М, 2018. – 290 с.(Высшее образование: Бакалавриат). - ISBN 978-5-369-01295-6

79. Гулай, Т.А. Теория вероятностей и математическая статистика [Электронный ресурс]: учебное пособие / Т.А. Гулай, А.Ф. Долгополова, Д.Б. Литвин, С.В. Мелешко. – 2-е изд., доп. – Ставрополь: АГРУС, 2013. – 260 с.

80. Пахунова, Р. Н. Общая и прикладная статистика : учебник для студентов высшего профессионального образования / П.Ф. Аскеров, Р.Н. Пахунова, А.В. Пахунов; под общ. ред. Р.Н. Пахуновой. – Москва : ИНФРА-М, 2022. — 272 с. + Доп. материалы [Электронный ресурс]. – (Высшее образование: Бакалавриат). – DOI 10.12737/748. – ISBN 978-5-16-006669-1

81. Прикладная математическая статистика : учебное пособие / сост. А. А. Мицель. - Томск : Томский государственный университет систем управления и радиоэлектроники, 2016. - 113 с. - Текст : электронный. – URL: https://znanium.com/catalog/product/1845838 (дата обращения: 27.03.2023)

82. Полякова, В. В. Прикладная статистика: методы анализа эмпирической информации : учебно-методическое пособие / В. В. Полякова, Н. В. Шаброва ; Министерство науки и высшего образования Российской Федерации, Уральский федеральный университет им. первого Президента России Б. Н. Ельцина. – Екатеринбург : Изд-во Уральского ун-та, 2020. – 188 с. – ISBN 978-5-7996-3021-8

83. Зайцев, В. М. Прикладная медицинская статистика: учебное пособие / В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин. – 2-е изд. – Санкт-Петербург : ООО «Издательство ФОЛИАНТ», 2006. – 432 с. - ISBN 5-93929-135-X

84. Григорьев, А. А. Методы и алгоритмы обработки данных : учебное пособие / А.А. Григорьев, Е.А. Исаев. –2-е изд., перераб. и доп. — Москва : ИНФРА-М, 2022. – 383 с. + Доп. материалы [Электронный ресурс]. – (Высшее образование: Бакалавриат). – DOI 10.12737/1032305. – ISBN 978-5-16-015581-4

85. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход/ЛемешкоБ.Ю., ЛемешкоС.Б., ПостоваловС.Н. и др. – Новосибирск : НГТУ, 2011. – 888 с.: ISBN 978-5-7782-1590-0.

86. Эконометрика: учебник / В.Н. Афанасьев, Т.В. Леушина, Т.В. Лебедева, А.П. Цыпин; под ред. проф. В.Н. Афанасьева; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2012. – 402 с.

87. Лемешко, Б. Ю. Критерии проверки гипотез о случайности и отсутствии тренда. Руководство по применению : монография / Б.Ю. Лемешко, И.В. Веретельникова. – Москва : ИНФРА-М, 2021. – 221 с. – (Научная мысль). – DOI 10.12737/1587437. – ISBN 978-5-16-017054-1

88. Теория статистики : учебник / под ред. проф. Г.Л. Громыко. – 4-е изд., перераб. и доп. — Москва : ИНФРА-М, 2021. – 465 с. – (Высшее образование: Бакалавриат). – DOI 10.12737/textbook_5d0734d6e23853.79720708. – ISBN 978-5-16-014914-1

89. Альсова, О. К. Исследование временных рядов в среде R : учебное пособие / О. К. Альсова. – Новосибирск : Изд-во НГТУ, 2021. – 88 с. – ISBN 978-5-7782-4337-8

90. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование: Учебник. – М.: Финансы и статистика, 2001 – 228 с. – ISBN 5-279-02419-8

91. Каяйкина М. С. Статистические методы изучения урожайности (на примере совхоза Ленинградской области). –Л.,1969. 106 с.

92. Манелля, А. И. Измерение устойчивости производства продукции земледелия // Статистический анализ развития АПК. – М.: Наука, 1992. С. 60-73.

93. Двойцова, И. Н. Основы математического моделирования социально-экономических процессов : учебное пособие / И. Н. Двойцова. – Железногорск : ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России, 2022. – 112 с.

94. Валентинов, В. А. Эконометрика / Валентинов В.А., – 3-е изд. – Москва :Дашков и К, 2016. – 436 с.: ISBN 978-5-394-02111-4

95. Крянев, А. В. Эконометрика (продвинутый уровень): Конспект лекций / Крянев А.В. – Москва :КУРС, НИЦ ИНФРА-М, 2017. – 62 с. –ISBN 978-5-906818-62-1

96. Балдин К.В., Быстров О.Ф., Соколов М.М. Эконометрика: Учеб. пособие для вузов. – 2-е изд., перераб. и доп. – М. : ЮНИТИ-ДАНА, 2017. – 254 с. – ISBN 978-5-238-00702-7

97. Середа, В. А. Эконометрика : учебное пособие / В. А. Середа, А. В. Литаврин, Н. Л. Собачкина. – Красноярск : Сиб. федер. ун-т, 2018. – 148 с. – ISBN 978-5-7638-3996-8

98. Едронова, В. Н. Экономический анализ развития территорий : учебник / В.Н. Едронова. – Москва : Магистр : ИНФРА-М, 2023. – 328 с. - ISBN 978-5-9776-0547-2

99. Пахунова, Р. Н. Общая и прикладная статистика : учебник для студентов высшего профессионального образования / П.Ф. Аскеров, Р.Н. Пахунова, А.В. Пахунов ; под общ. ред. Р.Н. Пахуновой. – Москва : ИНФРА-М, 2022. – 272 с. + Доп. материалы [Электронный ресурс]. – (Высшее образование: Бакалавриат). – DOI 10.12737/748. - ISBN 978-5-16-006669-1

100. Сергеева, И. И. Статистика : учебник / И.И. Сергеева, Т.А. Чекулина, С.А. Тимофеева. – 2-е изд., испр. и доп. – М.: ФОРУМ : ИНФРА-М, 2021. – 304 с. – (Среднее профессиональное образование). – ISBN 978-5-8199-0888-4

101. Лысенко, С. Н. Общая теория статистики : учебное пособие / С. Н. Лысенко, И. А. Дмитриева. – изд. испр. и доп. – М.: Вузовский учебник : ИНФРА-М, 2022. – 219 с. - ISBN 978-5-9558-0115-5

102. Ефимова, М. Р. Общая теория статистики: Учебник / М.Р. Ефимова, Е.В. Петрова, В.Н. Румянцев. – 2-е изд., испр. и доп. – М.: ИНФРА-М, 2011. – 416 с. (Высшее образование). ISBN 978-5-16-004265-7.

103. Рыжикова, Т. Н. Маркетинг: экономика, финансы, контроллинг : учебное пособие / Т.Н. Рыжикова. – Москва : ИНФРА-М, 2023. –

225 с. – (Высшее образование: Бакалавриат). — DOI 10.12737/24399. - ISBN 978-5-16-012515-2

104. Статистическое моделирование и прогнозирование : учеб. пособие / Е. М. Марченко [и др.] ; Владим. гос. ун-т им А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2018. – 100 с. ISBN 978-5-9984-0861-8

105. Дайитбегов, Д. М. Компьютерные технологии анализа данных в эконометрике: Монография / Д.М. Дайитбегов. – 3-е изд., испр. и доп. – М.: Вузовский учебник: НИЦ Инфра-М, 2018. – XIV, 587 с.: – (Научная книга). – ISBN 978-5-9558-0275-6

106. Айвазян, С. А. Эконометрика – 2: продвинутый курс с приложениями в финансах: Учебник / Айвазян С.А., Фантаццини Д. – М.: Магистр, НИЦ ИНФРА-М, 2018. – 944 с. – ISBN 978-5-9776-0333-1

107. Буреева Н.Н. Многомерный статистический анализ с использованием ППП «Statistica». Учебно-методический матриал по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики. – Нижний Новгород, 2007 – 112 с.

108. А.В. Кугаевских, Д.И. Муромцев, О.В. Кирсанова. Классические методы машинного обучения. – СПб: Университет ИТМО, 2022. – 53 с.

109. Чубукова, И. А. Data Mining : курс лекций / И. А. Чубукова. - Москва : ИНТУИТ, 2016. - 337 с. - ISBN 978-5-94774-819-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2136992 (дата обращения: 26.03.2024).

110. Обеспечение экономической безопасности предприятия [Электронный ресурс] : учеб. пособие / авт.-сост.: С. А. Грачев, М. А. Гундорова ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2022. – 420 с. – ISBN 978-5-9984-1666-8

ПРИЛОЖЕНИЯ

Приложение 1

Интегральная функция нормированного нормального распределения

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

Z	0,08	0,06	0,04	0,02	0
-3,5	0,00017	0,00019	0,0002	0,00022	0,00023
-3,4	0,00025	0,00027	0,00029	0,00031	0,00034
-3,3	0,00036	0,00039	0,00042	0,00045	0,00048
-3,2	0,00052	0,00056	0,0006	0,00064	0,00069
-3,1	0,00074	0,00079	0,00085	0,0009	0,00097
-3	0,00104	0,00111	0,00118	0,00126	0,00135
-2,9	0,0014	0,0015	0,0016	0,0017	0,0019
-2,8	0,002	0,0021	0,0023	0,0024	0,0026
-2,7	0,0027	0,0029	0,0031	0,0033	0,0035
-2,6	0,0037	0,0039	0,0041	0,0044	0,0047
-2,5	0,0049	0,0052	0,0055	0,0059	0,0062
-2,4	0,0066	0,0069	0,0073	0,0078	0,0082
-2,3	0,0087	0,0091	0,0096	0,0102	0,0107
-2,2	0,0113	0,0119	0,0125	0,0132	0,0139
-2,1	0,0146	0,0154	0,0162	0,017	0,0179
-2	0,0188	0,0197	0,0207	0,0217	0,0228
-1,9	0,0239	0,025	0,0262	0,0274	0,0287
-1,8	0,0301	0,0314	0,0329	0,0344	0,0359
-1,7	0,0375	0,0392	0,0409	0,0427	0,0446
-1,6	0,0465	0,0485	0,0505	0,0526	0,0548
-1,5	0,0571	0,0594	0,0618	0,0643	0,0668
-1,4	0,0694	0,0721	0,0749	0,0778	0,0808
-1,3	0,0838	0,0869	0,0901	0,0934	0,0968
-1,2	0,1003	0,1038	0,1075	0,1112	0,1151
-1,1	0,119	0,123	0,1271	0,1314	0,1357
-1	0,1401	0,1446	0,1492	0,1539	0,1587
-0,9	0,1635	0,1685	0,1736	0,1788	0,1841
-0,8	0,1894	0,1949	0,2005	0,2061	0,2119
-0,7	0,2177	0,2236	0,2297	0,2358	0,242
-0,6	0,2483	0,2546	0,2611	0,2676	0,2743
-0,5	0,281	0,2877	0,2946	0,3015	0,3085
-0,4	0,3156	0,3228	0,33	0,3372	0,3446
-0,3	0,352	0,3594	0,3669	0,3745	0,3821
-0,2	0,3897	0,3974	0,4052	0,4129	0,4207
-0,1	0,4286	0,4364	0,4443	0,4522	0,4602
0	0,4681	0,4761	0,484	0,492	0,5
z	0	0,02	0,04	0,06	0,08

0	0,5	0,508	0,516	0,5239	0,5319
0,1	0,5398	0,5478	0,5557	0,5636	0,5714
0,2	0,5793	0,5871	0,5948	0,6026	0,6103
0,3	0,6179	0,6225	0,6331	0,6406	0,648
0,4	0,6554	0,6628	0,67	0,6772	0,6844
0,5	0,6915	0,6985	0,7054	0,7123	0,719
0,6	0,7257	0,7324	0,7389	0,7454	0,7517
0,7	0,758	0,7642	0,7704	0,7764	0,7823
0,8	0,7881	0,7939	0,7995	0,8051	0,8106
0,9	0,8159	0,8212	0,8264	0,8315	0,8365
1	0,8413	0,8461	0,8505	0,8554	0,8599
1,1	0,8643	0,8686	0,8729	0,877	0,881
1,2	0,8849	0,8888	0,8925	0,8962	0,8997
1,3	0,9032	0,9066	0,9099	0,9131	0,9162
1,4	0,9192	0,9222	0,9251	0,9279	0,9306
1,5	0,9332	0,9357	0,9382	0,9406	0,9429
1,6	0,9452	0,9474	0,9495	0,9515	0,9535
1,7	0,9554	0,9573	0,9591	0,9608	0,9625
1,8	0,9641	0,9656	0,9671	0,9686	0,9699
1,9	0,9713	0,9726	0,9738	0,975	0,9761
2	0,9773	0,9783	0,9793	0,9803	0,9812
2,1	0,9821	0,983	0,9838	0,9846	0,9854
2,2	0,9861	0,9868	0,9875	0,9881	0,9887
2,3	0,9893	0,9898	0,9904	0,9909	0,9913
2,4	0,9918	0,9922	0,9927	0,9931	0,9934
2,5	0,9938	0,9941	0,9945	0,9943	0,9951
2,6	0,9953	0,9956	0,9959	0,9961	0,9963
2,7	0,9965	0,9967	0,9969	0,9971	0,9973
2,8	0,9974	0,9976	0,9977	0,9979	0,998
2,9	0,9981	0,9983	0,9984	0,9985	0,9986
3	0,99865	0,99874	0,99882	0,99889	0,99896
3,1	0,99903	0,9991	0,99915	0,99921	0,99926
3,2	0,99931	0,99936	0,9994	0,99954	0,99948
3,3	0,99952	0,99955	0,99958	0,99961	0,99964
3,4	0,99966	0,99969	0,99971	0,99973	0,99975
3,5	0,99977	0,99978	0,9998	0,99981	0,99983

Приложение 2

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0.10	0.05	0.02	0.01	0.002	0.001
1	6.31	12.7	31.82	63.7	318.3	637.0
2	2.92	4.30	6.97	9.92	22.33	31.6
3	2.35	3.18	4.54	5.84	10.22	12.9
4	2.13	2.78	3.75	4.60	7.17	8.61
5	2.01	2.57	3.37	4.03	5.89	6.86
6	1.94	2.45	3.14	3.71	5.21	5.96
7	1.89	2.36	3.00	3.50	4.79	5.40
8	1.86	2.31	2.90	3.36	4.50	5.04
9	1.83	2.26	2.82	3.25	4.30	4.78
10	1.81	2.23	2.76	3.17	4.14	4.59
11	1.80	2.20	2.72	3.11	4.03	4.44
12	1.78	2.18	2.68	3.05	3.93	4.32
13	1.77	2.16	2.65	3.01	3.85	4.22
14	1.76	2.14	2.62	2.98	3.79	4.14
15	1.75	2.13	2.60	2.95	3.73	4.07
16	1.75	2.12	2.58	2.92	3.69	4.01
17	1.74	2.11	2.57	2.90	3.65	3.95
18	1.73	2.10	2.55	2.88	3.61	3.92
19	1.73	2.09	2.54	2.86	3.58	3.88
20	1.73	2.09	2.53	2.85	3.55	3.85
21	1.72	2.08	2.52	2.83	3.53	3.82
22	1.72	2.07	2.51	2.82	3.51	3.79
23	1.71	2.07	2.50	2.81	3.59	3.77
24	1.71	2.06	2.49	2.80	3.47	3.74
25	1.71	2.06	2.49	2.79	3.45	3.72
26	1.71	2.06	2.48	2.78	3.44	3.71
27	1.71	2.05	2.47	2.77	3.42	3.69
28	1.70	2.05	2.46	2.76	3.40	3.66
29	1.70	2.05	2.46	2.76	3.40	3.66
30	1.70	2.04	2.46	2.75	3.39	3.65
40	1.68	2.02	2.42	2.70	3.31	3.55
60	1.67	2.00	2.39	2.66	3.23	3.46
120	1.66	1.98	2.36	2.62	3.17	3.37
∞	1.64	1.96	2.33	2.58	3.09	3.29
	0.05	0.025	0.01	0.005	0.001	0.0005
	Уровень значимости α (односторонняя критическая область)					

Приложение 3

Критические точки распределения Фишера

(k_1 и k_2 — число степеней свободы большей и меньшей дисперсии соответственно)

Уровень значимости α =0.01

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98.49	99.01	90.17	99.25	99.33	99.30	99.34	99.36	99.36	99.40	99.41	99.42
3	34.12	30.81	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37
5	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89
6	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47
8	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67
9	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11
10	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71
11	9.86	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40
12	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16
13	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45

Уровень значимости α =0.05

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18.5	19.00	19.16	19.25	19:30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38

Приложение 4

Таблица 5%-го и 1%-го уровней вероятности коэффициентов корреляции (r_a)

Размер выборки	Положительные значения		Отрицательные значения	
	5%-й уровень	1 %-й уровень	5%-й уровень	1 %-й уровень
5	0,253	0,297	-0,753	-0,798
6	0,354	0,447	-0,708	-0,863
7	0,370	0,510	-0,674	-0,799
8	0,371	0,531	-0,625	-0,764
9	0,366	0,533	-0,593	-0,737
10	0,360	0,525	-0,564	-0,705
11	0,353	0,515	-0,539	-0,679
12	0,348	0,505	-0,516	-0,655
13	0,341	0,495	-0,497	-0,634
14	0,335	0,485	-0,479	-0,615
15	0,328	0,475	-0,462	-0,597
20	0,299	0,432	-0,399	-0,524
25	0,276	0,398	-0,356	-0,473
30	0,257	0,370	-0,324	-0,433
35	0,242	0,347	-0,300	-0,401
40	0,229	0,329	-0,279	-0,376
45	0,218	0,313	-0,262	-0,256
50	0,208	0,301	-0,248	-0,339

Приложение 5

Распределение критерия Дарбина-Уотсона для положительной автокорреляции
(для 5%-го уровня значимости)

n	$V=1$		$V=2$		$V=3$		$V=4$		$V=5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
15	1,08	1,36	0,95	,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,89
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,63	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,55	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Учебное электронное издание

ГУНДОРОВА Марина Александровна
ГРАЧЕВ Сергей Александрович

ПАКЕТЫ ПРИКЛАДНЫХ СТАТИСТИЧЕСКИХ ПРОГРАММ

Учебное пособие

Издается в авторской редакции

Системные требования: Intel от 1,3 ГГц; Windows XP/7/8/10;
Adobe Reader; дисковод CD-ROM.

Тираж 25 экз.

Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых
Изд-во ВлГУ
rio.vlgu@yandex.ru

Институт экономики и туризма
кафедра экономики инноваций и финансов
mg82.82@mail.ru