Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени первого Президента России Б. Н. Ельцина»

УТВЕРЖДАЮ

Директор по образовательной деятельности

_____ С.Т. Князев
«7» сентября 2023 г.

Методы статистики для разработчиков

Учебно-методические материалы по направлению подготовки
09.03.03 Прикладная информатика
Образовательная программа «Прикладной искусственный интеллект»

Екатеринбург

РАЗРАБОТЧИКИ УЧЕБНО-МЕТОДИЧЕСКИХ МАТЕРИАЛОВ

Доцент, физ.-мат. наук  Белоусова Вероника Игоревна

СОДЕРЖАНИЕ

Введение в дисциплину

В современном сверхсвязанном мире данные генерируются и потребляются невиданными ранее темпами. Дата-сайентисты должны быть обучены использованию статистических методов не только для интерпретации цифр, но и для выявления таких злоупотреблений и защиты людей от введения в заблуждение. Немногие специалисты по статистике имеют формальную подготовку.

Рассмотрим основные вопросы курса:

- Что такое статистика?
- Статистика в отношении к машинному обучению.
- Зачем вам нужно осваивать статистику.
- Какому учебному плану следовать, чтобы освоить эти темы.
- Как изучать статистику, чтобы стать практиком, а не просто человеком, который правильно сдаёт тесты.
- Практические советы и обучающие ресурсы.

Что такое статистика?

Статистика — это набор математических методов и инструментов, позволяющих ответить на важные вопросы о данных.

Она делится на две категории:

Описательная статистика. Предлагает методы резюмирования данных путем преобразования необработанных наблюдений в значимую информацию, которую легко интерпретировать и распространять.

Логическая статистика. Предлагает методы изучения экспериментов, выполненных на маленьких образцах данных, и умозаключения для всей популяции (всего домена).

Сегодня статистика и машинное обучение — две тесно связанные между собой области. Статистика дает важные предпосылки для прикладного машинного обучения: она помогает выбирать, оценивать и интерпретировать модели прогнозирования.

Статистика в машинном обучении

В основе машинного обучения лежит статистика. Невозможно решить реальные проблемы с помощью машинного обучения, если вы не обладаете хорошим знанием основ статистики.

Конечно, имеются некоторые факторы, затрудняющие обучение статистике: математические уравнения, греческая нотация и тщательно выверенные понятия, затрудняющие развитие интереса к предмету. Можно решить эти проблемы с помощью простых и ясных объяснений, учебных пособий с соответствующим темпом и практических занятий — решения проблем с помощью прикладных методов статистики. От исследовательского анализа данных до разработки экспериментов для проверки гипотез статистика играет ключевую роль в решении проблем во всех основных отраслях и областях.

Тот, кто хочет развить глубокое понимание машинного обучения, должен узнать, как статистические методы формируют основу алгоритмов регрессии и классификации, как статистика позволяет учиться на основе данных и как она помогает извлекать смысл из немаркированных данных.

Зачем вам осваивать статистику?

Каждая организация стремится стать управляемой данными. Вот почему мы наблюдаем такой рост спроса на дата-сайентистов и аналитиков. Сегодня, чтобы решить проблемы, ответить на вопросы и наметить стратегию, нужно разобраться в данных. К счастью, статистика предлагает набор инструментов для получения этих знаний.

От данных к знаниям

Сами по себе сырые наблюдения — это просто данные. Чтобы трансформировать наблюдения в имеющие смысл идеи, применяется описательная статистика. Затем возможно применить логическую статистику, чтобы изучить небольшие выборки данных и дать схему с выводами для экстраполяции результатов на всю совокупность данных.

Статистика помогает ответить на вопросы, подобные этим:

- Какие из признаков наиболее важны?
- Как проектировать эксперимент, чтобы разработать стратегию продукта?
- Какие показатели производительности мы должны измерять?
- Какой самый распространенный и ожидаемый результат?
- Как отличить шум от достоверных данных?

Это важные и общие вопросы, на которые ежедневно приходится отвечать работающим с данными командами. Ответы на эти вопросы помогают эффективно принимать решения. Статистические методы помогают нам не только настраивать проекты прогнозного моделирования, но и интерпретировать результаты.

Статистика и проекты по машинному обучению

Почти каждый состоит из перечисленных ниже задач. И статистика играет в той или иной форме центральную роль во всех этих задачах.

Уточнение постановки проблемы

Наиболее важной частью прогностического моделирования является фактическое определение проблемы, дающее реальную цель, к которой мы должны стремиться. Это помогает определить тип проблемы, с которой мы имеем дело (то есть регрессия это или классификация), а также помогает в определении структуры и типов входных, выходных данных и метрик с учетом поставленной задачи. Но подстановка проблем не всегда проста. Если вы новичок в машинном

обучении, она может потребовать значительного изучения наблюдений в вашей области. Два основных понятия, которые необходимо освоить здесь — это экспериментальный анализ данных (EDA) и добыча данных (Data Mining).

Первоначальное исследование данных

Исследование данных включает в себя получение глубокого понимания как распределения переменных, так и отношений между переменными в ваших данных.

Отчасти знание домена помогает овладеть определённым типом переменных. Тем не менее как эксперты, так и новички в этой области извлекают пользу из реальной работы с реальными наблюдениями в домене. Важные связанные с этим понятия в статистике сводятся к изучению описательной статистики и визуализации данных.

Очистка данных

Часто точки данных, собранные из эксперимента или хранилища данных, являются нетронутыми. Данные могли быть подвергнуты процессам или манипуляциям, которые повредили их целостность. Это еще больше влияет на последующие процессы или использующие такие данные модели. Распространённые примеры — пропущенные значения, повреждение данных, ошибки в данных (из-за плохого датчика), а также не приведённые к единой форме данные (наблюдения с разными масштабами). Если вы хотите освоить методы очистки, изучите выявление отклонений и вменение отсутствующих значений.

Подготовка данных и настройка конвейера преобразования

Если данные содержат ошибки и несоответствия, часто нельзя применять их в моделировании. Во-первых, данным, возможно, придётся пройти через набор преобразований, чтобы изменить форму или структуру и сделать их более подходящими для определённой вами задачи, или используемых алгоритмов обучения. Затем можно разработать конвейер таких преобразований, который

будет применяться к данным для получения последовательных и совместимых входных данных для модели. Вы должны овладеть такими понятиями, как методы выборки данных и отбора признаков, преобразование данных, их масштабирование и кодирование.

Выбор и оценка модели

Ключевым шагом в решении прогностической проблемы являются выбор и оценка метода обучения. Оценочная статистика поможет вам оценить прогнозы модели на данных, которые модель не видела.

Проектирование экспериментов — это подраздел статистики, который управляет процессом выбора и оценки модели. Он требует хорошего понимания проверки статистических гипотез и оценочной статистики.

Тонкая настройка модели

Почти в каждом алгоритме машинного обучения имеется набор гиперпараметров, которые позволяют настроить метод обучения под выбранную вами постановку задачи. Эта гиперпараметрическая настройка часто носит эмпирический, но не аналитический характер. Для оценки влияния различных настроек гиперпараметра на производительность модели требуются большие наборы экспериментов.

Основные навыки в статистике

- Определение вопроса, на который можно ответить статистически, чтобы принимать эффективные решения.
- Вычисление и интерпретация общих статистических данных и использование стандартных методов визуализации данных для передачи результатов.

- Понимание того, как математическая статистика применяется в конкретной области, такие понятия, как центральная предельная теорема и закон больших чисел.
- Умение делать выводы из оценок местоположения и изменчивости (ANOVA).
- Определение связи между целевыми и независимыми переменными.
- Разработка экспериментов по проверке статистических гипотез, A/B тестирование и т. д.
- Вычисление и интерпретация метрик производительности, таких как p-значение, альфа, ошибки первого и второго рода и т. д.

Важные понятия статистики

- Приступая к освоению статистики, нужно понимать типы данных (данные в прямоугольной системе координат и другие данные), оценивать местоположение и вариабельность распределения данных, бинарные и категориальные данные, корреляцию, отношение между различными типами переменных.
- Статистические распределения — случайные числа, закон больших чисел, центральная предельная теорема, стандартная погрешность и т. д.
- Выборка и распределение данных — случайная выборка, смещение выборки, смещение выбора, распределение выборки, бутстрэп, доверительный интервал, нормальное распределение, t-распределение, биномиальное распределение, распределение «хи квадрат», F-распределение, распределение Пуассона и экспоненциальное распределение.
- Статистические эксперименты и и тестирование значимости — A/B тестирование, проведение проверки гипотез (нулевая и альтернативная гипотезы), ресемплирование, статистическая значимость, доверительный интервал, p-значение, альфа [прим. перев. — максимальный шанс допустить

ошибку первого рода], t-критерии, степени свободы, выводы из оценок местоположения и изменчивости, критические значения, ковариантность и корреляция, величина эффекта, статистическая мощность.

- Непараметрические статистические методы — ранжирование данных, критерии нормальности, нормализация данных, ранговая корреляция, критерии знаковых рангов, критерий независимости.

Введение в статистический анализ данных. Базовые понятия статистического анализа данных. Подходы к статистическому анализу данных

Этапы статистической обработки

– Предварительный статистический анализ данных

– Оценка закона распределения. Непараметрический подход

– Оценка закона распределения. Параметрический подход

– Восстановление пропущенных значений и анализ выбросов

– Унификация признакового описания

Базовые понятия статистического анализа данных

Этапы работ, предшествующие обработке экспериментальных данных

Всех специалистов, профессионально занимающихся обработкой статистических данных, условно можно разделить на три категории:

1) приверженцы классической математической статистики (объектами их исследований обычно являются некоторые разделы биологии или физики);

2) представители школы обработки экспериментальных данных в рамках идеологии исследования операций (предметом их разработок чаще всего бывают результаты активных экспериментов над сложной технической системой);

3) специалисты по прикладной статистике и анализу данных, ориентированные на исследование естественных и социальных систем в таких, например, областях, как геология, медицина, экономика и социология.

Характер данных и методологическое видение проблемного материала во всех трёх случаях столь различны, что в действительности эти три течения статистических исследований следовало бы признать самостоятельными. В настоящей лекции за основу принята концепция по отношению к прикладной статистике и анализу данных, окончательно сформировавшаяся к концу 80-х годов. Наиболее полно эта область прикладной математики изложена в трёхтомном справочном издании по прикладной статистике под редакцией С.А. Айвазяна. В текстах лекций использована концепция стиля подачи материала упомянутого выше справочника.

Прикладная статистика

Целесообразность введения термина прикладная статистика наряду с привычным понятием математическая статистика объясняется тем, что для внедрения метода статистической обработки необходимо дополнительно провести сложную и наукоемкую работу. Условно разобьем её на ряд этапов:

1) адекватно «приложить» исходные модельные допущения к реальной задаче;

2) представить имеющуюся исходную информации (физические сигналы, геологические срезы и др.) в стандартной форме;

3) разработать вычислительный алгоритм и его программное обеспечение;

4) организовать удобный режим общения с ЭВМ в процессе решения задачи. Весь комплекс вышеперечисленных действий и составляет содержание прикладной статистики

Прикладная статистика – это самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы

и модели, предназначенные для организации сбора, стандартной записи, обработки статистических данных с целью их удобного представления (в том числе и на ЭВМ), интерпретации и получения научных и практических выводов.

Заметим, что некоторые специалисты, в частности, французские, вместо введенного термина «прикладная статистика» используют понятие «анализ данных», трактуя его в расширительном смысле.

Идеи и методологические принципы многомерного статистического анализа данных

Эффект существенной многомерности. Статистический анализ должен опираться одновременно на совокупность взаимосвязанных свойств объектов.

Возможность лаконичного объяснения природы анализируемых многомерных структур. На нем построены такие важнейшие разделы математического аппарата классификации и снижения размерности, как метод главных компонент и факторный анализ, многомерное шкалирование, целенаправленное проецирование в разведочном анализе данных и др.

Максимальное использование «обучения» в настройке математических моделей многомерного статистического анализа данных.

Оптимизационная формулировка задач многомерного статистического анализа данных.

Цели эксперимента в науке и промышленности

Экспериментальные методы широко используются как в науке, так и в промышленности, однако нередко с весьма различными целями. Обычно основная цель научного исследования состоит в том, чтобы показать статистическую значимость эффекта воздействия определенного фактора на изучаемую зависимую переменную. В условиях промышленного эксперимента основная цель обычно заключается в извлечении максимального количества объективной информации о влиянии изучаемых факторов на производственный процесс с помощью наименьшего числа дорогостоящих наблюдений. Если в научных приложениях

методы дисперсионного анализа используются для выяснения реальной природы взаимодействий, проявляющейся во взаимодействии факторов высших порядков, то в промышленности учет эффектов взаимодействия факторов часто считается излишним в ходе выявления существенно влияющих факторов.

Указанное отличие приводит к существенному различию методов, применяемых в науке и промышленности. Если просмотреть классические учебники по дисперсионному анализу, то обнаружится, что в них, в основном, обсуждаются планы с количеством факторов не более пяти (планы же с более чем шестью факторами обычно оказываются бесполезными). Основное внимание в данных рассуждениях сосредоточено на выборе общезначимых и устойчивых критериев значимости. Однако если обратиться к стандартным учебникам по экспериментам в промышленности, то окажется, что в них обсуждаются, в основном, многофакторные планы (например, с 16-ю или 32-мя факторами), в которых нельзя оценить эффекты взаимодействия, и основное внимание сосредоточивается на том получении несмещенных оценок главных эффектов (или, реже, взаимодействий второго порядка) с использованием наименьшего числа наблюдений.

Подходы к статистическому анализу данных

Возможные подходы к статистическому анализу данных

Развитие теории и практики статистической обработки данных шло в двух параллельных направлениях. Первое включает методы математической статистики, предусматривающие возможность классической вероятностной интерпретации анализируемых данных и полученных статистических выводов (вероятностный подход). Второе направление содержит статистические методы, которые априори не опираются на вероятностную природу обрабатываемых данных, т.е. остаются за рамками научной дисциплины «математическая статистика» (логико-алгебраический подход). Ко второму подходу исследователь вынужден обращаться лишь тогда, когда условия сбора исходных данных не укладываются в рамки статистического ансамбля, т.е. в ситуации, когда не имеется практической

или хотя бы принципиально мысленно представимой возможности многократного тождественного воспроизведения основного комплекса условий, при которых производились измерения анализируемых данных.

Типы реальных ситуаций с позиции выполнения требований статистического ансамбля

Выделяют три типа реальных ситуаций: с высокой работоспособностью вероятностно-статистических методов; с допустимостью вероятностно-статистических приложений (при этом нарушатся требования сохранения неизменными условия эксперимента); с недопустимостью вероятностно-статистических приложений (в этом случае идея многократного повторения одного и того же эксперимента в неизменных условиях является бессодержательной).

Сравнение подходов к статистическому анализу данных

Основные отличительные особенности подходов на примере задачи классификации представим схематично в таблице 1.

Таблица 1– Отличительные особенности подходов

Составляющие	Первое направление	Второе направление
Цели исследования	Выделение классов, как инвариантов в потоке выборочных объектов	Выяснение распределения данных в системе
Объекты и признаки.	Независимы	Зависимость предполагается, ее нужно обнаружить
Выделяемые классы	Характеризуются эталоном и не пересекаются	Четко не выделяются, т.е. пересекаются

Аппарат исследования	Вероятностный - преобразование пространства признаков (даже в одномерную ось)	Логико-комбинаторный

Первое направление развития анализа данных, ориентированное на технические области знания, отстаивает идею простоты используемых моделей. В рамках этого направления неудовлетворительные результаты объясняют отсутствием информативных признаков.

Второе направление развития анализа данных ориентировано на социально-экономическую и социологическую информацию. При ее обработке появилось много новых идей, в частности, идея поэтапной группировки и коллектива решающих правил. Разработаны методы многомерного шкалирования, экспертных оценок.

В отличие от первого примера во втором примере невозможно: интерпретировать исходные данные в качестве случайной выборки генеральной совокупности (в связи с неприятием главной идеи понятия статистического ансамбля: идея многократного повторения одного и того же эксперимента в неизменных условиях теряет смысл); использовать вероятностную модель для построения и выбора наилучших методов статистической обработки; дать вероятностную интерпретацию выводам, основанным на статистическом анализе исходных данных.

Но в обоих случаях выбор наилучшего из всех возможных методов обработки данных производится в соответствии с некоторыми функционалами качества метода. Способ обоснования выбора этого функционала, а также его интерпретация различны. В первом случае выбор основан на допущении о вероятностной природе исходных данных и интерпретация тоже. Во втором случае исследователь не пользуется априорными сведениями о вероятностной природе исходных данных и при обосновании выбора оптимального критерия качества опирается на соображения содержательного (физического) плана - как именно и для чего получены данные. Когда критерий выбран, в обоих случаях используются

методы решения экстремальных задач. На этапе осмысления и интерпретации каждый из подходов имеет свою специфику.

При выборе типа модели следует понимать, что всякая модель является упрощенным (математическим) представлением изучаемой действительности. Мера адекватности модели и действительности является решающим фактором работоспособности используемых затем методов обработки. А так как ни одна модель не может идеально соответствовать реальной ситуации, то желательна многократная обработка исходных данных для разных вариантов модели.

Этапы статистической обработки. Основные этапы статистической обработки экспериментальных данных

Опишем общую логическую схему статистического анализа данных в виде семи этапов, перечислив их в хронологическом порядке (хотя они могут реализовываться в режиме итерационного взаимодействия).

Этап 1 Исходный (предварительный) анализ исследуемой системы. На этом этапе определяются: основные цели исследования на неформализованном, содержательном уровне; совокупность единиц (объектов), представляющая предмет статистического исследования; набор параметров-признаков $(x^1,\ldots,x^p)$ для описания обследуемых объектов; степень формализации соответствующих записей при сборе данных; время и трудозатраты, объем работ; выделение ситуаций, требующих предварительной проверки перед составлением детального плана исследований; формализованная постановка задачи; в каком виде осуществляется сбор первичной информации и введение в ЭВМ.

Если обработка проводится с помощью существующего пакета статистической обработки, то трудоемкость этого этапа бывает сравнима с суммарной трудоемкостью остальных этапов.

Этап 2 Составление плана сбора исходной информации. При составлении детального плана сбора первичной информации необходимо учитывать, как и

для чего данные анализируются, т.е. учитывать полную схему анализа. Этот этап называют «организационно-методической подготовкой», так как на нем планируется: какой должна быть выборка – случайной, пропорциональной, расслоенной (если используется аппарат общей теории выборочных обследований); объем и продолжительность исследования; схема проведения активного эксперимента (в случае, если он возможен) с привлечением методов планирования эксперимента и регрессионного анализа для определения некоторых входных переменных.

Этап 3 Сбор исходных данных, их подготовка и введение в ЭВМ. Сбор исходных данных и введение их в ЭВМ, а также внесение в ЭВМ полного и краткого определения используемых терминов. Существует два вида представления исходных данных: матрица «объект-признак»: со значениями k-го признака, характеризующего i-й объект в момент t (числа, текст): $x_i^{(k)}(t)$, $t = t_1, \ldots t_N$, $k = \overline{(1, p)}$, $i = \overline{(1, N)}$; и матрица «объект-объект» $\rho_{ij}(t)$ - характеристик попарной близости i-го и j-го объектов (при этом m=N) или признаков (при этом m=p) в момент t. Второй вид представления часто используется в социологии, где данные собираются с помощью специальных опросников, анкет. Примером характеристики попарной близости признаков может служить ковариационная матрица.

Этап 4 Первичная статистическая обработка данных. При первичной статистической обработке данных обычно решаются следующие задачи: отображение вербальных переменных в номинальную (с предписанным числом градаций) или ординальную (порядковую) шкалу; статистическое описание исходных совокупностей с определением пределов варьирования переменных; анализ резко выделяющихся переменных; восстановление пропущенных значений наблюдений; проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных; унификация типов переменных, когда с помощью различных приёмов добиваются унифицированной записи всех переменных; экспериментальный анализ закона распределения исследуемой гене-

ральной совокупности и параметризация сведений о природе изучаемых распределений (эту разновидность первичной статистической обработки называют иногда процессом составления сводки и группировки); вычислительная реализация учета сложности задачи и возможностей ЭВМ; формулировка задачи на входном языке пакета статистической обработки.

Этап 5 Выбор основных методов и алгоритмов статистической обработки данных, составление детального плана вычислительного анализа материала. Составление детального плана вычислительного анализа. Определяются основные группы, для которых будет проводиться дальнейший анализ. Пополняется и уточняется тезаурус содержательных понятий. Описывается блок-схема анализа с указанием привлекаемых методов. Формируется оптимизационный критерий, по которому выбирается один из альтернативных методов.

Этап 6 Реализация плана вычислительного анализа исходных данных (непосредственная эксплуатация ЭВМ

Исследователь на этом этапе осуществляет управление вычислительным процессом, формирует задачу обработки и описания данных на входном языке пакета. Учитываются размерность задачи, алгоритмическая сложность вычислительного процесса, возможности ЭВМ, и особенности данных (обусловленность операций, надежность используемых оценок параметров).

Этап 7 Подведение итогов. Строится формальный отчет о проведенном исследовании. Интерпретируются результаты применения статистических процедур (оценки параметров, проверки гипотез, отображения в пространство меньшей размерности, классификации). При интерпретации могут использоваться методы имитационного моделирования.

Если исследование проводится в рамках первого подхода (см. п.1.2), то выводы формируются в терминах оценок неизвестных параметров, или в виде отчета о справедливости гипотез с указанием количественной степени достоверности. В случае второго подхода вероятностная интерпретация не делается.

Работа завершается содержательной формулировкой новых задач, вытекающих из проведенного исследования.

Основная цель разведочного анализа данных

Этап разведочного анализа данных (РАД) зачастую игнорируется или реализуется поверхностно в ходе прикладных статистических исследований. Одна из главных причин – отсутствие необходимой научно-методологической литературы. Большое внимание этим вопросам уделено в третьем томе справочника по прикладной статистике Айвазяна С.А. и др. Основная цель РАД – построить некоторую статистическую модель в виде эмпирического описания структуры данных, которую необходимо будет потом в ходе статистического исследования верифицировать. Основная задача РАД – переход к компактному описанию данных при возможно более полном сохранении существенных аспектов информации, содержащихся в данных.

Методы разведочного анализа данных

Методы разведочного (предмодельного) статистического анализа данных, направлены на «прощупывание» вероятностной и геометрической природы обрабатываемых данных и предназначены для формирования адекватных реальности рабочих исходных допущений, на которых строится дальнейшее исследование. РАД является необходимым и естественным моментом первичной статистической обработки и применяется, когда отсутствует априорная информация о статистическом или причинном механизме порождения имеющихся у исследователя данных.

Важнейшим элементом РАД является широкое использование визуального представления многомерных данных. Его возможности возросли благодаря появлению динамических форм визуального представления. Для этого многомерные данные отображаются в пространство низкой размерности с сохранением существенных структурных особенностей. При этом структура данных может

оказаться такой сложной, что небольшого числа проекций недостаточно для их представления. Тогда структуру описывают за счет агрегирования информации, содержащейся в большом числе низкоразмерных проекций.

К РАД относятся методы, дающие наглядное представление о структуре многомерных данных в пространствах малой размерности. В случае, если размерность пространства, куда отображаются данные, меньше или равно трем, то эти методы относятся к собственно разведочному анализу, когда по некоторому критерию при помощи вычислительной процедуры оптимизации ищут отображения, дающие наиболее выразительные проекции, а окончательное решение принимается визуально путем анализа (в одномерном случае – это гистограмма, на плоскости – диаграмма рассеивания).

К РАД относятся также методы, связанные с линейным проецированием, упрощением описания с помощью компонентного анализа и многомерного шкалирования, кластер-анализа, анализа соответствий (для неколичественных переменных).

Модели структуры многомерных данных в разведочном анализе данных

Пусть данные заданы в виде матрицы данных. Объекты можно представить в виде точек в многомерном (p-мерном) пространстве. Для описания структуры этого множества точек в РАД используется одна из следующих статистических моделей:

1- модель облака точек примерно эллипсоидальной конфигурации;

2- кластерная модель, т.е. совокупность нескольких «облаков» точек, достаточно далеко отстающих друг от друга;

3- модель «засорения» (компактное облако точек и при этом присутствуют дальние выбросы);

4- эмпирический образ данных в виде покрытия выборочных точек многомерного признакового пространства сетью гиперпараллелепипедов с оцененной плотностью распределения (многомерный аналог гистограммы);

5- модель носителя точек как многообразия (линейного или нелинейного) более низкой размерности, чем исходное: типичным примером является выборка из вырожденного распределения; в рамках этой модели можно рассматривать и регрессионную модель, когда соответствующие многообразие допускает функциональное представление $X_{11} = F(X_1) + \varepsilon$, где X_{11}- прогнозируемые, X_1-предсказывающие признаки, $F(X_1)$- функция регрессии, ε- ошибка.

6- дискриминантная модель, когда точки разделены на несколько групп и дана информация о их принадлежности к той или иной группе.

.

Предварительный статистический анализ данных

Любое экспериментальное исследование содержит этапы постановки задачи, планирования и проведения эксперимента, а также анализа и интерпретация результатов. Главной трудностью на этапе постановки задачи является переход с языка специальности на язык планирования эксперимента, на язык математики.

Содержательная постановка задач статистического **описания и прогноза** является переходной формулировкой, позволяющей перейти к математической, на основании выявленной цели исследования. Математическая постановка задач статистического **описания и прогноза** предполагает то, что формулировка задачи будет сделана в терминах, используемых в конкретной формальной дедуктивной системе.

Математическая постановка задач статистического описания предназначена для описания структуры множества выборочных точек и для формирования адекватных реальности рабочих исходных допущений, на которых строится дальнейшее исследование. В вероятностно-статистическом подходе математическая по-

становка задач статистического описания может состоять в оценке закона распределения. В логико-комбинаторном подходе, или в РАД используется одна из первых четырех статистических моделей: модель облака точек, кластерная модель, модель «засорения» и эмпирический образ данных

В общем виде задачу классификации исследуемой совокупности N объектов O={O_i}, $i = \overline{1,N}$, где для каждого объекта замерены значения p параметров, т.е. каждый объект O_i описан вектором $X_i = (x_i^1,\ldots,x_i^r)$, можно сформулировать как задачу поиска такого разбиения S заданной совокупности на непересекающиеся классы $S_1,\ldots,S_k$: $\cup S_i = O$, $S_i \cap S_j = Æ$, i¹j, при котором функционал качества Q(S) достигает экстремального значения на множестве A допустимых правил классификации. В качестве Q(S) используют критерии, минимизирующие межгрупповое сходство и одновременно максимизирующее внутригрупповое сходство. Состав множества A зависит от предварительной (априорной) выборочной информации об этих классах. Итак, задача классификации формально сводится к нахождению разбиения S*: $Q(S^*) = \min Q(S)$ для S Î A. Заметим, что при этом число k может быть и неизвестно. При любых трактовках кластеров и для различных методов классификаций неизбежно возникает проблема измерения близости объектов. С этой проблемой связаны следующие трудности: неоднозначность выбора способа нормировки и определения расстояния между объектами.

Содержательная и математическая постановка задачи статистического прогноза

Построение математической модели, например. Технологического процесса в зависимости от поставленной задачи может преследовать следующие цели: минимизировать расход материала на единицу выпускаемой продукции при сохранении качества, произвести замену дорогостоящих материалов на более дешевые или дефицитных на распространение; сократить время обработки в целом или на отдельных операциях, перевести отдельные режимы в некритические зоны, сни-

зить трудовые затраты на единицу продукции и т.п.; улучшить частные показатели и общее количество готовой продукции, повысить однородность продукции, улучшить показатели надежности и т.п.; увеличить надежность и быстродействие управления, увеличить эффективность контроля качества, создать условия для автоматизации процесса управления и т.п. Прежде всего, необходимо выбрать зависимую переменную Y, которую обычно называют целевой функцией или параметром оптимизации, за который принимают один из показателей качества продукции либо по каждой технологической операции отдельно, либо по всему технологическому процессу сразу. Параметр оптимизации должен соответствовать следующим требованиям: параметр должен измеряться при любом изменении (комбинации) режимов технологического процесса; параметр должен быть статистически эффективным, то есть измеряться с наибольшей точностью; параметр должен быть информационным, то есть всесторонне характеризовать технологический процесс (операцию); параметр должен иметь физический смысл, то есть должна быть возможность достижения полезных результатов при соответствующих условиях процесса; параметр должен быть однозначным, т.е. должно минимизироваться или максимизироваться только одно свойство изделия.

Для достоверного отображения объективно существующих процессов необходимо выявить существенные взаимосвязи и не только выявить, но и дать им количественную оценку. Этот подход требует вскрытия причинных зависимостей. Под причинной зависимостью понимается такая связь между процессами, когда изменение одного из них является следствием изменения другого.

Сформулируем математическую постановку задачи статистического прогноза на примере задачи регрессионного анализа в п.3.

Схема взаимодействия переменных при статистическом исследовании зависимостей

Основная цель статистического исследования зависимостей (СИЗ) состоит в том, чтобы на основании частных результатов статистического наблюдения за показателями двух или трех различных явлений, происходящих с исследуемым объектом, выявить и описать существующие взаимосвязи. В случае численного выражения такие показатели называют переменными.

Рамки применения аппарата СИЗ определяются двумя условиями: - стохастичность интересующей нас взаимосвязи между переменными (т.е. реализация явления или события А одной переменной может повлечь за собой событие В другой переменной с вероятностью р); - взаимосвязь между переменными выявляется на основе статистических наблюдений по выборкам из соответствующих генеральных совокупностей событий.

Опишем функционирование изучаемого реального объекта набором переменных, среди которых выделим: $x^{(1)},..., x^{(p)}$ - «входные» переменные, описывающие условия или причинные компоненты функционирования (поддаются контролю или частичному управлению); для них используются такие термины как факторы-аргументы, факторы-причины, экзогенные, предикторные (предсказательные), объясняющие; $y^{(1)},..., y^{(m)}$ - «выходные», характеризующие поведение объекта или результат (эффективность) функционирования; обычно их называют отклики, эндогенные, результирующие, объясняемые, факторы-следствия, целевые факторы; $e^{(1)},..., e^{(m)}$ - латентные (скрытые, не поддающиеся непосредственному измерению) случайные «остаточные» компоненты, отражающие влияние на $y^{(1)},..., y^{(m)}$ неучтенных «на входе» факторов, а также случайные ошибки в измерении анализируемых показателей; остатки.

Используя введенный набор переменных, задача СИЗ может быть сформулирована следующим образом: по результатам N измерений

$$(x_i^{(1)},...,x_i^{(p)},y_i^{(1)},...,y_i^{(m)}),\ i=\overline{1,N}$$

исследуемых переменных на N объектах построить такую (векторно-значимую) функцию

$$f(x^{(1)},\dots,x^{(p)}) = \begin{Bmatrix} f^{(1)}(x^{(1)},\dots,x^{(p)}) \\ \dots\dots\dots\dots\dots \\ f^{(m)}(x^{(1)},\dots,x^{(p)}) \end{Bmatrix},$$

которая позволила бы наилучшим образом восстановить значения переменных $Y = (y^{(1)},\dots,y^{(m)})'$ по заданным значениям объясняющих переменных $X = (x^{(1)},\dots,x^{(p)})'$.

Математический инструментарий СИЗ

Методы СИЗ составляют содержание отдельных частей многомерного статистического анализа, которые можно определить как раздел математической статистики, посвященный построению оптимальных планов сбора, систематизации и обработки многомерных статистических данных, нацеленных на выявление характера и структуры взаимосвязей между компонентами (X,Y) и предназначенных для получения практических и научных выводов. Среди p+m компонент могут быть: количественные, порядковые (ординальные), классификационные (номинальные).

Методы СИЗ формировались с учетом специфики моделей, обусловленных природой изучаемых переменных. Схематично всю совокупность методов приведем в таблице 2.

Таблица 2 – Математический инструментарий СИЗ

Природа результирующих показателейY	Природа объясняющих переменных X	Названия обслуживающих разделов многомерного статистического анализа
Количественная	Количественная	Регрессионный и корреляционный анализ

Количественная	Одна количественная переменная, интерпретируемая, как время	Анализ временных рядов
Количественная	Неколичественная (ординальные или номинальные переменные)	Дисперсионный анализ
Количественная	Смешанная (количественные и неколичественные переменные)	Ковариационный анализ, модели типологической регрессии
Неколичественная (порядковые переменные)	Неколичественная (ординальные или номинальные переменные)	Анализ ранговых корреляций и таблиц сопряженности
Неколичественная (номинальные переменные)	Количественная	Дискриминантный анализ, кластер-анализ, расщепление смесей распределения
Смешанная (количественные и неколичественные переменные)	Смешанная (количественные и неколичественные переменные)	Аппарат построения логических решающих функций и эмпирического образа данных

Краткая характеристика математического инструментария

Корреляционный анализ оценивает степень тесноты статистической взаимосвязи и обосновывает целесообразность регрессионного анализа. Регрессионный анализ позволяет получить прогноз количественных значений результирующей

переменной по значениям входных. Анализ временных рядов занимается исследованием поведения результирующих переменных во времени. Дисперсионный анализ выявляет наличие взаимосвязи между качественными показателями и результирующей переменной.

Оценка закона распределения. Непараметрический подход

Разновидности первичной статистической обработки

При первичной статистической обработке данных обычно решаются следующие задачи: отображение вербальных переменных в номинальную (с предписанным числом градаций) или ординальную (порядковую) шкалу; статистическое описание исходных совокупностей с определением пределов варьирования переменных; анализ резко выделяющихся переменных; восстановление пропущенных значений наблюдений; проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных; унификация типов переменных, когда с помощью различных приёмов добиваются унифицированной записи всех переменных; экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризация сведений о природе изучаемых распределений (эту разновидность первичной статистической обработки называют иногда процессом составления сводки и группировки); вычислительная реализация учета сложности задачи и возможностей ЭВМ; формулировка задачи на входном языке пакета статистической обработки.

Параметрическое и непараметрическое оценивание закона распределения

Первичные данные, полученные при наблюдении, обычно трудно обозримы. Для того, чтобы начать анализ, в них надо внести некоторый порядок и придать им удобный для исследователя вид. В частности, для начала желательно получить представление об одномерных распределениях случайных величин, входящих в данные.

Существуют два типа задач аппроксимации распределений. Если вид функции распределения известен, но не известны ее параметры, тогда задача сводится к параметрическому оцениванию. Бывают ситуации, когда конкретный вид функции распределения неизвестен и о виде распределения можно сделать лишь самые общие предположения. При таких условиях аппроксимацию неизвестной функции распределения на основе выборки $(x_1, x_2, \ldots, x_N)$ называют непараметрической.

Равноинтервальная гистограмма и полигон частот

Классическими методами статистической аппроксимации функции плотности являются гистограмма (равноинтервальная и равнонаполненная) и полигон частот.

Выборочная функция плотности распределения $f_N(x)$ или гистограмма (равноинтервальная) строится следующим образом. Делим промежуток [a,b], на котором сосредоточены данные выборки на S интервалов $\Delta_1, \Delta_2, \ldots, \Delta_S$, равной длины h=(b-a)/S. Подсчитываем число наблюдений $m_1, m_2, \ldots, m_S$, попавших в интервал $\Delta_1, \Delta_2, \ldots, \Delta_S$, соответственно. Полагаем Полигон частот $\varphi_N(x)$ получают путем сглаживания гистограммы

$$\varphi_N(x) = \frac{m_k + m_{k+1}}{2Nh} + (x - a_k)\frac{m_{k+1} - m_k}{Nh^2}, \quad x \in [x_k, x_{k+1}],$$

где $x_k (k = \overline{1,S})$ - середина промежутка Δ_k, a_k-правый конец промежутка Δ_k.

Очевидно, что $\varphi_N(x_k) = \frac{m_k}{Nh}, \varphi_N(x_{k+1}) = \frac{m_{k+1}}{Nh}$.

Равнонаполненная гистограмма и полигон частот

Выборочная функция плотности распределения $f_N(x)$ или гистограмма (равнонаполненная) строится исходя из предположения, что вся площадь под графи-

ком оценки функции $f_N(x)$ разбивается на k равных частей. Тогда площадь каждой части равна $\Delta_1 h_1 = \Delta_2 h_2 = \cdots = \Delta_i h_i = \cdots = 1/s$, $h_i = 1/(s \cdot \Delta_i)$. Для конкретной выборки рассчитываются длины интервалов Δ_i, а затем по формуле $h_i = 1/(s \cdot \Delta_i)$, определяется h_i. На основании полученных значений длины и высоты каждого прямоугольника гистограммы получаем оценку $f_N(x)$.

Метод прямоугольных вкладов

Для малых выборок (N<30) гистограмма и полигон частот оказываются обычно искаженными за счет тех или иных случайных локальных отклонений, связанных с отсутствием необходимого числа объектов. Одним из способов частично ликвидировать этот пробел явилась «ядерная» аппроксимация, которая путем «размазывания» имеющихся точек заполняет на гистограмме «впадины» и срезает «пики». Отметим, что «ядерное» сглаживание учитывает особенность функции плотности распределения $\int_a^b f(x)dx = 1$ и потому из всех методов сглаживания является наиболее корректным.

Ядерная аппроксимация закона распределения. Оценка плотности распределения для большинства методов «ядерного» типа обобщенно может быть выражена линейной суммой двух компонент: априорной и эмпирической:

$$f(x) = \alpha_0 f_0(x) + \frac{1-\alpha_0}{N} \sum_{i=1}^{N} p(x - x_i),$$

где $f_0(x)$ - априорная компонента; $p(x - x_i)$- составляющая эмпирической компоненты, связанная с i- ой реализацией выборки (заметим, что x_i играет роль параметра); α_0 - вес априорной компоненты.

Различным методам исследования соответствуют разные значения $\alpha_0 \in [0,1]$ и разные виды функции $p(x - x_i)$. Широко известны оценки «ядерного» типа для f(x) при значении $\alpha_0 = 0$.

В методе прямоугольных вкладов (МПВ)

$$\alpha_0 = \frac{1}{N+1}, \quad f_0(x) = \begin{cases} \frac{1}{b-a}, x \in [a,b] \\ 0 \quad , x \notin [a,b] \end{cases},$$

где [a,b] -интервал изменения случайной величины x; d - ширина функции вклада.

В качестве d может быть взято, например: $d = \frac{1}{s}(b-a)$, где $s \in [5,9]$.

Алгоритм ядерной аппроксимации функции плотности распределения имеет следующий вид.

Этап 1. Задается множество точек $\tilde{y}_{(i)}$: $\{x_i - d/2, x_i + d/2\}, i = \overline{1,N}$;

Этап 2. Полученное множество точек $\tilde{y}_{(i)}$ упорядочивается по возрастанию: $y_{(1)}, y_{(2)}, \ldots, y_{(2N)}$;

Этап 3. Определяется «ядерная» аппроксимация функции плотности распределения:

$$f(x) = \begin{cases} 0 \\ \frac{s-1}{(N+1)(b-a)s} + \left(\frac{N}{N+1} \cdot b_j \right) \Big/ d \\ 0 \end{cases} \begin{matrix} , x < y_1, \\ , y_j \le x < y_{j+1}, \\ , x \ge y_{s+1}. \end{matrix}$$

где b_j - количество точек исходной выборки, попавших в интервал $[y_j, y_{j+1})$, а y_j ($j = \overline{1,s+1}$) –некоторое подмножество точек из множества $y_{(1)}, y_{(2)}, \ldots, y_{(2N)}$.

Оценка закона распределения. Параметрический подход

Нормальная вероятностная бумага

Пусть даны N наблюдений $x_1,\ldots,x_N$, извлеченные из генеральной совокупности с функцией распределения F(t). Пусть $x^{(1)},\ldots,x^{(N)}$- упорядоченный по возрастанию ряд наблюдений. Тогда за оценку F(t) принимают $\hat{F}(t) = i/N$, где $i \to \max_j x^{(j)}$; $x^{(j)} \le t$

В тех случаях, когда требуется проверить гипотезу о том, что случайная величина имеет функцию распределения G(t), принадлежащую семейству вида F((t-m)/s), где F(.) известная непрерывная функция распределения, при построении оценки $\hat{F}(.)$ часто используют специальную шкалу, откладывая по оси ординат вместо $\hat{F}(t)$ величину $V = F^{-1}(\hat{F}(t))$, где F^{-1} - функция, обратная к F. В этом случае в координатах (t,v) график G(t) превращается в прямую линию, по положению которой можно легко оценить параметры m и s. Заметим, что наибольшее распространение на практике получила нормальная вероятностная бумага, для которой $V = \Phi^{-1}$, где Ф(.)- стандартная функция нормального распределения.

Опишем алгоритм оценки с помощью вероятностной бумаги параметров центра $\hat{\mu}$ и разброса $\hat{\sigma}$. Работа осуществляется в несколько этапов.

Этап 1. Строится вероятностная бумага. Для этого внизу окна графика на оси абсцисс (см.рис.) откладывается интервал $[x_{min}, x_{max}] = [x^{(1)}, x^{(N)}]$. Масштаб подбирается так, чтобы интервал занял ширину окна, за исключением левого отступа 7-8 см. Ось ординат проводится с отступом от левого края 4-5 см. При этом пунктиром отделяется шкала величины $V=\Phi^{-1}$, которая равномерно изменяется от -3 до 3. Таким образом, точка V=-3.0 будет находиться на оси абсцисс, а точка V=3.0 будет находиться в верхнем левом углу. Слева от пунктирной оси делаются отметки шкалы V, а между осью V и $\hat{F}(t)$ отметки вероятности p: 0.01;0.05;0.1;0.25;0.5;0.75;0.9;0.95;0.99.

Шкала вероятностей является неравномерной. Засечка вероятности осуществляется следующим образом. Берется вероятность p=0.01. По таблицам нормальной функции распределения находится значение $V = \Phi^{-1}(p)$. Напротив полученного значения V ставится засечка 0.01 на шкале вероятностей. Далее берется вероятность p=0.05 и т.д.

Этап 2. Исходная выборка значений упорядочивается по возрастанию. В результате получается последовательность $x^{(1)}$,…, $x^{(N)}$.

Этап 3. Для каждого значения $x^{(i)}$, $i=\overline{1,N}$ на плоскости $(t,\hat{F}(t))$ отмечается точка ($x^{(i)}$, i/N). Для того, чтобы определить расположение этой точки, находится значение $V_i=\Phi^{-1}(i / N)$, которое откладывается по оси V .

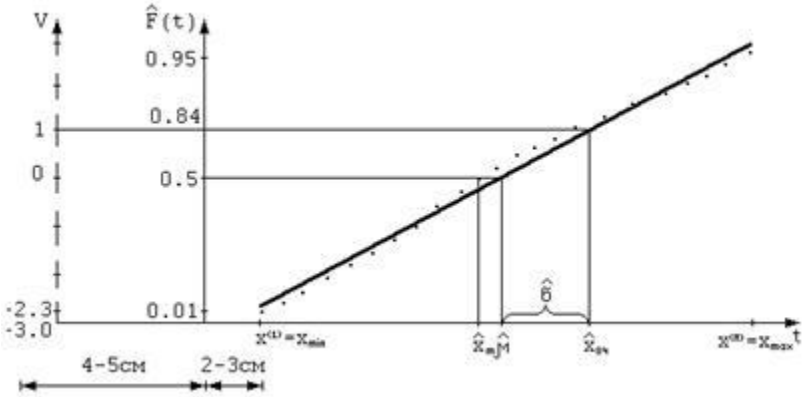


Рисунок 1 – Оценка параметров нормального распределения на нормальной вероятностной бумаге

Этап 4. Если точки ($x^{(i)}$,i/N) в какой-то мере ложатся вдоль некоторой прямой, то можно грубо считать генеральную совокупность, из которой извлечена данная выборка, нормальной. В противном случае надо подыскать преобразование переменной, например, логарифмирование, извлечение корня и т.п., в результате которого выборка бы соответствовала нормальному распределению.

Этап 5. В случае принятия гипотезы о нормальности распределения осуществляется оценка параметров распределения. В качестве оценки центра $\hat{\mu}$ берется медиана выборки, которая соответствует вероятности p=0.5. Оценка стандартного отклонения $\hat{\sigma}=\hat{x}_{84}-\bar{\mu}$, где $\hat{x}_{84}$ - оценка 0.84 квантиля распределения, полученного при V=1.

Параметрическое оценивание

Построение гистограммы, полигона частот и ядерной аппроксимации основано на локальной интерполяции. Другой подход к аппроксимации заключается в интерполировании закона распределения на всем интервале [a,b]. К методам

этого типа относится аппроксимация с помощью системы кривых Пирсона. Систему кривых Пирсона получают путем выравнивания дискретного гипергеометрического распределения непрерывной кривой. При этом для выбора подходящей кривой используют четыре первых момента выборочного распределения. Отметим, что практического распространения данный способ аппроксимации распределения не получил в связи с неустойчивостью моментов первого порядка и невозможностью интерпретации механизма генерации выборки полученного типа распределения.

Критерием согласия χ^2

Более популярными среди интегральных методов аппроксимации оказались параметрические методы оценки распределения путем **проверки на согласие** данного эмпирического распределения с конкретным теоретическим распределением, например, нормальным, экспоненциальным и т.д. В реальной ситуации тип распределения часто бывает известен. Кроме того, просмотрев гистограмму или полигон частот, пользователь для себя уже принимает общепризнанную гипотезу H_0 о типе распределения (или наоборот, отвергает ее из-за сильного засорения выборки, смешения в ней двух или более подвыборок из разных генеральных совокупностей). Математический аппарат в виде критерия согласия используется здесь с целью подтверждения и оформления решения пользователя.

Воспользуемся χ^2 - критерием согласия. Процедура проверки гипотезы H_0 в данном случае будет состоять из следующих этапов.

Этап 1. Область изменения выборки [a,b] делим на S равных интервалов, как при построении гистограммы. Если в каком-то интервале частота m_s слишком мала (меньше 5), то этот интервал объединяется с соседним интервалом. Таким образом количество интервалов может уменьшиться и стать равным S'.

Этап 2. По выборке вычисляют оценки параметров теоретического распределения (тем самым теоретическое распределение будет полностью определено). Теперь по теоретическому распределению вычислим вероятности p_s того, что

$\sum_{s=1}^{s'} p_s = 1$

случайная величина X принимает значение из s-го интервала, при этом

. Затем найдем теоретические частоты $n_s = N \cdot p_s$.

Этап 3. Гипотеза H_0 верна, если теоретические и эмпирические частоты n_s и m_s достаточно мало отличаются друг от друга. Для проверки гипотезы H_0 используем следующую статистику:

$$Q^2 = \sum_{s=1}^{s'} \frac{(m_s - n_s)^2}{n_s}$$

Этап 4. Случайная величина Q^2 имеет $\chi^2(v)$ распределение с числом степеней свободы v=S'-r-1, где S' - количество интервалов, r -количество параметров теоретического распределения, оценки которого вычислялись по выборке. Чем больше Q^2. тем хуже согласованы теоретическое и эмпирическое распределения. При достаточно большом значении Q^2 нужно отвергнуть гипотезу H_0. Поэтому используем только правостороннюю критическую область. P - значением является площадь области под функцией плотности распределения $\chi^2(v)$ справа от точки Q^2 (см. таблицу процентилей распределения $\chi^2(v)$). Если P < α, то мы отвергаем H_0 и принимаем гипотезу H_1: теоретическое и эмпирическое распределения не согласованы. Здесь α — это уровень значимости, который обычно принимается равным 0.05.

Восстановление пропущенных значений и анализ выбросов

Восстановление пропущенных значений

Непараметрический подход к оценке пропусков в матрице данных. Наряду с подходом, требующим аналитического задания закона распределения, существует и другой, основанный на использовании расстояния между параметрами объектов (в некоторой метрике), определяемого по значениям признаков, измеренных у обоих объектов. Постулируется, что, если два объекта близки в пространстве измеренных признаков, то они должны быть близки и в пространстве

по неизмеренным признакам. Метрика и пороговое значение расстояния, определяющие близость объектов, вводятся в зависимости от условий задачи (шкалы, количества признаков).

Алгоритм ZET

Рассмотрим схематично конкретизацию этого подхода в известном алгоритме ZET. Пусть у объекта X_i требуется оценить значение пропущенного признака $x^{(j)}$, т.е. оценить $x_i^{(j)}$ в матрице X. Для этого в X выделяется подмножество объектов, у которых измерено значение j-го признака. В этом подпространстве выделяется однородная группа объектов наиболее близких к X_i в подпространстве признаков, полученном из исходного пространства исключением j-го признака. Неизмеренное значение $x_i^{(j)}$ заменяется средним по выделенной группе объектов. Для оценки качества заполнения пропусков ввести формализованный критерий трудно. Приближенно его оценивают, например так: из матрицы X случайным образом исключается часть измеренных значений, затем исключенные пропуски заполняются. Мера качества заполнения определяется с помощью меры заполнения истинных значений от полученных.

Анализ выбросов

При наличии таких данных возникает вопрос: чем объяснить обнаруженные резкие отклонения в исходных данных? Например, объясняются ли они природой анализируемой генеральной совокупности? Если случайные колебания выборочных значений обусловлены искажениями стандартных условий сбора статистических данных или прямыми ошибками регистрации и записи, то их надо исключить. Наиболее надежным способом решения вопроса об исключении данных из рассмотрения является изучение условий регистрации и сбора данных. Если невозможен анализ условий, при которых регистрировалось аномальное наблюдение, то обращаются к статистическим методам. Их общая логическая

схема: исходя из исходных предложений о природе анализируемой совокупности данных, исследователь задается функцией $\Psi(X^*, X)$ (X - все имеющие наблюдения, X^* - подозрительные наблюдения), характеризующей степень аномальности, определяет значение Ψ и сравнивает с пороговым значением Ψ. При $\Psi > \Psi_0$ подозрительное наблюдения исключается, или для него определяется весовой коэффициент. В вероятностной постановке Ψ_0 определяется из стандартных статистических таблиц с учётом закона распределения статистики Ψ в предположении необоснованности относительно X^*. В других случаях Ψ_0 определяется из содержательных соображений.

Проверка гипотез

Статистические процедуры анализа резко выделяющихся наблюдений основаны на предположении однородности данных. При этом выбросы рассматриваются как наблюдения, нетипично удаляющиеся от центра распределения. Основная трудность при использовании имеющихся аналитических процедур состоит в том, что реальная доля «засорения» не известна, а оценивается по тем же данным, по которым проверяется значимость отклонения. Наиболее устойчивы к отклонениям от предположения нормальности основной части выборки графические процедуры. При использовании статистических методов выделения выбросов следует иметь в виду, что выбросы могут оказаться наиболее существенной частью выборки, проясняющей, например то, как собирались данные (например, изменение условий эксперимента, не замеченное исследователем). Данная задача распадается на два этапа: выделение подозрительных наблюдений; проверка статистической значимости отличий от основной массы данных. Оба этапа основываются на определенных предположениях о распределении основной (не засоренной) части наблюдений и выбросов (засорений). Обычно предполагают, что не засоренная часть наблюдений имеет одно или многомерное нормальное распределение с неизвестными параметрами $N(\mu, \sigma^2)$, а засоренная: $N(\mu + d, \sigma^2)$ или $N(\mu, \gamma\sigma^2), \gamma \geq 1$.

Унификация признакового описания

Отношение, признаки, измерения

Для описания разнородных задач первичной статистической обработки помимо обычного языка математической статистики удобно использовать терминологию теории бинарных отношений. Опишем кратко основные понятия.

Отношения. Бинарное отношение P на множестве объектов A - подмножество упорядоченных пар объектов (a, в) декартового произведения A на A: $A \times A$

.

У некоторых особо важных отношений есть специальные названия.

Отношение эквивалентности разбивает все множество объектов на не пересекающиеся классы, в каждом из которых объекты признаются тождественными, неразличимыми, а из разных классов – нетождественными.

Квазипорядок (нестрогий порядок) определяет отношение «быть не меньше». Если исключить из него возможность равенства элементов, то оно превратится в порядок.

Толерантностью называется отношение «похожести». В анализе данных оно имеет особую роль, так как объединение объектов происходит по похожести. Здесь в отличие от эквивалентности из a=в, в=с не следует a=с.

Метризованное отношение. Каждому отношению на множестве объектов $a_1,\ldots,a_n$ можно сопоставить матрицу $N*N$ из бинарных значений $r_{ij}=\{0,1\}$, где $r_{ij}=1$, для $(a_i,a_j)\in P$, $r_{ij}=0$, иначе. Понятие «отношение «можно расширить, распространив его на количественные признаки. В 1977 Б. Г. Литваком введено понятие «метризованного отношения». «Метризованным отношением» называется пара <W(P),P> , где P – отношение, W(P) – множество чисел (весов), характеризующих «степень принадлежности» пары к данному «метризованному отношению». Вместо булевских матриц (2.2) вводятся матрицы с вещественными элементами p_{ij}, которые определяются (для линейных отношений порядка).

$$p_{ij} = \begin{cases} W_{ij}, if (a_i, a_j) \in P \\ -W_{ij}, if (a_j, a_i) \in P \end{cases}$$

Признаки. Отношения определены на парах объектов. Признак – это свойство, измеренное на каждом объекте. Может случиться, что отношение существует, а измеримые признаки им не отвечают. Так, отношению толерантности нельзя сопоставить признак, определенный на каждом объекте.

Измерение. Рассмотрим способы измерения признаков. Обычно под процедурой измерения какого-либо свойства понимается приписывание некоторых числовых значений отдельным уровням этого свойства в определенных единицах. При этом важно знать в какой мере условность в выборе единиц измерения повлияет на значение показателя. Например, если стоимость продукции измерить в рублях, а потом в тысячах рублей, то изменится лишь число единиц измерения, суть же останется прежней. Здесь возможно умножение, деление на константу, т. Е. масштабирование. Бессмысленно задавать масштаб для температуры по Цельсию, так как мы не можем сказать во сколько раз -5 $^{\circ}C$ меньше +10 $^{\circ}C$. Таким образом разные типы признаков имеют разное множество допустимых преобразований $f(x)$ своих значений, которое определяет тип шкалы.

Типовые структуры признаков

Признаки, описывающие объекты получаются по-разному. В зависимости от того, как измеряют или оценивают значение признака, они могут быть первичными или вторичными. Замер берётся за значение признака. Можно выделить шесть типов признаков:

К первому типу относится прямое измерение, т.е. измерение с использованием приборов (например, измерение длины стола линейкой, измерение скорости машины спидометром, измерение температуры воздуха градусником, измерение силы тока амперметром, измерение глубины моря тахометром и т.д.) или при помощи счета (например, сосчитать количество книг на полке, количество фруктов в ящике, количество рыб в аквариуме и т.д.).

Ко второму типу относится прямое измерение с последующим аналитическим преобразованием, зависящим от параметров (они вносят случайный разброс в значение). Это измерение подразделяется на одноуровневое, т.е. измерение на объекте и двухуровневое – на группе объектов (например, измерение дозы облучения – человека помещают в некоторую камеру, где одновременно измеряется его вес, количество радиационных частиц, содержащихся в нем и получают представление о дозе внутреннего облучения).

К третьему типу относится аналитическая комбинация: $\sum x_i / n = \bar{x}$ нескольких первого типа или нескольких первого и второго типов (характеристика группы людей – имеется некоторое количество детей в группе, известен их вес, рост, нужно определить средние характеристики по группе, например, средний вес, процент девочек в группе).

К четвёртому типу относится прямая экспертная оценка (например, уровень подготовленности студента, пригодность продуктов для употребления, возможность использования природных ресурсов и т.д.).

К пятому – прямая экспертная оценка с последующим аналитическим преобразованием (например, в зависимости от компетентности эксперимента, т.е. от степени доверия к оценке, полученной экспертом, получается результирующая оценка путём умножения исходной оценки на некоторый коэффициент, который является функцией от компетентности).

К шестому – аналитическая комбинация экспертных оценок (например, берётся несколько экспертных оценок и у каждой есть своя компетентность, и вычисляется средняя оценка).

Типы шкал

Интегрированная информация о шкалах приведена в таблице 3.

Таблица 3 – Интегрированная информация о шкалах

Наименование шкалы	Множество допустимых преобразований F(x)	Отношения, отвечающие шкале	Допустимые числовые операции с измерениями	Примеры измерения
Качественная шкала				
Наименований (номинальная, классификационная)	Взаимно-однозначные	Эквивалентность	Сравнения: x=y, x<>y	Национальность, пол, профессия, вид оплаты труда
Порядковая (ранговая, ординальная)	Монотонно-неубывающие функции	Квазипорядок (нестрогая ранжировка)	Сравнения: x<=y	В строгом смысле примеров шкалы нет. Условно: шкала твердости минералов, экспертные ранжировки, оценки предпочтений
Количественная шкала				
Разностей (балльная)	F(x)=d+x	Аддитивное метризованное	Сравнения: x-y<=z-v; x+y, x-y	Квалификационные разряды, балльные оценки
Интервалов (интервальная)	F(x)=d+kx, k>0	4-арное мультипликативное метризованное	(x-y)/(z-v), x+y, x-y	Любые показатели, значения которых могут быть отрицательными: температура по Цельсию, летоисчисление, прибыль (при наличии

				убытков), высота над уровнем моря
Отношений (относительная)	F(x)=kx, k>0	Мультипликативное метризованное	X/y, x*y, x+y, x-y	Температура по Кельвину, возраст, производительность труда

Шкалы. Отображение $\Psi: A \to R^1$, называется шкалой наименований, если его допустимым преобразованием является взаимно однозначное отображение $\eta: \Psi(A) \to R^1$. Шкальные значения играют роль имен объектов. Здесь определено отношение равенства, которое соответствует отношению эквивалентности. Оно индуцирует на А разбиение на непересекающиеся классы. Эти признаки называют классификационными или номинальными. Примеры: профессия, национальность, пол, место рождения.

Отображение $\Psi: A \to R^1$ называется шкалой порядка, если его допустимым преобразованием является монотонно возрастающее непрерывное отображение $\eta_2 : \Psi(A) \to R^1$. Определены отношения равенства и порядка. Первое соответствует эквивалентности объектов, второе - порядку. Отношение эквивалентности индуцирует разбиение А на классы, а отношение порядка задает линейный порядок на множестве классов эквивалентности. Соответствующее отношение порядка задает порядок на множестве различных значений признака $x^{(j)}$, которые называются градациями шкалы порядка. Эти признаки называют порядковыми или ординальными. В строгом смысле примеров шкалы нет. Условно примерами шкалы являются: сила ветра в баллах, образование, оценка на экзамене, шкала твердости минералов.

Отображение $\Psi(A) \to R^1$ называется количественной шкалой: а) интервалов; б) отношений; в) разностей; г) абсолютной, если допустимым преобразованием является положительное линейное преобразование вида:

$\eta_2 : \psi(A) \to \alpha\psi(A) + \beta$,

где для каждого подвида количественной шкалы а) $\alpha \in R^+, \beta \in R^1$; б) $\alpha \in R^-, \beta = 0$; в) $\alpha = 1, \beta \in R^1$; г) $\alpha = 1, \beta = 0$. Примеры: а) любые показатели, значение которых может быть отрицательным: температура по Цельсию, летоисчисление, убытки - прибыль; б) возраст, вес, длина; в) квалификационные разряды, балльные оценки; г) количество элементов некоторого множества, адрес в памяти ЭВМ.

Унификация типа переменных

Одна из сложностей автоматизированного анализа информации заключается в том, что среди признаков могут быть количественные и качественные (порядковые или классификационные), а большинство методов статистической обработки предполагают их однотипность. Поэтому и возникает вопрос об унификации записи единичного наблюдения.

1-й вариант решения. Наблюдение представляют в виде вектора размерности $m_1 + \ldots + m_p$, m_i - число градаций (интервалов группирования, уровней качества или однородных групп) признака $x^{(i)}$. Компоненты этого вектора принимают значение 0 или 1. Недостатки: субъективизм в выборе способов разбиения диапазонов количественных признаков, потеря информативности при переходе от индивидуальных к групповым значениям.

2-й вариант. Преобразование качественных переменных в количественные с помощью «оцифровки» (шкалирование).

3-вариант. Сведение классификационных и количественных данных к порядковым.

Критерии сравнения групп

Статистические критерии

Параметрические и непараметрические критерии

Для формальной проверки статистических гипотез существуют различные статистические критерии. Их можно разделить на две большие группы: параметрические и непараметрические.

Параметрические критерии основаны на том, что распределение данных известно. То есть, при применении какого-нибудь параметрического критерия нужно всегда следить за тем, что главное допущение критерия – тип распределения – выполняется. Как правило, многие параметрические критерии предполагают нормальность распределения данных. Во многом это связано с тем, что нормальное распределение широко распространено. Кроме того, часто все, что мы можем сказать о распределении данных, это то, является ли оно нормальным или нет, потому что задача определения типа распределения довольно сложна и существующие формальные тесты могут определить лишь общий класс распределения или показать, “между какими” распределениями находится интересующее нас распределение.

Непараметрические критерии исходят из того, что распределение данных неизвестно. Поэтому при использовании этих критериев часто действия производятся не с самими значениями в выборке/выборках, а с их рангами.

То, что при применении тех или иных критериев нужно думать о распределении данных, не всегда означает, что перед их использованием нужно обязательно проверять распределение данных на нормальность. Иногда формальный критерий может показать, что гипотезу о нормальном распределении нужно отвергнуть, но распределение интересующего нас показателя может быть очень близким к нормальному. Поэтому главное исходить из формы распределения и тщательно анализировать данные на качественном уровне: правда ли, что показатель слишком часто принимает минимальное или максимальное значение (распределение скошено вправо или влево), правда ли, что из теоретических знаний

об исследуемом показателе следует, что его распределение похоже на нормальное?

Сравнение средних значений. Сравнение средних значений в двух группах. Критерий Стьюдента для двух выборок (t-test)

Предположение: выборки взяты из нормального распределения.

Нулевая гипотеза: средние двух генеральных совокупностей равны.

Варианты: есть два варианта критерия Стьюдента: для независимых выборок (в двух выборках содержатся значения показателя для разных объектов) и для связных выборок (в двух выборках содержатся значения показателя для одних и тех же объектов, например, в разные периоды времени). Пример использования критерия Стьюдента для независимых выборок: сравнение средних значений ВВП на душу населения в демократиях и автократиях. Пример использования критерия Стьюдента для связных выборок: сравнение средней заработной платы в одних и тех же регионах до экономической реформы и после нее.

Строго говоря, есть еще одно деление внутри критерия Стьюдента для независимых выборок: при условии, что дисперсии генеральных совокупностей, из которых взяты выборки, равны, и при условии, что эти дисперсии не равны. В R по умолчанию встроена реализация второго варианта, так как это более реалистичное условие.

Реализация в R:

Сравним средний уровень детской смертности в католических и протестантских кантонах Швейцарии (данные за 1888 год, встроенная в R база swiss). Предполагаем, что уровень смертности в католических и протестантских кантонах имеет нормальное распределение.

```
# переменная religion - по которой будем делить кантоны на 2 группы
swiss$religion <- ifelse(swiss$Catholic > 50, "catholic", "protestant")
# сам тест
# через ~ указывается показатель, по которому делим наблюдения в базе на 2 гр
уппы
```

```
t.test(data = swiss, Infant.Mortality ~ religion)
##
## Welch Two Sample t-test
##
## data: Infant.Mortality by religion
## t = 1.0863, df = 31.717, p-value = 0.2855
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.8620975 2.8310630
## sample estimates:
## mean in group catholic mean in group protestant
##      20.55000      19.56552
```

Здесь p-value = 0.2855, значит, вероятность того, что наша нулевая гипотеза о равенстве средних верна (при условии имеющихся данных), равна 0.2855. На 5%-ном уровне значимости есть основания не отвергать нулевую гипотезу о равенстве средних значений (0.2855 > 0.05). Средний уровень детской смертности в католических и протестантских кантонах можно считать одинаковым.

Если вдруг в базе данных показатели, средние по которым нужно сравнить, просто находятся в двух разных столбцах, то t-test выглядит так:

```
# вместо векторов могут быть столбцы базы через $
set.seed(123)
index.a <- rnorm(100, mean = 2, sd = 6)
index.b <- rnorm(100, mean = 10, sd = 10)
t.test(index.a, index.b)
##
## Welch Two Sample t-test
##
## data: index.a and index.b
## t = -5.7428, df = 156.59, p-value = 4.705e-08
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -8.577204 -4.186989
## sample estimates:
## mean of x mean of y
## 2.542435 8.924532
```

Тут p-value = 4.705e-08, то есть почти 0. На 5%-ном уровне значимости есть основания отвергнуть нулевую гипотезу о равенстве средних значений.

Для связных выборок (парных показателей), в качестве аргумента функции t.test нужно дописать paired=TRUE).

Критерий Уилкоксона (Манна-Уитни) для двух групп

Предположение: выборки взяты не из нормального распределения (из какого – неизвестно).

Нулевая гипотеза: выборки взяты из одного и того же распределения (можно использовать как аналог t-test и говорить о равенстве средних двух независимых выборок, но изначально критерий Уилкоксона не об этом).

Варианты: как и в случае с критерием Стьюдента для двух выборок, есть вариант для независимых и связных выборок.

Реализация в R:

Сравним средний уровень детской смертности в католических и протестантских кантонах Швейцарии (данные за 1888 год, встроенная в R база swiss). Предполагаем, что уровень смертности в католических и протестантских кантонах имеет распределение, отличное от нормального.

```
# через ~ указывается показатель, по которому делим наблюдения в базе на 2 гр
уппы
wilcox.test(data = swiss, Infant.Mortality ~ religion)
## Warning in wilcox.test.default(x = c(22.2, 20.2, 26.6, 23.6, 24.9, 21,
## 24.4, : cannot compute exact p-value with ties
##
```

```
## Wilcoxon rank sum test with continuity correction
##
## data: Infant.Mortality by religion
## W = 286.5, p-value = 0.5841
## alternative hypothesis: true location shift is not equal to 0
```

Если вылезает предупреждение “не могу подсчитать точное p-значение при наличии повторяющихся наблюдений”, можно добавить аргумент exact=FALSE, что будет говорить R о том, что мы это понимаем, и не ждем от него втаком случае точного расчета p-value.

Случай с двумя столбцам:

```
wilcox.test(index.a, index.b)
##
## Wilcoxon rank sum test with continuity correction
##
## data: index.a and index.b
## W = 2923, p-value = 3.902e-07
## alternative hypothesis: true location shift is not equal to 0
```

Сравнение средних значений в трех и более группах

ANOVA

ANOVA – analysis of variance, дисперсионный анализ.

Предположение: выборки взяты из нормального распределения.

Нулевая гипотеза: средние значения k генеральных совокупностей равны (где k – число исследуемых выборок).

Реализация в R:

Сравним средний вес цыплят в пяти группах – в зависимости от того, каким кормом их кормили (данные из встроенной в R базы chickwts). Предполагаем, что вес цыплят, относящихся к разным группам, имеет нормальное распределение.

```
anova.res <- aov(data = chickwts, weight ~ feed) # ANOVA, выдает сумму квадратов
summary(anova.res) # все статистики + p-value
```

```
##     Df Sum Sq Mean Sq F value Pr(>F)
## feed   5 231129 46226 15.37 5.94e-10 ***
## Residuals 65 195556  3009
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Критерий Краскела-Уоллиса

Предположение: выборки взяты из распределения, отличного от нормального (из какого – неизвестно).

Нулевая гипотеза: выборки взяты из одного и того же распределения (можно говорить о равенстве средних k независимых выборок, но изначально критерий не об этом).

Реализация в R:

Сравним средний вес цыплят в пяти группах – в зависимости от того, каким кормом их кормили (данные из встроенной в R базы chickwts). Предполагаем, что вес цыплят, относящихся к разным группам, имеет распределение, отличное от нормального.

```
kruskal.test(data = chickwts, weight ~ feed)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: weight by feed
## Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

ПОСТРОЕНИЕ ПРОСТЕЙШИХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Этапы решения задачи моделирования

Часто на практике возникает следующая задача. Имеется объект исследования (ОИ), который характеризуется набором переменных: входных ($x_i, i = 1,2,...,k$) и выходной y.

Рис. 3.1. Схема объекта исследования

Требуется найти зависимость выходной переменной от входных

$$y = f(x_1, x_2, ..., x_k) \qquad (3.1)$$

При этом считается, что механизмы процессов, протекающих внутри объекта исследования, неизвестны, а имеются только соответствующие значения входных и выходных параметров. Такая задача носит название задачи «черного ящика».

Рассмотрим простейший случай, когда на вход действует только одна переменная x и требуется найти

$$y = f(x) \qquad (3.2)$$

Решение задачи моделирования в этом случае состоит из 4 этапов:

1) Проведение эксперимента.
2) Выбор вида экспериментальной зависимости.
3) Нахождение параметров выбранной зависимости.
4) Проверка адекватности модели и выводы.

На первом этапе задаем значения входной переменной x из возможного диапазона и замеряем соответствующие значения выходной переменной y. Получаем таблицу

x	x_1	…	x_n

y	y_1	…	y_n

Еслиn велико, то для удобства работы экспериментальные данные можно сгруппировать, не забывая при этом, что группировка вносит погрешности в результаты вычислений.

Тогда результаты опытных данных будут представлены в виде корреляционной таблицы

X / Y	Δ_1	Δ_2	…	Δ_k
Δ_{k+1}	n_{11}	n_{12}	…	n_{1k}
…	…	…	…	…
Δ_{k+m}	n_{m1}	n_{m2}	…	n_{mk}

ЗдесьΔ_i– интервалы, в которые попали соответствующие значения переменной $X(i=\overline{1,k})$ и функции $Y(i=\overline{k+1,k+m})$, n_{ij} - частота появления пары (x_i ,y_j).

Обычно вместо самих интервалов берут значения их середины. Получают таблицу

X / Y	x_1	x_2	…	x_k	
y_1	n_{11}	n_{12}	…	n_{1k}	p_1
…	…	…	…	…	
y_m	n_{m1}	n_{m2}	…	n_{mk}	p_m
	w_1	w_2		w_k	

В этой таблице $w_j=\sum_{i=1}^{m} n_{ij}$ - частота признака x_j, $p_i=\sum_{j=1}^{k} n_{ij}$ - частота признака y_i ,

$n=\sum_{i=1}^{m} p_i=\sum_{j=1}^{k} w_j=\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}$ - объем выборки.

На втором этапе исследования возможны два случая: когда форма экспериментальной кривой известна, и когда она неизвестна.

Во последнем случае могут помочь рекомендации, приведенные в [1, 2], подсказки в справке Excel о выборе линии тренда, метод средних точек для выбора между некоторыми видами зависимости, а также интуитивные представления и опыт решения подобных задач другими исследователями [3, 4].

На практике чаще всего подходящий вид уравнения регрессии выбирают по виду корреляционного поля [5].

В основе регрессионного анализа лежит принцип наименьших квадратов, в соответствии с которым в качестве уравнения регрессии y=f(x) выбирается функция, доставляющая минимум сумме квадратов разностей $K = \sum_{i=1}^{n} [f(x_i) - y_i]^2$, а неизвестные коэффициенты сглаживающей кривой y=f(x) находят из условия ее минимума. Так, если мы ищем кривую в виде $y = a \cdot e^{-bx}$, то из условия minK мы должны найти неизвестные коэффициенты a и b.

Геометрически критерий метода наименьших квадратов означает: из всех кривых заданного вида выбирают ту, у которой сумма площадей квадратов отклонений – наименьшая.

Если аргументом считать y, а x – функцией (то есть если искомую кривую ищут в виде x=g(y)), то говорят о регрессии X на Y. Отклонения в этом случае откладывают по оси X.

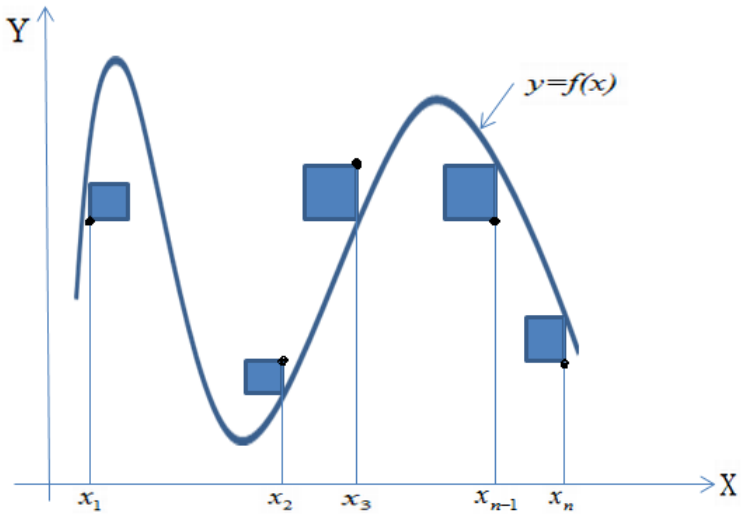


Рис. 3.2. Регрессия Y на X

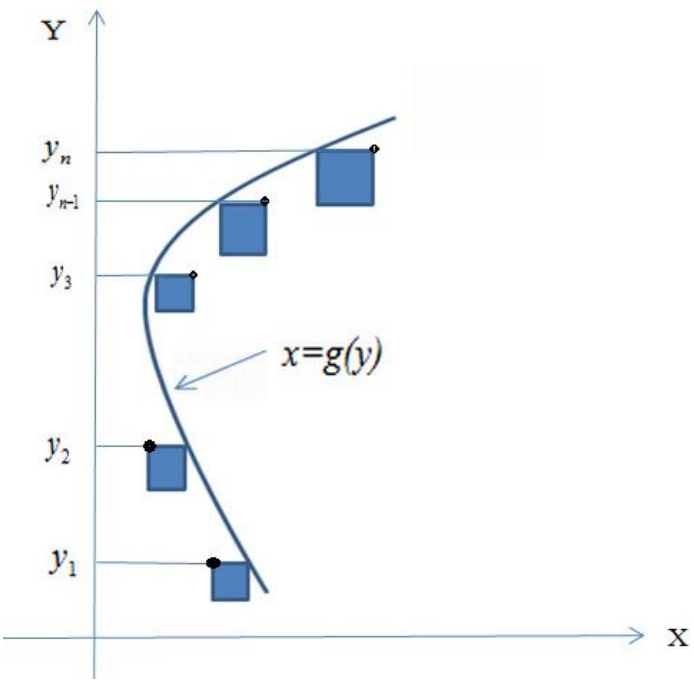


Рис. 3.3. Регрессия X на Y

Количественной мерой рассеяния значений y_i вокруг регрессии f(x) является дисперсия $D = \frac{1}{n-q}\sum_{i=1}^{n}[f(x_i) - y_i]^2$, где q – число коэффициентов, входящих в аналитическое выражение регрессии [6].

Если искомое уравнение – алгебраический полином, то есть

$$f(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_p x^p = Q(x, c_j), \qquad (3.3)$$

то задача поиска минимума Ксводится к составлению и решению системы нормальных уравнений (3.5).

При этом степень аппроксимирующего полинома p и число узлов таблицы n связаны соотношением:

$$p \le n-1 \qquad (3.4)$$

Так, если функция задана в виде таблицы из пяти точек, то аппроксимировать ее можно полиномами до 4 степени включительно ($p \le 4$).

$$\begin{cases} \sum_{i=1}^{n} y_i = n c_0 + c_1 \sum_{i=1}^{n} x_i + \ldots + c_p \sum_{i=1}^{n} x_i^p \\ \sum_{i=1}^{n} x_i y_i = c_0 \sum_{i=1}^{n} x_i + c_1 \sum_{i=1}^{n} x_i^2 + \ldots + c_p \sum_{i=1}^{n} x_i^{p+1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n} x_i^p y_i = c_0 \sum_{i=1}^{n} x_i^p + c_1 \sum_{i=1}^{n} x_i^{p+1} + \ldots + c_p \sum_{i=1}^{n} x_i^{2p} \end{cases} \qquad (3.5)$$

Существуют и другие подходы к поиску коэффициентов c_i: метод наименьших модулей, минимаксный подход к задаче аппроксимации и др. [6].

После того, как модель построена, то есть найдены значения коэффициентов c_i , необходимо удостовериться в её качестве. С этой целью выполняют проверку адекватности модели объекту исследования, для которого она построена.

Проверить адекватность модели – значит установить, насколько хорошо она описывает реальный процесс и насколько качественно ее можно будет использовать для прогнозирования развития данного процесса.

Для того, чтобы проверить адекватность модели, необходима некоторая экспериментальная информация, полученная на этапе функционирования системы или при проведении специального эксперимента, в ходе которого наблюдается интересующий нас процесс.

Проверка адекватности заключается в доказательстве факта, что точность результатов, полученных по модели, сопоставима с точностью расчетов, произведенных на основании экспериментальных данных.

Процедура оценки адекватность разработанной модели реально существующей системе основана на сравнении измерений на реальной системе и результатов экспериментов на модели и может проводиться различными способами. Наиболее распространенные из них [7]:

– по средним значениям откликов модели и системы;

– по дисперсиям отклонений откликов модели от среднего значения откликов системы;

– по максимальному значению относительных отклонений откликов модели от откликов системы.

Адекватность математической модели в простейших случаях может быть установлена визуально путем сравнения экспериментальных значений y_iсо значениями $f(x_i)$ модельной функции в тех же точках таблицы.

Определенную информацию об адекватности уравнения регрессии дает исследование остатков вида$e_i=y_i-f(x_i)$. Наличие грубых отклонений (промахов, выбросов), не связанных с естественным разбросом, может приводить к существенным ошибкам при построении регрессии, что, в свою очередь, может привести к грубым ошибкам прогноза. Некоторые методы выявления выбросов: критерии Эктона, Титьена-Мура-Бекмана, Прескотта-Лунда и другие рассмотрены в [6].

Одной из наиболее эффективных оценок адекватности регрессионной модели, мерой качества уравнения регрессии является **коэффициент детерминации**, определяемый по формуле

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3.6)$$

где $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

В случае линейной связи между X и Y, учитывая, что

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(f(x_i) - \bar{y})^2 + \sum_{i=1}^{n}(y_i - f(x_i))^2$$

R^2 можно вычислить по формуле:

$$R^2 = \frac{\sum_{i=1}^{n}(f(x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3.7)$$

R^2 показывает, насколько предсказание по модели лучше, чем предсказание по среднему значению отклика [1]. R^2 характеризует долю разброса отклика, описываемую регрессией, и лежит в пределах от 0 до 1. Чем ближе R^2 к единице, тем лучше модель описывает экспериментальные данные.

В более сложных случаях, в частности, когда данные заданы корреляционной таблицей, адекватность может быть установлена применением различных статистических критериев.

Чаще всего для оценки адекватности регрессионной модели применяют критерий Фишера-Снедекора [6,8].

Пояснение.Говорят, что случайная величина распределена по закону Фишера-Снедекора, если ее плотность распределения вычисляется по формуле:

$$I_{\nu_1,\nu_2}(x) = \frac{1}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}\frac{\nu_2}{2}-1} x\left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}, \quad x>0$$

где v_1 иv_2 – параметры распределенимя, B(y,z) – бета-функция [2].

Математическое ожидание и дисперсия равны соответственно

$$M(X) = \frac{\nu_2}{\nu_2-2} \quad \text{при } \nu_2 > 2$$

$$D(X) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \quad \text{при } \nu_2 > 4$$

Графики функции плотности распределения Фишера-Снедекора при различных значениях v_1иv_2 приведены на рисунке 3.4.

Программа для его построения:

```
clc
clf()
function y=fish(x,v1,v2)
    y=1/beta(v1/2,v2/2)*(v1/v2)^(v1/2*v2/2-1)*x.*(1+v1/v2*x)^(-(v1+v2)/2)
endfunction
x=0:.1:6;
plot(x,fish(x,3,5),x,fish(x,2,3),x,fish(x,2,5)),xgrid()
legend('v1=3, v2=5','v1=2, v2=3', 'v1=2, v2=5')
```

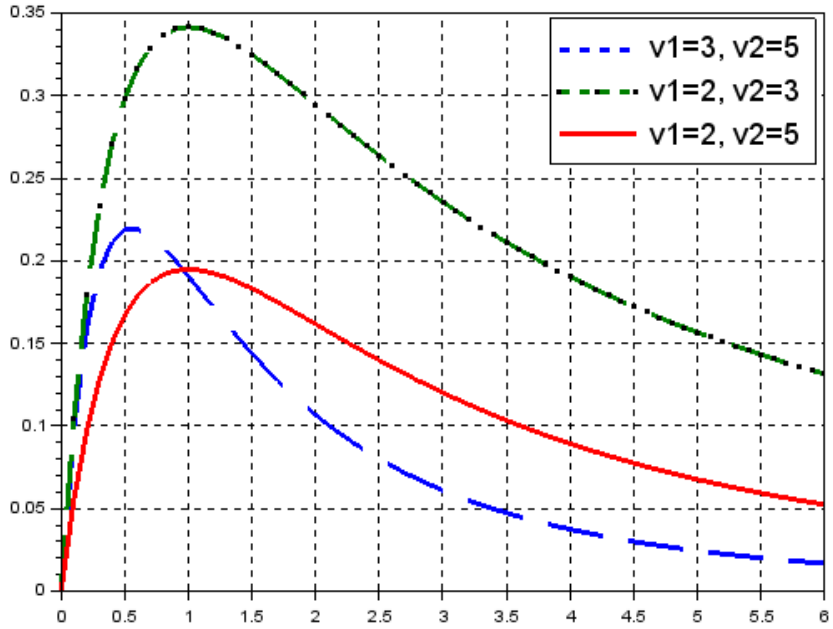


Рис. 3.4. Графики функции плотности распределения Фишера-Снедекора при различных значениях v_1 и v_2

Адекватность полученного уравнения будет тем выше, чем меньше будет разброс точек корреляционного поля точек вокруг построенной кривойy=f(x).

Оценим величину этого разброса. Для этого подсчитаем сумму квадратов отклонений ординат y_i всех точек корреляционного поля от сглаживающей линии регрессии y=f(x):

$$Q = \sum_{j=1}^{k}\sum_{i=1}^{m}\left(y_i - f(x_j)\right)^2 n_{ij} \qquad (3.8)$$

Для того, чтобы оценить разброс выборочных значений y_iвокруг выборочных средних $\overline{y}_j$ $\left(\overline{y}_j = \frac{\sum_{i=1}^{m} y_i n_{ij}}{w_j} \quad j = 1, 2, \ldots, k\right)$ при проведении повторных опытов для различных x_iвычисляют $Q_{повт}$

$$Q_{повт} = \sum_{j=1}^{k}\sum_{i=1}^{m}\left(y_i - \overline{y}_j\right)^2 n_{ij} \qquad (3.9)$$

Она определяет степень влияния на величину Y различных неучтенных факторов (помех), не связанных с величиной X. Эта сумма зависит только от экспериментальных данных.

Кроме этой суммы вычисляют $Q_{адекв}$

$$Q_{адекв} = \sum_{j=1}^{k} \left(f(x_j) - \overline{y_j} \right)^2 w_j \qquad (3.10)$$

$Q_{адекв}$ зависит от вида уравнения y=f(x). Она характеризует меру отклонений сглаживающих средних f(x) от реальных (выборочных) средних $\overline{y}_j$. И чем эта сумма меньше, тем более адекватным будет, очевидно, сглаживающее уравнение регрессии y=f(x).

Можно доказать, что $Q=Q_{повт}+Q_{адекв}$:

$$\sum_{j=1}^{k} \left(f(x_j) - \overline{y}_j \right)^2 w_j + \sum_{j=1}^{k} \sum_{i=1}^{m} \left(y_i - \overline{y}_j \right)^2 n_{ij} = \sum_{j=1}^{k} \sum_{i=1}^{m} \left(y_i - f(x_j) \right)^2 n_{ij} \quad (3.11)$$

Естественно, что если$Q_{адекв}=0$, то сглаживающее уравнение регрессии y=f(x)полностью адекватно выборочным данным.

Если$Q_{адекв}>0$, как обычно и бывает на самом деле, то сравниваяQ$_{адекв}$ с $Q_{повт}$, выясняют, достаточно ли мала $Q_{адекв}$,чтобы для заданного уровня значимости α можно было бы принять нулевую гипотезу H_0об адекватности полученного уравнения y=f(x) при альтернативной гипотезе H_1 об ее неадекватности. Это можно сделать по критериюФишера-Снедекора, если известно, что зависимая случайная величина Y при любом значении x_i величины X распределена по нормальному закону и имеет независящую от X постоянную дисперсию.

Для проверки гипотезы об адекватности находят дисперсию повторности $S^2_{повт}$ и дисперсию адекватности $S^2_{адекв}$:

$$S^2_{повт} = \frac{Q_{повт}}{n-k}; \quad S^2_{адекв} = \frac{Q_{адекв}}{k-q} \qquad (3.12)$$

Здесь n– объем выборки, k – количество различных значений, принимаемых переменной X,q- число параметров регрессионной модели.

После этого находят выборочное значение критерия F Фишера-Снедекора

$$f_{выб} = \frac{S^2_{адекв}}{S^2_{повт}} = \left(\frac{\sum_{j=1}^{k} \left(f(x_j) - \bar{y}_j \right)^2 w_j}{\sum_{j=1}^{k} \sum_{i=1}^{m} \left(y_i - \bar{y}_j \right)^2 n_{ij}} \right) : \frac{k-q}{n-k} \qquad (3.13)$$

Сравнивая его с критическим значением $f_{кр}=f_{кр}(\alpha, k-q, n-k)$, делают вывод об адекватности математической модели[1]. Здесь$f_{кр}(\alpha, k-q, n-k)$ квантиль распределения Фишера-Снедекора ck-q, n-kстепенями свободы.

Чаще всего на практике строят линейные, гиперболические и квадратичные сглаживающие кривые. Рассмотрим построение каждой из них.

Линейная регрессия Y на X и X на Y

Предположим, что из каких-то соображений известно, что между количественными признаками X и Y существует линейная корреляционная зависимость. Уравнение линейной регрессии Y на Xзаписывают в виде y=a+bx.

Тогда условие для нахождения неизвестных оценок коэффициентов регрессии перепишется в виде: $\min\limits_{a,b} \sum\limits_{i=1}^{n} [a + bx_i - y_i]^2$.

Задача поиска минимума может быть сведена к получению и решению системы двух линейных уравнений с двумя неизвестными:

$$\begin{cases} \sum\limits_{i=1}^{n} y_i = na + b\sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i y_i = a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 \end{cases} \qquad (3.14)$$

Если данные представлены в виде корреляционной таблицы, то система нормальных уравнений выглядит так:

$$\begin{cases} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{k} n_{ij} y_i = na + b\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{m} n_{ij} x_j \\ \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k} n_{ij} x_j y_i = a\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{m} n_{ij} x_j + b\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{m} n_{ij} x_j^2 \end{cases}$$

или

$$\begin{cases} \sum_{i=1}^{m} p_i y_i = na + b\sum_{j=1}^{k} w_j x_j \\ \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} x_j y_i = a\sum_{j=1}^{k} w_j x_j + b\sum_{j=1}^{k} w_j x_j^2 \end{cases} \quad (3.15)$$

Данную задачу можно решить и методами корреляционного анализа. Уравнение при этом перепишется в виде

$$y = a + bx = \bar{y} + r\frac{s_Y}{s_X}(x - \bar{x}) \quad (3.16)$$

Коэффициенты a и bназываются выборочными коэффициентами регрессии и вычисляются по формулам

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

$$a = \bar{y} - b\bar{x} \quad (3.17)$$

Здесь

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (3.18)$$

Если использовать коэффициент корреляции, то можно вести вычисления по формулам

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\ s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3.19)$$

В случае корреляционной таблицы имеем

$$b = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} x_j y_i - n\bar{x}\bar{y}}{\sum_{j=1}^{k} w_j x_j^2 - \frac{\left(\sum_{j=1}^{k} w_j x_j\right)^2}{n}} \qquad (3.20)$$

$$a = \bar{y} - b\bar{x}$$

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} w_j x_j \,,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{m} p_i y_i \,,$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{j=1}^{k} w_j (x_j - \bar{x})^2}\ \ s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{m} p_i (y_i - \bar{y})^2} \qquad (3.21)$$

$$r = \hat{\rho}_{xy} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} x_j y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{j=1}^{k} w_j (x_j - \bar{x})^2 \sum_{i=1}^{m} p_i (y_i - \bar{y})^2}} \qquad (3.22)$$

При построении линейной регрессии X на Y уравнение ищут в виде

$x=a_1+b_1y$

При этом формулы для нахождения a_1 и b_1 приобретают вид:

$$\begin{cases} \sum_{i=1}^{n} x_i = n a_1 + b_1 \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i = a_1 \sum_{i=1}^{n} y_i + b_1 \sum_{i=1}^{n} y_i^2 \end{cases}$$

$$x = a_1 + b_1 y = \bar{x} + r\frac{s_X}{s_Y}(y - \bar{y})$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \qquad a_1 = \bar{x} - b_1\bar{y}$$

В случае корреляционной таблицы

$$b_1 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} x_j y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{m} p_i y_i^2 - \frac{\left(\sum_{j=1}^{m} p_i y_i\right)^2}{m}}$$

$$a_1 = \bar{x} - b_1\bar{y}$$

В экономической теории и других дисциплинах вводится понятие среднего коэффициента эластичности $К_э$, который показывает, на сколько процентов в среднем изменится показатель y от своего среднего значения при изменении фактора x на 1% от своей средней величины:

$$K_э = f'(x)\frac{\bar{x}}{\bar{y}} \qquad (3.23)$$

В случае линейной модели (f(x)=a+bx)

$$K_э = b\frac{\bar{x}}{\bar{y}}$$

Значимость коэффициентов регрессии

Статистические выводы относительно коэффициента β истинного уравнения регрессии y=θ+βxмогут быть получены с помощью статистики $t_\beta = \frac{b-\beta}{S_\beta}$, где

$$S_\beta = \frac{S}{s_x\sqrt{n-1}}; \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2;$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \ .$$

β – истинное значение коэффициента регрессии, b – выборочное значение коэффициента регрессии.

Статистика t_β при справедливости нулевой гипотезы $H_0 : \beta = b$ имеет t -распределение с n-2 степенями свободы [6].

С помощью квантилей распределения Стьюдента можно проверить гипотезу о равенстве коэффициента β выборочному значению, гипотезу о значимости коэффициента регрессии (существенности его отклонения от нуля), построить доверительный интервал для коэффициента β.

Значение коэффициента β является значимым с достоверностью α, если $|b| > t_{\frac{1+\alpha}{2}} S_\beta$.

Гипотеза о равенстве коэффициента β заданному значению β_0 принимается, если $|\beta - b| < t_{\frac{1+\alpha}{2}} S_\beta$.

Двусторонний $\alpha \cdot 100\% -$й доверительный интервал для β:

$$b - t_{\frac{1+\alpha}{2}} S_\beta < \beta < b + t_{\frac{1+\alpha}{2}} S_\beta$$

Статистические выводы о коэффициенте θмогут быть получены с помощью статистики

$$t_\alpha = \frac{a-\theta}{S_\alpha}, \quad \text{где} \quad S_\alpha = S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{(n-1)s_x^2}},$$

θ – истинное значение коэффициента регрессии, а – выборочное значение коэффициента регрессии.

Статистика t_α при справедливости нулевой гипотезы $H_0 : \theta = a$ имеет t -распределение с n-2 степенями свободы [6].

Проверка гипотезы о значимости коэффициента θи построение доверительного интервала для него выполняются по аналогии с коэффициентом β.

Значимость коэффициента корреляции

Для проверки значимости коэффициента корреляции выдвигается нулевая гипотеза, которая состоит в том, что коэффициент корреляции равен нулю при альтернативной гипотезе, что он отличен от нуля:

$$H_0 : \quad \rho = 0$$
$$H_1 : \quad \rho \neq 0$$

Очевидно, что чем больше по абсолютной величине значение величины r , тем больше оснований опровергнуть нулевую гипотезу.

Возникает вопрос.

Насколько большое должно быть абсолютное значение величины r?

Для того чтобы проверить гипотезуH_0, мы должны знать распределение величины r.

Проверка гипотезы основана на том факте, что величина $t = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}$ имеет распределение Стьюдента с n-2 степенями свободы[5].

При заданном уровне значимости α определяют критическое значение $t_{кр.}$

Если $t>t_{кр}$, то гипотеза H_0 отклоняется.

Если $t<t_{кр}$, то гипотеза H_0 принимается.

Доверительные интервалы для коэффициента корреляции ρ

Пусть выборка (x_i, y_i) получена из ГС, имеющей двумерное нормальное распределение и r - выборочный коэффициент корреляции. При достаточно больших n статистика

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = Arth(r)$$

имеет приближенно нормальное распределение $N(Arth(r), \frac{1}{\sqrt{n-3}})$.

Здесь Arth(w) – арктангенс гиперболический w .

Доверительный интервал для Arth(ρ) имеет вид:

$$Arth(r) - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} < Arth(\rho) < Arth(r) + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}$$

Здесь $u_{1-\frac{\alpha}{2}}$ – соответствующая квантиль нормального распределения.

Доверительный интервал для ρ вычисляется с помощью таблиц гиперболического тангенса ρ=th(z).

Некоторые полезные формулы. Arth(-r)=- Arth(r), еслиArth(y)=t, то $y = \frac{e^{2t}-1}{e^{2t}+1} = th(t)$. Значения функций Arth(y) и th(t) можно вычислить с помощью математических функций EXCEL (atanh(y) итanh(t) соответственно).

Построение модели линейной регрессии по несгруппированным данным

Проиллюстрируем изложенный материал на примерах. Решение задач проведено с помощью системы компьютерной математики Scilab версии 5.5.2 [9-11] и средств Microsoft Excel.

Однофакторный дисперсионный анализ

Многие приложения связаны с экспериментами, в которых рассматривается несколько групп или уровней одного фактора. Некоторые факторы, например температура обжига керамики, могут иметь несколько числовых уровней (т.е. 300°, 350°, 400° и 450°). Другие факторы, например местоположение товаров в супермаркете, могут иметь категориальные уровни (например, первый поставщик, второй поставщик, третий поставщик, четвертый поставщик). Однофакторные эксперименты, в ходе которых экспериментальные единицы случайным образом распределяются по группам или уровням фактора, называются полностью рандомизированными.

Использование F-критерия для оценки разностей между несколькими математическими ожиданиями

Если числовые измерения фактора в группах являются непрерывными и выполняются некоторые дополнительные условия, для сравнения математических ожиданий нескольких групп применяется дисперсионный анализ (ANOVA — **An**alysis **o**f **Va**riance). Дисперсионный анализ, использующий полностью рандомизированные планы, называется однофакторной процедурой ANOVA. В некотором смысле термин дисперсионный анализ является неточным, поскольку при этом анализе сравниваются разности между математическими ожиданиями групп, а не между дисперсиями. Однако сравнение математических ожиданий осуществляется именно на основе анализа вариации данных. В процедуре ANOVA полная вариация результатов измерений подразделяется на межгруппо-

вую и внутригрупповую (рис. 1). Внутригрупповая вариация объясняется ошибкой эксперимента, а межгрупповая — эффектами условий эксперимента. Символ с обозначает количество групп.

Рис. 1. Разделение вариации в полностью рандомизированном эксперименте

Предположим, что с групп извлечено из независимых генеральных совокупностей, имеющих нормальное распределение и одинаковую дисперсию. Нулевая гипотеза заключается в том, что математические ожидания генеральных совокупностей одинаковы: H_0: $\mu_1 = \mu_2 = ... = \mu_c$. Альтернативная гипотеза гласит, что не все математические ожидания одинаковы: H_1: не все μ_j одинаковы j = 1, 2, …, c).

На рис. 2 представлена истинная нулевая гипотеза о математических ожиданиях пяти сравниваемых групп при условии, что генеральные совокупности имеют нормальное распределение и одинаковую дисперсию. Пять генеральных совокупностей, связанных с разными уровнями фактора, идентичны. Следовательно, они накладываются одна на другую, имея одинаковые математическое ожидание, вариацию и форму.

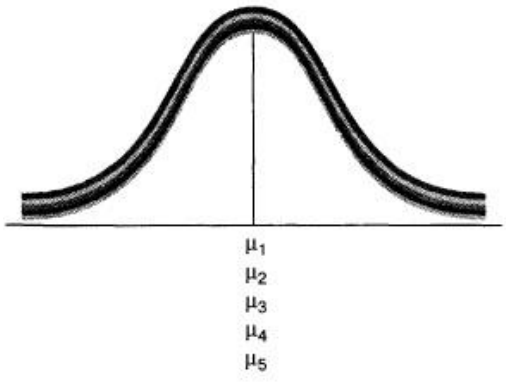


Рис. 2. Пять генеральных совокупностей имеют одинаковое математическое ожидание: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

С другой стороны, предположим, что на самом деле нулевая гипотеза является ложной, причем четвертый уровень имеет наибольшее математическое ожидание, первый уровень — чуть меньшее математическое ожидание, а остальные уровни — одинаковые и еще меньшие математические ожидания (рис. 3). Обратите внимание на то, что за исключением величины математических ожиданий все пять генеральных совокупностей идентичны (т.е. имеют одинаковую изменчивость и форму).

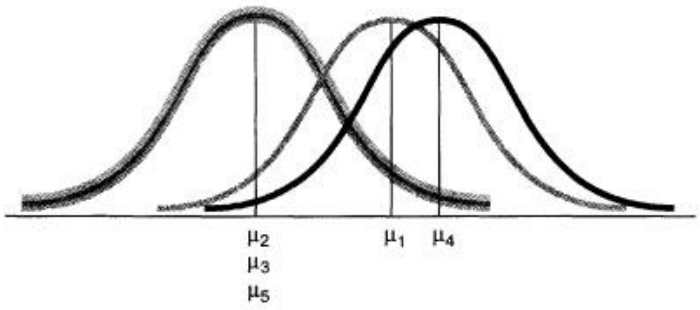


Рис. 3. Наблюдается эффект условий эксперимента: $\mu_4 > \mu_1 > \mu_2 = \mu_3 = \mu_5$

При проверке гипотезы о равенстве математических ожиданий нескольких генеральных совокупностей полная вариация разделяется на две части: межгрупповую вариацию, обусловленную разностями между группами, и внутригрупповую, обусловленную разностями между элементами, принадлежащими одной группе. Полная вариация выражается полной суммой квадратов (SST – sum of squares total). Поскольку нулевая гипотеза заключается в том, что математические ожидания всех с групп равны между собой, полная вариация равна сумме квадратов разностей между отдельными наблюдениями и общим средним (среднее средних) $\bar{\bar{X}}$, вычисленным по всем выборкам. Полная вариация:

$$(1)\quad SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \bar{\bar{X}})^2$$

где $\bar{\bar{X}} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j}X_{ij}$ — общее среднее, X_{ij} — i-е наблюдение в j-й группе или уровне, n_j — количество наблюдений в j-й группе, n — общее количество наблюдений во всех группах (т.е. $n = n_1 + n_2 + ... + n_c$), c — количество изучаемых групп или уровней.

Межгрупповая вариация, называемая обычно межгрупповой суммой квадратов (SSA – sum of squares among groups), равна сумме квадратов разностей между выборочным средним каждой группы $\bar{X}_j$ и общим средним $\bar{\bar{X}}$, умноженных на объем соответствующей группы n_j:

$$(2)\quad SSA = \sum_{j=1}^{c} n_j(\bar{X}_j - \bar{\bar{X}})^2$$

где c — количество изучаемых групп или уровней, n_j — количество наблюдений в j-й группе, $\bar{X}_j$ — среднее значение j-й группы, $\bar{\bar{X}}$ — общее среднее.

Внутригрупповая вариация, называемая обычно внутригрупповой суммой квадратов (SSW – sum of squares withing groups), равна сумме квадратов разностей между элементами каждой группы и выборочным средним этой группы $\bar{X}_j$:

$$(3)\quad SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$$

где X_{ij} — i-й элемент j-й группы, $\bar{X}_j$ — среднее значение j-й группы.

Поскольку сравнению подвергаются c уровней фактора, межгрупповая сумма квадратов имеет c – 1 степеней свободы. Каждый из c уровней обладает n_j – 1 степенями свободы, поэтому Внутригрупповая сумма квадратов имеет n – c степеней свободы, и

$$\sum_{j=1}^{c}(n_j - 1) = n - c$$

Кроме того, общая сумма квадратов имеет n – 1 степеней свободы, поскольку каждое наблюдение X_{ij} сравнивается с общим средним $\bar{\bar{X}}$, вычисленным по всем n наблюдениям. Если каждую из этих сумм разделить на соответствующее количество степеней свободы, возникнут три вида дисперсии: межгрупповая (mean square among — MSA), внутригрупповая (mean square within — MSW) и полная (mean square total — MST):

(4а) $MSA = \frac{SSA}{c-1}$ (4б) $MSW = \frac{SSW}{n-c}$ (4в) $MST = \frac{SST}{n-1}$

Несмотря на то, что основное предназначение дисперсионного анализа — сравнить математические ожидания с групп, чтобы выявить эффект условий эксперимента, его название обусловлено тем, что главным инструментом является анализ дисперсий разного типа. Если нулевая гипотеза является истинной, и между математическими ожиданиями с групп нет существенных различий, все три дисперсии — MSA, MSW и MST — являются оценками дисперсии σ^2, присущей анализируемым данным. Таким образом, чтобы проверить нулевую гипотезу H_0: $\mu_1 = \mu_2 = ... = \mu_c$ и альтернативную гипотезу H_1: не все μ_j одинаковы j = 1, 2, …, c), необходимо вычислить статистику F-критерия, представляющую собой отношение двух дисперсий, MSA и MSW. Тестовая F-статистика в однофакторном дисперсионном анализе

$$(5)\quad F = \frac{MSA}{MSW}$$

Статистика F-критерия подчиняется F-распределению с c – 1 степенями свободы в числителе MSA и n – c степенями свободы в знаменателе MSW. При заданном уровне значимости α нулевая гипотеза отклоняется, если вычисленная F-статистика больше верхнего критического значения F_U, присущего F-распределению с c – 1 степенями свободы в числителе и n – c степенями свободы в знаменателе. Таким образом, как показано на рис. 4, решающее правило формулируется следующим образом: нулевая гипотеза H_0 отклоняется, если $F > F_U$; в противном случае она не отклоняется.

Рис. 4. Критическая область дисперсионного анализа при проверке гипотезы H_0

Если нулевая гипотеза H_0 является истинной, вычисленная F-статистика близка к 1, поскольку ее числитель и знаменатель являются оценками одной и

той же величины — дисперсии σ^2, присущей анализируемым данным. Если нулевая гипотеза H_0 является ложной (и между математическими ожиданиями разных групп существует значительная разница), вычисленная F-статистика будет намного больше единицы, поскольку ее числитель, MSA, помимо естественной изменчивости данных, оценивает эффект условий эксперимента или разности между группами, в то время как знаменатель MSW оценивает лишь естественную изменчивость данных. Таким образом, процедура ANOVA представляет собой F-критерий, в котором при заданном уровне значимости α нулевая гипотеза отклоняется, если вычисленная F-статистика больше верхнего критического значения F_U, присущего F-распределению с c – 1 степенями свободы в числителе и n – c степенями свободы в знаменателе, как показано на рис. 4.

Для иллюстрации однофакторного дисперсионного анализа вернемся к сценарию, изложенному в начале заметки. Цель эксперимента — определить, имеют ли парашюты, сотканные из синтетического волокна, полученного от разных поставщиков, одинаковую прочность. В каждой из групп соткано по пять парашютов. Группы разделены по поставщикам— Поставщик 1, Поставщик 2, Поставщик 3 и Поставщик 4. Прочность парашютов измеряется с помощью специального устройства, испытывающего ткань на разрыв с двух сторон. Сила, необходимая для разрыва парашюта, измеряется по особой шкале. Чем выше сила разрыва, тем прочнее парашют. Пакет анализа Excel позволяет провести анализ F-статистики одним кликом. Пройдите по меню Данные → Анализ данных, и выберите строку Однофакторный дисперсионный анализ, заполните открывшееся окно (рис. 5). Результаты эксперимента (сила разрыва), некоторые описательные статистики и результаты однофакторного дисперсионного анализа представлены на рис. 6.

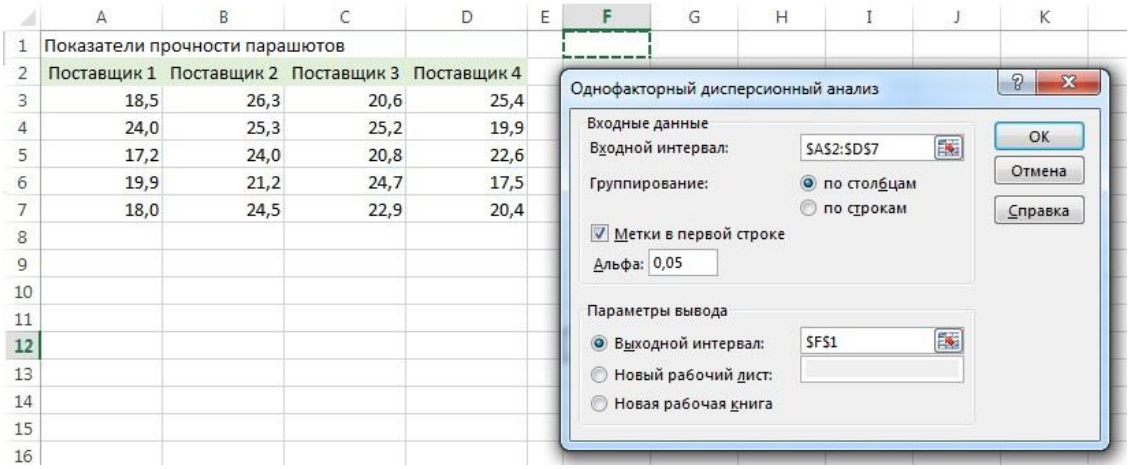


Рис. 5. Окно Однофакторный дисперсионный анализ Пакета анализа Excel

	A	B	C	D	E	F	G
1	**Показатели прочности парашютов**						
2	Поставщик 1	Поставщик 2	Поставщик 3	Поставщик 4			
3	18,5	26,3	20,6	25,4			
4	24,0	25,3	25,2	19,9			
5	17,2	24,0	20,8	22,6			
6	19,9	21,2	24,7	17,5			
7	18,0	24,5	22,9	20,4			
8							
9	**Однофакторный дисперсионный анализ**						
10	ИТОГИ						
11	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
12	Поставщик 1	5	97,6	19,52	7,237		
13	Поставщик 2	5	121,3	24,26	3,683		
14	Поставщик 3	5	114,2	22,84	4,553		
15	Поставщик 4	5	105,8	21,16	8,903		
16							
17	Дисперсионный анализ						
18	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
19	Между группами	63,286	3	21,095	3,462	0,041	3,239
20	Внутри групп	97,504	16	6,094			
21	Итого	160,790	19				

Рис. 6. Показатели прочности парашютов, сотканных из синтетических волокон, полученных от разных поставщиков, описательные статистики и результаты однофакторного дисперсионного анализа

Анализ рисунка 6 показывает, что между выборочными средними наблюдается некоторая разница. Средняя прочность волокон, полученных от первого поставщика, равна 19,52, от второго — 24,26, от третьего — 22,84 и от четвертого — 21,16. Можно ли назвать эту разницу статистически значимой? Распределение силы разрыва продемонстрировано на диаграмме разброса (рис. 7). На ней ясно видны разности как между группами, так и внутри них. Если бы объем каждой

группы был больше, для их анализа можно было бы применить диаграмму «ствол и листья», блочную диаграмму или график нормального распределения.

	A	B
1	**Показатели прочности парашютов**	
2	Поставщик	Прочность на разрыв
3	1	18,5
4	1	24,0
5	1	17,2
6	1	19,9
7	1	18,0
8	2	26,3
9	2	25,3
10	2	24,0
11	2	21,2
12	2	24,5
13	3	20,6
14	3	25,2
15	3	20,8
16	3	24,7
17	3	22,9
18	4	25,4
19	4	19,9
20	4	22,6
21	4	17,5
22	4	20,4

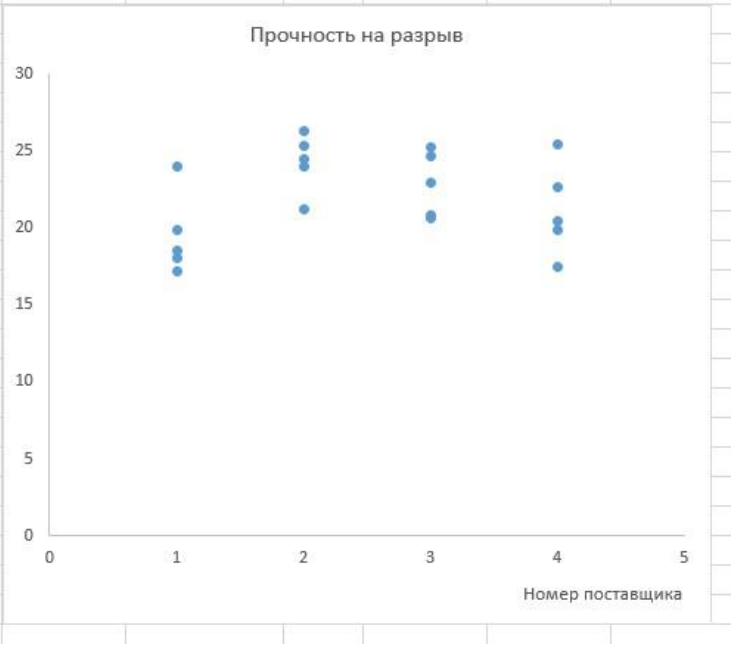


Рис. 7. Диаграмма разброса прочности парашютов, сотканных из синтетических волокон, полученных от четырех поставщиков

Нулевая гипотеза утверждает, что между средними показателями прочности нет существенных различий: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Альтернативная гипотеза заключается в том, что существует по крайней мере один поставщик, у которого средняя прочность волокон отличается от других: H_1: не все μ_j одинаковы j = 1, 2, …, c).

Общее среднее (см. рис. 6) $\bar{\bar{X}}$ =СРЗНАЧ(D12:D15) = 21,945; для определения $\bar{\bar{X}}$ также можно усреднить все 20 исходных чисел: $\bar{\bar{X}}$ =СРЗНАЧ(A3:D7). Значения дисперсий рассчитываются Пакетом анализа и отражаются в табличке Дисперсионный анализ (см. рис. 6): SSA = 63,286, SSW = 97,504, SST = 160,790 (см. колонку SS таблицы Дисперсионный анализ рисунка 6). Средние значения вычисляются путем деления этих сумм квадратов на соответствующее количество степеней свободы. Поскольку c = 4, а n = 20, получаем следующие значения степеней свободы; для SSA: c – 1 = 3; для SSW: n – c = 16; для SST: n – 1 = 19

(см. колонку df). Таким образом: MSA = SSA / (c – 1) = 21,095; MSW = SSW / (n – c) = 6,094; MST = SST / (n – 1) = 8,463 (см. колонку MS). F-статистика = MSA / MSW = 3,462 (см. колонку F).

Верхнее критическое значение F_U, характерное для F-распределения, определяется по формуле =F.ОБР(0,95;3;16) = 3,239. Параметры функции =F.ОБР(): α = 0,05, числитель имеет три степени свободы, а знаменатель — 16. Таким образом, вычисленная F-статистика, равная 3,462, превышает верхнее критическое значение F_U = 3,239, нулевая гипотеза отклоняется (рис. 8).

Рис. 8. Критическая область дисперсионного анализа при уровне значимости, равном 0,05, если числитель имеет три степени свободы, а знаменатель —16

p-значение, т.е. вероятность того, что при истинной нулевой гипотезе F-статистика не меньше 3,46, равно 0,041 или 4,1% (см. колонку p-Значение таблицы Дисперсионный анализ рисунка 6). Поскольку эта величина не превышает уровень значимости α = 5%, нулевая гипотеза отклоняется. Более того, p-значение свидетельствует о том, что вероятность обнаружить такую или большую разность между математическими ожиданиями генеральных совокупностей при условии, что на самом деле они одинаковы, равна 4,1%.

Итак. Между четырьмя выборочными средними существует разница. Нулевая гипотеза заключалась в том, что все математические ожидания четырех генеральных совокупностей равны между собой. В этих условиях мера полной изменчивости (т.е. полная вариация SST) прочности всех парашютов вычисляется

путем суммирования квадратов разностей между каждым наблюдением X_{ij} и общим средним $\bar{\bar{X}}$. Затем полная вариация разделялась на два компонента (см. рис. 1). Первый компонент представлял собой межгрупповую вариацию SSA, а второй — внутригрупповую SSW.

Чем объясняется изменчивость данных? Иначе говоря, почему все наблюдения не одинаковы? Одна из причин заключается в том, что разные фирмы поставляют волокна разной прочности. Это частично объясняет, почему группы имеют разные математические ожидания: чем сильнее эффект условий эксперимента, тем больше разность между математическими ожиданиями групп. Другой причиной изменчивости данных является естественная изменчивость любого процесса, в данном случае — производства парашютов. Даже если бы все волокна приобретались у одного и того же поставщика, их прочность была бы неодинаковой при прочих равных условиях. Поскольку этот эффект проявляется в каждой из групп, он называется внутригрупповой вариацией.

Разности между выборочными средними называются межгрупповой вариацией SSA. Часть внутригрупповой вариации, как уже указывалось, объясняется принадлежностью данных разным группам. Однако даже если бы группы были совершенно одинаковыми (т.е. нулевая гипотеза была бы истинной), межгрупповая вариация все равно существовала. Причина этого заключается в естественной изменчивости процесса производства парашютов. Поскольку выборки разные, их выборочные средние отличаются друг от друга. Следовательно, если нулевая гипотеза является истинной, как межгрупповая, так и внутригрупповая изменчивость представляют собой оценку изменчивости генеральной совокупности. Если нулевая гипотеза является ложной, межгрупповая гипотеза будет больше. Именно этот факт лежит в основе F-критерия для сравнения разностей между математическими ожиданиями нескольких групп.

После выполнения однофакторного дисперсионного анализа и обнаружения значительной разницы между фирмами остается неизвестным, какой же из

поставщиков существенно отличается от остальных. Нам известно лишь, что математические ожидания генеральных совокупностей не равны. Иначе говоря, по крайней мере одно из математических ожиданий существенно отличается от других. Чтобы определить, какой из поставщиков отличается от других, можно воспользоваться процедурой Тьюки, использующей попарное сравнение между поставщиками. Эта процедура была разработана Джоном Тьюки. Впоследствии он и К. Крамер независимо друг от друга модифицировали эту процедуру для ситуаций, в которых объемы выборок отличаются друг от друга.

Множественное сравнение: процедура Тьюки-Крамера

В нашем сценарии для сравнения прочности парашютов использовался однофакторный дисперсионный анализ. Обнаружив значительные различия между математическими ожиданиями четырех групп, необходимо определить, какие именно группы отличаются друг от друга. Хотя существует несколько способов решить эту задачу, мы опишем лишь процедуру множественного сравнения Тьюки-Крамера. Этот метод является примером процедур апостериорного сравнения (post hoc comparison), поскольку проверяемая гипотеза формулируется после анализа данных. Процедура Тьюки-Крамера позволяет одновременно сравнить все пары групп. На первом этапе вычисляются разности $X_j - X_{j'}$, где $j \neq j'$, между математическими ожиданиями $c(c-1)/2$ групп. Критический размах процедуры Тьюки-Крамера вычисляется по формуле:

(6) Критический размах $= Q_U \sqrt{\frac{MSW}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$

где Q_U — верхнее критическое значение распределения стьюдентизированного размаха, имеющего c степеней свободы в числителе и n – c степеней свободы в знаменателе.

Если объемы выборок не одинаковы, критический размах вычисляется для каждой пары математических ожиданий отдельно. На последнем этапе каждая из

c(c – 1)/2 пар математических ожиданий сравнивается с соответствующим критическим размахом. Элементы пары считаются значимо различными, если модуль разности $|X_j – X_{j'}|$ между ними превышает критический размах.

Применим процедуру Тьюки-Крамера к задаче о прочности парашютов. Поскольку компания, производящая парашюты, имеет четыре поставщика, следует проверить 4(4 – 1)/2 = 6 пар поставщиков (рис. 9).

	A	B	C	D	E
1	**Показатели прочности парашютов**				
2	№ испытания	Поставщик 1	Поставщик 2	Поставщик 3	Поставщик 4
3	1	18,5	26,3	20,6	25,4
4	2	24,0	25,3	25,2	19,9
5	3	17,2	24,0	20,8	22,6
6	4	19,9	21,2	24,7	17,5
7	5	18,0	24,5	22,9	20,4
8	Среднее	19,5	24,3	22,8	21,2
9					
10	Попарные сравнения выборочных средних				
11	Пары поставщиков	X_j	$X_{j'}$	$\|X_j - X_{j'}\|$	
12	1 и 2	19,5	24,3	4,74	
13	1 и 3	19,5	22,8	3,32	
14	1 и 4	19,5	21,2	1,64	
15	2 и 3	24,3	22,8	1,42	
16	2 и 4	24,3	21,2	3,10	
17	3 и 4	22,8	21,2	1,68	
18					

Рис. 9. Попарные сравнения выборочных средних

Поскольку все группы имеют одинаковый объем (т.е. все $n_j = n_{j'}$), достаточно вычислить только один критический размах. Для этого по таблице Дисперсионного анализа (рис. 6) определим величину MSW = 6,094. Затем найдем величину Q_U при α = 0,05, c = 4 (число степеней свободы в числителе) и n – c = 20 – 4 = 16 (число степеней свободы в знаменателе). К сожалению, я не нашел соответствующей функции в Excel, так что воспользовался таблицей (рис. 10).

Верхние 5% значений (α=0,05)

Знаменатель, df_2	Числитель, df_1											
	2	3	4	5	6	7	8	9	10	11	12	13
1	18,00	27,00	32,80	37,10	40,40	43,10	45,40	47,40	49,10	50,60	52,00	53,20
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55
15	3,01	3,67	4,08	4,37	4,60	4,78	4,94	5,08	5,20	5,31	5,40	5,49
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44
17	2,98	3,63	4,02	4,30	4,52	4,71	4,86	4,99	5,11	5,21	5,31	5,39
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27	5,35
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,32

Рис. 10. Критическое значение стьюдентизированного размаха Q_U

Получаем:

$$\text{Критический размах} = 4{,}05\sqrt{\frac{6{,}094}{2}\left(\frac{1}{5}+\frac{1}{5}\right)} = 4{,}47$$

Поскольку лишь 4,74 > 4,47 (см. нижнюю таблицу рис. 9), статистически значимая разница существует между первым и вторым поставщиком. Все остальные пары имеют выборочные средние, которые не позволяют говорить о их различии. Следовательно, средняя прочность парашютов, сотканных из волокон, приобретенных у первого поставщика, значимо меньше, чем у второго.

Необходимые условия однофакторного дисперсионного анализа

При решении задачи о прочности парашютов мы не проверяли, выполняются ли условия, при которых можно использовать однофакторный F-критерий. Как же узнать, можно ли применять однофакторный F-критерий при анализе конкретных экспериментальных данных? Однофакторный F-критерий можно применять, только если выполняются три основных предположения: экспериментальные данные должны быть случайными и независимыми, иметь нормальное распределение, а их дисперсии должны быть одинаковыми.

Первое предположение — случайность и независимость данных — должно выполняться всегда, поскольку корректность любого эксперимента зависит от случайности выбора и/или процесса рандомизации. Чтобы избежать искажения результатов, необходимо, чтобы данные извлекались из с генеральных совокупностей случайно и независимо друг от друга. Аналогично данные должны быть случайным образом распределенными по с уровням интересующего нас фактора (экспериментальным группам). Нарушение этих условий может серьезно исказить результаты дисперсионного анализа.

Второе предположение — нормальность — означает, что данные извлечены из нормально распределенных генеральных совокупностей. Как и для t-критерия, однофакторный дисперсионный анализ на основе F-критерия относительно мало чувствителен к нарушению этого условия. Если распределение не слишком значительно отличается от нормального, уровень значимости F-критерия изменяется мало, особенно если объем выборок достаточно велик. Если же условие о нормальности распределения нарушается серьезно, следует применять непараметрические процедуры дисперсионного анализа (будут рассмотрены позже).

Третье предположение — однородность дисперсии — означает, что дисперсии каждой генеральной совокупности равны между собой (т.е. $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_j^2$). Это предположение позволяет решить, разделять или объединять внутригрупповые дисперсии. Если объемы групп совпадают, условие однородности дисперсии слабо влияет на выводы, полученные с помощью F-критерия. Однако, если объемы выборок неодинаковы, нарушение условия о равенстве дисперсий может серьезно исказить результаты дисперсионного анализа. Таким образом, следует стремиться к тому, чтобы объемы выборок были одинаковыми. Одним из методов проверки предположения об однородности дисперсии является критерий Левенэ, описанный ниже.

Если из всех трех условий нарушается лишь условие об однородности дисперсии, можно применять процедуру, аналогичную t-критерию, использующему

раздельную дисперсию. Однако, если предположения о нормальном распределении и однородности дисперсии нарушаются одновременно, необходимо выполнить нормализацию данных и уменьшить разности между дисперсиями или применить непараметрическую процедуру.

Критерий Левенэ для проверки однородности дисперсии

Несмотря на то, что F-критерий относительно устойчив к нарушениям условия о равенстве дисперсий в группах, грубое нарушение этого предположения существенно влияет на уровень значимости и мощность критерия. Возможно, одним из наиболее мощных является критерий Левенэ. Для проверки равенства дисперсий с генеральных совокупностей проверим следующие гипотезы:

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_j^2$

H_1: не все σ_j^2 одинаковы (j = 1, 2, …, c)

Модифицированный критерий Левенэ основан на утверждении, что если изменчивость в группах одинакова, для проверки нулевой гипотезы о равенстве дисперсий можно применить анализ дисперсии абсолютных величин разностей между наблюдениями и медианами групп. Итак, сначала следует вычислить абсолютные величины разностей между наблюдениями и медианами в каждой группе, а затем выполнить однофакторный дисперсионный анализ полученных абсолютных величин разностей. Для иллюстрации критерия Левенэ вернемся к сценарию, изложенному в начале заметки. Используя данные, представленные на рис. 6, проведем аналогичный анализ, но в отношении модулей разниц исходных данных и медиан по каждой выборке отдельно (рис. 11).

	A	B	C	D	E	F	G
1	**Показатели прочности парашютов**						
2	№ испытания	Поставщик 1	Поставщик 2	Поставщик 3	Поставщик 4		
3	1	18,5	26,3	20,6	25,4		
4	2	24,0	25,3	25,2	19,9		
5	3	17,2	24,0	20,8	22,6		
6	4	19,9	21,2	24,7	17,5		
7	5	18,0	24,5	22,9	20,4		
8	Медиана	18,5	24,5	22,9	20,4		
9							
10	Модули разности исходных данных и медианного значения						
11	№ испытания	Поставщик 1	Поставщик 2	Поставщик 3	Поставщик 4		
12	1	0,0	1,8	2,3	5,0		
13	2	5,5	0,8	2,3	0,5		
14	3	1,3	0,5	2,1	2,2		
15	4	1,4	3,3	1,8	2,9		
16	5	0,5	0,0	0,0	0,0		
17							
18	Однофакторный дисперсионный анализ						
19	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
20	Поставщик 1	5	8,70	1,74	4,75		
21	Поставщик 2	5	6,40	1,28	1,71		
22	Поставщик 3	5	8,50	1,70	0,94		
23	Поставщик 4	5	10,60	2,12	4,01		
24							
25	Дисперсионный анализ						
26	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
27	Между группами	1,770	3	0,590	0,207	0,890	3,239
28	Внутри групп	45,648	16	2,853			
29							
30	Итого	47,418	19				
31							

Таблица 11. Абсолютные величины разностей между медианами прочности волокон и наблюдениями в каждой группе поставщиков

Двухфакторный дисперсионный анализ

Ранее был рассмотрен полностью рандомизированный эксперимент и связанный с ним однофакторный дисперсионный анализ. В настоящей заметке будет изучен двухфакторный дисперсионный анализ, в ходе которого одновременно оцениваются два фактора. Мы рассмотрим лишь ситуации, в которых выборки имеют одинаковый объем n'.[1]

Применение статистики в этой заметке будет показано на сквозном примере. Предположим, что вы руководитель производства в компании Perfect

Parachute («Идеальный парашют»). Парашюты изготавливаются из синтетических волокон, поставляемых четырьмя разными поставщиками. Совершенно очевидно, что одной из основных характеристик парашюта является его прочность. Вам необходимо убедиться, что все поставляемые волокна обладают одинаковой прочностью. Более того, на фабрике используется два вида ткацких станков: Jetta и Turk. Можно ли утверждать, что парашюты, изготовленные на станке фирмы Jetta, так же прочны, как и парашюты, произведенные на станках компании Turk? Существует ли разница между прочностью парашютов, сотканных из синтетических волокон разных поставщиков на разных станках? Чтобы ответить на этот вопрос, следует разработать схему эксперимента, в ходе которого измеряется прочность парашютов, сотканных из синтетических волокон разных поставщиков на разных станках. Информация, полученная в ходе этого эксперимента, позволит определить, какой поставщик и какой тип станка обеспечивают наибольшую прочность парашютов.

Вследствие сложности вычислений, особенно при большом количестве уровней каждого фактора и реплик, для двухфакторного анализа следует применять либо Excel, либо специализированное программное обеспечение. В двухфакторном эксперименте факторы A и B считаются взаимодействующими, если эффект фактора A зависит от уровня фактора B. Напомним, что в полностью рандомизированном плане полная сумма квадратов (SST) подразделяется на межгрупповую сумму квадратов (SSA) и внутригрупповую сумму квадратов (SSW). В двухфакторном эксперименте с одинаковым количеством реплик в каждой ячейке полная вариация (SST) подразделяется на сумму квадратов, соответствующую фактору A (SSA), сумму квадратов, соответствующую фактору B (SSB), сумму квадратов, учитывающую взаимодействие факторов A и B (SSAB), и сумму квадратов, возникающую вследствие случайной ошибки (SSE) (рис. 1).

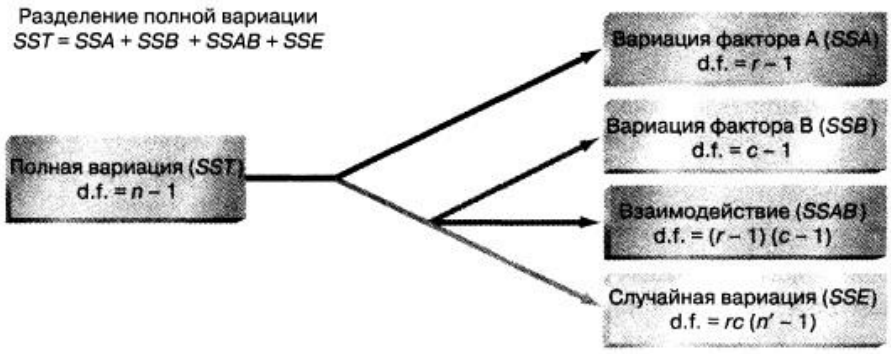


Рис. 1. Разделение полной вариации в двухфакторном эксперименте

В двухфакторном дисперсионном анализе применяются три разных критерия:

1. Для проверки гипотезы об отсутствии эффекта фактора A
2. Для проверки гипотезы об отсутствии эффекта фактора B
3. Для проверки гипотезы об отсутствии эффекта взаимодействия факторов A и B (рис. 2).

Каждая из трех нулевых гипотез отклоняется, если при заданном уровне значимости α соответствующая F-статистика (см. последнюю колонку рис. 2) больше верхнего критического значения F-распределения F_U.

Источник вариации	Количество степеней свободы	Сумма квадратов	Дисперсия	F
A	$r-1$	SSA	$MSA = SSA/(r-1)$	$F=MSA/MSE$
B	$c-1$	SSB	$MSB = SSB/(c-1)$	$F=MSB/MSE$
AB	$(r-1)(c-1)$	$SSAB$	$MSAB = SSAB/(r-1)(c-1)$	$F=MSAB/MSE$
Ошибка	$rc(n'-1)$	SSE	$MSE = SSE/rc(n'-1)$	
Всего	$n-1$	SST		

Рис. 2. Дисперсионный анализ в двухфакторном эксперименте

Для иллюстрации двухфакторного дисперсионного анализа вернемся к нашему сценарию. Допустим, что, будучи руководителем производства, вы решили сравнить поставщиков синтетических волокон, и оценить, на каком из станков выпускаются более прочные парашюты: Jetta или Turk. Кроме того, необходимо определить, зависит ли разница между четырьмя поставщиками от типа станков, на которых производятся парашюты. Итак, необходимо разрабо-

тать план эксперимента, в котором каждому поставщику и типу станка соответствует пять парашютов (рис. 3). Для проведения анализа пройдите по меню Данные → Анализ данных и выберите строку Двухфакторный дисперсионный анализ с повторениями.

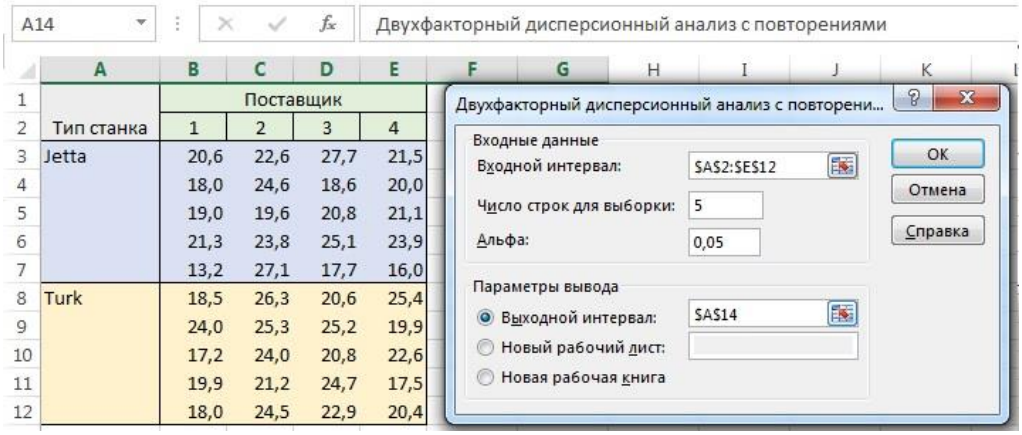


Рис. 3. Двухфакторный дисперсионный анализ с повторениями в Пакете анализа Excel

На рис. 4 показаны результаты двухфакторного дисперсионного анализа данных: объем выборки, сумма, арифметическое среднее и дисперсия каждой комбинации типа станка и поставщика. В первых двух таблицах приведены результаты дисперсионного анализа для всех типов станка, а в третьей — для каждого поставщика. В сводной таблице дисперсионного анализа идентификатор df обозначает количество степеней свободы, SS — сумму квадратов, MS — среднее квадратичное отклонение, F — вычисленную F-статистику.

	A	B	C	D	E	F	G
13							
14	**Двухфакторный дисперсионный анализ с повторениями**						
15	ИТОГИ	1	2	3	4	Итого	
16	*Jetta*						
17	Счет	5	5	5	5	20	
18	Сумма	92,10	117,70	109,90	102,50	422,20	
19	Среднее	18,42	23,54	21,98	20,50	21,11	
20	Дисперсия	10,20	7,57	18,40	8,36	13,13	
21							
22	*Turk*						
23	Счет	5	5	5	5	20	
24	Сумма	97,60	121,30	114,20	105,80	438,90	
25	Среднее	19,52	24,26	22,84	21,16	21,95	
26	Дисперсия	7,24	3,68	4,55	8,90	8,46	
27							
28	*Итого*						
29	Счет	10	10	10	10		
30	Сумма	189,70	239,00	224,10	208,30		
31	Среднее	18,97	23,90	22,41	20,83		
32	Дисперсия	8,09	5,14	10,41	7,79		
33							
34	**Дисперсионный анализ**						
35	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
36	Выборка	6,97	1	6,97	0,81	0,37	4,15
37	Столбцы	134,35	3	44,78	5,20	0,00	2,90
38	Взаимодействие	0,29	3	0,10	0,01	1,00	2,90
39	Внутри	275,59	32	8,61			
40							
41	Итого	417,20	39				

Рис. 4. Результат двухфакторного дисперсионного анализа прочности парашютов

Чтобы проанализировать эти результаты, сначала следует проверить, существует ли взаимодействие между факторами A (типами станка) и B (поставщиками). Если эффект взаимодействия является значительным, дальнейший анализ ограничивается лишь оценкой этого эффекта. С другой стороны, если эффект взаимодействия незначителен, необходимо сосредоточиться на главных эффектах — потенциальных различиях между типами станков (фактор A) и поставщиками (фактор B).

Чтобы определить наличие эффекта взаимодействия при уровне значимости, равном 0,05, применяется следующее решающее правило: нулевая гипотеза об отсутствии эффекта взаимодействия отклоняется, если вычисленное значение F-статистики (см. таблицу Дисперсионный анализ, строку Взаимодействие столбец F на рис. 4), больше верхнего критического значения F-распределения (там

же, столбец F-критическое). Поскольку $F = 0{,}01 < F_U = 2{,}90$, а p-значение равно 0,998, гипотеза H_0 не отклоняется. Следовательно, у нас недостаточно оснований утверждать, что факторы станка и поставщика взаимодействуют друг с другом. Следовательно, необходимо проанализировать главные эффекты.

При заданном уровне значимости, равном 0,05, в основе проверки разности между двумя станками (фактор A) лежит следующее решающее правило: нулевая гипотеза отклоняется, если вычисленное значение F-статистики больше верхнего критического значения F-распределения (см. таблицу Дисперсионный анализ, строку Выборка на рис. 4). Поскольку $F = 0{,}81 < F_U = 4{,}15$, а p-значение равно 0,37 и больше уровня значимости $\alpha = 0{,}05$, гипотеза H_0 не отклоняется. Следовательно, у нас недостаточно оснований утверждать, что между прочностью парашютов, произведенных на разных станках, существует значимая разница.

При заданном уровне значимости, равном 0,05, в основе проверки разности между поставщиками (фактор B) лежит следующее решающее правило: нулевая гипотеза отклоняется, если вычисленное значение F-статистики больше верхнего критического значения F-распределения (см. таблицу Дисперсионный анализ, строку Столбцы на рис. 4). Поскольку $F = 5{,}20 > F_U = 2{,}92$, а p-значение равно 0,005 и меньше уровня значимости, гипотеза H_0 отклоняется. Следовательно, можно утверждать, что между прочностью парашютов, произведенных из волокна, приобретенного у разных поставщиков, существует значимая разница.[2]

Интерпретация эффектов взаимодействия

Чтобы лучше разобраться во взаимодействии факторов, следует построить график средних значений в ячейках (т.е. средних значений, соответствующих конкретным уровням факторов), как показано на рис. 5 (в качестве данных для построения графика использованы области B19:E19 и B25:E25 рис. 4). Из графика средней прочности для каждой комбинации станок–поставщик следует, что две линии, соответствующие разным станкам, проходят почти параллельно друг

другу. Это означает, что разности между средними величинами прочности парашютов, произведенных на разных станках, практически одинаковы для всех четырех поставщиков. Иначе говоря, между этими двумя факторами нет связи, что полностью подтверждается F-критерием.

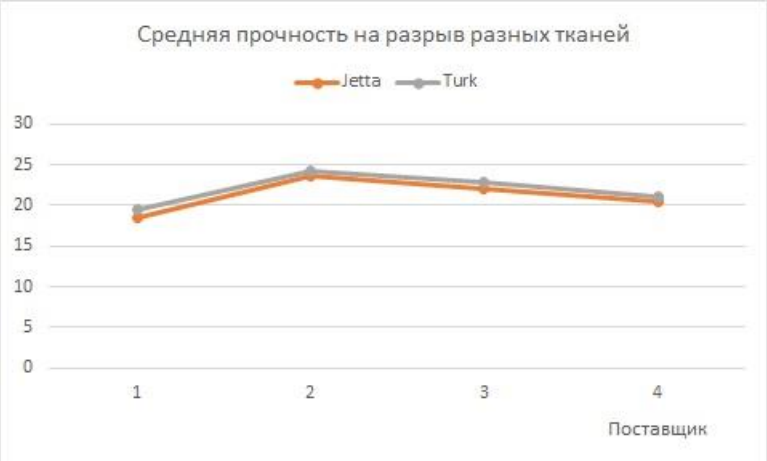


Рис. 5. График средних значений прочности парашютов в зависимости от станков и поставщиков

В чем проявляется эффект взаимодействия? В некоторых ситуациях определенные уровни фактора А могут оказаться связанными с конкретными уровнями фактора В. Например, предположим, что некоторые парашюты оказываются более прочными, если они сотканы из определенных волокон на станках Jetta, а другие — если они сотканы из волокон других поставщиков на станках Turk. Если бы это было правдой, линии на рис. 5 не были бы параллельными и взаимодействие между факторами было бы статистически значимым. Следовательно, в этих ситуациях разница между станками не будет одинаковой при разных поставщиках. Это усложняет интерпретацию главных эффектов, поскольку разности, соответствующие одному фактору (например, типу станка), не согласуются с другим фактором (например, поставщиком). Проиллюстрируем эту ситуацию следующим примером.

Пример.1. Интерпретация статистически значимых эффектов взаимодействия. Данные, приведенные на рис. 6а, характеризуют продолжительность ра-

боты подшипников под воздействием двух факторов: автоколебания и нагревания. Как влияют автоколебания и нагревание на продолжительность работы подшипников? Результаты двухфакторного дисперсионного анализа продолжительности работы подшипников, полученные с помощью Пакета анализа в Excel приведены на рис. 6б. Обратите внимание на то, что, кроме сводной таблицы дисперсионного анализа, Excel вычисляет среднее значение для каждой комбинации двух факторов: степени автоколебаний и нагревания, а также среднее значение для каждого уровня факторов. Для того чтобы проанализировать эти результаты, сначала необходимо определить, наблюдается ли статистически значимый эффект взаимодействия факторов автоколебания (фактор А) и нагревания (фактор В). При уровне значимости $\alpha = 0{,}05$ нулевую гипотезу об отсутствии эффекта взаимодействия следует отклонить, поскольку p-значение равно 0,0018, т.е. меньше 0,05. Кроме того, F-статистика равна 53,78 и превышает величину 7,71 — верхнее критическое значение F-распределения с одной степенью свободы в числителе и четырьмя степенями свободы в знаменателе.

	A	B	C	D	E	F	G
1	**Продолжительность работы подшипников при автоколебании и нагревании**						
2		Слабое нагревание	Сильное нагревание				
3	Слабое автоколебание	12	26				
4		24	16				
5	Сильное автоколебани	18	101				
6		28	113				
7							
8	**Двухфакторный дисперсионный анализ с повторениями**						
9	ИТОГИ	Слабое нагревание	Сильное нагревание	Итого			
10	*Слабое автоколебание*						
11	Счет	2	2	4			
12	Сумма	36	42	78			
13	Среднее	18	21	19,5			
14	Дисперсия	72	50	43,667			
15							
16	*Сильное автоколебание*						
17	Счет	2	2	4			
18	Сумма	46	214	260			
19	Среднее	23	107	65			
20	Дисперсия	50	72	2392,667			
21							
22	*Итого*						
23	Счет	4	4				
24	Сумма	82	256				
25	Среднее	20,5	64				
26	Дисперсия	49	2506				
27							
28	Дисперсионный анализ						
29	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
30	Выборка	4 140,5	1	4140,5	67,877	0,001	7,709
31	Столбцы	3 784,5	1	3784,5	62,041	0,001	7,709
32	Взаимодействие	3 280,5	1	3280,5	53,779	0,002	7,709
33	Внутри	244	4	61			
34							
35	Итого	11 449,5	7				
36							

а

б

Рис. 6. (а) Продолжительность работы подшипников при автоколебании и нагревании; (б) Результаты двухфакторного дисперсионного анализа продолжительности работы подшипников

Значимый эффект взаимодействия между автоколебанием и нагреванием можно проследить на рис. 7. Поскольку графики средних значений продолжительности работы подшипников при слабом и сильном нагревании, соответствующие двум степеням автоколебаний, не параллельны, разности между средними значениями продолжительности работы при двух типах автоколебаний и двух степенях нагревания неодинаковы. Наличие эффекта взаимодействия факторов усложняет анализ основных эффектов. Теперь невозможно определить, суще-

ствует ли статистически значимая разница между средними продолжительностями работы подшипников при слабых и сильных автоколебаниях, поскольку при разных степенях нагревания эта разность неодинакова. Аналогично невозможно определить, существует ли статистически значимая разница между средними продолжительностями работы подшипников при слабом и сильном нагревании, поскольку при разных степенях автоколебаний эта разность неодинакова.

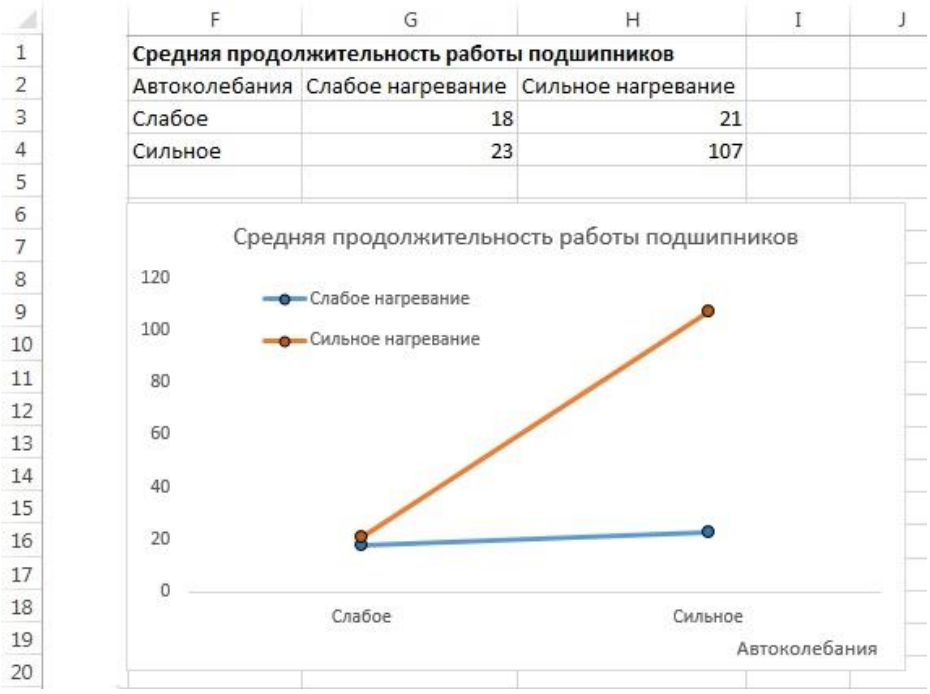


Рис. 7. График средних значений продолжительности работы подшипников по ячейкам

Множественные сравнения

Если эффект взаимодействия факторов не важен, для множественного сравнения нескольких факторов можно применять процедуру Тьюки-Крамера.

Критический размах процедуры Тьюки-Крамера для фактора A

(1) Критический размах = $Q_U \sqrt{\frac{MSE}{cn'}}$

где Q_U — верхнее критическое значение распределения стьюдентизированного размаха, имеющего r степеней свободы в числителе и rc(n' – 1) степеней свободы в знаменателе.

Критический размах процедуры Тьюки-Крамера для фактора B

(2) Критический размах = $Q_U \sqrt{\frac{MSE}{rn'}}$

где Q_U — верхнее критическое значение распределения стьюдентизированного размаха, имеющего c степеней свободы в числителе и rc(n' – 1) степеней свободы в знаменателе.

Применим процедуру Тьюки-Крамера к задаче о прочности парашютов (см. рис. 3). Анализ сводной таблицы дисперсионного анализа, представленной на рис. 4, показывает, что статистически значимым является лишь один главный эффект. При уровне значимости, равном 0,05, нет оснований утверждать, что между двумя типами станков (Jetta и Turk) существует значимая разница (фактор A), однако между четырьмя поставщиками (фактор B) эта разница существует. Таким образом, дальнейший анализ должен концентрироваться на разностях между поставщиками.

Поскольку компания, производящая парашюты, имеет четыре фирмы-поставщика, следует проверить 4(4 – 1)/2 = 6 пар поставщиков (рис. 8а). Вычислим модули разности между соответствующими средними значениями по выборкам отдельных поставщиков (рис. 8б).

	A	B	C	D	E	F
1		Поставщик				
2	Тип станка	1	2	3	4	
3	Jetta	20,6	22,6	27,7	21,5	
4		18,0	24,6	18,6	20,0	
5		19,0	19,6	20,8	21,1	
6		21,3	23,8	25,1	23,9	а
7		13,2	27,1	17,7	16,0	
8	Turk	18,5	26,3	20,6	25,4	
9		24,0	25,3	25,2	19,9	
10		17,2	24,0	20,8	22,6	
11		19,9	21,2	24,7	17,5	
12		18,0	24,5	22,9	20,4	
13	Среднее	18,97	23,90	22,41	20,83	
14						
15	Попарные сравнения выборочных средних					
16	Пары поста	X_j	$X_{j'}$	$\lvert X_j - X_{j'} \rvert$		
17	1 и 2	18,97	23,90	4,93		
18	1 и 3	18,97	22,41	3,44		б
19	1 и 4	18,97	20,83	1,86		
20	2 и 3	23,90	22,41	1,49		
21	2 и 4	23,90	20,83	3,07		
22	3 и 4	22,41	20,83	1,58		
23						

Рис. 8. (а) Исходные данные о прочности парашютов; (б) попарные сравнения средних значений по выборкам отдельных поставщиков

Чтобы вычислить критический размах, обратимся к данным на рис. 4: MSE = 8,61, r = 2, c = 4, n’ = 5, rc(n' – 1) = 32. При α = 0,05, c = 4 и rc(n' – 1) = 32 по таблицам размаха (рис. 9) определим, что Q_U — верхнее критическое значение F-статистики с двумя степенями свободы в числителе и 32 степенями свободы в знаменателе — приближенно равно 3,84. Используя формулу (2), получаем:

$$\text{Критический размах} = Q_U \sqrt{\frac{\text{MSE}}{\text{rn}'}} = 3{,}84 \sqrt{\frac{8{,}61}{2*5}} = 3{,}56$$

	Верхние 5% значений (α=0,05)										
	Числитель, df_1										
Знаменатель, df_2	2	3	4	5	6	7	8	9	10	11	12
15	3,01	3,67	4,08	4,37	4,60	4,78	4,94	5,08	5,20	5,31	5,40
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35
17	2,98	3,63	4,02	4,30	4,52	4,71	4,86	4,99	5,11	5,21	5,31
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10
30	2,89	3,49	3,84	4,10	4,30	4,46	4,60	4,72	4,83	4,92	5,00
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,74	4,82	4,91
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81

Рис. 9. Критическое значение стьюдентизированного размаха Q_U; к сожалению, в Excel нет функции, рассчитывающей такой размах

Дисперсионный анализ

Дисперсионный анализ — метод, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. Дисперсионный анализ позволяет сравнивать средние значения двух и более групп.

Основную задачу дисперсионного анализа можно сформулировать следующим образом: оказывает ли значимое влияние на значение некоторой количественной переменной интересующий нас признак, измеренный на номинальном или порядковом уровне?

В терминах метода дисперсионного анализа та переменная, которая, как мы считаем, должна оказывать влияние на конечный результат, называется фактором. Например, если мы хотим объяснить различия в средних доходах респондентов тем, что респонденты проживают в различных населенных пунктах, то

переменная «место проживания респондента» - будет выступать фактором. Конкретное значение фактора (например, определенный населенный пункт) называют уровнем фактора. Значение измеряемого признака (в нашем примере — величину среднего дохода) называют откликом.

Если исследуется зависимость отклика только от одного фактора, то такой дисперсионный анализ называется однофакторным, если исследуется зависимость от двух и более факторов, то такой дисперсионный анализ называется многофакторным.

Само название - дисперсионный анализ (analysis of variance – сокращенно ANOVA) происходит из того, что метод проверки статистической гипотезы о равенстве средних значений в нескольких непересекающихся группах, основан на сопоставлении двух оценок дисперсии анализируемой количественной переменной.

Однофакторный дисперсионный анализ

В **однофакторной модели дисперсионного анализа** исходят из следующей модели порождения данных:

$$x_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad i = \overline{1, n_j}, \quad j = \overline{1, k},$$

где: x_{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, соответствующий j-му уровню фактора;

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

Выражение $x_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ можно представить в виде

$$x_{ij} = \mu + (\mu_j - \mu) + (x_{ij} - \mu_j),$$

или:

$$x_{ij} - \mu = (\mu_j - \mu) + (x_{ij} - \mu_j).$$

Данное соотношение говорит о том, что отклонение наблюдаемого значения отклика для j-ой группы складывается из суммы двух слагаемых: отклонения отклика от среднего значения j-ой группы: $(x_{ij} - \mu_j)$, и отклонения среднего значения j-ой группы от среднего значения всей совокупности: $(\mu_j - \mu)$. Что, по сути, означает, что дисперсия отклика может быть представлена в виде суммы двух дисперсий, одна из которых характеризует внутригрупповую изменчивость, а вторая межгрупповую.

Разложение общей дисперсии на составляющие для выборочных данных обычно записывается в виде равенства сумм квадратов соответствующих отклонений:

$$SS_T = SS_B + SS_R,$$

где:

$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{\mu})^2$ – общая, или полная, сумма квадратов отклонений;

$SS_B = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{\mu}_j - \overline{\mu})^2 = \sum_{j=1}^{k} n_j(\overline{\mu}_j - \overline{\mu})^2$ – сумма квадратов отклонений групповых средних от общего среднего, или межгрупповая (межуровневая факторная) сумма квадратов отклонений, также называемая суммой квадратов эффекта фактора или просто эффектом фактора;

$SS_R = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{\mu}_j)^2$ – сумма квадратов отклонений наблюдений от групповых средних, или внутригрупповая (остаточная) сумма квадратов отклонений, также называемая остаточным эффектом или эффектом ошибок;

k – число уровней фактора,

n_j – число наблюдений для j-го уровня фактора,

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

В разложении дисперсии на составляющие заключена основная идея дисперсионного анализа: общая вариация переменной, порожденная влиянием фактора и измеренная суммой SS_T, складывается из двух компонент: SS_B и SS_R, характеризующих изменчивость этой переменной между уровнями фактора (SS_B) и внутри уровней фактора (SS_R).

В дисперсионном анализе анализируются не сами суммы квадратов отклонений, а так называемые средние квадраты, которые получаются делением сумм квадратов отклонений на соответствующее число степеней свободы. Число степеней свободы для суммы квадратов случайных величин определяется как общее число линейно независимых слагаемых.

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

Для суммы квадратов эффекта фактора $SS_B = \sum_{j=1}^{k} n_j(\bar{\mu}_j - \bar{\mu})^2$ 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 $SS_R = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu}_j)^2$ 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 несмещенными оценками соответствующих дисперсий, имеют вид:

$$MS_T = \frac{1}{n-1}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu})^2,$$

$$MS_B = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{\mu}_j - \bar{\mu})^2,$$

$$SS_R = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu}_j)^2.$$

В случае нормального распределения остатков ε_{ij}, при условии истинности $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ (что равносильно: $\mu_1 = \mu_2 = \ldots = \mu_k$), статистика

$$F = \frac{MS_B}{MS_R} = \frac{n-k}{k-1}\frac{SS_B}{SS_R}$$

имеет распределение Фишера с $\nu_1 = k-1$ и $\nu_2 = n-k$ числом степеней свободы.

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 считается, что средние для различных уровней фактора значимо различаются.

Условия применимости данной модели дисперсионного анализа:

1) нормальность распределения данных для каждого уровня фактора;

2) однородность (равенство) дисперсий для различных уровней фактора.

Рассмотренная модель дисперсионного анализа предполагает, что данные измерены в количественной шкале.

Для порядковых данных непараметрической альтернативой однофакторного дисперсионного анализа являются ранговый дисперсионный анализ Краскела–Уоллиса и медианный тест.

В основе метода **дисперсионного анализа Краскела — Уоллиса** лежит однофакторный дисперсионный анализ, в котором вместо значений переменных используется ранг переменных.

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

При условии истинности гипотезы H_0 равенства средних рангов групп, статистика

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

будет иметь приближенно распределение Хи-квадрат с $k-1$ степенью свободы.

Если наблюдаемое значение статистики $H_{набл} \geq H_{кр}$, где $H_{кр}$ - критическая точка распределения 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 фактора значимо различаются.

Многофакторный дисперсионный анализ

Если анализируется одновременное влияние двух и более различных факторов на результаты наблюдений, то используется **многофакторный дисперсионный анализ**. Например, двухфакторная модель нам потребуется, если мы будем строить модель объяснения различий в средних доходах респондентов не только с учетом места проживания респондента, но и с учетом пола респондента.

Пусть мы исследуем влияние на величину X двух факторов A и B, имеющих, соответственно k и m уровней. В двухфакторной модели дисперсионного анализа обычно исходят из следующей модели порождения данных:

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

где: x_{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.

Заметим, что эффекты α_i, β_j, γ_{ij} удовлетворяют условиям: $\sum_{i=1}^{k} \alpha_i = 0$, $\sum_{j=1}^{m} \beta_i = 0$, $\sum_{i=1}^{k} \gamma_{ij} = 0$, $\sum_{j=1}^{m} \gamma_{ij} = 0$.

Выражение $x_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}$ можно представить в виде:

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

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 A и B ($x_{ijl} - \mu_{ij}$), главных эффектов i-го уровня фактора A и j-го уровня фактора B и эффекта взаимодействия. Что, означает, с учетом указанных выше условий на эффекты, что дисперсия отклика может быть представлена в виде суммы четырех дисперсий,

одна из которых характеризует внутригрупповую изменчивость для i, j-го набора уровней факторов A и B, а остальные соответствующие эффекты.

Разложение общей дисперсии на составляющие для выборочных данных обычно записывается в виде равенства сумм квадратов соответствующих отклонений (которое, вообще говоря, справедливо только в случае выполнения условия пропорциональности $n_{ij} = n_{i*}n_{*j}/n$):

$$SS_T = SS_A + SS_B + SS_{AB} + SS_R,$$

где:

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

$SS_A = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(\bar{\mu}_{i*} - \bar{\mu})^2 = \sum_{i=1}^{k} n_{i*}(\bar{\mu}_{i*} - \bar{\mu})^2$ – сумма квадратов отклонений средних по уровням фактора A от общей средней, или сумма квадратов главных эффектов A;

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

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

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

Число степеней свободы сумм квадратов SS_A и SS_B равно соответственно $\nu_A = k-1$ и $\nu_B = m-1$.

Число степеней свободы сумм квадратов взаимодействия эффектов SS_{AB} равно $\nu_{AB} = km - (\kappa - 1) - (m-1) - 1 = (\kappa - 1)(m-1)$.

Число степеней свободы сумм квадратов остатков SS_R равно $\nu_R = n - km$

Соответственно средние суммы квадратов будут равны:

$$MS_A = \frac{SS_A}{k-1},\ MS_B = \frac{SS_B}{m-1},\ MS_A = \frac{SS_{AB}}{(k-1)(m-1)},\ MS_R = \frac{SS_R}{n-km}.$$

Поскольку двухфакторная модель учитывает различные эффекты влияния факторов, то и статистический анализ для двухфакторной модели предполагает проверку гипотез о значимости различных эффектов. В качестве статистик критериев проверки гипотез о значимости соответствующих эффектов используются отношения средней суммы квадратов эффектов к средней сумме квадратов остатков. При условии истинности H_0:«эффект незначим» и нормальном распределении остатков данные статистики имеют распределение Фишера с параметрами степеней свободы, определяемыми числами степеней свободы соответствующих сумм, участвующих в отношении. В табл. 1 приведены основные рассматриваемые гипотезы, статистики критериев для проверки данных гипотез и соответствующие числа степеней свободы данных статистик.

Табл. 1. Статистики для проверки гипотез двухфакторного дисперсионного анализа

Основная гипотеза:	Все $\alpha_i = 0$	Все $\beta_j = 0$	Все $\gamma_{ij} = 0$
Статистика критерия	MS_A / MS_R	MS_B / MS_R	MS_{AB} / MS_R
Числа степеней свободы	$\nu_1 = k-1$ $\nu_2 = n-nk$	$\nu_1 = m-1$ $\nu_2 = n-nk$	$\nu_1 = (k-1)(m-1)$ $\nu_2 = n-nk$

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.

Апостериорные множественные сравнения средних

Результат дисперсионного анализа, указывающий, что средние значения отклика для разных уровней фактора, различаются, не является окончательным результатом анализа изучаемого явления. Это скорее промежуточный результат, который подразумевает дальнейшее раскрытие 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.

Самый очевидный и простой вариант решения данной задачи - провести серию по парных сравнений при помощи 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:

$$t = \frac{\overline{\mu}_i - \overline{\mu}_j}{\sqrt{MS_R(1/n_i + 1/n_j)}}.$$

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 $H_1 : \mu_1 \neq \mu_2$.

Однако, такой подход является не совсем корректным. Если задать, скажем, 5% уровень значимости, то при каждом сравнении вероятность отклонить нулевую гипотезу будет равна 5%, а при серии по парных сравнений вероятность отклонить хотя бы одну нулевую гипотезу в таком случае существенно превысит 5%. Например, при по парном сравнении средних 4 групп, эта вероятность составит 26,5 %.

Существуют разные подходы к решению данной проблемы. Один из них – уменьшить уровень значимости при по парном сравнении так, чтобы вероятность хотя бы одного отклонения нулевой гипотезы равнялось заданному уровню значимости. Такой подход реализуется в **методе Бонферрони** (правильнее говорить о принципе Бонферрони) множественных сравнений, в котором при каждом по парном 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 как бы самые крайние 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:

$$F = \frac{\left(\overline{\mu}_i - \overline{\mu}_j\right)^2}{(k-1)MS_R(1/n_i + 1/n_j)},$$

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 $H_1 : \mu_i \neq \mu_j$.

Заметим, что в отличии от LSD критерия, где статистика $\left(\overline{\mu}_i - \overline{\mu}_j\right)^2 / MS_R(1/n_i + 1/n_j)$ имеет одну степень свободы, в критерии Шеффе предполагается, что статистика имеет $k-1$ степень свободы. 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:

$$t_R = \frac{\left|\overline{\mu}_i - \overline{\mu}_j\right|}{\sqrt{MS_R(1/n_i + 1/n_j)/2}},$$

значения которой 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:

$$t_R = \frac{\left|\overline{\mu}_i - \overline{\mu}_j\right|}{\sqrt{MS_R / \min(n_i, n_j)}}.$$

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

$$t_R = \frac{\left|\overline{\mu}_i - \overline{\mu}_j\right|}{\sqrt{MS_R \frac{1}{k}\sum_{l=1}^{k}\frac{1}{n_l}}}.$$

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 $c_1 = 1, c_2 = -1$, $c_3 = c_4 = \ldots = c_k = 0$, 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:

$$t = \frac{\sum_{i=1}^{k} c_i \bar{\mu}_i}{\sqrt{MS_R \sum_{i=1}^{k} (c_i^2 / n_i)}}.$$

Критическими точками статистики, по-прежнему, будут являться квантили распределения Стьюдента уровня $1 - \alpha/2$ с числом степеней свободы $\nu = n - k$.

Оценивание классификационных методов

Оценивание методов следует проводить, исходя из следующих характеристик [21]: скорость, робастность, интерпретируемость, надежность.

Скорость характеризует время, которое требуется на создание модели и ее использование.

Робастность, т.е. устойчивость к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

Свойства классификационных правил:

- размер дерева решений;
- компактность классификационных правил.

Надежность методов классификации предусматривает возможность работы этих методов при наличии в наборе данных шумов и выбросов.

Задача кластеризации

Только что мы изучили задачу классификации, относящуюся к стратегии "обучение с учителем".

В этой части лекции мы введем понятия кластеризации, кластера, кратко рассмотрим классы методов, с помощью которых решается задача кластеризации, некоторые моменты процесса кластеризации, а также разберем примеры применения кластерного анализа.

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены.

Синонимами термина " кластеризация " являются "автоматическая классификация ", "обучение без учителя" и "таксономия".

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

Цель кластеризации - поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Само понятие " кластер " определено неоднозначно: в каждом исследовании свои " кластеры ". Переводится понятие кластер (cluster) как "скопление", "гроздь".

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Вопрос, задаваемый аналитиками при решении многих задач, состоит в том, как организовать данные в наглядные структуры, т.е. развернуть таксономии.

Наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. Для решения экономических задач кластеризация длительное время мало использовалась из-за специфики экономических данных и явлений.

В таблице 5.2 приведено сравнение некоторых параметров задач классификации и кластеризации.

Таблица 5.2. Сравнение классификации и кластеризации

Характеристика	Классификация	Кластеризация
Контролируемость обучения	Контролируемое обучение	Неконтролируемое обучение
Стратегия	Обучение с учителем	Обучение без учителя
Наличие метки класса	Обучающее множество сопровождается меткой, указывающей	Метки класса обучающего множества неизвестны

	класс, к которому относится наблюдение	
Основание для классификации	Новые данные классифицируются на основании обучающего множества	Дано множество данных с целью установления существования классов или кластеров данных

На рис. 5.7 схематически представлены задачи классификации и кластеризации.

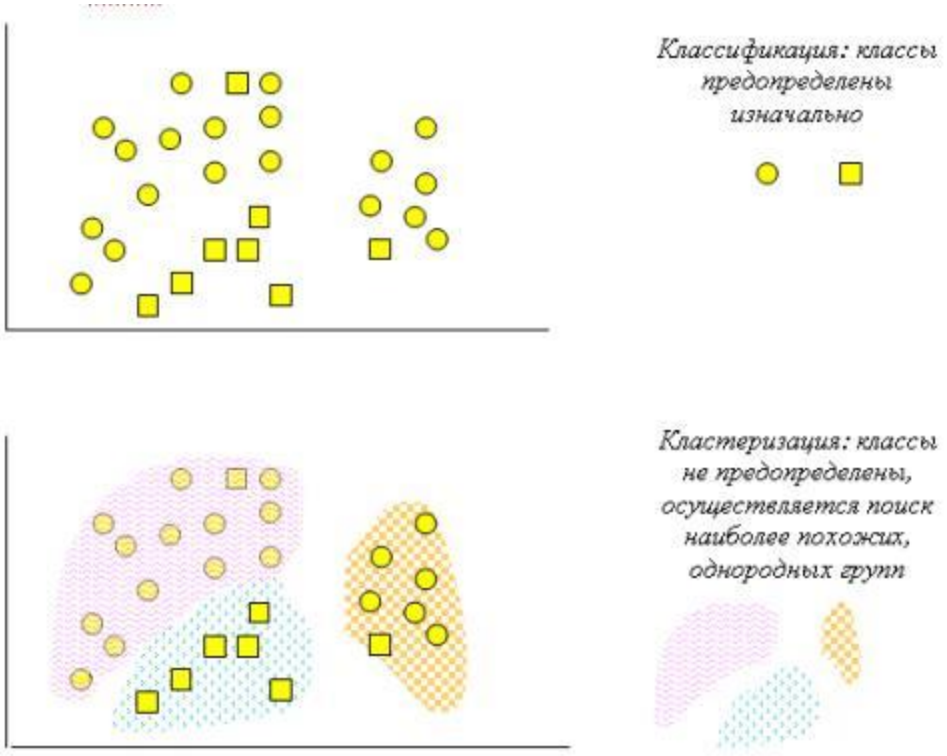


Рис. 5.7. Сравнение задач классификации и кластеризации

Кластеры могут быть непересекающимися, или эксклюзивными (non-overlapping, exclusive), и пересекающимися (overlapping). Схематическое изображение непересекающихся и пересекающихся кластеров дано на рис. 5.8.

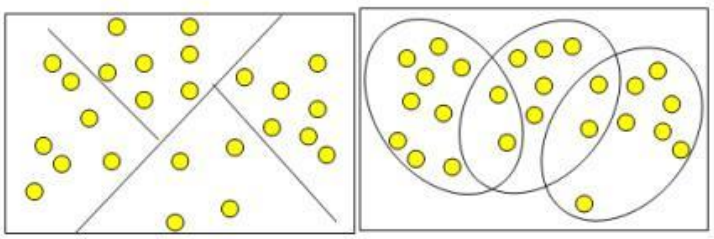


Рис. 5.8. Непересекающиеся и пересекающиеся кластеры

Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры "цепочного" типа, когда кластеры представлены длинными "цепочками", кластеры удлиненной формы и т.д., а некоторые методы могут создавать кластеры произвольной формы.

Различные методы могут стремиться создавать кластеры определенных размеров (например, малых или крупных) либо предполагать в наборе данных наличие кластеров различного размера.

Некоторые методы кластерного анализа особенно чувствительны к шумам или выбросам, другие - менее.

В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма.

Данные особенности следует учитывать при выборе метода кластеризации.

Подробнее обо всех свойствах кластерного анализа будет рассказано в лекции, посвященной его методам.

На сегодняшний день разработано более сотни различных алгоритмов кластеризации. Некоторые, наиболее часто используемые, будут подробно описаны во втором разделе курса лекций.

Приведем краткую характеристику подходов к кластеризации.

- Алгоритмы, основанные на разделении данных (Partitioning algorithms), в т.ч. итеративные:
    - разделение объектов на k кластеров;
    - итеративное перераспределение объектов для улучшения кластеризации.
- Иерархические алгоритмы (Hierarchy algorithms):
    - агломерация: каждый объект первоначально является кластером, кластеры, соединяясь друг с другом, формируют больший кластер и т.д.

- Методы, основанные на концентрации объектов (Density-based methods):
    - основаны на возможности соединения объектов;
    - игнорируют шумы, нахождение кластеров произвольной формы.
- Грид-методы (Grid-based methods):
    - квантование объектов в грид-структуры.
- Модельные методы (Model-based):
    - использование модели для нахождения кластеров, наиболее соответствующих данным.

Приложение

Корреляционный анализ

Корреляционная связь – это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменением другого признака.

Коэффициент корреляции – это показатель степени связи между двумя переменными или измерениями. Коэффициент корреляции измеряется от -1 до+1.

Величина коэффициента корреляции по модулю показывает степень зависимости. Корреляционны связи разливаются по величине следующим образом:

r=0 нет связи;

r=0,01 – 0,3 – слабая связь;

r=0,31 – 0,7 – умеренная связь;

r=0,71 – 0,99 – сильная связь;

r=1 – совершенная связь.

Коэффициент корреляции Пирсона.

H_0: корреляция между переменными А и Б не отличается от нуля.

H_1: корреляция между переменными А и Б достоверно отличается от нуля.

$$rx = \frac{n \times \sum (x_i \times y_i) - (\sum x_i \times \sum y_i)}{\sqrt{\left[n \times \sum x_i^2 - (\sum x_i)^2\right] \times \left[n \times \sum y_i^2 - (\sum y_i)^2\right]}}$$

X - значение одной переменной;

Y - значение другой переменной;

n - число пар данных взятых для анализа.

Задача.

Продавец мороженного интересуется, есть ли связь между температурой воздуха и количеством пачек мороженного, купленных у него в ларьке.

День недели	**Температура воздуха, X**	**Количество**	**X^2**	**Y^2**	**XY**

		куплен-ных пачек мороже-ного, Y			
Пн	7	1			
Вт	4	3			
Ср	13	5			
Чт	16	7			
Пт	10	9			
Сб	22	11			
Вс	19	13			
N=7	$\sum$	$\sum$	$\sum$	$\sum$	$\sum$

Вычислив значения для каждого из столбцов, просуммируем им (для каждого столбца отдельно) и подставим получавшиеся результаты в формулу.

Для определения гипотезы (H_0 или H_1) $r_{эмп}$ (значение получаемое после расчетов) сравнивается с $r_{кр}$ (табличное значение, см. таблицу критических значение коэффициента корреляции Пирсона).

Для определения значения k

k= n(общее количество испытуемых, в данном примере n=7) – 2.

Таблица критических значение коэффициента корреляции Пирсона

$k = n-2$	P		$k = n-2$	P	
	0,05	0,01		0,05	0,01
5	0,75	0,87	27	0,37	0,47
6	0,71	0,83	28	0,36	0,46
7	0,67	0,80	29	0,36	0,46
8	0,63	0,77	30	0,35	0,45
9	0,60	0,74	35	0,33	0,42
10	0,58	0,71	40	0,30	0,39
11	0,55	0,68	45	0,29	0,37
12	0,53	0,66	50	0,27	0,35
13	0,51	0,64	60	0,25	0,33
14	0,50	0,62	70	0,23	0,30
15	0,48	0,61	80	0,22	0,28
16	0,47	0,59	90	0,21	0,27
17	0,46	0,58	100	0,20	0,25
18	0,44	0,56	125	0,17	0,23
19	0,43	0,55	150	0,16	0,21
20	0,42	0,54	200	0,14	0,18
21	0,41	0,53	300	0,11	0,15
22	0,40	0,52	400	0,10	0,13
23	0,40	0,51	500	0,09	0,12
24	0,39	0,50	700	0,07	0,10
25	0,38	0,49	900	0,06	0,09
26	0,37	0,48	1000	0,06	0,09

Решите задачу, укажите, какая гипотеза принимается, сформулируйте ее, нарисуйте ось значимости, сделайте содержательный вывод.

Коэффициент корреляции Спирмена

Гипотезы:

H_0: корреляция между переменными А и Б не отличается от нуля.

H_1: корреляция между переменными А и Б достоверно отличается от нуля.

$$r_{rank} = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

D – разность рангов;

n – количество ранжируемых пар.

Задача.

Знания десяти студентов проверены по двум теста А и Б. Оценки по стобальной системе оказались следующими:

Правила ранжирования:

1. Выписываем значения в ряд по возрастанию (от меньшего к большему) (в нашем примере каждый столбец ранжируется отдельно);
2. Наименьшему значению присваиваем 1 ранг.

Если в числовом ряду попадаются **одинаковые значения**
10, 13, **14 ,14** ,16 ,20 ,22 , то:
10 (1. ранг),13 (2 ранг),**14 (3,5 ранг), 14 (3,5 ранг),**16 (5 ранг), 20 (6 ранг), 22 (7 ранг).
Числа 14 должны были бы получить ранги 3 и 4, но, поскольку они равны, то получают средний ранг:

$$\frac{3+4}{2} = 3{,}5$$

Первым ранжирует столбец 1 (Оценки по тесту А, X):

50 (1 ранг, так как это наименьшее значение), 57 (2 ранг), 60 (3 ранг), 62 (4 ранг), 70 (5 ранг), 75 (6 ранг), 84 (7 ранг), 86 (8 ранг), 90 (9 ранг), 95 (10 ранг).

Данные значения вписываем в столбец 3 (Ранг X).

Оценки по тесту А, X	**Оценки по тесту Б, Y**	**Ранг X**	**Ранг Y**	**Разность рангов, D (Ранг X - Ранг Y)**	**Квадрат разности рангов, D^2**
95	92	10			
90	93	9			
86	83	8			
84	80	7			
75	55	6			
70	60	5			
62	45	4			
60	72	3			
57	62	2			
50	70	1			

					$\sum$

Вычислив значения для каждого из столбцов, просуммируем значения последнего столбца (D^2) и подставим получавшиеся результаты в формулу.

Для определения гипотезы (H_0 или H_1) $r_{эмп}$ (значение получаемое после расчетов) сравнивается с $r_{кр}$ (табличное значение, см. таблицу критических значений Спирмена).

Критические значения коэффициента корреляции рангов Спирмена

n	ρ		n	ρ		n	ρ	
	0,05	0,01		0,05	0,01		0,05	0,01
5	0,94	-	17	0,48	0,62	29	0,37	0,48
6	0,85	-	18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,49	0,51	37	0,33	0,43
14	0,54	0,68	26	0,39	0,50	38	0.32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40

Решите задачу, укажите, какая гипотеза принимается, сформулируйте ее, нарисуйте ось значимости, сделайте содержательный вывод.

Построение модели линейной регрессии

1. Построить корреляционное поле. По характеру расположения точек в корреляционном поле выбрать вид регрессии.
2. Вычислить числовые характеристики $\bar{x}, \bar{y}, s_x, s_y, r$.
3. Определить значимость коэффициента корреляции ги найти для него доверительный интервал с надежностью γ=0,95.

4. Найти эмпирическое уравнение регрессий Y на X и X на Y. Проверить гипотезы о значимости коэффициентов регрессии и построить доверительные интервалы для них.

5. Вычислить коэффициент детерминации R^2 и объяснить его смысловое значение.

6. Проверить адекватность уравнения регрессии Y на X.

1. Исходные данные

Зависимость теплоемкости C_p фторида магния от температуры T выражается следующими данными:

T, K	300	400	500	600	700	800	900	1000
Ср, Дж/(моль • К)	70,35	75,38	80.53	85,81	91,26	96,83	102,53	108,27

Примечание. Для упрощения расчетов можно сделать замену $t = \frac{T - 300}{100}$. Мы этого не делали.

Нанесем исходные данные на график.

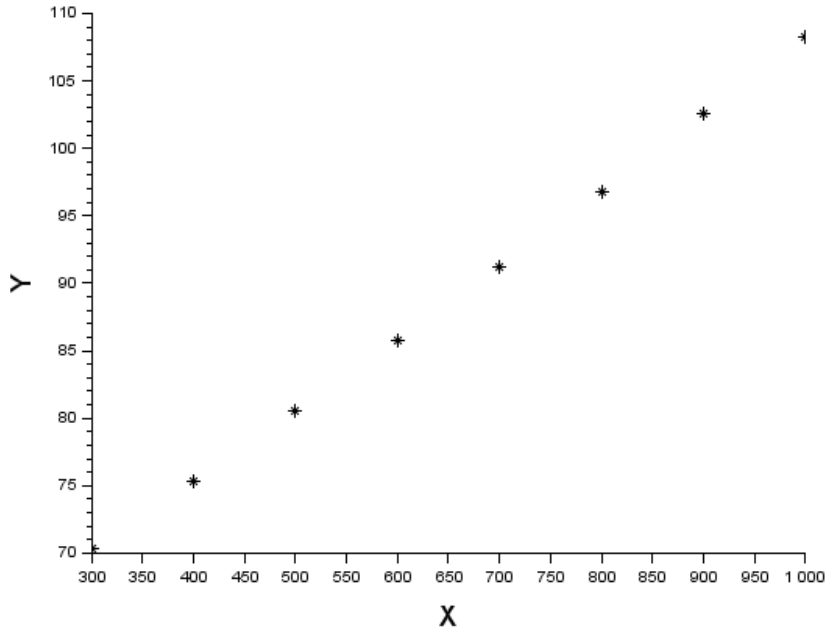


Рис. 3.5. Выявление вида зависимости

На этом графике ХэтоТ, а Y – этоС. Из графика видно, что можно провести прямую,которая пройдет рядом с точками. Это означает, что между X и Y существет линейная зависимость.

2. Определение требуемых характеристик

Составим расчетную таблицу.

В нейхэтоТ, а y – этоС. Десятая строка – соответствующие суммы.

В соответствии с формулами (3.18) и (3.19) имеем:

$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{5200}{8} = 650$ - средняя температура,

$\bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{710{,}96}{8} = 88{,}87$ средняя теплоемкость фторида магния.

	A	B	C	D	E	F
1	T (x)	C (y)	(x-x cp)^2	(y-y cp)^2	x^2	xy
2	300	70,35	122500	342,9904	90000	21105
3	400	75,38	62500	181,9801	160000	30152
4	500	80,53	22500	69,5556	250000	40265
5	600	85,81	2500	9,3636	360000	51486
6	700	91,26	2500	5,7121	490000	63882
7	800	96,83	22500	63,3616	640000	77464
8	900	102,53	62500	186,5956	810000	92277
9	1000	108,27	122500	376,36	1000000	108270
10	**5200**	**710,96**	**420000**	**1235,919**	**3800000**	**484901**
11						
12	x cp	**650**				
13	y cp	**88,87**				

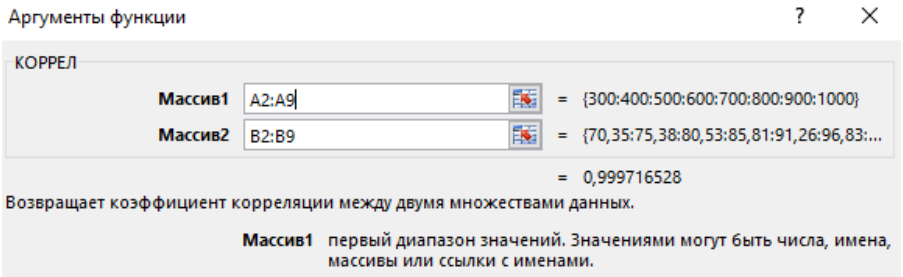
$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{8-1}\sum_{i=1}^{8}(x_i - \bar{x})^2} = \sqrt{\frac{1}{7}\cdot 420000} = 244{,}949$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{8-1}\sum_{i=1}^{8}(y_i - \bar{y})^2} = \sqrt{\frac{1}{7}\cdot 1235{,}919} = 13{,}28758$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} . = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$= \frac{484901 - 8 \cdot 650 \cdot 88{,}87}{\sqrt{420000 \cdot 1235{,}919}} = 0{,}999717$$

Тот же результат можно получить, используя стандартные статистические функции Excel: среднее значение x вычисляется как =СРЗНАЧ(A2:A9); среднее значение y=СРЗНАЧ(B2:B9); S_x =СТАНДОТКЛОН.В(A2:A9); S_y =СТАНДОТКЛОН.В(B2:B9); и r =КОРРЕЛ(A2:A9;B2:B9):



На Scilabрасчет необходимых характеристик можно произвести так:

```
clc
x=[300 400 500 600 700 800 900 1000];
y=[70.35 75.38 80.53 85.81 91.26 96.83 102.53 108.27];
n=length(x);
xcp=mean(x)
ycp=mean(y)
r=correl(x,y);
sx=stdev(x)
sy=stdev(y)
printf("среднее значение x равно %4.5f", xcp)
disp("")
printf("среднее значение y равно %4.5f", ycp)
disp("")
printf("sx=%3.5f",sx)
disp("")
printf("sy=%3.5f",sy)
disp("")
printf("r=%3.5f",r)
```

В результате получим:

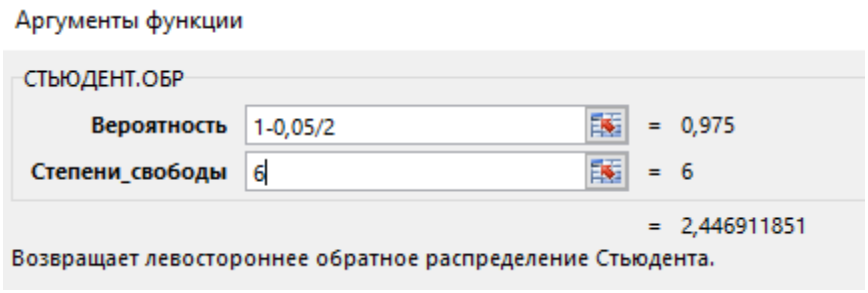
```
среднее значение x равно 650.00000
среднее значение y равно 88.87000
sx=244.94897
sy=13.28758
r=0.99972
```

3. Значимость коэффициента корреляции и доверительный интервал для него

Вычислим статистику

$$t = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2} = \frac{0{,}999717}{\sqrt{1-0{,}999717^2}}\sqrt{8-2} = 102{,}8521$$

По таблицам критических точек распределения Стьюдента или с помощью статистической функции ExcelСТЬЮДЕНТ.ОБР, входом в которую является вероятность 1-α/2 (α=1-γ) и n-2=6 степеней свободы находим критическое значение $t_{\alpha, n-2}$=2,447.



Так как t=102,8521>$t_{\alpha, n-2}$=2,447, делаем вывод, что выборочный коэффициент корреляции значимо отличается от нуля. Следовательно, можно предположить, что теплоемкостьC_p фторида магния и температураTсвязаны линейной корреляционной зависимостью.

Для проверки значимости коэффициента корреляции при уровне значимости α=0,05 составлена программа

```
clc
x=[300 400 500 600 700 800 900 1000];
y=[70.35 75.38 80.53 85.81 91.26 96.83 102.53 108.27];
n=length(x);
r=correl(x,y);
alfa=0.05;
t=cdft("T", n-2,1- alfa/2,alfa/2);
printf("r=%1.5f",r) ;
disp("")
 printf("Коэффициент r   ")
if abs(r)/sqrt(1-r^2)*sqrt(n-2)>t then
   printf("значимо")
```

```
else
   printf("незначимо")
end
printf("   отличается от нуля")
disp("")
```

В результате получим:

r=0.99972

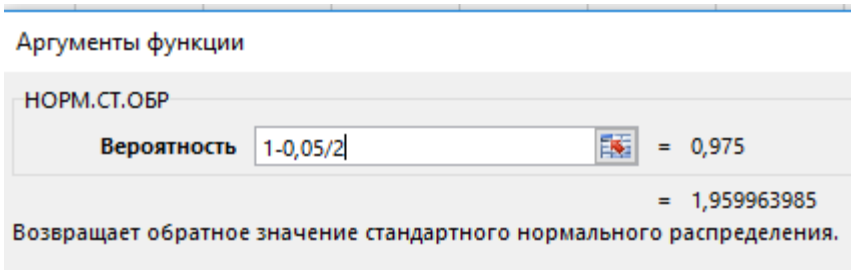
Коэффициент r значимо отличается от нуля

Построим сначала доверительный интервал для гиперболического арктангенса коэффициента корреляции ρ.

Доверительный интервал для Arth(ρ) имеет вид:

$$Arth(r)-\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}<Arth(\rho)<Arth(r)+\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}$$

Здесь α=1-γ. В нашем примере α=0,05. $u_{1-\frac{\alpha}{2}}$ - (1-α/2)*100% квантиль нормального распределения, которую можно найти по таблицам нормального распределения или с помощью статистической функции ExcelНОРМ.СТ.ОБР:



atanh(r)=atanh(0,999717)=4,430701

Тогда $4{,}430701-\frac{1{,}95996}{\sqrt{5}}<arth(\rho)<4{,}430701+\frac{1{,}95996}{\sqrt{5}}$

3,554179<arth(ρ)<5,307224; tanh(3,554179)=0,998365: tanh(5,307224)=0,999951

Окончательно получаем:

0,998365<ρ<0,999951

Для построения доверительного интервала в Scilabвоспользуемся программой

```
clc
x=[300 400 500 600 700 800 900 1000];
y=[70.35 75.38 80.53 85.81 91.26 96.83 102.53 108.27];
n=length(x);
r=correl(x,y);
disp("доверительный  интервал для Arth(ro) ")
Gamma=0.95;//доверительная вероятность
P=(1+Gamma)/2,Q=1-P;u=cdfnor("X",0,1,P,Q)
d=u/sqrt(n-3);t1=atanh(r)-d ; t2=atanh(r)+d;
printf("      %3.5f",t1) ;printf("<Arth(ro)<%3.5f", t2);
disp("")
disp("доверительный  интервал для ro ")
printf("      %3.5f",tanh(t1)) ;printf("<ro<%3.5f",tanh(t2)) ;
disp("")
```

```
доверительный интервал для Arth(ro)
    3.55418<Arth(ro)<5.30722
доверительный интервал для ro
    0.99836<ro<0.99995
```

4. Эмпирическое уравнение линейных регрессий Y на X и X на Y. Проверка гипотез о значимости коэффициентов регрессии и построение доверительных интервалов для них

Уравнение линейной регрессии Y на X имеет вид: y=a+bx. Чтобы определить неизвестные коэффициенты a и b, необходимо составить и решить систему нормальных уравнений. В нашем случае неизвестных коэффициентов - два, поэтому надо получить систему из двух линейных уравнений:

$$\begin{cases} \sum_{i=1}^{8} y_i = 8a + b\sum_{i=1}^{8} x_i \\ \sum_{i=1}^{8} x_i y_i = a\sum_{i=1}^{8} x_i + b\sum_{i=1}^{8} x_i^2 \end{cases}$$

В этой системе уравнений хэтоТ, а у – этоС .

Получилась система уравнений

$$\begin{cases} 710{,}96 = 8a + 5200b \\ 484901 = 5200a + 3800000b \end{cases}$$

В Excel ее можно решить с помощью надстройки Поиск решения. Получим

a	b
53,61988	0,054231
Левая	Правая
710,96	710,96
484901	484901

В исходных переменных искомое уравнение $C_p = 53{,}61988 + 0{,}054231T$

Решим эту же задачу методами корреляционного анализа. Уравнение при этом перепишем в виде

$$y = a + bx = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$$

Подставляя сюда ранее найденные значения, убеждаемся в том, что мы получили одно и то же уравнение

$y = 53{,}61988 + 0{,}054231x$ или $C_p = 53{,}61988 + 0{,}054231T$

Коэффициент b можно вычислить и с помощью функции НАКЛОН(диапазон Y; диапазон X).

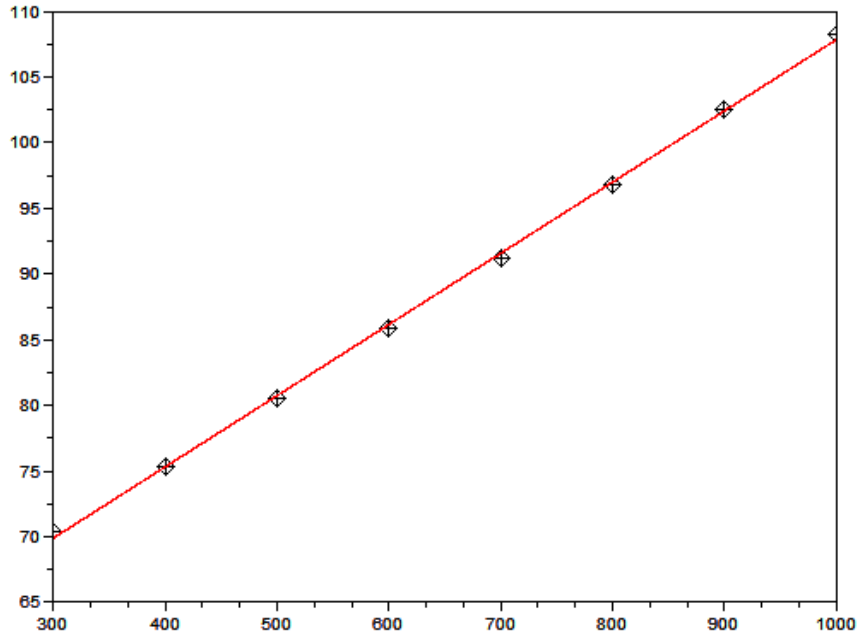
Программа в среде Scilab может выглядеть так:

```
clc
clf()
x=[300 400 500 600 700 800 900 1000];
y=[70.35 75.38 80.53 85.81 91.26 96.83 102.53 108.27];
```

```
[b,a]=reglin(x,y);
plot2d(x,y,-8)//знак - означает, что точки не соединены прямыми,
//цифра задает  вид маркера для обозначения исходных данных
t=min(x):.1:max(x);
q=a+b*t;
plot2d(t,q,5)
printf("ЛинейнаярегрессияYнаXy=%3.5f",a) ;
printf("+(%3.5f",b) ;
printf(")*x") ; disp("");
```

Результат работы программы:

ЛинейнаярегрессияYнаXy=53.61988+(0.05423)*x



Для проверки гипотез о значимости коэффициентов регрессиии построения 95% доверительных интервалов можно воспользоваться приведеннойниже программой.

```
clc
x=[300 400 500 600 700 800 900 1000];
y=[70.35 75.38 80.53 85.81 91.26 96.83 102.53 108.27];
n=length(x);
```

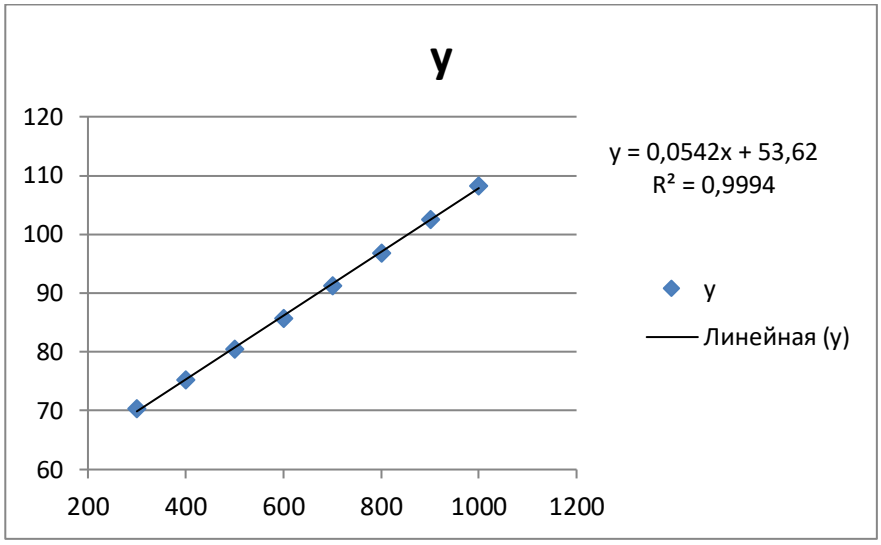
```
[b,a]=reglin(x,y);
printf("Линейная регрессияYнаXy=%3.5f",a) ;
printf("+(%3.5f",b) ;
printf(")*x") ; disp("");
S=(1/(n-2)*sum((y-a-b*x).^2))^0.5;
xcp=mean(x);
sx=stdev(x);
Sb=S/(sx*(n-1)^0.5)
Sa=S*sqrt(1/n+xcp^2/((n-1)*sx^2));
alfa=0.95;
t=cdft("T", n-2, (1+alfa)/2,(1-alfa)/2);
printf("b=%3.5f",b) ;
disp("")
 printf("коэффициент b является  ")
if abs(b)>t*Sb then
printf("значимым")
else
printf("незначимым")
end
disp("")
printf("a=%3.5f",a) ;
disp("")
printf("коэффициентаявляется  ")
if abs(a)>t*Sa then
printf("значимым")
else
   printf("не значимым")
end
disp("")
```

```
disp("доверительный  интервал для a ")
printf("      %3.5f",a-t*Sa) ;printf("<b<%3.5f",a+t*Sa) ;
disp("")
disp("доверительный  интервал для b ")
printf("      %3.5f",b-t*Sb) ;printf("<b<%3.5f",b+t*Sb) ;
```

Результат работы программы:

```
Линейная регрессияYнаXy=53.61988+(0.05423)*x
b=0.05423
 коэффициент b является  значимым
a=53.61988
 коэффициент a является  значимым
    доверительный  интервал для a
      52.73068<b<54.50908
  доверительный  интервал для b
      0.05294<b<0.05552
```

Если же нам хочется просто побыстрей получить уравнение регрессии, не вдаваясь в математические расчеты, самый простой способ – воспользоваться возможностями Excel и добавить искомую линию тренда к исходным данным, отредактировав при необходимости полученный график:



Здесь R^2 – выборочный коэффициент детерминации.

В заключение приведем еще один пример «быстрых» расчетов, с помощью статистической функции Excel ЛИНЕЙН. В силу ее некоторой специфики опишем работу с ней немного подробнее.

Задаем диапазон значений Y и диапазон значений X. Два остальные поля оставляем пустыми. После нажатия кнопки ОК в ячейке с формулой видим только значение коэффициента **b.** Чтобы увидеть значение **a**, нужно проделать некоторые дополнительные действия. Выделяем две ячейки: ячейку с формулой и соседнюю справа. Нажимаем клавишу F2, затем комбинацию клавиш Ctrl, Shift и Enter. Во второй выделенной ячейке появляется значение **b.**

b	**a**
0,054231	53,61988

Для построения линейной регрессии X на Yвнесем в программу небольшие изменения:

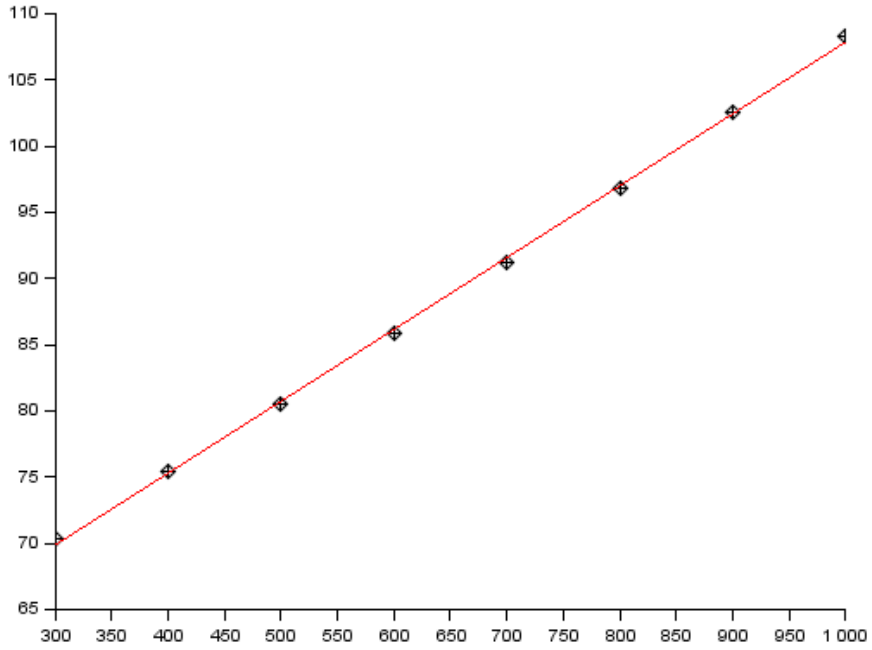
```
clc
clf()
x=[300 400 500 600 700 800 900 1000];
y=[70.35 75.38 80.53 85.81 91.26 96.83 102.53 108.27];
[d,c]=reglin(y,x);
plot2d(x,y,-8)//знак - означает, что точки не соединены прямыми,
        //цифра задает вид маркера для обозначения исходных данных
t=min(x):.1:max(x);
q=(t-c)/d;
plot2d(t,q,5)
printf("Линейная регрессия X на  Y  x=%3.5f",c) ;
```

printf("+(%3.5f",d) ;

printf(")*y") ; disp("");

В результате получим:

Линейная регрессия X на Y x=-987.80312+(18.42920)*y



5. Вычисление коэффициента детерминации R^2

Убедимся в том, что значение коэффициента детерминации на графике(R^2=0,9994) совпадает с вычислениями по формулам (3.6) и (3.7). Необходимые расчеты проведены в Excel.

x	y	Y=53,61988+0,054231x	$(y-Y)^2$	$(y-y\ ср)^2$
300	70,35	69,88918	0,21236	342,9904
400	75,38	75,31228	0,00459	181,9801
500	80,53	80,73538	0,04218	69,5556
600	85,81	86,15848	0,12144	9,3636
700	91,26	91,58158	0,10341	5,7121
800	96,83	97,00468	0,03051	63,3616

900	102,53	102,42778	0,01045	186,5956
1000	108,27	107,85088	0,17566	376,36
		Сумма	**0,7006**	**1235,919**
y ср	**88,87**	**R^2**	**0,99943**	

$\bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{(70{,}53 + 75{,}38 + 80{,}53 + \ldots + 108{,}27)}{8} = 88{,}845 = y\,cp$ вычислено с помощью функции СРЗНАЧ.

$$R^2 = 1 - \frac{\sum_{i=1}^{8}(y_i - f(x_i))^2}{\sum_{i=1}^{8}(y_i - \bar{y})^2} = 1 - \frac{0{,}07006}{1235{,}919} = 0{,}99943$$

Или

$$R^2 = \frac{\sum_{i=1}^{8}(\bar{y} - f(x_i))^2}{\sum_{i=1}^{8}(y_i - \bar{y})^2} = \frac{1235{,}221}{1235{,}919} = 0{,}99943$$

Такое большое значение коэффициента детерминации говорит о том, практически весь разброс значений величины y объясняется линейной корреляционной зависимостью междутеплоемкостьюC_p фторида магния и температуройT.

6. Проверка гипотезы об адекватности уравнения регрессии Y на X

Проверим гипотезу об адекватности полученной сглаживающей прямой исходным данным по критерию Фишера при уровне значимости α=0,05.

Для этого вычислим статистику

$$F_{выб} = \frac{R^2(n-2)}{1-R^2} = \frac{0{,}99943 \cdot (8-2)}{1-0{,}99943} = 10578{,}55483$$

Здесь R^2 – коэффициент детерминации, n=8. По числу степеней свободы k_1=1 и k_2=n-2=6 найдем критическое значение $F_{кр}$с помощью статистической функции F.ОБР.ПХ (Microsoft Excel 2010, 2016).

$F_{кр}$=5,987378

Так как $F_{выб}>F_{кр}$, делаем вывод о том, что полученное уравнение линейной регрессии $C_p = 53{,}61988 + 0{,}054231T$ статистически значимо описывает результаты эксперимента.

Дисперсионный анализ социологических признаков в пакете STATISTICA.

Пример 1. Результаты ответов 400 респондентов на вопросы анкеты «Томск 400» «Есть ли у вас хронические заболевания: 1) сердечно-сосудистые; 2) бронхо-легочные; 3) желудочно-кишечного тракта; 4) эндокринологические; 5) опорно-двигательной системы; 6) невралгические (в том числе слух, зрение); 7) урологические (гинекологические)» с вариантами ответов: “Да”, “Нет” оформлены в виде 7 числовых выборок кодов ответов с названиями «ЗБ1» - «ЗБ7». Код ответа соответствует номеру ответа. Также имеется выборка «НП» числовых кодов, соответствующих месту проживания респондента (1 – «Томск», 2 - «Северск», 3 – «Томский район», 4 - «Асино», 5 – «Асиновский район», 6 - «Каргасокский район», 7 – «Каргасок», 8 - «Тегульдет»). Используя дисперсионный анализ, установить, одинаков ли уровень различных заболеваний в различных населенных пунктах.

Используя имеющиеся данные, можно сформулировать различные задачи дисперсионного анализа в рамках анализа уровня заболеваний в различных населенных пунктах. Можно проверить гипотезу о различии уровня заболеваний (по всем заболеваниям) по населенным пунктам – это будет в данном случае задача многомерного однофакторного дисперсионного анализа. Можно проверить гипотезы о различии уровней заболеваний по каждому заболеванию в отдельности по различным населенным пунктам. В этом случае мы получим совокупность задач, каждая из которых относится к одномерному однофакторному дисперсионному анализу.

Поскольку мы имеем дело с дихотомическими данными, анализ различий в данном случае равносилен проверке гипотез о различии частот заболеваний.

Чтобы воспользоваться параметрическим аппаратом статистики, необходимо чтобы коды ответов содержали значения “1” и “0”. В этом случае среднее арифметическое значение признака будет являться его относительной частотой, и задача сравнения частот сводится к задаче сравнения средних, для которой можно использовать параметрические методы. Поскольку в нашем случае коды ответов иные, необходимо перекодировать данные, либо вручную, либо так, как это сделано в примере 6. В результате ответам “Да”, “Нет” у нас буду соответствовать коды «1» и «0».

Рассмотрим самую простую реализацию однофакторного дисперсионного анализа в пакете статистика, используя соответствующий модуль в меню «Basic Statisics/Tables». Запускаем в головном меню модуль «Statistics», в стартовой панели выбираем пункт «Basic Statisics/Tables».

В меню модуля «Basic Statistics and Tables» (рис. 1) выбираем пункт «Breakdown & one-way ANOVA» («Классификация и одномерный дисперсионный анализ») и в появившемся окне модуля выбора зависимых и группирующих переменных (рис. 2) выбираем в качестве зависимых переменных (откликов) переменные «ЗБ1» - «ЗБ7», а в качестве группирующей переменной (фактора) - переменную «НП».

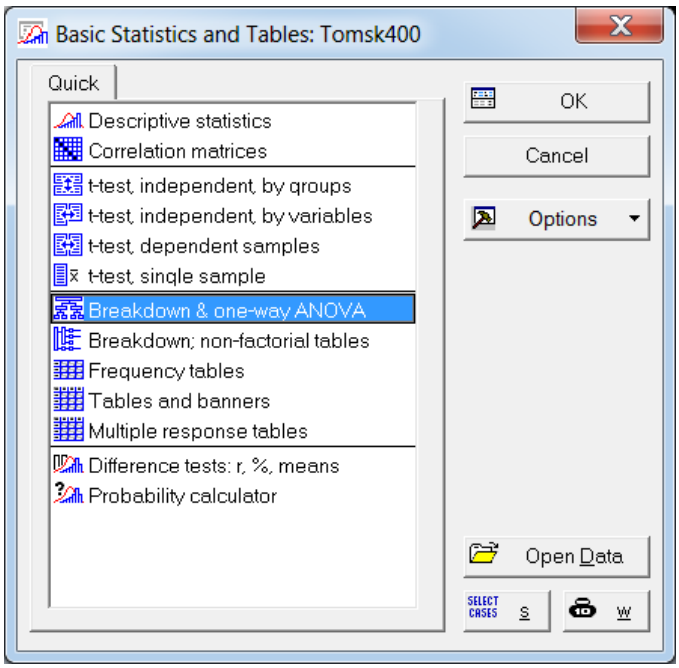


Рис. 1. Выбор метода однофакторного дисперсионного анализа

Выбор нескольких зависимых переменных в данном случае означает, что дисперсионный анализ мы будем проводить для каждой из них. Можно выбрать и несколько группирующих переменных, например помимо переменной «НП», задать еще переменную «Пол». Тем самым мы увеличиваем число градаций фактора. Сам фактор становится комбинированным, он одновременно будет учитывать и место проживания и пол респондента. Подчеркнем, еще раз, что выбор в данном случае более, чем одного фактора, не означает построение многофакторной модели, а просто увеличивает число уровней фактора.

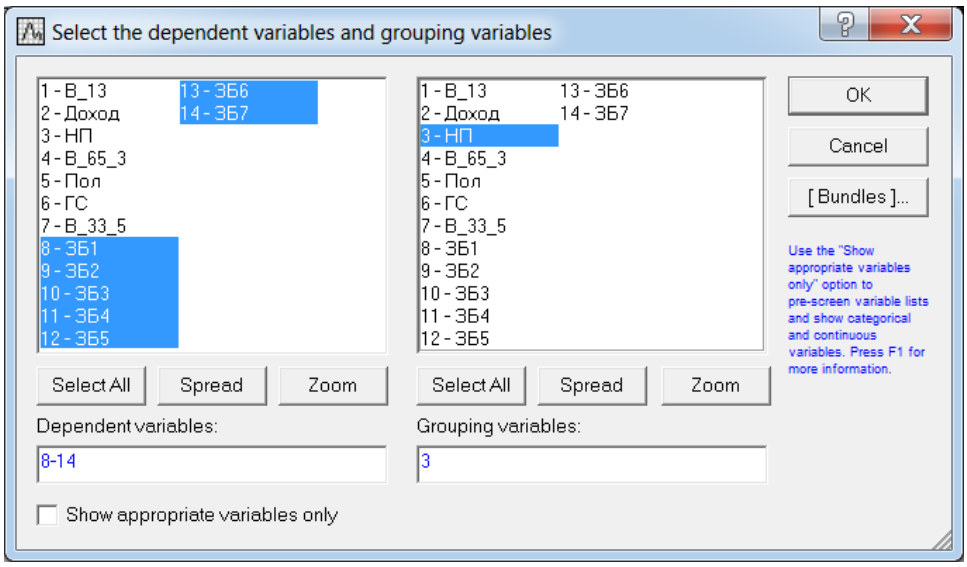


Рис. 2. Выбор зависимых и группирующей переменной для дисперсионного анализа

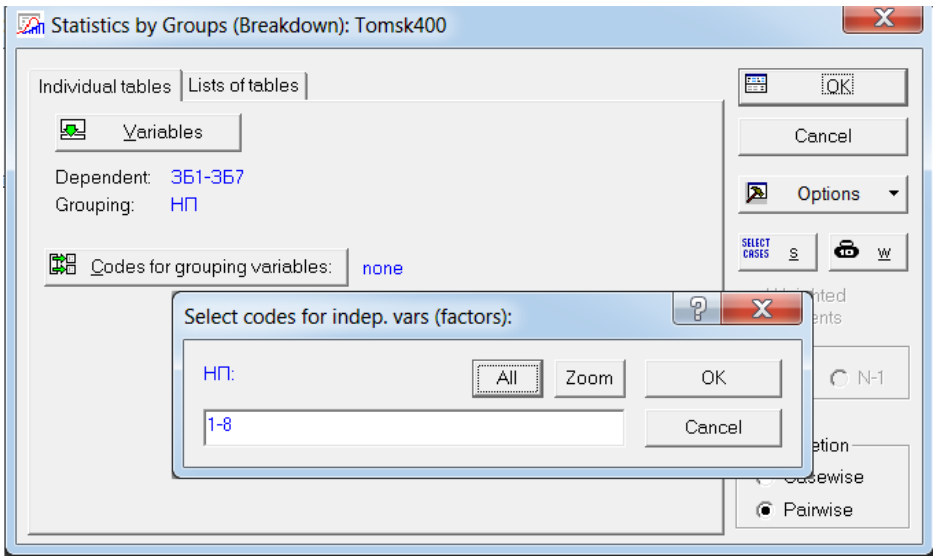


Рис. 3. Задание уровней фактора

Можно провести дисперсионный анализ не по всем уровням группирующей переменной (фактора), а только по заданным уровням. Для этого в окне выбора переменных для дисперсионного анализа (рис. 3) надо указать требуемые коды фактора.

После нажатия на клавишу «OK» переходим в окно результатов дисперсионного анализа – «Statistics by Groups - Results». Выберем вкладку «Quick» и нажмем на кнопку «Summary: Table of statistics». Получим таблицу описательной статистики исходных данных, изображенную на рис. 4.

Breakdown Table of Descriptive Statistics (Tomsk400)
N=400 (No missing data in dep. var. list)

НП	ЗБ1 Means	ЗБ1 N	ЗБ1 Std.Dev.	ЗБ2 Means	ЗБ2 N	ЗБ2 Std.Dev.	ЗБ3 Means	ЗБ3 N	ЗБ3 Std.Dev.	ЗБ4 Means	ЗБ4 N	ЗБ4 Std.Dev.
1	0,560000	200	0,497633	0,155000	200	0,362813	0,300000	200	0,459408	0,230000	200	0,421889
2	0,527273	55	0,503857	0,127273	55	0,336350	0,145455	55	0,355808	0,145455	55	0,355808
3	0,586667	75	0,495748	0,133333	75	0,342224	0,320000	75	0,469617	0,293333	75	0,458356
4	0,640000	25	0,489898	0,320000	25	0,476095	0,240000	25	0,435890	0,200000	25	0,408248
5	0,384615	13	0,506370	0,307692	13	0,480384	0,230769	13	0,438529	0,153846	13	0,375534
6	0,833333	12	0,389249	0,166667	12	0,389249	0,333333	12	0,492366	0,250000	12	0,452267
7	0,583333	12	0,514929	0,500000	12	0,522233	0,250000	12	0,452267	0,333333	12	0,492366
8	0,500000	8	0,534522	0,000000	8	0,000000	0,375000	8	0,517549	0,125000	8	0,353553
All Grps	0,567500	400	0,496043	0,170000	400	0,376103	0,277500	400	0,448326	0,227500	400	0,419743

Рис. 4. Описательная статистика исходных данных

По каждой из выбранных переменных в таблице приведены значения среднего, количества наблюдений и стандартного отклонения.

Результаты дисперсионного анализа получим, если на вкладке «Quick» нажмем на кнопку «Analysis of Variance» (рис. 5).

Analysis of Variance (Tomsk400)
Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
ЗБ1	1,581486	7	0,225927	96,59601	392	0,246418	0,916841	0,493178
ЗБ2	2,593345	7	0,370478	53,84666	392	0,137364	2,697054	0,009699
ЗБ3	1,381777	7	0,197397	78,81572	392	0,201061	0,981778	0,444004
ЗБ4	1,010495	7	0,144356	69,28700	392	0,176753	0,816715	0,573774
ЗБ5	3,498497	7	0,499785	96,41150	392	0,245948	2,032079	0,050103
ЗБ6	7,803397	7	1,114771	80,97410	392	0,206567	5,396667	0,000006
ЗБ7	1,865682	7	0,266526	49,83182	392	0,127122	2,096616	0,042975

Рис. 5. Результаты дисперсионного анализа

В каждой строке таблицы представлены результаты дисперсионного анализа по соответствующей зависимой переменной. В столбцах таблицы отображены: сумма квадратов межгруппового разброса (эффект фактора), число степеней свободы эффекта, средний эффект, остаточная сумма квадратов отклонений (сумма квадратов внутригруппового разброса), число степеней свободы для остаточной суммы квадратов, средняя остаточная сумма квадратов (оценка внутригрупповой дисперсии), значение статистики Фишера, наблюдаемый уровень значимости. В таблице выделены строки, где уровень значимости $p < 0{,}05$, то есть для той переменной, для которой значимо влияние различных уровней фактора «НП».

Таким образом, по результатам дисперсионного анализа мы можем утверждать, что уровень таких заболеваний, как «ЗБ2» – бронхо-легочные, «ЗБ6» - невралгические, «ЗБ7» - урологические (гинекологические) различен в различных населенных пунктах. Кроме того, слабо значимое различие уровней заболевания по различным населенным пунктам можно отметить и для заболевания «ЗБ5» - заболевания опорно-двигательной системы.

Если на вкладке «Quick» нажать на кнопку «Interaction plots», то получим графики зависимостей средних значений выбранных переменных от уровней фактора с указанием 95% доверительных интервалов. На рис. 6 приведен такой график для переменной «ЗБ6» - частоты невралгических заболеваний.

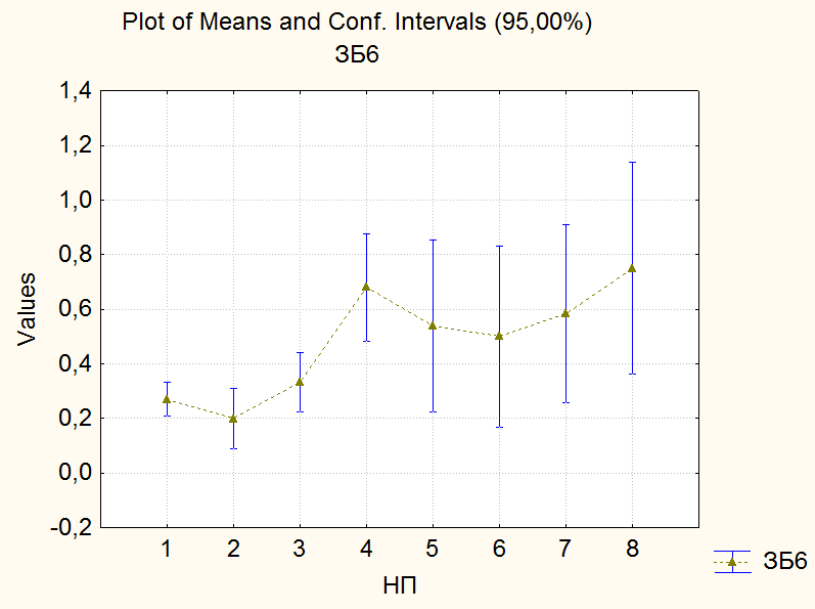


Рис. 6. Зависимость уровня заявленных невралгических заболеваний (переменная «ЗБ6») от уровней фактора «НП» (места проживания)

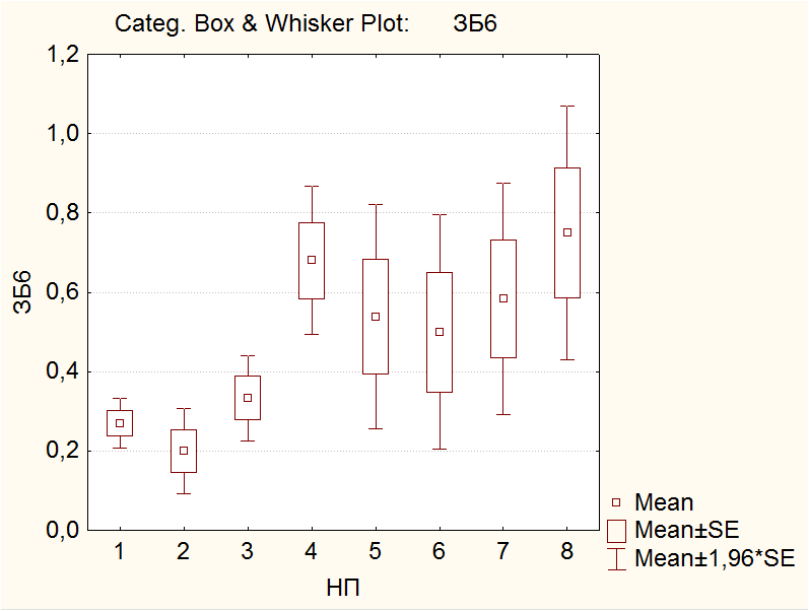


Рис. 7. Диаграммы размаха типа «ящики-усы» для уровня заявленных невралгических заболеваний (переменная «ЗБ6») в зависимости от уровней фактора «НП» (места проживания)

Если на вкладке «Quick» нажать на кнопку «Categorized box & whisker plot», то получим аналогичные графики в виде диаграммы типа «ящики-усы» (рис. 7).

Как уже отмечалось ранее, дисперсионный анализ позволяет установить факт зависимости средних значений одной величины от уровней другой величины, но не позволяет сделать вывод о различии каких-либо средних между собой. Если установлен факт различия средних, то для выяснения какие из средних различаются, следует перейти на вкладку апостериорных сравнений средних «Post-hoc» и выбрать один из методов множественного сравнения (рис. 8).

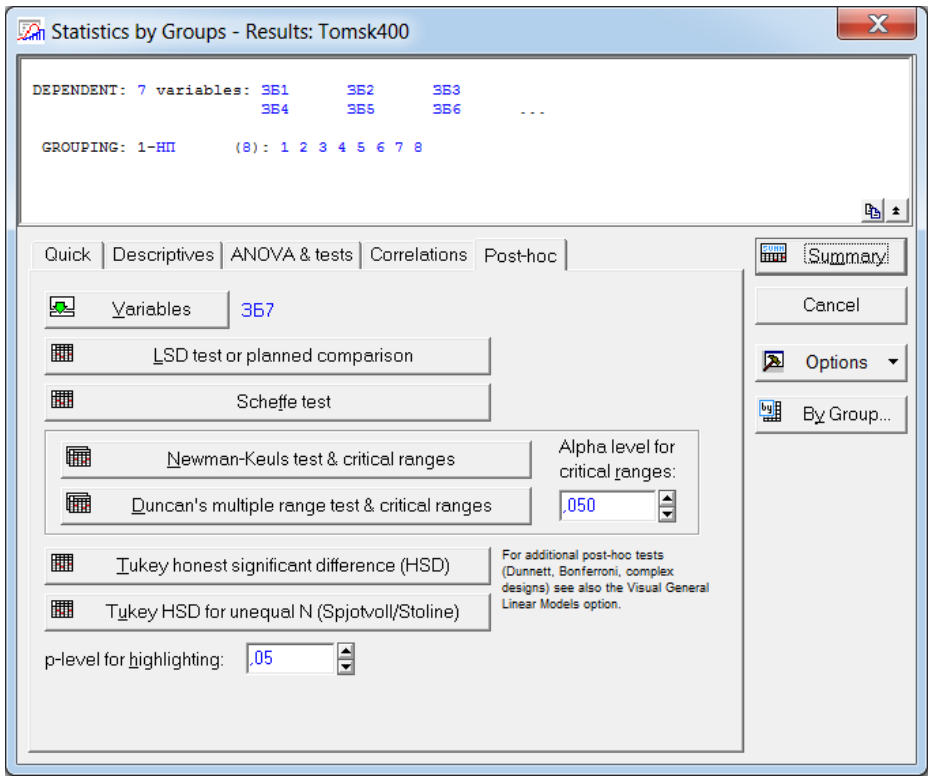


Рис. 8. Окно выбора теста множественного сравнения средних

Результаты множественного сравнения средних для переменной «ЗБ6» (уровня заявленных невралгических заболеваний) по критериям наименьшей значимой разности (LSD), Ньюмана-Келса, достоверно значимой разности Тьюки (HSD), Шеффе приведены на рис. 9-12.

		LSD Test; Variable: ЗБ6 (Tomsk400) Marked differences are significant at p < ,05000							
НП		**{1} M=,27000**	{2} M=,20000	{3} M=,33333	{4} M=,68000	{5} M=,53846	{6} M=,50000	{7} M=,58333	{8} M=,75000
1	**{1}**		0,312370	0,304041	0,000026	0,039703	0,089419	0,020875	0,003598
2	{2}	0,312370		0,099228	0,000015	0,016202	0,038947	0,008444	0,001496
3	{3}	0,304041	0,099228		0,001045	0,133825	0,238934	0,077642	0,014134
4	{4}	0,000026	0,000015	0,001045		0,362992	0,260127	0,545112	0,704772
5	{5}	0,039703	0,016202	0,133825	0,362992		0,832691	0,805328	0,300945
6	{6}	0,089419	0,038947	0,238934	0,260127	0,832691		0,653592	0,228884
7	{7}	0,020875	0,008444	0,077642	0,545112	0,805328	0,653592		0,422222
8	{8}	0,003598	0,001496	0,014134	0,704772	0,300945	0,228884	0,422222	

Рис. 9. Результаты множественного сравнения по критерию LSD

НП		Newman-Keuls test; Variable: ЗБ6 (Tomsk400) Marked differences are significant at p < ,05000							
		{1} M=,27000	{2} M=,20000	{3} M=,33333	{4} M=,68000	{5} M=,53846	{6} M=,50000	{7} M=,58333	{8} M=,75000
1	**{1}**		0,644294	0,676143	0,074348	0,287503	0,282853	0,234642	0,025928
2	{2}	0,644294		0,653307	0,025928	0,167716	0,195991	0,115851	0,006939
3	{3}	0,676143	0,653307		0,149154	0,365893	0,271639	0,351131	0,066132
4	{4}	0,074348	0,025928	0,149154		0,619022	0,634930	0,523742	0,644294
5	{5}	0,287503	0,167716	0,365893	0,619022		0,799758	0,767265	0,502278
6	{6}	0,282853	0,195991	0,271639	0,634930	0,799758		0,846659	0,465970
7	{7}	0,234642	0,115851	0,351131	0,523742	0,767265	0,846659		0,514490
8	{8}	0,025928	0,006939	0,066132	0,644294	0,502278	0,465970	0,514490	

Рис. 10. Результаты множественного сравнения по критерию Ньюмана-Келса

НП		Tukey HSD test; Variable: ЗБ6 (Tomsk400) Marked differences are significant at p < ,05000							
		{1} M=,27000	{2} M=,20000	{3} M=,33333	{4} M=,68000	{5} M=,53846	{6} M=,50000	{7} M=,58333	{8} M=,75000
1	**{1}**		0,972880	0,970139	0,000575	0,438735	0,685596	0,282672	0,066913
2	{2}	0,972880		0,717990	0,000339	0,233948	0,433427	0,139273	0,030080
3	{3}	0,970139	0,717990		0,021485	0,806569	0,937946	0,641139	0,210616
4	{4}	0,000575	0,000339	0,021485		0,985150	0,950916	0,998819	0,999948
5	{5}	0,438735	0,233948	0,806569	0,985150		0,999999	0,999997	0,969057
6	{6}	0,685596	0,433427	0,937946	0,950916	0,999999		0,999836	0,930716
7	{7}	0,282672	0,139273	0,641139	0,998819	0,999997	0,999836		0,992986
8	{8}	0,066913	0,030080	0,210616	0,999948	0,969057	0,930716	0,992986	

Рис. 11. Результаты множественного сравнения по критерию HSD Тьюки

НП		Scheffe Test; Variable: ЗБ6 (Tomsk400) Marked differences are significant at p < ,05000							
		{1} M=,27000	{2} M=,20000	{3} M=,33333	{4} M=,68000	{5} M=,53846	{6} M=,50000	{7} M=,58333	{8} M=,75000
1	**{1}**		0,994357	0,993720	0,012947	0,749050	0,893483	0,613942	0,287293
2	{2}	0,994357		0,908078	0,008720	0,560323	0,745153	0,429949	0,179650
3	{3}	0,993720	0,908078		0,146308	0,943747	0,985648	0,872008	0,531908
4	{4}	0,012947	0,008720	0,146308		0,997083	0,989024	0,999799	0,999992
5	{5}	0,749050	0,560323	0,943747	0,997083		1,000000	1,000000	0,993467
6	{6}	0,893483	0,745153	0,985648	0,989024	1,000000		0,999974	0,983700
7	{7}	0,613942	0,429949	0,872008	0,999799	1,000000	0,999974		0,998697
8	{8}	0,287293	0,179650	0,531908	0,999992	0,993467	0,983700	0,998697	

Рис. 12. Результаты множественного сравнения по критерию Шеффе

Как и ожидалось, наиболее консервативные результаты показал критерий Шеффе – различия всего в двух парах, а наименее консервативные результаты - критерий LSD - различия в 11 парах. Критерий Ньюмана-Келса в случае выборок равного объема более чувствителен, чем критерий Тьюки. Но в данном случае объемы выборок для различных уровней фактора сильно различаются, в этом

случае модифицированный критерий Ньюмана-Келса лучше не использовать. Наверное, в данном случае, следует ориентироваться на результаты критерия Тьюки, согласно которому, в нашем случае, различие средних в первую очередь обусловлено различием средних для уровней фактора 1 и 4, 2 и 4, 3 и 4, 2 и 8. Что означает, что существенно различается уровень заявленных невралгических заболеваний в г. Асино по сравнению с г. Томском, г. Северском и Томским районом, а также в пос. Тегульдет по сравнению с г. Северском.

Для достоверности полученных результатов дисперсионного анализа необходимо проверить предположения о нормальном распределении сравниваемых групп и об однородности дисперсий в группах. Гипотезу об однородности дисперсий можно проверить на вкладке «ANOVA & tests», используя критерии Левене и Брауна-Форсайта. Гипотезу о нормальности можно визуально проверить на вкладке «Descriptives», построив категоризованные гистограммы. Однако, в случае частотных данных, для неравных частот, дисперсии должны различаться. Сравнение на нормальность для дихотомических данных также лишено смысла. Если есть сомнения в полученных результатах, можно обратиться к непараметрическому дисперсионному анализу Краскела-Уоллиса.

Мы рассмотрели наиболее простую реализацию однофакторного дисперсионного анализа в пакете STATISTICA. Более “продвинутый вариант” реализован в модуле «ANOVA» в меню «Statistics» головного меню. Для выбора данного варианта запускаем в головном меню модуль «Statistics» и в стартовой панели выбираем пункт «ANOVA». В появившемся окне (рис. 13) выбираем тип анализа («One-way ANOVA» - однофакторный дисперсионный анализ) и задаем метод («Quick specs dialog - диалог быстрых спецификаций»).

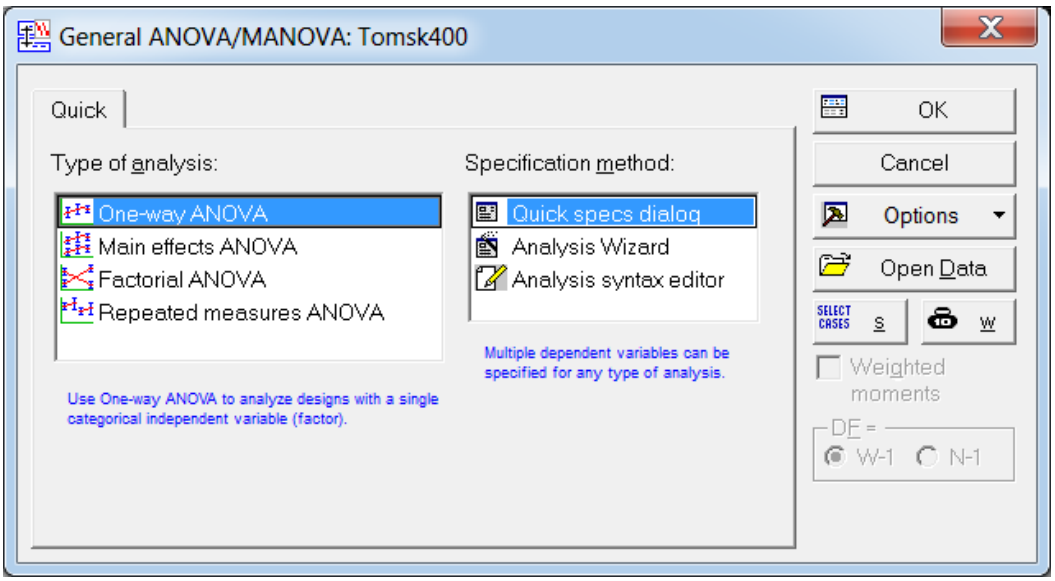


Рис. 5.13. Выбор метода дисперсионного анализа

После нажатия на «OK», попадаем в окно выбора переменных для анализа (рис. 14). Выбираем в качестве зависимых переменных переменные «ЗБ1» - «ЗБ7», а в качестве группирующей переменной (фактора) - переменную «НП». Можно также выбрать уровни (коды) группирующей переменной (фактора), по которым будет проводиться анализ. Если коды не задавать, анализ будет проводиться по всем уровням группирующей переменной. После нажатия на клавишу «OK» переходим в окно результатов дисперсионного анализа – «ANOVA Results 1» и выбираем вкладку «Summary» (рис. 15).

Для просмотра описательной статистики на вкладки «Summary» следует выбрать «Cell statistics». Для просмотра результатов дисперсионного анализа выбираем «Univariate results», в результате получаем таблицу, изображенную на рис. 16.

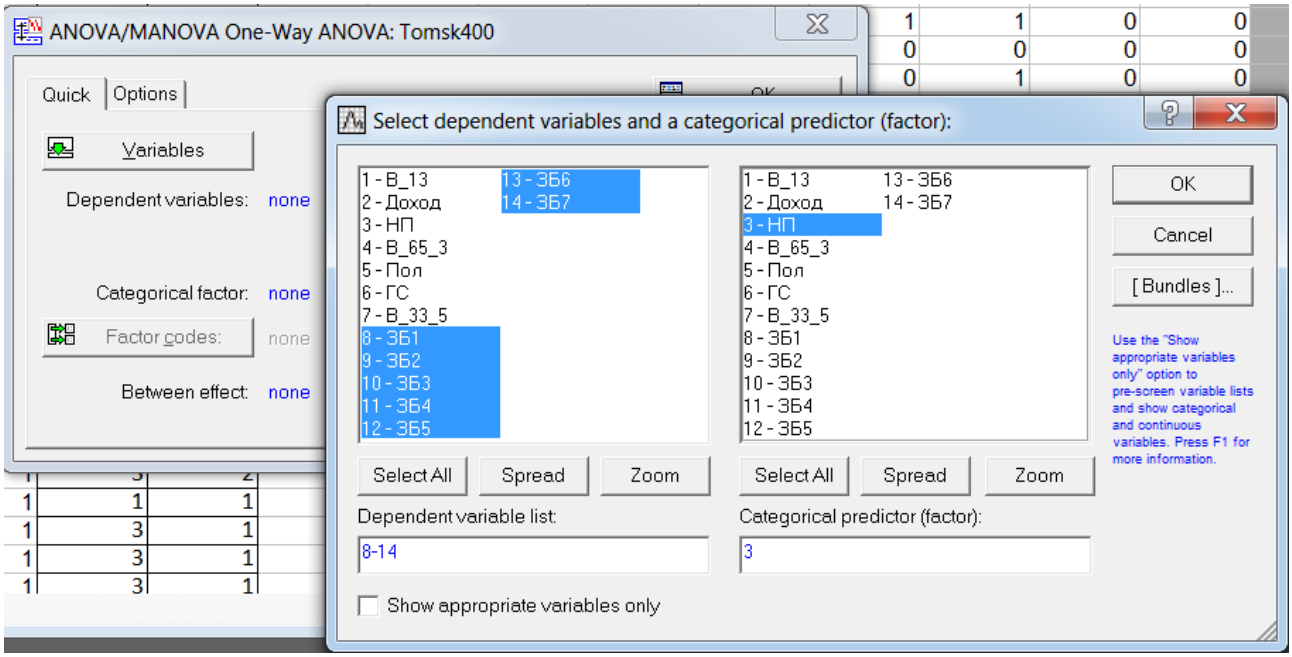


Рис. 14. Выбор переменных для дисперсионного анализа

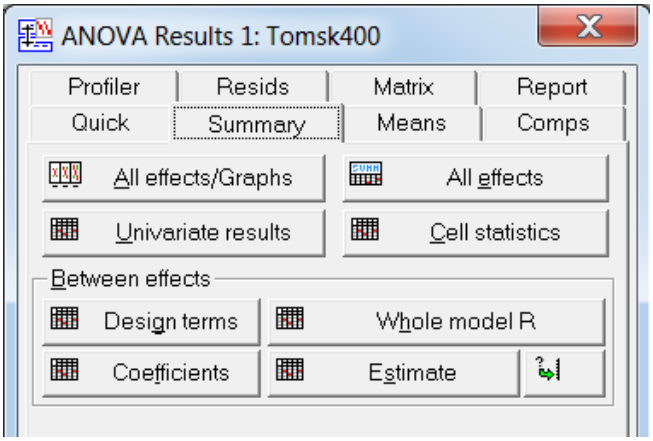


Рис. 15. Часть окна результатов дисперсионного анализа

Univariate Results for Each DV (Tomsk400)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Degr. of Freedom	ЗБ1 SS	ЗБ1 MS	ЗБ1 F	ЗБ1 p	ЗБ2 SS	ЗБ2 MS	ЗБ2 F	ЗБ2 p
Intercept	1	47,85449	47,85449	194,2002	0,000000	6,56919	6,569194	47,82329	0,000000
НП	7	1,58149	0,22593	0,9168	0,493178	2,59334	0,370478	2,69705	0,009699
Error	392	96,59601	0,24642			53,84666	0,137364		
Total	399	98,17750				56,44000			

Рис. 16. Результаты дисперсионного анализа, включая анализ различий между выборками

Первую строку таблицы (эффект «Intercept») можно проигнорировать. Во второй строке таблицы для каждой из переменных «ЗБ1», «ЗБ2», …, «ЗБ7», приводятся суммы квадратов отклонений (SS), средние суммы квадратов отклонений (MS) для межгруппового разброса (эффекта фактора «НП») с указанием значения статистики Фишера F и уровня значимости. В третьей строке таблицы приводятся суммы квадратов отклонений (SS), средние суммы квадратов отклонений (MS) для остатков или внутригруппового разброса. В последней строке указаны полные суммы квадратов отклонений по каждой переменной «ЗБ1», «ЗБ2», …, «ЗБ7». Можно убедиться, что данная таблица, за исключением формы отображения эквивалентна таблице, изображенной на рис. 5.

Для графического отображения результатов дисперсионного анализа можно также нажать на кнопку «All effects/Graphs». В появившемся окне далее следует нажать кнопку «OK» lkzвыбора переменных, и выбрать переменные, для которых будут построены графики средних с доверительными интервалами (рис. 17).

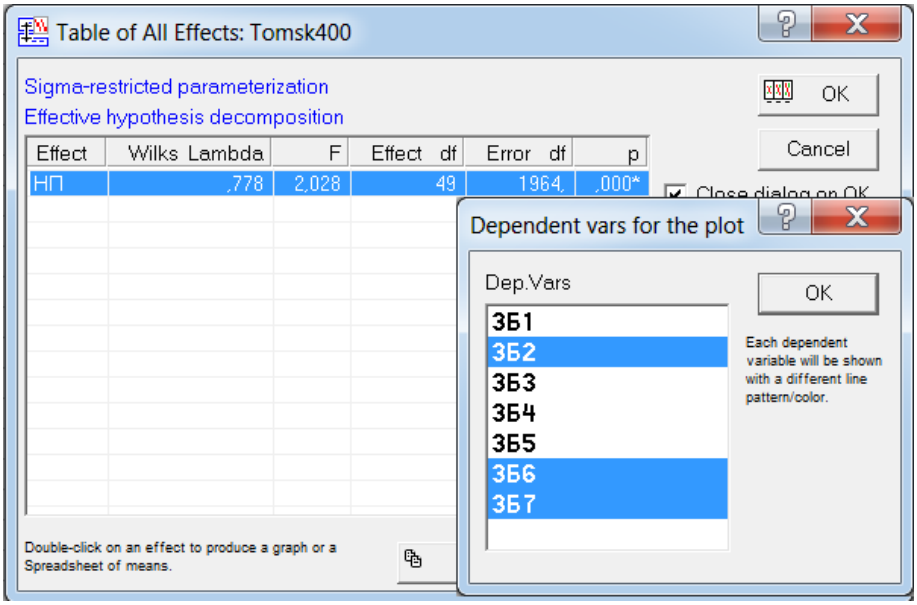


Рис. 17. Окно для выбора отображения результатов дисперсионного анализа в графическом/табличном виде

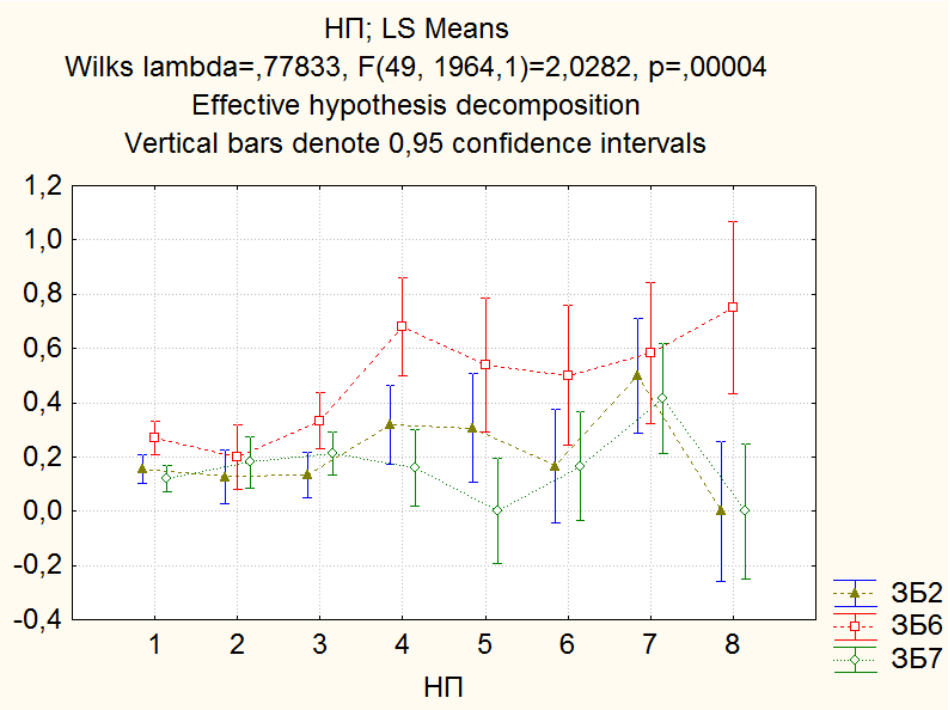


Рис. 18. Графики средних для переменных «ЗБ2», «ЗБ6», «ЗБ7» в зависимости от уровней фактора «НП»

В результате получим графики средних (рис. 18), аналогичные изображенному на рис. 6. Заметим, что в таблице, изображенной на рис. 17, и на графике, изображенном на рис. 18 отображаются значение и уровень значимости статистики лямбда Уилкса, которая характеризует различие векторов средних по всем переменным. Указанное значение статистики высоко значимо ($p = 0{,}00004$), это означает, что уровни заболеваний по различным заболеваниям существенно различаются, что является достаточно очевидным фактом и не является целью данного исследования.

Чтобы получить результаты множественного сравнения, следует в модуле результатов дисперсионного анализа – «ANOVA Results 1» выбрать расширенный режим путем нажатия кнопки «More results», перейти на вкладку апостериорных сравнений средних «Post-hoc» и выбрать один из методов множественного сравнения (рис. 19). Для режима отображения (параметр «Display») устанавливаем «Significant differences» (значимые разности).

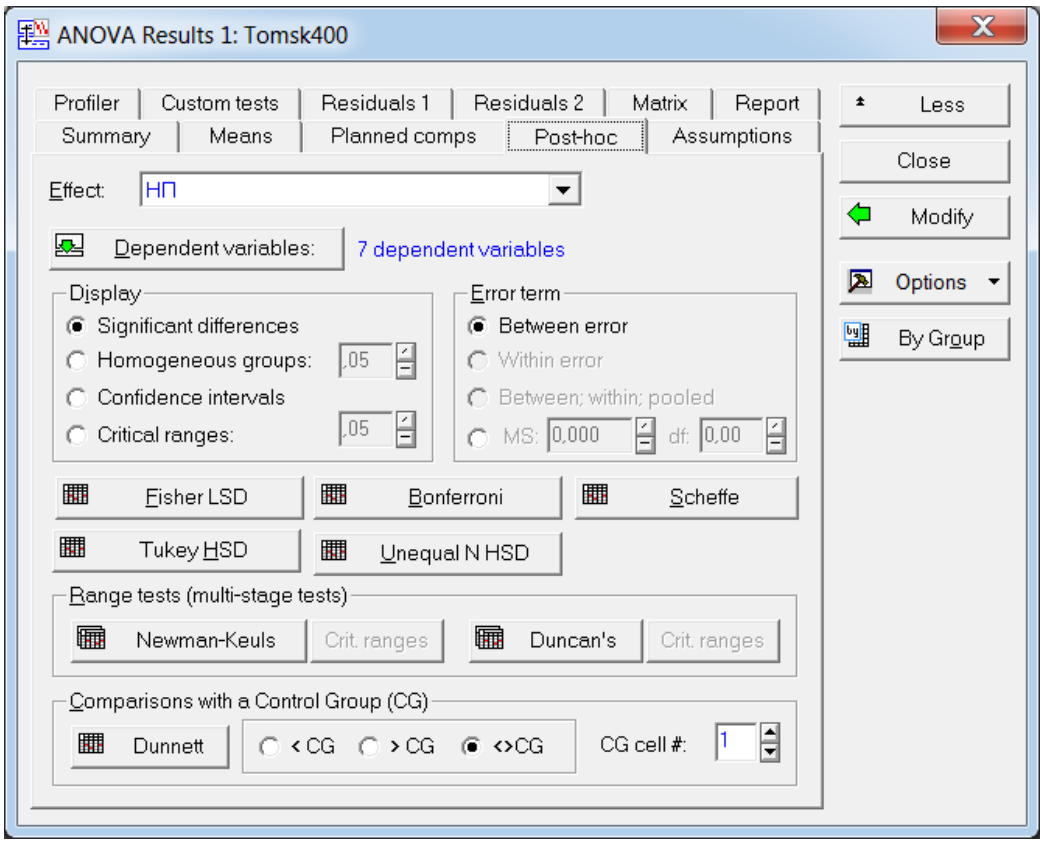


Рис. 19. Вкладка выбора метода апостериорных сравнений

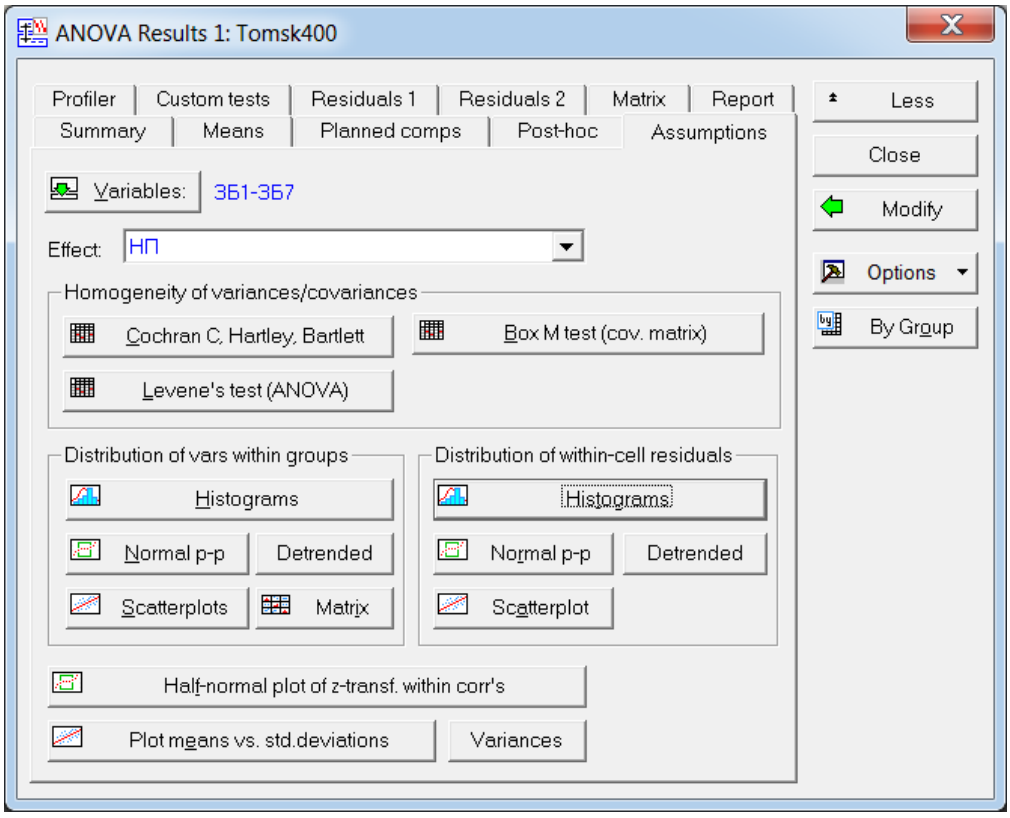


Рис. 20. Вкладка «Assumptions» - проверка предположений об однородности дисперсий и нормальности распределений

Проверку гипотез однородности дисперсий можно осуществить на вкладке «Assumptions» (рис. 20), здесь же можно визуально проверить нормальность распределения, построив гистограммы, как для переменных, так и для остатков (хотя в случае дихотомических данных особого смысла в этих графиках нет).

Помимо множественного сравнения средних, в модуле «ANOVA Results 1» на вкладке «Planed comps» можно проверять гипотезы о равенстве нулю контрастов, то есть сравнивать средние для любых сочетаний групп. Перейдем на вкладку «Planed comps» и нажмем на кнопку «Specify contrasts for LS means» для построения контраста (рис. 21).

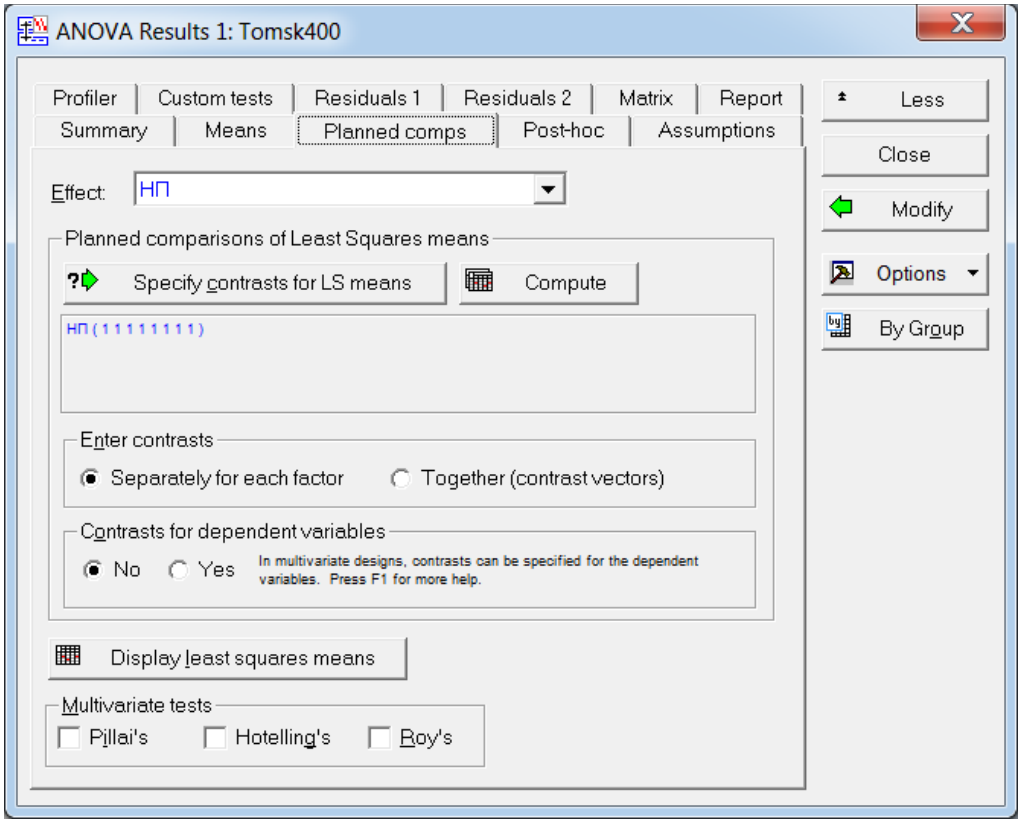


Рис. 21. Вкладка «Planed comps» - проверка гипотез о контрастах

В открывшемся окне «Specify Contrasts for this Factor» строим контраст, задавая коэффициенты, как показано на рис. 22. Значения коэффициентов можно вводить вручную, можно использовать панели, содержащие значения 0, ±1, ±2 справа.

С учетом того, что контраст использует средние значения по группам, мы создали контраст вида: $C = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \frac{1}{5}(\mu_4 + \mu_5 + \mu_6 + \mu_7 + \mu_8)$ (с точностью до постоянного множителя). Соответственно, проверяя гипотезу $H_0 : C = 0$, мы будем проверять гипотезу о равенстве средних двух групп, первая из которых содержит значения фактора «НП» 1 - 3 (г. Томск, г. Северск, Томский район), а вторая содержит значения фактора «НП» 4 - 8 (остальные населенные пункты).

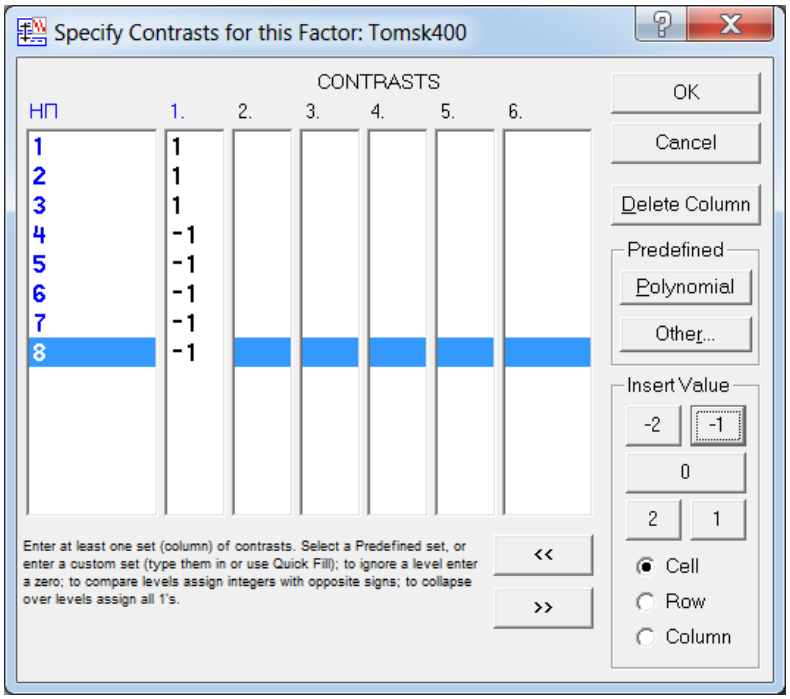
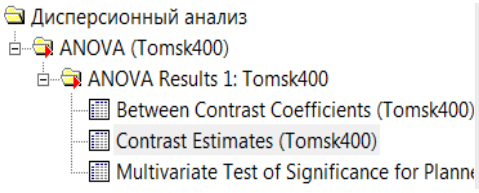


Рис. 22. Построение контраста

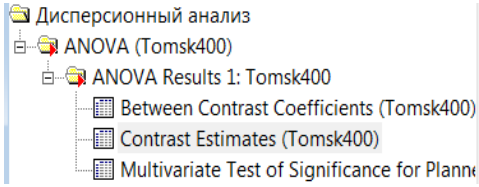
После построения контраста (контрастов), нажав «OK» возвращаемся на вкладку «Planed comps» и нажимаем на кнопку «Compute» для выполнения теста. В результате, в рабочей книге в разделе «ANOVA Results 1» на странице «Contrast Estimates» получим результаты тестирования. На рис. 23 приведены результаты тестирования для переменной «ЗБ1», а на рис. 24 для переменной «ЗБ2».



Contrast Estimates (Tomsk400)
Contrast estimates for dependent variables

Contrast	ЗБ1 Estimate	ЗБ1 Std.Err	ЗБ1 t	ЗБ1 p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
CNTRST1	-0,113537	0,265884	-0,427019	0,669600	-0,636274	0,409199

Рис. 23. Проверка значимости контраста для переменной «ЗБ1»



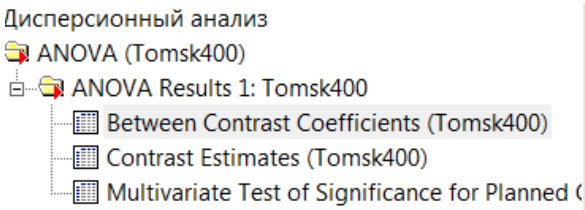
Contrast Estimates (Tomsk400)
Contrast estimates for dependent variables

Contrast	ЗБ2 Estimate	ЗБ2 Std.Err	ЗБ2 t	ЗБ2 p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
CNTRST1	-0,451262	0,198514	-2,27320	0,023555	-0,841547	-0,060976

Рис. 24. Проверка значимости контраста для переменной «ЗБ2»

В столбцах таблицы последовательно приведены: значения контраста, стандартная ошибка контраста, значение статистики LSD, уровень значимости статистики, границы 95% доверительного интервала для контраста. Как видим для переменной «ЗБ1», значение статистики не значимо ($p = 0{,}6696$), а для переменной «ЗБ2», значение статистики значимо ($p = 0{,}23555$). Это означает, что для переменной «ЗБ1» (сердечно-сосудистые заболевания) частоты заболеваний в двух группах не различается, а для переменной «ЗБ2» (бронхолегочные заболевания) частоты заболеваний в двух группах различаются статистически значимо.

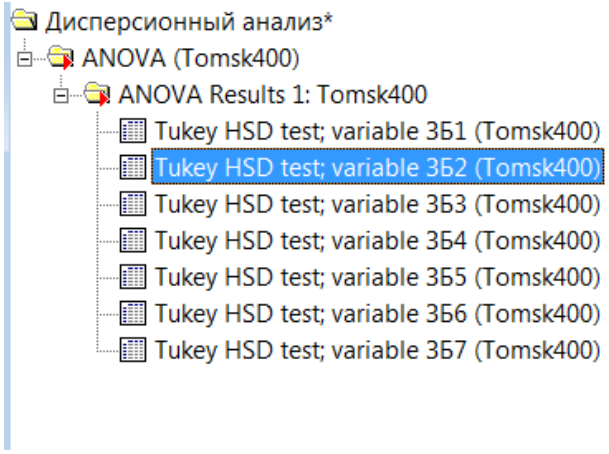
Также в рабочей книге в разделе «ANOVA Results 1», на странице «Between Contrast Coefficients» можно посмотреть значения коэффициентов для контраста, которые выбрала STATISTICA (рис. 25). Можно убедиться, что данные коэффициенты, с точностью до постоянного множителя, совпадают с коэффициентами {1/3, 1/3, 1/3, -1/5, -1/5, -1/5, -1/5, -1/5}.



Between Contrast Coefficients (Tomsk400)
Coefficients for each cell in the selected effect

Cell No.	НП	Cell N	CNTRST1
1	1	200	1,250000
2	2	55	1,250000
3	3	75	1,250000
4	4	25	-0,750000
5	5	13	-0,750000
6	6	12	-0,750000
7	7	12	-0,750000
8	8	8	-0,750000

Рис. 25. Коэффициенты контраста CNTRS1



Tukey HSD test; variable ЗБ2 (Tomsk400)
Homogenous Groups, alpha = ,10000
Error: Between MS = ,13736, df = 392,00

Cell No.	НП	ЗБ2 Mean	1	2
8	8	0,000000	****	
2	2	0,127273	****	
3	3	0,133333	****	
1	1	0,155000	****	
6	6	0,166667	****	****
5	5	0,307692	****	****
4	4	0,320000	****	****
7	7	0,500000		****

Рис. 26. Однородные кластеры групп в соответствии с выбранным критерием множественного сравнения (HSD Тьюки) и заданным уровнем значимости

Если на вкладке «Post-hoc» для режима отображения (параметр «Display») установить «Homogeneous groups» (однородные группы), то будут выделены однородные (различающиеся незначимо в соответствии с выбранным критерием множественного сравнения) кластеры групп, расположенные в порядке возрастания средних значений. Полученные группы для различных переменных располагаются на различных страницах в рабочей книге результатов дисперсионного анализа (рис. 26).

Как видим, из рис. 26. для переменной «ЗБ2» на уровне значимости 0,1 можно сформировать два кластера населенных пунктов. Первый содержит населенные пункты {«НП8», «НП2», «НП3», «НП1», «НП6», «НП5», «НП4»}, а второй населенные пункты {«НП6», «НП5», «НП4», «НП7»}. Заметим, что чем

больше уровень значимости, тем более близкие группы будут выделены и, соответственно возрастет количество групп.

Пример 2. Результаты ответов 400 респондентов на вопросы анкеты «Томск 400» «Как Вы оцениваете Ваше здоровье в сравнении со здоровьем Ваших сверстников» (варианты ответов: “Очень хорошее”, “Хорошее”, “Среднее”, “Плохое”, “Очень плохое”, “Затрудняюсь ответить”) представлены в виде числовой выборки кодов ответов со значениями, соответственно, {1,2,3,4,5,6}. Также имеется выборка числовых кодов, соответствующих месту проживания респондента (1 – «Томск», 2 - «Северск», 3 – «Томский район», 4 - «Асино», 5 – «Асиновский район», 6 - «Каргасокский район», 7 – «Каргасок», 8 - «Тегульдет»). Используя дисперсионный анализ, установить, одинаково ли оценивают свое здоровье респонденты в различных населенных пунктах.

Поскольку зависимая переменная (варианты ответов на вопрос «Как Вы оцениваете Ваше здоровье в сравнении со здоровьем Ваших сверстников») категориального типа, то для выявления различия в ответах на вопросы респондентов различных населенных пунктов используем непараметрический дисперсионный анализ Краскела-Уоллиса.

Выборочные данные представлены в нашей таблице данных под именами «B_13» и «НП». Чтобы исключить из рассмотрения респондентов, давших на вопрос «Как Вы оцениваете Ваше здоровье в сравнении со здоровьем Ваших сверстников» ответ «Затрудняюсь ответить», незабываем указать код категории, которые мы исключаем из анализа. Для этого в таблице данных кликаем дважды на имени переменной «B_13» и в раскрывшемся окне свойств переменной устанавливаем значение параметра «MD code» равным значению 6 (код ответа «Затрудняюсь ответить»).

Предварительно можно качественно оценить различие средних, построив диаграммы размаха в соответствующем разделе модуля «Descriptive statistics». Однако, это можно будет сделать и непосредственно в модуле непараметрического дисперсионного анализа.

Для проведения непараметрического дисперсионного анализа рангов Краскела-Уоллиса проделаем следующее. Запускаем в головном меню модуль «Statistics», в стартовой панели выбираем пункт «Nonparametrics». В меню модуля «Nonparametric Statistics» (рис. 27) выбираем раздел «Comparing multiple indep. Samples (groups)» («Сравнение нескольких независимых выборок)».

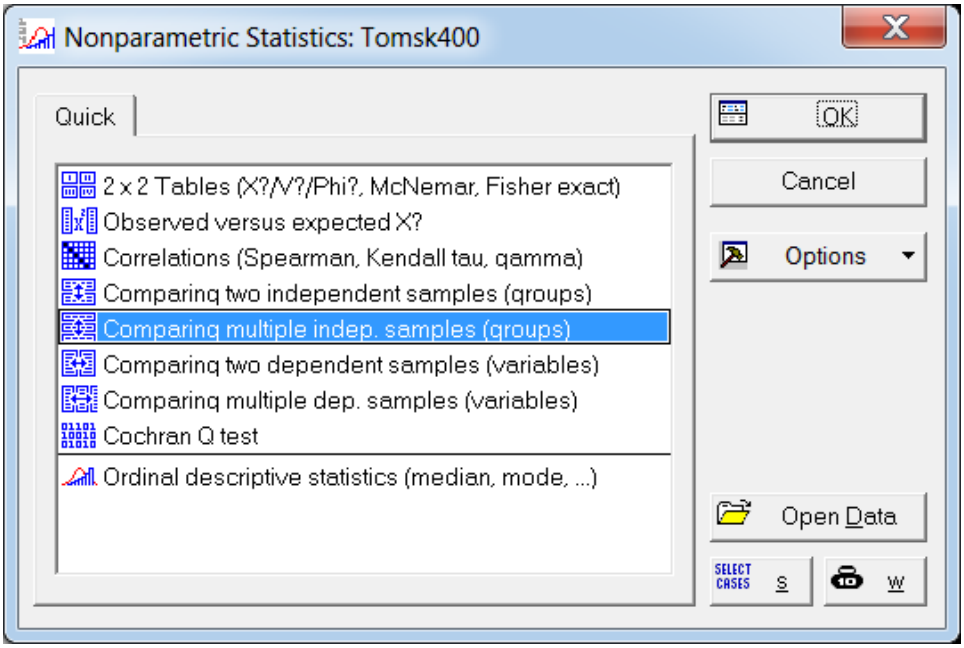


Рис. 27. Выбор метода непараметрического дисперсионного анализа в стартовом окне модуля «Nonparametric Statistics»

В появившемся окне модуля «Kruskal-Wallis ANOVA and Median Test» (рис. 28), выбираем переменные, нажав на кнопку «Variables». В качестве зависимой переменной выбираем переменную «B_13» а в качестве группирующей – переменную «НП».

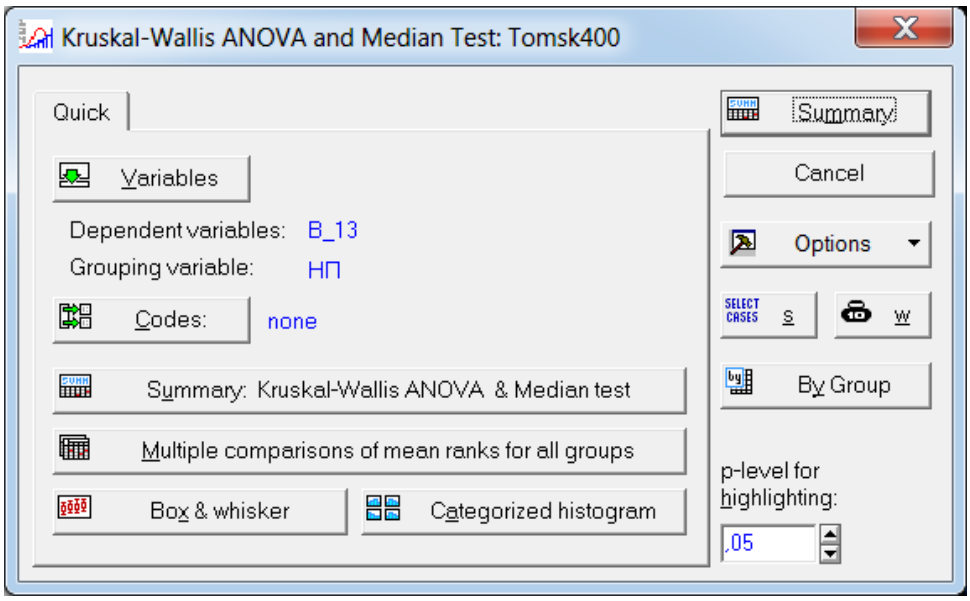


Рис. 28. Окно модуля непараметрического дисперсионного анализа

Оценим предварительно качественно различие средних по уровням фактора. Для этого нажимаем на кнопку «Box & whisker», в качестве переменной выбираем «B_13», а в качестве параметров для диаграммы типа «ящики-усы» параметры «Median / Quart / Range» («Медиана / Квартильный размах / Полный размах». В результате получаем диаграмму, изображенную на рис. 29. Как видим, на основе данной диаграммы трудно что-либо сказать о различии средних. Количество уровней зависимой переменной невелико, поэтому медианы для всех категорий переменной «НП» совпадают, и, соответственно, все интервалы размаха перекрываются. Но совпадение самих значений медиан, еще не означает, что число значений больших (меньших) медианы для разных уровней фактора одинаково. Парадокс, но мы проверяем гипотезу о “различии” медиан, при условии их “равенства”! Дело в том, что со статистической точки зрения, медиана просто делит всю совокупность в определенном соотношении (причем не обязательно 50% на 50% - смотри внимательно определение медианы). И если эти соотношения для выборок различаются, это и означает различие медиан двух совокупностей.

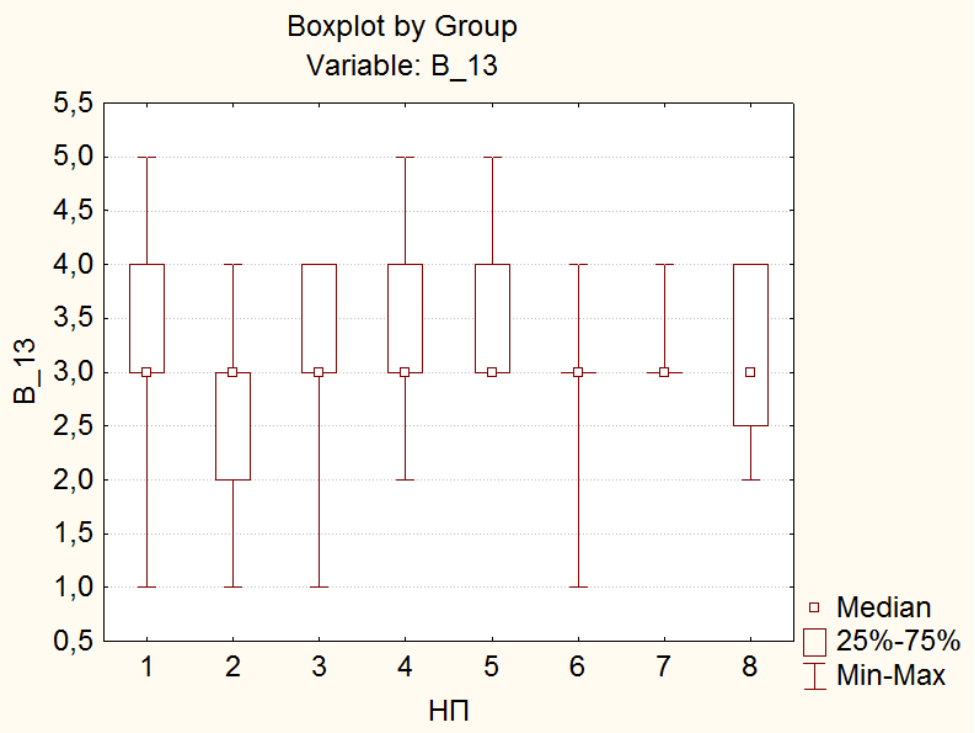


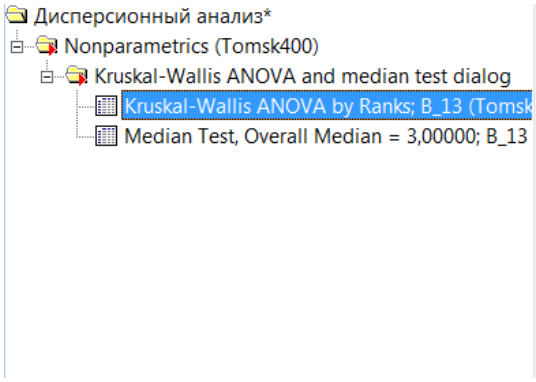
Рис. 29. Диаграммы размаха по категориям переменной «НП»

Вернемся в окно непараметрического дисперсионного анализа (рис. 5.30). Нажав на кнопку «Summary», в рабочей книге в разделе «Kruskal-Wallis ANOVA and median test dialog» на странице «Kruskal-Wallis ANOVA by Ranks» получим результаты дисперсионного анализа Краскела-Уоллиса, а на странице «Median Test» результаты медианного теста.

Согласно результатам дисперсионного анализа Краскела-Уоллиса (рис. 30), существует статистически значимое ($p = 0{,}0341$) влияние уровней фактора «НП» на значения переменной «B_13». Другими словами, респонденты в различных населенных пунктах по-разному оценивают свое здоровье.

Результаты медианного теста также показывают различие в ответах для различных населенных пунктов (рис. 31) на уровне $p = 0{,}0275$. В медианном тесте выдается также информация о наблюдаемом числе («observed») значений, меньше либо равных медианы (и, соответственно, больше медианы), и о ожидаемом числе («expected») тех же значений, вычисленном при условии истинности нулевой гипотезы о равенстве медиан признака «B_13» при различных уровнях

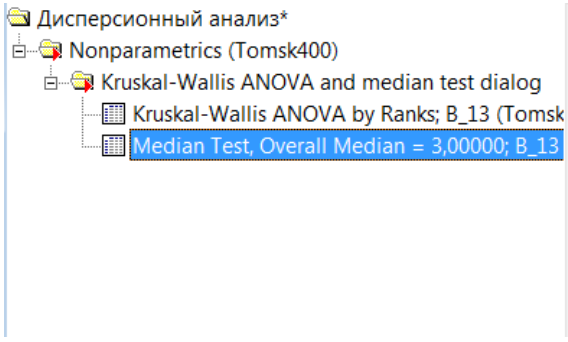
фактора «НП». Ориентируясь на эти значения, можно сделать выводы о том какие группы и как различаются.



Kruskal-Wallis ANOVA by Ranks; B_13 (Tomsk400)
Independent (grouping) variable: НП
Kruskal-Wallis test: H (7, N= 395) =15,14943 p =,0341

Depend.: B_13	Code	Valid N	Sum of Ranks
1	1	200	41042,50
2	2	51	7717,00
3	3	74	14827,50
4	4	25	5212,00
5	5	13	3176,50
6	6	12	2297,50
7	7	12	2317,50
8	8	8	1619,50

Рис. 30. Результаты дисперсионного анализа Краскела-Уоллиса



Median Test, Overall Median = 3,00000; B_
Independent (grouping) variable: НП
Chi-Square = 15,74775 df = 7 p = ,0275

Dependent: B_13	1	2	3	4
<= Median: observed	143,0000	48,00000	54,00000	18,00000
expected	150,3797	38,34684	55,64051	18,79747
obs.-exp.	-7,3797	9,65316	-1,64051	-0,79747
> Median: observed	57,0000	3,00000	20,00000	7,00000
expected	49,6203	12,65316	18,35949	6,20253
obs.-exp.	7,3797	-9,65316	1,64051	0,79747
Total: observed	200,0000	51,00000	74,00000	25,00000

Рис. 31. Результаты медианного теста

Так, для г. Томска (НП = 1) число наблюдаемых значений больших медианы (57) больше ожидаемого (49,6203). Это предположительно означает, что респонденты г. Томска хуже оценивают свое здоровье, чем, например, респонденты г. Северска (НП = 2), для которых число наблюдаемых значений больших медианы (3) меньше ожидаемого (12,65316).

Статистически определить между какими группами наблюдается значимое различие можно, используя множественное апостериорное сравнение средних рангов. Чтобы получить результаты множественного сравнения рангов в окне модуля «Kruskal-Wallis ANOVA and Median Test» нажимаем кнопку «Multiple comparisons of mean ranks for all groups», в результате получаем таблицу, изображенную на рис. 32. Как видим, только для пары г. Томск – г. Северск можно

считать, что существует слабо значимое различие ($p = 0{,}0733$). Поскольку данный результат был получен после значимого результата дисперсионного анализа, следует признать, что эта пара и определила результат дисперсионного анализа.

Depend.: B_13	Multiple Comparisons p values (2-tailed); B_13 (Tomsk400) Independent (grouping) variable: НП Kruskal-Wallis test: H (7, N= 395) =15,14943 p =,0341 **1** **R:205,21**	2 R:151,31	3 R:200,37	4 R:208,48	5 R:244,35	6 R:191,46	7 R:193,13	8 R:202,44
1		0,073283	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
2	0,073283		0,510295	1,000000	0,244274	1,000000	1,000000	1,000000
3	1,000000	0,510295		1,000000	1,000000	1,000000	1,000000	1,000000
4	1,000000	1,000000	1,000000		1,000000	1,000000	1,000000	1,000000
5	1,000000	0,244274	1,000000	1,000000		1,000000	1,000000	1,000000
6	1,000000	1,000000	1,000000	1,000000	1,000000		1,000000	1,000000
7	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000		1,000000
8	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	

Рис. 32. Результаты множественного сравнения средних рангов

Таким образом, окончательный результат дисперсионного анализа: есть значимое различие в оценке своего здоровья респондентами г. Томска и г. Северска - респонденты г. Томска хуже оценивают свое здоровье, чем респонденты г. Северска. Различие в оценках своего здоровья респондентами других населенных пунктов, как между собой, так и в сравнении с г. Томск и г. Северск статистически незначимо.

Пример 3. Используя двухфакторный дисперсионный анализ, установить значимость совместного влияния таких факторов, как пол и место проживания респондента на уровень заявленных в ходе анкетирования хронических невралгических (в том числе слух, зрение) заболеваний.

В примере 1 был проведен однофакторный дисперсионный анализ, согласно которому была установлено различие заявленного уровня некоторых хронических заболеваний (в том числе невралгических) в различных населенных пунктах. Аналогичный однофакторный анализ можно было бы провести, чтобы выяснить различаются ли уровни заявленных хронических заболеваний в зависимости от пола респондентов.

Можно провести анализ влияния одновременно двух факторов (места проживания и пола) на уровень заболеваний без учета взаимодействия факторов. Такой факторный анализ является частным случаем многофакторного дисперсионного анализа и называется дисперсионным анализом главных эффектов (Main effects ANOVA).

Классический же многомерный анализ в отличии от анализа главных эффектов предполагает, кроме того, анализ эффектов взаимодействия факторов.

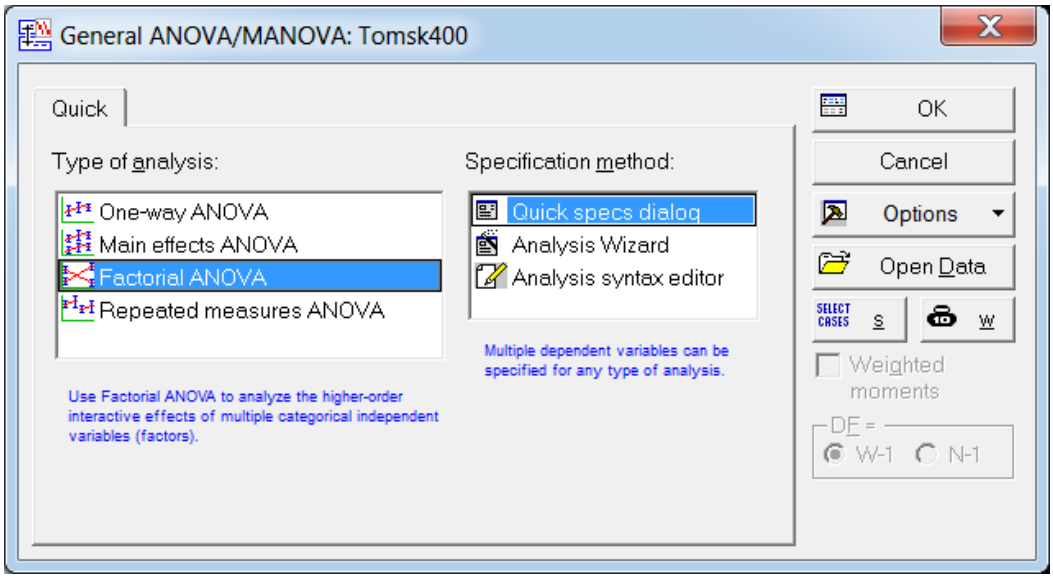


Рис. 33. Выбор метода дисперсионного анализа

Для проведения двухфакторного анализа запускаем в головном меню модуль «Statistics» и в стартовой панели выбираем пункт «ANOVA». В появившемся окне (рис. 33) выбираем тип анализа («Factorial ANOVA» - многофакторный дисперсионный анализ) и задаем метод («Quick specs dialog - диалог быстрых спецификаций»). После нажатия на «OK», попадаем в окно выбора переменных для анализа (рис. 34).

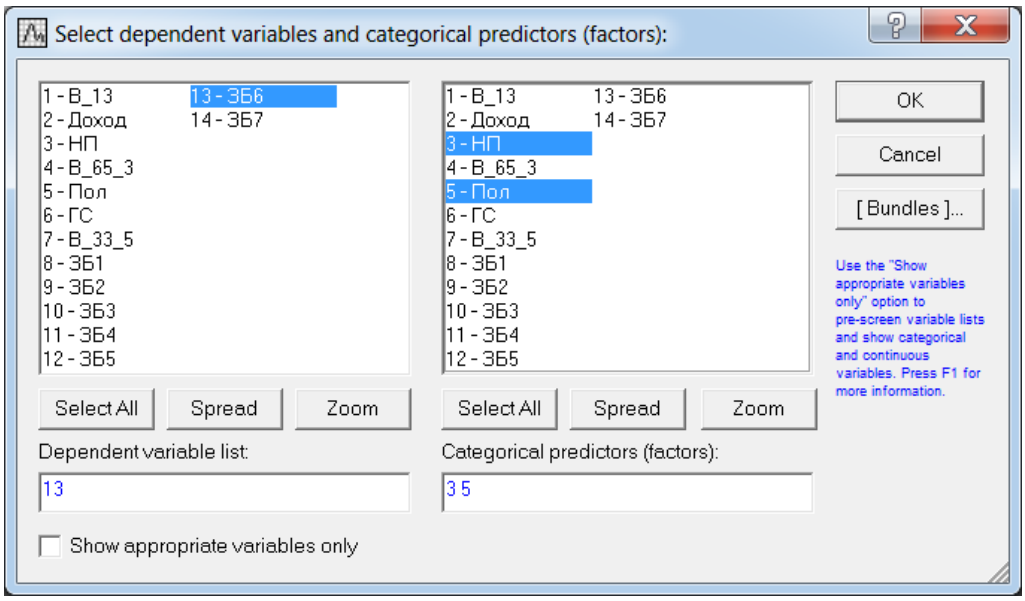


Рис. 34. Выбор переменных для дисперсионного анализа

Выбираем в качестве зависимой переменной переменную «ЗБ6» (которая содержит коды «1» и «0», соответствующие наличию или отсутствию заболевания), а в качестве группирующих переменных (факторов) - переменные «НП» и «Пол». Можно также выбрать уровни (коды) группирующих переменных, по которым будет проводиться анализ. Если коды не задавать, анализ будет проводиться по всем уровням группирующих переменных. После нажатия на клавишу «OK» переходим в окно результатов дисперсионного анализа – «ANOVA Results 1» и выбираем вкладку «Summary» (рис. 15).

Для просмотра описательной статистики на вкладке «Summary» следует выбрать «Cell statistics». Для просмотра результатов дисперсионного анализа выбираем «Univariate results», в результате получаем таблицу, изображенную на рис. 35.

Univariate Results for Each DV (Tomsk400)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Degr. of Freedom	ЗБ6 SS	ЗБ6 MS	ЗБ6 F	ЗБ6 p
Intercept	1	26,08174	26,08174	129,5018	0,000000
НП	7	5,74254	0,82036	4,0733	0,000251
Пол	1	0,36225	0,36225	1,7986	0,180670
НП*Пол	7	3,13988	0,44855	2,2272	0,031414
Error	384	77,33780	0,20140		
Total	399	88,77750			

Рис. 35. Результаты многофакторного дисперсионного анализа

Первую строку таблицы (эффект «Intercept») можно проигнорировать. Во второй и третьих строках таблицы приводятся эффекты факторов «НП» и «Пол» - суммы квадратов отклонений (SS), средние суммы квадратов отклонений (MS) с указанием значения статистики Фишера F и наблюдаемого уровня значимости. В четвертой строке таблицы приводится эффект взаимодействия факторов «НП» и «Пол», также с указанием значения статистики Фишера F и наблюдаемого уровня значимости. В пятой строке таблицы приводятся суммы квадратов отклонений (SS), средние суммы квадратов отклонений (MS) для остатков или внутригруппового разброса. В последней строке указана полная сумма квадратов отклонений.

Как видим из таблицы результатов дисперсионного анализа, значимыми эффектами является эффект фактора «НП» и эффект взаимодействия факторов «НП» и «Пол», при этом эффект фактора «Пол» не является значимым.

Для построения графиков средних разных эффектов на вкладке «Summary» нажимаем на кнопку «All effects/Graphs» и в появившемся окне выбираем эффект, для которого будут построены графики средних с доверительными интервалами. На рис. 36 приведен график средних для эффекта «НП», а на рис. 37 графики средних для эффекта взаимодействия факторов «НП» и «Пол».

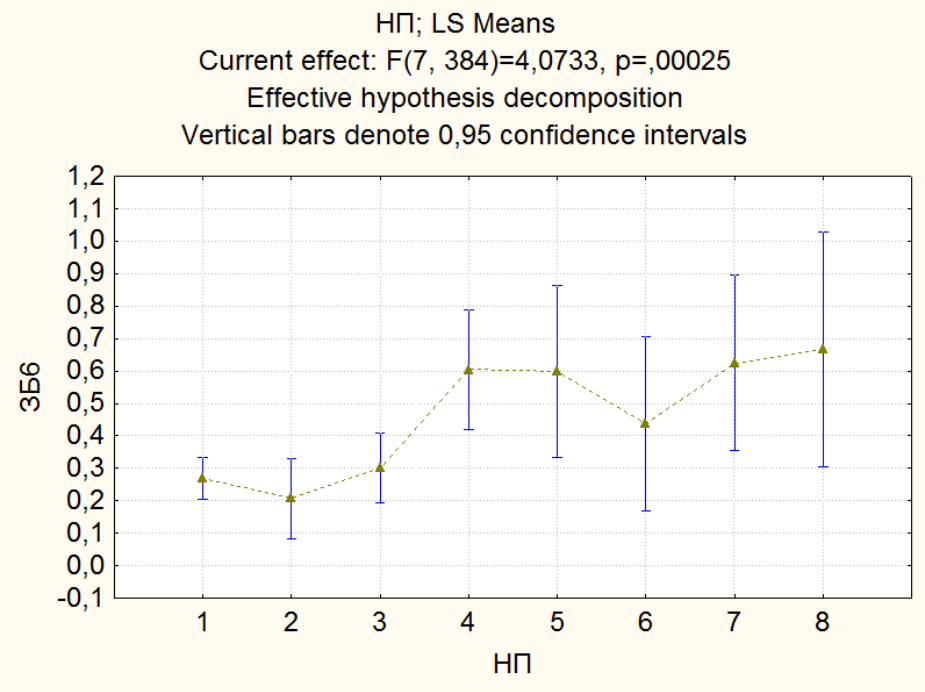


Рис. 36. График средних для эффекта «НП»

Сравнивая графики, можно сделать вывод, что наблюдаемое различие уровня заболеваний для населенных пунктов 1 и 4, 2 и 4, 3 и 4 обусловлено в первую очередь, различием уровня заболеваний для женщин данных населенных пунктов. Для мужчин же, судя по графикам, уровень заболеваний для данных населенных пунктов вряд ли значимо различается.

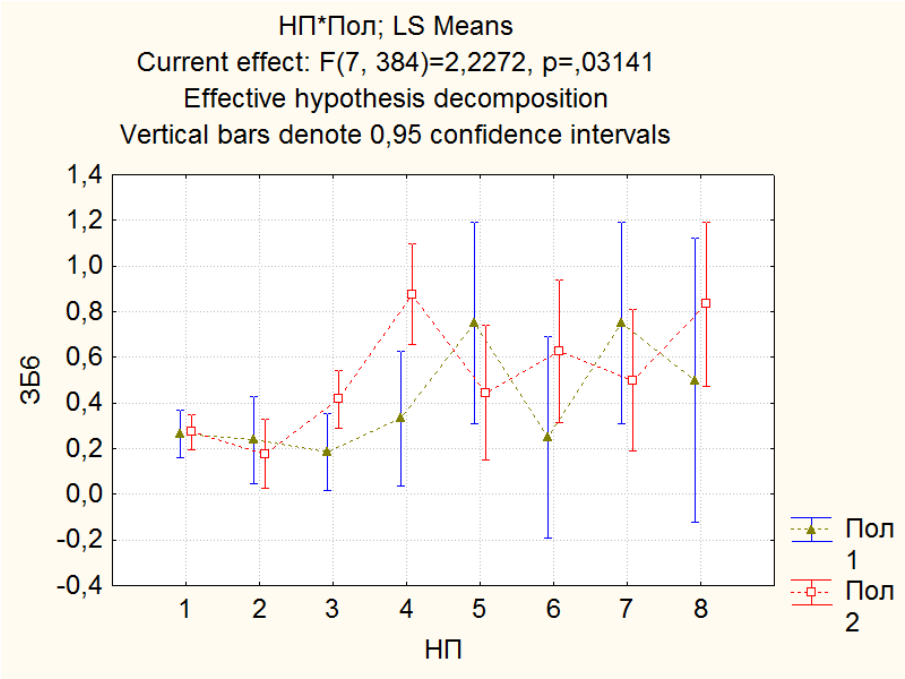


Рис. 37. Графики средних для эффектов «НП*Пол»

Для выявления значимо различающихся средних эффекта взаимодействия используем метод множественных сравнений.

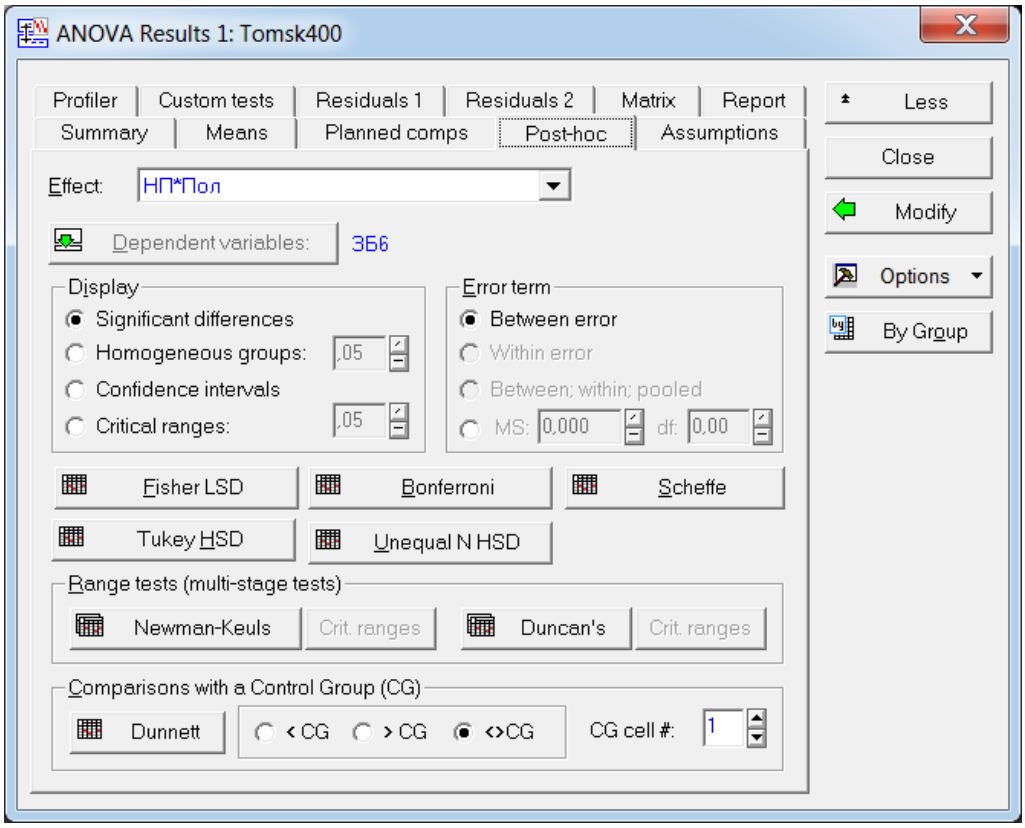


Рис. 38. Выбор метода множественных сравнений для эффекта взаимодействия «НП*Пол»

Для этого в модуле результатов дисперсионного анализа – «ANOVA Results 1», путем нажатия кнопки «More results», выбираем расширенный режим, переходим на вкладку апостериорных сравнений средних «Post-hoc», выбираем эффект «НП*Пол» и выбираем один из методов множественного сравнения, например, средний по консервативности метод HSD Тьюки (рис. 38). Для режима отображения (параметр «Display») устанавливаем «Significant differences» (значимые разности).В результате получим таблицу уровней значимости попарных различий средних для всех комбинаций уровней факторов «НП» и «Пол», часть которой приведена на рис. 39.

Tukey HSD test; variable ЗБ6 (Tomsk400)
Approximate Probabilities for Post Hoc Tests
Error: Between MS = ,20140, df = 384,00

Cell No.	НП	Пол	{1} ,26389	{2} ,27344	{3} ,23810	{4} ,17647	{5} ,18519	{6} ,41667	{7} ,33333	{8} ,87500	{9} ,75000
1	1	1		1,000000	1,000000	0,999916	0,999993	0,907044	1,000000	0,000119	0,759939
2	1	2	1,000000		1,000000	0,999250	0,999924	0,882907	1,000000	0,000075	0,771039
3	2	1	1,000000	1,000000		1,000000	1,000000	0,980361	1,000000	0,002029	0,771374
4	2	2	0,999916	0,999250	1,000000		1,000000	0,556914	0,999920	0,000059	0,533968
5	3	1	0,999993	0,999924	1,000000	1,000000		0,736277	0,999973	0,000148	0,586474
6	3	2	0,907044	0,882907	0,980361	0,556914	0,736277		1,000000	0,034700	0,989397
7	4	1	1,000000	1,000000	1,000000	0,999920	0,999973	1,000000		0,216368	0,977245
8	4	2	0,000119	0,000075	0,002029	0,000059	0,000148	0,034700	0,216368		1,000000
9	5	1	0,759939	0,771039	0,771374	0,533968	0,586474	0,989397	0,977245	1,000000	
10	5	2	0,999095	0,999359	0,998935	0,969953	0,982669	1,000000	1,000000	0,621864	0,999140

Рис. 39. Уровни значимости для попарных различий средних для всех комбинаций уровней факторов «НП» и «Пол»

Из таблицы видно, что значимое различие средних (заявленных частот заболеваний) существует между женщинами, проживающими в г. Асино и респондентами обеих полов, проживающих в г. Северске, в г. Томске и Томском районе.

Tukey HSD test; variable ЗБ6 (Tomsk400)
Homogenous Groups, alpha = ,10000 (Non-Exhaustive Search)
Error: Between MS = ,20140, df = 384,00

Cell No.	НП	Пол	ЗБ6 Mean	1	2	3
4	2	2	0,176471	****		
5	3	1	0,185185	****		
3	2	1	0,238095	****	****	
11	6	1	0,250000	****	****	****
1	1	1	0,263889	****	****	
2	1	2	0,273437	****	****	
7	4	1	0,333333	****	****	****
6	3	2	0,416667	****	****	
10	5	2	0,444444	****	****	****
15	8	1	0,500000	****	****	****
14	7	2	0,500000	****	****	****
12	6	2	0,625000	****	****	****
13	7	1	0,750000	****	****	****
9	5	1	0,750000	****	****	****
16	8	2	0,833333		****	****
8	4	2	0,875000			****

Рис. 40. Однородные кластеры групп в соответствии с выбранным критерием множественного сравнения (HSD Тьюки) и заданным уровнем значимости

Можно также, как это было сделано в примере 1, выделить однородные группы, статистически не различающиеся по уровню заболеваний. На вкладке

«Post-hoc» для режима отображения (параметр «Display») устанавливаем значение «Homogeneous groups» (однородные группы). Задаем уровень значимости, например, $p = 0{,}1$ (чем больше уровень, тем более близкие группы будут выделены) и выбираем вновь критерий множественного сравнения HSD Тьюки. В результате получаем однородные кластеры групп, расположенные в порядке возрастания средних значений (рис. 40).

Как видим для данного значения уровня значимости, на основе критерия Тьюки, можно выделить три однородные группы, содержащие сочетания факторов в соответствии с таблицей на рис. 39. Заметим, что для некоторых населенных пунктов мужчины и женщины могут быть отнесены к разным группам однородности.

Рекомендуемая литература:

1. Клячкин, В. Н. Статистические методы анализа данных : учебное пособие / В. Н. Клячкин, Ю. Е. Кувайскова, В. А. Алексеева .— Москва : Финансы и статистика, 2021 .— 240 с. : ил. — ISBN 978-5-00184-057-2. https://biblioclub.ru/index.php?page=book&id=683694 (Электронное издание).

2. Неделько, В. М. Основы статистических методов машинного обучения : учебное пособие / В. М. Неделько .— Основы статистических методов машинного обучения. Новосибирск : Новосибирский государственный технический университет, 2010 .— 72 с. —. ISBN 978-5-7782-1385-2 (Электронное издание).

3. Пролубников, А. В. Математические методы распознавания образов : учебное пособие / А. В. Пролубников .— Математические методы распознавания образов.— Омск : Издательство Омского государственного университета, 2020 .— 110 с. — ISBN 978-5-7779-2461-2 (Электронное издание).

4. Hastie, T. The elements of statistical learning : data mining, inference, and prediction / T Hastie, R. Tibshirani, J. Friedman, 2009 https://hastie.su.domains/Papers/ESLII.pdf (Электронное издание).